Sélection de modèle pour la classification non supervisée. Choix du nombre de classes.

Jean-Patrick Baudry

N° d'ordre: 9673

THÈSE

Présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS-SUD XI

Spécialité : Mathématiques

par

Jean-Patrick BAUDRY

Sélection de Modèle pour la Classification Non Supervisée. Choix du Nombre de Classes.

Soutenue le 3 décembre 2009 devant la Commission d'examen:

M.	Gérard BIAU	(Rapporteur)
M.	Gilles CELEUX	(Directeur de thèse)
M.	Gérard GOVAERT	(Examinateur)
M.	Christian HENNIG	(Rapporteur)
M.	Jean-Michel MARIN	(Examinateur)
M.	Pascal MASSART	(Président du jury)

Thèse préparée au
Département de Mathématiques d'Orsay
Laboratoire de Mathématiques (UMR 8628), Bât. 425
Université Paris-Sud 11
91 405 Orsay CEDEX

Remerciements

Mes premiers remerciements s'adressent naturellement à mon directeur de thèse : vous êtes à l'origine de mon intérêt pour les sujets traités dans cette thèse ; d'une disponibilité sans faille, vous avez dirigé, encadré, accompagné, chaque étape de ces quelques années de découverte de la recherche. Vous m'avez appris, outre bien des connaissances et des idées en statistiques, ce qu'est la collaboration scientifique, et elle a quelque chose d'enthousiasmant. Merci Gilles. Vivement notre prochain déjeuner : j'ai besoin de conseils pour choisir une pièce de théâtre !

J'ai eu la chance aussi de profiter de l'encadrement de Jean-Michel Marin lors de ses dernières années parisiennes : pour tout ce que cette thèse te doit ; pour le dynamisme et la richesse que tu as apportées aux nombreuses discussions qui en ont ponctué les débuts ; pour ton soutien ; merci. Je te suis de plus reconnaissant d'avoir accepté de participer au jury de ma soutenance.

Je remercie Pascal Massart, pour tous les repères que tu m'as donnés. Repères pour trouver la thèse et le directeur qu'il me fallait. Repères mathématiques, par les réunions qui ont constitué la trame des résultats théoriques de ma thèse. Et enfin, repères géographiques, en m'évitant de perdre la moitié du labo, en deux tentatives, dans le parc de la Gatineau (Canada) puis dans les environ de Fréjus ! Un grand merci pour avoir accepté de présider mon jury de thèse.

Je suis reconnaissant à Gérard Biau et Christian Hennig de l'honneur qu'ils m'ont fait en acceptant d'être rapporteurs de mon manuscrit et de participer au jury de ma soutenance. J'exprime toute ma gratitude à Gérard Govaert pour sa participation à mon jury de thèse.

Merci à Bertrand (un verre nous attend toujours, dans un bar, quelque part...) et Cathy (ma « sœur de thèse » !) pour vos conseils précieux et toute votre aide, notamment lors de la relecture de certaines parties de ce rapport. C'est un plaisir de collaborer avec vous : la pente est raide mais nous en voyons le bout !

Je dois beaucoup au milieu de statisticiens très riche que j'ai pu côtoyer avant et pendant de ma thèse : à l'ENS (merci notamment à Yannick Baraud, Patricia Reynaud-Bouret), à Orsay et au cours de congrès et séminaires (je tiens à saluer ici notamment Avner Bar-Hen, Raphael Gottardo, Jean-Michel Poggi, Adrian Raftery). Je me réjouis de découvrir cette année le laboratoire MAP5 et son équipe, qui me fait un accueil riche et sympathique (merci en particulier pour leurs conseils précieux à Antoine Chambaz, Servane Gey, Adeline Samson).

J'ai pu découvrir, en tant que moniteur à Orsay, l'enseignement des statistiques dans des conditions idéales, notamment sous l'encadrement de Marie-Anne Poursat.

Je poursuis avec plaisir cette découverte cette année comme ATER à l'IUT de Paris, où je suis ravi de travailler avec une équipe sympathique, dynamique et accueillante (merci notamment pour leur accueil et pour leur aide à Guillaume Bordry, François-Xavier Jollois, Mohamed Mellouk, Marc Métivier, Florence Muri-Majoube, Élisabeth Ottenwaelter).

J'ai dû apprendre à rendre les bons documents au bon moment... et si ma formation « administrative » a progressé, c'est grâce à la gentillesse et la patience des personnels administratifs des différents services d'Orsay et de Paris. Je pense notamment à Valérie Lavigne et Katia Evrat à Orsay, et à Clarisse Pantin de la Guere et Sylvie Viguier à l'IUT.

Mes amitiés à tous les doctorants (anciens ou acutels) dont la fréquentation dans les laboratoires, les séminaires et les congrès a été enrichissante, scientifiquement et humainement : Alain, Baba, Camille, Claire, Cyprien, Delphine, Fanny, Fred, Jairo, Jonas, Ken, Mahendra, Marianne (je t'écris, hein ? !), Mohamed, Nathalie (ma collègue !), Nicolas, Pierre, Robin, Sylvain, Vincent... Salut aussi aux autres occupants du bureau des doctorants au MAP5, qui y créent une ambiance si particulière : Arno, Benjamin, Djenaba, Émeline, Maël, Mélina, Sandra, Stefan...

Si j'écris un jour une autre thèse, je connais un endroit douillet où les conditions sont les meilleures pour passer un mois d'août studieux. En fait, les conditions y sont formidables pour passer toute une jeunesse ! Ce n'est évidemment pas seulement pour votre aide et votre soutien inébranlables, irremplaçables, au cours de ma thèse et de toutes mes études, que je vous suis reconnaissant, Maman, Papa, c'est aussi pour m'avoir offert la chance de pouvoir faire, chaque jour, ce que je veux.

J'ai été fier de soutenir ma thèse devant mes deux grands-pères. Merci Nonno, pour tout ce que tu m'as appris, pour tes encouragements : tu vois, j'ai « tapé dans la butte » ! Merci, Papy, de ton soutien et de ta gentillesse inestimable. Je pense bien sûr avec tendresse à mes grand-mères.

J'ai la chance d'avoir pour petit frère et pour ami le meilleur psy du monde. Merci de ne jamais être bien loin.

Merci pour tout Émilie, notamment pour ta présence et ton aide lors de ma soutenance.

Merci aussi à tous mes autres amis, que je ne peux pas tous citer et qui ont de plus belles choses à lire que ces lignes.

Avertissement.

Certaines pages en couleurs de ce document sont en noir et blanc dans cette version imprimée. La compréhension n'en devrait pas être perturbée. Le lecteur intéressé pourra toutefois consulter la version électronique, disponible sur la page Internet de l'auteur.

Foreword.

Some of the colored pages of this document have been printed in black and white. This should not be harmful to the understading. The interested reader may however refer to the electronic version, available on the author's Web page.

Contents

Présentation de la thèse

Classification Non Supervisée par Modèles de Mélange. L'essentiel du travail présenté porte sur la classification non supervisée. Étant donné $x_1, \ldots, x_n$ dans $\mathbb{R}^d$, réalisation d'un échantillon $X_1, \ldots, X_n$ i.i.d. de loi inconnue de densité $f^{\wp}$ par rapport à la mesure de Lebesgue, le problème est de répartir ces observations en classes. La particularité de la classification non supervisée est qu'aucune information sur la classe à laquelle appartient chaque observation n'est disponible a priori. Cela la distingue, par les objectifs à atteindre et les méthodes employées, de la classification supervisée qui consiste à étudier la structure de classes à partir d'observations dont l'appartenance à chacune des classes en question est connue. L'objectif est typiquement de comprendre la structure des classes afin de former une règle pour prédire la classe d'une nouvelle observation non étiquetée. Dans un cadre non supervisé, la structure des classes n'est pas nécessairement supposée préexister à l'étude. Elle est inhérente à la méthode employée : dans le cadre des modèles de mélange présenté ci-après, elle dépend des modèles — notamment de la loi des composantes des mélanges — et de la méthode d'estimation retenue. Cette absence de solution *objective* et indépendante des choix de modélisation rend délicates l'évaluation et la comparaison de méthodes. C'est certainement pour bonne part la raison pour laquelle la théorie de la classification supervisée est plus étudiée et mieux connue d'un point de vue mathématique que celle de la classification non supervisée. De nombreuses applications appellent en effet des solutions dans le cadre de cette dernière, et des méthodes pratiques ont été développées qui combinent des approches géométriques, informatiques, statistiques. La compréhension des notions sous-jacentes à ces méthodes ainsi qu'une contribution au développement de la théorie de certaines d'entre elles, sont des enjeux motivant en partie l'essentiel des travaux de cette thèse. Un autre objectif — indissociable du premier — est de contribuer au développement de telles méthodes et de solutions pour leur mise en pratique.

L'approche statistique pour la classification non supervisée la plus répandue et la plus étudiée repose sur les *modèles de mélanges*. On s'intéresse essentiellement aux mélanges gaussiens. La densité gaussienne de paramètres $\omega = (\mu, \Sigma)$ est notée $\phi(\,.\,;\omega)$. Soit $K \in \mathbb{N}^*$. L'ensemble des *lois de mélange* à K composantes gaussiennes forme le modèle

$$\mathcal{M}_K = \Bigg\{ f(\,.\,;\theta) = \sum_{k=1}^{K} \pi_k \phi(\,.\,;\omega_k) : \theta = (\pi_1, \ldots, \pi_K, \omega_1, \ldots, \omega_k) / \sum_{k=1}^{K} \pi_k = 1, \forall \omega_1, \ldots, \omega_K \Bigg\}.$$

L'approche qui semble la plus courante consiste à supposer que la loi de l'échantillon est gaussienne — bien approchée par une loi gaussienne — conditionnellement à la classe

de chaque variable X_i. La méthode correspondante est dans un premier temps d'estimer au mieux la loi de l'échantillon sous la forme d'un mélange. C'est en effet possible car, sous des hypothèses raisonnables garantissant l'identifiabilité, estimer la loi de chaque composante équivaut à estimer la loi du mélange. L'estimateur du maximum de vraisemblance $\widehat{\theta}^{\mathrm{MLE}}$ est un candidat naturel et intéressant. Elle consiste ensuite à s'appuyer sur la connaissance obtenue de la loi de l'échantillon pour en déduire la structure des classes et finalement la classification des observations. Cette dernière tâche est habituellement accomplie par la règle du *maximum a posteriori* (MAP) : la probabilité conditionnelle d'appartenance à la classe k de chaque observation x_i sous la loi définie par $\widehat{\theta}^{\mathrm{MLE}}$ est

$$\tau_{ik} = \frac{\pi_k \phi(x_i; \widehat{\omega}_k^{\mathrm{MLE}})}{f(x_i; \widehat{\theta}^{\mathrm{MLE}})}.$$

La règle de classification du MAP, en notant $\widehat{z}_i^{\mathrm{MAP}}(\widehat{\theta}^{\mathrm{MLE}})$ le label estimé de x_i, est définie par

$$\widehat{z}_i^{\mathrm{MAP}}(\widehat{\theta}^{\mathrm{MLE}}) = \operatorname*{argmax}_{k \in \{1,\dots,K\}} \tau_{ik}(\widehat{\theta}^{\mathrm{MLE}}).$$

Remarquons qu'une classe est alors identifiée à chaque composante gaussienne.

Dans le chapitre 1, ces notions sont définies précisément et discutées. La section 1.2.2 présente l'algorithme EM, qui a rendu possible l'estimation par maximum de vraisemblance dans le cadre des modèles de mélange.

Choix du nombre de classes : critères classiques et ICL. Il est même courant en classification non supervisée de ne pas connaître a priori le nombre de classes à former et donc le nombre de composantes du modèle à ajuster. Les approches pour le choisir font intervenir différentes notions de *classe*, *composante* et *cluster* : le point de vue adopté pour chacune de ces notions dans cette thèse est précisé en introduction du chapitre 2. Une méthode relativement simple et populaire de sélection de modèle consiste à minimiser un critère de la forme vraisemblance pénalisée. Notons $\widehat{\theta}_K^{\mathrm{MLE}}$ le maximum de vraisemblance dans le modèle $\mathcal{M}_K$ et D_K le nombre de paramètres libres dans ce modèle. Les critères les plus courants sont le critère AIC :

$$\hat{K}^{\mathrm{AIC}} = \operatorname*{argmin}_K \Big\{ -\log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}}) + D_K \Big\},$$

connu pour être asymptotiquement efficace[1] dans certains cadres de sélection de modèle — par exemple en régression —, et le critère BIC :

$$\hat{K}^{\mathrm{BIC}} = \operatorname*{argmin}_K \Big\{ -\log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}}) + \frac{\log n}{2} D_K \Big\},$$

connu pour être consistant, notamment pour les modèles de mélange sous des hypothèses de régularité : $\hat{K}^{\mathrm{BIC}}$ converge vers le plus petit K tel que le modèle $\mathcal{M}_K$ minimise la distance à $f^{\wp}$, au sens de la divergence de Kullback-Leibler. Il se trouve que AIC sous-pénalise manifestement les modèles dans le cadre des modèles de mélange et sélectionne souvent un nombre exagérément grand de composantes, au moins pour les

[1] Une procédure de sélection de modèle $\widehat{K}$ est dite efficace si elle se comporte presque aussi bien que l'*oracle*. Voir la section 3.1.2 pour une définition précise de cette notion.

tailles d'échantillons — raisonnables — que nous avons considérées. BIC se comporte à la hauteur de ce que prévoit la théorie et sélectionne souvent un modèle permettant de bien approcher la loi de l'échantillon. Cependant, cet objectif est discutable : ce faisant, BIC donne parfois lieu à un nombre de classes plus grand que le nombre jugé pertinent. Cette situation se présente notamment lorsque certaines des composantes gaussiennes de $\widehat{\theta}^{\mathrm{MLE}}_{K^{\mathrm{BIC}}}$ sont proches ou se recouvrent sensiblement : cela peut traduire la présence de groupes de données que l'on aurait souhaité voir identifiés comme des classes mais dont la distribution conditionnelle n'est pas gaussienne. Plusieurs composantes gaussiennes peuvent alors être nécessaires pour approcher correctement un tel groupe de données. Ici intervient la notion de *cluster* : la notion de composante gaussienne rend rarement compte intégralement de l'objectif de classification. Une notion intuitive de classe comporte l'idée de classes *compactes* et bien *séparées* les unes des autres. Il est nécessaire de préciser cette notion. Le critère ICL (Biernacki et al., 2000) semble choisir un nombre de classes satisfaisant un compromis intéressant entre les notions de composante gaussienne et de cluster. Il repose sur un terme dit d'*entropie* :

$$\mathrm{ENT}(\theta) = -\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}(\theta)\log\tau_{ik}(\theta),$$

qui est une mesure de la confiance que l'on peut accorder à la classification obtenue par MAP sous θ. ICL est alors défini par

$$\hat{K}^{\mathrm{ICL}} = \operatorname*{argmin}_{K}\Big\{-\log \mathrm{L}(\widehat{\theta}^{\mathrm{MLE}}_{K}) + \mathrm{ENT}(\widehat{\theta}^{\mathrm{MLE}}_{K}) + \frac{\log n}{2} D_K\Big\}.$$

Étudier d'un point de vue théorique ce critère a été le principal objectif de cette thèse. L'un des enjeux en est une meilleure compréhension de la notion de classe sous-jacente : puisque les résultats obtenus avec ICL semblent rencontrer en pratique une notion intéressante de classe, il est intéressant de tenter de découvrir la notion « théorique » correspondante.

Voir la première section du chapitre 2 pour des rappels plus complets sur ces trois critères pénalisés et les raisonnements qui y ont mené.

Étude théorique d'ICL : minimisation de contraste pour la classification non supervisée. L'étude des propriétés théoriques d'ICL peut difficilement être abordée de façon analogue à l'étude des critères AIC ou BIC. En effet, il n'y a pas de lien manifeste entre le maximum de vraisemblance et la valeur du terme d'entropie. Un indice que l'approche usuelle n'est pas satisfaisante pour étudier ICL est que ce critère n'est pas consistant, au sens où BIC l'est : même asymptotiquement, $\hat{K}^{\mathrm{ICL}}$ n'a aucune raison d'être égal au nombre de composantes pour lequel la distance à $f^{\wp}$ est minimale, dès lors que les composantes ne sont pas bien séparées. Le cadre théorique général qui a permis de mieux comprendre le critère ICL et d'obtenir des résultats théoriques est celui de la minimisation de contraste. Il est rappelé en détail dans la section 3.1 et appliqué au problème de la classification non supervisée au chapitre 4. On y montre en effet que choisir pour contraste à minimiser (l'opposé de) la log-vraisemblance classifiante conditionnelle

$$-\log \mathrm{L}_{\mathrm{cc}}(\theta) = -\log \mathrm{L}(\theta) + \mathrm{ENT}(\theta)$$

est un objectif intéressant pour la classification non supervisée, correspondant à une notion de classe combinant les notions de classe gaussienne — c'est-à-dire de forme plus ou moins ellispoïde — et de cluster. Un nouvel estimateur peut alors être défini pour chaque modèle $\mathcal{M}_K$ par analogie au maximum de vraisemblance : le maximum de vraisemblance classifiante conditionnelle (voir section 4.3)

$$\widehat{\theta}^{\mathrm{MLccE}} \in \operatorname*{argmax}_{\theta} \log \mathrm{L}_{\mathrm{cc}}(\theta).$$

Cet estimateur ne vise pas la vraie loi des données mais une loi intéressante pour la classification non supervisée. Dans le même esprit, les critères de sélection de modèle que l'on peut alors définir par analogie aux critères de la forme vraisemblance pénalisée visent à sélectionner un modèle — un nombre de composantes — permettant une classification à la fois peu hésitante et raisonnable par rapport à la loi des données. La pénalité de ces critères est étudiée et on montre que la gamme de pénalités permettant de définir des critères convergents — au sens du contraste L_{cc} — est analogue à celle des critères convergents dans le cadre de la vraisemblance habituelle. Cette étude s'appuie sur des résultats de contrôle de processus empiriques par des mesures de la complexité des familles de fonctions en termes d'entropie à crochets. Les modèles étant considérés d'un point de vue paramétrique, le calcul des entropies à crochets est assez direct. Mais l'adaptation des résultats existant pour le maximum de vraisemblance au contraste considéré nécessite des conditions qui n'ont pu être garanties qu'au prix d'hypothèses supplémentaires, en raison des propriétés de la fonction de contraste. Il est notamment nécessaire de garantir qu'aucune composante ne tend vers 0 et que le contraste reste borné. ICL peut alors être expliqué comme une approximation d'un critère analogue à BIC — et notamment, consistant — dans le cadre de ce nouveau contraste :

$$\hat{K}^{\mathrm{L_{cc}\text{-}ICL}} = \operatorname*{argmin}_{K} \Big\{ -\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}) + \frac{\log n}{2} D_K \Big\}.$$

La principale approximation, qui peut ne pas être mineure, est le remplacement de l'estimateur $\widehat{\theta}^{\mathrm{MLccE}}$ par $\widehat{\theta}^{\mathrm{MLE}}$.

La section 4.4 est consacrée à l'étude de ces critères de sélection de modèle pour la classification non supervisée. Des simulations illustrent les comportements pratiques des critères considérés.

Nous proposons dans la première section du chapitre 5 des solutions pour le calcul de l'estimateur $\widehat{\theta}^{\mathrm{MLccE}}$, qui pose des difficultés analogues, bien que plus difficiles, au calcul de $\widehat{\theta}^{\mathrm{MLE}}$ dans les modèles de mélange. Un algorithme adapté de l'algorithme EM est proposé et étudié. L'étape essentielle du choix des paramètres pour initialiser l'algorithme est également traitée : une nouvelle méthode est notamment proposée, qui s'avère intéressante pour l'initialisation de l'algorithme EM également. Des solutions sont aussi proposées pour choisir et imposer des bornes sur l'espace des paramètres.

Robustesse de la nouvelle procédure. Le chapitre 6 traite des propriétés de robustesse de la procédure L_{cc}-ICL. La notion de robustesse étudiée est celle du point d'effondrement défini et étudié notamment pour le critère BIC (dans le cadre du maximum de vraisemblance habituel) par Hennig (2004). Elle consiste à évaluer la proportion de valeurs qu'il est nécessaire d'ajouter à un échantillon pour qu'on ne puisse plus retrouver

de composante semblable à chacune de celles de la solution originale, parmi les composantes de l'estimateur obtenu avec le nouvel échantillon. C'est-à-dire essentiellement que les nouvelles observations suffisent à masquer la structure des composantes obtenues avec l'échantillon d'origine. Un cas notable d'effondrement de la solution d'origine est celui où la solution avec le nouvel échantillon comporte moins de composantes.

Une condition est obtenue, sous laquelle le point d'effondrement est minoré. Cette condition ne peut pas être comparée directement à celle de Hennig (2004) pour BIC, mais semble plus forte, ce qui suggère que $\mathrm{L_{cc}}$-ICL est moins robuste. C'est cependant faux au moins dans le cas où $\hat{K}^{\mathrm{L_{cc}\text{-}ICL}} < \hat{K}^{\mathrm{BIC}}$.

Alors que des résultats analogues semblent difficiles à obtenir directement pour ICL, à cause de l'absence de lien direct entre $\widehat{\theta}^{\mathrm{MLE}}$ et la valeur de l'entropie, ils peuvent l'être pour $\mathrm{L_{cc}}$-ICL : c'est un intérêt du travail du chapitre 4. Les résultats sont illustrés par des exemples.

Heuristique de pente. L'heuristique de pente est une méthode pratique de calibration des critères pénalisés guidée par les données. Les fondements théoriques en sont rappelés dans la section 3.2 : elle repose essentiellement d'une part sur l'existence d'une forme de pénalité $\mathrm{pen}_{\mathrm{shape}}$ telle que $\kappa_{\min}\,\mathrm{pen}_{\mathrm{shape}}$ est une *pénalité minimale*, au sens où la complexité des modèles sélectionnés avec une valeur plus faible de κ explose, mais reste raisonnable pour cette valeur $\kappa_{\min}$. Et d'autre part sur l'observation que la pénalité $2\kappa_{\min}\,\mathrm{pen}_{\mathrm{shape}}$ est efficace. L'heuristique de pente consiste alors, connaissant la forme de pénalité $\mathrm{pen}_{\mathrm{shape}}$, à estimer $\kappa_{\min}$ sur la base des données, et à en déduire le critère à considérer. Cette méthode permet d'extraire des données des informations sur des facteurs a priori inconnus du problème, typiquement la variance.

Deux approches sont présentées pour sa mise en œuvre :

- le *saut de dimension*, qui consiste plus ou moins à choisir pour $\kappa_{\min}$ la plus petite valeur de κ pour laquelle le modèle sélectionné a une complexité raisonnable ;
- l'*estimation directe de la pente*, qui utilise le fait que $\kappa_{\min}$ est aussi la pente de la relation linéaire (pour les grands K)

$$\mathrm{pen}_{\mathrm{shape}}(K) \mapsto -\max_{\theta \in \Theta_K} \frac{1}{n} \sum_{i=1}^{n} \gamma(X_i; \theta),$$

  où γ est le constraste considéré.

Cette dernière approche, bien que suggérée par les auteurs de l'heuristique de pente et du saut de dimension (voir Birgé and Massart, 2006 et Arlot and Massart, 2009), ne semble pas avoir été véritablement considérée jusqu'à présent. Les deux approches sont présentées en détail et comparées. Elles sont illustrées par l'application de l'heuristique de pente au maximum de vraisemblance habituel (section 3.3) et au nouveau contraste considéré au chapitre 4 (sections 4.4.5 et 4.4.6).

Un travail en cours avec Bertrand Michel et Cathy Maugis vise à proposer des solutions pour la mise en pratique de l'heuristique de pente par l'approche de l'estimation directe de la pente, et à développer un code Matlab pour rendre son application aisée. Ce travail est expliqué dans la section 5.2. Des simulations sont proposées en 5.2.4

pour illustrer l'usage de ce petit logiciel et une situation où le saut de dimension et l'estimation directe de la pente ne se comportent pas de la même façon.

Mélanges de mélanges. Une autre approche permettant de concilier les points de vue « classes gaussiennes » et « cluster » consiste à accepter de modéliser chaque classe elle-même par un mélange gaussien. La loi de l'échantillon est alors modélisée par un *mélange de mélanges*. Cette approche présente l'avantage de permettre de profiter des bonnes propriétés d'approximation des mélanges gaussiens pour modéliser la loi conditionnellement à chaque classe, y compris lorsque celle-ci n'est pas gaussienne, mais de former un nombre limité et pertinent de classes. Voir la section 2.2 pour une présentation générale de cette approche.

Un travail commun avec Adrian Raftery, Gilles Celeux, Kenneth Lo et Raphael Gottardo consiste à proposer une procédure dans ce cadre. Il fait l'objet d'un article à paraître. Le chapitre 7 est constitué de cet article, dont les notations ont été modifiées. Mais cet article est parallèle au travail présenté au chapitre 4 et le point de vue sous lequel y est présenté ICL, notamment, correspond au point de vue « original ».

La procédure proposée consiste dans un premier temps à choisir le nombre total de composantes gaussiennes $\hat{K}^{\mathrm{BIC}}$ par le critère BIC et à en tirer une classification à $\hat{K}^{\mathrm{BIC}}$ classes. Ensuite, ces classes sont regroupées hiérarchiquement. Les classes regroupées à chaque étape sont choisies de façon à maximiser la baisse d'entropie à chaque étape. L'ensemble de la hiérarchie peut alors être intéressante pour le scientifique. Elle permet non seulement d'obtenir des classifications correspondant à chaque nombre de classes inférieur à $\hat{K}^{\mathrm{BIC}}$, mais aussi de comprendre quelle est l'importance relative des différents regroupements, lesquels semblent absolument nécessaires et lesquels doivent être effectués prudemment. Des outils graphiques sont proposés pour aider cette lecture des résultats. Ils peuvent éventuellement servir à choisir le nombre de classes à former, lorsque l'utilisateur le souhaite. En effet, toutes les solutions obtenues reposent sur la même loi de mélange, quel que soit le nombre de classes considéré : celle du maximum de vraisemblance pour $\hat{K}^{\mathrm{BIC}}$ composantes gaussiennes. Les critères habituels ne peuvent donc pas être appliqués dans ce cadre et aucune méthode d'inférence statistique usuelle ne peut permettre de sélectionner le nombre de classes à former.

L'intérêt de la procédure est illustré par des simulations dans des situations variées et un exemple d'application à des données réelles de cytologie (sections 7.4 et 7.5).

Classification éclairée par une partition externe. Le critère ICL — et à sa suite la plupart des travaux présentés dans cette thèse (voir notamment le chapitre 4) — repose sur l'idée qu'un critère de sélection de modèle peut être spécialement choisi en fonction de l'objectif de l'étude, à l'opposé du choix qui consiste à fonder toute étude statistique sur une estimation préliminaire de la densité des données. Cette idée est appliquée dans un travail en cours avec Gilles Celeux et Ana Sousa Ferreira rapporté au chapitre 8, au problème suivant.

Supposons qu'au-delà des observations à classer, nous disposons d'une classification — dite « externe » — de ces données qui n'est pas le produit de l'étude de celles-ci, mais provient par exemple d'une variable supplémentaire. Une question intéressante est de savoir si la classification que l'on obtient par l'approche fondée sur les modèles de

mélange, sans tenir compte de la classification externe[2], peut être reliée à celle-ci. Nous proposons un critère de sélection de modèle, produit d'une heuristique analogue à celle menant à ICL, qui permet de choisir le nombre de classes en tenant compte du lien entre la classification obtenue pour chaque modèle considéré, et la classification externe. Ce lien est mesuré par un terme reposant sur la table de contingence entre les deux classifications.

Des simulations illustrent le comportement de ce critère dans des situations connues, et son intérêt est mis en évidence par l'étude d'un jeu de données réelles portant sur les motivations et possibilités d'évolution et de formation professionnelles des enseignants au Portugal.

[2]Il ne s'agit donc pas de classification supervisée.

Chapitre 1

Model-Based Clustering

Contents

In this chapter are introduced basics of the model-based clustering. Clustering — or unsupervised classification — consists of splitting data among several classes, when no label is known. Many approaches are available for clustering. Model-based clustering is a statistical fruitful approach which relies on the fitting of mixture models to the data. The estimated distribution is then used to design classes. Assume a sample consists of observations arising from several classes, but the classes are unobserved. The usual underlying idea is that each observation arises from a class, and that recovering the distribution of the whole sample (which is the mixture distribution) should enable one to recover the distribution of each class (which is a component of the mixture) and then to recover the classes themselves. In this chapter, it is assumed that the number of classes to design from the data is known. The following chapters deal with the choice of the number of classes to design when it is unknown. Model-based clustering is particularly helpful to this aim.

The most common approach to this end is the classical maximum likelihood approach, with Gaussian mixture components. This is a natural model in this framework, since it means roughly assuming the classes to have ellipsoid shapes. Gaussian mixture models, their interpretation in this framework, and the particular difficulty of their identifiability are introduced in Section 1.1. It is also explained there how they are interpreted as missing data models. In Section 1.2, after a short historical presentation of the estimation methods which were once employed in this context, the now widespread maximum likelihood estimation is discussed. The EM algorithm, which made the computation of the maximum likelihood estimator tractable with the rise of computers, is introduced, and some solutions to perform it and overcome its difficulties are discussed.

Once an estimator is computed, there still remains to design the classes. This is typically done through the Maximum A Posteriori rule, which is recalled in Section 1.3.

Finally, an alternative approach to the usual likelihood maximization is recalled in Section 1.4: the classification likelihood maximization. It is not based on the usual statistical point of view, in that it does not handle separately the estimation of the data distribution and the final aim, which is clustering. This is done by considering the "classification likelihood", which is defined there, and can be maximized through an iterative algorithm analogous to EM and called Classification EM (CEM). Moreover, the link between the usual and classification likelihoods and the "entropy", are derived in this section. Both this link and this quantity will be essential in the following.

Let us stress that only Gaussian mixture models are considered in this thesis. Other mixture distributions may be considered, notably Student mixtures to cope with outliers, or mixtures of Gaussian and uniform components (see for example McLachlan and Peel, 2000 or Hennig, 2004). Most methods presented in this thesis may presumably be adapted to such models.

References for this chapter are McLachlan and Peel (2000) and Fraley and Raftery (2002).

1.1 Gaussian Mixtures

1.1.1 Definition

Of concern here are the Gaussian mixture models. The Gaussian density function with respect to the Lebesgue measure λ on $\mathbb{R}^d$ is denoted by ϕ:

$\forall x \in \mathbb{R}^d, \forall \mu \in \mathbb{R}^d, \forall \Sigma \in \mathbb{S}_+^d$,

$$\phi(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det\Sigma}}\ \mathrm{e}^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)},$$

where $\mathbb{S}_+^d$ is the set of symmetric positive definite matrices on $\mathbb{R}^d$. Let us denote $\omega = (\mu,\Sigma)$.

The distribution of a random variable on $\mathbb{R}^d$ is a *Gaussian mixture* if and only if its density function with respect to Lebesgue measure is expressible as

$$f(x) = \int \phi(x;\omega)d\nu(\omega) \quad d\lambda - \text{a.s.},$$

where ν is a probability distribution on the parameters space: the *mixture distribution*.

We restrict attention to *finite mixtures*, i.e. mixtures such that the mixture distribution support is finite. Let us then denote $\nu = \sum_{k=1}^K \pi_k \delta_{\omega_k}$, where $\{\omega_1,\ldots,\omega_K\}$ is the support of ν, δ_ω is the Dirac measure at ω, and $\pi_k \in [0,1]$ are such that $\sum_{k=1}^K \pi_k = 1$ (the set of all K-tuples $(\pi_1,\ldots,\pi_K)$ for which this condition is fulfilled is denoted by Π_K).

The Gaussian distributions $\phi(.;\omega_k)$ are the *components* of the mixture and the π_k's the *mixing proportions*.

Generally, the set of all parameters of a finite Gaussian mixture is denoted θ (that is, θ contains all mixing proportions and all component parameters). The mixture density function is then denoted $f(\,.\,;\theta)$ (the dependency on K is often omitted in the notation):

$$\forall K \in \mathbb{N}^*, \forall \theta \in \left(\Pi_K \times \left(\mathbb{R}^d\right)^K \times \left(\mathbb{S}_+^d\right)^K\right), f(\,.\,;\theta) = \sum_{k=1}^K \pi_k \phi(\,.\,;\omega_k).$$

Let us fix the number K of components and define a *Gaussian mixture model* as generally done (see for example McLachlan and Peel, 2000). It is the set of distributions which density functions belong to[1]

$$\mathcal{M}_K = \left\{ \sum_{k=1}^K \pi_k \phi(\,.\,;\omega_k) \,\middle|\, (\pi_1,\ldots,\pi_K,\omega_1,\ldots,\omega_K) \in \Theta_K \right\}, \tag{1.1}$$

with $\Theta_K \subset \Pi_K \times \left(\mathbb{R}^d \times \mathbb{S}_+^d\right)^K$.

Those models will have to be restricted for technical reasons, which will be discussed later. But they may also be restricted for modeling necessities. Banfield and Raftery

[1] As usual, we shall identify the model (the set of distributions) and the set of corresponding densities.

(1993), and then Celeux and Govaert (1995) in an analogous way, suggested the following decomposition of the covariance matrix Σ_K of each mixture component:

$$\Sigma_k = \lambda_k D_k A_k D_k', \tag{1.2}$$

with λ_k the greatest eigenvalue of Σ_k in Banfield and Raftery (1993) and $\lambda_k = \det \Sigma_k{}^{\frac{1}{d}}$ in Celeux and Govaert (1995). We will follow this last convention. λ_k is the volume of Σ_k. $\lambda_k A_k$ is the diagonal matrix with the eigenvalues of Σ_k on its diagonal (ranked in decreasing order): it defines the component form (its iso-densities being ellipsoids which can be more or less elongated along each eigenvector direction, according to the corresponding eigenvalue). D_k is the matrix of eigenvectors of Σ_k and defines the orientation of the component. Celeux and Govaert (1995) define and study this way 28 different models. Indeed, different models are obtained by imposing — or not — the mixing proportions, volumes, shapes and/or orientations to be equal from a component to another. Other possible restrictions are $A_k = Id$ (in which case the components — and the corresponding models — are said to be *spherical*) or to impose each D_k to be a permutation matrix (the components and the corresponding models defined this way are called *diagonal*: the components are parallel to the axes). Those constraints are interesting in the model-based clustering framework since they enable to model geometrical assumptions about the shape of the classes to be designed.

Some of the functions we shall consider in the study of mixture models (see Chapter 4 and the definition of ICL Section 2.1.4 notably), require the knowledge of each component density. Therefore, they cannot be defined over the model $\mathcal{M}_K$ as is. They are rather defined over the set

$$\widetilde{\mathcal{M}}_K = \left\{ \Big(\pi_1 \phi(\,.\,;\omega_1), \ldots, \pi_K \phi(\,.\,;\omega_K) \Big) \;\Big|\; (\pi_1, \ldots, \pi_K) \in \Pi_K, \right. \\ \left. (\omega_1, \ldots, \omega_K) \in \left(\mathbb{R}^d * \mathbb{S}_+^d \right)^K \right\}.$$

This set provides more information than a model as usually defined. Not only defines an element of this set a mixture density (the same as the corresponding element in $\mathcal{M}_K$ would), it also provides each component density and the corresponding proportion (just remark that $\pi_k = \int \pi_k \phi(x;\omega_k) dx$). However, we shall consider the mixture models as parametric models and work with the parameters. Since there is actually a one-to-one correspondence between $\widetilde{\mathcal{M}}_K$ and the parameters space, there will be no difficulty in the definition of those functions. But rigorously, they are not really defined over the model $\mathcal{M}_K$, but in the particular case of an identifiable parametrization.

1.1.2 Identifiability

It is well-known that mixture models, as parametrized in (1.1), encounter identifiability problems.

Every estimation situation will be set in a fixed K framework: the identifiability difficulties we are interested in are then those which are met in this situation. Let us fix K and cite two remarkable reasons why the $\mathcal{M}_K$ mixture model is not identifiable:

- Up to $K!$ different parameters may result in the same mixture density because of the so-called "label switching" phenomenon. It refers to the fact that a mixture density stays the same if its components are permuted:

$$f\big(\,.\,;(\pi_1,\ldots,\pi_K,\mu_1,\ldots,\mu_K,\Sigma_1,\ldots,\Sigma_K)\big) \\ = f\big(\,.\,;(\pi_{\alpha(1)},\ldots,\pi_{\alpha(K)},\mu_{\alpha(1)},\ldots,\mu_{\alpha(K)},\Sigma_{\alpha(1)},\ldots,\Sigma_{\alpha(K)})\big)$$

  for any permutation α of the set $\{1,\ldots,K\}$.

- Whenever $K \geq 2$ and at least two components of the mixture are identical, an infinite number of parameters result in the same mixture distribution. If, for example, $K = 2$ and $\theta = (1, 0, \mu, \mu, \Sigma, \Sigma)$, then $f(\,.\,;\theta) = f\big(\,.\,;(\pi, 1-\pi, \mu, \mu, \Sigma, \Sigma)\big)$ for any $\pi \in [0, 1]$.

The first one of those difficulties can be fixed by restricting the parameters space. McLachlan and Peel (2000) suggest for example to define an order on the parameters, and to keep only the smallest parameter corresponding to each distribution. The second one (Aitkin and Rubin, 1985) can be fixed by imposing that each component must be different from the others ($\forall k, k', \omega_k \neq \omega_{k'}$). A K-component mixture should actually be considered as a $(K-1)$-component mixture, whenever two of its components are identical. Moreover, in the same spirit, null mixing proportions should not be allowed.

Yakowitz and Spragins (1968) define a weak notion of identifiability. According to this notion, a parametrization is identifiable if the components parameter may be recovered from the density, up to the order. It simply accepts label switching. It is sufficient for the needs of the applications in this thesis, since the quantities of interest never depend on the order of the components. This property has been reformulated in Keribin (2000) as:

$$\sum_{k=1}^{K} \pi_k \phi(\,.\,;\omega_k) = \sum_{k=1}^{K} \pi'_k \phi(\,.\,;\omega'_k) \Longleftrightarrow \sum_{k=1}^{K} \pi_k \delta_{\omega_k} = \sum_{k=1}^{K} \pi'_k \delta_{\omega'_k}.$$

This property have been proved to hold for multivariate finite Gaussian mixtures (and even for a union of the models $\mathcal{M}_\mathcal{K}$ for several values of K) by Yakowitz and Spragins (1968) under the assumptions that $\pi_k > 0$ and that the ω_k's (and the ω'_k's) are distinct from each other, as stated above. We shall assume those conditions to hold here and hereafter. Those are quite weak restrictions on the models. However, they entail difficulties to choose the involved constants as the parameter space have to be assumed compact. For instance, to avoid null proportions under the compactness assumption, a bound $\pi_{\min}$ has to be chosen so as to impose $\pi_k \geq \pi_{\min}$.

Another equivalent point of view (Redner, 1981) consists of considering the quotient topological parameter space obtained by the equivalence relation $\theta_1 \equiv \theta_2 \Leftrightarrow f(\,.\,;\theta_1) = f(\,.\,;\theta_2)$.

Finally, let us notice that to tackle the non-identifiability of mixture models and to be able to construct a likelihood test, Dacunha-Castelle and Gassiat (1999) define a parametrization called "locally conic parametrization", which allows to separate the identifiable part of the parameters from the non-identifiable one. Keribin (2000), who

notably derived the consistency of the BIC criterion in particular mixture models situations (see Section 2.1.3), rely on this parametrization. Those methods are not presented here.

But, notably in Chapter 4, where we make use of the Gaussian mixture models, the identifiability assumption will be seen not to be needed, since the natural model of our study is rather $\widetilde{\mathcal{M}}_K$: see Section 4.1.1.

1.1.3 Interpretation and Properties of the Gaussian Mixture Model

Mixture models typically model mixtures of populations. They are relevant to model populations which are supposed to be composed of several distinct subpopulations, each one of them having its own particularities (biological, physical, sociological,...). And those particularities are assumed to be reflected through the observations. Each subpopulation is modeled by one of the mixture components, which parameters are linked to its characteristics. The distribution of the whole population is then the mixture of those components, weighted by the proportions corresponding to the proportion of individuals in each subpopulation.

An interpretation of a mixture distribution is that several causes (which may typically be subpopulations) contribute to the observations according to their respective weights and properties, which are modeled by the components parameters.

The simulation of a mixture model may be interpreted in the following way — and it provides a straightforward algorithm to implement it. For each observation, a component k is first chosen among $\{1, \dots, K\}$ according to the probabilities $(\pi_1, \dots, \pi_K)$, and the observation X itself is then simulated according to the distribution of the k^{th} component.

Let us now introduce the notion of *size* of a component: it refers to the number of observations among the n-sample, which have arisen from this component. This is expected to be about $\pi_k \times n$.

Mixture models then provide a natural tool in classification or clustering frameworks.

They also can be very useful in the density estimation framework: when the number of components is allowed to be great enough, mixture models enjoy good approximation properties (See Titterington et al., 1985, Section 3.3.3, for references). McLachlan and Peel (2000) illustrate this through some examples of (univariate) distributions shapes (skewed, multimodal or not, etc.) that could obviously not be well approached by a Gaussian density (nor would it be by any classical density family) but that Gaussian mixtures enable to.

1.1.4 Mixture Models as Missing Data Models

Let us highlight that mixture models may then be naturally interpreted as missing data models (typically in a clustering framework). From this point of view, the missing data is the "label", i.e. the component k from which the observation X arose (see Section 1.1.3). Let us denote by a vector $Z \in \mathbb{R}^d$ this label. All its values are null, but the k^{th} which

value is 1[2]. This variable is often a part of the quantities of interest, if not the quantity of interest itself. But it is not observed in the clustering framework and has to be guessed through the observations X_i.

The *complete data* is then the pair (X, Z). If the density function of this pair is defined by

$$f(x, z; \theta) = \prod_{k=1}^{K} \big(\pi_k \phi(x; \omega_k)\big)^{z^k}, \tag{1.3}$$

it can easily be checked that X is marginally distributed according to the Gaussian mixture model with corresponding parameters.

When considering the (i.i.d.) sample $X_1, \ldots, X_n$, the corresponding labels $Z_1, \ldots, Z_n$ constitute an i.i.d. sample which marginal distribution is the distribution over $\{1, \ldots, K\}$ with probabilities $\pi_1, \ldots, \pi_K$.

This "missing data model" point of view is the key to the definition of the EM algorithm (Section 1.2.2), to the definition of the ICL criterion (Section 2.1.4) and to the new contrast that is proposed in Chapter 4.

Remark that mixture models also can be natural and precious tools as some or all of the labels are available (semi-supervised or supervised classification). Those frameworks are not discussed here: see for example Chapelle et al. (2006) for further material upon this topic.

1.2 Estimation in Gaussian Mixture Models

It is now assumed that a sample $X_1, \ldots, X_n$ in $\mathbb{R}^d$, from an unknown distribution $f^{\wp}$, has been observed. We shall model it through a Gaussian mixture to be chosen in model $\mathcal{M}_K$. K is fixed in this section.

The usual point of view of statistical estimation in a model is that the "true" distribution of the sample is assumed to belong to the model. But we do not wish here to assume generally that $f^{\wp} \in \mathcal{M}_K$. This would be in contradiction with the approximation point of view, and we will introduce in following chapters situations where it is clearly wrong (typically, as the number of components is too small). Remark that such an assumption would imply that K is exactly the true number of components of $f^{\wp}$ (it cannot be smaller of course, but neither can it be greater because we exclude $\pi_k = 0$ and mixtures with two identical components so as to guarantee the identifiability). However some theoretical results cited in the following assume $f^{\wp} \in \mathcal{M}_K$. When this assumption is not necessary, the target distribution is the one in model $\mathcal{M}_K$ which approximates "at best" $f^{\wp}$. In the maximum likelihood framework (Section 1.2.1), "at best" means the distribution which Kullback-Leibler divergence (see below) to the true distribution $f^{\wp}$ is the smallest and the maximum likelihood estimator is actually a candidate estimator to this distribution (Huber, 1967; White, 1982; Leroux, 1992).

Recall the definition of the *Kullback-Leibler* divergence between F and G, when F and G are two distributions absolutely continuous with respect to the same measure λ

[2]By a slight abuse of the notation, Z will often be identified with the corresponding component index k.

(taken here as the Lebesgue measure) with densities f and g:

$$d_{\mathrm{KL}}(F,G) = \int \log \frac{f}{g}\, f d\lambda.$$

Recall this is a non-negative quantity, which is null if and only if $F = G$ (i.e. $f = g$ except perhaps on a λ-zero measure set), and which measures the "distance" between F and G, although it does not verify the properties of a mathematical distance (it is notably not symmetric). It will often be written $d_{\mathrm{KL}}(f,g)$.

Lots of methods to fit a mixture model to data have been proposed. Redner and Walker (1984) and Titterington (1990) establish a long review upon this topic. The quantity of efforts that have been made to design and to perfect methods to fit mixture models in various situations illustrates the interest they have been generating to for a long time.

The first fruitful reported trials in this direction seem to go back to Pearson (Pearson, 1894), who computed some estimators of the parameters of a univariate Gaussian mixture through the moments method. It consists of resolving equations obtained by identifying moments linked to the distribution to be estimated (and then depending on its parameters) with the empirical moments computed from the observations. Pearson (1894) obtains this way for each estimator of this problem, expressions which depend on a well-chosen root of a ninth-degree polynomial. As Redner and Walker (1984) show, the methods which were developed until the rise of the computers in the 1960's were mainly inspired from this approach... It is actually affordable, at least in numerous particular cases, with low computation capabilities. It should nevertheless be thought of how difficult it might be to solve (by hand!) the equations as the dimension of the data is greater than one, for example: it is then necessary to restrict drastically the model.

The rise and progresses of computers from the 1960's and the possibility to achieve more and more complex computations allowed another method to appear and to become the most studied and employed one (Redner and Walker, 1984). The computing and the optimization of the likelihood became actually possible and they took the place of the moments method. This holds all the more since the use of the EM algorithm, which makes those computations easier (Dempster et al., 1977, and others: see Section 1.2.2). From then, mixture models became even more popular and employed. From now on, this section is to deal with maximum likelihood estimation and the EM algorithm.

However, estimating the parameters of a mixture model is still a difficult task, practically and theoretically. First of all, the usual approach of likelihood maximizing, which consists of trying to optimize the likelihood equations obtained by identifying the log likelihood first derivative with 0, does not enable to obtain explicit expressions of the estimators. Their expressions are nonlinear with respect to the parameters and much too complex to this end... Those equations then typically have to be solved by iterative algorithms. The EM algorithm presented below is an example of iterative algorithm which directly aims at maximizing the likelihood. But anyway, as Redner and Walker (1984) highlight, we should be aware of the potentially poor behavior of maximum likelihood estimators in mixture models. This is all the more true when the components of the mixture are not well separated. This is not surprising: actually, when the components are clearly separated, the problem is almost the same as fitting each component of

the mixture to the corresponding observations, independently from the others. Anyway, there is, at this time and to our knowledge, no other approach which would outperform likelihood maximization in this framework. The maximum classification likelihood estimator will be considered (see Section 1.4) but it will be seen that it answers an other objective.

1.2.1 Definition and Consistence Properties of the Maximum Likelihood Estimator

Gaussian mixture models are now usually fitted through maximum likelihood estimation. Maximum likelihood estimation, apart from heuristics and intuition, is essentially justified through asymptotic results. We shall see what particular difficulties this approach encounters in the framework of mixture models. First of all, we have to tackle the unboundedness of the likelihood at the parameter space boundary. Theoretical asymptotic results then apply, but there still remains the special case of spurious maxima to handle, as the available sample is actually finite.

The *likelihood* of the distributions in model $\mathcal{M}_K$ is

$$\forall \theta \in \Theta_K, \mathrm{L}(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \phi(x_i; \omega_k)$$

$(x_1, \ldots, x_n)$ will often be omitted in the notation and it will then be written $\mathrm{L}(\theta)$.

The maximum likelihood estimator in model $\mathcal{M}_K$ for the sample $(X_1, \ldots, X_n)$ would then at first sight be defined as

$$\hat{\theta}_K \in \underset{\theta \in \Theta_K}{\operatorname{argmax}} \mathrm{L}(\theta; X_1, \ldots, X_n).$$

But a major difficulty of Gaussian mixture models is encountered there: this maximum does not generally exist. The likelihood is actually not even always bounded, as illustrated by the

Lemma 1
$\forall x_1, \ldots, x_n \in \mathbb{R}^d, \forall K \geq 2, \forall (\pi_1, \ldots, \pi_K) \in \Pi_K$ *such that* $\pi_1 = 10^{-10^{137}}$,
$\forall \mu_2, \ldots, \mu_K \in \mathbb{R}^d, \forall \Sigma_2, \ldots, \Sigma_K \in \mathbb{S}_+^d, \forall i \in \{1, \ldots, n\}$,

$$L\big((\pi_1, \ldots, \pi_K, x_i, \mu_2, \ldots, \mu_K, \sigma^2 Id_d, \Sigma_2, \ldots, \Sigma_K); \mathbf{x}\big) \xrightarrow[\sigma^2 \to 0]{} +\infty$$

The value of π_1 *has been suggested in Le Cam (1991)!*

Proof *Let* $(\pi_1, \ldots, \pi_K) \in \Pi_K$ *such that* $\pi_1 > 0, \mu_2, \ldots, \mu_K \in \mathbb{R}^d, \Sigma_2, \ldots, \Sigma_K \in \mathbb{S}_+^d$ *and* i_0 *be fixed. Let us denote for any* $\sigma > 0$,

$\theta(\sigma) = (\pi_1, \ldots, \pi_K, x_{i_0}, \mu_2, \ldots, \mu_K, \sigma^2 Id_d, \Sigma_2, \ldots, \Sigma_K)$.

$$\begin{aligned}\log L(\theta; \mathbf{X}) &= \sum_{i \neq i_0} \log \sum_{k=1}^{K} \pi_k \phi(x_i; \omega_k) + \log \left(\pi_1 \phi(x_{i_0}; x_{i_0}, \sigma^2 Id_d) + \sum_{k=2}^{K} \pi_k \phi(x_{i_0}; \omega_k) \right) \\ &\geq \sum_{i \neq i_0} \log \sum_{k=2}^{K} \pi_k \phi(x_i; \omega_k) + \log \left(\pi_1 \phi(x_{i_0}; x_{i_0}, \sigma^2 Id_d) \right)\end{aligned}$$

The first term in this sum does not depend on σ and the second one tends to infinity as σ^2 tends to zero. ■

This unboundedness of the likelihood at the boundary of the parameter space is a real difficulty, which has been widely discussed (Redner and Walker, 1984; McLachlan and Peel, 2000, ...). It seems that two main approaches were proposed to tackle it.

But before discussing them, let us first recall why the maximum likelihood estimation is used and what properties it is expected to fulfill. Wald (1949) has proved in a general setting, under a few assumptions that we do not discuss now, but which notably include identifiability of the model (its assumption 4), and that $f^{\wp}$ belongs to the considered model (i.e. $f^{\wp} = f(\,.\,; \omega_0)$ for a certain ω_0), that:

Theorem 1 (Theorem 1 in Wald, 1949)
Let $\tilde{\Omega}$ be any closed subset of the parameter space Ω which does not contain the true parameter ω_0. Then

$$\mathbb{P}\left[\lim_{n \to \infty} \frac{\sup_{\omega \in \tilde{\Omega}} f(X_1; \omega) f(X_2; \omega) \ldots f(X_n; \omega)}{f(X_1; \omega_0) f(X_2; \omega_0) \ldots f(X_n; \omega_0)} = 0 \right] = 1.$$

(Although we employ here the notation we elsewhere reserve for Gaussian densities, Wald (1949) proved this theorem for much more general models and densities f).

This theorem shows that, under the assumptions of Wald (1949), as soon as an estimator $\hat{\omega}$ is defined such that its likelihood is greater than a constant times the likelihood of the true distribution, this estimator is consistent (this is Wald's Theorem 2), in the sense that $\hat{\omega} \longrightarrow \omega_0$ with probability 1. Maximizing the likelihood is an obvious — but often difficult — way to guarantee this property. This justifies the attempts to obtain consistent estimators through maximum likelihood.

Redner (1981) extended those results to non-identifiable models, and in particular to mixture families. His results lie on the same assumptions as Wald (1949) but the identifiability. This is why, by analogy with the mixtures situation, we denote back parameters as θ instead of ω in the identifiable case. Just think of θ as a mixture parameter, but remember those of Redner (1981) are more general. To overcome the non-identifiability, Redner (1981) defines Θ_0 as the set of parameters which distribution is the same as θ_0 ($\Theta_0 = \{\theta : f(\,.\,; \theta) = f(\,.\,; \theta_0)\}$), and obtains a result identical to Theorem 1, with $\tilde{\Theta}$ any closed subset of Θ not intersecting Θ_0. Actually, he proves this way the consistency of the maximum likelihood estimator in the quotient topological space already defined to tackle the lack of identifiability (Section 1.1.2). He applies this

to mixture families when the existence of a maximum likelihood estimator is guaranteed by the assumption that the parameter space is compact (and contains θ_0).

Now, let us remark that those results suggest that exactly maximizing the likelihood might not be necessary to find a consistent estimator. This is what allows to use likelihood methods in the mixture framework, even when the likelihood is not bounded over the parameter space. It suggests that other solutions should be affordable than assuming the parameter space to be compact. But anyway, there is still the difficulty to find a good estimate in the parameter space. Practically, likelihood maximization is not used to design a sequence of estimators converging to the true parameter: n is given by the data, and one has to choose one estimator. Maximum likelihood theoretical results give heartening tracks to build it. Let us now go back to the approaches to overcome the unboundedness of mixture models.

The one consists of restricting the model so that a global maximizer of the likelihood exists. An obvious way of doing so is to choose a compact parameter space, in which situation the results of Redner (1981) directly apply. It actually suffices to have the parameter space closed. This implies notably a difficulty concerning the mixing proportions if one chooses to exclude zero proportions and for the covariance matrices: the bound must be chosen such that the condition is true for the true corresponding parameters. Remark that it is not necessary to impose bounds on the mean parameters, since $L \to -\infty$ as $\mu_k \to \infty$ (this is proved for example for the univariate case in Redner's proof of Theorem 2.1). Hathaway (1985), in the univariate case, defines another constraint of this kind which is less restrictive: he imposes that

$$\min_{k,k'}\left(\frac{\sigma_k}{\sigma_{k'}}\right) \geq c > 0,$$

with c a constant to be chosen. This has the advantage that the corresponding defined model is scale-invariant. This choice of c is in fact the first main difficulty with this approach since it is needed that this constraint be verified by the true distribution (Redner (1981) assumes that it belongs to the model) — which is of course unknown! The second difficulty is that it is not easy to compute the maximum likelihood estimator under this constraint (see Hathaway (1986), who proposed a constrained EM algorithm adapted to this situation). But under this condition, Hathaway (1985) shows that a global maximizer of the likelihood exists and that it provides a converging estimator of the true parameter (assuming the observations arise from a Gaussian mixture). This idea may be generalized in the multivariate case. McLachlan and Peel (2000) highlights for example that a global maximum likelihood estimator exists (and is strongly consistent) if the covariance matrices of the components are imposed to be equal (this is an analogous situation as the one-dimensional case with $c = 1$): the situation encountered in Lemma 1 may actually not occur under this restriction. This constraint could even be relaxed slightly thanks to the decomposition exposed in Section 1.1.1: it suffices for example that the components covariance matrices determinants λ_k be imposed to be equal ($\forall k \neq k', \lambda_k = \lambda_{k'}$) (Biernacki et al., 2003).

Let us cite the nice approach of Lindsay (1983). He states the problem of the existence and the uniqueness (among others) of the maximum likelihood estimator in the mixtures framework, from a geometrical point of view. He does not suppose the number of components to be known and lets it to be free. By considering the convex hull of the likelihood set (i.e. $\{(f(x_1;\theta),\ldots,f(x_n;\theta));\theta \in \cup_{K=1}^{\infty}\Theta_K\}$), he transforms

the problem in a convex optimization problem. And so he proves the existence of the maximum likelihood estimator as a mixture with less components than $\tilde{n}$, if $\tilde{n}$ is the number of distinct observations in $(x_1, \ldots, x_n)$, under the assumption that the likelihood set is compact (which essentially means for us assuming $\det \Sigma_k \geq \sigma_0 > 0$ and the parameter space is closed). Working in the convex hull of the likelihood set is an interesting choice, in that it enables the resolution of some difficult aspects of the mixtures (notably the non-identifiability of the simple parametrization). It means working with densities instead of parameters.

Another approach could be based on the results of Peters and Walker (1978). They assume the same regularity conditions about the mixture densities with respect to the parameters as Chanda (1954) defined in a more general maximum likelihood framework, and assume that the Fisher information matrix at the true parameter is positive-definite. They are then able to prove (see their Appendix A) that a neighborhood of the true parameter can be chosen such that, with probability one, as $N \longrightarrow \infty$, there exists a unique solution of the likelihood equations in that neighborhood, and that this solution is a maximum likelihood estimator, and then the unique strongly consistent maximum likelihood estimator. One should then define the maximum likelihood estimator in a different way than before so as to take into account the unboundedness of the likelihood, hoping to catch this unique strongly consistent estimator:

Definition 1 *In model $\mathcal{M}_K$, the maximum likelihood estimator $\widehat{\theta}_K^{MLE}$ for the sample $X_1, \ldots, X_n$ is defined as*

$$\widehat{\theta}_K^{MLE} \in \operatorname{argmax} \left\{ L(\tilde{\theta}; \mathbf{X}) \,\middle|\, \tilde{\theta} : \tilde{\theta} \in \mathcal{O} \subset \Theta_K \text{ open and } \tilde{\theta} = \operatorname*{argmax}_{\theta \in \mathcal{O}} L(\theta; \mathbf{X}) \right\}.$$

$\widehat{\theta}_K^{MLE}$ then reaches the largest local maximum of the likelihood over Θ_K.

If the true distribution lies in the interior of Θ_K, then the conditions of Peters and Walker (1978) apply (assuming the Fisher information matrix is positive-definite).

But there still remains in any case a practical difficulty. When maximizing the likelihood for a given sample, it may occur that the obtained solution corresponds to a *spurious local maximum* . Suppose there are in the sample $X_1, \ldots, X_n$ a few observations — say n_1 — very close from each other, or which almost lie in a same subspace which dimension is smaller than d. They constitute a little cluster that could be fitted through one of the mixture components with a little covariance determinant. Then the corresponding mixture distribution might win against the local maximum near the true distribution, and the likelihood maximization be misleading... Let us highlight that observations may occur such that this is possible, in any situation (notably even when the parameter space is compact). This is a difficulty different than the first-mentioned unboundedness of the likelihood at the parameter space boundary: if the covariance matrix is "decreased", the likelihood is, too. Theoretical asymptotic results guarantee that when the number of observations increases, this spurious estimator will not be able to win against another estimator closer to the true distribution any more. As a matter of fact such clusters have small probability to occur as $n \longrightarrow \infty$. But in practice, only finite samples are available!

From a practical point of view, it will be seen that the optimization of likelihood through the EM algorithm often leads to solutions near a likelihood local maximum,

in the interior of the parameter space. And one then aims at finding the one which likelihood is the greatest. Maxima linked to the unboundedness of the likelihood may be avoided by imposing constraints on the parameters, or by verifying that the obtained solution does not contain a component with very small covariance matrix determinant. This last verification seems to be adapted also to avoid spurious maxima. But no easy systematic solution seems to be available, which would not imply the human judgment... In practice however, this is not too difficult as long as K is not exaggerated as compared with the true number of components and as n is not tiny.

1.2.2 EM

According to McLachlan and Peel (2000), a huge majority of papers dealing with mixture modeling published after 1978 use the EM algorithm. Another evidence of the impact of the EM algorithm might be the fact that Dempster et al. (1977) was, in June 2009 and according to *Google Scholar*, cited over 19000 times... even if such figures must obviously be handled with care. Its advent really made the fitting of mixtures tractable. It is an iterative algorithm to maximize the likelihood. Several authors (see the overview of Redner and Walker (1984) upon this topic) proposed as soon as in the 1960's iterative algorithms to compute the maximum likelihood estimator with mixtures, lot of which have been shown later to be particular applications of the EM algorithm (See for example Shlezinger, 1968). Peters and Walker (1978) for example consider an iterative generalized deflected gradient algorithm to solve the likelihood equations and obtain a particular case of the EM algorithm, and a generalized version of it. But it was first formulated and studied in a general setting — which is incomplete data setting — in Dempster et al. (1977). We will follow their derivation of the EM algorithm, which arose from the missing data interpretation of mixture models, on the contrary to most of the preceding derivations of particular cases of this algorithm, which were deduced from attempts to set the log likelihood derivative to 0 by iterative algorithms. The approach suggested by Dempster et al. (1977) is more general, and moreover provides an interesting theoretical framework. It enables to study the EM algorithm from a theoretical point of view and obtain results quite naturally. The fundamental property of the EM algorithm (Theorem 2 below: this is the property of monotonicity of the likelihood along the iterations of the algorithm) notably, was expected to be true but not proved (Peters and Walker, 1978 suggest it from their experience as a "conjecture" but "have been unable to prove [it]"). The application of the EM algorithm to mixture models is also widely discussed in Redner and Walker (1984). The EM algorithm is based on the observation that, considering the missing data interpretation of mixture models (see Section 1.1.4), it is much easier to maximize the likelihood of the complete data than that of the observed data. According to the density expression in (1.3), the *complete log likelihood* is

$$\log \mathrm{L_c}\big(\theta; (x_1, z_1), \ldots, (x_n, z_n)\big) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \pi_k \phi(x_i; \omega_k). \tag{1.4}$$

Actually, maximizing this complete log likelihood amounts to maximizing each component independently from the others, with respect to the observations X_i assigned to it (i.e. such that $Z_{ik} = 1$). In the situation we consider, where the components are Gaussian distributed, this is an easy task.

Of course, the labels Z_i are unobserved. So that one has to maximize actually $\mathbb{E}_{\tilde{\theta}}\left[\log L_c(\theta; \mathbf{X}, \mathbf{Z})|\mathbf{X}\right]$, with $\tilde{\theta}$ to be precised. This will involve the *conditional probabilities* of each component k:

$$\forall k, \forall x, \forall \theta \in \Theta_K, \tau_k(x;\theta) = \frac{\pi_k \phi(x;\omega_k)}{\sum_{k'=1}^{K} \pi_{k'} \phi(x;\omega_{k'})}. \tag{1.5}$$

$\tau_k(X_i;\theta)$ is the probability that the observation arose from the k^{th} component, conditionally to X_i, under the distribution defined by θ. It will also be written $\tau_{ik}(\theta)$.

Since one has to choose a parameter $\tilde{\theta}$ under which to compute the conditional expectations, Dempster et al. (1977) proposed an iterative procedure, in which $\tilde{\theta}$ is the current parameter. They named it EM from the respective names of each one of its two steps.

Let us describe the two steps of the algorithm when θ^j is the current parameter. θ^{j+1} is to be computed.

E step Compute for any $\theta \in \Theta_K$, $Q(\theta, \theta^j) = \mathbb{E}_{\theta^j}\left[\log L_c(\theta; \mathbf{X}, \mathbf{Z})|\mathbf{X}\right]$. According to (1.4), this amounts to computing each $\tau_{ik}^j = \tau_k(X_i;\theta^j)$.

M step Maximize $Q(\theta, \theta^j)$ with respect to $\theta \in \Theta_K$ to get θ^{j+1}.

There is no difficulty about the E step. The M step is much more tractable than the maximization of $\log \mathrm{L}(\theta)$ since $Q(\theta, \theta^j)$ is a weighted sum of Gaussian log-densities. In many cases (it depends on the chosen model and the constraints), a closed-form expression of the solution for the M step can be obtained: exactly as when maximizing the complete likelihood, the M step consists of maximizing each component independently, but each observation contributes to each component according to its corresponding posterior probability. The M step may be more difficult typically when the model at hand introduces dependencies between the different components parameters (for instance as the covariance matrices are imposed to be equal). See for example Biernacki et al. (2006), where the M step is detailed for all the 28 models defined by Celeux and Govaert (1995). Most of those M steps are closed-form. In case they are not, iterative procedures to approximate them are proposed.

Some theoretical as well as numerical results about the convergence properties of the solutions obtained by the EM algorithm justify the appeal for the EM algorithm. See Redner and Walker (1984) for an extensive synthesis of those results. We shall only cite (see for example the Theorem 1 in Dempster et al. (1977)) the

Theorem 2 (Fundamental Property of the EM Algorithm)
$\forall \theta, \theta' \in \Theta_K$

$$Q(\theta', \theta) \geq Q(\theta, \theta) \Longrightarrow L(\theta') \geq L(\theta),$$

with equality if and only if $Q(\theta', \theta) = Q(\theta, \theta)$ *and* $\tau_k(x;\theta) = \tau_k(x;\theta')$ *for any* k *and almost every* x.

The condition in Theorem 2 is even weaker than the one imposed in the M step (in which $Q(\theta^{j+1}, \theta^j) \geq Q(\theta, \theta^j)$ for any $\theta \in \Theta_K$). It then applies and guarantees the

monotonicity of the likelihood along the iterations of the EM algorithm, which is an appealing property.

Redner and Walker (1984) even prove a linear convergence of the sequence of estimators obtained through the EM algorithm, to the consistent maximum likelihood estimator (Theorem 4.3 in Redner and Walker (1984) in a general missing data framework and Theorem 5.2 for an application to mixtures). It mainly assumes the Fisher information matrix at the true parameter to be positive definite, and that the initialization parameter of the algorithm, say θ^0, is sufficiently near the maximum likelihood estimator $\widehat{\theta}_K^{\text{MLE}}$.

This last assumption is actually very important: the monotonicity of the likelihood is appealing, but the log likelihood is not convex. A consequence is that the algorithm can at best approach a local maximum which lies near the initialization parameter: whenever this initialization parameter is near a spurious maximum, there is no hope that the algorithm leaves its attraction to approach the greatest local maximum... This highlights the importance and the difficulty of the initialization step, i.e. of the choice of the first parameter θ^0. It is necessary to provide to the EM algorithm a good initial parameter. Several strategies were proposed to this end. The most classical and simple one consists of choosing K observations at random and to initialize the algorithm with their coordinates as means, the estimated variance (along each dimension) of the whole sample as (diagonal) covariance matrix and the proportions being equal. The EM algorithm is then run for a while, and this whole procedure is repeated several times. The parameter among those obtained which has the highest likelihood is chosen as an initial parameter. Actually, this strategy can be improved by choosing the initial means at random according to a normal distribution (for example) centered at the sample mean and with covariance matrix the empirical covariance. This introduces a greater variability of the candidate initial parameters. More elaborated strategies have been proposed (see Biernacki et al., 2003), notably:

Small_EM Called short runs of EM in Biernacki et al. (2003). Choose several random starts and run a very little number of iterations of EM each time. Choose the one of those solutions maximizing the likelihood to initialize a long run of EM. Repeat all of this procedure several times, and choose the solution maximizing the likelihood.

CEM The CEM algorithm will be introduced in Section 1.4.1. The corresponding initialization strategy consists of a few runs of CEM from random positions followed by a long run of EM initialized at the solution maximizing the complete likelihood among those obtained by CEM.

SEM The SEM algorithm, which is a Stochastic EM (Celeux and Diebolt, 1985), provides a Markov chain expected to spend much time near local maxima, and particularly near the greatest local maximum. Biernacki et al. (2003) then suggest two initialization steps involving SEM: first, a run of SEM may be followed by a run of EM initialized at the parameter obtained by computing the mean of the parameters obtained along the sequence provided by SEM. Second, a run of SEM may be followed by a run of EM initialized at the solution of the sequence provided by SEM which reaches the highest likelihood value.

The idea underlying those initialization steps is to explore the parameter space at best

and to expect finding a sensible solution to initialize EM at nearest from the optimal maximum. According to the authors, those three procedures often beat the simple random starts procedure, but none can be declared to be better than the others independently of the data and the model.

Other procedures have been proposed. Ueda and Nakano (1998) for example, proposed an annealing algorithm called DAEM (Deterministic Annealing EM). It consists of choosing a temperature $\frac{1}{\beta}$, and maximizing through EM an *energy*, which is a quantity linked to the likelihood, and is even smoother that β is close to 0. This provides a new initialization value for the next E step, to be used to maximize once more the energy through a new M step, but at a greater temperature, and so on until the reached temperature is 1 (in which case the maximized energy is the quantity which is involved in the classical EM algorithm). The hope of such a procedure is that, smoothing the quantity to be maximized, the EM algorithm has the possibility to leave the attraction of a spurious cluster, and to pass through the likelihood valley that would exist between it and a better solution. McLachlan and Peel (2000, Chapter 2) and Biernacki et al. (2003) do not recommend the use of this algorithm since it is quite long to perform and may be benefit only in some particular situations.

Finally, the main difficulty concerning the EM algorithm consists of being able to provide it a sensible initialization parameter. This is particularly a difficult task when the number of components to be fitted is greater than the true number of components, and situations are introduced subsequently where such a case is of interest. This notably happens as the slope heuristics is applied: see Section 3.3 for example. As we mainly used the MIXMOD software (Biernacki et al., 2006) to run the EM algorithm when computing maximum likelihood estimators, we mainly used the initialization strategies proposed by Biernacki et al. (2003), and which the MIXMOD software provides. Among those strategies, the small_EM was used at most, but in a little different way than the software does: we inserted the MIXMOD runs in a loop to strengthen its results and chose the best solution obtained.

1.3 The MAP Classification Rule

Recall we want to interpret the fitted components as classes: in view of the clustering task, we hope the Gaussian components shall reflect the subpopulations. This is an assumption on the shape of the classes, which are expected to look like Gaussian densities (i.e. classes should have ellipsoid shapes). The choice of the model is then an assumption on the expected shape of the classes: choosing for instance a spherical model (say $[p_k\_L\_I]$ in the notation of Celeux and Govaert, 1995) means to be looking for spherical classes with equal volumes, but perhaps different sizes, etc. From this point of view, the constraints imposed on the model are also assumptions about the classes. It is expected that the obtained mixture will be informative, perhaps bring informations about the structure of the subpopulations, and help decide which observation should be assigned to which class. This is why we introduced Gaussian mixture models. Suppose a number K of components has been chosen, and an estimator $\hat{\theta}_K$ (typically $\widehat{\theta}_K^{\mathrm{MLE}}$) in model $\mathcal{M}_K$ is available. There still remains to define a classification rule, so as to split the observations among K classes.

1.3.1 Definition

We define to this aim the usual MAP (for *Maximum A Posteriori*) rule. We have already defined in (1.5) the *conditional* probability of component k, which is conditional to the observation of X:

$$\tau_k(x;\theta) = \frac{\pi_k \phi(x;\omega_k)}{\sum_{k'=1}^{K} \pi_{k'} \phi(x;\omega_{k'})}.$$

This rule has a Bayesian flavor and $\tau_k(x;\theta)$ is sometimes called the *a posteriori* probability of component k. Remark that

$$\tau_k(x;\theta) = \frac{f(x,k;\theta)}{f(x;\theta)}.$$

$\tau_k(X;\widehat{\theta}^{\mathrm{MLE}})$ is the maximum likelihood estimator of the probability that X arose from component k, conditionally to X. It is then natural to assign to x the label $\widehat{z}^{\mathrm{MAP}}$ (which dependency on $\widehat{\theta}_K^{\mathrm{MLE}}$ is omitted in the notation when not ambiguous) maximizing this probability:

$$\widehat{z}^{\mathrm{MAP}}(\widehat{\theta}^{\mathrm{MLE}}) = \operatorname*{argmax}_{k=1,\dots,K} \tau_k(x;\widehat{\theta}^{\mathrm{MLE}})$$

(This maximum exists and is unique except perhaps on a set of measure zero).

The same can be done with any estimator $\hat{\theta}$:

$$\widehat{z}^{\mathrm{MAP}}(\hat{\theta}) = \operatorname*{argmax}_{k=1,\dots,K} \tau_k(x;\hat{\theta}).$$

From a decision-theoretic point of view, this rule may be justified as the maximum likelihood estimation of the Bayes rule: see McLachlan and Peel (2000, Section 1.15.2).

1.4 Classification Likelihood

An alternative point of view about mixture models arises from the missing data approach. The likelihood as defined in Section 1.2 is linked to the model $\mathcal{M}_K$, as defined in (1.1). This likelihood will be named *observed likelihood* when necessary. Considering mixture models as missing data models, as exposed in Section 1.1.4, we saw that an other density function is associated with the model, and then the corresponding likelihood is to be different, too. This leads to the so-called *classification likelihood* approach, which is also called *complete data likelihood* for obvious reasons.

1.4.1 Definition and Maximization through the CEM Algorithm

Denoting $(\mathbf{X},\mathbf{Z}) = ((X_1,Z_1),\dots,(X_n,Z_n))$ a complete data sample, the corresponding likelihood is :

$$\forall \theta \in \Theta_K,\ \mathrm{L_c}(\theta;(x_1,z_1),\dots,(x_n,z_n)) = \prod_{i=1}^{n}\prod_{k=1}^{K}\big(\pi_k\phi(x_i;\omega_k)\big)^{z_{ik}}. \qquad \text{((1.4) recalled)}$$

Some authors tried to take advantage of the obvious link of this likelihood with the clustering purpose (Scott and Symons, 1971; Celeux and Govaert, 1995). The labels $\mathbf{Z}$ being unobserved, the classification likelihood then has to be maximized both with respect to the parameter θ and to the labels, which are so considered themselves as parameters of the problem. This means optimizing the partitioning of the data and the parameters of each fitted component at the same time (since the assignments are deterministic here). This approach is then somewhat intermediate between model-based clustering and non-model-based approaches. As Celeux and Govaert (1992) have shown, maximizing the classification likelihood when the models at hand are spherical (the covariance matrices are Id) and the components are imposed to have equal volumes, is equivalent to the k-means approach. They have also shown (Celeux and Govaert, 1995) the equivalence of this method with several other more or less well-known inertia type criteria, according to the chosen model (among those defined in Section 1.1.1). This is interesting first since it provides a generalization of those criteria and the possibility to both unify them and extend them to different models. And it also allows to study those geometrical criteria from a statistical point of view.

Celeux and Govaert (1992) also proposed an algorithm derived from EM, which they called Classification Expectation Maximization (CEM), to compute the maximum classification likelihood estimator. It suffers from the same kind of initialization difficulties as EM, but in a perhaps less difficult fashion. Actually, it is seemingly like the k-means algorithm from this point of view: when the partition of the data (or equivalently, the labels) is chosen, the maximization is quite easy since it reduces to K Gaussian likelihood maximizations, a Gaussian component being fitted to the data assigned to it, independently from the other observations. Then it would "suffice" to try all possible partitions of the data to be sure to find the maximum classification likelihood. Obviously such an exhaustive search is definitely intractable as soon as n is not tiny.

Let us describe this CEM algorithm. Let θ^j be the current parameter. Here is how θ^{j+1} is updated:

E step Compute all $\tau_{ik}^j = \tau_k(x_i; \theta^j)$.

C step "C" stands for Classification step. Assign to each observation its most probable label, derived from the τ_{ik}^j's. With the notation already defined, this is $\widehat{z}_{ik}^{\mathrm{MAP}}(\theta^j)$.

M step Maximize $\mathrm{L_c}(\theta; \mathbf{x}, \widehat{\mathbf{z}}^{\mathrm{MAP}}(\theta^j))$ with respect to $\theta \in \Theta_K$ to get θ^{j+1}. This is equivalent to

$$\pi_k^{j+1} = \frac{\mathrm{card}(\{i : \widehat{z}_{ik}^{\mathrm{MAP}}(\theta^j) = 1\})}{n};$$

$$\omega_k^{j+1} = \operatorname*{argmax}_{\omega} \sum_{\{i:\widehat{z}_{ik}^{\mathrm{MAP}}(\theta^j)=1\}} \log \phi(x_i; \omega).$$

The difference with the EM algorithm then lies in the "C" step, which imposes a deterministic assignment of the labels, instead of weighting each one according to the conditional probability of each component. Remark that whatever the initialization, the algorithm converges to a stationary state within a finite number of iterations. This makes it easier to try a great number of initialization parameters.

1.4.2 From L to L_c

We shall stress an important algebraic relation between the observed and classification likelihoods, first presented in Hathaway (1986):

$$\begin{aligned}\forall \theta \in \Theta_K,\ \log \mathrm{L_c}(\theta) &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log \pi_k \phi(x_i;\omega_k) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log \underbrace{\frac{\pi_k \phi(x_i;\omega_k)}{\sum_{j=1}^{K}\pi_j\phi(x_i;\omega_j)}}_{\tau_k(x_i;\theta)} + \underbrace{\sum_{i=1}^{n}\overbrace{\sum_{k=1}^{K} z_{ik}}^{1} \log \sum_{j=1}^{K}\pi_j\phi(x_i;\omega_j)}_{\log \mathrm{L}(\theta)} \\ &= \log \mathrm{L}(\theta) + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log \tau_k(x_i;\theta).\end{aligned} \tag{1.6}$$

1.4.3 Entropy

The difference between $\log \mathrm{L}$ and $\log \mathrm{L_c}$ shall often be considered, when the labels $\mathbf{Z}$ are replaced by their conditional probabilities $\boldsymbol{\tau}$. It can be expressed through the *entropy* function:

$$\forall \theta \in \Theta_K, \forall x \in \mathbb{R}^d,\ \mathrm{ENT}(\theta;x) = -\sum_{k=1}^{K} \tau_k(x;\theta)\log \tau_k(x;\theta). \tag{1.7}$$

Then (1.6), with the labels $\mathbf{Z}$ replaced by their conditional probabilities $\boldsymbol{\tau}$, becomes

$$\log \mathrm{L_c}(\theta;\mathbf{x},\boldsymbol{\tau}(\theta)) = \log \mathrm{L}(\theta;\mathbf{x}) - \mathrm{ENT}(\theta;\mathbf{x}),$$

where $\mathrm{ENT}(\theta;\mathbf{x}) = \sum_{i=1}^{n}\mathrm{ENT}(\theta;x_i)$.

The definition of the criterion ICL (see Section 2.1.4) and the new contrast involved in Chapter 4 are based on this quantity. Its properties are further discussed in Section 4.2.2.

1.5 Example

Let us consider a simple simulation study to illustrate the preceding definitions, properties and discussions. The treated example shall be continued in Chapter 3.3.3 and Chapter 4.4.6.

Example 1 *Consider a sample from a four-component Gaussian mixture in* $\mathbb{R}^2$*. The covariance matrices are imposed to be diagonal, which corresponds to the* $[p_k\_ L_k\_ B_k]$ *model of Celeux and Govaert (1995). The sample has size* $n = 200$*. The true complete data (which is known from the method of simulation) is represented in Figure 1.1. This example then takes place in a situation with four subpopulations which have Gaussian shape each. The "observed" data is then what appears in Figure 1.2: we have to "forget"*

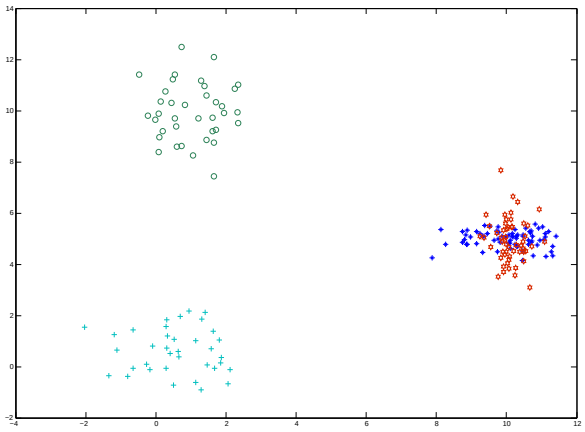


Figure 1.1: The true complete data

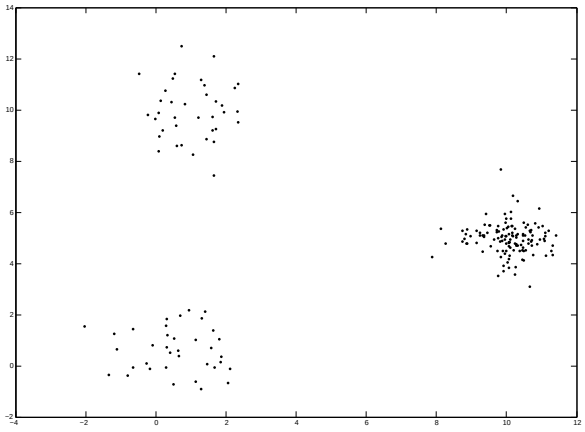


Figure 1.2: The observed data

the labels to obtain it. We suppose here that we know we want to design four classes from this data.

The usual model-based clustering approach consists then, as described in this chapter, in fitting a mixture model through maximum likelihood estimation to the data, and then designing classes according to the MAP rule.

Let us fit the model $\mathcal{M}_4$ with form $[p_k\_L_k\_B_k]$, by the EM algorithm. We did so by the use of the MIXMOD *software with a small_EM initialization step (see Section 1.2.2). The obtained Gaussian mixture and the corresponding MAP classification are reported in Figure 1.3.*

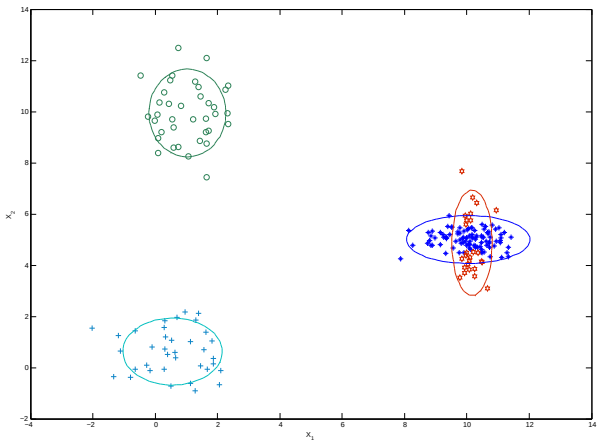


Figure 1.3: Classification through MAP under $\widehat{\theta}_4^{\text{MLE}}$
Ellipses represent isodensity curves corresponding to each component of the mixture under $\widehat{\theta}_4^{\text{MLE}}$

The obtained classification looks like the true one (Figure 1.1). But there are a few errors (see Figure 1.4), all of which obviously lie in the region of the intersection of the two oblate components: in this region, it is never clear to which component an observation should be assigned. The component probabilities conditionally to those observations have value about one half for each one of the two oblate fitted components and the label assigned by MAP is then quite doubtful. The inspection of the conditional probabilities should warn the statistician against this situation. He shall be confident in the classification of all observations assigned to the upper left class, since their conditional probabilities are about 1 *for this class, and* 0 *for the others. But he will be careful with the classification of the observations at the intersection of the "cross", since their posterior probabilities never have value close to* 1*...*

Let us finally remark that the maximum likelihood estimates well the true parameter.

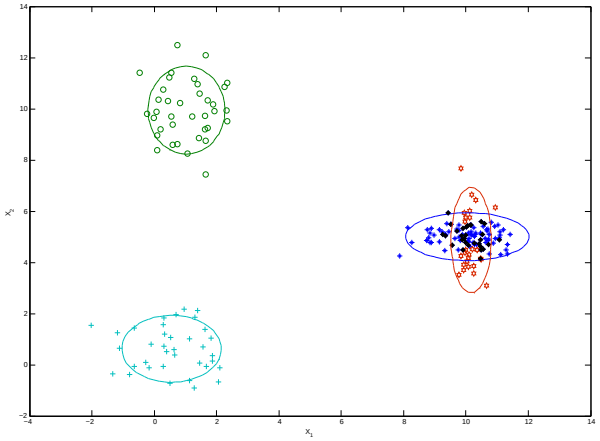


Figure 1.4: In black diamonds, the observations misclassified through MAP under $\widehat{\theta}_4^{\mathrm{MLE}}$
Ellipses represent isodensity curves corresponding to each component of the mixture under $\widehat{\theta}_4^{\mathrm{MLE}}$

Chapter 2

How Many Components? How Many Clusters? How Many Classes?

Contents

Model-based clustering basics have been introduced in Chapter 1. The number of classes to be designed was assumed to be known there. Now, as clustering consists of designing classes when no label has been observed, it may definitely be — and it currently occurs in practice — that the number of classes to be designed is unknown. This chapter introduces the main matter of this thesis, which is the choice of the number of classes. The main reference for all which is concerned in this chapter, with a lot of further material and discussions, is McLachlan and Peel (2000).

Gaussian mixture models can be used in a clustering purpose. The usual model-based clustering approach consists of fitting the model $\mathcal{M}_K$ for each one of the possible number of classes $K \in \{K_{\min}, \ldots, K_{\max}\}$ through maximum likelihood estimation, as explained in Section 1.2, to select a number of classes on the basis of the results of this estimation step, and to assign classes to the observations according to the MAP rule (see Section 1.3). Remark that this point of view is tightly related to the implicit choice that each class has to be identified with one of the mixture Gaussian component, as already discussed in the introduction of Section 1.3. One possible number of components corresponds to each model. Therefore selecting the number of components can be thought of as a model selection question. We shall see in Section 2.1 how the number of classes is usually chosen in this framework through the so-called penalized likelihood criteria. The classical criteria AIC and BIC are introduced respectively in Section 2.1.2 and Section 2.1.3. The ICL criterion has been derived especially in a clustering purpose. Its study is the first motivation of this thesis, and it is introduced in Section 2.1.4.

Now, the choice of identifying one class with each component might be questionable. Actually, in many situations, the classes' shapes have no reason to be "Gaussian". In this case, other component densities might be employed, but it may also be wanted to take advantage of the nice approximation properties of Gaussian mixtures. That is, one may want to model each class itself through a Gaussian mixture. The distribution of the entire data would then be a mixture of those mixtures. We will show in the second section of this chapter how it can be done by identifying as "clusters" (see the discussion about components, clusters and classes below) the Gaussian components to be merged. Classes are then assigned according to those clusters, and not to Gaussian components anymore, and their number is smaller than the number of components.

Classes: Components, Clusters?

Let us stress the differences between what is meant here by those three different notions of component, cluster and class.

The aim of clustering — as well as classification, by the way — is to design *classes*. A class is a group of data. The observations gathered in a same class are expected to come from a same subpopulation with specific features. A class can be designed from different points of view. The two main approaches are to design classes either by identifying one to each *component* of a fitted mixture (model-based clustering: see Section 2.1 and Chapter 4), or from a more intuitive and informal notion of *cluster*. Hybrid approaches may also be considered (see Section 2.2 and Chapter 7).

The notion of (mixture) *component* is a probability notion, and it has been rigorously defined in Section 1.1.1: it is one of the Gaussian (e.g.) densities which is involved in a

mixture.

We have not met the notion of *cluster* yet: it is rather linked to the geometrical idea of a group of observations which are close to each other and that we intuitively would not separate. This notion is not well defined: all authors do not agree, for example, about the question whether two much overlapping Gaussian components with very different shapes should be considered as one cluster or two, and how to decide it (see for example Hennig, 2009). Non-model-based clustering approaches such as the k-means types clustering methods, are based on this notion.

We would then rather defend the name of "unsupervised classification" for the purpose of designing classes from data when no label is available. The usual term of "clustering" is however employed in this thesis, as it is the usual dedicated term, in spite of the fact that it rather suggests a classification from the "cluster" point of view.

2.1 Penalized Criteria

Penalized criteria are introduced by considering one model for each considered number of components, notably for the sake of simplicity of the notation. However, the same criteria may be useful to select between several kinds of models for each considered number of components (homoscedastic or not, Gaussian components or not, etc.). It might for example be of interest to know whether the classes have equal volumes or not, etc. See for example Biernacki and Govaert (1999) upon this. It might even be necessary to use such criteria to select at the same time the variables which are interesting to design the clustering: see Maugis and Michel (2009), Maugis et al. (2009).

It should be cautioned that the AIC (Section 2.1.2) and the BIC (Section 2.1.3) criteria we are to describe were not designed in a clustering purpose. They were rather designed in a density estimation purpose. The philosophy of model-based clustering when using those criteria is then to look for the best estimation and approximation of the observations' distribution, and to trust this distribution to recover classes. One of the aims of this work is to show that the point of view which allows to understand the ICL criterion (Section 2.1.4), which has been derived in a clustering purpose, is radically different. This does not seem to be commonly accepted and is defended in Chapter 4. But we shall first explain how those criteria were elaborated.

Other approaches have been proposed and studied to evaluate the number of components of a mixture. Let us quickly introduce them. One consists of evaluating the number of modes of the distribution to assess the number of components. But a mixture of two normals might be unimodal. Moreover, this approach does not have the flexibility of mixture models (for example, in terms of shape of the fitted models), and does not provide as much interpretation possibilities. Titterington et al. (1985) explain and discuss this approach and how to apply it, particularly about the problem of assessing bimodality against unimodality. Another interesting method is based on likelihood ratio tests. This is probably the most natural statistical approach, but it does not allow to compare all the models at once as the penalized likelihood criteria do, and it is even only valid in nested models situations. Moreover it suffers from the non-identifiability of mixture models as the number of components is too great. The general asymptotic distribution of the likelihood ratio statistic had been long unknown, and was only known

in particular cases, or conjectured through simulations. McLachlan and Peel (2000) discuss those results, but also suggest to work around this difficulty by bootstrapping the likelihood ratio test statistic. Finally, Dacunha-Castelle and Gassiat (1999) introduced a "locally conic parametrization" of mixture models and succeeded in computing this asymptotic distribution, under quite strong regularity assumptions which are fulfilled by Gaussian mixture models, at least in the spherical case (covariance matrices proportional to the identity), all components having the same covariance matrix, when the parameter space is compact and $\pi_k > 0$ (Keribin, 2000). This partly solves the problem since the asymptotic distribution, although explicit, is difficult to evaluate. The authors suggest it could be itself bootstrapped.

2.1.1 Definition

The general idea of *penalized likelihood criteria* is the following. There is no reason to choose among the available models according to the maximum likelihood value reached in each one of them ($\mathrm{L}(\widehat{\theta}_{\mathcal{M}}^{\mathrm{MLE}})$). As a matter of fact, the likelihood is expected to mostly increase with the model complexity, notably as the models are nested. A relevant criterion can be based on the likelihood values, but the model selection paradigm suggests that it should be *penalized*. The criterion typically looks like

$$\mathrm{crit}(K) = \log L(\widehat{\theta}_K^{\mathrm{MLE}}) - \mathrm{pen}(K)$$

where $\mathrm{pen}(K) > 0$ may depend on n and even on the data (in which case the penalty is said to be *data-driven*).

The selected number of components is then

$$\hat{K} = \underset{K \in \{K_{\min}, \ldots, K_{\max}\}}{\operatorname{argmax}} \mathrm{crit}(K).$$

Note that K may be replaced by any model index if several models were to be compared for each K.

Let us now see how such classical penalties have been defined. They come from various considerations, but have in common to be fully justified only under identifiability conditions which may break down for mixture models. Let us then assume what is necessary, so that their derivation is justified. We already discussed those conditions when considering the identifiability of the Gaussian mixture models and the definition of maximum likelihood estimators.

It is interesting to consider what can be said from a general point of view, about the procedure. As already mentioned, we do not wish to assume that $f^{\wp}$ is a mixture distribution from any $\mathcal{M}_{K^*}$. White (1982) proved consistency and asymptotic normality of the maximum likelihood estimator in a general model (under regularity conditions), as the true distribution does not lie in the considered model (*misspecified model*). In this case, he proved the target distribution $f_{\mathcal{M}}^0$ to be the distribution in this model which minimizes the Kullback-Leibler divergence to the true distribution. Nishii (1988) continued on those works by studying, in the misspecified model case (which means here that none of the considered models contains the true distribution $f^{\wp}$), which penalties define a strongly (resp. weakly) *consistent* procedure, in the sense that $\hat{K}$ converges

almost surely (resp. in probability) to K_0, which is the less complex model among those considered minimizing $d_{\mathrm{KL}}(f^{\wp}, \mathcal{M}_K)$:

$$d_{\mathrm{KL}}(f^{\wp}, \mathcal{M}_K) = \min_{f \in \mathcal{M}_K} d_{\mathrm{KL}}(f^{\wp}, f) = d_{\mathrm{KL}}(f^{\wp}, f_K^0);$$

$$\text{Let } \theta_K^0 \in \Theta_K \text{ s.t. } f_K^0 = f(\,.\,; \theta_K^0);$$

$$K_0 = \min \operatorname*{argmin}_{K \in \{K_{\min}, \ldots, K_{\max}\}} d_{\mathrm{KL}}(f^{\wp}, \mathcal{M}_K).$$

In case $f^{\wp} \in \mathcal{M}_{K^*}$, K_0 is the smallest K such that $f^{\wp}$ belongs to $\mathcal{M}_K$ (and then $K_0 \leq K^*$).

The results of Nishii (1988) we are to discuss were extended in the mixture framework under quite strong regularity conditions by Leroux (1992) for the underestimation. Keribin (2000) dealt both with the under- and the overestimation, basing her results on the locally conic parametrization of Dacunha-Castelle and Gassiat (1999). She proved her assumptions to be fulfilled in particular by finite Gaussian mixtures, under the same identifiability conditions we assume, with means parameters lying in a compact space, and covariance matrices spherical and lower-bounded ($\sigma^2 Id$, with $\sigma^2 \geq \sigma_0^2 > 0$).

Nishii (1988) considers penalties of the form $pen(K) = c_n D_K$, with D_K the *dimension* of the model (the number of free parameters needed to parametrize it) and where c_n does not depend on the data. This kind of penalties include AIC and BIC. He computes the order of convergence of the involved maximum likelihood estimators. And so, he proves that, as soon as $c_n = o(n)$, the penalized criterion cannot asymptotically underestimate K_0. Actually, $\big(\log L(\widehat{\theta}_K^{\mathrm{MLE}}) - \log L(\widehat{\theta}_{K'}^{\mathrm{MLE}})\big)$ must be of order $n\big(d_{\mathrm{KL}}(f^{\wp}, f_K^0) - d_{\mathrm{KL}}(f^{\wp}, f_{K'}^0)\big)$, as soon as $\big(d_{\mathrm{KL}}(f^{\wp}, f_K^0) - d_{\mathrm{KL}}(f^{\wp}, f_{K'}^0)\big) \neq 0$. Suppose $K < K_0$. Then, asymptotically, $\mathrm{crit}(K_0) - \mathrm{crit}(K)$ is of the same order as $n\big(d_{\mathrm{KL}}(f^{\wp}, f_{K_0}^0) - d_{\mathrm{KL}}(f^{\wp}, f_K^0)\big)$ which converges to $-\infty$ as n does.

Now, the overestimation case ($K > K_0$) is more difficult to handle since it requires to have a control over the fluctuations of the difference between $\frac{1}{n}\log L(\widehat{\theta}_{K_0}^{\mathrm{MLE}})$ and $\frac{1}{n}\log L(\widehat{\theta}_K^{\mathrm{MLE}})$ which now converges to zero. Nishii (1988) proves a penalty such that $\frac{c_n}{\log\log n} \to \infty$ suffices to guarantee the strong convergence, and that the weak convergence is guaranteed as soon as $c_n \to \infty$.

Of course, those results once more only allow to justify the criteria asymptotically. But they offer a theoretical framework in which to compare the expected behavior of those criteria. The results of Nishii (1988) illustrate how much work it remains to (define and) design "optimal" penalties: they restrict the possible penalties to a still quite large family (between $\log\log n$ and $o(n)$). Penalties verifying those conditions can be very different and obviously have quite different behaviors for finite n. Notably, Nishii (1988) assumed the penalty to be proportional to the dimension. Keribin (2000) does not assume such a shape of the penalty, and the results she obtains does not provide any insight about this. Nonasymptotic oracle results (Massart, 2007, for example) allow to define more precisely the necessary shape of the penalty, (typically, up to a constant: see Section 3.1.2 and Section 3.2.1).

We are now to consider classical particular penalties.

2.1.2 AIC

Akaike (1973) defined the first one of those criteria, by considering $\frac{1}{n}\log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}})$ as an estimator of

$$\int f^{\wp} \log f(\,.\,;\widehat{\theta}_K^{\mathrm{MLE}}) d\lambda = -d_{\mathrm{KL}}(f^{\wp}, f(\,.\,;\widehat{\theta}_K^{\mathrm{MLE}})) + \int f^{\wp} \log f^{\wp} d\lambda,$$

which he aims at maximizing (the $\int f^{\wp} \log f^{\wp} d\lambda$ term does not depend on K and can be forgotten). But since the same data are used to design the maximum likelihood estimator $\widehat{\theta}^{\mathrm{MLE}}$ and to estimate its divergence to $f^{\wp}$, a bias is induced. The empirical distribution $\mathbb{F}_n$ of the data is indeed typically closer to $f(\,.\,;\widehat{\theta}^{\mathrm{MLE}})$ than the true distribution $f^{\wp}$. The intuition about this may be shaped through the (very!) particular case with $n = 1$: if the sample is X_1, then $\mathbb{F}_n = \delta_{X_1}$ and

$$\begin{aligned}\int \mathbb{F}_n \log f(\,.\,;\widehat{\theta}_K^{\mathrm{MLE}}) &= \log f(X_1;\widehat{\theta}_K^{\mathrm{MLE}}) \\ &= M \\ &> \int f^{\wp}(x) \log f(x;\widehat{\theta}_K^{\mathrm{MLE}}) dx,\end{aligned}$$

where $M = \max_{\theta\in\Theta_K} \log f(x;\theta)$ does not depend on x, except perhaps if there are constraints about the means in Θ_K: we assumed here there are not.

It is then necessary to estimate the bias

$$B(K) = \mathbb{E}_{f^{\wp}}\left[\frac{1}{n}\log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}}) - \int f^{\wp} \log f(\,.\,;\widehat{\theta}_K^{\mathrm{MLE}})\right],$$

where the expectation $\mathbb{E}_{f^{\wp}}$ concerns both L and $\widehat{\theta}_K^{\mathrm{MLE}}$.

By the way, this suggests a quite natural cross-validation procedure (Smyth, 2000), since there would be no bias if independent samples were employed to compute $\widehat{\theta}_K^{\mathrm{MLE}}$ on the one hand and to estimate the expectation $\mathbb{E}_{f^{\wp}}\left[\log f(X;\widehat{\theta}_K^{\mathrm{MLE}})\right]$ by the law of large numbers, with $\widehat{\theta}_K^{\mathrm{MLE}}$ considered as deterministic, on the other hand.

Akaike's fashion was to show that $B(K)$ is asymptotically (approximately) equal to D_K, which is the number of free parameters necessary to describe the model Θ_K. This justifies the definition of An Information Criterion (AIC, now called Akaike's Information Criterion):

$$\mathrm{crit}_{\mathrm{AIC}}(K) = \log L(\widehat{\theta}_K^{\mathrm{MLE}}) - D_K.$$

A rigorous derivation of this criterion under mild regularity condition can be found in Ripley (1995, Section 2.2). This yields a criterion which has been proved to be asymptotically efficient (see for example Yang, 2005 for definitions of those notions and further references).

However, even when the distribution of the data is available at least in one of the models at hand, it has often been observed that the AIC criterion tends to overestimate the true number of components. Indeed, this criterion (asymptotically) aims at minimizing the Kullback-Leibler divergence to the true distribution. And it should do so:

the Nishii's condition $\text{pen}_{\text{AIC}}(K) = o(n)$ is fulfilled, and it should then not underestimate the true number of components. But the penalty is not heavy enough to fulfill the other Nishii's condition, and there is no guarantee that it does not overestimate the true model. This is not an "identification" criterion (Yang, 2005), and it is thus not adapted to recover the true number of components, even when available. Actually, the simulations clearly show that it is not interesting for the purpose of selecting a relevant number of classes in clustering.

2.1.3 BIC

Schwarz (1978) derived a penalized criterion from Bayesian considerations.

He considers the *integrated likelihood* in model $\mathcal{M}$:

$$\begin{aligned} p(\mathbf{X}) &= \int p(\theta, \mathbf{X}) d\theta \\ &= \int \underbrace{L(\theta)}_{p(\mathbf{X}|\theta)} p(\theta) d\theta, \end{aligned}$$

where $p(\theta)$ is the prior distribution of θ. This is an interesting choice: instead of only taking into account the likelihood of what a model can do at best, the "mean" behavior of the whole model, with respect to the data, is integrated this way into the study. A wide, complex model of course has greater chances to include distributions very close to the empirical distribution related to the data, but it also includes more poor-fitting distributions than a more parsimonious model: considering the integrated likelihood is then in some sense an approach to deal with the complexity of the models. But its computation is obviously mostly intractable, particularly with as complex models as mixtures. Let us see how Schwarz (1978) derived a penalized criterion from this approach. Writing $\tilde{\theta}$ for the *mode* of $p(\theta, x)$,

$$\log p(\theta, \mathbf{X}) \approx \log p(\tilde{\theta}, \mathbf{X}) - \frac{1}{2}(\theta - \tilde{\theta})' H(\tilde{\theta})(\theta - \tilde{\theta})$$

and (assuming that the posterior is well concentrated around its mode)

$$\begin{aligned} p(\mathbf{X}) &\approx p(\tilde{\theta}, \mathbf{X}) \int \mathrm{e}^{-\frac{1}{2}(\theta-\tilde{\theta})' H(\tilde{\theta})(\theta-\tilde{\theta})} \\ &\approx p(\tilde{\theta}, \mathbf{X}) \frac{(2\pi)^{\frac{D}{2}}}{\sqrt{\det H(\tilde{\theta})}}. \end{aligned}$$

Which provides the Laplace's approximation

$$\log p(\mathbf{X}) \approx \log L(\tilde{\theta}) + \log p(\tilde{\theta}) + \frac{D}{2} \log 2\pi - \frac{1}{2} \log \det H(\tilde{\theta}).$$

Now, replace $\tilde{\theta}$ by $\widehat{\theta}^{\text{MLE}}$ and $H(\tilde{\theta}) = \frac{\partial^2}{\partial \theta^2} \left(\log p(\theta, \mathbf{X}) \right)_{|\tilde{\theta}}$ by $I(\widehat{\theta}^{\text{MLE}}) = \frac{\partial^2}{\partial \theta^2} \left(\log L(\theta) \right)_{|\widehat{\theta}^{\text{MLE}}}$, since it is assumed that the prior distribution $p(\theta)$ is very diffuse and non-informative and then, as a consequence, that the maximum likelihood is a good approximation of the mode $\tilde{\theta}$. This yields

$$\log p(\mathbf{X}) \approx \log L(\widehat{\theta}^{\text{MLE}}) + \log p(\widehat{\theta}^{\text{MLE}}) + \frac{D}{2} \log 2\pi - \frac{1}{2} \log \det I(\widehat{\theta}^{\text{MLE}}).$$

As

$$\det I(\widehat{\theta}^{\mathrm{MLE}}) = O(n^D),$$

it follows the Bayesian Information Criteria (BIC) by considering the terms of main order (and assigning the same prior probability to each model):

$$\mathrm{crit}_{\mathrm{BIC}}(K) = \log L(\widehat{\theta}_K^{\mathrm{MLE}}) - \frac{D}{2}\log n.$$

Schwarz (1978) obtained this criterion in the particular case of exponential families, and under assumptions related to the introduced Bayesian framework. The necessary regularity conditions break down when considering mixture models, but many authors advocate even though the use of the BIC in this framework, arguing theoretical and experimental reasons: Leroux (1992) proves that the BIC does not underestimate the number of components (under regularity conditions for mixtures); Roeder and Wasserman (1995) justify the use of the BIC approximation in a Bayesian framework with mixture models by the theoretical results of Kass and Wasserman (1995) (which strengthen Schwarz's results but also break down for mixtures) and their own simulation results; see also Fraley and Raftery (1998)... However, Biernacki et al. (2000) show by simulations that the BIC may overestimate the number of components in misspecified models situations (the true mixture distribution is available in none of the considered models). This may be understood as the tendency in this situation of the BIC to select a model which minimizes the Kullback-Leibler divergence to the true distribution (with, say, K^* components): if this is not available in the K^*-components considered mixture model, then the BIC tends to select a model with more components to get a better approximation of the true distribution. This is very well illustrated by the simulations of Biernacki et al. (2000) where the true distribution is a mixture of a Gaussian component and a uniform component. BIC tends to overestimate the number of classes (2) since it fits several Gaussian components to approximate the uniform one.

Finally, remark that both conditions of Nishii (1988) are fulfilled for the BIC: it is then expected to be consistent in the sense that it should converge to the true number of components and model shape, as the true distribution is available. This is what Keribin (2000) proved for Gaussian mixtures in the particular first-mentioned situation, under regularity conditions, with the locally conic parametrization.

BIC is a good "identification" criterion.

2.1.4 ICL

Biernacki et al. (2000) attempt to tackle the difficulty encountered by the BIC to select a relevant number of classes under the misspecified situation. Their idea is to mimic the BIC approach but replacing the observed likelihood by the complete data likelihood (also called the classification likelihood, see Section 1.1.4). They expect this way to find a criterion which would take into account the clustering quality and to avoid the overestimation phenomenon from which the BIC might suffer.

Consider the integrated complete data log likelihood to which the same approxima-

tion as the one leading to the BIC is applied:

$$\begin{aligned}\log p(\mathbf{X},\mathbf{Z}) &= \log \int f(\mathbf{X},\mathbf{Z};\theta)p(\theta)d\theta \\ &\approx \max_{\theta} \log f(\mathbf{X},\mathbf{Z};\theta) - \frac{D_K}{2}\log n + O(1).\end{aligned}$$

But the label $\mathbf{Z}$ is unobserved and Biernacki et al. (2000) choose to replace it with $\widehat{\mathbf{Z}}^{\mathrm{MAP}}(\widehat{\theta}^{\mathrm{MLE}})$. Moreover, they replace $\max_\theta \log f(\mathbf{X},\mathbf{Z};\theta)$ with $\log f(\mathbf{X},\mathbf{Z};\widehat{\theta}^{\mathrm{MLE}})$, arguing that $\widehat{\theta}^{\mathrm{MLE}} \approx \operatorname{argmax}_\theta \log f(\mathbf{X},\mathbf{Z};\theta)$ as n is large enough. We think this is highly questionable: see Chapter 4. But following those choices, and keeping the relation between the complete data likelihood and the observed likelihood (1.6) in mind, we get

$$\begin{aligned}\mathrm{crit}_{\mathrm{ICL}^1}(K) &= \log f(\mathbf{X},\widehat{\mathbf{Z}}^{\mathrm{MAP}};\widehat{\theta}_K^{\mathrm{MLE}}) - \frac{D_K}{2}\log n \\ &= \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}}) + \sum_{i=1}^{n}\sum_{k=1}^{K} \widehat{Z}_{i,k}^{\mathrm{MAP}} \log \tau_{ik}(\widehat{\theta}_K^{\mathrm{MLE}}) - \frac{D_K}{2}\log n.\end{aligned}$$

McLachlan and Peel (2000) propose to rather replace $\mathbf{Z}_i$ by $\tau_i(\widehat{\theta}^{\mathrm{MLE}})$. This yields

$$\begin{aligned}\mathrm{crit}_{\mathrm{ICL}^2}(K) &= \log f(\mathbf{X},\boldsymbol{\tau}(\widehat{\theta}_K^{\mathrm{MLE}});\widehat{\theta}_K^{\mathrm{MLE}}) - \frac{D_K}{2}\log n \\ &= \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}}) - \mathrm{ENT}(\widehat{\theta}_K^{\mathrm{MLE}}) - \frac{D_K}{2}\log n\end{aligned}$$

from (1.4.3). The *entropy* term, first defined in Section 1.4.3, is further studied in Section 4.2.2: it is a measure of the *assignment confidence*.

McLachlan and Peel (2000) rather follow the lines of the derivation of the criterion in Biernacki et al. (1998) (a first version of Biernacki et al., 2000), which is slightly different and more precise, but is essentially based on the same assumptions and choices as ICL^1. Remark however that Biernacki et al. (1998) do not plug in the maximum likelihood estimate everywhere:

$$\mathrm{crit}_{\mathrm{ICL}^3}(K) = \operatorname*{argmax}_{\theta} \log f(\mathbf{X},\widehat{\mathbf{Z}}^{\mathrm{MAP}};\theta) - \frac{D_K}{2}\log n.$$

This seems to be a more reliable point of view, which is intermediate between ICL^1 and the criterion we propose in Chapter 4. In this version, the BIC-like approximation is further justified. However, $\widehat{\mathbf{Z}}^{\mathrm{MAP}}$ is still based on the maximum likelihood estimator (or any other estimator). Moreover, the authors seemingly chose to use the maximum likelihood estimator as an approximation of $\operatorname{argmax}_\theta \log f(\mathbf{X},\widehat{\mathbf{Z}}^{\mathrm{MAP}};\theta)$ in the practice, and notably in the simulations, and finally exposed the ICL^1 version in the final paper.

The two versions ICL^1 and ICL^2 differ in that the first one is based on a "hard" assignment of the labels, whereas the second one is based on a "soft" assignment, where each component is weighted by its probability conditionally to the observation at hand. They only differ for mixtures for which there are observations which are assigned with

uncertainty, i.e. for which several conditional probabilities are different from zero. Now, since

$$\begin{aligned}\forall\theta, \forall x \in \mathbb{R}^d,\ -\mathrm{ENT}(\theta; x) &= \sum_{k=1}^{K} \tau_k(x;\theta) \log \tau_k(x;\theta) \\ &\leq \sum_{k=1}^{K} \tau_k(x;\theta) \log \max_{j\in\{1,\ldots,K\}} \tau_j(x;\theta) \\ &= \sum_{k=1}^{K} \widehat{z}_k^{\mathrm{MAP}}(x;\theta) \log \tau_k(x;\theta),\end{aligned}$$

$\mathrm{crit}_{\mathrm{ICL}^1} \geq \mathrm{crit}_{\mathrm{ICL}^2}$: ICL^1 penalizes more the models giving rise to uncertain clustering than ICL^2. However, they are expected to behave analogously and we will from now on refer to second version under the general name of ICL.

The ICL uses to be presented as a log likelihood penalized criterion, which penalization includes an entropy term ($\mathrm{pen}_{\mathrm{ICL}}(K) = \mathrm{ENT}(\widehat{\theta}_K^{\mathrm{MLE}}) + \frac{D_K}{2}$). This term penalizes the models giving rise to uncertain classifications and the complexity of the models is penalized the same way as by the BIC.

There are no further theoretical results about the ICL criterion up to now. Chapter 4 is an attempt to set a convenient theoretical framework to better understand the behavior of the ICL and to justify it.

Biernacki et al. (2000) claim that the ICL is more robust to model misspecifications than the BIC, when selecting the number of classes. This is obviously a nice feature for such a criterion when the purpose is clustering. Biernacki et al. (2000) and McLachlan and Peel (2000) illustrated this through simulated and real data examples. McNicholas and Murphy (2008) found that ICL reached analogous performances as BIC, as applied to some data with the parsimonious Gaussian mixture models they propose.

Actually, the key here is that ICL is not "consistent" in the sense that BIC is. First, it does not even avoid underestimation of the true number of components, as AIC does: it should first be cautioned that the ICL does not fulfill the first Nishii's condition: $\mathrm{pen}_{\mathrm{ICL}} \neq o(n)$ (from the law of large numbers, $\frac{1}{n}\mathrm{ENT}(\mathbf{X};\theta) \longrightarrow \mathbb{E}_{f^\wp}[\mathrm{ENT}(X;\theta)]$ and $\mathrm{ENT}(\mathbf{X};\widehat{\theta}^{\mathrm{MLE}})$ is then rather expected to be of the same order as n and should not be considered as a part of the penalty). This has to be linked with the BIC's tendency to overestimate the true number of components under the model misspecification. The BIC selects a model which correctly approximates the true distribution. On the contrary, ICL is so designed that it shall not select an optimal model from the approximation (Kullback-Leibler) viewpoint but that it selects a model which the corresponding obtained classification is relevant. We have here to clearly distinguish the notions of components and clusters: whereas the BIC should (asymptotically) select the true number of (Gaussian) components, ICL selects a relevant number of classes, based on a certain notion of cluster taking the fit and the assignment confidence into account. This point of view is further developed in Chapter 4. As we chose to identify the number of classes with the number of Gaussian components, this means selecting $\mathcal{M}_K$ with K a relevant number of classes, even if $\mathcal{M}_K$ does not have the best approximation properties with respect to the observations. BIC and ICL have the same behavior (and actually almost the same values) for models which provide well-separated components, but ICL penal-

izes models which do not, on the contrary to BIC which only penalizes the complexity of the models.

This will be illustrated through simulations and further discussed in Chapter 4.

Conclusion

Many other criteria to select the number of classes (mostly not of the penalized-likelihood form) were defined but we are mainly interested in those presented above since, on the one hand, the AIC and the BIC are very popular and widely used. But they were not elaborated in a particular clustering purpose, and we saw that their use in such a framework might be questionable. ICL, on the other hand, was designed for a clustering purpose. But no theoretical studies or results were available up to now, to help understanding what aim it targets. Its practical results are however appreciated, at least in some situations, because they meet what people need and what they "intuitively" like to obtain. See Section 4.1.2 for examples of applications where ICL met the objective of the users. We then propose an attempt to set a theoretical framework in which the behavior of ICL can be described (Chapter 4). This is then an attempt to define a theoretical counterpart to this intuition. We will see that it will give rise to a criterion which is closely linked to ICL, but might be calibrated in a data-driven fashion. Such kind of criteria including a criterion in the usual likelihood framework, and which rather behaves like BIC, are derived from the slope heuristics of Birgé and Massart (2006). They will be first introduced in Section 3.3 where the contrast minimization framework is introduced, and then further studied in chapter Chapter 4.

2.2 Components Are Not Always Classes!

Model-based clustering, specially when based on Gaussian mixtures, obviously suffers from situations where clusters do not have Gaussian-like (namely ellipsoid) shape. In this situation, when embracing the "one component = one class" rule, two points of view have been presented: the density approximation point of view of BIC, which suffers from overestimation of the number of classes, and the component/cluster point of view of ICL, which might suffer from quite poor approximation results, which could be a drawback when the distribution of the data is of interest, too. A radically different point of view is clustering through non model-based clustering. An intermediate solution consists of fitting a Gaussian mixture to the data, typically by the BIC solution, to get a nice fit to the data and then to possibly merge some of the fitted Gaussian components which are considered together as a single cluster. Finally classes are formed according to those clusters' conditional probabilities.

The idea of merging Gaussian components to get a better fit had already being applied in the classification framework for a while (see Hastie and Tibshirani, 1996). Then, it was applied in a clustering framework by, for instance, Tantrum et al. (2003) and Li (2005). Hennig (2009) proposes an overview of the existing methods based on this idea and some improvements on them, as well as some new ones and a new visualization method. We also proposed a solution to this approach (Baudry et al., 2008b): this is essentially Chapter 7. Those different methods differ by the way they choose which

components to merge. Some give solutions or at least tracks to choose the number of classes to be designed (typically, the number of components is chosen — with good reason — through the BIC).

Let us explain how some Gaussian components of a mixture may be merged to design mixture classes, and give insight into how our method enables to choose which components to merge. Our method is notably partly compared to the related methods presented in Hennig (2009).

2.2.1 Mixtures as Classes

The starting point is a good estimation of the data distribution. The Gaussian mixture models $\mathcal{M}_K$ have nice approximation properties, provided that a relevant number of components has been chosen. For example, the BIC criterion, based on the maximum likelihood estimations is expected to select a number of components corresponding to the best model from the approximation point of view, at least asymptotically (see Section 2.1.3). Let us consider the solution it yields (with $\hat{K}$ the number of components it selects and $\hat{\theta} = \widehat{\theta}^{\text{MLE}}_{\hat{K}}$):

$$f(\,.\,;\hat{\theta}) = \sum_{k=1}^{\hat{K}} \hat{\pi}_k \underbrace{\phi(\,.\,;\hat{\omega}_k)}_{\hat{\phi}_k(\cdot)}.$$

As already mentioned, there is no reason that this solution should be relevant from a clustering point of view. Of course, it depends on the application at hand, but when some of the components overlap, the classification obtained through the MAP rule with respect to the Gaussian components may not provide a relevant clustering. This is a strictly component-based notion of classes, whereas the user may often want to involve a *cluster* notion in the study. But it is however an interesting approach to base a clustering study of the data on what statistics can offer at best to estimate its distribution. It can be done by *merging* some of the obtained Gaussian components. That is, to consider several components have to be handled as a single class. For any number $K \leq \hat{K}$ of classes to design, the estimated distribution may be rewritten as

$$f(\,.\,;\hat{\theta}) = \sum_{k=1}^{K} \hat{\pi}_{J_k^K} \underbrace{\sum_{j \in J_k^K} \frac{\hat{\pi}_j}{\hat{\pi}_{J_k^K}} \hat{\phi}_j}_{\hat{f}_k^K},$$

with $\{J_1^K, \ldots, J_K^K\}$ a partition of $\{1, \ldots, \hat{K}\}$ and $\hat{\pi}_{J_k^K} = \sum_{j \in J_k^K} \hat{\pi}_j$ for any $k \in \{1, \ldots, K\}$. This is exactly the same density (and fit to the data) as before, but this writing as a "mixture of Gaussian mixtures" emphasizes that the components which labels belong to the same set J_k^K are considered together. The only consequence concerns the conditional probabilities, and then the entropy value (see below) and the MAP rule classification: for any observation x_i, the conditional probabilities are computed with

respect to the obtained classes. We denote[1] them by $\hat{\tau}^K$:

$$\forall k \in \{1, \ldots, K\}, \hat{\tau}_k^K(x_i) = \frac{\hat{\pi}_{J_k^K} \hat{f}_k^K(x_i)}{\sum_{k'=1}^K \hat{\pi}_{J_{k'}^K} \hat{f}_{k'}^K(x_i)}$$

and the estimated labels are assigned according to those conditional probabilities:

$$\hat{z}_i^K = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \hat{\tau}_k^K(x_i).$$

The merging process is performed hierarchically.

Two observations which were assigned, under $\hat{\theta}$, labels corresponding to components which have been merged are then mostly assigned the same label, corresponding to the new class. The reverse is not always true since it may happen that, by merging some components, the obtained class is assigned to an observation which did not "belong" to any of the merged components (think of the example an observation has around one third conditional probability of belonging to any of three components and those to which it was not assigned are merged...). This is a sensible behavior: this is a consequence of the fact that the classification is designed according to the merged classes and not to the original components.

2.2.2 Merging Criterion

Now, it has to be chosen which components have to be merged. Hennig (2009) establishes an overview of methods proposed to this aim. According to his classification, two main groups of methods emerge, depending on the general cluster notion embraced. The ones are based on a notion of unimodal clusters, and the others are based on estimations of the misclassification probabilities, which means a notion of cluster related to the quality of the yielded classification. All of those methods are performed hierarchically: starting from the initial mixture, two components are chosen according to the embraced criterion and possibly merged. A mixture of components is obtained, to which the same is performed, and so on until a stopping rule — which is specified in each case — is reached.

In a joint work with G. Celeux, A.E. Raftery, R. Gottardo and K. Lo, we actually proposed such a merging method in Baudry et al. (2008b). It is based on the *entropy* measurement of the assignment confidence (please refer to Section 1.4.3 and Section 4.2.2), and then rather belongs to the second family of methods. This work is the subject of Chapter 7. Let us quickly introduce the embraced criterion. At any step K, the solution at hand can be defined by $\{J_1^K, \ldots, J_K^K\}$ (see the subsequent remark about the definition of the entropy). We propose to choose the two classes J_a^K and J_b^K

[1] The dependency of $\hat{\tau}^K$ on $\{J_1^K, \ldots, J_K^K\}$ is omitted in the notation.

to be merged by maximizing the entropy decrease:

$$\begin{aligned}\{a,b\} &= \underset{\alpha,\beta\in\{1,\dots,K\}}{\operatorname{argmax}} -\sum_{i=1}^{n}\Bigg(\sum_{k=1}^{K}\hat{\tau}_{ik}^{K}\log\hat{\tau}_{ik}^{K} \\ &\qquad - \sum_{k\in\{1,\dots,K\}\setminus\{\alpha,\beta\}}\hat{\tau}_{ik}^{K}\log\hat{\tau}_{ik}^{K} \\ &\qquad - (\hat{\tau}_{i\alpha}^{K}+\hat{\tau}_{i\beta}^{K})\log(\hat{\tau}_{i\alpha}^{K}+\hat{\tau}_{i\beta}^{K})\Bigg) \\ &= \underset{\alpha,\beta\in\{1,\dots,K\}}{\operatorname{argmax}} -\sum_{i=1}^{n}\Big(\hat{\tau}_{i\alpha}^{K}\log\hat{\tau}_{i\alpha}^{K}+\hat{\tau}_{i\beta}^{K}\log\hat{\tau}_{i\beta}^{K}-(\hat{\tau}_{i\alpha}^{K}+\hat{\tau}_{i\beta}^{K})\log(\hat{\tau}_{i\alpha}^{K}+\hat{\tau}_{i\beta}^{K})\Big).\end{aligned}$$

And then,

$$\{J_1^{K-1},\dots,J_{K-1}^{K-1}\} = \{J_1^K,\dots,J_K^K\}\setminus\{J_a^K,J_b^K\}\bigcup\{J_a^K\cup J_b^K\}.$$

We do not advocate any definitive stopping rule for the merging process. Actually, the whole hierarchy may often be of interest in applications. A careful analysis of which classes are merged at each step, and notably the order in which they are, is of interest, and brings much more informations than any single solution with a given K. The plot of the resulting entropy with respect to the number of classes, possibly rescaled by the number of observations involved in the corresponding merging step, is a helpful tool to point out at which steps something particularly interesting occurs. In case a method to choose the final number of classes is really needed, we suggest an approach based on a piecewise linear regression fit to the (possibly rescaled) entropy plot.

Remark that the entropy could only be rigorously defined here over a set similar to $\widetilde{\mathcal{M}}_K$ for any K (see Section 1.1.1), with the "components" being themselves mixtures. Only an element of such a set brings enough information — notably, the definition of each class — so as to compute the entropy: with $\hat{f}$ only, only the usual entropy, with respect to each component may be computed.

This approach then relies on a combination of the notions of components (to fit the data) and cluster (first merge components such that the resulting classes are assigned with greatest confidence). The involved notion of cluster relies on the entropy point of view: i.e. an assignment confidence notion. See Section 4.2.2 for further discussion on this cluster notion.This is interesting for applications where the user wants to be able to design classes with great confidence, which is not always the case of course: Hennig (2009) proposes several approaches, which correspond to different cluster notions, and then which fit different applications needs. Ours is seemingly quite linked to his so-called "ordered posterior plot" method. This visualization method relies on the plots of the conditional probabilities values, for each initial component, of the observations with nonnegligible conditional probability (see Hennig, 2009, Section 5). Actually, this method seems to enable to make visual the notion which roughly corresponds to what the entropy measures: which components are assigned with great confidence or not, and when not, which are those which overlap the most.

Our method is also obviously linked to the "Bhattacharyya Distance" and the "Direct Estimated Misclassification Probabilities" (*DEMP*) methods presented in Hennig

(2009). The first one merges components according to the Bhattacharyya distance between them, which has the property to be (the logarithm of) a lower-bound of their overall Bayes misclassification probability. As mentioned by the author, it is not necessarily a sharp bound, however. Moreover, to avoid computing the Bhattacharyya distance between mixtures, each class is represented by a Gaussian density. This may be an important difference with our method when some obtained mixtures have very non-Gaussian shape: think of the "Cross" experiment example (see for example Section 7.4.4, which could presumably be modified to highlight the different behavior of the two methods). On the contrary to this method, ours does take into account the mixing proportions (indirectly, through the number of observations assigned to each class, which is tightly linked to it) when choosing which classes to merge: two classes which overlap much, one of them being a small-size class, may be merged later in the hierarchy than two classes which overlap less (in mean) but with great sizes. This is also the case of the DEMP method. Its name is quite explicit: it consists of estimating the probability that an observation arising from a mixture corresponding to a class, be assigned the label of another class. The maximum for any pair of classes is considered. And the pair maximizing this quantity is merged. This approach presumably mostly behaves similarly to ours.

Chapter 3

Contrast Minimization, Model Selection and the Slope Heuristics

Contents

In this chapter is first recalled the framework of contrast minimization. This very general framework covers the study of well-known and popular estimation and model selection methods, as shown in Massart (2007). Several goals are distinguished and the difference between identification and efficiency is highlighted.

It is also reported how some penalized criteria can be known up to a multiplying constant. This motivates the slope heuristics of Birgé and Massart (2006), which is recalled: the notions of oracle, optimal and minimal penalties are discussed. Two data-driven approaches to practically take advantage of this heuristics are introduced: the dimension jump and the so-called data-driven slope estimation. Their introduction is done so as to make as obvious as possible their links and differences.

An application of those concepts and methods for the choice of the number of components in the clustering with Gaussian mixture models framework is finally proposed. Considering (minus) the usual likelihood as a contrast, criteria are derived which are tightly linked to AIC and BIC. Chapter 4 will be devoted to the study of ICL by an application of those ideas to a contrast which is directly linked to clustering.

3.1 Contrast Minimization

3.1.1 General Setting

Let us give notation for the very general contrast minimization framework, as introduced for example in Massart (2007), which should be the reader's handbook for the topic of this section.

Recall the i.i.d. sample $X_1, \dots, X_n \in \mathbb{R}^d$ arises from an (unknown!) distribution with density $f^{\wp}$. To avoid ambiguities, we will denote $\mathbb{E}_{X_1,\dots,X_n}$ the expectation taken with respect to the sample, and $\mathbb{E}_X$ the expectation taken with respect to $X \sim f^{\wp}$.

The quantity of interest, which lies in a set $\mathcal{U}$ ("$\mathcal{U}$" stands for "universe"), is not necessarily this distribution, and we shall denote it by s: it is somehow related to $f^{\wp}$. The method is based on the existence of a *contrast* function[1] $\gamma : \mathbb{R}^d \times \mathcal{U} \longrightarrow \mathbb{R}$ fulfilling the fundamental property that

$$s = \underset{t \in \mathcal{U}}{\operatorname{argmin}} \, \mathbb{E}_X \left[\gamma(X; t)\right] \tag{3.1}$$

(in good settings, s is expected to be unique). This settles γ as the track to s. The associated *loss function* is used to evaluate each element of $\mathcal{U}$ in this light:

$$\forall t \in \mathcal{U}, \; l(s,t) = \mathbb{E}_X \left[\gamma(t)\right] - \mathbb{E}_X \left[\gamma(s)\right].$$

This function is nonnegative and is zero if and only if $t = s$. Let $\mathcal{M}$ be a model (i.e. a subset of $\mathcal{U}$). (One of) the best *approximations* of s in $\mathcal{M}$ is

$$s_{\mathcal{M}} \in \underset{t \in \mathcal{M}}{\operatorname{argmin}} \, \mathbb{E}_X \left[\gamma(t)\right].$$

[1] X will often be implicit in the notation: $\gamma(t) = \gamma(X; t)$. We might then write $\mathbb{E}_X \left[\gamma(t)\right]$ even if X does not appear in the notation.

A natural *estimator* of $s_{\mathcal{M}}$ in model $\mathcal{M}$ is then

$$\hat{s}_{\mathcal{M}} \in \underset{t \in \mathcal{M}}{\operatorname{argmin}} \gamma_n(t),$$

where

$$\forall t \in \mathcal{U},\ \gamma_n(t) = \frac{1}{n} \sum_{i=1}^{n} \gamma(X_i; t)$$

is the *empirical contrast* drawn from the sample. It is expected that, γ_n being close to $\mathbb{E}_X [\gamma(X; .)]$, $\hat{s}_{\mathcal{M}}$ will be a good estimator of $s_{\mathcal{M}}$. The quality of an estimator is measured by its *risk*

$$R(\hat{s}_m) = \mathbb{E}_{X_1,\ldots,X_n} [l(s, \hat{s}_m)],$$

or even, as we shall see, by its ability to minimize the loss function for a particular sample, with great probability (rather than in expectation).

Typical and classical examples of contrast minimization situations are maximum likelihood or least squares estimations. Numerous applications may be found in Massart (2007). We consider in this thesis two particular applications of it in the framework of mixture models: the maximum likelihood method which is first considered in Section 3.1.2 and a new contrast, which purpose is clustering, defined and studied in Chapter 4.

3.1.2 Model Selection

Suppose now a family of models $(\mathcal{M}_m)_{\{m \in M\}}$ is considered. The notion of *complexity* of the models shall be referred to subsequently: The suitable measure of the complexity of the models depends on the particular situation and has to be guessed through the theoretical study. It is typically the dimension of the model in a finite-dimensional siuation.

The question is how to choose among the corresponding family of estimators $(\hat{s}_m)_{\{m \in M\}}$? Let $\hat{m}$ be a model selection procedure. The final estimator is then $\hat{s}_{\hat{m}}$, where both $\hat{s}_m$ (for any m) and $\hat{m}$ are build from the same sample $X_1, \ldots, X_n$. Such a procedure may be evaluated from either an asymptotic or a non-asymptotic point of view (see for example Yang, 2005 or Arlot, 2007).

A "best" ideal model $\mathcal{M}_{m^*}$ for a given n and a given dataset is such that

$$m^* \in \underset{m \in M}{\operatorname{argmin}}\, l(s, \hat{s}_m).$$

Since it depends on the distribution $f^{\wp}$, the corresponding estimator $\hat{s}_{m^*}$ is called the *oracle*. It is a benchmark for a selection model procedure.

Let us first give examples of asymptotic points of view. An *asymptotic optimality* with respect to the loss function l and to the oracle can be defined by

$$\frac{l(s, \hat{s}_{\hat{m}})}{\inf_{m \in M} l(s, \hat{s}_m)} \xrightarrow[n \to \infty]{\text{p.s.}} 1.$$

A procedure fulfills this property if it reaches as good results — as measured by the loss function — as the oracle up to a multiplying factor going to 1 as n goes to infinity. This is an *efficiency* goal.

Another important asymptotic property is the "consistency". This is an *identification* goal, in that it consists of being able to recover the best model from the loss (approximation) point of view, rather than being able to mimic the oracle:

$$\hat{m} \xrightarrow[n\to\infty]{\text{p.s.}} \operatorname*{argmin}_{m\in M} l(s, s_m).$$

This point of view only makes sense if it is assumed that a model minimizes the loss function. This is a "true" model assumption, but this does not mean that the corresponding s_m is the distribution $f^{\wp}$ (see remark below) neither that it is s. But for instance, in the situation where $s = f^{\wp}$ (this is the case typically as the considered contrast is the usual log likelihood) and $\exists m_0 \in M / \; s \in \mathcal{M}_{m_0}$, a model selection procedure is *consistent* if it recovers the model containing $f^{\wp}$. When several models contain it, the procedure is expected to recover the less complex one. This is the most considered point of view in this thesis: this is a natural goal when trying to recover the "good" (if not true) number of classes, from the loss function point of view (notably when it is identified to the number of components of a mixture model).

See McQuarrie and Tsai (1998, Chapter 1) for a discussion about the different points of view underlying consistent criteria on the one hand and efficient criteria on the other hand.

The definition of the minimax property is beyond the scope of this work. Roughly speaking, an estimator is minimax if it is uniformly efficient over a given class of values of s. Let us however notice that is has been proved in the regression framework (with least-squares loss) that a model selection procedure cannot be simultaneously consistent and minimax. The interested reader shall refer to Yang (2005).

Now, from a non-asymptotic and efficiency point of view, a "good" procedure $\hat{m}$ is expected to fulfill an *oracle inequality*:

$$l(s, \hat{s}_{\hat{m}}) \leq C \inf_{m\in M} l(s, \hat{s}_m) + \eta_n,$$

with C a constant as close to 1 as possible, and η_n a remainder term negligible with respect to $l(s, \hat{s}_m)$. No "true" model has to be assumed. This inequality is expected to hold either with great probability or in expected value, or even, when such results are too difficult to achieve, as a weaker result[2]:

$$\mathbb{E}_{X_1,\ldots,X_n}\left[l(s, \hat{s}_{\hat{m}})\right] \leq C \inf_{m\in M} \mathbb{E}_{X_1,\ldots,X_n}\left[l(s, \hat{s}_m)\right] + \eta_n. \tag{3.2}$$

Such a procedure achieves non-asymptotic results in that the constant C does not depend on n.

Such a non-asymptotic viewpoint provides better understanding and evaluation of the true situation, mainly as a great number of models are considered, or as the complexity of those models is high. Reciprocally, this point of view is interesting for small-size samples: this is indeed a question of scale between the sample size n and the number and/or complexity of the considered models which is involved. Think for example of the case where models with dimension greater than the number of observations are considered. Or of the case where a huge number of models is considered for a given dimension:

[2]Indeed $\mathbb{E}_{X_1,\ldots,X_n}\left[\inf_{m\in M} l(s, \hat{s}_m)\right] \leq \inf_{m\in M} \mathbb{E}_{X_1,\ldots,X_n}\left[l(s, \hat{s}_m)\right]$.

the behavior of the best estimator among those obtained for each one of those models is to be quite different from the behavior of each single estimator. This is the role of the concentration inequalities for maximum of empirical processes derived in Massart (2007) to control such situations.

Let us stress how different this paradigm is from the usual "consistency" point of view. Once more, a first particularity is that the objective is not always the true distribution $f^{\wp}$, but s. This holds both for the identification and the efficiency points of view. It will be seen for example with the clustering contrast introduced in Chapter 4 that this target, if related to $f^{\wp}$, might be quite different from it. And above all, from the efficiency point of view, even if there exists m_0 such that $s \in \mathcal{M}_{m_0}$, there is no reason that $m^* = m_0$. Indeed, m^* has to take into account the complexity of the models: the decomposition of the loss into an approximation and an estimation parts

$$l(s, \hat{s}_m) = l(s, s_m) + \mathbb{E}_X \left[\gamma(s_m) - \gamma(\hat{s}_m)\right],$$

illustrates that a bias/variance compromise has to be reached. The more complex the model, the larger the (expected value of the) second term in this decomposition. Even if $\mathcal{M}_{m_0}$ exists, it might be too complex: the estimator in this model might then have a too great variance and shall not be preferred to an estimator in a smaller model, with greater bias but in which s_m is to be better estimated since the variance is smaller. It is not helpful (and might be worse) that a model contains a very good approximation of s if the available data do not enable to recover it!

The main approaches to design such model selection procedures are hold-out and cross-validation procedures (see Arlot, 2007), or penalized criteria. We consider in this thesis such criteria, which are of the form

$$\hat{m} \in \underset{m \in M}{\operatorname{argmin}} \Big\{ \underbrace{\gamma_n(\hat{s}_m) + \operatorname{pen}(m)}_{\operatorname{crit}(m)} \Big\}, \tag{3.3}$$

with $\operatorname{pen} : M \longrightarrow \mathbb{R}^+$. The reasons of the necessity of this penalty have been discussed in Section 2.1. It will be further explained in this framework in Section 3.2.

How to choose pen? Some penalized criteria we already introduced were designed from essentially asymptotic considerations (see Section 2.1). We shall come back to ICL later. AIC and BIC are such criteria in the likelihood contrast framework with penalties proportional to the dimension of the model, which is the measure of the complexity in the mixtures framework. AIC has been proved to be asymptotically efficient and minimax from an efficiency point of view in some frameworks (see for example Yang, 2005 and Section 2.1.2), while BIC is known to be consistent (see Nishii, 1988; Keribin, 2000 and Section 2.1.3). Now, concentration results can be used to design penalties which perform from a non-asymptotic point of view almost as well as the oracle in a wide range of situations (see Massart, 2007). But such penalties might be known from theory up to a multiplying factor κ. Consider as an example the Mallows' Cp criteria (Mallows, 1973), in a histogram regression framework, which is proportional to the variance σ^2. It is typically unknown. Another case example of such a situation is Theorem 7.11 in Massart (2007) (recalled in Section A.1) which provides a penalty up to an unknown multiplying constant, in a general maximum likelihood framework. This theorem is general and guarantees the existence of a constant κ_{opt} and a penalty shape $\text{pen}_{\text{shape}}$ (following the notation of Arlot and Massart, 2009) such that the procedure

based on $\text{pen}_{\text{opt}} = \kappa_{\text{opt}} \, \text{pen}_{\text{shape}}$ follows an oracle inequality (if the oracle risk can be linked to the minimal Kullback-Leibler divergence to the true distribution among the models: see Section A.1). This result is applied and discussed to the problem of choosing the number of components in the Gaussian mixture framework, in Section 3.3. But the optimal constant κ_{opt} in each particular situation is unknown. Actually, a value may be deduced from the theory, but this would be far from being optimal in any application, since it has to be pessimistic enough to be general: the theorems should rather be used as guidelines to choose the penalty shape, but the "good" constant κ depends on the application at hand.

Birgé and Massart (2001) introduced the idea of trying to estimate the best constant in each particular case by a study which is based on the data. This is why those procedures are said to be *data-driven*. A possible approach is to design resampling penalties (Arlot, 2007). Another approach is the data-driven slope heuristics of Birgé and Massart (2006), that we present now.

3.2 Slope Heuristics

3.2.1 Minimal Penalty, Optimal Penalty

Let us here recall the heuristics of Birgé and Massart (2006), which is also discussed in Arlot and Massart (2009).

From Section 3.1, the ideal penalty, which we call the *oracle penalty* since it would select the oracle is $\text{pen}^*(m) = l(s, \hat{s}_m) - \gamma_n(\hat{s}_m)$. This penalty is equivalent to pen^* as defined below since $\mathbb{E}_X\left[\gamma(s)\right] = l(s, \hat{s}_m) - \mathbb{E}_X\left[\gamma(\hat{s}_m)\right]$ does not depend on m. Remark that it also depends on $f^{\wp}$ and hence is unavailable, too.

$$\begin{aligned}\text{pen}^*(m) &= \mathbb{E}_X\left[\gamma(\hat{s}_m)\right] - \gamma_n(\hat{s}_m) \\ &= \underbrace{\mathbb{E}_X\left[\gamma(\hat{s}_m) - \gamma(s_m)\right]}_{v_m} + \underbrace{\mathbb{E}_X\left[\gamma(s_m)\right] - \gamma_n(s_m)}_{-\delta_n(s_m)} + \underbrace{\gamma_n(s_m) - \gamma_n(\hat{s}_m)}_{\hat{v}_m}, \end{aligned} \tag{3.4}$$

where v_m, $\hat{v}_m$ and δ_n are defined by the braces. Note that the AIC has been built (Akaike, 1973: see Section 2.1.2) by choosing as a penalty pen_{AIC} an approximation of the expectation (with respect to the sample) of this pen^* so as to obtain an unbiased estimator of the risk. Here our goal is to estimate this ideal penalty from the data to build a penalty which would perform almost as well as pen^*: an *optimal* penalty is a penalty which corresponding model selection procedure is optimal, from one of the viewpoints introduced in Section 3.1.2. Actually, if $\hat{m}$ is such as defined in (3.3), since for all m (those lines follow exactly those of Arlot and Massart, 2009)

$$\begin{aligned} l(s, \hat{s}_m) &= \mathbb{E}_X\left[\gamma(\hat{s}_m) - \gamma(s)\right] \\ &= \gamma_n(\hat{s}_m) + \underbrace{\mathbb{E}_X\left[\gamma(\hat{s}_m) - \gamma(s_m)\right]}_{v_m} + \underbrace{\mathbb{E}_X\left[\gamma(s_m)\right] - \gamma_n(s_m)}_{-\delta_n(s_m)} \\ &\qquad + \underbrace{\gamma_n(s_m) - \gamma_n(\hat{s}_m)}_{\hat{v}_m} - \mathbb{E}_X\left[\gamma(s)\right] \\ &= \gamma_n(\hat{s}_m) + \text{pen}^*(m) - \mathbb{E}_X\left[\gamma(s)\right], \end{aligned}$$

and from (3.3)

$$\gamma_n(\hat{s}_{\hat{m}}) + \text{pen}(\hat{m}) \leq \gamma_n(\hat{s}_m) + \text{pen}(m),$$

we get

$$l(s, \hat{s}_{\hat{m}}) + \big(\text{pen}(\hat{m}) - \text{pen}^*(\hat{m})\big) \leq \inf_{m \in M} \Big\{ l(s, \hat{s}_m) + \big(\text{pen}(m) - \text{pen}^*(m)\big) \Big\}, \qquad (3.5)$$

so that it suffices to have $\text{pen}(m)$ close to $\text{pen}^*(m)$ for any m to derive an oracle inequality. Note that, if the penalty is too heavy ($\text{pen}(m) > \text{pen}^*(m)$), an oracle inequality can still be derived from (3.5), since the risk of $\hat{s}_{\hat{m}}$ can still be upper-bounded by the minimal risk up to an additive constant. Refer to Arlot and Massart (2009) for much more details.

Minimal Penalty If the chosen penalty were $\text{pen}_{\min}(m) = \hat{v}_m$,

$$\begin{aligned} \text{crit}_{\min}(m) &= \gamma_n(\hat{s}_m) + \hat{v}_m \\ &= \gamma_n(s_m), \end{aligned}$$

which should concentrate around its expectation $\mathbb{E}_X[\gamma(s_m)]$. Hence, this procedure would select a model minimizing the bias. The variance is not taken into account: such a criterion has great probability to select a too complex model. If the penalty is chosen such that $\text{pen}(m) = \kappa \hat{v}_m$ with $\kappa < 1$, the situation becomes disastrous since then $\text{crit}(m) = (1-\kappa)\gamma_n(\hat{s}_m) + \kappa\gamma_n(s_m)$. The second term of this sum is about the bias, which is expected to be at its minimum value for the most complex models; the first one decreases as the complexity increases. This criterion is then decreasing: the selected model is for sure one of the most complex ones. But if $\text{pen}(m) = \kappa \hat{v}_m$ with $\kappa > 1$, for those most complex models for which the bias is (almost) the same, the criterion is expected to increase with the complexity (since $\text{crit}(m) = (1-\kappa)\gamma_n(\hat{s}_m) + \kappa\gamma_n(s_m)$ as well...) and the most complex models are ruled out. This suggests that $\hat{v}_m$ is a *minimal* penalty, namely that lighter penalties give rise to a selection of the most complex models, whereas higher penalties should select models with "reasonable" complexity.

Oracle Penalty: Twice Minimal Penalty The heuristics then relies on the assumption that

$$v_m \approx \hat{v}_m.$$

One reason to believe in such an assumption is that $\hat{v}_m$ is the empirical counterpart of v_m: the one is obtained from the other when the respective roles of $f^\wp$ and of the empirical measure probability $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ are reversed.

Then, since it is expected that the fluctuations of $\delta_n(s_m)$ around its zero expectation can be controlled through concentration results[3]:

$$\begin{aligned} \text{pen}^*(m) &\approx v_m + \hat{v}_m \\ &\approx 2\hat{v}_m. \end{aligned}$$

The oracle penalty is then about twice the minimal penalty, which is a fundamental point of the present heuristics.

[3] v_m and $\hat{v}_m$ are also expected to be concentrated close to their expectation, but it is not zero, and this involves a much more difficult analysis, at least concerning $\hat{v}_m$.

Oracle Penalty, Optimal Penalty As already mentioned this heuristics is useful when an optimal penalty $\text{pen}_{\text{opt}} = \kappa_{\text{opt}} \text{pen}_{\text{shape}}$ is known up to a multiplying factor: we stress that an optimal penalty is not necessarily the oracle penalty. This is a penalty which corresponding selected estimator fulfills an oracle inequality.

Some results of Massart (2007) have already been mentioned, which provide such penalties. They do not directly provide an estimation of the oracle penalty as defined in (3.4). They rather provide a penalty shape and prove that there exists a constant κ_{opt} such that the procedure based on $\kappa_{\text{opt}} \text{pen}_{\text{shape}}(m)$ almost fulfills an oracle inequality.

Arlot (2007, Chapter 6) suggests another approach which consists of directly estimating the penalty shape by resampling methods. His approach also provides a penalty shape $\text{pen}_{\text{shape}}$ which yields optimal criteria (i.e. which fulfills an oracle inequality) $\kappa_{\text{opt}} \text{pen}_{\text{shape}}$ up to a multiplying factor κ_{opt}. He suggests this constant to be chosen by applying the slope heuristics (Arlot, 2007, Chapter 6, page 165).

Finally, in the fixed design and homoscedastic regression framework, Mallows' Cp is known to be asymptotically optimal. Its penalty $\frac{2\sigma^2 D_m}{n}$ (with D_m the dimension of the model $\mathcal{M}_m$) is the expectation of the oracle penalty. It is then an optimal penalty from the asymptotic point of view. But it is known up to the variance σ^2. This last one can either directly be estimated from the data, or be considered as a constant to be chosen in front of the penalty shape $\frac{2D_m}{n}$.

In the previous three examples, an optimal penalty shape $\text{pen}_{\text{shape}}$ is known and it is also known that an optimal (or almost optimal) criterion $\kappa_{\text{opt}} \text{pen}_{\text{shape}}$ can be derived from it up to the multiplying constant κ_{opt}. The slope heuristics is to be applied to estimate this constant. Remark that the slope heuristics is derived by considering the oracle penalty, whereas it is applied to an optimal penalty shape. It is not necessarily guaranteed that the oracle penalty itself is of the shape $\kappa^* \text{pen}_{\text{shape}}$. This is a further assumption that a given optimal penalty fulfills the same property as the oracle penalty, i.e. that half this optimal penalty is a minimal penalty. Of course, the hope is that the results are fine enough so that the derived optimal penalty is very close to the oracle penalty. It can then be expected that $\frac{\kappa_{\text{opt}}}{2} \text{pen}_{\text{shape}}(m)$ is a good estimate of $\hat{v}_m$ and therefore that it is a minimal penalty, which is a keystone of the slope heuristics.

Conclusion Let us restate the two main points of the slope heuristics of Birgé and Massart (2006) we derived:

SH1 there exists a minimal penalty pen_{min} such that any lighter penalty selects models with clearly too high complexities and such that heavier penalties select models with reasonable complexity;

SH2 twice the minimal penalty is an optimal penalty.

Actually, those assumptions which derive from heuristics also receive the support of theoretical results: Birgé and Massart (2006) proved such results in a homoscedastic Gaussian regression framework. They also proved that the risk of estimators obtained with lighter penalties than the minimal one explodes. The penalty shape they derive is proportional to the dimension of the model when the considered family of models is not too large, but involves a logarithmic term when the family of models is huge.

Arlot and Massart (2009) extended those results to the heteroscedastic regression with random design framework, without assuming that the data are Gaussian. They had to restrict the considered models to histograms, but conjecture that this is only due to technical reasons and that the heuristics remains valid for the general least squares regression framework at least. They consider the case of reasonably rich families of models (namely the number of models grows as a power of n) and derive penalties proportional to the dimension. In a density estimation framework, Lerasle (2009) validates the slope heuristics (with the dimension jump approach: see below) and proves oracle inequalities for both independent (Lerasle, 2009, Chapter 2) and mixing (Lerasle, 2009, Chapter 4) data. He derives penalties proportional to the "dimension", too. Moreover the conjecture that the slope heuristics may be valid in a wider range of frameworks is supported by some partial further theoretical results and by the results of some encouraging simulations studies. Indeed, the slope heuristics was successfully applied in various model selection situations:

- estimation of oil reserves (Lepez, 2002);
- change-points detection in a Gaussian least squares framework (Lebarbier, 2005);
- selection of the number of non-zero mean components in a Gaussian framework with application to genomics (Villers, 2007);
- simultaneous variable selection and clustering in a Gaussian mixture models setting with applications to the study of oil production through curve clustering and to genomics (Maugis and Michel, 2008);
- selection of the suitable neighborhood in a Gaussian Markov random field framework (Verzelen, 2009);
- estimation of the number of interior knots in a B-spline regression model (Denis and Molinari, 2009);
- choice of a simplicial complex in the computational geometry field in Caillerie and Michel (2009);
- and the simulations we are to present in this thesis in both the frameworks of Gaussian mixture models likelihood (Section 3.3) and clustering (Section 4.4.5), some of which were already introduced in Baudry et al. (2008a).

This (probably not thorough) enumeration illustrates that the slope heuristics brings solution to real needs and the good results reported in those simulated and real data experiments contribute to confirm its usefulness. This is an enthusiastic evidence on how fruitful the efforts of Birgé, Massart and Arlot to fill in the gap between the theory of non-asymptotic model selection and the practical applications are.

Sections 3.2.2 and 3.2.3 introduce two practical approaches to apply the slope heuristics. Recall that an optimal penalty is assumed to be known up to a multiplying factor: $\mathrm{pen}_{\mathrm{opt}} = \kappa_{\mathrm{opt}}\, \mathrm{pen}_{\mathrm{shape}}(m)$. The slope heuristics' second point SH2 guarantees that this constant might be recovered as twice that of the minimal penalty, which existence is stated in point SH1. The two presented approaches differ by the way they estimate the minimal penalty.

The first one, which is the most studied and applied, is the so-called *dimension jump* method introduced in Birgé and Massart (2001), further studied in Birgé and Massart (2006) and Arlot and Massart (2009), and notably applied in Lebarbier (2005). The second one, suggested for example as a possibility in Arlot (2007), consists of directly estimating the "slope" κ_{opt} in a data-driven fashion: up to our knowledge, it was up to now applied by Baudry et al. (2008a), Maugis and Michel (2008), and more recently by Denis and Molinari (2009). We introduce it in Section 3.2.3, and discuss the difficulties encountered when applying it and the solutions we propose in a joint work with C. Maugis and B. Michel in Section 5.2. We notably propose a Matlab package for the practical use of the slope heuristics in which those solutions are embedded. Hopefully will this package contribute to a wider use of the slope heuristics!

Let us sum up what is assumed for both those methods before presenting them in Sections 3.2.2 and 3.2.3. First, the penalty of a penalized model selection criterion as described in Section 3.1.2 is known up to a constant:

$$\exists \kappa_{\text{opt}} > 0 \text{ s.t. } \operatorname{pen}_{\text{opt}}(m) = \kappa_{\text{opt}} \operatorname{pen}_{\text{shape}}(m).$$

We further assume both point SH1 and point SH2 of the slope heuristics specified above. The goal is to get from the data an estimate $\hat{\kappa}$ of κ_{opt} to design a criterion which performs almost as well as $\operatorname{pen}_{\text{opt}}$ (i.e. reaches an oracle inequality).

3.2.2 Dimension Jump

We describe the so-called *dimension jump* introduced by Birgé and Massart (2001) and further discussed in Arlot and Massart (2009). This is a practical method to take advantage of the slope heuristics derived in the previous section 3.2.1 when the purpose is to calibrate a penalty.

It has been quite successfully applied in the already mentioned practical works of Lepez (2002), Lebarbier (2005), Arlot (2007), Villers (2007), Verzelen (2009).

Let us denote D_m a measure of the complexity of each model m. D_m is typically the model dimension when it is a finite dimensional linear subspace. Remark that although mostly $\operatorname{pen}_{\text{shape}}(m) = \operatorname{pen}_{\text{shape}}(D_m)$ only depends on the complexity of the model, there is no need to assume it here. This is interesting since this does not hold in any cases: for example, the resampling penalties of Arlot (2007, Chapter 6) are not defined as functions of the models complexity.

Note that several models may have the same complexity. Then only the one minimizing $\gamma_n(\hat{s}_m)$ amongst them is of interest. The concentration arguments needed to derive the slope heuristics require the family of models to remain controlled.

Then, when considering the sequence of selected models as κ grows from zero to infinity, it is expected that:

- the complexity of the selected model with respect to κ is a non-increasing and piecewise constant function;
- models amongst the most complex are selected as κ is close to 0;

- models with "reasonable" complexities are selected as κ reaches quite great values;

so that a quite abrupt fall of the complexity of the selected model is expected as κ grows, around a value which is then chosen as $\frac{\hat{\kappa}}{2}$ (see Figure 3.1 for an illustration). This is expected to be close to $\frac{\kappa_{\text{opt}}}{2}$ and then, from SH2 of the slope heuristics:

$$\text{pen}(m) = \hat{\kappa}\, \text{pen}_{\text{shape}}(m)$$

and select

$$\hat{m} \in \underset{m \in M}{\text{argmin}} \left\{ \gamma_n(\hat{s}_m) + \hat{\kappa}\, \text{pen}_{\text{shape}}(m) \right\}.$$

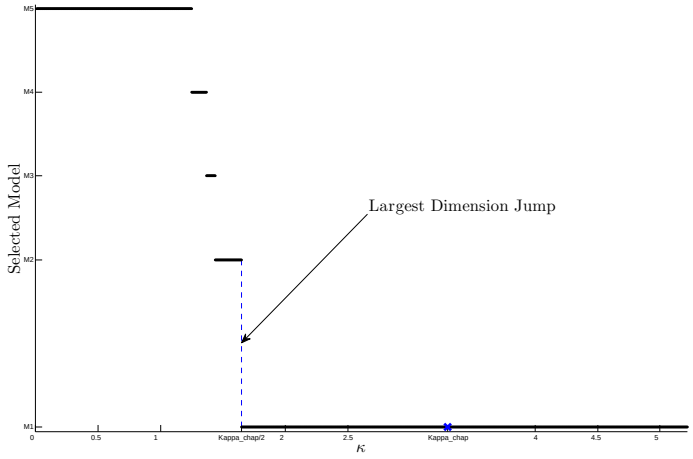


Figure 3.1: Illustration of the Dimension Jump

Arlot and Massart (2009) deduce the following algorithm from this (see Figure 3.1):

1. compute $\hat{m}(\kappa)$ with respect to $\kappa > 0$:

$$\forall \kappa > 0,\ \hat{m}(\kappa) \in \underset{m \in M}{\text{argmin}} \left\{ \gamma_n(\hat{s}_m) + \kappa\, \text{pen}_{\text{shape}}(m) \right\};$$

2. find $\hat{\kappa}$ such that $\hat{m}(\kappa)$ is among the most complex models for $\kappa < \frac{\hat{\kappa}}{2}$ and $\hat{m}(\kappa)$ have reasonable complexity for $\kappa > \frac{\hat{\kappa}}{2}$;

3. select $\hat{m} = \hat{m}(\hat{\kappa})$.

Arlot and Massart (2009) propose an algorithm which makes the first step computationally tractable since it only requires at most $(\text{card}(M) - 1)$ steps, and actually probably much less.

Of course, the definition of $\hat{\kappa}$ has to be further specified. Arlot and Massart (2009) propose two possibilities:

- Choose for $\frac{\hat{\kappa}}{2}^{\mathrm{dj}}$ ("dj" stands for "dimension jump") the value of κ corresponding to the greatest jump of dimension (recall that $D_{\hat{m}(\kappa)}$ is a piecewise constant function with respect to κ)[4]:

$$\frac{\hat{\kappa}}{2}^{\mathrm{dj}} \in \operatorname*{argmax}_{\kappa>0} \left\{ D_{\hat{m}(\kappa^-)} - D_{\hat{m}(\kappa^+)} \right\}.$$

  If several values of κ reach the maximum value, Lebarbier (2005) suggests to choose the smallest one.

- Define a complexity D_{thresh} (for "threshold") such as smaller complexities are considered as reasonable but not larger ones (Arlot and Massart (2009) suggest for example, in the regression framework they consider, to choose D_{thresh} of order $\frac{n}{\log n}$ or $\frac{n}{\log^2 n}$) and choose $\frac{\hat{\kappa}}{2}$ as the smallest value of κ which corresponding penalty selects a smaller complexity than D_{thresh}:

$$\frac{\hat{\kappa}}{2}^{\mathrm{thresh}} = \min\{\kappa > 0 : D_{\hat{m}(\kappa)} \leq D_{\mathrm{thresh}}\}.$$

Those two definitions are not equivalent. Arlot and Massart (2009) show that they should yield the same selection as the dimension jump is clear or as there are several dimension jumps close to each other, but might not otherwise. They report simulations according to which it could happen quite seldom. When the selected model is the same for both definitions — it does not really matter whether $\hat{\kappa}^{\mathrm{thresh}} = \hat{\kappa}^{\mathrm{dj}}$ or not —, the method can be confidently automatically applied. When the selected models differ — which seldom occurred —, Arlot and Massart (2009) recommend that the user looks at the graphic himself.

3.2.3 Data-driven Slope Estimation

Let us now introduce another method to take advantage of the slope heuristics so as to calibrate the penalty. This method consists of directly estimating the constant κ_{opt} by the "slope" of the expected linear relation of $-\gamma_n(\hat{s}_m)$ with respect to the penalty shape (see below). This method is being rather less employed than the "dimension jump" up to now. This might be due to some difficulties related to its application: Lebarbier (2005) partly presents this method and discusses it, but chooses the dimension jump approach notably because of the lack of stability she encountered when trying to estimate the slope. The solutions we propose to address these and so as to make possible and reliable the application of the slope heuristics through this approach are presented and discussed in Section 5.2. They arise from a joint work with C. Maugis and B. Michel: we attempted to overcome the practical difficulties to implement a Matlab package. Our aim is to provide an easy-to-use solution to anyone who would like to apply the slope heuristics. This method has already been presented and studied by Baudry et al. (2008a) and Maugis and Michel (2008).

Note that with this approach, it is not required that $\mathrm{pen}_{\mathrm{shape}}(m) = \mathrm{pen}_{\mathrm{shape}}(D_m)$ neither. And it is actually even not required to exhibit an explicit measure of the complexity

[4] With $m(\kappa^-) = \lim_{\substack{\widetilde{\kappa}\to\kappa \\ \widetilde{\kappa}<\kappa}} m(\widetilde{\kappa})$ and $m(\kappa^+) = \lim_{\substack{\widetilde{\kappa}\to\kappa \\ \widetilde{\kappa}>\kappa}} m(\widetilde{\kappa})$.

of the models (but the penalty itself, which might be considered as such a measure...), on the contrary to the previous method, since the identification of the dimension jump is based on such a measure (which might actually be the penalty itself...).

It is based on the following two equalities that we recall before identifying them:

$$\mathrm{pen}_{\mathrm{opt}}(m) = \kappa_{\mathrm{opt}}\, \mathrm{pen}_{\mathrm{shape}}(m)$$

on the one hand, and

$$\begin{aligned}\mathrm{pen}_{\mathrm{opt}}(m) &\approx 2\hat{v}_m \\ &= 2(\gamma_n(s_m) - \gamma_n(\hat{s}_m)),\end{aligned}$$

on the other hand. But we already mentioned that the bias is expected to be constant for the most complex models. And so does its estimator $\gamma_n(s_m)$, up to its fluctuations around its mean. Therefore,

$$-\gamma_n(\hat{s}_m) \approx -\gamma_n(s_m) + \frac{\kappa_{\mathrm{opt}}}{2}\, \mathrm{pen}_{\mathrm{shape}}(m)$$

is expected to behave linearly with respect to $\mathrm{pen}_{\mathrm{shape}}(m)$ (see Figure 3.2), with slope $\frac{\hat{\kappa}}{2}$ around $\frac{\kappa_{\mathrm{opt}}}{2}$. $\frac{\hat{\kappa}}{2}\, \mathrm{pen}_{\mathrm{shape}}$ is then an estimator of the minimal penalty, to be doubled from the heuristics to get an estimate of the optimal penalty.

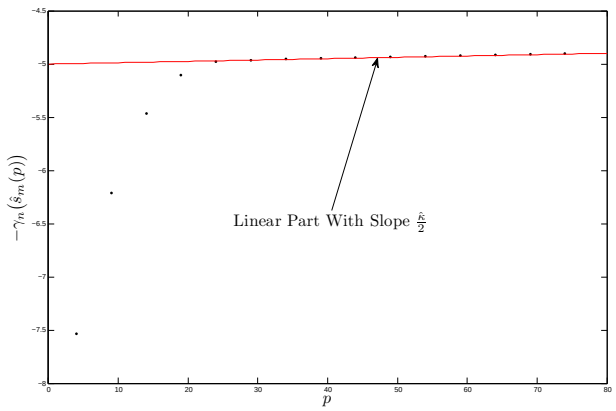


Figure 3.2: Illustration for the Data-driven Slope Estimation

The first step of the method is a validation step: it consists of verifying that the relation between $-\gamma_n(\hat{s}_m)$ and $\mathrm{pen}_{\mathrm{shape}}(m)$ is linear for the most complex models. An obviously different behavior should warn the user that the assumptions are not fulfilled and that the slope heuristics should probably not be applied. We elaborate further about this in Section 5.2: in such a situation it should first be verified that complex enough models have been involved in the study. Then, the shape of the oracle penalty should be questioned.

As this validation step confirms that the method can be applied, the method then simply amounts to choosing $\hat{\kappa}$ as twice the slope of this linear relation.

There remains a difficulty to be tackled: there may be several models reaching the same value of the penalty. In this case, define:

$$\forall p \in \{\mathrm{pen}_{\mathrm{shape}}(m) : m \in M\},\ m(p) \in \operatorname{argmin}\left\{\gamma_n(\hat{s}_m) \,\middle|\, m \in M : \mathrm{pen}_{\mathrm{shape}}(m) = p\right\}$$

(this is the only model(s) reaching this penalty value which might be selected...). Then the function

$$p \in \{\mathrm{pen}_{\mathrm{shape}}(m) : m \in M\} \longmapsto -\gamma_n(\hat{s}_{m(p)})$$

is expected to be linear for the largest values of p, with slope $\frac{\hat{\kappa}}{2}$. Remark that since $\gamma_n(\hat{s}_{m(p)}) = \min_{\{m \in M : \mathrm{pen}_{\mathrm{shape}}(m) = p\}} \gamma_n(\hat{s}_m)$, further concentration arguments are necessary to guarantee that this function is actually linear: this requires the models family not to be too rich (at least, that not too many models reach the same penalty value).

We are now in position to state the algorithm corresponding to the approach introduced in this section:

1. choose $p_{\mathrm{lin}} \in \{\mathrm{pen}_{\mathrm{shape}}(m) : m \in M\}$ such that $p \geq p_{\mathrm{lin}} \longmapsto -\gamma_n(\hat{s}_{m(p)})$ can be considered as linear;
2. estimate the slope $\frac{\hat{\kappa}}{2}$ of this linear relation;
3. select $\hat{m} = \operatorname{argmin}_{m \in M} \left\{\gamma_n(\hat{s}_m) + \hat{\kappa}\,\mathrm{pen}_{\mathrm{shape}}(m)\right\}$.

Section 5.2 will be devoted to a study on how to put this algorithm into practice. The main difficulties are related to step 1: the choice on how to identify the models for which the linear relation holds will be crucial. The results of the procedure may highly depend on a sensitive choice at this critical step.

3.2.4 Comparison Between Both Approaches

We have presented two practical approaches to take advantage of the slope heuristics of Birgé and Massart (2006) in order to calibrate at best a penalized criterion from the data. Let us sum up some points which seem relevant to compare both of them.

First of all, both procedures are data-driven: this is of course the reason to involve the slope heuristics in the choice of the penalty. This is quite different from a plug-in method, where for example an unknown variance term would be first estimated, and then plugged into a criterion which would depend on its value. This is also obviously different from a fixed penalty procedure such as AIC. The underlying idea is that the data contains information about the model minimizing the loss. When considering a fixed penalty, the fluctuations of $\gamma_n(\hat{s}_m)$ are handled as a difficulty and have to be controlled uniformly, to design the penalty, whatever their actual values. With a data-driven penalty, they are taken into account to help evaluating the best penalty in the particular case at hand. And this might improve the results. Lebarbier (2005) for example reports experiments in which the data-driven procedure applied without using the knowledge of the variance reaches smaller risks than when using an estimator of the variance or even its true value: Data-driven penalties tend to compensate the possible imperfections of the chosen penalty shape (for example, the choice of constants c_1 and c_2 in Lebarbier, 2005), which could not occur with a fixed penalty.

Both procedures allow to calibrate penalties which reach non-asymptotic results. They have been proposed in this framework since some of the available theoretical results provide a penalty shape, but not a good choice of the constant κ. The motivations of the non-asymptotic paradigm have already been discussed.

Both procedures depend on the choice of a parameter. The dimension jump, applied with the choice of the greatest jump, is sensitive to the choice of the most complex model ($D_{\max}$) included in the study (see Lebarbier, 2005). The other dimension jump approach obviously depends on the choice of D_{thresh}. The estimation of the slope depends on the way the "most complex" models, for which the relation between $\gamma_n(\hat{s}_m)$ and $\text{pen}_{\text{shape}}(m)$ is considered to be linear, are identified. When applying the method we propose in Section 5.2, this amounts to the choice of the minimal number of dimensions defining a "plateau". So that a more or less arbitrary choice has to be made for each method. This choice has to be handled with care, with respect to the particular situation at hand, since it can be crucial in some cases. Let us however stress that both the parameters ($D_{\max}$ and D_{thresh}) needed in the methods based on the dimension jump rely on a choice which seems to be tightly linked to the particular situation at hand, whereas the choice of the size of a "plateau" (in terms of percentage of the total number of considered models for example) seems to be possibly more generally defined, even if it needs some care in certain cases.

Both methods encounter difficulties in some cases. This is not surprising that such data-driven procedures may hesitate. The dimension jump may face several jumps of same order, and perhaps not close enough to each other to give rise to the same selected model. The estimation of the slope sometimes faces situations where it is really hard to identify which part of the graph of $\gamma_n(\hat{s}_{m(p)})$ is linear. And this may lead to different selected models, too.

The estimation of the slope enjoys an interesting feature, which should however be applied even when using the dimension jump: it should be verified that a linear part exists. This would reinforce the choice of the penalty shape... Now, once the slope is estimated, it suffices to add a few points corresponding to highly complex models to check they are on the same line as the previous most complex models (see Section 5.2). This would confirm — or not — that the complexity of models needed to obtain the linear behavior has been reached. This verification does not seem to be so straightforward with the dimension jump approach.

Finally, both procedures are obviously linked. Assume the linear part of the graph of $\gamma_n(\hat{s}_{m(p)})$ is actually quite linear. Then, the slope of this linear part (times the penalty shape...) is exactly the amount of penalty necessary to avoid the most complex models...and is exactly the slope which should correspond to the largest jump of dimension in the dimension jump approach in this case. Actually, the ideal situation is as both (or rather the three) procedures provide the same selected model. This model may in such a case be selected confidently. We are not able for now to provide a definite opinion about which method to prefer in general: more experiments would notably be necessary to go further in this direction.

3.3 Application to the Maximum Likelihood

We present in this section the application of the methods introduced in the previous section to the particular framework of Gaussian mixture models. The main motivation of this attempt has been to better understand the behavior of the penalized maximum likelihood criteria we introduced in Section 2.1 and particularly of the BIC and the ICL. The chosen contrast is then (minus) the log likelihood. Please refer to Section 1.2.1 for the notation about the maximum likelihood framework if needed. This section illustrates both the theoretical and the practical interests of the contrast minimization framework.

3.3.1 Contrast: Likelihood

We consider the models $\mathcal{M}_K$, for $K \in \{K_{\min}, \dots, K_{\max}\}$, as defined in (1.1) (namely Gaussian mixture models). Once more, to keep the notation simple, it is assumed that only one model is available for each dimension K, even if the results of this section could be applied with a much more complex family of models (weights may then have to be defined, according to the results of Massart (2007), and the optimal penalty may be different). Several model types could be compared for each K, and Corollary 1 notably would still hold as is, if the number of models is finite. The contrast function is imposed here by the object of our study: it is minus the log likelihood[5]:

$$\forall x \in \mathbb{R}^d, \forall K, \forall \theta \in \Theta_K, \gamma(x; f(\,.\,;\theta)) = -\log f(x;\theta).$$

The universe should be defined according to the objective: in a density estimation framework, $\mathcal{U}$ should be the set of every densities. The target is then

$$\begin{aligned} s &= \operatorname*{argmin}_{t \in \mathcal{U}} \Big\{ -\mathbb{E}_X \left[\log t(X)\right] \Big\} \\ &= \operatorname*{argmin}_{t \in \mathcal{U}} d_{\mathrm{KL}}(f^{\wp}, t) \\ &= f^{\wp}. \end{aligned}$$

The loss function is the Kullback-Leibler divergence, and the best approximation of $f^{\wp}$ in model $\mathcal{M}_K$ is

$$\theta_K = \operatorname*{argmin}_{\theta \in \Theta_K} d_{\mathrm{KL}}(f^{\wp}, f(\,.\,;\theta)),$$

which is estimated by the maximum likelihood estimator

$$\begin{aligned} \hat{\theta}_K &= \operatorname*{argmin}_{\theta \in \Theta_K} \Big\{ -\log L(\theta) \Big\} \\ &= \widehat{\theta}_K^{\mathrm{MLE}}. \end{aligned}$$

Finally, the risk of an estimator is its mean Kullback-Leibler divergence to $f^{\wp}$.

Let us highlight that it may quite be that $f^{\wp}$ belongs to none of the models $\mathcal{M}_K$ at hand. The Gaussian mixture models are however appreciated for their nice approximation properties, with many true distribution $f^{\wp}$ forms, as already discussed in Section 1.1.

[5] In this parametric setting, it will often be written $\gamma(x;\theta)$ for $\gamma(x; f(\,.\,;\theta))$ and the parameter will generally be identified with the corresponding density in the notation.

Introducing the Kullback-Leibler loss, it is clear that we have set a density estimation framework. Let us consider what kind of model selection criterion can be derived in this setting.

We shall consider penalized criteria as defined in (3.3). The question is how to choose the penalty. As a matter of fact, it is easy to conclude from the derivation of AIC (see Section 2.1.2) that it was derived by an approximation to reduce to zero the bias of $\log \mathrm{L}(\widehat{\theta}^{\mathrm{MLE}})$ as an estimator of the risk.

We attempted to define penalized criteria in this framework which would reach oracle-type inequalities. We first simply conjectured that such a criterion should have a penalty of the form[6]

$$\mathrm{pen}(K) = \alpha D_K, \tag{3.6}$$

i.e. we conjectured an optimal penalty shape should be proportional to the dimension of the model. This is quite natural since this is often the case, particularly in such cases as here, where the number of models for each dimension is low. This conjecture was also supported by the shape of the asymptotic criteria: AIC and BIC. The conjecture was also about the implicit assumption that the good measure of complexity in this framework actually is the dimension of the model so defined (namely the number of free parameters needed to describe the model). Our first experiments tended to confirm this conjecture since $\gamma_n(\hat{s}_K)$ appeared to be linear with respect to the dimension D_K for high dimensional models, for simulated experiments. See Section 3.3.3 for examples of such simulations.

Actually, since then, results of Maugis and Michel (2009) enable to derive further theoretical justification of this penalty shape. They notably computed the bracketing entropy of Gaussian mixture models (in both the cases of general and diagonal Gaussian mixture models, restricting the means μ_k in $[-\mu_{\max}; \mu_{\max}]$, the covariance matrices to have eigenvalues between $\lambda_{\min} > 0$ and $\lambda_{\max}$, and $\pi_k > 0$). This is the key to the application of Theorem 7.11[7] in Massart (2007) we already mentioned, which is the tool to define penalized criteria reaching almost oracle inequalities in the maximum likelihood framework. It introduces the bracketing entropy of the model with respect to the Hellinger distance to be the good measurement of the complexity in this situation[8]. The results of Maugis and Michel (2009) show that this this can be linked to the model dimension when considering Gaussian mixture models. They hold in a more general situation than ours since they consider both clustering and variable selection. We shall employ them to derive the penalty shape we need. Since the following result is a direct application of Theorem 7.11 of Massart (2007) and actually a particular case of the results of Maugis and Michel (2009), we shall call it a corollary. We propose a proof of it in Section 3.3.2 since it is easier to prove than in the more general situation of Maugis and Michel (2009).

Those results involve the Hellinger distance between two probability densities f and g with respect to the measure μ:

$$d_{\mathrm{hel}}(f,g) = \frac{1}{\sqrt{2}} \left\| \sqrt{f} - \sqrt{g} \right\|_2.$$

[6]Recall D_K is the "dimension" of $\mathcal{M}_K$: see Section 2.1.

[7]Recalled in Section A.1

[8]The definition of those notions as well as the results employed are recalled in Section 3.3.2

Let us state this result for a general or diagonal (i.e. whose components have diagonal covariance matrices) Gaussian mixture model family $(\mathcal{M}_K)_{K\in\{1,\dots,K_M\}}$.

Corollary 1 *Let $(\mathcal{M}_K)_{K\in\{1,\dots,K_M\}}$ be some collection of Gaussian mixture models over $\mathbb{R}^d$ such that*

$$\forall f=\sum_{k=1}^{K}\pi_k\phi(\,.\,;\mu_k,\Sigma_k)\in\mathcal{M}_K,\begin{cases}\forall k,\forall j\in\{1,\dots,d\},-\mu_{max}\leq\mu_k^j\leq\mu_{max}\\ \forall k,\text{ any eigenvalue of }\Sigma_K\text{ belongs to }[\lambda_{min};\lambda_{max}]\end{cases}$$

and let $\left(\hat{f}_K\right)_{\{1,\dots,K_M\}}$ be the corresponding family of maximum likelihood estimators.

Let $\mathrm{pen}:\{1,\dots,K_M\}\longrightarrow\mathbb{R}^+$ *such that*

$$\forall K\in\{1,\dots,K_M\},\mathrm{pen}(K)\geq\alpha\frac{D_K}{n}\left(1+\log\frac{1}{1\wedge\sqrt{\frac{D_K}{n}}}\right),$$

where $\alpha>0$ is an unknown constant depending on d, μ_{max}, λ_{min} and λ_{max}, and define the penalized log likelihood criterion crit *as*

$$\mathrm{crit}(m)=-\log L(\hat{f}_K)+\mathrm{pen}(K).$$

Then some random variable $\hat{K}$ minimizing crit *over $\{1,\dots,K_M\}$ exists and*

$$\mathbb{E}_{f^\wp}\left[d_{hel}^2(f^\wp,\hat{f}_{\hat{K}})\right]\leq C\left(\inf_{K\in\{1,\dots,K_M\}}\left(d_{KL}(f^\wp,\mathcal{M}_K)+\mathrm{pen}(K)\right)+\frac{K_M}{n}\right),\tag{3.7}$$

with $C>0$ an absolute constant.

A few remarks:

1. This result would still hold with ρ-maximum likelihood estimators, i.e. estimators such that
$$L(\hat{s}_m)\leq L(s_m)+\rho,$$
with $\rho>0$, at the price of a supplementary $C\rho$ term in the oracle inequality. This latitude is quite comforting in view of the difficulties of the definition and the computation of maximum likelihood estimators in the mixture models framework: see Section 1.2.1.

2. The main interest of this result for us is the shape of the penalty. It is almost proportional to the dimension D_K, which confirms the conjecture (3.6) about this, and strengthen the justification of the application of the slope heuristics. Only the $\log\sqrt{\frac{D_K}{n}}$ term is disappointing. We could not avoid it, exactly for the same reasons as Maugis and Michel (2009), who also get such a term. This is a consequence of the global evaluation of the bracketing entropy of the models: probably a local control, which would bound the entropy of $\{f\in\mathcal{M}_K:d_{\mathrm{hel}}(\tilde{f},f)\leq\varepsilon\}$ for any ε and a given $\tilde{f}$, may enable to avoid this term. The same techniques as in Chapter 4 may be applied. It would consists of taking advantage of the parametric

situation, and the regularity of the involved functions to derive (more easily) local bracketing entropy bounds. Corollary 4, for example, as well as most reasonings in Section 4.4, would still hold for the contrast considered here (the log likelihood). As the results reported there suggest, this $\log D_K$ term would then be expected not to appear. No oracle inequality is derived there, however: the link to Theorem 7.11 of Massart (2007) is not completed yet. This would require to assume the densities to be bounded (but we shall see in the following remark that such an assumption is necessary here, too). But it would above all yield a result with "constants" depending on quantities which dependency on the dimension of the models and the bounds on the parameters should be further studied. Maugis and Michel (2009) really counted the entropies, which is more complicated, but thus get a control of those dependencies. Anyway, ignoring this term is not a bad approximation, and it should be hardly detected in practice.

3. Actually, inequality (3.7) would be even more relevant if it linked only quantities involving Hellinger distances, instead of comparing the Hellinger distance on the one hand, to Kullback-Leibler divergences, on the other hand. As noticed in Maugis and Michel (2009), Lemma 7.23 in Massart (2007) allows a uniform control of the Kullback-Leibler divergence by the Hellinger distance, provided that $\ln \|\frac{t}{u}\|_\infty$ is bounded uniformly with respect to $t, u \in \mathcal{M}_K$ for any K. This is actually a strong assumption. First, this quantity is generally defined only if the densities are restricted to a compact subset of $\mathbb{R}^d$. Secondly, it might only be bounded uniformly over each model if the parameters are restricted to lie in compact spaces, too (this notably concerns the variances). This is an equivalent assumption (namely that the contrast is bounded) as the one which is necessary in Chapter 4.

4. Even then, (3.7) would not be an oracle inequality such as for example (3.2), since it would be necessary to control the right-hand side of the inequality by the risks of the estimators $\left(\hat{f}_K\right)_{K\in\{1,\dots,K_M\}}$.

5. Finally, this corollary provides a penalty which enables to derive an (almost) optimal procedure, in that it fulfills an (almost) oracle inequality. This is not sufficient to assess that no smaller penalty exists, which would also yield optimal procedures...

The main point to recall about this is that a penalty proportional to the dimension of the models should be (almost) optimal, but that the involved multiplying factor is unknown: (3.6) is further justified. This is then a typical situation in which the slope heuristics is needed. Let us consider simulations which illustrate how it works and to compare its results with that of the classical criteria AIC and BIC in Section 3.3.3.

3.3.2 Proof of Corollary 1

Proof (Corollary 1) *Let us denote* $H\left(\varepsilon, \sqrt{\mathcal{M}_K}\right) = \mathcal{E}_{[]}\left(\varepsilon, \sqrt{\mathcal{M}_K}, \|\cdot\|_2\right)$ *the bracketing entropy of the family of functions* $\left\{\sqrt{f} : f \in \mathcal{M}_K\right\}$ *with respect to the* $L_2(\lambda)$*-norm (see Section 4.3.2 below for the definition of the bracketing entropy). Proposition 2 and Corollary 1 in Maugis and Michel (2009) (applied in the case there is no irrelevant*

variable: $\alpha = Q$ with their notation) yield (since they define the Hellinger-distance without $\sqrt{2}$ factor, on the contrary to Massart, 2007):

$$\forall \varepsilon \in]0,1], \mathcal{E}_{[]}(\varepsilon, \mathcal{M}_K, d_{hel}) \leq \alpha D_K + D_K \log \frac{1}{\varepsilon},$$

with $\alpha = \alpha(\lambda_{min}, \lambda_{max}, \mu_{max}, d)$, for any (diagonal or general) Gaussian mixture model such that

$$\forall f = \sum_{k=1}^{K} \pi_k \phi(\,.\,; \mu_k, \Sigma_k) \in \mathcal{M}_K, \begin{cases} \forall k, \forall j \in \{1, \ldots, d\}, -\mu_{max} \leq \mu_k^j \leq \mu_{max} \\ \forall k, \textit{any eigenvalue of } \Sigma_K \textit{ belongs to } [\lambda_{min}; \lambda_{max}]. \end{cases}$$

Remark that this implies that

$$\forall \varepsilon > 1, \mathcal{E}_{[]}(\varepsilon, \mathcal{M}_K, d_{hel}) \leq \alpha D_K.$$

Then,

$$\begin{aligned} \int_0^\sigma \sqrt{H\big(u, \sqrt{\mathcal{M}_K}\big)}\, du &= \int_0^{1\wedge\sigma} \sqrt{H\big(u, \sqrt{\mathcal{M}_K}\big)}\, du + \int_{1\wedge\sigma}^{\sigma} \sqrt{H\big(u, \sqrt{\mathcal{M}_K}\big)}\, du \\ &\leq \left(\sqrt{\alpha}\sigma + \int_0^{1\wedge\sigma} \sqrt{\log \frac{1}{u}}\, du\right) \sqrt{D_K} \\ &\leq \left(\sqrt{\alpha}\sigma + \sqrt{1\wedge\sigma}\sqrt{\int_0^{1\wedge\sigma} \log \frac{1}{u}\, du}\right) \sqrt{D_K} \quad \textit{(Cauchy-Schwarz)} \\ &\leq \underbrace{\left(\sqrt{\alpha} + \sqrt{\log \frac{e}{1\wedge\sigma}}\right) \sqrt{D_K}\sigma}_{\psi_K(\sigma)}, \end{aligned}$$

The function ψ_K fulfills the properties required for the application of Theorem 7.11 of Massart (2007) (see Section A.1): it is nondecreasing and $\sigma \mapsto \frac{\psi_K(\sigma)}{\sigma}$ is nonincreasing.

Theorem 7.11 of Massart (2007) is based on the existence, for any K, of $\sigma_K > 0$ such that

$$\psi_K(\sigma_K) = \sqrt{n}\sigma_K^2 \Longleftrightarrow \sigma_K = \sqrt{\frac{D_K}{n}}\Big(\sqrt{\alpha} + \sqrt{\log \frac{e}{1\wedge\sigma_K}}\Big).$$

Let us check this existence. Remark that

$$\sqrt{\frac{D_K}{n}}\sqrt{\alpha} \leq \sqrt{\frac{D_K}{n}}\left(\sqrt{\alpha} + \sqrt{\log \frac{e}{1\wedge\Big(\sqrt{\frac{D_K}{n}}\sqrt{\alpha}\Big)}}\right)$$

always holds. And then, since the right-hand side of the equation decreases as σ increases, this implies that $\exists! \sigma_K \geq \sqrt{\alpha}\sqrt{\frac{D_K}{n}}$ such that $\psi_K(\sigma_K) = \sqrt{n}\sigma_K^2$ such that

$$\sigma_K^2 < \frac{D_K}{n}\left(2\alpha + 2\log \frac{e}{1\wedge\Big(\sqrt{\frac{D_K}{n}}\sqrt{\alpha}\Big)}\right).$$

Theorem 7.11 then applies with $\mathrm{pen}(K) \geq \alpha' \frac{D_K}{n}\left(1 + \log \frac{1}{1\wedge\Big(\sqrt{\frac{D_K}{n}}\Big)}\right)$. Let us stress that uniform weights are suitable in our setting, since there are few models per dimension. Moreover, the result is sharp (with respect to the entropy evaluation) up to the logarithm term, since $\sigma_K > \sqrt{\frac{D_K}{n}}\sqrt{\alpha}$. ■

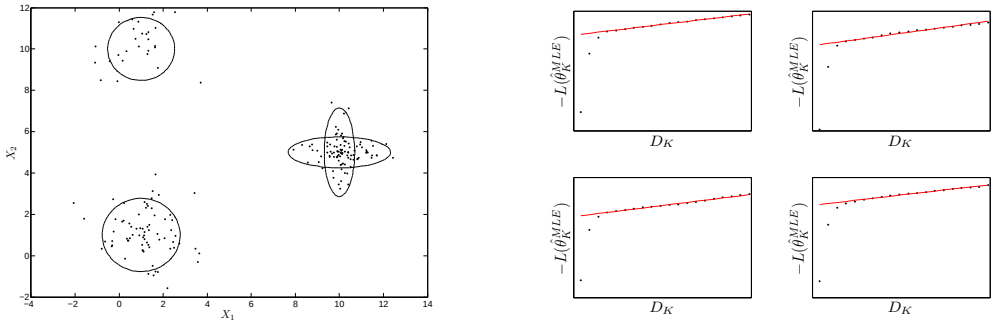
3.3.3 Simulations

The presented simulations all take place in the clustering framework with Gaussian mixtures, with one model for each considered dimension. The aim is the choice of the number of components. The contrast is here minus the log likelihood and it is minimized by the use of the EM algorithm, which is implemented in the MIXMOD software (Biernacki et al., 2006). Both the Small_EM and the Km1 initialization methods have been involved (see Sections 1.2.2 and 5.1.3)[9].

The "Cross" Experiment

We simulated data in $\mathbb{R}^d$ such as illustrated in Figure 3.3(a). The sample size is $n = 200$. The true distribution $f^{\wp}$ is a four-component diagonal Gaussian mixture: all covariance matrices are diagonal. This is an interesting example since it is quite simple and enables to highlight the differences of behavior between the different criteria we consider (see also Section 4.4.6).

We fitted to this data diagonal Gaussian mixture models, with one to twenty components.



(a) (b)

Figure 3.3: "Cross" Experiment.

An example of dataset (a) and a few examples of the graphs of $D_K \mapsto \log L(\widehat{\theta}_K^{\mathrm{MLE}})$ (b).

Such samples were simulated a hundred times.

It is interesting to have a look at some of the graphs $D_K \longmapsto \log L(\hat{\theta}_K)$ to check whether a linear part actually appears: this is Figure 3.3(b). Their seemingly appears a clear linear part. This confirms what was expected from the theoretical results (Corollary 1) and allows to make use of the slope heuristics with confidence.

In Table 3.1 are reported the results of 100 experiments with such datasets. Table 3.1 sums up the number of times each criterion selected each number of components (which is here to be chosen as the number of classes).

[9]Details on the simulation settings and the applied algorithms may be found in Section A.2.

Selected Number of Components	2	3	4	5	6	7	8	9	10–20
AIC	0	0	1	1	2	2	3	3	88
BIC	0	4	91	5	0	0	0	0	0
Slope Heuristics	0	2	84	10	3	0	0	0	1

Table 3.1: "Cross" Experiment Results.

According to those results, BIC behaves very well in such a situation: it mostly recovers the true model. This is a well-known behavior of BIC. The interesting fact is that the slope heuristics-based criterion behaves analogously as BIC, with a tendency to rather overestimate the number of components. The efficiency results (Table 3.2) confirm that AIC does not provide a heavy enough penalty in this mixture framework: it would be expected to be a good criterion from an efficiency point of view, but it appears that it is much worse than the slope heuristics, or even than BIC in this context.

	Risk of the criterion $\times 10^3$	$\frac{\text{Risk of the criterion}}{\text{Risk of the oracle}}$
Oracle	59	1
AIC	506	8.03
BIC	65	1.10
Slope Heuristics	69	1.17

Table 3.2: "Cross" Experiment Results.

Risk of each criterion in terms of Kullback-Leibler divergence to the true distribution, estimated by Monte Carlo simulations. The oracle results reported in the table correspond to the trajectory oracle

$$K_{\text{oracle}} = \underset{1 \leq K \leq 20}{\operatorname{argmin}}\, d_{\text{KL}}(f^{\wp}, f(\,.\,; \widehat{\theta}_K^{\text{MLE}}))$$

for each dataset. The expected oracle number of components

$$K_{\text{oracle}} = \underset{1 \leq K \leq 20}{\operatorname{argmin}}\, \mathbb{E}_{f^{\wp}} \left[d_{\text{KL}}(f^{\wp}, f(\,.\,; \widehat{\theta}_K^{\text{MLE}})) \right]$$

is four (see Figure 3.4). The true number of components is four.

The Misspecified Models Experiment

This experiment has been designed to highlight the behavior of the considered model selection criteria in a misspecified models situation, namely none of the considered models contains the true distribution. Indeed $f^{\wp}$ is a general Gaussian mixture with two non-diagonal components: see Figure 3.5 (a). The data size is $n = 200$.

But the fitted models are still diagonal Gaussian mixture models. So that the bias is zero in none of the considered models, and it must be quite far from zero.

Once more, we wish to check that a linear part appears in the graph of $D_K \mapsto$

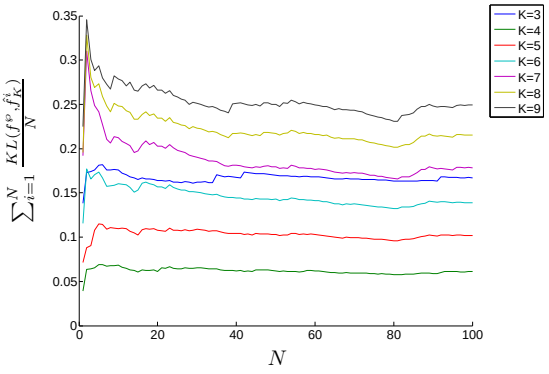
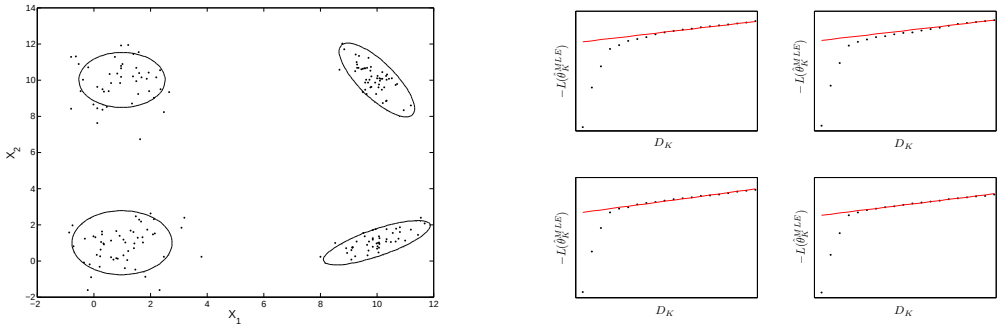


Figure 3.4: "Cross" Experiment.

Convergence of the Monte Carlo Simulations for the Computation of $K_{\text{oracle}} = \underset{1 \leq K \leq 20}{\text{argmin}} \mathbb{E}_{f^{\wp}} \left[d_{\text{KL}}(f^{\wp}, f(\,.\,; \widehat{\theta}_K^{\text{MLE}})) \right]$



(a) (b)

Figure 3.5: Misspecified Models Experiment.

A dataset example (a) and a few examples of the graphs of $D_K \mapsto \log L(\widehat{\theta}_K^{\text{MLE}})$ (b).

$\log L(\hat{\theta}_K)$: Figure 3.5 (b) shows it seems to be the case.

Results are reported in Table 3.3.

Selected number of components	4	5	6	7	8	9–16	17	18	19	20
Oracle	4	10	30	43	12	1	0	0	0	0
AIC	0	0	0	0	0	20	14	12	26	28
BIC	3	43	38	13	3	0	0	0	0	0
Slope Heuristics	2	19	26	32	11	10	0	0	0	0

Table 3.3: Misspecified Models Experiment Results.

This experiment confirms that BIC must recover the "true" number of components: since $f^{\wp}$ is not a diagonal mixture, more than four diagonal components are needed to correctly approximate it and BIC rather selects five or six components (see Table 3.3).

From Table 3.3, the slope heuristics based criterion behaves like the oracle: it mostly selects a great number of components, which are necessary to get the best possible approximation of the data distribution. However, it achieves risk results a little worse than BIC (see Table 3.4): both are good, as compared to the oracle.

Figure 3.6 illustrates the difficulty of the problem in this setting: it is not clear what value of K should be chosen as the oracle, even as Monte Carlo simulations are available.

	Risk of the criterion $\times 10^3$	$\frac{\text{Risk of the criterion}}{\text{Risk of the oracle}}$
Oracle	206	1
AIC	712	3.45
BIC	240	1.16
Slope Heuristics	249	1.21

Table 3.4: Misspecified Models Experiment Results.

Risk of each criterion in terms of Kullback-Leibler divergence to the true distribution, estimated by Monte Carlo simulations. The oracle results reported in the table correspond to the trajectory oracle:

$$K_{\text{oracle}} = \underset{1 \leq K \leq 20}{\operatorname{argmin}}\, d_{\text{KL}}(f^{\wp}, f(\,.\,; \widehat{\theta}_K^{\text{MLE}}))$$

for each dataset. The expected oracle number of components:

$$K_{\text{oracle}} = \underset{1 \leq K \leq 20}{\operatorname{argmin}}\, \mathbb{E}_{f^{\wp}} \left[d_{\text{KL}}(f^{\wp}, f(\,.\,; \widehat{\theta}_K^{\text{MLE}})) \right]$$

is six or seven (see Figure 3.6). The "true" number of components is four, but the model is misspecified.

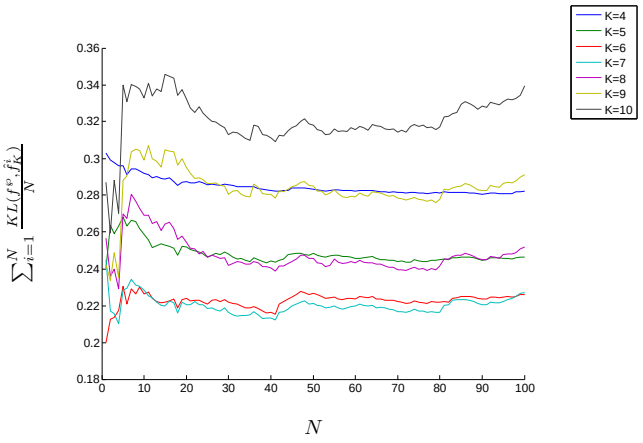


Figure 3.6: Misspecified Model Experiment.

Convergence of the Monte Carlo simulations for the computation of $K_{\text{oracle}} = \underset{1 \leq K \leq 20}{\operatorname{argmin}} \mathbb{E}_{f^{\wp}} \left[d_{\text{KL}}(f^{\wp}, f(\,.\,; \widehat{\theta}_K^{\text{MLE}})) \right]$.

Conclusion

We introduced in this chapter the contrast minimization framework, and how it is convenient to study model selection procedures: the application to the likelihood contrast in the Gaussian mixture model is interesting for our main purpose, which is choosing the number of components. It enabled to derive some optimal model selection procedures, but which may depend on unknown constants. The slope heuristics of Birgé and Massart (2006) seems to be a powerful practical tool to recover those constants.

We shall recast the problem of choosing the number of classes with a clustering purpose in the contrast minimization framework, with a convenient constrast, in Chapter 4. This shall yield a new criterion and shed new light on the ICL criterion. Moreover, the slope heuristics may then be applied in this framework, and the first results reported in Chapter 4 suggest a rather relevant behavior of the corresponding criterion.

Finally, we noticed the practical difficulties encountered when practically applying the slope heuristics. Jointly with C. Maugis and B. Michel, we propose solutions to overcome those difficulties with the "estimation of the slope" approach. This work is reported in Section 5.2. It was mainly motivated by the design of a software — actually a Matlab package — that we wish to make available to enable a wider use of the slope heuristics.

Chapter 4

Estimation and Model Selection for Model-Based Clustering with the Conditional Classification Likelihood

Contents

What does ICL do? This chapter is dedicated to an attempt to answer the question which has motivated this thesis: better understanding the ICL criterion.

The ICL criterion has been proposed by Biernacki et al. (2000) as a criterion to select the number of classes with Gaussian mixture models for a clustering purpose. Better than other classical criteria, ICL seems to yield a relevant number of classes for a clustering purpose. Since it "penalizes" the solutions with large overlap, it does not tend to overestimate the number of classes, on the contrary to BIC, for example. This nice feature, first mainly studied through simulations, has met the interest of statisticians in various applications.

However, almost no theoretical results about ICL were available. It was rather criticized for not being "consistent", in the sense that BIC is: it does not recover the true number of components of a mixture, even if the true distribution is available in one of the considered models, and even asymptotically, if the true components overlap.

There seemed to be a gap between the practical interest of ICL and the available theoretical analysis, which was rather negative. It is then interesting to try to understand what people like, in a intuitive manner, with the hope that it matches a notion of what a class should be.

Actually, we embraced several approaches to try to better understand the behavior of ICL. The contrast minimization framework proved to be the most fruitful. It enabled to fully understand that ICL was not a penalized likelihood criterion, as opposed to the usual point of view. It is rather linked to another contrast: the conditional classification likelihood. Involving this quantity in a clustering purpose is definitely not a new idea: Biernacki and Govaert (1997) for example, considered it as a model selection criterion for its own, and it is actually a part of the ICL itself. But it had never been considered as a contrast, i.e. a concurrent to the likelihood itself: mostly, the maximum likelihood estimator is plugged-in. Doing so much improves the understanding of ICL, which is a penalized conditional classification likelihood criterion.

Considering this contrast enables to derive theoretical results about a new penalized criterion which is almost ICL. Indeed, it is the same function, with the same penalty, but evaluated at a different estimator. This estimator, to be consistent with the ICL point of view, is the maximum conditional classification likelihood estimator. So that both a new criterion and a new estimator, which purpose is clustering, are introduced. The estimator is proved to be consistent under usual regularity conditions, and a result on sufficient conditions about the penalty which guarantee the consistency of the corresponding model selection procedure is provided. ICL is a consistent criterion, from this point of view. The embraced approach for doing so is mimicked from the techniques of Massart (2007) in a non-asymptotic general likelihood framework. Although no oracle inequality is derived here, this gives a hint about the optimal penalties shapes. This partly justifies the application of the slope heuristics to derive another criterion.

An interest of this approach, is that it enables to further study the notion of class underlying ICL. This is nor a simple notion of cluster — as for example for the k-means procedure — neither a pure notion of "component" — as underlying the MLE/BIC approach in this framework — but a compromise between both. Remark that in this chapter, the "one component=one class" rule is followed.

This notion of class underlying ICL seems to match a widespread intuitive notion of

what it should be. Simulations strengthen those considerations.

In this chapter, sketches of proofs are mostly given to help understand how to bring results together to get the final result, but for the sake of readability the most technical parts of the proofs are given in sections apart.

4.1 Introduction

The notions presented in this section have mainly been introduced in the three first chapters already. They are quickly recalled so that the reader may simply refer to the corresponding sections, which are specified each time it is necessary.

Let an i.i.d. sample $X_1, \ldots, X_n$ from an unknown distribution $f^{\wp}$.

4.1.1 Gaussian Mixture Models

$\mathcal{M}_K$ is the Gaussian mixture model with K components:

$$\mathcal{M}_K = \left\{ \sum_{k=1}^{K} \pi_k \phi(\,.\,;\omega_k) \,\middle|\, (\pi_1, \ldots, \pi_K) \in \Pi_K, (\omega_1, \ldots, \omega_K) \in \Theta_K \right\}, \tag{4.1}$$

which entire parameter space is $\Theta_K = \Pi_K \times \left(\mathbb{R}^d \times \mathbb{S}_+^d\right)^K$ (see Section 1.1.1). Recall that $\Pi_K = \{(\pi_1, \ldots, \pi_K) \in [0,1]^K : \sum_{k=1}^K \pi_k = 1\}$. Constraints on the model may be imposed by restricting Θ_K. The type of constraints which are typically considered here follow the decomposition suggested by Celeux and Govaert (1995): the covariance matrices are written[1] $\Sigma_k = \lambda_k D_k A_k D_k'$, and each one of the three factors of the decomposition (corresponding respectively to the component volume, direction and shape), may be constrained to be equal or not for all components, as well as the mixing proportions may be constrained to be equal or not. See Section 1.1.1, where other kinds of constraints are suggested, based on this decomposition.

Those are studied here as parametric models. It is then assumed the existence of a parametrization $\varphi : \Theta_K \subset \mathbb{R}^{D_K} \to \mathcal{M}_K$. It is assumed that Θ_K and φ are "optimal", in the sense that D_K is minimal. D_K is then the number of *free parameters* in model $\mathcal{M}_K$ and is called the *dimension* of $\mathcal{M}_\mathcal{K}$. For example, only $(K-1)$ mixing proportions have to be parametrized, since the K^{th} proportion can be deduced from them.

It shall not be needed to assume the parametrization to be identifiable, i.e. that φ is injective. Our purpose in the following is twofold: identifying a relevant number of classes to be designed; and actually design those classes. Theorem 7 below justifies that the first task can be achieved under a weak identifiability assumption, namely that the loss function is minimized at a single value of the conditional classification likelihood, as a function of x. This single value could be represented by several values of the parameter. Then, the estimation theorem (Theorem 4) guarantees that the defined

[1] See Section 1.1.1 for the $\Sigma_k = \lambda_k D_k A_k D_k'$ decomposition: $\lambda_k = \left(\det \Sigma_k\right)^{\frac{1}{d}}$, A_k is the diagonal matrix with the eigenvalues of Σ_k (divided by λ_k) on the diagonal and D_k is the (orthogonal) matrix of eigenvectors of Σ_k.

estimator converges to one of the best parameters from the embraced point of view. The classes can finally be defined through the MAP rule (which is recalled a few lines below), from the estimated parameter. The injectivity of φ is never involved. Identifiability of the mixture models is not the good notion here: two parameters defining the same mixture distribution may even be different from the conditional classification point of view (think of the example: $\phi(\,.\,;\omega) = \pi\phi(\,.\,;\omega)+(1-\pi)\phi(\,.\,;\omega)$, where the first parameter yields one class, whereas the second yields two classes). There is therefore no reason to identify them. Label switching may be prevented without damage, but it is not needed to do so neither. Writing that this chapter takes place in the Gaussian mixture model framework is then — conveniently — abusing a little the definitions. The conditional classification likelihood — as the ICL criterion besides — is not even defined over the models $\mathcal{M}_K$ if they are not assumed to be identifiable (the knowledge of each component distribution is needed to define the entropy: see the remark about $\widetilde{\mathcal{M}}_K$ in Section 1.1.1). Recall (from Section 1.1.2) that the identifiability of $\mathcal{M}_K$ may be guaranteed under the conditions of Yakowitz and Spragins (1968) that $\pi_k > 0$ and $\omega_k \neq \omega_{k'}$ as soon as $k \neq k'$ (see Section 1.1.2). Since we shall assume the parameter spaces to be compact, those conditions are quite unpleasant: lower bounds on the π_k's and on $\|\omega_k - \omega_{k'}\|$ have to be specified. But we need not such identifiability assumption and the $\omega_k \neq \omega_{k'}$ condition is avoided. Mixture distributions which do not fulfill this condition are never of interest anyway, because of the considered contrast. But the technical condition on the proportions seems difficult to be avoided. The reason shall be apparent from the study of the entropy term in Section 4.2.2 and be further discussed in Section 4.3.2.

Let us introduce two examples of usual Gaussian mixture models, which will be of concern subsequently. Since compactness assumption will be needed, it is described what conditions are sufficient to guarantee this assumption to hold.

Example 2 (General Gaussian Mixture Model)
No constraint is imposed on the Gaussian mixtures. Mixing proportions are allowed to be different, and each covariance matrix may be any positive definite symmetric matrix. Θ_K *is a subset of* $\Pi_K \times \left(\mathbb{R}^d \times \mathbb{S}_+^d\right)^K$ *and may be imposed for example to be compact with constraints like:*
$\forall k \in \{1, \ldots, K\}$,

$$\begin{gathered}
\pi_k \geq \pi_{min} \\
\forall j \in \{1, \ldots, d\}, \mu_{min} \leq \mu_k^j \leq \mu_{max} \\
\lambda_{min} \leq \lambda_k \leq \lambda_{max} \\
\forall j \in \{1, \ldots, d\}, a_{min} \leq A_k^j \leq a_{max},
\end{gathered}$$

where A_k *is the diagonal matrix with diagonal* $(A_k^1, \ldots, A_k^d)$*. This model has dimension* $(d-1) + Kd + K\frac{d(d+1)}{2} = K\frac{d(d+3)}{2} + d - 1$ *and a corresponding parametrization is given in Section 4.3.3.*

Example 3 (Diagonal Gaussian Mixture Model)
The covariance matrices are imposed to be diagonal. This corresponds to Gaussian components which are parallel to the axis. Θ_K *is a subset of* $\Pi_K \times \left(\mathbb{R}^d \times \mathbb{R}^d\right)^K$*, and may*

be imposed to be compact with: $\forall k \in \{1, \ldots, K\}$,

$$\begin{gathered}\pi_k \geq \pi_{min} \\ \forall j \in \{1, \ldots, d\}, \mu_{min} \leq \mu_k^j \leq \mu_{max} \\ \forall j \in \{1, \ldots, d\}, a_{min} \leq A_k^j \leq a_{max}.\end{gathered}$$

where $\lambda_k A_k$ *is the diagonal matrix with diagonal* $(A_k^1, \ldots, A_k^d)$ *(those eigenvalues here are then not normalized by the determinant). This model has dimension* $(d-1)+Kd+Kd = 2Kd + d - 1$ *and a corresponding parametrization is given in Section 4.3.3.*

This chapter is uniquely devoted to the question of clustering through Gaussian mixture models. Though all results could be extended to other mixture models of exponential families.

The adopted process is the usual one (Chapter 2):

- fit each considered mixture model;
- select a model and a number of components on the basis of the first step;
- classify the observations through the MAP rule (recalled below) with respect to the mixture distribution fitted in the selected model.

Notably, the usual choice is made in this chapter, to identify a class with each fitted Gaussian component. The number of classes to be designed is then chosen at the second step. See also Chapter 2 for a discussion upon this choice.

Let us recall the MAP classification rule (see Section 1.3). It involves the conditional probabilities of the components

$$\forall \theta \in \Theta_K, \forall k, \forall x,\ \tau_k(x;\theta) = \frac{\pi_k \phi(x;\omega_k)}{\sum_{k'=1}^K \pi_{k'} \phi(x;\omega_{k'})}. \qquad \text{((1.5) recalled)}$$

$\tau_k(X;\theta)$ is the probability that the observation X arose from the k^{th} component, conditionally to X, under the distribution defined by θ (i.e. $\tau_k(x;\theta)$ is the *conditional* probability of component k under θ as x has been observed)[2]. The MAP classification rule is then defined by

$$\widehat{z}^{\text{MAP}}(\theta) = \underset{k=1,\ldots,K}{\operatorname{argmax}}\, \tau_k(x;\theta).$$

The usual maximum likelihood estimator (with respect to the sample $\mathbf{x}$) in model $\mathcal{M}_K$ is written $\widehat{\theta}_K^{\text{MLE}}$.

4.1.2 ICL

The motivation of the works reported in this chapter was to better understand, mainly from a theoretical point of view, the ICL model selection criterion. Let us

[2]It will also be written $\tau_{ik}(\theta)$ as $x = x_i$.

first recall how it was originally derived (see Section 2.1.4). Attempting to mimic the derivation of the BIC criterion (Section 2.1.3) in a clustering framework, Biernacki et al. (2000) approximate the integrated classification likelihood (Section 1.4: $\mathrm{L_c}(\theta;(\mathbf{x},\mathbf{z}))=\prod_{i=1}^{n}\prod_{k=1}^{K}\left(\pi_k\phi(x_i;\omega_k)\right)^{z_{ik}}$, where $\mathbf{z}$ are the unobserved labels, specifying the component from which the corresponding observations arose, when $f^{\wp}$ is actually a mixture distribution) through a Laplace's approximation. Further, they replace the unobserved z_{ik}'s by their MAP estimate under $\widehat{\theta}_K^{\mathrm{MLE}}$ (namely[3] $\widehat{z}_i^{\mathrm{MAP}}(\widehat{\theta}_K^{\mathrm{MLE}})$), and assume that the mode of the classification likelihood can be identified with the maximum (observed) likelihood estimator as n is large enough, which is a questionable choice which shall be discussed in this chapter. They obtain the ICL criterion:

$$\begin{aligned}\mathrm{crit_{ICL}}(K) &= \log f(\mathbf{X},\widehat{\mathbf{Z}}^{\mathrm{MAP}};\widehat{\theta}_K^{\mathrm{MLE}})-\frac{\log n}{2}D_K\\ &= \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}})+\sum_{i=1}^{n}\sum_{k=1}^{K}\widehat{Z}_{i,k}^{\mathrm{MAP}}(\widehat{\theta}_K^{\mathrm{MLE}})\log\tau_{ik}(\widehat{\theta}_K^{\mathrm{MLE}})-\frac{\log n}{2}D_K.\end{aligned}$$

As mentioned in Section 2.1.4, McLachlan and Peel (2000) rather replace the unobserved z_{ik}'s by their posterior respective probabilities with respect to the maximum likelihood estimator $\tau_{ik}(\widehat{\theta}_K^{\mathrm{MLE}})$:

$$\mathrm{crit_{ICL}}(K)=\mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}})+\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}(\widehat{\theta}_K^{\mathrm{MLE}})\log\tau_{ik}(\widehat{\theta}_K^{\mathrm{MLE}})-\frac{\log n}{2}D_K. \tag{4.2}$$

Both those versions of the criterion appear to behave analogously, and the latter is now considered.

The ICL differs from the classical and widely used BIC criterion of Schwarz (1978) through the *entropy* term (see Section 1.4.3 and Section 4.2.2 below):

$$\forall\theta\in\Theta_K,\ \mathrm{ENT}(\theta;\mathbf{x})=-\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}(\theta)\log\tau_{ik}(\theta). \qquad \text{((1.7) recalled)}$$

The BIC criterion is known to be consistent for the number of components, at least when the true distribution is actually a (Gaussian, here) mixture distribution and lies in one of the considered models (Keribin (2000) or Nishii (1988): see Section 2.1.3). This nice property may however not be adapted for a clustering purpose. In many applications, there is no reason to assume that the classes to be designed have a Gaussian shape. The BIC in this case tends to overestimate the number of components since several Gaussian components may be needed to approximate each non-Gaussian component of the true mixture distribution $f^{\wp}$. And the user may rather be interested in a cluster notion — as opposed to this strictly "component" approach — which also includes a separation notion and which be quite robust to non-Gaussian components. Of course, it depends on the application needs, and on the idea that the user has of a class (see Hennig, 2009 for a discussion about the notion of cluster). It may be of interest to discriminate into two different classes a group of observations which the best fit is reached with a mixture of two Gaussian components having quite different parameters

[3] By abusing the notation, we identify the $0-1$ label vector with the index of the corresponding component: $''z=k_0''\Leftrightarrow z_k=1$ if $k=k_0$ and $z_k=0$ else.

(we particularly think of the covariance matrices parameters). BIC tends to do so. But it may also be more relevant, at least for many applications, and it may conform to a current intuitive notion of cluster, which should however be specified, to identify two very close — or largely overlapping — Gaussian components as a single non-Gaussian shaped cluster (see for example Figure 4.3)...

The ICL has been derived with this viewpoint, to overcome this limitation of the BIC. It is widely understood and explained (for instance in Biernacki et al., 2000) as the BIC criterion with a further penalization term, which is the entropy. Since this entropy term penalizes models which maximum likelihood estimator yields an uncertain MAP classification (see Section 4.2.2 below), the ICL is expected to be more robust than BIC to non-Gaussian components. Though, we do not think that the entropy should be considered as a penalty term and an other point of view about ICL will be developed in this chapter.

This nice feature of the ICL criterion has been studied and confirmed through simulations and real data studies by Biernacki et al. (2000), McLachlan and Peel (2000, Section 6.11), and in several simulation studies that we have performed, some of which are reported in Section 4.4 below. Besides, it has met the needs of some applications and several authors successfully chose to use it for the mentioned reasons in various applications area: among others,

- Goutte et al. (2001) use it in a study of fMRI images;
- Hamelryck et al. (2006) for the problem of predicting a protein structure from its sequence;
- Pigeau and Gelgon (2005) introduce ICL for a method which aims at building an hierarchical structure on an image collection;
- De Granville et al. (2006) introduce ICL in an approach which purpose is the learning by robots of the possible grasps they can apply to an object, based on the sight of a human manipulating it;
- etc.

This practical interest for the ICL lets us think that it meets an interesting notion of cluster, corresponding to what — at least — some users expect. But almost no theoretical studies nor results are available about ICL. This has been the main motivation of the work reported in this chapter — and actually of the whole thesis — to go further in this direction. It will notably be seen in the following how it led to considering a new contrast in the contrast minimization framework, and hence to a new procedure to fit a mixture model and to a new model selection criterion for clustering, similar to ICL but for which the development of the underlying logic is driven to its conclusion, from the estimation step to the model selection step, instead of introducing the maximum likelihood estimator like Biernacki et al. (2000) did. Moreover, the presented point of view will enable to derive theoretical results about ICL, which the maximum likelihood framework did not enable. Typically, ICL is sometimes criticized for not being "consistent" (for the number of components), on the contrary to BIC, in "good" situations. It will be shown that, as a matter of fact, it is, in a sense to be specified.

4.2 A New Contrast: Conditional Classification Likelihood

4.2.1 Definition, Origin

We already introduced the classification likelihood (see (1.4) in Section 1.4.1):

$$\forall \theta \in \Theta_K,\ \mathrm{L_c}\big(\theta; (x_1, z_1), \ldots, (x_n, z_n)\big) = \prod_{i=1}^{n} \prod_{k=1}^{K} \big(\pi_k \phi(x_i; \omega_k)\big)^{z_{ik}}. \qquad \text{((1.4) recalled)}$$

This is the likelihood of the complete data $(\mathbf{x}, \mathbf{z})$. It has been seen in Section 2.1.4 that it was involved in the initial derivation of the ICL, possibly (in the version of McLachlan and Peel, 2000) giving rise to a term $\log L(\widehat{\theta}_K^{\mathrm{MLE}}) - \mathrm{ENT}(\widehat{\theta}_K^{\mathrm{MLE}})$. We shall here derive it from a slightly different point of view. This will shed new light on its link with clustering and be a starting point for the subsequent theoretical study.

Of course, in our clustering context, neither the labels $\mathbf{z}$ are observed, nor we assume that they even exist (think of the case several models with different number of components are fitted: then at most one can correspond to the true number of classes, when it exists). A possibility to involve the classification likelihood even though is to consider its conditional expectation with respect to the observations $\mathbf{x}$. In the case there exists a true classification and a model with the true number of classes is considered, this conditional expectation may actually be interpreted as the quantity the closest to the classification likelihood, which can be obtained given the available information.

Recall the fundamental equality (1.6) (Section 1.4.2)

$$\forall \theta \in \Theta_K,\ \log \mathrm{L_c}(\theta) = \log \mathrm{L}(\theta) + \sum_{i=1}^{n} \sum_{k=1}^{K} z_i^k \log \tau_k(x_i; \theta). \qquad \text{((1.6) recalled)}$$

Denoting the conditional expectation of the classification log likelihood $\log \mathrm{L_{cc}}(\theta) = \log \mathrm{L_{cc}}(\theta; \mathbf{x})$ (for Conditional Classification log Likelihood)[4],

$$\begin{aligned}
\log \mathrm{L_{cc}}(\theta) &= \mathbb{E}_\theta \left[\log \mathrm{L_c}(\theta) | \mathbf{X} = \mathbf{x}\right] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}(\theta) \log \pi_k \phi(x_i; \omega_k) \\
&= \log \mathrm{L}(\theta) + \underbrace{\sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}(\theta) \log \tau_{ik}(\theta)}_{-\mathrm{ENT}(\theta; \mathbf{x})}.
\end{aligned}$$

This quantity, deriving from the classification likelihood, is obviously linked to our clustering objective, and it is tempting to involve it in our study. Let us however highlight that it is not completely justified by the previous heuristics. Notably, it would be more relevant to consider the conditional expectation with respect to $f^{\wp}$ instead of θ...of course, it is unknown, and a first attempt may be to plug the maximum likelihood

[4] Of course, in the following, $\mathbb{E}_\theta$ stands for $\mathbb{E}_{f(\,.\,;\theta)}$

estimator. The presented derivation of the conditional classification likelihood is nevertheless interesting since it highlights its link with the clustering. And the reason why we introduce this quantity is that it is related to ICL. We will derive a model selection criterion directly by considering the conditional classification likelihood as a contrast (following the contrast minimization approach introduced in Chapter 3). We will then be able to derive some theoretical properties of this new criterion and of the related estimator, which will have been defined first. Yet, we will link this criterion to ICL and then deduce a better understanding of the ICL criterion, which is our main purpose.

4.2.2 Entropy

The conditional classification log likelihood differs from the usual log likelihood through the entropy term. It is then necessary to further study the entropy.

$$\forall \theta \in \Theta_K,\ \mathrm{ENT}(\theta; \mathbf{x}) = -\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}(\theta) \log \tau_{ik}(\theta). \qquad \text{((1.7) recalled)}$$

The behavior of the entropy is based on the properties of the function

$$h : t \in [0,1] \longmapsto \begin{cases} -t \log t & \text{if } t > 0 \\ 0 & \text{if } t = 0. \end{cases}$$

This nonnegative function (see Figure 4.1) takes zero value if and only if $t = 0$ or $t = 1$.

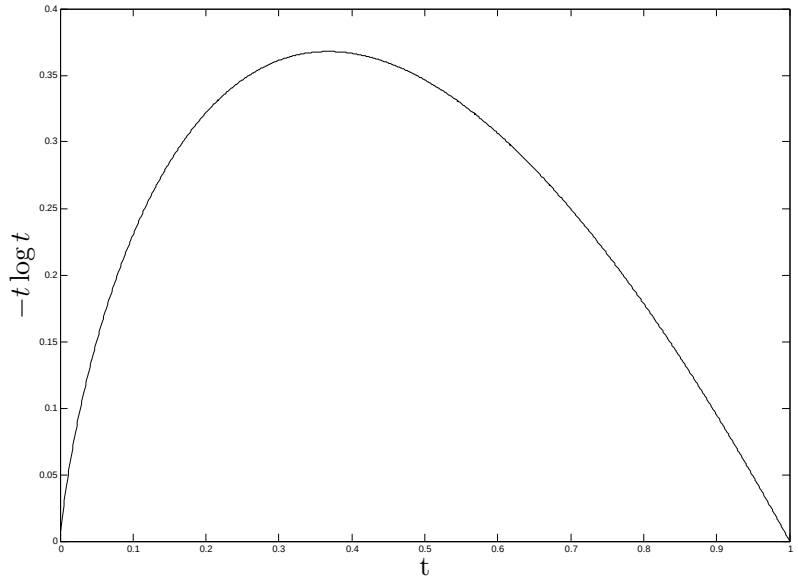


Figure 4.1: The function $h : t \mapsto -t \log t$

It is continuous but not differentiable at 0 ($h'(t) \xrightarrow[t \to 0^+]{} +\infty$), and in particular it is not

Lipschitz over $[0,1]$, which will be a cause of analysis difficulties. Let us also introduce the function

$$h_K : (t_1, \ldots, t_K) \in \Pi_K \longmapsto \sum_{k=1}^{K} h(t_k).$$

This nonnegative function (see Figure 4.2) then takes zero value if and only if there

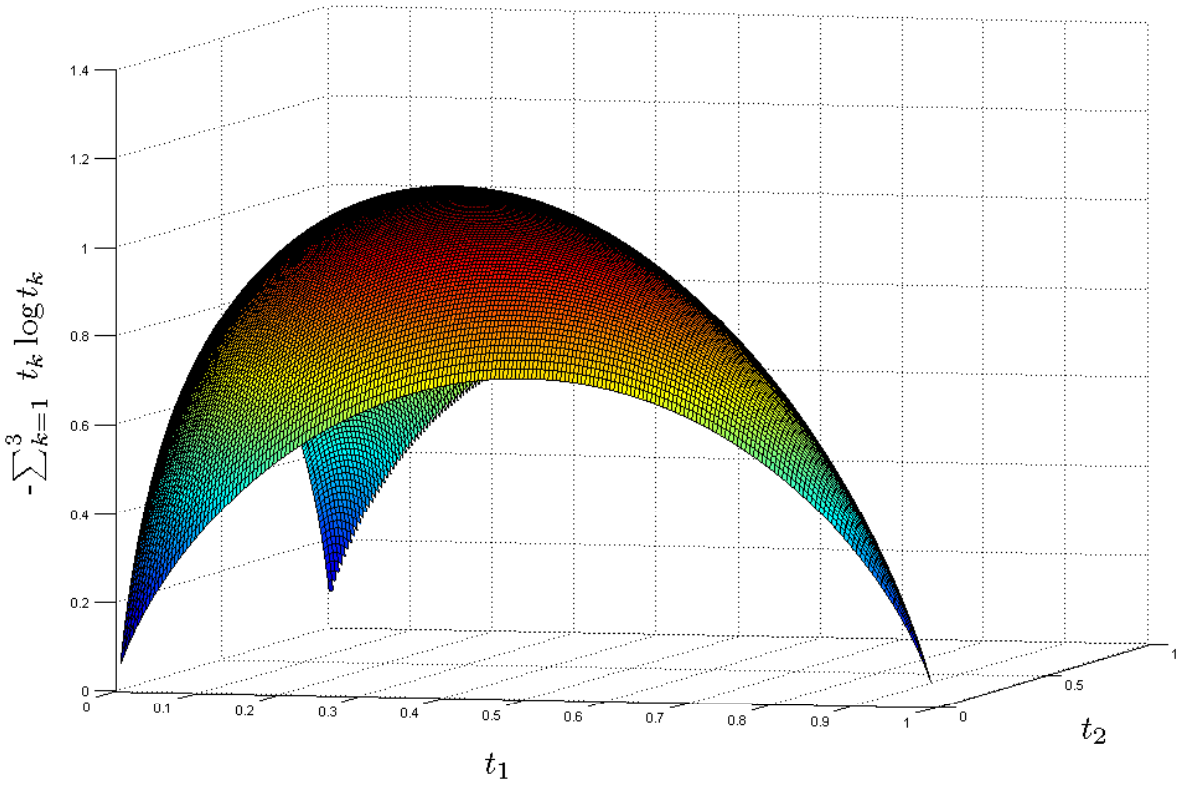


Figure 4.2: The function $h_K : (t_1, \ldots, t_K) \mapsto \sum_{k=1}^{K} h(t_k)$ when $K = 3$ (in Π_3, $t_3 = 1 - t_1 - t_2$)

exists $k_0 \in \{1, \ldots, K\}$ such that $t_{k_0} = 1$ and $t_k = 0$ for $k \neq k_0$. It reaches its maximum value $\log K$ at $(t_1, \ldots, t_K) = (\frac{1}{K}, \ldots, \frac{1}{K})$.

Proof *If h_K reaches a maximum value at $(t_1^0, \ldots, t_K^0)$ under the constraint $\sum_{k=1}^{K} t_k^0 = 1$, then, with $S : (t_1, \ldots, t_K) \mapsto \sum_{k=1}^{K} t_k$, the following must hold:*

$$\exists \lambda \in \mathbb{R} / dh_K(t_1^0, \ldots, t_K^0) = \lambda dS(t_1^0, \ldots, t_K^0).$$

This is equivalent to

$$\forall k, \log t_k^0 + 1 = \lambda.$$

Then, $\forall k, \forall k', t_k^0 = t_{k'}^0$ and since $\sum_{k=1}^{K} t_k^0 = 1$, this yields $t_k^0 = \frac{1}{K}$. ■

Now, the contribution $\mathrm{ENT}(\theta; x_i)$ of a single observation (let us call it its individual entropy) to the total entropy $\mathrm{ENT}(\theta; \mathbf{x})$ is considered. Figure 4.3 illustrates the following remarks: two observations x_{i_1} and x_{i_2} are successively considered. They correspond to opposite situations. The dataset arises from a four-component Gaussian mixture model which each component isodensity is colored together with the observations arising from it. At this stage and for this illustration in Figure 4.3, the entropy with respect to the true parameter $\theta^{\wp}$ is considered, since it is available. From the remarks about h and h_K, the following two remarks can be stated.

- The individual entropy of x_i is about zero if there exists k_0 such that $\tau_{ik_0} \approx 1$ and $\tau_{ik} \approx 0$ whenever $k \neq k_0$. Remark that there is no difficulty to classify x_i through MAP in such a case: it will confidently be assigned the class corresponding to component k_0. In Figure 4.3, this is the situation of the observation x_{i_1} and the component k_{green}.

- The individual entropy of x_i is all the greater that $(\tau_{i1}, \ldots, \tau_{iK})$ is closer to the uniform repartition $(\frac{1}{K}, \ldots, \frac{1}{K})$, i.e. that the classification through the MAP rule is uncertain. The worst case is reached as the posterior probabilities are uniformly distributed among the classes $1, \ldots, K$. The individual entropy would then take the value $\log K$ and there would be no available information about to which class it should be assigned or not. The observation x_{i_2} in the example Figure 4.3 has about the same posterior probability $\frac{1}{2}$ to arise from each one of the components k_{cyan} and k_{blue}. Its individual entropy is about $\log 2$.

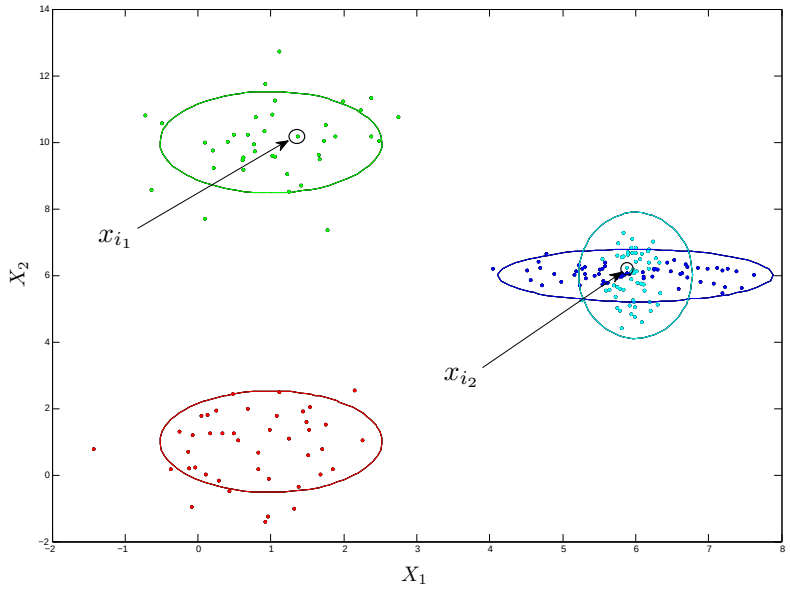


Figure 4.3: An example dataset

In conclusion the individual entropy is a measure of the *assignment confidence* of the considered observation through the MAP classification rule under the distribution $f(\,.\,; \theta)$.

The total entropy $\mathrm{ENT}(\theta;\mathbf{x})$ of a sample $\mathbf{x}$ is the empirical mean entropy (times the sample size) and is an estimator of the expected entropy $\mathbb{E}_{f^\wp}[\mathrm{ENT}(\theta;X)]$ (up to the factor n). Remark that the definition of $\mathbb{E}_{f^\wp}[\mathrm{ENT}(\theta;X)]$ is not a difficulty since $0 \leq \mathrm{ENT}(\theta;x) \leq \log K$ for all x. $\mathrm{ENT}(\theta;\mathbf{x})$ is the empirical mean assignment confidence, and then measures the quality of the classification obtained through the MAP rule over the whole sample.

Involving this quantity in a clustering study means that the chosen cluster notion involves that one expects the classification to be quite confident. This is obviously linked to the notion of well-separated clusters. The conditional classification likelihood shall be used as a contrast from Section 4.2.3. The notion of class underlying this choice is then a compromise between the fit (and then the idea of a Gaussian shape in the Gaussian mixtures framework) because of the log likelihood term on the one hand ("component" point of view), and the assignment confidence because of the entropy term on the other hand (which is rather a "cluster" point of view).

These remarks about the entropy then confirm the link of the $-\log \mathrm{L_{cc}}$ contrast to the clustering and allow to better understand the notion of class which is implied by this choice.

4.2.3 $\log \mathbf{L_{cc}}$ as a Contrast

Choosing $-\log \mathrm{L_{cc}}$ as a contrast when the study is set in the contrast minimization framework (Section 3.1), is then choosing a clustering point of view, with the notion of class as explained in Section 4.2.2. It can be further illustrated by considering the corresponding best distribution from this point of view in a model $\mathcal{M}_m = \{f(\,.\,;\theta) : \theta \in \Theta_m\}$, namely the distribution minimizing the corresponding *loss function*

$$\theta_m \in \underbrace{\underset{\theta\in\Theta_m}{\operatorname{argmin}}\left\{d_{\mathrm{KL}}\left(f^\wp, f(\,.\,;\theta)\right) + \mathbb{E}_{f^\wp}\left[\mathrm{ENT}(\theta;X)\right]\right\}}_{=\underbrace{\underset{\theta\in\Theta_m}{\operatorname{argmin}}\,\mathbb{E}_{f^\wp}\left[-\log \mathrm{L_{cc}}(\theta)\right]}_{\text{this set is denoted by } \Theta_m^0}}.$$

The definition of $\mathbb{E}_{f^\wp}[-\log \mathrm{L_{cc}}(\theta)]$ is no problem as soon as the distribution $f^\wp$ does not have huge tails. If $f^\wp$ is absolutely continuous with respect to the Lebesgue measure, it suffices that its density be negligible with respect to $\frac{1}{\|x\|^3}$. See the discussion after Theorem 7 for further justification. Other reasonable assumptions would be that the support of $f^\wp$ is compact or that the contrast is bounded from above (consider $-\log \mathrm{L_{cc}} \wedge M$ for a constant $M > 0$ well chosen and large enough: see also the discussion after Theorem 7). Those assumptions will be considered in Section 4.4: one or the other will have to be imposed there. The nonemptiness of Θ_m^0 — the set of parameters minimizing the risk over Θ_m — may be guaranteed for example by assuming Θ_m to be compact ($\theta \longmapsto \mathbb{E}_{f^\wp}[-\log \mathrm{L_{cc}}(\theta)]$ is then continue over the compact set Θ_m according to the dominated convergence theorem, and then reaches its infimum over this set). We are interested in applying this contrast to a Gaussian mixture model. Let K be fixed and consider the model $\mathcal{M}_K$ as defined in Section 4.1.1. First of all, remark that $\log \mathrm{L_{cc}} = \log \mathrm{L}$ if $K = 1$: Θ_K^0 is then the set of parameters of the distributions which minimize the Kullback-Leibler divergence to the data distribution $f^\wp$ (let us denote

this last parameters set Θ_K^{KL}). Now, if $K > 1$, $\theta_K^0 \in \Theta_K^0$ may be close to minimizing the Kullback-Leibler divergence if the corresponding components do not overlap since then, the entropy is about zero. But as those components overlap, this is not the case anymore. Example 4 illustrates this.

In order to really define the loss function, and to fully understand this framework, it is necessary to consider the "best element of the universe" $\mathcal{U}$ (see Section 3.1.1):

$$\underset{\theta \in \mathcal{U}}{\operatorname{argmin}} \, \mathbb{E}_{f^{\wp}} \left[-\log \mathrm{L}_{\mathrm{cc}}(\theta) \right].$$

The universe $\mathcal{U}$ must be chosen with care. There is no natural relevant choice, as for example in the density estimation framework where the set of every densities of the world may be chosen. First, as already mentioned, the considered contrast is well-defined in a parametric mixture setup, and not necessarily over any mixtures densities set because of the definition of the entropy term involving the definition of each component. However, this would still enable to consider mixtures much more general than mixtures of Gaussian components. The ideas developed in Chapter 7 may for example suggest to involve in the universe mixtures which components are themselves Gaussian mixtures. The contrast may be well-defined over such sets of mixtures, if the parameterization is well chosen. But this would not make sense. The mixture with one component which consists of a mixture of K Gaussian components, and which then yields a single class which has a non-Gaussian shape, always has a smaller contrast value than the corresponding Gaussian mixture, yielding K classes: the likelihood is exactly the same since the mixture distribution is the same, but the entropy is null when the mixture is considered as one non-Gaussian component while it is never null when the mixture is considered as K Gaussian components. This illustrates how carefully the components involved in the study must be chosen: involving for example any mixture of Gaussian mixtures means considering that a class may be almost anything, and it may notably contain for example two Gaussian shaped clusters of observations very far from each other! The components should in any case be chosen with respect to the corresponding cluster shape. The most natural is then to involve in the universe only Gaussian mixtures: $\mathcal{U}$ may be chosen as $\cup_{1 \leq K \leq K_M} \mathcal{M}_K$ (i.e. the union of all the models involved in the study: see Section 4.4).

Example 4 *$f^{\wp}$ is the normal density $\mathcal{N}(0,1)$ ($d = 1$).*
The model $\mathcal{M} = \{\frac{1}{2}\phi(\,.\,; -\mu, \sigma^2) + \frac{1}{2}\phi(\,.\,; \mu, \sigma^2); \mu \in \mathbb{R},\ \sigma^2 > 0\}$ is considered. There is no reason to impose any further condition on the model here.

It is interesting to consider Θ_2^0 in this (probably) most simple situation. Actually, we were not even able to calculate it by hand and to get an expression of it, even with σ fixed! It illustrates the analysis difficulty of the problem and of the contrast $\log L_{cc}$. Therefore, there will be no hope to calculate it in a more general case, particularly as the dimension and/or the number of components of the model are greater.

But it can be computed by numerical evaluations. We then obtain that $\Theta_K^0 = \{(-\mu_0, \sigma_0^2), (\mu_0, \sigma_0^2)\}$, so that, up to a label switch, there exists a unique minimizer of $\mathbb{E}_{f^{\wp}} [-\log L_{cc}(\mu, \sigma^2)]$ in Θ_K in this case (see Figure 4.4). By the way, there is no need to impose any condition on the model to get this result here. We numerically found that $\mu_0 \approx 0.83$ and $\sigma_0^2 \approx 0.31$. This solution is obviously not the same as the one minimizing the Kullback-Leibler divergence, which is $\mu_{KL} = 0$ and $\sigma_{KL}^2 = 1$: $f(\,.\,; \theta_K^{KL}) = f^{\wp}$ (see Figure 4.5).

This illustrates that the objective that we choose by introducing the $-\log L_{cc}$ *contrast is not to recover the true distribution, even when it is available in the considered model. In this example, the distribution corresponding to* θ_K^{KL} *does not enable to define a rule to design two relevant classes from a "typical" dataset that would arise from* $f^{\wp}$*, in that they would exactly and completely overlap each other, and the class assignment through MAP would be completely arbitrary...instead,* θ_K^0 *reaches a compromise between the divergence to the true distribution and the assignment confidence, as illustrated in Figure 4.5.*

The necessity of choosing a relevant model is striking in this example: this two-component model should obviously not be used for a clustering purpose, at least for datasets with size great enough so that the true distribution is well estimated by the empirical distribution.

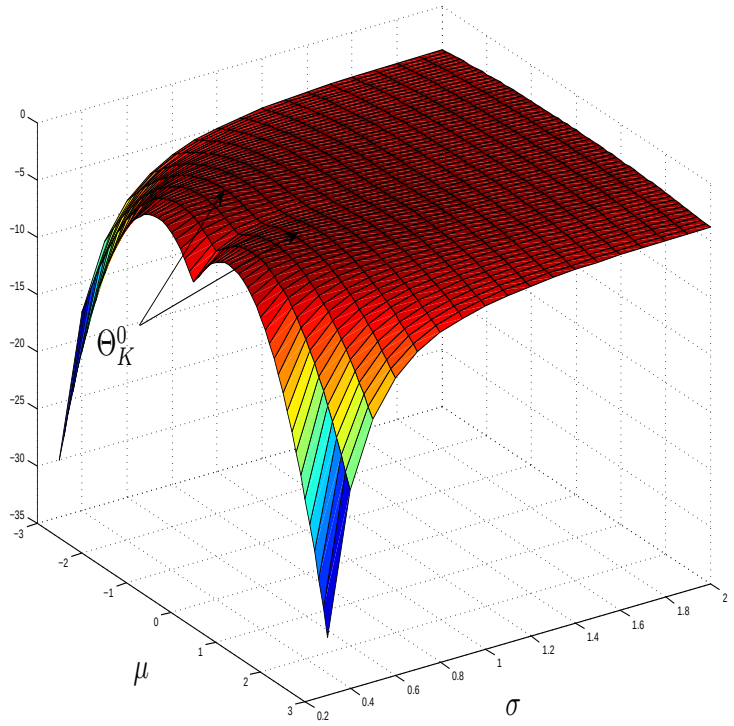


Figure 4.4: $\mathbb{E}_{f^{\wp}}\left[\log \mathrm{L_{cc}}(\mu, \sigma^2)\right]$ w.r.t. μ and σ, and Θ_K^0, for Example 4

The estimator which results from this $-\log \mathrm{L_{cc}}$ contrast is considered in the following section.

4.3 Estimation: MLccE

$-\log \mathrm{L_{cc}}$ is then a contrast tightly linked to a clustering purpose, and to a particular notion of class. This last notion presumably conforms a widespread notion of cluster.

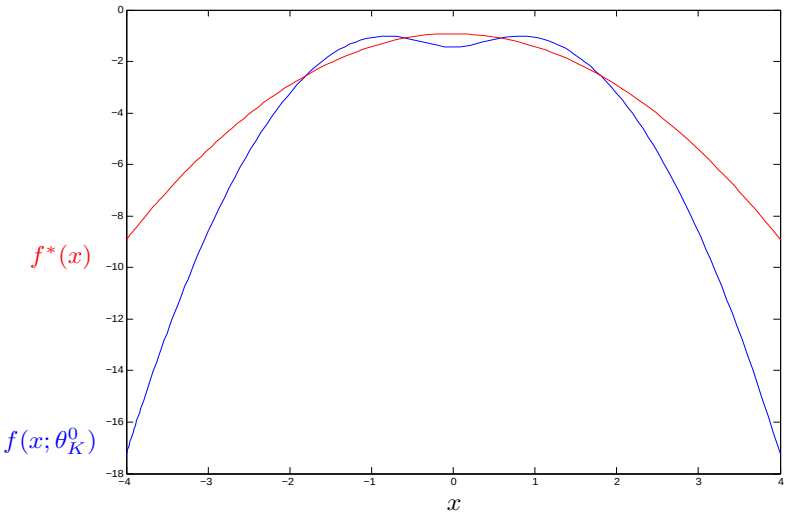


Figure 4.5: $\log f^{\wp}$ (red, which is also $\log f(\,.\,;\theta_K^{\mathrm{KL}})$) and $\log f(\,.\,;\theta_K^0)$ (blue) for Example 4

Therefore, it is worth extending the reasoning and trying to apply this contrast from the estimation step. Let us fix the number of components K and the model $\mathcal{M}_K$ in this section, and study the new estimator defined this way. The subscript K is omitted in the notation of this section.

A new estimator is then defined as the minimum contrast estimator corresponding to $-\log \mathrm{L_{cc}}$. General conditions ensuring its consistency are given in Theorem 3. They notably involve the Glivenko-Cantelli property of the class of functions $\{\gamma(\theta) : \theta \in \Theta\}$. This property is recalled in Section 4.3.2 and the bracketing entropy of this class of functions is studied in the same section, since it enables to assess the Glivenko-Cantelli property. Several situations are considered, since these results will be useful in Section 4.4, too. These results brought together provide the consistency of the estimator in Gaussian mixture models: this is Theorem 4 in Section 4.3.1. The proofs are given in Section 4.3.3 and a few illustrative simulations in Section 4.3.4.

We aim at deriving results adapted to the considered Gaussian mixture models framework, and to the $-\log \mathrm{L_{cc}}$ contrast. Most results will be stated in a general parametric model setting with a general contrast γ. The generic model is denoted by $\mathcal{M}$ and is assumed to have parameter set $\Theta \subset \mathbb{R}^D$, with D the "dimension" of the model, namely the number of free parameters as defined in Section 4.1.1. The assumptions involved in those general results will be discussed in the particular framework we are interested in, as well as the supplementary conditions we may have to impose to our model setting to guarantee they hold, when they are stated.

Some notation is introduced for this section and the following one. All expectations $\mathbb{E}$ and probabilities $\mathbb{P}$ are taken with respect to the distribution $f^{\wp}d\lambda$. X is a random variable in $\mathbb{R}^d$ with distribution $f^{\wp}d\lambda$ and $X_1, \ldots, X_n$ an i.i.d. sample from the same distribution. For a general contrast γ, we write γ_n its empirical version: $\gamma_n(\theta) = \frac{1}{n}\sum_{i=1}^n \gamma(\theta; X_i)$. $\mathbb{R}^D$ is equipped with the infinite norm: $\forall \theta \in \mathbb{R}^D, \|\theta\|_\infty =$

$\max_{1\leq i\leq D}|\theta_i|$, where θ_i is the i^{th} coordinate of the decomposition of θ over the canonical basis of $\mathbb{R}^D$. Consequently, for any $r \in \mathbb{N}^* \cup \{\infty\}$ and for any function $g : \mathbb{R}^d \to \mathbb{R}$, $\|g\|_r$ is the L_r-norm of g with respect to the $f^{\wp}d\lambda$ distribution: $\|g\|_r = \mathbb{E}_{f^{\wp}}\left[|g(X)|^r\right]^{\frac{1}{r}}$ if $r < \infty$ and[5] $\|g\|_\infty = \operatorname{ess\,sup}_{X\sim f^{\wp}}|g(X)|$; for any linear form $l : \mathbb{R}^D \to \mathbb{R}$, $\|l\|_\infty$ is the usual norm of a linear map over a finite-dimensional normed vector space: $\|l\|_\infty = \max_{\theta\in\mathbb{R}^D}\frac{l(\theta)}{\|\theta\|_\infty} = \max_{\theta\in\mathbb{R}^D:\|\theta\|_\infty=1} l(\theta)$.

4.3.1 Definition, Consistency

The minimum contrast estimator, called MLccE (Maximum conditional classification log Likelihood Estimator) by direct analogy with the maximum likelihood estimator, is written $\widehat{\theta}^{\text{MLccE}}$:

$$\widehat{\theta}^{\text{MLccE}} \in \operatorname*{argmin}_{\theta\in\Theta} \gamma_n(\theta).$$

To ensure its existence, we impose that Θ is compact. This is a heavy assumption, but it will be natural and necessary for the following results to hold. Assuming that the covariance matrices are bounded from below is a reasonable and necessary assumption in the Gaussian mixture framework: without this assumption, neither the log likelihood, nor the conditional classification likelihood would be bounded. Insights to choose the lower bounds on the proportions and the covariance matrices are suggested in Section 5.1 below. The upper bound on the covariance matrices and the condition on the means, although not necessary in the usual likelihood framework, do not seem to be avoidable in this framework (see Section 4.3.2). This is a consequence of the behavior of the entropy term in the contrast function as a component goes to zero. Remark that the results of this chapter still hold if the estimator only maximizes $-\log \mathrm{L}_{\mathrm{cc}}$ up to $o_{\mathbb{P}}(1)$.

The following theorem, which is directly adapted from van der Vaart (1998, Section 5.2), gives sufficient conditions for the consistency of the estimator $\widehat{\theta}^{\text{MLccE}}$. We write $\forall\theta\in\Theta, \forall\widetilde{\Theta}\subset\Theta, d(\theta,\widetilde{\Theta}) = \inf_{\widetilde{\theta}\in\widetilde{\Theta}}\|\theta-\widetilde{\theta}\|_\infty$.

Theorem 3
Let $\Theta \subset \mathbb{R}^D$ *and* $\gamma : \Theta \times \mathbb{R}^d \longrightarrow \mathbb{R}$.
Assume $\exists\theta^0 \in \Theta$ *such that*

$$\mathbb{E}_{f^{\wp}}\left[\gamma(\theta^0)\right] = \min_{\theta\in\Theta}\mathbb{E}_{f^{\wp}}\left[\gamma(\theta)\right] \tag{A1}$$

(i.e. Θ^0 *is not empty).*
Assume

$$\forall\varepsilon>0,\ \inf_{\{\theta;d(\theta,\Theta^0)>\varepsilon\}}\mathbb{E}_{f^{\wp}}\left[\gamma(\theta)\right] > \mathbb{E}_{f^{\wp}}\left[\gamma(\theta^0)\right]. \tag{A2}$$

Assume

$$\sup_{\theta\in\Theta}\left|\gamma_n(\theta) - \mathbb{E}_{f^{\wp}}\left[\gamma(\theta)\right]\right| \xrightarrow{\mathbb{P}} 0. \tag{A3}$$

Define $\forall n$,

$$\hat{\theta} = \hat{\theta}(X_1,\ldots,X_n) \in \Theta \text{ such that } \gamma_n(\hat{\theta}) \leq \gamma_n(\theta^0) + o_{\mathbb{P}}(1).$$

[5] Recall that $\operatorname{ess\,sup}_{Z\sim\mathbb{P}} Z = \inf\{z : \mathbb{P}[Z\leq z] = 1\}$. We then notably have $\operatorname{ess\,sup}_{X\sim f^{\wp}}|g(X)| \leq \sup_{x\in\operatorname{supp} f^{\wp}}|g(x)|$.

Then

$$d(\hat{\theta}, \Theta^0) \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Under the assumptions (A1), (A2), (A3), which will be of concern in the three following paragraphs, the minimum contrast estimator is then consistent (in probability).

Remark that the strong consistency of this theorem holds if (A3) is replaced by a convergence almost sure (this is the case under the conditions we are to define), and if the inequality in the definition of $\hat{\theta}$ holds almost surely instead of in probability.

Sketch of Proof *There is no difficulty with this theorem. The assumptions guarantee a convenient situation. With great probability as n grows, from* (A3)*, $\gamma_n(\theta)$ is uniformly close to $\mathbb{E}_{f^\wp}[\gamma(\theta)]$. This particularly holds for $\widehat{\theta}^{MLccE}$ and θ^0. Then, from the definition of $\widehat{\theta}^{MLccE}$, $\mathbb{E}_{f^\wp}\left[\gamma(\widehat{\theta}^{MLccE})\right]$ cannot be very larger than $\mathbb{E}_{f^\wp}[\gamma(\theta^0)]$, which reaches the minimal value...Yet* (A2) *makes it impossible to have almost minimal $\mathbb{E}_{f^\wp}[\gamma(\theta)]$ value while lying far from Θ^0_K.* ■

Let us subsequently discuss this result and the corresponding assumptions in the Gaussian mixture model context, with $\gamma = -\log \mathrm{L_{cc}}$ and $\Theta = \Theta_K$. This will yield the

Theorem 4 (Weak Consistency of MLccE, compact case)
Let $\mathcal{M}$ be a Gaussian mixture model, as defined in Section 4.1.1, with parameter space $\Theta \subset \mathbb{R}^D$.
Assume that Θ is compact.
Let $\Theta^0 = \underset{\theta\in\Theta}{\operatorname{argmin}} \mathbb{E}_{f^\wp}[-\log L_{cc}(\theta; X)]$.
Let $\Theta^{\mathcal{O}}$ be an open subset of $\mathbb{R}^D$ over which $\log L_{cc}$ is well-defined, such that $\Theta \subset \Theta^{\mathcal{O}}$ and assume

$$L'(x) = \sup_{\theta\in\Theta^{\mathcal{O}}} \left\| \left(\frac{\partial \log L_{cc}}{\partial \theta} \right)_{(\theta;x)} \right\|_\infty < \infty \quad f^\wp d\mu - a.s.$$
$$\|L'\|_1 < \infty.$$

For any $\theta^0 \in \Theta^0$, let, for any n, $\widehat{\theta}^{MLccE}$ be an estimator (almost) maximizing the conditional classification likelihood: $\widehat{\theta}^{MLccE} = \widehat{\theta}^{MLccE}(X_1, \ldots, X_n) \in \Theta$ such that

$$-\log L_{cc}(\widehat{\theta}^{MLccE}; \mathbf{X}) \leq -\log L_{cc}(\theta^0; \mathbf{X}) + o_{\mathbb{P}}(n).$$

Then

$$d(\widehat{\theta}^{MLccE}, \Theta^0) \xrightarrow[n\longrightarrow\infty]{\mathbb{P}} 0.$$

The assumption about L' is a consequence of lemma 3 (see below) and shall be discussed in Section 4.3.2.
Under the compactness assumption, $\widehat{\theta}^{\mathrm{MLccE}}$ is then consistent (in probability). It is even strongly consistent if it minimizes the empirical contrast almost surely (instead of up to a $o_{\mathbb{P}}(1)$). Let us highlight that it then converges to the set of parameters minimizing the loss function (i.e. the expected contrast), which has no reason to contain the true

distribution — but in the particular case $K = 1$ — even if the last lies in the model $\mathcal{M}_K$. Only in cases where the data arise from a true Gaussian mixture which components are well separated, and K is the true number of components, $f^{\wp}$ can be expected to lie close to $\{\gamma(\theta) : \theta \in \Theta^0\}$.

Let us discuss the assumptions of Theorem 3 in this framework so as to derive this last result.

Assumption (A1) is the least that can be expected! It is guaranteed if the parameter space is assumed to be compact. Indeed, $\theta \in \Theta \mapsto \mathbb{E}_{f^{\wp}}[\gamma(\theta)]$ is then continue from the dominated convergence theorem, since then $x \in \mathbb{R}^d \mapsto \sup_{\theta\in\Theta} \gamma(\theta; x)$ behaves for the largest values of x roughly as $\gamma(\theta; x)$ for any θ, and is then integrable with respect to $f^{\wp}d\lambda$ as soon as this distribution does not have heavy tails: remark that, for fixed θ and for large values of x, $-\log \mathrm{L}_{\mathrm{cc}}(\theta; x)$ increases roughly as $-\log f(x; \theta)$ ($\mathrm{ENT}(\theta; x)$ is bounded), and then, $f^{\wp}(x) = o(\frac{1}{x^3})$ should suffice. It would be quite technical to write this completely.

Assumption (A2) holds, too, under this compactness assumption: since $\theta \in \Theta_K \mapsto \mathbb{E}_{f^{\wp}}[\gamma(\theta)]$ reaches its minimum value on the compact set $\Theta_K \backslash \{\theta \in \Theta_K : d(\theta, \Theta_K^0) > \varepsilon\}$ (closed and bounded if Θ_K is), this minimum value is necessarily strictly greater than $\mathbb{E}_{f^{\wp}}[\gamma(\theta^0)]$.

Remark that assumption (A3) is strong and could be appreciably relaxed (see for instance van der Vaart (1998): it suffices that the variances $\mathrm{var}(\gamma(\theta))$ be bounded from above uniformly in θ, which guarantees a control of $(\gamma_n(\theta) - \mathbb{E}_{f^{\wp}}[\gamma(\theta)])$). However, there is no reason to do so here since it can be guaranteed under the compactness assumption which has been already stated. This will be proved (in the stronger a.s. version) in Section 4.3.2 through bracketing entropy arguments. Actually, it could be proved with more direct reasoning, but we are interested in this approach since it will be helpful in the theoretical study of the model selection step (Section 4.4).

In conclusion, Theorem 3 easily applies when Θ_K is assumed to be compact. Deriving theoretical results under this assumption helps understanding the behavior of the estimator (and next, of the model selection procedure).

Let us now prove assumption (A3).

4.3.2 Bracketing Entropy and Glivenko-Cantelli Property

The notions of *Glivenko-Cantelli* classes of functions and of *entropy with bracketing* of a class of functions with respect to a distribution $\mathbb{P}$ over $\mathbb{R}^d$ are first recalled.

Definition 2
A class $\mathcal{G}$ of measurable functions $g : \mathbb{R}^d \to \mathbb{R}$ is $\mathbb{P}$-Glivenko-Cantelli iff:

$$\left\| \frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}[g(X)] \right\|_{\mathcal{G}} := \sup_{g\in\mathcal{G}} \left| \frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}[g(X)] \right| \xrightarrow{a.s.} 0, \tag{4.3}$$

where $X_1, \ldots, X_n$ is a sample from the distribution $\mathbb{P}$ and the expectation over X is taken with respect to the distribution $\mathbb{P}$.

A class $\mathcal{G}$ of functions is then $\mathbb{P}$-Glivenko-Cantelli if it fulfills a uniform law of large numbers for the distribution $\mathbb{P}$.

A sufficient condition for a family $\mathcal{G}$ to be $\mathbb{P}$-Glivenko-Cantelli if that it is not too complex, in a sense that can be measured through the *entropy with bracketing*:

Definition 3
Let $r \in \mathbb{N}^*$ and $l, u \in L_r(\mathbb{P})$.
The bracket $[l, u]$ is the set of all functions $g \in \mathcal{G}$ with $l \leq g \leq u$ (i.e. $\forall x \in \mathbb{R}^d, l(x) \leq g(x) \leq u(x)$).
$[l, u]$ is an ε-bracket if $\|l - u\|_r = \mathbb{E}\left[|l - u|^r\right]^{\frac{1}{r}} \leq \varepsilon$.
The bracketing number $N_{[]}(\varepsilon, \mathcal{G}, L_r(\mathbb{P}))$ is the minimum number of ε-brackets needed to cover $\mathcal{G}$.
The entropy with bracketing $\mathcal{E}_{[]}(\varepsilon, \mathcal{G}, L_r(\mathbb{P}))$ of $\mathcal{G}$ with respect to $\mathbb{P}$ is the logarithm of the bracketing number.

The bracketing entropy is a L_r-measure of the complexity of the class $\mathcal{G}$. It is quite natural that the behavior of all functions lying inside an ε-bracket can be controlled by the behavior of the extrema of the bracket (remark that those extrema are not assumed to belong to $\mathcal{G}$ themselves). If those endpoints belong to $L_1(\mathbb{P})$, they fulfill a law of large numbers, and if the number of them needed to cover $\mathcal{G}$ is finite, then this is no surprise that $\mathcal{G}$ can be proved to fulfill a uniform law of large numbers:

Theorem 5 *Every class $\mathcal{G}$ of measurable functions such that $\mathcal{E}_{[]}(\varepsilon, \mathcal{G}, L_1(\mathbb{P})) < \infty$ for every $\varepsilon > 0$ is $\mathbb{P}$-Glivenko-Cantelli.*

See van der Vaart (1998, Chapter 19) for a proof of this result. This is a generalization of the usual Glivenko-Cantelli theorem, which states the a.s. uniform convergence of the empirical distribution function to the distribution function.

We shall then prove that the class of functions $\{\gamma(\,.\,;\theta) : \theta \in \Theta_K\}$ has finite ε-bracketing entropy for any $\varepsilon > 0$ and the assumption (A3) will be ensured.

Recall $\mathbb{R}^D$ is equipped with the infinite norm $\|\theta\|_\infty = \max_{i \in \{1,\ldots,D\}} |\theta_i|$, which turns out to be convenient for the bracketing entropy calculations. From now on, it is assumed that $\Theta \subset \Theta^{\mathcal{O}}$, with $\Theta^{\mathcal{O}}$ an open subset of $\mathbb{R}^D$ over which γ is well-defined and C^1 for $f^{\wp} d\lambda$-almost all $x \in \mathbb{R}^d$. This is a natural assumption: recall Θ will typically be assumed to be compact. It must be ensured that it is included in an open set so that the differential of γ, which will be involved next, is well-defined. This assumption is no problem in the Gaussian mixture model framework with the conditional classification likelihood (or the usual likelihood by the way), for example with the general model with no constraint on the covariance matrices, or with the model with diagonal covariance matrices. But it requires (with the conditional classification likelihood as contrast) the proportions to be positive. Actually, this could be avoided from this point of view, but this assumption is to be necessary anyway because of the definition of L' (see Lemma 2). We already mentioned that, because of the behavior of the differential of the function h (see Section 4.2.2) at zero, components going to zero must be avoided. Moreover, for the same technical reason, we have to assume the mean parameters to be bounded. This may perhaps be avoided, but probably not with our approach, which relies on the mean

value theorem, and then requires the differential of the contrast to be controlled. This can be done only if the component densities are kept away from zero.

The following lemma guarantees that the bracketing entropy of $\{\gamma(\,.\,;\theta) : \theta \in \Theta_K\}$ is finite for any ε, if Θ_K is convex and bounded. The assumption about the differential of the contrast is not a difficulty in our framework, provided that non-zero lower bounds over Θ on the proportions and the covariance matrices are imposed. The lemma is written for any $\widetilde{\Theta}$ bounded and included in Θ_K (which would not necessarily be bounded in this case) since it will be applied locally around θ^0 in Section 4.4.

For any $\widetilde{\Theta}$ bounded $\subset \mathbb{R}^D$, $\operatorname{diam}\widetilde{\Theta} = \sup_{\theta_1,\theta_2\in\widetilde{\Theta}} \|\theta_1 - \theta_2\|_\infty$.

Lemma 2 (Bracketing Entropy, Convex Case)
Let $r \in \mathbb{N}^*$.
Let $D \in \mathbb{N}^*$ and $\Theta \subset \mathbb{R}^D$ assumed to be convex.
Let $\Theta^{\mathcal{O}}$ be an open subset of $\mathbb{R}^D$ such that $\Theta \subset \Theta^{\mathcal{O}}$ and $\gamma : \Theta^{\mathcal{O}} \times \mathbb{R}^d \longrightarrow \mathbb{R}$.
$\theta \in \Theta^{\mathcal{O}} \longmapsto \gamma(\theta; x)$ is assumed to be C^1 over $\Theta^{\mathcal{O}}$ for $f^{\wp}d\lambda$-almost all $x \in \mathbb{R}^d$.
Assume[6]

$$L'(x) = \sup_{\theta\in\Theta} \left\| \left(\frac{\partial\gamma}{\partial\theta}\right)_{(\theta;x)} \right\|_\infty < \infty \quad f^{\wp}d\mu\text{-a.s.}$$
$$\|L'\|_r = \mathbb{E}_{f^{\wp}}\left[L'(X)^r\right]^{\frac{1}{r}} < \infty.$$

Then,

$$\forall \widetilde{\Theta} \text{ bounded } \subset \Theta, \forall \varepsilon > 0,\ N_{[]}(\varepsilon, \{\gamma(\theta) : \theta \in \widetilde{\Theta}\}, \|\cdot\|_r) \leq \left(\frac{\|L'\|_r \ diam\,\widetilde{\Theta}}{\varepsilon}\right)^D \vee 1.$$

Remark that Θ does not have to be compact. It is however a sufficient condition for $L'(x) < \infty$ to hold a.s. The proof of this result is a calculation. It relies on the mean value theorem, hence the convexity assumption. The natural parameter space corresponding to the Gaussian mixture model with diagonal covariance matrices (each covariance matrix is parametrized by its diagonal values...), for instance, is convex. The parameter space corresponding to the natural parametrization of the model with diagonal covariance matrices and equal volumes between components is convex, too (if $d > 1$...). The mixture model with general covariance matrices has a convex natural parameter space, too, since the set of definite positive matrices is convex. See the proofs section below (Section 4.3.3) for further justification. However, in the Gaussian mixture framework there is no reason in general that the model parameter space Θ should be convex. It is then useful to generalize Lemma 2.

It can be done at the price of assuming Θ to be compact, and included in an open set over which the property about the supremum of the contrast differential still holds. This is no difficulty for the mixture models we consider, under the same lower bounds constraints as before (since $\Theta^{\mathcal{O}}$ itself can be chosen to be included in a compact subset of the set of possible parameters...). The entropy is then increased by a multiplying

[6]Let us stress that the « ' » symbol is not a differentiation symbol here.

constant factor Q, which only depends on Θ and roughly measures its "nonconvexity". Since only the exponential behavior of the entropy with respect to ε is of concern, this does not make the result worse. On the one hand, it is still finite for any ε, from which the family $\{\gamma(\theta) : \theta \in \Theta\}$ is Glivenko-Cantelli and assumption (A3) is guaranteed. On the other hand, the integrated square root of the entropy in a neighbourhood of 0 is finite too, and the family $\{\gamma(\theta) : \theta \in \Theta\}$ is $f^{\wp}$-Donsker[7], which will be essential in Section 4.4, together with the form of the entropy of the family $\{\gamma(\theta) : \theta \in \widetilde{\Theta}\}$ with respect to ε and to $\operatorname{diam}\widetilde{\Theta}$ for any $\widetilde{\Theta} \subset \Theta$.

Lemma 3 (Bracketing Entropy, Compact Case)
Let $r \in \mathbb{N}^*$.
Let $D \in \mathbb{N}^*$ and $\Theta \subset \mathbb{R}^D$ assumed to be compact.
Let $\Theta^{\mathcal{O}}$ be an open subset of $\mathbb{R}^D$ such that $\Theta \subset \Theta^{\mathcal{O}}$ and $\gamma : \Theta^{\mathcal{O}} \times \mathbb{R}^d \longrightarrow \mathbb{R}$.
$\theta \in \Theta^{\mathcal{O}} \longmapsto \gamma(\theta; x)$ is assumed to be C^1 over $\Theta^{\mathcal{O}}$ for $f^{\wp}d\lambda$-almost all $x \in \mathbb{R}^d$.
Assume

$$L'(x) = \sup_{\theta \in \Theta^{\mathcal{O}}} \left\| \left(\frac{\partial \gamma}{\partial \theta} \right)_{(\theta;x)} \right\|_{\infty} < \infty \quad f^{\wp}d\mu\text{-a.s.}$$
$$\|L'\|_r = \mathbb{E}_{f^{\wp}} \left[L'(X)^r \right]^{\frac{1}{r}} < \infty.$$

Then,

$$\exists Q \in \mathbb{N}^*, \forall \widetilde{\Theta} \subset \Theta, \forall \varepsilon > 0,\ N_{[]}(\varepsilon, \{\gamma(\theta) : \theta \in \widetilde{\Theta}\}, \|\cdot\|_r) \leq Q \left(\frac{\|L'\|_r\ diam \widetilde{\Theta}}{\varepsilon} \right)^D \vee 1.$$

Q is a constant which depends on the geometry of Θ ($Q = 1$ if Θ is convex).

The proof of this lemma is done by applying Lemma 2 since Θ is still locally convex. Since it is compact, it can be covered with a finite number Q of open balls, which are convex... Lemma 2 then applies to the convex hull of the intersection of Θ with each one of them. The supremum of L' is taken over $\Theta^{\mathcal{O}}$ — instead of Θ — to be sure that the assumptions of Lemma 2 are fulfilled over those entire balls, which may not be included in Θ...

Actually, the result we need for Section 4.4 is slightly different, and is obtained from Lemma 2 by a little modification. Since it is applied locally there, the convexity assumption is no problem. A supplementary and strong assumption is made: $\|L\|_{\infty} < \infty$. This assumption is not fulfilled in general in the Gaussian mixture model framework: a sufficient condition for this to hold is that the support of $f^{\wp}$ is bounded. This is false of course for most usual distributions we may have in mind. But this is a reasonable modelling assumption: it may even mostly be rather the only reasonable modelling assumption since most modelled phenomena are bounded (as mentioned for example in Bickel and Doksum, 2001, Chapter 1, page 4). Another sufficient condition to guarantee this assumption is that the contrast is bounded from above. This is actually not the

[7]Recall that a family $\mathcal{G}$ of functions is $\mathbb{P}$-*Donsker* if it fulfills a "uniform" Central Limit Theorem, namely if the process $\left(\sqrt{n}(\frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}\left[g(X)\right])\right)_{g \in \mathcal{G}}$ converges in distribution. A sufficient condition is that $\sqrt{\mathcal{E}_{[]}(\varepsilon, \mathcal{G}, L_2(\mathbb{P}))}$ is integrable at zero. See van der Vaart (1998, Chapter 19).

case of the contrast $-\log \mathrm{L_{cc}}$ as we defined it (fix θ and let $x \longrightarrow \infty$), but this can be imposed: simply replace $-\log \mathrm{L_{cc}}$ by $(-\log \mathrm{L_{cc}} \wedge M)$ and, provided that M is large enough, this new contrast shall behave like $\log \mathrm{L_{cc}}$, since we are interested in the regions where the contrast is minimized. Of course, this is a supplementary difficulty in practice to choose a relevant M value.

Lemma 4 (Bracketing Entropy, Convex Case)
Let $r \geq 2$.
Let $D \in \mathbb{N}^*$ and $\Theta \subset \mathbb{R}^D$ assumed to be convex.
Let $\Theta^{\mathcal{O}}$ be an open subset of $\mathbb{R}^D$ such that $\Theta \subset \Theta^{\mathcal{O}}$ and $\gamma : \mathbb{R}^D \times \Theta^{\mathcal{O}} \longrightarrow \mathbb{R}$.
$\theta \in \Theta^{\mathcal{O}} \longmapsto \gamma(\theta; x)$ is assumed to be C^1 over $\Theta^{\mathcal{O}}$ for $f^{\wp} d\lambda$-almost all $x \in \mathbb{R}^d$.
Assume

$$L(x) = \sup_{\theta \in \Theta} |\gamma(\theta; x)| < \infty \quad f^{\wp} d\mu\text{-a.s.}$$
$$\|L\|_\infty = \operatorname*{ess\,sup}_{X \sim f^{\wp}} L(X) < \infty.$$

and

$$L'(x) = \sup_{\theta \in \Theta} \left\| \left(\frac{\partial \gamma}{\partial \theta} \right)_{(\theta; x)} \right\|_\infty < \infty \quad f^{\wp} d\mu\text{-a.s.}$$
$$\|L'\|_2 = \mathbb{E}_{f^{\wp}} \left[L'(X)^2 \right]^{\frac{1}{2}} < \infty.$$

Then,

$$\forall \widetilde{\Theta} \subset \Theta, \forall \varepsilon > 0, \ N_{[]}(\varepsilon, \{\gamma(\theta) : \theta \in \widetilde{\Theta}\}, \|\cdot\|_r) \leq \left(\frac{2^{r-2} \|L\|_\infty^{\frac{r-2}{2}} \|L'\|_2 \ diam \widetilde{\Theta}}{\varepsilon^{\frac{r}{2}}} \right)^D \vee 1.$$

Finally, let us remark that those results apply with various contrasts. We are interested in this chapter in the application to the conditional classification likelihood, but they hold even more in the usual likelihood framework. Maugis and Michel (2009) already provide bracketing entropy results in this framework. Our results cannot be directly compared to theirs since the distance they consider is the Hellinger distance. Let us however notice that their results are more precise in the sense that their dependency on the bounds imposed on the parameter space and the dimension of the variable space d is explicit. This is helpful when deriving an oracle inequality: it suffices to impose the same bounds on all models, whatever their dimension, so that the penalty shape can be justified. This is an explicit condition. But it was not possible to derive a local control of the entropy in this Hellinger distance situation, hence an unpleasant logarithm term in the expression of the optimal penalty in Maugis and Michel (2009) (see Section 3.3). Finally, they derive controls of the bracketing entropy without assuming the contrast to be bounded but their results however finally lead to this assumption, as mentioned in the discussion following the statement of Theorem 7 (page 113). The results presented above achieve the same rate with respect to ε. They depend on more opaque quantities ($\|L\|_\infty$ and $\|L'\|_2$). This notably implies, from this first step already, the assumption that the contrast is bounded — over the true distribution support. However, it could be expected to control those quantities with respect to the bounds on the parameter space. Moreover, beside their simplicity, they have the advantage that it is straightforward to derive a local control of the entropy.

4.3.3 Proofs

Proof (Theorem 3: Minimum Contrast Estimator Convergency) *The assumptions ensure a convenient situation. Let $\varepsilon > 0$ and let*

$$\eta = \inf_{d(\theta,\Theta^0)>\varepsilon} \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right] - \mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right] > 0,$$

from assumption (A2)*. For n large enough and with large probability, from assumption* (A3) *and the definition of $\hat{\theta}$,*

$$sup_{\theta\in\Theta}|\gamma_n(\theta) - \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right]| < \frac{\eta}{3}$$
$$\gamma_n(\hat{\theta}) \leq \gamma_n(\theta^0) + \frac{\eta}{3}.$$

Then

$$\begin{aligned}\mathbb{E}_{f^\wp}\left[\gamma(\hat{\theta})\right] - \mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right] \leq \mathbb{E}_{f^\wp}\left[\gamma(\hat{\theta})\right] &- \gamma_n(\hat{\theta}) \\ &+ \gamma_n(\hat{\theta}) - \gamma_n(\theta^0) \\ &+ \gamma_n(\theta^0) - \mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right] \\ < \eta.\end{aligned}$$

And then $d(\hat{\theta}, \Theta^0) < \varepsilon$ with great probability, as n is large enough. ■

Proof (Lemma 2: Bracketing Entropy, Convex Case) *Let $\varepsilon > 0$, and $\widetilde{\Theta} \subset \Theta$, with $\widetilde{\Theta}$ bounded. Let $\widetilde{\Theta}_\varepsilon$ be a grid in Θ which "ε-covers" $\widetilde{\Theta}$ in any dimension with step ε. $\widetilde{\Theta}_\varepsilon$ is for example $\widetilde{\Theta}_\varepsilon^1 \times \cdots \times \widetilde{\Theta}_\varepsilon^D$ with*

$$\forall i \in \{1,\ldots,D\}, \widetilde{\Theta}_\varepsilon^i = \left\{\widetilde{\theta}_{\min}^i, \widetilde{\theta}_{\min}^i + \varepsilon, \ldots, \widetilde{\theta}_{\max}^i\right\},$$

where

$$\forall i \in \{1,\ldots,D\}, \left\{\theta^i : \theta \in \widetilde{\Theta}\right\} \subset \left[\widetilde{\theta}_{\min}^i - \frac{\varepsilon}{2}, \widetilde{\theta}_{\max}^i + \frac{\varepsilon}{2}\right].$$

This is always possible since Θ is convex. For the sake of simplicity, it is assumed without loss of generality, that $\widetilde{\Theta}_\varepsilon \subset \widetilde{\Theta}$. The interest of the $\|\cdot\|_\infty$ norm is that the step of the grid $\widetilde{\Theta}_\varepsilon$ is the same as the step over each dimension, ε. Indeed,

$$\forall \tilde{\theta} \in \widetilde{\Theta}, \exists \tilde{\theta}_\varepsilon \in \widetilde{\Theta}_\varepsilon / \|\tilde{\theta} - \tilde{\theta}_\varepsilon\|_\infty \leq \frac{\varepsilon}{2}.$$

Yet the cardinal of $\widetilde{\Theta}_\varepsilon$ is at most

$$\prod_{i=1}^{D} \frac{(sup_{\theta\in\widetilde{\Theta}}\theta^i - inf_{\theta\in\widetilde{\Theta}}\theta^i)}{\varepsilon} \vee 1 \leq \left(\frac{diam\,\widetilde{\Theta}}{\varepsilon}\right)^D \vee 1.$$

Now, let θ_1 and θ_2 in Θ_K and $x \in \mathbb{R}^d$.

$$\left|\gamma(\theta_1;x) - \gamma(\theta_2;x)\right| \leq \sup_{\theta\in[\theta_1;\theta_2]} \left\| \left(\frac{\partial\gamma}{\partial\theta}\right)_{(\theta;x)} \right\|_\infty \|\theta_1 - \theta_2\|_\infty,$$

since Θ_K *is convex. Moreover,* $L'(x) = \sup_{\theta\in\Theta_K} \left\| \left(\frac{\partial\gamma}{\partial\theta}\right)_{(\theta;x)} \right\|_\infty < \infty$ *for any* $x \in \mathbb{R}^d$ *and*

$$\left|\gamma(\theta_1;x) - \gamma(\theta_2;x)\right| \leq \sup_{\theta\in\Theta_K} \left\| \left(\frac{\partial\gamma}{\partial\theta}\right)_{(\theta;x)} \right\|_\infty \|\theta_1 - \theta_2\|_\infty.$$

Let $\tilde{\theta} \in \widetilde{\Theta}$ *and choose* $\tilde{\theta}_\varepsilon \in \widetilde{\Theta}_\varepsilon$ *such that* $\|\tilde{\theta} - \tilde{\theta}_\varepsilon\|_\infty \leq \frac{\varepsilon}{2}$. *Then,*

$$\forall x \in \mathbb{R}^d, \left|\gamma(\tilde{\theta}_\varepsilon;x) - \gamma(\tilde{\theta};x)\right| \leq L'(x)\frac{\varepsilon}{2}$$

and

$$\gamma(\tilde{\theta}_\varepsilon;x) - \frac{\varepsilon}{2}L'(x) \leq \gamma(\tilde{\theta};x) \leq \gamma(\tilde{\theta}_\varepsilon;x) + \frac{\varepsilon}{2}L'(x).$$

The set of $\varepsilon\|L'\|_r$*-brackets (with respect to the* $\|\cdot\|_r$*-norm)* $\left\{[\gamma(\tilde{\theta}_\varepsilon) - \frac{\varepsilon}{2}L';\gamma(\tilde{\theta}_\varepsilon) + \frac{\varepsilon}{2}L'] : \tilde{\theta}_\varepsilon \in \widetilde{\Theta}_\varepsilon\right\}$ *then has cardinal at most* $\left(\frac{diam\,\widetilde{\Theta}}{\varepsilon}\right)^D \vee 1$ *and covers* $\left\{\gamma(\tilde{\theta}) : \tilde{\theta} \in \widetilde{\Theta}\right\}$, *which yields the result.* ■

Example 5 (Diagonal Gaussian Mixture Model Parameter Space is Convex) *Following the notation of Celeux and Govaert (1995), we write* $[p\lambda_k B_k]$ *for the general diagonal model: the model of Gaussian mixtures which components have diagonal covariance matrices. The mixing proportions are assumed to be equal. No other constraint is imposed. To keep simple notation, let us consider the case* $d = 2$ *and* $K = 2$ *(*$d = 1$ *or* $K = 1$ *are obviously particular cases!). A natural parametrization of this model (which dimension is* 8*) consists of*

$$\theta \in \mathbb{R}^4 \times \mathbb{R}^{+*4} \overset{\varphi}{\longmapsto} \frac{1}{2}\phi\left(\,.\,;\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix},\begin{pmatrix}\theta_5 & 0\\ 0 & \theta_6\end{pmatrix}\right) + \frac{1}{2}\phi\left(\,.\,;\begin{pmatrix}\theta_3\\ \theta_4\end{pmatrix},\begin{pmatrix}\theta_7 & 0\\ 0 & \theta_8\end{pmatrix}\right)$$

With $d = 2$ *and* $K = 2$, $[p\lambda_k B_k] = \varphi(\mathbb{R}^4 \times \mathbb{R}^{+*4})$, *and the parameter space* $\mathbb{R}^4 \times \mathbb{R}^{+*4}$ *is convex.*

Example 6 (The Same Model with Equal Volumes is Convex, too...) $[p\lambda B_k]$ *is the same model as in the previous example, but with the supplementary constraint that the covariance matrices volumes (namely, their respective determinant) have to be equal between components. With* $d = 2$ *and* $K = 2$, *a natural parametrization of this model with dimension* 7 *is*

$$\theta \in \mathbb{R}^4 \times \mathbb{R}^{+*3} \overset{\varphi}{\longmapsto} \frac{1}{2}\phi\left(\,.\,;\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix},\sqrt{\theta_7}\begin{pmatrix}\theta_5 & 0\\ 0 & \frac{1}{\theta_5}\end{pmatrix}\right) + \frac{1}{2}\phi\left(\,.\,;\begin{pmatrix}\theta_3\\ \theta_4\end{pmatrix},\sqrt{\theta_7}\begin{pmatrix}\theta_6 & 0\\ 0 & \frac{1}{\theta_6}\end{pmatrix}\right)$$

With $d = 2$ *and* $K = 2$, $[p\lambda B_k] = \varphi(\mathbb{R}^4 \times \mathbb{R}^{+*3})$, *and the parameter space* $\mathbb{R}^4 \times \mathbb{R}^{+*3}$ *is convex.*

Proof (Lemma 3: Bracketing Entropy, Compact Case) *Let* $O_1, \ldots, O_Q$ *be a finite covering of* Θ *consisting of open balls such that* $\cup_{q=1}^Q O_q \subset \Theta^{\mathcal{O}}$. *Such a covering always exists since* Θ *is assumed to be compact. Remark that*

$$\Theta = \cup_{q=1}^Q (O_q \cap \Theta) \subset \cup_{q=1}^Q conv(O_q \cap \Theta).$$

Now, for any q, conv$(O_q \cap \Theta)$ *is convex and* $\sup_{\theta \in conv(O_q \cap \Theta)} \left\| \left(\frac{\partial \gamma}{\partial \theta}\right)_{(\theta;x)} \right\|_\infty \leq L'(x)$ *since* $conv(O_q \cap \Theta) \subset O_q \subset \Theta^{\mathcal{O}}$*. Therefore, Lemma 2 applies to* $O_q \cap \widetilde{\Theta} \subset conv(O_q \cap \Theta)$*:*

$$N_{[]}(\varepsilon, \{\gamma(\theta) : \theta \in \widetilde{\Theta} \cap O_q\}, \|\cdot\|_r) \leq \left(\frac{\|L\|_r \ diam \widetilde{\Theta}}{\varepsilon} \right)^D \vee 1,$$

Yet, $N_{[]}(\varepsilon, \{\gamma(\theta) : \theta \in \widetilde{\Theta}\}, \|\cdot\|_r) \leq N_{[]}(\varepsilon, \cup_{q=1}^{Q} \{\gamma(\theta) : \theta \in \widetilde{\Theta} \cap O_q\}, \|\cdot\|_r)$ *and the result follows.* ■

Proof (Lemma 4: Bracketing Entropy, Compact Case) *Lemma 2 has to be adapted. Simply replace the last lines of its proof by the following: Assume the same grid* $\widetilde{\Theta}_\varepsilon$ *has been built.*

Now, let θ_1 *and* θ_2 *in* Θ *and* $x \in \mathbb{R}^d$*.*

$$\left|\gamma(\theta_1; x) - \gamma(\theta_2; x)\right|^r \leq \sup_{\theta \in [\theta_1; \theta_2]} \left\| \left(\frac{\partial \gamma}{\partial \theta}\right)_{(\theta;x)} \right\|_\infty^2 \|\theta_1 - \theta_2\|_\infty^2 \big(2 \sup_{\theta \in \{\theta_1, \theta_2\}} |\gamma(\theta; x)|\big)^{r-2},$$

since Θ *is convex. From the definitions of L and L′:*

$$\left|\gamma(\theta_1; x) - \gamma(\theta_2; x)\right|^r \leq L'(x)^2 \|\theta_1 - \theta_2\|_\infty^2 (2\|L\|_\infty)^{r-2}.$$

Let $\tilde{\theta} \in \widetilde{\Theta}$ *and choose* $\tilde{\theta}_\varepsilon \in \widetilde{\Theta}_\varepsilon$ *such that* $\|\tilde{\theta} - \tilde{\theta}_\varepsilon\|_\infty \leq \frac{\varepsilon}{2}$*. Then,*

$$\forall x \in \mathbb{R}^d, \left|\gamma(\tilde{\theta}_\varepsilon; x) - \gamma(\tilde{\theta}; x)\right| \leq L'(x)^{\frac{2}{r}} \left(\frac{\varepsilon}{2}\right)^{\frac{2}{r}} (2\|L\|_\infty)^{\frac{r-2}{r}}$$

and

$$\gamma(\tilde{\theta}_\varepsilon; x) - \varepsilon^{\frac{2}{r}} L'(x)^{\frac{2}{r}} \|L\|_\infty^{\frac{r-2}{r}} 2^{1-\frac{4}{r}} \leq \gamma(\tilde{\theta}; x) \leq \gamma(\tilde{\theta}_\varepsilon; x) + \varepsilon^{\frac{2}{r}} L'(x)^{\frac{2}{r}} \|L\|_\infty^{\frac{r-2}{r}} 2^{1-\frac{4}{r}}.$$

The set of brackets

$$\left\{ \left[\gamma(\tilde{\theta}_\varepsilon; x) - \varepsilon^{\frac{2}{r}} L'(x)^{\frac{2}{r}} \|L\|_\infty^{\frac{r-2}{r}} 2^{1-\frac{4}{r}}; \gamma(\tilde{\theta}_\varepsilon; x) + \varepsilon^{\frac{2}{r}} L'(x)^{\frac{2}{r}} \|L\|_\infty^{\frac{r-2}{r}} 2^{1-\frac{4}{r}} \right] : \tilde{\theta} \in \widetilde{\Theta}_\varepsilon \right\}$$

(of $\|\cdot\|_r$*-norm length* $(2^{2-\frac{4}{r}}) \|L\|_\infty^{\frac{r-2}{r}} \|L'\|_2^{\frac{2}{r}} \varepsilon^{\frac{2}{r}}$*) has cardinal at most* $\left(\frac{diam \widetilde{\Theta}}{\varepsilon}\right)^D \vee 1$ *and covers* $\left\{\gamma(\tilde{\theta}) : \tilde{\theta} \in \widetilde{\Theta}\right\}$*, which yields Lemma 4.*

■

4.3.4 Simulations

A typical difficulty with Gaussian mixtures is that the minimum value of $\gamma_n(\theta)$ could be reached at a boundary value of Θ_K. It happens typically, as one of the Gaussian components is centered at one observation ($\mu_k = X_{\tilde{i}}$) or at a very little (and accidental)

cluster of observations and the variance determinant reaches its minimum value. This $\widehat{\theta}^{\text{MLccE}}$ could have a very low γ_n value (perhaps much lower than $\gamma_n(\theta^0)$), even if it is far from Θ_K^0 and if $\mathbb{E}_{f^*}\left[\gamma(\widehat{\theta}^{\text{MLccE}}; X)\right]$ is much greater than $\mathbb{E}_{f^*}\left[\gamma(\theta_K^0; X)\right]$. Theorem 3 guarantees that this is not a problem from an asymptotic point of view. Actually, because of the lower bound on the covariance matrices, $\gamma(\theta; X_{\tilde{i}})$ cannot get small enough to compensate the improvement of contrast reached on the other observations with a "good" solution (i.e. close to Θ_0) when the total number of observations, and then of "$\gamma(\theta; X_i)$" factors in the likelihood expression, grows to infinity...but we have n fixed! We practically tackle this difficulty by choosing a lower bound on $\det(\Sigma_k)$ great enough such that a solution with a single-observation component cannot compete with a "good" solution (Section 5.1). This requires to know the order of the contrast at its minimum value. Note that if the variances could go to zero, the contrast $\gamma_n(\theta)$ would not even be bounded, if $K > 1$: if a component mean is $X_{\tilde{i}}$ for any $\tilde{i}$, and if the corresponding variance goes to zero, then the contrast goes to minus infinity. Hence the lower bound on the covariance matrices is, once more, a minimal condition.

Example 7 [Example 4 continued]

We can practically check the convergence foreseen with Theorem 3 for the simple Example 4. Figure 4.6 illustrates the convergence (in probability) of $\widehat{\theta}^{MLccE}$ in this setting.

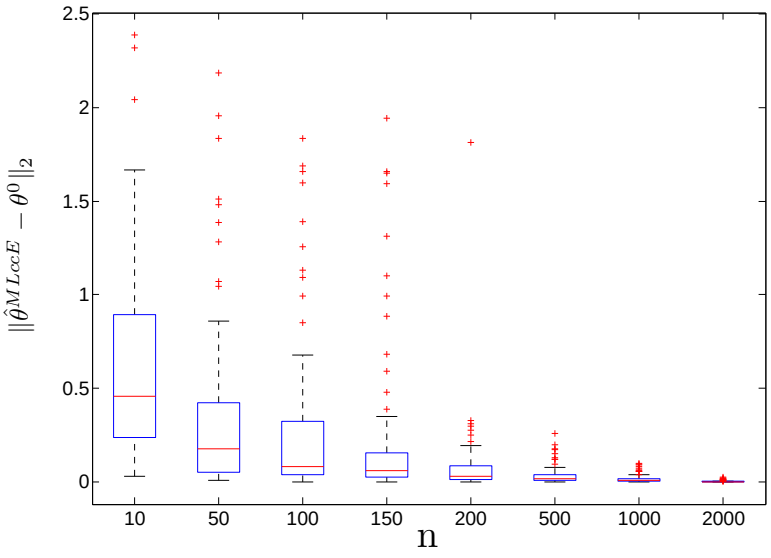


Figure 4.6: $\|\widehat{\theta}^{\text{MLccE}} - \theta^0\|_2$ boxplots for 100 experiences for different values of n, with the Example 4 settings

4.4 Model Selection

As illustrated by Example 4, model selection is a crucial step in this framework. Actually, the number of classes (and then the number of components in the current setting) may

even be the quantity of interest of the study. Anyhow, a relevant number of classes must obviously be chosen so as to design a good unsupervised classification. Moreover, we are especially interested in this step since our main interest in this chapter is the ICL criterion.

Model selection procedures of concern here are the penalized conditional classification likelihood criteria. They are of the form considered in the general contrast minimization framework introduced in Section 3.1.2, applied to the contrast at hand in this chapter:

$$\mathrm{crit}(K) = -\log \mathrm{L_{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}) + \mathrm{pen}(K).$$

The collection of models at hand is $\{\mathcal{M}_K\}_{1\leq K\leq K_M}$. Each $\mathcal{M}_K$ is a parametric mixture model with K components and the corresponding parameter space is Θ_K. It is assumed to be optimal, like in the previous section: its dimension D_K is the number of free parameters of the model. The same notation is used as in the previous section. Likewise, the application we have in view is the $-\log \mathrm{L_{cc}}$ contrast in the Gaussian mixture model framework. However, most results are stated for a general contrast γ in a general parametric models collection. Then the general assumptions of those results are discussed in the particular framework we consider to derive sufficient conditions which guarantee they hold.

In Section 4.4.1, the "consistency" of such a model selection procedure for the number of components minimizing the loss function ("identification") is proved for a class of penalties, which, not surprisingly, are not different from those defined by Nishii (1988) or Keribin (2000) in a maximum usual likelihood context. Sufficient conditions are given in Theorem 6. The heaviest condition (B4) may be guaranteed under regularity and (weak) identifiability assumptions in our framework. This involves a study relying on results of Massart (2007) and will be discussed and proved in Section 4.4.2. Remark that this approach has the advantage that it is the first step of the work to do to reach a non-asymptotic result, which we however have not proved yet. Assumptions about Gaussian mixture models ensuring the consistency of the model selection procedure are given in Theorem 7, Section 4.4.1. Proofs are given in Section 4.4.3.

4.4.1 Consistent Penalized Criteria

Assume that K_0 exists such that

$$\text{and} \quad \begin{array}{l} \forall K < K_0, \inf\limits_{\theta\in\Theta_{K_0}} \mathbb{E}_{f^{\wp}}\left[-\log \mathrm{L_{cc}}(\theta)\right] < \inf\limits_{\theta\in\Theta_K} \mathbb{E}_{f^{\wp}}\left[-\log \mathrm{L_{cc}}(\theta)\right] \\ \forall K \geq K_0, \inf\limits_{\theta\in\Theta_{K_0}} \mathbb{E}_{f^{\wp}}\left[-\log \mathrm{L_{cc}}(\theta)\right] \leq \inf\limits_{\theta\in\Theta_K} \mathbb{E}_{f^{\wp}}\left[-\log \mathrm{L_{cc}}(\theta)\right] \end{array}$$

which means that the bias of the models is stationary from the model $\mathcal{M}_{K_0}$. There exists a "best" model from the approximation point of view, and the models which are more complex than it reach at best as good approximation results as it. Remark that the last property should hold mostly: if the models were not constrained — or if the constraints are chosen consequently — it would be expected that the approximation properties improve as the complexity of the model grows. Under this assumption, a model selection procedure is expected to recover asymptotically K_0, i.e. to be *consistent*. This is an *identification* goal. Indeed, it would obviously be disastrous for a model

selection criterion to select a model which does not minimize (or almost minimize...) the bias. But, from this identification point of view, it is besides assumed that the model $\mathcal{M}_{K_0}$ contains all the interesting information about the distribution of the data (typically, the structure of classes in our framework), and that choosing a more complex model only increases the variance, without providing any further information: this is why exactly K_0, and not any larger K, should be recovered. The efficiency point of view would be different (see Section 3.1.2).

Let us stress at this stage that the "true" number of components of $f^\wp$ is not directly of concern in the statement of this problem: it is in particular not assumed that it equals K_0, and is not even assumed to be defined ($f^\wp$ does not have to be a Gaussian mixture). We shall derive sufficient conditions for a penalized model selection procedure to be consistent.

Most of the ideas and techniques employed in this section are from the books of Massart (2007) and van der Vaart (1998), and most of the presented results either directly come from those books (it is then specified) or are inspired from theirs.

Theorem 6
$\{\Theta_K\}_{1\leq K\leq K_M}$ a collection of parametric models. For any K, $\Theta_K \subset \mathbb{R}^{D_K}$. Assume the models are ranked by increasing complexity: $D_1 \leq \cdots \leq D_{K_M}$.
For any K, $\Theta_K^0 = \underset{\theta\in\Theta_K}{\operatorname{argmin}} \mathbb{E}_{f^\wp}[\gamma(\theta)]$. Let $\theta_K^0 \in \Theta_K^0$.

Assume that:

$$K_0 = \min \underset{1\leq K\leq K_M}{\operatorname{argmin}} \mathbb{E}_{f^\wp}\left[\gamma(\Theta_K^0)\right]. \tag{B1}$$

$$\begin{gathered}\forall K,\ \hat{\theta}_K \in \Theta_K,\ \textit{defined such that } \gamma_n(\hat{\theta}_K) \leq \gamma_n(\theta_K^0) + o_{\mathbb{P}}(1),\\ \textit{fulfills } \gamma_n(\hat{\theta}_K) \xrightarrow{\mathbb{P}} \mathbb{E}_{f^\wp}\left[\gamma(\theta_K^0)\right].\end{gathered} \tag{B2}$$

$$\forall K, \begin{cases} \operatorname{pen}(K) > 0 \textit{ and } \operatorname{pen}(K) = o_{\mathbb{P}}(1) & \textit{when } n \to +\infty \\ n\big(\operatorname{pen}(K) - \operatorname{pen}(K')\big) \xrightarrow[n\to+\infty]{\mathbb{P}} \infty & \textit{when } K > K' \end{cases} \tag{B3}$$

$$n\big(\gamma_n(\hat{\theta}_{K_0}) - \gamma_n(\hat{\theta}_K)\big) = O_{\mathbb{P}}(1) \textit{ for any } K \in \underset{1\leq K\leq K_M}{\operatorname{argmin}} \mathbb{E}_{f^\wp}\left[\gamma(\Theta_K^0)\right]. \tag{B4}$$

Define $\hat{K}$ such that

$$\hat{K} = \underset{1\leq K\leq K_M}{\operatorname{argmin}} \Big\{ \underbrace{\gamma_n(\hat{\theta}_K) + \operatorname{pen}(K)}_{\operatorname{crit}(K)} \Big\}.$$

Then

$$\mathbb{P}[\hat{K} \neq K_0] \xrightarrow[n\longrightarrow\infty]{} 0.$$

Sketch of Proof
It is first proved that $\hat{K}$ cannot asymptotically "underestimate" K_0. Suppose $\mathbb{E}_{f^\wp}[\gamma(\theta_K^0)] > \mathbb{E}_{f^\wp}\left[\gamma(\theta_{K_0}^0)\right]$. Then, from (B2), *$\big(\gamma_n(\hat{\theta}_K) - \gamma_n(\hat{\theta}_{K_0})\big)$ is asymptotically of order $\mathbb{E}_{f^\wp}[\gamma(\theta_K^0)] - \mathbb{E}_{f^\wp}\left[\gamma(\theta_{K_0}^0)\right] > 0$. Since the penalty is $o_{\mathbb{P}}(1)$ from* (B3), *$\operatorname{crit}(K_0) < \operatorname{crit}(K)$ asymptotically and $\hat{K} > K$.*

The proof that $\hat{K}$ does not asymptotically "overestimate" K_0 involves the heaviest assumption (B4). *It is more involved since then* $\left(\mathbb{E}_{f^\wp}\left[\gamma(\theta^0_K)\right] - \mathbb{E}_{f^\wp}\left[\gamma(\theta^0_{K_0})\right]\right)$ *is zero. The fluctuations of* $\left(\gamma_n(\hat{\theta}_K) - \gamma_n(\hat{\theta}_{K_0})\right)$ *then have to be evaluated to calibrate a penalty which be large enough to cancel them. According to* (B4), *a penalty larger than* $\frac{1}{n}$ *should be enough.* (B3) *guarantees this condition.* ■

Assumption (B1) is necessary so that this identification point of view makes sense.

Assumption (B3) defines the range of possible penalties.

Assumption (B2) is guaranteed under assumption (A3) of Theorem 3 and from the definition of $\hat{\theta}_K$:

Lemma 5
$\Theta \subset \mathbb{R}^D$ *and* $\gamma : \Theta \times \mathbb{R}^d \to \mathbb{R}$. $\theta^0 \in \Theta^0 = \underset{\theta\in\Theta}{\operatorname{argmin}} \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right]$. *Assume*

$$\gamma_n(\hat{\theta}) \leq \gamma_n(\theta^0) + o_{\mathbb{P}}(1)$$

and

$$\sup_{\theta\in\Theta}\left|\gamma_n(\theta) - \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right]\right| \xrightarrow{\mathbb{P}} 0$$

(this is assumption (A3) *of Theorem 3).*

Then (B2) *holds:*

$$\gamma_n(\hat{\theta}) \xrightarrow{\mathbb{P}} \mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right].$$

The proof of this Lemma is a uniform convergence argument which relies on assumption (A3): asymptoticaly, minimizing $\theta \mapsto \gamma_n(\theta)$ must not be very different from minimizing $\theta \mapsto \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right]$ since they are uniformly close to each other.

Assumption (B4) is the heaviest assumption. Section 4.4.2 is devoted to deriving sufficient conditions so that it holds. It will justify the following result:

Theorem 7
Let $\left(\mathcal{M}_K\right)_{1\leq K\leq K_M}$ *be the collection of Gaussian mixture models introduced in Section 4.1.1, with corresponding parameter spaces* $\left(\Theta_K\right)_{K\in\{1,\ldots,K_M\}}$. Θ_K *is assumed to be compact for any* K.
Let for any K

$$\Theta^0_K = \underset{\theta\in\Theta_K}{\operatorname{argmin}}\, \mathbb{E}_{f^\wp}\left[-\log L_{cc}(\theta)\right].$$

Define

$$K_0 = \min \underset{1\leq K\leq K_M}{\operatorname{argmin}}\, \mathbb{E}_{f^\wp}\left[-\log L_{cc}(\Theta^0_K)\right].$$

Assume, $\forall K, \forall\theta \in \Theta_K, \forall\theta^0_{K_0} \in \Theta^0_{K_0}$,

$$\mathbb{E}_{f^\wp}\left[-\log L_{cc}(\theta)\right] = \mathbb{E}_{f^\wp}\left[-\log L_{cc}(\Theta^0_{K_0})\right] \Longleftrightarrow -\log L_{cc}(\theta; x) = -\log L_{cc}(\theta^0_{K_0}; x) \quad f^\wp d\lambda - a.s.$$

Let for any K, $\Theta_K^{\mathcal{O}}$ be an open subset of $\mathbb{R}^{D_K}$ over which $\log L_{cc}$ is well-defined, such that $\Theta_K \subset \Theta_K^{\mathcal{O}}$ and assume

$$\forall K \in \{1,\dots,K_M\}, \begin{cases} L_K(x) = \sup\limits_{\theta\in\Theta_K^{\mathcal{O}}} \left|\log L_{cc}(\theta;x)\right| < \infty \quad f^{\wp}d\mu - a.s. \\ \|L_K\|_\infty < \infty. \end{cases}$$

and

$$\forall K \in \{1,\dots,K_M\}, \begin{cases} L'_K(x) = \sup\limits_{\theta\in\Theta_K^{\mathcal{O}}} \left\| \left(\frac{\partial \log L_{cc}}{\partial\theta}\right)_{(\theta;x)} \right\|_\infty < \infty \quad f^{\wp}d\mu - a.s. \\ \|L'_K\|_2 < \infty. \end{cases}$$

Assume that $\forall K, \forall \theta_K^0 \in \Theta_K^0$,

$$I_{\theta_K^0} = \frac{\partial^2}{\partial\theta^2}\left(\mathbb{E}_{f^{\wp}}\left[-\log L_{cc}(\theta)\right]\right)_{|\theta_K^0} \text{ is nonsingular.}$$

Let for any K and n, $\widehat{\theta}_K^{MLccE} = \widehat{\theta}_K^{MLccE}(X_1,\dots,X_n) \in \Theta_K$ such that

$$-\log L_{cc}(\widehat{\theta}_K^{MLccE};\mathbf{X}) \le -\log L_{cc}(\theta_K^0;\mathbf{X}) + o_{\mathbb{P}}(n).$$

Let $\text{pen} : \{1,\dots,K_M\} \longrightarrow \mathbb{R}^+$ (which depends on n, Θ_K, and may depend on the data) such that

$$\forall K \in \{1,\dots,K_M\}, \begin{cases} \text{pen}(K) > 0 \text{ and } \text{pen}(K) = o_{\mathbb{P}}(n) & \text{when } n \to +\infty \\ \big(\text{pen}(K) - \text{pen}(K')\big) \xrightarrow[n\to+\infty]{\mathbb{P}} \infty & \text{for any } K' < K. \end{cases}$$

Select $\hat{K}$ such that

$$\hat{K} = \underset{1\le K\le K_M}{\text{argmin}} \left\{-\log L_{cc}(\widehat{\theta}_K^{MLccE}) + \text{pen}(K)\right\}.$$

Then

$$\mathbb{P}[\hat{K} \neq K_0] \xrightarrow[n\longrightarrow\infty]{} 0.$$

It is implicitely assumed that $\mathbb{E}_{f^{\wp}}\left[\left|\log \mathrm{L}_{\mathrm{cc}}(\theta)\right|\right] < \infty$ for any θ in any Θ_K. This is a very mild assumption.

Remark that, if Θ_K is convex, L and L' can be defined as suprema over Θ_K instead of $\Theta_K^{\mathcal{O}}$ and there is no need to involve the sets $\Theta_K^{\mathcal{O}}$. This follows from the proofs of Corollaries 2 and 3 below.

A new identifiability assumption is introduced:

$$\forall K, \forall\theta \in \Theta_K, \forall \theta_{K_0}^0 \in \Theta_{K_0}^0,$$
$$\mathbb{E}_{f^{\wp}}\left[-\log \mathrm{L}_{\mathrm{cc}}(\theta)\right] = \mathbb{E}_{f^{\wp}}\left[-\log \mathrm{L}_{\mathrm{cc}}(\Theta_{K_0}^0)\right] \Longleftrightarrow -\log \mathrm{L}_{\mathrm{cc}}(\theta;x) = -\log \mathrm{L}_{\mathrm{cc}}(\theta_{K_0}^0;x)$$
$$f^{\wp}d\lambda - \text{a.s.}$$

It is a reasonable assumption: as expected, no further identifiability assumption is imposed on the parameter directly, and notably, the label switching phenomenon is no problem here. But it is necessary for this identification point of view to make sense, that a single value of the *contrast function* $x \longmapsto \gamma(\theta; x)$ minimizes the loss function. Remark that in the maximum likelihood framework, such an assumption is guaranteed to hold at least if any model contains the true distribution of the data, since then the loss function — the Kullback-Leibler divergence — is uniquely minimized at the density corresponding to the true distribution. Obviously, several values of the parameter, perhaps in different models, may give rise to this density, besides the label switching. We do not know any such result with the $-\log \mathrm{L_{cc}}$ contrast, and simply hypothesize that the assumption holds.

The assumption about the nonsingularity of I_{θ^0} is unpleasant, since it is hard to be guaranteed. Hopefully, it could be weakened. The result of Massart (2007) which inspires this, and is available in a general maximum (usual) likelihood context, does not require such an assumption. This is Theorem 7.11 (recalled in Section A.1), which is a much stronger result since it provides a non-asymptotic oracle inequality if the oracle risk can be linked to the minimal Kullback-Leibler divergence to the true distribution among the models. It relies on a clever choice of the involved distances (Hellinger distances between the density functions instead of the parametric point of view embraced here), and on particular properties of the log function. However, this is an usual assumption and the works of Nishii (1988) and Keribin (2000), which are discussed below, for example, rely on similar assumptions.

This result of Massart (2007) moreover does not require the contrast (i.e. the likelihood) to be bounded, as we have to. Remark however that this result involves both Hellinger distances and Kullback-Leibler divergence. To make it homogeneous from this point of view, it seems that an assumption similar to the boundedness of the contrast is necessary: Lemma 7.23 in Massart (2007) then enables to conclude and obtain an oracle inequality from the first-mentioned Theorem 7.11, which only involves Hellinger distances. Maugis and Michel (2009), who apply this result in a framework combining clustering and variable selection in a Gaussian mixture model selection framework, point this difficulty out, too. So that it seems reasonable that the assumptions about L and L' (the last is much milder than the former) be necessary. They are typically ensured if either the contrast is bounded (replace $-\log \mathrm{L_{cc}}$ by $(-\log \mathrm{L_{cc}}) \wedge M$ for $M > 0$ well-chosen and large enough) or if the support of $f^{\wp}$ is bounded.

Remark that the conditions about the penalty form we derive are analogous to that of Nishii (1988) or Keribin (2000), which are both derived in the maximum likelihood framework. As those of Keribin (2000), they can be regarded as generalizing those of Nishii (1988) when the considered models are Gaussian mixture models. Indeed, Nishii (1988) considers penalties of the form $c_n D_K$ and proves the model selection procedure to be weakly consistent if $\frac{c_n}{n} \to 0$ and $c_n \to \infty$. Those results are given in a general maximum likelihood framework, with perhaps misspecified models. Note that Nishii (1988) assumes the parameter space to be convex. He moreover notably assumes the uniqueness of the quasi true parameter (this is assuming $\Theta_K^0 = \{\theta_K^0\}$) and that $I_{\theta_K^0}$ is nonsingular (with the usual likelihood contrast), together with other regularity assumptions. Those results are not particularly designed for mixture models. Instead, as we do, Keribin (2000) considers general penalty forms and proves the procedure to be con-

sistent if $\frac{\text{pen}(K)}{n} \xrightarrow[n\to\infty]{} 0$, $\text{pen}(K) \xrightarrow[n\to\infty]{} \infty$ and $\liminf_{n\to\infty} \frac{\text{pen}(K)}{\text{pen}(K')} > 1$ if $K > K'$. These conditions are equivalent to Nishii's if $\text{pen}(K) = c_n D_K$. In a general mixture model framework, she assumes the model family to be well-specified, the same notion of identifiability as we do, and a condition which does not seem to be directly comparable to ours about $I_{\theta_K^0}$ but which tastes roughly the same. It might be milder. Those assumptions are proved to hold with the usual likelihood contrast in the Gaussian mixture model framework if the covariance matrices are proportional to the identity and are the same for all components, and if the parameter space is compact. Our conditions about the penalty are a little weaker than Keribin's, but they still are quite analogous. Moreover, as compared to those results, we notably have to keep the proportions away from zero. This is necessary because of the entropy term in the $-\log \mathrm{L}_{\mathrm{cc}}$ contrast.

The strong version of Theorem 7, which would state the almost sure consistency of $\hat{K}$ to K_0, would then probably involve penalties a little heavier, as Nishii (1988) and Keribin (2000) proved in their respective frameworks. Both had to assume besides that $\frac{\text{pen}(K)}{\log\log n} \to \infty$.

Theorem 7 is a direct consequence of Theorem 6, Lemma 5, Theorem 4, which can be applied under those assumptions, and of Corollary 3 below and the discussion about its assumptions along the lines of Section 4.4.2.

4.4.2 Sufficient Conditions to Ensure Assumption (B4)

Let us introduce the notation $S_n\gamma(\theta) = n\gamma_n(\theta) - n\mathbb{E}_{f^\wp}[\gamma(\theta; X)]$.

The main result of this section is Lemma 6. Some intermediate results which enable to link Lemma 6 to Theorem 6 via Assumption (B4) are stated as corollaries and proved subsequently. Lemma 6 povides a control of

$$\sup_{\theta\in\Theta} \frac{S_n\big(\gamma(\theta^0) - \gamma(\theta)\big)}{\|\theta^0 - \theta\|_\infty^2 + \beta^2}$$

(with respect to β) and then of

$$\frac{S_n\big(\gamma(\theta^0) - \gamma(\hat{\theta})\big)}{\|\theta^0 - \hat{\theta}\|_\infty^2 + \beta^2}.$$

With a good choice of β, and if $S_n\big(\gamma(\theta^0) - \gamma(\hat{\theta})\big)$ can be linked to $\|\theta^0 - \hat{\theta}\|_\infty^2$, it is proved in Corollary 2 that it may then be assessed that

$$n\|\hat{\theta} - \theta^0\|_\infty^2 = O_{\mathbb{P}}(1).$$

Plugging this last property back into the result of Lemma 6 yields (Corollary 3)

$$n\big(\gamma_n(\theta_K^0) - \gamma_n(\hat{\theta}_K)\big) = O_{\mathbb{P}}(1)$$

for any model $K \in \operatorname*{argmin}_{1\le K\le K_M} \mathbb{E}_{f^\wp}[\gamma(\theta_K^0)]$ and then, under mild identifiability condition,

$$n\big(\gamma_n(\theta_{K_0}^0) - \gamma_n(\hat{\theta}_K)\big) = O_{\mathbb{P}}(1),$$

which is Assumption (B4).

We shall prove the

Lemma 6
Let $D \in \mathbb{N}^*$ and $\Theta \subset \mathbb{R}^D$ assumed to be convex.
Let $\Theta^{\mathcal{O}}$ be an open subset of $\mathbb{R}^D$ such that $\Theta \subset \Theta^{\mathcal{O}}$ and $\gamma : \Theta^{\mathcal{O}} \times \mathbb{R}^d \longrightarrow \mathbb{R}$.
$\theta \in \Theta^{\mathcal{O}} \longmapsto \gamma(\theta; x)$ is assumed to be C^1 over $\Theta^{\mathcal{O}}$ for $f^{\wp}d\lambda$-almost all $x \in \mathbb{R}^d$.
Let $\theta^0 \in \Theta$ such that $\mathbb{E}_{f^{\wp}}[\gamma(\theta^0)] = \inf_{\theta\in\Theta} \mathbb{E}_{f^{\wp}}[\gamma(\theta)]$.
Assume

$$L(x) = \sup_{\theta\in\Theta} |\gamma(\theta; x)| < \infty \quad f^{\wp}d\mu - a.s.$$
$$\|L\|_\infty = \operatorname*{ess\,sup}_{X\sim f^{\wp}} L(X) < \infty.$$

and

$$L'(x) = \sup_{\theta\in\Theta} \left\| \left(\frac{\partial\gamma}{\partial\theta}\right)_{(\theta;x)} \right\|_\infty < \infty \quad f^{\wp}d\mu - a.s.$$
$$\|L'\|_2 = \mathbb{E}_{f^{\wp}}\left[L'(X)^2\right]^{\frac{1}{2}} < \infty.$$

Then $\exists\alpha > 0 / \forall n, \forall\beta > 0, \forall\eta > 0$,

$$\sup_{\theta\in\Theta} \frac{S_n(\gamma(\theta^0) - \gamma(\theta))}{\|\theta^0 - \theta\|_\infty^2 + \beta^2} \leq \frac{\alpha}{\beta^2}\Bigg(\|L'\|_2\beta\sqrt{nD} + \big(\|L\|_\infty + \|L'\|_2\beta\big)D + \|L'\|_2\sqrt{n\eta}\beta + \|L\|_\infty\eta \Bigg)$$

holds with probability greater than $(1 - e^{-\eta})$.

Note that α is an absolute constant which notably does not depend on θ^0.

Sketch of Proof *The proof of Lemma 6 relies on the results of Massart (2007) presented below and on the evaluation of the bracketing entropy of the class of functions at hand we derived in Section 4.3.3. Lemma 4 provides a local control of the entropy and hence, through Theorem 8, a control of the supremum of $S_n(\gamma(\theta^0) - \gamma(\theta))$ as $\|\theta - \theta^0\|_\infty^2 < \sigma$, with respect to σ. The "peeling" Lemma 9 then enables to take advantage of this local control to derive a fine global control of $\sup_{\theta\in\Theta} \frac{S_n(\gamma(\theta^0)-\gamma(\theta))}{\|\theta-\theta^0\|^2+\beta^2}$, for any β^2. This control in expectation, which can be derived conditionaly to any event A, yields a control in probability thanks to Lemma 8, which can be thought of as an application of Markov's inequality.* ■

The proof of Lemma 6 can be refined a little along the lines of the proof of Theorem 7.11 of Massart (2007) to get the

Lemma 7
Under the same assumptions as Lemma 6, define $\sigma_0 = \sqrt{\frac{D}{n}}$. *Then,*

$$\exists \alpha > 0 / \forall n, \forall \eta > 0, \forall \beta > \sigma_0,$$
$$\sup_{\theta \in \Theta} \frac{S_n(\gamma(\theta^0) - \gamma(\theta))}{\|\theta^0 - \theta\|_\infty^2 + \beta^2} \leq \frac{\alpha}{\beta^2} \Bigg(\big(\|L'\|_2 + \|L\|_\infty \big) \beta \sqrt{nD} + \|L'\|_2 \sqrt{\eta n} \beta + \|L\|_\infty \eta \Bigg)$$

holds with probability larger than $1 - e^{-\eta}$.

α *is an absolute constant.*

This lemma is a refinement of Lemma 6 which is not necessary for the following results. However, by analogy with the results of Massart (2007), it is interesting since, together with further calculations, it provides clues that the optimal penalty to choose to obtain an oracle result would probably be proportional to D. The feature we are interested in, beside the form of the upper bound, is that it does not involve D terms, but only $\sqrt{Dn}$. We did not derive such an oracle inequality up to now, though. This shall be further discussed in Section 4.4.4. The proof of this version of the lemma is given apart from the previous one, for the sake of readability of the proofs.

Corollary 2
Let $D \in \mathbb{N}^*$ *and* $\Theta \subset \mathbb{R}^D$.
Let $\Theta^{\mathcal{O}}$ *be an open subset of* $\mathbb{R}^D$ *such that* $\Theta \subset \Theta^{\mathcal{O}}$ *and* $\gamma : \Theta^{\mathcal{O}} \times \mathbb{R}^d \longrightarrow \mathbb{R}$.
$\theta \in \Theta^{\mathcal{O}} \longmapsto \gamma(\theta; x)$ *is assumed to be* C^1 *over* $\Theta^{\mathcal{O}}$ *for* $f^{\wp} d\lambda$*-almost all* $x \in \mathbb{R}^d$.
Let $\theta^0 \in \Theta$ *such that* $\mathbb{E}_{f^{\wp}}[\gamma(\theta^0)] = \inf_{\theta \in \Theta} \mathbb{E}_{f^{\wp}}[\gamma(\theta)]$.
Assume

$$L(x) = \sup_{\theta \in \Theta^{\mathcal{O}}} |\gamma(\theta; x)| < \infty \quad f^{\wp} d\mu - a.s.$$
$$\|L\|_\infty < \infty.$$

and

$$L'(x) = \sup_{\theta \in \Theta^{\mathcal{O}}} \left\| \left(\frac{\partial \gamma}{\partial \theta} \right)_{(\theta; x)} \right\|_\infty < \infty \quad f^{\wp} d\mu - a.s.$$
$$\|L'\|_2 < \infty.$$

Assume moreover that $I_{\theta^0} = \frac{\partial^2}{\partial \theta^2} \left(\mathbb{E}_{f^{\wp}}[\gamma(\theta)] \right)_{|\theta^0}$ *is nonsingular.*
Let $(\hat{\theta}_n)_{n \geq 1}$ *such that* $\hat{\theta}_n \in \Theta$,

$$\gamma_n(\hat{\theta}_n) \leq \gamma_n(\theta^0) + O_{\mathbb{P}}(\frac{1}{n})$$

and

$$\hat{\theta}_n \xrightarrow[n \to \infty]{\mathbb{P}} \theta^0.$$

Then,

$$n\|\hat{\theta}_n - \theta^0\|_\infty^2 = O_{\mathbb{P}}(1).$$

The constant involved in $O_{\mathbb{P}}(1)$ *depends on* D, $\|L\|_\infty$, $\|L'\|_2$ *and* I_{θ^0}.

This is a direct consequence of Lemma 6: it suffices to choose β well (see Section 4.4.3). The dependency of $O_{\mathbb{P}}(1)$ in D, $\|L\|_\infty$, $\|L'\|_2$ and I_{θ^0} is not a problem since we aim at deriving an asymptotic result: the order of $\|\theta - \theta^0\|_\infty^2$ with respect to n when the model is fixed is of concern.

The assumption that I_{θ^0} is nonsingular plays an analogous role as Assumption (A2) in Theorem 3: it ensures that $\mathbb{E}_{f^\wp}[\gamma(\theta)]$ cannot be close to $\mathbb{E}_{f^\wp}[\gamma(\theta^0)]$ if θ is not close to θ^0. But this stronger assumption is necessary to strengthen the conclusion: the rate of the relation between $\mathbb{E}_{f^\wp}[\gamma(\theta)] - \mathbb{E}_{f^\wp}[\gamma(\theta^0)]$ and $\|\theta - \theta^0\|$ can then be controlled...

Remark that, should this assumption fail, $\exists \tilde{\theta} \in \Theta / \tilde{\theta}' I_{\theta^0} \tilde{\theta} = 0 \Rightarrow \mathbb{E}_{f^\wp}\left[\gamma(\theta^0 + \lambda\tilde{\theta})\right] = \mathbb{E}_{f^\wp}[\theta^0] + o(\lambda^2)$ and then there is no hope to have $\alpha > 0$ such that $\mathbb{E}_{f^\wp}[\gamma(\theta)] - \mathbb{E}_{f^\wp}[\gamma(\theta^0)] > \alpha\|\theta - \theta^0\|^2$... The result cannot be derived in this case. This means that this approach cannot be applied without this — admittedly unpleasant — assumption. It might perhaps be an other approach (with different distances, not involving the parameters but rather directly the contrast values) which would enable to avoid it, as Massart (2007) did in the likelihood framework.

Corollary 3
$\{\Theta_K\}_{1 \le K \le K_M}$ *a collection of parametric models. For any* K, $\Theta_K \subset \mathbb{R}^{D_K}$. *Assume the models are ranked by increasing complexity:* $D_1 \le \cdots \le D_{K_M}$.
For any K, *assume there exists an open set* $\Theta_K^{\mathcal{O}} \subset \mathbb{R}^{D_K}$ *such that* $\Theta_K \subset \Theta_K^{\mathcal{O}}$ *and such that with* $\Theta^{\mathcal{O}} = \Theta_1^{\mathcal{O}} \cup \cdots \cup \Theta_{K_M}^{\mathcal{O}}$, $\gamma : \Theta^{\mathcal{O}} \times \mathbb{R}^d \longrightarrow \mathbb{R}$ *is well-defined and* C^1 *over* $\Theta^{\mathcal{O}}$ *for* $f^\wp d\lambda$*-almost all* $x \in \mathbb{R}^d$.
Assume

$$L(x) = \sup_{\theta \in \Theta^{\mathcal{O}}} |\gamma(\theta; x)| < \infty \quad f^\wp d\mu - a.s.$$
$$\|L\|_\infty < \infty.$$

and

$$L'(x) = \sup_{\theta \in \Theta^{\mathcal{O}}} \left\| \left(\frac{\partial \gamma}{\partial \theta} \right)_{(\theta;x)} \right\|_\infty < \infty \quad f^\wp d\mu - a.s.$$
$$\|L'\|_2 < \infty.$$

For any K, $\Theta_K^0 = \operatorname{argmin}_{\theta \in \Theta_K} \mathbb{E}_{f^\wp}[\gamma(\theta)]$. *Let* $\theta_K^0 \in \Theta_K^0$.
Let $K_0 = \min \operatorname{argmin}_{1 \le K \le K_M} \mathbb{E}_{f^\wp}[\gamma(\Theta_K^0)]$.
Assume $\forall K, \forall \theta \in \Theta_K$,

$$\mathbb{E}_{f^\wp}[\gamma(\theta)] = \mathbb{E}_{f^\wp}\left[\gamma(\theta_{K_0}^0)\right] \Longleftrightarrow \gamma(\theta) = \gamma(\theta_K^0) \quad f^\wp d\lambda - a.s.$$

Let $\mathcal{K} = \left\{K \in \{1, \dots, K_M\} : \mathbb{E}_{f^\wp}\left[\gamma(\theta_K^0)\right] = \mathbb{E}_{f^\wp}\left[\gamma(\theta_{K_0}^0)\right]\right\}$.
For any $K \in \mathcal{K}$, *let* $\hat{\theta}_K = \hat{\theta}_K(X_1, \dots, X_n) \in \Theta_K$ *such that*

$$\gamma_n(\hat{\theta}_K) \leq \gamma_n(\theta_K^0) + O_{\mathbb{P}}(\frac{1}{n})$$
$$\hat{\theta}_K \xrightarrow[n\to\infty]{\mathbb{P}} \theta_K^0.$$

Moreover, assume that

$$I_{\theta_K^0} = \frac{\partial^2}{\partial\theta^2}\left(\mathbb{E}_{f^\wp}\left[\gamma(\theta)\right]\right)_{|\theta_K^0}$$

is nonsingular for any $K \in \mathcal{K}$.

Then

$$\forall K \in \mathcal{K}, \quad n\big(\gamma_n(\hat{\theta}_{K_0}) - \gamma_n(\hat{\theta}_K)\big) = O_{\mathbb{P}}(1).$$

This last corollary states conditions under which assumption (B4) of Theorem 6 is ensured.

Let us go back to the proof of Lemma 6. It is inspired from the proof of Theorem 7.11 in Massart (2007). It relies on three results from the same book, which are recalled now. They are not proved in the proofs section: the reader is referred to Massart (2007).

Lemma 8 (Lemma 2.4 in Massart, 2007)
Let $Z \in L_1(\mathbb{R})$. *Let* $\varphi : \mathbb{R}^+ \longmapsto \mathbb{R}$ *increasing such that for all measurable set* A *with* $\mathbb{P}[A] > 0$,

$$\mathbb{E}^A[Z] \leq \varphi\left(\log\frac{1}{\mathbb{P}[A]}\right),$$

where $\mathbb{E}^A[Z] = \frac{\mathbb{E}[Z1_A]}{\mathbb{P}[A]}$. *Then:*

$$\forall x > 0, \mathbb{P}\left[Z \geq \varphi(x)\right] \leq e^{-x}.$$

This is a simple but clever application of Markov's inequality, which will be helpful to derive results with large probability from techniques which yield results in expectation.

The next lemma, which is an essential ingredient of the result, is the

Lemma 9 ("Pealing Lemma", 4.23 in Massart, 2007)
Let S *be a countable set,* $u \in S$, $a : S \longrightarrow \mathbb{R}^+$ *such that* $a(u) = \inf_{t\in S} a(t)$. Z *a process indexed by* S. *Assume* $\forall\sigma > 0, \mathbb{E}\left[\sup_{t\in\mathcal{B}(\sigma)} Z(t) - Z(u)\right] < \infty$, *with* $\mathcal{B}(\sigma) = \{t \in S; a(t) \leq \sigma\}$. *Then, for any function* ψ *on* $\mathbb{R}^+$ *such that* $\frac{\psi(x)}{x}$ *is nonincreasing on* $\mathbb{R}^+$ *and fulfills*

$$\forall\sigma \geq \sigma_0, \quad \mathbb{E}\left[\sup_{t\in\mathcal{B}(\sigma)} Z(t) - Z(u)\right] \leq \psi(\sigma),$$

one has for any $x > \sigma_0$:

$$\mathbb{E}\left[\sup_{t\in S}\frac{Z(t) - Z(u)}{a^2(t) + x^2}\right] \leq 4x^{-2}\psi(x).$$

This lemma enables to derive a finer control of the global increments of a process from a control of its local increments. Actually, it allows to link the increments of the process from a point to another with respect to the distance between those two points, when the increments at any given distance from the first point are controlled. It will be applied to the (centered) difference of the empirical contrast between θ^0 and an estimator of it. This difference is expected to go to zero at a rate at least $\frac{1}{\sqrt{n}}$ because of the Central Limit Theorem. Actually, it will be proved thanks to this lemma that this convergence is accelerated because of the convergence of the estimator to θ^0 and that its rate is of order $\frac{1}{n}$.

Finally, the following Theorem is a keystone of the proof. Let us denote after Massart (2007):

$$\forall A \text{ measurable with } \mathbb{P}\big[A\big] > 0, \forall \varphi : \mathbb{R}^d \to \mathbb{R} \text{ measurable, } \mathbb{E}^A\big[\varphi(X)\big] = \frac{\mathbb{E}\big[\varphi(X)1_A(X)\big]}{\mathbb{P}\big[A\big]}.$$

Theorem 8 (Theorem 6.8 in Massart, 2007) *Let $\mathcal{F}$ be a countable class of real valued and measurable functions. Assume*

$$\exists \sigma > 0, \exists b > 0 / \forall f \in \mathcal{F}, \forall k \geq 2, \mathbb{E}\left[|f(X_i)|^k\right] \leq \frac{k!}{2}\sigma^2 b^{k-2},$$

and

$$\forall \delta > 0, \exists C_\delta \text{ a set of brackets covering } \mathcal{F} /$$
$$\forall [g_l, g_u] \in C_\delta, \forall k \in \mathbb{N}^* \backslash \{1\}, \mathbb{E}[(g_u - g_l)^k(X_i)] \leq \frac{k!}{2}\delta^2 b^{k-2}.$$

Let $e^{H(\delta)}$ be the minimal cardinality of such a covering. Then:

$$\exists \kappa \text{ absolute constant } / \forall \varepsilon \in]0,1], \forall A \text{ measurable with } \mathbb{P}\big[A\big] > 0,$$
$$\mathbb{E}^A\left[\sup_{f\in\mathcal{F}} S_n(f)\right] \leq E + (1+6\varepsilon)\sigma\sqrt{2n\log\frac{1}{\mathbb{P}[A]}} + 2b\log\frac{1}{\mathbb{P}[A]},$$
$$\text{where } E = \frac{\kappa}{\varepsilon}\sqrt{n}\int_0^{\varepsilon\sigma}\sqrt{H(u)\wedge n}\, du + 2(b+\sigma)H(\sigma).$$

This theorem gives a control of the supremum of the empirical process over a class of functions with respect to upper bounds on the moments of the functions in $\mathcal{F}$ and of a set of brackets covering $\mathcal{F}$, and with respect to the minimal number of brackets of such a covering. This theorem links the behavior of the supremum of the empirical process and the complexity of the class of functions considered.

Lemmas 8, 9, and Theorem 8, together with the calculation of the entropy with bracketing of the class of functions obtained for $-\log \mathrm{L}_{\mathrm{cc}}$ over the considered parameter space (Lemma 4) are the ingredients of the proof of Lemma 6.

4.4.3 Proofs

Proof (Theorem 6) *Let* $\mathcal{K} = \mathrm{argmin}_{K\in\{1,\dots,D_K\}} \mathbb{E}_{f^\wp}\left[\gamma(\theta_K^0)\right]$. *By assumption,* $K_0 = \min \mathcal{K}$.

It is first proved that $\hat{K}$ does not asymptotically "underestimate" K_0. Let $K \notin \mathcal{K}$. Let $\varepsilon = \frac{1}{2}\big(\mathbb{E}_{f^\wp}[\gamma(\theta_K^0)] - \mathbb{E}_{f^\wp}\big[\gamma(\theta_{K_0}^0)\big]\big) > 0$. From (B2) *and* (B3) *($\text{pen}(K) = o_{\mathbb{P}}(1)$), with great probability and for n large enough:*

$$\begin{aligned}\Big|\gamma_n(\hat{\theta}_K) - \mathbb{E}_{f^\wp}\big[\gamma(\theta_K^0)\big]\Big| &\leq \frac{\varepsilon}{3}\\ \Big|\gamma_n(\hat{\theta}_{K_0}) - \mathbb{E}_{f^\wp}\big[\gamma(\theta_{K_0}^0)\big]\Big| &\leq \frac{\varepsilon}{3}\\ \text{pen}(K_0) &\leq \frac{\varepsilon}{3}.\end{aligned}$$

Then

$$\begin{aligned}\text{crit}(K) &= \gamma_n(\hat{\theta}_K) + \text{pen}(K)\\ &\geq \mathbb{E}_{f^\wp}\big[\gamma(\theta_K^0)\big] - \frac{\varepsilon}{3} + 0\\ &= \mathbb{E}_{f^\wp}\big[\gamma(\theta_{K_0}^0)\big] + \frac{5\varepsilon}{3}\\ &\geq \underbrace{\gamma_n(\hat{\theta}_{K_0}) + \text{pen}(K_0)}_{\text{crit}(K_0)} + \varepsilon.\end{aligned}$$

Then, with large probability and for n large enough, $\hat{K} \neq K$.

Let now $K \in \mathcal{K}$, with $K > K_0$. This part of the result is more involved than the first one but at this stage, it is not more difficult to derive: all the job is hidden in the assumption (B4)*...Indeed, it implies that $\exists V > 0$, such that for n large enough and with large probability,*

$$n\big(\gamma_n(\hat{\theta}_{K_0}) - \gamma_n(\hat{\theta}_K)\big) \leq V.$$

Increase n enough so that $n\big(\text{pen}(K) - \text{pen}(K_0)\big) > V$ with great probability (which is possible from assumption (B4)*). Then, for n large enough and with large probability,*

$$\begin{aligned}\text{crit}(K) &= \gamma_n(\hat{\theta}_K) + \text{pen}(K)\\ &\geq \gamma_n(\hat{\theta}_{K_0}) - \frac{V}{n} + \text{pen}(K)\\ &> \text{crit}(K_0).\end{aligned}$$

And then, with large probability and for n large enough, $\hat{K} \neq K$.

Finally, since $\mathbb{P}[\hat{K} \neq K_0] = \sum_{K \notin \mathcal{K}} \mathbb{P}[\hat{K} = K] + \sum_{K \in \mathcal{K},\ K \neq K_0} \mathbb{P}[\hat{K} = K]$, the result follows. ■

Proof (Lemma 5) *For any $\varepsilon > 0$, with large probability and for n large enough:*

$$\gamma_n(\hat{\theta}) - \mathbb{E}_{f^\wp}\big[\gamma(\theta^0)\big] = \underbrace{\gamma_n(\hat{\theta}) - \gamma_n(\theta^0)}_{\leq \varepsilon} + \underbrace{\gamma_n(\theta^0) - \mathbb{E}_{f^\wp}\big[\gamma(\theta^0)\big]}_{\leq \varepsilon},$$

on the one hand. And

$$\gamma_n(\hat{\theta}) - \mathbb{E}_{f^\wp}\big[\gamma(\theta^0)\big] = \underbrace{\gamma_n(\hat{\theta}) - \mathbb{E}_{f^\wp}\Big[\gamma(\hat{\theta})\Big]}_{\geq -\varepsilon} + \underbrace{\mathbb{E}_{f^\wp}\Big[\gamma(\hat{\theta})\Big] - \mathbb{E}_{f^\wp}\big[\gamma(\theta^0)\big]}_{\geq 0},$$

on the other hand. ■

Proof (Lemma 6) *Actually, the proof as it is written below holds for an at most countable model (because this assumption is necessary for Lemma 9 and Theorem 8 to hold). But it can be checked that both those results may be applied to a dense subset of* $\{\gamma(\theta) : \theta \in \Theta\}$ *containing* θ^0 *and their respective conclusions generalized to the entire set: choose* Θ^{count} *a countable dense subset of* Θ*. Then, for any* $\theta \in \Theta$*, let* $\theta_n \in \Theta^{count} \xrightarrow[n\to\infty]{} \theta$*. Then,* $\gamma(\theta_n; X) \xrightarrow[n\to\infty]{a.s.} \gamma(\theta; X)$*. Yet, whatever* $g : \mathbb{R}^D \times \left(\mathbb{R}^d\right)^n \to \mathbb{R}$ *such that* $\theta \in \mathbb{R}^D \mapsto g(\theta, \mathbf{X})$ *continue a.s.,* $\sup_{\theta\in\Theta} g(\theta; \mathbf{X}) = \sup_{\theta\in\Theta^{count}} g(\theta; \mathbf{X})$ *a.s. Hence,* $\mathbb{E}_{f^\wp}\left[\sup_{\theta\in\Theta} g(\theta; \mathbf{X})\right] = \mathbb{E}_{f^\wp}\left[\sup_{\theta\in\Theta^{count}} g(\theta; \mathbf{X})\right]$*. Remark however that this is quite artificial to do so: the models which are actually considered are discrete, because of the computation limitations.*

Let us introduce the centered empirical process

$$\begin{aligned} S_n\gamma(\theta) &= n\gamma_n(\theta) - n\mathbb{E}_{f^\wp}\left[\gamma(\theta; X)\right] \\ &= \sum_{i=1}^{n}\big(\gamma(\theta; X_i) - \mathbb{E}_{f^\wp}\left[\gamma(\theta; X)\right]\big). \end{aligned}$$

Here and hereafter, α *stands for a generic absolute constant, which may differ from a line to an other; All* $\mathbb{E}$ *and* $\mathbb{P}$ *symbols are understood under* $f^\wp d\mu$*.*

Let $\theta^0 \in \Theta$ *such that* $\mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right] = \inf_{\theta\in\Theta} \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right]$*. Let us define*

$$\forall \sigma > 0, \Theta(\sigma) = \{\theta \in \Theta : \|\theta - \theta^0\|_\infty \leq \sigma\}.$$

On the one hand,

$$\begin{aligned} &\forall r \in \mathbb{N}^*\backslash\{1\}, \forall \theta \in \Theta(\sigma), \\ &\qquad \left|\gamma(\theta^0; x) - \gamma(\theta; x)\right|^r \leq L'(x)^2 \|\theta^0 - \theta\|_\infty^2 (2L(x))^{r-2} \quad f^\wp d\mu\text{-a.s.}, \end{aligned}$$

since $\Theta(\sigma) \subset \Theta$ *is convex (this is a consequence of its definition because* Θ *is). Thus,*

$$\begin{aligned} \forall r \in \mathbb{N}^*\backslash\{1\}, \forall \theta \in \Theta(\sigma), \mathbb{E}_{f^\wp}\left[|\gamma(\theta^0) - \gamma(\theta)|^r\right] &\leq \|L'\|_2^2 \, \|\theta^0 - \theta\|_\infty^2 (2\|L\|_\infty)^{r-2} \\ &\leq \frac{r!}{2}(\|L'\|_2\sigma)^2 \left(\frac{2\|L\|_\infty}{2}\right)^{r-2}. \end{aligned} \tag{4.4}$$

And on the other hand, from Lemma 4 which applies here, for any $r \in \mathbb{N}^*\backslash\{1\}$*, for any* $\delta > 0$*, there exists* C_δ *a set of brackets which covers* $\{(\gamma(\theta^0) - \gamma(\theta)) : \theta \in \Theta(\sigma)\}$ *(deduced from a set of brackets which covers* $\{\gamma(\theta) : \theta \in \Theta(\sigma)\}$*...) such that:*

$$\forall r \in \mathbb{N}^*\backslash\{1\}, \forall [g_l, g_u] \in C_\delta, \|g_u - g_l\|_r \leq \left(\frac{r!}{2}\right)^{\frac{1}{r}} \delta^{\frac{2}{r}} \left(\frac{4\|L\|_\infty}{3}\right)^{\frac{r-2}{r}}$$

and such that, writing $e^{H(\delta,\Theta(\sigma))}$ *the minimal cardinal of such a* C_δ*,*

$$e^{H(\delta,\Theta(\sigma))} \leq \left(\frac{\overbrace{diam\,\Theta(\sigma)}^{\leq 2\sigma}\,\|L'\|_2}{\delta}\right)^D \vee 1. \tag{4.5}$$

Then, according to Theorem 8,
$\exists\alpha, \forall\varepsilon \in]0,1], \forall A$ *measurable such that* $\mathbb{P}[A] > 0$,

$$\begin{aligned}\mathbb{E}^A\left[\sup_{\theta\in\Theta(\sigma)} S_n\big(\gamma(\theta^0)-\gamma(\theta)\big)\right] \le \frac{\alpha}{\varepsilon}\sqrt{n}\int_0^{\varepsilon\|L'\|_2\sigma}\sqrt{H\big(u,\Theta(\sigma)\big)}\,du \\ + 2\Big(\frac{4}{3}\|L\|_\infty + \|L'\|_2\sigma\Big)H\big(\|L'\|_2\sigma,\Theta(\sigma)\big) \\ + (1+6\varepsilon)\|L'\|_2\sigma\sqrt{2n\log\frac{1}{\mathbb{P}[A]}} \\ + \frac{8}{3}\|L\|_\infty\log\frac{1}{\mathbb{P}[A]}.\end{aligned} \tag{4.6}$$

Now, we have

$$\begin{aligned}\forall t\in\mathbb{R}^+, \int_0^t\sqrt{\log\frac{1}{u}\vee 0}\,du &= \int_0^{t\wedge 1}\sqrt{\log\frac{1}{u}}\,du \\ &\le \sqrt{t\wedge 1}\sqrt{\int_0^{t\wedge 1}\log\frac{1}{u}\,du} \quad \textit{from Cauchy-Schwarz inequality} \\ &= (t\wedge 1)\sqrt{\log\frac{e}{t\wedge 1}}.\end{aligned}$$

From which, together with (4.5),

$$\begin{aligned}\forall t\in\mathbb{R}^+, \int_0^t\sqrt{H(u,\Theta(\sigma))}du &\le \sqrt{D}\int_0^t\sqrt{\log\frac{2\|L'\|_2\sigma}{u}\vee 0}\,du \\ &\le \sqrt{D}\big(t\wedge 2\|L'\|_2\sigma\big)\sqrt{\log\frac{e}{\frac{t}{2\|L'\|_2\sigma}\wedge 1}},\end{aligned} \tag{4.7}$$

after a simple change of variable.

Next, let us apply Lemma 9: From (4.5), (4.6) *and* (4.7),

$$\forall\sigma>0, \mathbb{E}_{f^\wp}\left[\sup_{\theta\in\Theta(\sigma)} S_n\big(\gamma(\theta^0)-\gamma(\theta)\big)\right] \le \varphi(\sigma),$$

with

$$\begin{aligned}\varphi(t) = \frac{\alpha}{\not{\varepsilon}}\sqrt{n}\sqrt{D}\,\not{\varepsilon}\|L'\|_2 t\sqrt{\log\frac{2e}{\varepsilon}} + 2\Big(\frac{4}{3}\|L\|_\infty + \|L'\|_2 t\Big)D\log 2 \\ + (1+6\varepsilon)\|L'\|_2 t\sqrt{2n\log\frac{1}{\mathbb{P}[A]}} + \frac{8}{3}\|L\|_\infty\log\frac{1}{\mathbb{P}[A]}.\end{aligned}$$

As required for Lemma 9 to hold, $\frac{\varphi(t)}{t}$ *is nonincreasing. It follows*

$$\forall\beta>0, \mathbb{E}^A\left[\sup_{\theta\in\Theta}\frac{S_n\big(\gamma(\theta^0)-\gamma(\theta)\big)}{\|\theta^0-\theta\|_\infty+\beta^2}\right] \le 4\beta^{-2}\varphi(\beta).$$

We then choose $\varepsilon = 1$ *and apply Lemma 8: for any* $\eta > 0$ *and any* $\beta > 0$*, with probability larger than* $1 - e^{-\eta}$*,*

$$\sup_{\theta\in\Theta} \frac{S_n\big(\gamma(\theta^0) - \gamma(\theta)\big)}{\|\theta^0 - \theta\|_\infty^2 + \beta^2} \leq \frac{\alpha}{\beta^2}\bigg(\sqrt{nD}\|L'\|_2\beta\sqrt{\log 2e} + \big(\|L\|_\infty + \|L'\|_2\beta\big)D\log 2 + \|L'\|_2\beta\sqrt{n\eta} + \|L\|_\infty\eta\bigg).$$

■

Proof (Lemma 7) (4.4) *may be rewritten as*

$$\forall r \in \mathbb{N}^*\backslash\{1\}, \forall\theta\in\Theta(\sigma), \quad \mathbb{E}_{f^\wp}\left[|\gamma(\theta^0) - \gamma(\theta)|^r\right] \leq \frac{r!}{2}\Big(\underbrace{\|L'\|_2\sigma \wedge 4\|L\|_\infty}_{\sigma'}\Big)^2 \left(\|L\|_\infty\right)^{r-2}.$$

This follows from $|\gamma(\theta^0) - \gamma(\theta)| \leq 2\|L\|_\infty$*, which always holds, and is required to have* σ' *bounded from above, which will be useful in the following. It does not damage the result since the upper bound with respect to* σ *is interesting locally, i.e. when* σ *is small.*

Then, from (4.7)*,*

$$\int_0^{\|L'\|_2\sigma} \sqrt{H\big(u, \Theta(\sigma)\big)}du \leq \underbrace{\sqrt{\log 2e}\sqrt{D}\|L'\|_2\sigma}_{\psi(\sigma)}.$$

Let $\sigma_0 = \sqrt{\frac{D}{n}}$*. Then,*

$$\begin{aligned} H\big(\|L'\|_2\sigma, \Theta(\sigma)\big) &\leq D\log 2 \quad \textit{from (4.5)} \\ &= \frac{D\log 2}{\sigma}\sigma \quad \forall\sigma > 0 \\ &\leq \log 2\sqrt{Dn}\sigma \quad \forall\sigma > \sigma_0. \end{aligned}$$

Moreover, $H\big(\sigma', \Theta(\sigma)\big) \leq H\big(\|L'\|_2\sigma, \Theta(\sigma)\big)$*. Indeed,*

- *If* $\|L'\|_2\sigma \leq 4\|L\|_\infty$*, this holds.*
- *If* $\|L'\|_2\sigma \geq 4\|L\|_\infty$*, remark that, since* $\forall\theta\in\Theta(\sigma), |\gamma(\theta^0) - \gamma(\theta)| \leq 2\|L\|_\infty$ *for* $f^\wp d\lambda -$*almost all* x*,*

$$\forall\theta\in\Theta(\sigma), \gamma(\theta^0) - 2\|L\|_\infty \leq \gamma(\theta) \leq \gamma(\theta^0) + 2\|L\|_\infty \quad f^\wp d\lambda - a.s.,$$

and then, $H\big(4\|L\|_\infty, \Theta(\sigma)\big) = 0$ *and* $H\big(\sigma', \Theta(\sigma)\big) = 0$ *in this case.*

Now, it follows, from Theorem 8,

$$\exists \alpha, \forall \varepsilon \in]0,1], \forall A \text{ measurable such that } \mathbb{P}[A] > 0,$$

$$\mathbb{E}^A\left[\sup_{\theta\in\Theta(\sigma)} S_n\big(\gamma(\theta^0)-\gamma(\theta)\big)\right] \leq \frac{\alpha}{\varepsilon}\sqrt{n}\int_0^{\varepsilon\sigma'}\sqrt{H\big(u,\Theta(\sigma)\big)}\,du + 2\Big(\frac{4}{3}\|L\|_\infty+\sigma'\Big)H\big(\sigma',\Theta(\sigma)\big) + (1+6\varepsilon)\sigma'\sqrt{2n\log\frac{1}{\mathbb{P}[A]}} + \frac{8}{3}\|L\|_\infty\log\frac{1}{\mathbb{P}[A]}.$$

And then, for any $\sigma > \sigma_0$, since $\sigma' \leq \|L'\|_2\sigma$, for $\varepsilon = 1$,

$$\mathbb{E}^A\left[\sup_{\theta\in\Theta(\sigma)} S_n\big(\gamma(\theta^0)-\gamma(\theta)\big)\right] \leq \alpha\sqrt{n}\sqrt{\log 2e}\|L'\|_2\sqrt{D}\sigma + 2\left(\frac{4}{3}\|L\|_\infty+4\|L\|_\infty\right)\sqrt{n}\log 2\sqrt{D}\sigma + (1+6\varepsilon)\|L'\|_2\sigma\sqrt{2n\log\frac{1}{\mathbb{P}[A]}} + \frac{8}{3}\|L\|_\infty\log\frac{1}{\mathbb{P}[A]}.$$

This function of σ divided by σ is nonincreasing. Then, the conclusion follows from Lemma 9 and Lemma 8 exactly as for Lemma 6, but under the condition that $\beta > \sigma_0$:

$$\exists \alpha > 0, \forall \beta > \sigma_0, \forall \eta > 0,$$

$$\sup_{\theta\in\Theta}\frac{S_n\big(\gamma(\theta^0)-\gamma(\theta)\big)}{\|\theta^0-\theta\|_\infty^2+\beta^2} \leq \frac{\alpha}{\beta^2}\left[\left(\Big(\|L'\|_2+\|L\|_\infty\Big)\sqrt{D}+\|L'\|_2\sqrt{\eta}\right)\sqrt{n}\beta + \|L\|_\infty\eta\right]$$

holds with probability larger than $1-e^{-\eta}$.

■

Proof (Corollary 2) *Let $\varepsilon > 0$ such that $B(\theta^0,\varepsilon) \subset \Theta^{\mathcal{O}}$. Then, since $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta^0$, there exists $n_0 \in \mathbb{N}^*$ such that, with large probability, for $n \geq n_0$, $\hat{\theta}_n \in B(\theta^0,\varepsilon)$.*

Now, $B(\theta^0,\varepsilon)$ is convex and the assumptions of the corollary guarantee that Lemma

6 applies. Let us apply it to $\hat{\theta}_n$*:* $\forall n \geq n_0, \forall \beta > 0$*, with great probability as* η *is great,*

$$\frac{S_n\big(\gamma(\theta^0) - \gamma(\hat{\theta}_n)\big)}{\|\theta^0 - \hat{\theta}_n\|_\infty^2 + \beta^2} \leq \frac{\alpha}{\beta^2}\Bigg(\sqrt{nD}\|L'\|_2\beta + \big(\|L\|_\infty + \|L'\|_2\beta\big)D + \|L'\|_2\beta\sqrt{n\eta} + \|L\|_\infty\eta\Bigg). \tag{4.8}$$

But since I_{θ^0} *is supposed to be nonsingular, it can be written that*

$$\forall \theta \in B(\theta^0, \varepsilon), \mathbb{E}_{f^\wp}[\theta] - \mathbb{E}_{f^\wp}\left[\theta^0\right] = (\theta - \theta^0)' I_{\theta^0}(\theta - \theta^0) + r(\|\theta - \theta^0\|_\infty^2)\|\theta - \theta^0\|_\infty^2 \geq \big(2\alpha' + r(\|\theta - \theta^0\|_\infty^2)\big)\|\theta - \theta^0\|_\infty^2,$$

where $\alpha' > 0$ *depends on* I_{θ^0} *and* $r : \mathbb{R}^+ \longrightarrow \mathbb{R}$ *fulfills* $r(x) \xrightarrow[x\to 0]{} 0$ *(we do not introduce the o notation at this stage to avoid ambiguities with* $o_{\mathbb{P}}$ *to appear below...). Then, for* $\|\theta - \theta^0\|_\infty$ *small enough (*ε *may be decreased...),*

$$\forall \theta \in B(\theta^0, \varepsilon), \mathbb{E}_{f^\wp}[\theta] - \mathbb{E}_{f^\wp}\left[\theta^0\right] \geq \alpha'\|\theta - \theta^0\|_\infty^2. \tag{4.9}$$

Since

$$S_n\big(\gamma(\theta^0) - \gamma(\hat{\theta}_n)\big) = n\big(\gamma_n(\theta^0) - \gamma_n(\hat{\theta}_n)\big) + n\mathbb{E}_{f^\wp}\left[\gamma(\hat{\theta}_n) - \gamma(\theta^0)\right] \geq O_{\mathbb{P}}(1) + n\mathbb{E}_{f^\wp}\left[\gamma(\hat{\theta}_n) - \gamma(\theta^0)\right],$$

(4.8) *together with* (4.9) *leads (with great probability) to*

$$n\|\hat{\theta}_n - \theta^0\|_\infty^2 \leq \frac{\|L'\|_2(\sqrt{nD} + \sqrt{\eta n} + D)\beta + \|L\|_\infty(D + \eta) + O_{\mathbb{P}}(1)}{\frac{\alpha'}{\alpha} - \frac{1}{n\beta^2}\left(\|L'\|_2(\sqrt{nD} + \sqrt{\eta n} + D)\beta + \|L\|_\infty(D + \eta)\right)},$$

as soon as the denominator of the right-hand side is positive.

It then suffices to choose β *such that this condition is fulfilled and such that the right-hand side is upper-bounded by a quantity which does not depend on* n *to get the result. Let us try* $\beta = \frac{\beta_0}{\sqrt{n}}$ *with* β_0 *independent of* n*:*

$$n\|\hat{\theta}_n - \theta^0\|_\infty^2 \leq \frac{\|L'\|_2(\sqrt{D} + \sqrt{\eta} + \frac{D}{\sqrt{n}})\beta_0 + \|L\|_\infty(D + \eta) + O_{\mathbb{P}}(1)}{\frac{\alpha'}{\alpha} - \frac{1}{\beta_0^2}\left(\|L'\|_2(\sqrt{D} + \sqrt{\eta} + \frac{D}{\sqrt{n}})\beta_0 + \|L\|_\infty(D + \eta)\right)} \leq \frac{\|L'\|_2(\sqrt{D} + \sqrt{\eta} + D)\beta_0 + \|L\|_\infty(D + \eta) + O_{\mathbb{P}}(1)}{\frac{\alpha'}{\alpha} - \frac{1}{\beta_0^2}\left(\|L'\|_2(\sqrt{D} + \sqrt{\eta} + D)\beta_0 + \|L\|_\infty(D + \eta)\right)}$$

This only holds if the denominator is positive. Choose β_0 *large enough so as to guarantee this, which is always possible. The result follows: with large probability and for* n *larger than* n_0*,*

$$n\|\hat{\theta}_n - \theta^0\|_\infty^2 = CO_{\mathbb{P}}(1),$$

with C *depending on* D*,* $\|L\|_\infty$*,* $\|L'\|_2$*,* I_{θ^0} *and* η*.* ■

Proof (Corollary 3) *This is a direct application of Corollary 2. Let* $K \in \mathcal{K}$*:* $\mathbb{E}_{f^\wp}\left[\gamma(\theta_K^0)\right] = \mathbb{E}_{f^\wp}\left[\gamma(\theta_{K_0}^0)\right]$*.* Θ_K *can be assumed to be convex: if it is not,* $\hat{\theta}_K$ *lies in* $B(\theta_{K_0}^0, \varepsilon)$ *with large probability for large* n*. Choose* ε *small enough to guarantee* $B(\theta_{K_0}^0, \varepsilon) \subset \Theta^{\mathcal{O}}$*. Then,* Θ_K *may be replaced by* $B(\theta_{K_0}^0, \varepsilon)$*. According to Lemma 6 and under the assumptions we made, with probability larger than (1-*$e^{-\eta}$*) for* n *large enough,* $\forall \beta > 0$*,*

$$S_n\big(\gamma(\theta_K^0) - \gamma(\hat{\theta}_K)\big) \leq \alpha \frac{\|\theta_K^0 - \hat{\theta}_K\|_\infty^2 + \beta^2}{\beta^2} \left(\|L'\|_2 \Big(\sqrt{nD_K} + \sqrt{\eta n} + D_K \Big) \beta + \|L\|_\infty (D_K + \eta) \right),$$

which yields with $\beta = \frac{\beta_0}{\sqrt{n}}$ *for any* $\beta_0 > 0$*:*

$$S_n\big(\gamma(\theta_K^0) - \gamma(\hat{\theta}_K)\big) \leq \alpha \frac{n\|\theta_K^0 - \hat{\theta}_K\|_\infty^2 + \beta_0^2}{\beta_0^2} \left(\|L'\|_2 \Big(\sqrt{D_K} + \sqrt{\eta} + \frac{\overbrace{D_K}^{\leq D_K}}{\sqrt{n}} \Big) \beta_0 + \|L\|_\infty (D_K + \eta) \right).$$

But, according to Corollary 2, and because we imposed the corresponding assumptions, $n\|\theta_K^0 - \hat{\theta}_K\|_\infty^2 = O_{\mathbb{P}}(1)$*. Moreover, by definition,*

$$S_n\big(\gamma(\theta_K^0) - \gamma(\hat{\theta}_K)\big) = n\big(\gamma_n(\theta_K^0) - \gamma_n(\hat{\theta}_K)\big) + n\underbrace{\Big(\mathbb{E}_{f^\wp}\left[\gamma(\hat{\theta}_K)\right] - \mathbb{E}_{f^\wp}\left[\gamma(\theta_K^0)\right]\Big)}_{\geq 0}$$
$$\geq n\big(\gamma_n(\theta_K^0) - \gamma_n(\hat{\theta}_K)\big).$$

Thus,

$$n\big(\gamma_n(\theta_K^0) - \gamma_n(\hat{\theta}_K)\big) = O_{\mathbb{P}}(1).$$

This holds for any $K \in \mathcal{K}$ *and then in particular for* K_0 *and* K*. Besides,* $\gamma_n(\theta_K^0) = \gamma_n(\theta_{K_0}^0)$ *since, by assumption,* $\gamma(\theta_K^0; x) = \gamma(\theta_{K_0}^0; x)$ *for* $f^\wp d\lambda$*-almost all* x*. Hence*

$$n\big(\gamma_n(\hat{\theta}_{K_0}) - \gamma_n(\hat{\theta}_K)\big) = O_{\mathbb{P}}(1).$$

■

4.4.4 A New Light on ICL

The previous section suggests links between model selection penalized criteria with the usual likelihood on the one hand and with the conditional classification likelihood we defined on the other hand. Indeed penalties with the same form as those given by Nishii (1988) or Keribin (2000) with the usual likelihood are proved to be "consistent" in our framework. Moreover, the form of the results obtained (notably, Lemma 7) and some further calculations suggest that the reasoning of Massart (2007) when deriving an almost oracle inequality in the likelihood framework might be further mimicked and that an oracle inequality might be derived with quite comparable forms of penalties: they are notably expected to be proportional to the dimension of the models. But our

current results would at best enable to derive an oracle-like inequality with a multiplying factor depending on the parameters of the problem (typically, $\|L\|_\infty$ and $\|L'\|_2$), instead of an absolute constant.

Therefore, from those results and by analogy with the usual likelihood framework, it is expected that penalties proportional to D_K are optimal from the efficiency point of view (think of AIC), and that penalties proportional to $D_K \log n$ are optimal for an identification purpose (think of BIC). This possibility to derive an identification procedure from an efficient procedure by a $\log n$ factor is notified for example by Arlot (2007).

Let us then consider by analogy with BIC the penalized criterion

$$\begin{aligned}\mathrm{crit}(K) &= \log \mathrm{L_{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}) - \frac{\log n}{2} D_K \\ &= \log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLccE}}) - \mathrm{ENT}(\widehat{\theta}_K^{\mathrm{MLccE}}; \mathbf{X}) - \frac{\log n}{2} D_K.\end{aligned}$$

The point is that we almost recover ICL (see (4.2)). This is actually ICL evaluated at $\widehat{\theta}_K^{\mathrm{MLccE}}$ instead of $\widehat{\theta}_K^{\mathrm{MLE}}$ and we shall call it $\mathrm{L_{cc}}$-ICL. We already criticized the approximation of the mode of the integrated classification likelihood by $\widehat{\theta}_K^{\mathrm{MLE}}$ in the derivation of ICL (Section 4.1.2). There is no reason that this approximation should be good but perhaps in cases the components of $\widehat{\theta}_K^{\mathrm{MLE}}$ are well separated. Such a derivation of ICL is then hardly interpretable: only the resulting criterion itself can be interpreted, by the study of the entropy term, and by a comparison with known criteria, such as the BIC. But ICL may be regarded as an approximation of $\mathrm{L_{cc}}$-ICL. The corresponding penalty is $\frac{\log n}{2} D_K$, and the derivation of $\mathrm{L_{cc}}$-ICL in Section 4.4.1 illustrates that the entropy term should not be considered as a part of the penalty. This notably justifies why ICL does not select the same number of components as BIC or any consistent criterion in the usual likelihood framework, even asymptotically. Actually, it is not expected to do so.

As mentioned in Section 4.3.4, $\widehat{\theta}_K^{\mathrm{MLccE}}$ may be quite different from $\widehat{\theta}_K^{\mathrm{MLE}}$. This particularly holds when the components of $\widehat{\theta}_K^{\mathrm{MLE}}$ overlap. In such a case, $\widehat{\theta}_K^{\mathrm{MLccE}}$ provides more separated clusters, because of the entropy term. The compromise between the Gaussian component and the cluster viewpoints achieved by ICL in the model selection step is already embraced with $\widehat{\theta}_K^{\mathrm{MLccE}}$ in the estimation step. This means the user is provided a solution which aims at reaching this compromise for each number of classes K. Obviously, the number of classes selected through $\mathrm{L_{cc}}$-ICL may then differ from the one selected by ICL. It may for example be that for a given K which is not selected by ICL because of the entropy term, $\widehat{\theta}_K^{\mathrm{MLccE}}$ reaches smaller enough entropy than $\widehat{\theta}_K^{\mathrm{MLE}}$ that $\mathrm{L_{cc}}$-ICL selects this solution. However the likelihood is worsened in such a case, too, and this situation occurred seldom in simulations (see Section 4.4.6).

Finally, $\mathrm{L_{cc}}$-ICL is quite close to ICL and enables to better understand the concepts underlying ICL. Section 4.4.6 illustrates that ICL remains attractive, notably because of its ease to be computed as compared to the practical difficulties involved when evaluating $\widehat{\theta}_K^{\mathrm{MLccE}}$.

4.4.5 Slope Heuristics

The slope heuristics first introduced by Birgé and Massart (2006), has been presented and discussed in Section 3.2. It has been applied to the choice of the number of components in the Gaussian mixture models framework and the usual likelihood in Section 3.3.

We shall consider this heuristics to calibrate penalties of the form suggested in Section 4.4.1, with the $-\log \mathrm{L_{cc}}$ contrast. As mentioned in the Section 4.4.4 above, we have no definitive theoretical justification that the penalty should be chosen proportional to the dimension of the model, but serious hints in that direction. Simulations (Section 4.4.6) confirm this choice. Criteria of the following from are then considered and expected to have an oracle-like behavior:

$$\mathrm{crit}(K) = -\log \mathrm{L_{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}) + \kappa D_K,$$

where κ is unknown. The slope heuristics, which would by the way require even more than an oracle inequality to be theoretically justified in this framework (see Section 3.2), provides a practical approach to choose κ. Remark that it comes with a convenient method to verify that it can be reasonably assumed to be justified. It is recalled below. Further explanation, justification, and discussion (in a general framework) are given in Section 3.2. The data-driven slope estimation approach (Section 3.2.3) is applied in this section.

The slope heuristics relies on the assumption that the bias of the models decreases as the complexity of the models increases and is stationary for the most complex models. In our framework, this requires the family of models to be roughly embedded. With general or diagonal models for example, this requires the lower-bound on the covariance matrices and/or on the proportions, to be small enough. The models are always embedded if the proportions can be equal to zero, but this situation should be avoided. This heuristics should be handled with care in a framework with models such as the equal-volumes covariance matrices: models with different number of components are then definitely not embedded if the proportions are kept away from zero, and there is no reason why the minimum value of the contrast should even decrease as the model's complexity increases.

The procedure is the following. It is illustrated in the simulations section (Section 4.4.6). First compute $\widehat{\theta}_K^{\mathrm{MLccE}}$ for each $K \in \{1, \dots, K_M\}$ (cf. for example Figure 4.12). This is not an easy step actually: it is discussed in Section 5.1. Next, plot the value of $-\log \mathrm{L_{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}})$ with respect to D_K (cf. Figure 4.7). There should appear a linear part in this graph, for the greatest dimensional models. In case not, either K_M has been chosen too small and more complex models should be involved in the study to be able to apply the slope heuristics, or the slope heuristics does not apply. Actually, when the optimization of the contrast for each model — i.e. the computation of $\widehat{\theta}_K^{\mathrm{MLccE}}$ — was not too hard, we almost always observed such a linear part. Then, compute the slope $\frac{\hat{\kappa}}{2}$ of this linear part and choose $\hat{\kappa}$ as a constant in the criterion presented just above ($\kappa = \hat{\kappa}$). Finally, select $\hat{K}$ according to

$$\hat{K} = \min \operatorname*{argmax}_{K \in \{1,\dots,K_M\}} \left\{ \log \mathrm{L_{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}) - \hat{\kappa} D_K \right\}.$$

The main practical challenge is the choice of the points of the graph of $D_K \mapsto$

$-\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}})$ which should be considered to belong to the linear part, and thus involved in the computation of the slope. The procedure result might depend quite largely on this choice in some situations. We applied in a first time a quite simple rule:

- check that a linear part occurs "by eye";
- for each $K_m \in \{1, \ldots, K_M - 1\}$, compute the slope $\frac{\hat{\kappa}_{K_m}}{2}$ of the linear regression of the points $\big(D_K, -\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}})\big)_{K \in \{K_m, \ldots, K_M\}}$ and the corresponding variation of (twice the) slope
$$\delta_{K_m - 1} = \hat{\kappa}_{K_m} - \hat{\kappa}_{K_m - 1}$$
(for $K_m > 1$);
- choose K_m^0 as the smallest K_m such that $\delta_{K_m} \leq q \min_{K_m \in \{1, \ldots, K_M - 2\}} \delta_{K_m}$:
$$K_m^0 = \min\Big\{K_m \in \{1, \ldots, K_M - 2\} : \delta_{K_m} \leq q \min_{K_m \in \{1, \ldots, K_M - 2\}} \delta_{K_m}\Big\},$$
for a given $q > 1$;
- Finally, $\hat{\kappa} = \hat{\kappa}_{K_m^0}$.

The choice of q is obviously problematic. But there is, up to our knowledge, no method in this context which would not depend on a tuning parameter to be chosen. Some are more problematic than others, and in the simulations we performed, this one seemed to offer the advantage that the final solution reasonably depends on the choice of q. The underlying idea it that the slope should be quite "stable" once the linear part has been reached. What "stable" in a general setting means, is quite difficult to quantify. This method enables to quantify it while taking account of the data at hand, with respect to the minimal reached "stability", which seems much more reliable than trying to assess a general value of what "stable" would be, independently of the particular situation. Other measures could have been involved, such as the variance or the R^2 of the regression...Remark that this method with q encounters troubles when some of the successive slopes are almost the same...

A more involved and presumably more reliable method is proposed in Section 5.2. It has been implemented in a Matlab package, which is available for the practice of the slope heuristics. It yields roughly the same results as those reported below, but warns the user as the heuristics should not be applied — or be applied with care —, where our first procedure rather returned a bad solution in such a case...

4.4.6 Simulations

In this section, several simulated experiments[8] are reported, which illustrate and complete the preceding considerations. For each simulation setting, at least 100 datasets have been simulated. As a matter of fact, the estimation softwares sometimes encounter difficulties and stop before yielding a result. Those cases have been removed from the

[8] Details on the simulation settings and the applied algorithms may be found in Section A.2.

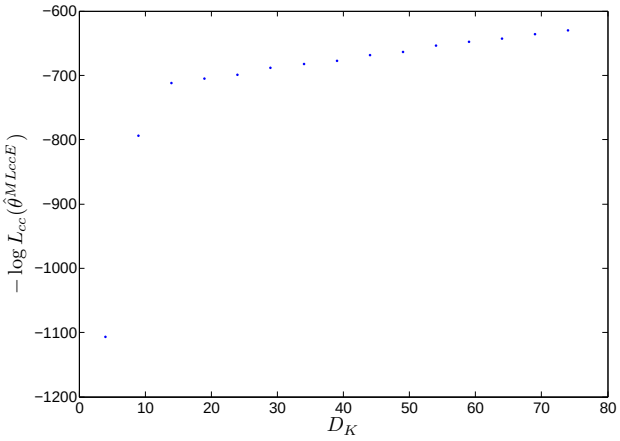


Figure 4.7: The "Slope" graph

study, so that the considered criteria can be compared on the basis of interesting examples. $\widehat{\theta}_K^{\mathrm{MLE}}$ and $\widehat{\theta}_K^{\mathrm{MLccE}}$ have been computed for each $K \in \{1, \ldots, K_M\}$, and the percentage of selection of each possible number of class is reported for each one of the following criteria:

- AIC: $\mathrm{crit}_{\mathrm{AIC}}(K) = \log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}}) - D_K$;
- BIC: $\mathrm{crit}_{\mathrm{BIC}}(K) = \log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}}) - \frac{\log n}{2} D_K$;
- Slope Heuristics applied to $\left(D_K, \log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}})\right)_{K \in \{1,\ldots,K_M\}}$: $\mathrm{crit}_{\mathrm{SHL}}(K) = \log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}}) - 2 \times \widehat{\mathrm{slope}} \times D_K$ (see Section 3.3);
- ICL: $\mathrm{crit}_{\mathrm{ICL}}(K) = \log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLE}}) - \frac{\log n}{2} D_K$;
- L_{cc}-ICL: $\mathrm{crit}_{\mathrm{L_{cc}-ICL}}(K) = \log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}) - \frac{\log n}{2} D_K$;
- Slope Heuristics applied to $\left(D_K, \log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}})\right)_{K \in \{1,\ldots,K_M\}}$: $\mathrm{crit}_{\mathrm{SHL_{cc}}}(K) = \log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}) - 2 \times \widehat{\mathrm{slope}} \times D_K$ (see Section 4.4.5).

Besides, for each criterion and each dataset type, typical examples of the obtained classifications are given. A few graphs $D_K \mapsto -\log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}})$ and $D_K \mapsto -\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}})$ are given for each setting, to testify a linear part occurs.

The "Cross" Experiment

In this experiment, $f^{\wp}$ is a four-component Gaussian mixture in $\mathbb{R}^2$: see Figure 4.8. The top-left and the bottom-left components are spherical: their covariance matrices are proportional to the identity matrix. The two components which together form the

"cross" on the right are "diagonal": their covariance matrices are diagonal matrices. Diagonal Gaussian mixture models are fitted: the true distribution is then available in the model with four components.

This experiment is particularly illustrative for our purpose. The point is that the "cross" is clearly made of two different components, but that it may be considered that it has to be split into two different classes or not, depending on the embraced classification point of view. The point of view adopted when using BIC is that every group of observations which requires a Gaussian component to be fitted deserves to be considered as a class for itself. There are in this dataset clearly four classes from this point of view. And the results of this experiment (Table 4.1) confirm that BIC performs well in this task.

But presumably a widespread notion of "cluster" would make most people unthinkingly design three classes in this dataset. This notion is at least partly based on the idea that observations which cannot be confidently discriminated from each other, should actually belong to the same class. ICL does well fit this notion, thanks to the entropy term: the results are striking with this experiment, too. Of course, ICL also takes somehow into account the Gaussian shape of the designed classes, since it relies on Gaussian mixture models. This corresponds to the notion that clusters should be more or less ellipsoid-shaped.

Remark that a strictly cluster-based approach, as the k-means or the hierarchical Ward's approach, for example, would never split the cross into its two Gaussian components, even when applied to design four classes.

$\mathrm{L_{cc}}$-ICL behaves exactly as ICL in this experiment. Moreover, for the number of components both of them select ($\hat{K}_{\mathrm{ICL}} = \hat{K}_{\mathrm{L_{cc}-ICL}} = 3$), $\widehat{\theta}_K^{\mathrm{MLE}}$ and $\widehat{\theta}_K^{\mathrm{MLccE}}$ give rise to very close estimations (actually, because of the precision of the machine, they cannot be distinguished from each other), and to exactly the same classes. See Figure 4.9 (K=3) and Figure 4.12 (K=3).

Now, both slope heuristics methods are interesting. The first one (SHL) behaves like BIC. They are based on the same contrast, and this is no surprise. Remark that AIC seems clearly not to penalize enough in this experiment, although it might be expected, as an "efficient" criterion (in other areas), to behave rather like SHL than like BIC. This holds also from an efficiency point of view (see Table 4.2). The slope heuristics based on $-\log \mathrm{L_{cc}}$ (denoted by $\mathrm{SHL_{cc}}$) behaves roughly like ICL. It is however more scattered. This might be partly due to optimization difficulties, which currently occur while maximizing $\mathrm{L_{cc}}$. Remark however that they reach almost the same results, from an efficiency point of view, as compared to the oracle (see Table 4.3): it may be surprising that the slope heuristics does not behaves better than ICL, but both reach very good results.

A few graphs (chosen at random) $D_K \mapsto \log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}})$ and $D_K \mapsto \log \mathrm{L_{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}})$ are represented on Figure 4.10 and Figure 4.13. Those graphics are important: the slope heuristics should only be applied as a linear part appears for the highest dimensional models. The Linear Regression is plotted, too. Mostly, the number of components K from which no bias is to be decreased anymore can almost be identified "by eye" in this experiment. The penalty values obtained through slope heuristics for both contrasts are represented in Figure 4.11 and Figure 4.14. They illustrate that those procedures are

data-driven: the value of the penalty they yield vary markedly in this experiment.

All solutions $\widehat{\theta}_K^{\mathrm{MLE}}$ and $\widehat{\theta}_K^{\mathrm{MLccE}}$ for every K have been represented, for an example dataset (Figure 4.9 and Figure 4.12). One of the most illustrative datasets has been chosen. Remark that, even for the true number of Gaussian components ($K = 4$ here), $\widehat{\theta}_K^{\mathrm{MLccE}}$ does not match the true distribution (which belongs to $\mathcal{M}_4$, yet), on the contrary to $\widehat{\theta}_K^{\mathrm{MLE}}$. It clearly avoids solutions with overlapping components. Of course, according to the particular repartition of the data, it may happen that the non-gaussianity of one cluster is worse than the overlapping of the two components and that $\widehat{\theta}_K^{\mathrm{MLccE}}$ chooses a solution close to $\widehat{\theta}_K^{\mathrm{MLE}}$, even in this experiment. It is however seldom. Anyone classifying by hand would probably do the same. The point, of course, is where the limit lies.

Tables 4.2 and 4.3 report comparisons of the risk of each criterion with the corresponding oracle. Remark that ICL should rather be compared to the $\mathrm{L_{cc}}$ oracle. The most interesting information in those results is that the oracle (expected) number of components is four for L and three for $\mathrm{L_{cc}}$, with quite large difference to the other numbers of components in each case (see Figure 4.15 (a) and (b)).

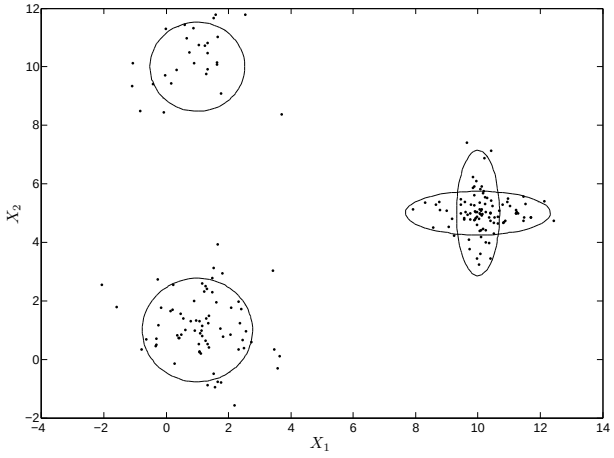


Figure 4.8: "Cross" Experiment.

Simulated observations example with isodensity contours of the true distribution. $n = 200$

An Experiment with Misspecified Models

In this experiment, $f^{\wp}$ is a four-component Gaussian mixture in $\mathbb{R}^2$: See Figure 4.16. All components are very well separated. The two left components are diagonal, but the two right components are not. The bottom-right component is rotated through angle $-\frac{\pi}{6}$ from horizontal, and the top-right component is rotated through angle $\frac{\pi}{3}$ from horizontal. Since fitted models are still diagonal mixture models, this experiment takes place in a misspecified models situation.

It enables to easily study the behavior of the criteria we are interested in in this misspecified situation.

Selected Number of Components	2	3	4	5	6	7	8	9	10–20
AIC	0	0	1	1	2	2	3	3	88
BIC	0	4	91	5	0	0	0	0	0
SHL (ddes)	0	2	84	10	3	0	0	0	1
SHL (dj)	0	3	85	10	2	0	0	0	0
ICL	0	96	3	1	0	0	0	0	0
L_{cc}-ICL	0	99	1	0	0	0	0		
SHL_{cc}	2	79	8	8	3	0	0		

Table 4.1: "Cross" Experiment Results.

"ddes" indicates the data-driven slope estimation approach for the slope heuristics application and "dj" the dimension jump approach. Some boxes are left blank to recall the criteria related to L_{cc} have been computed with $K \in \{1, \ldots, 8\}$ only.

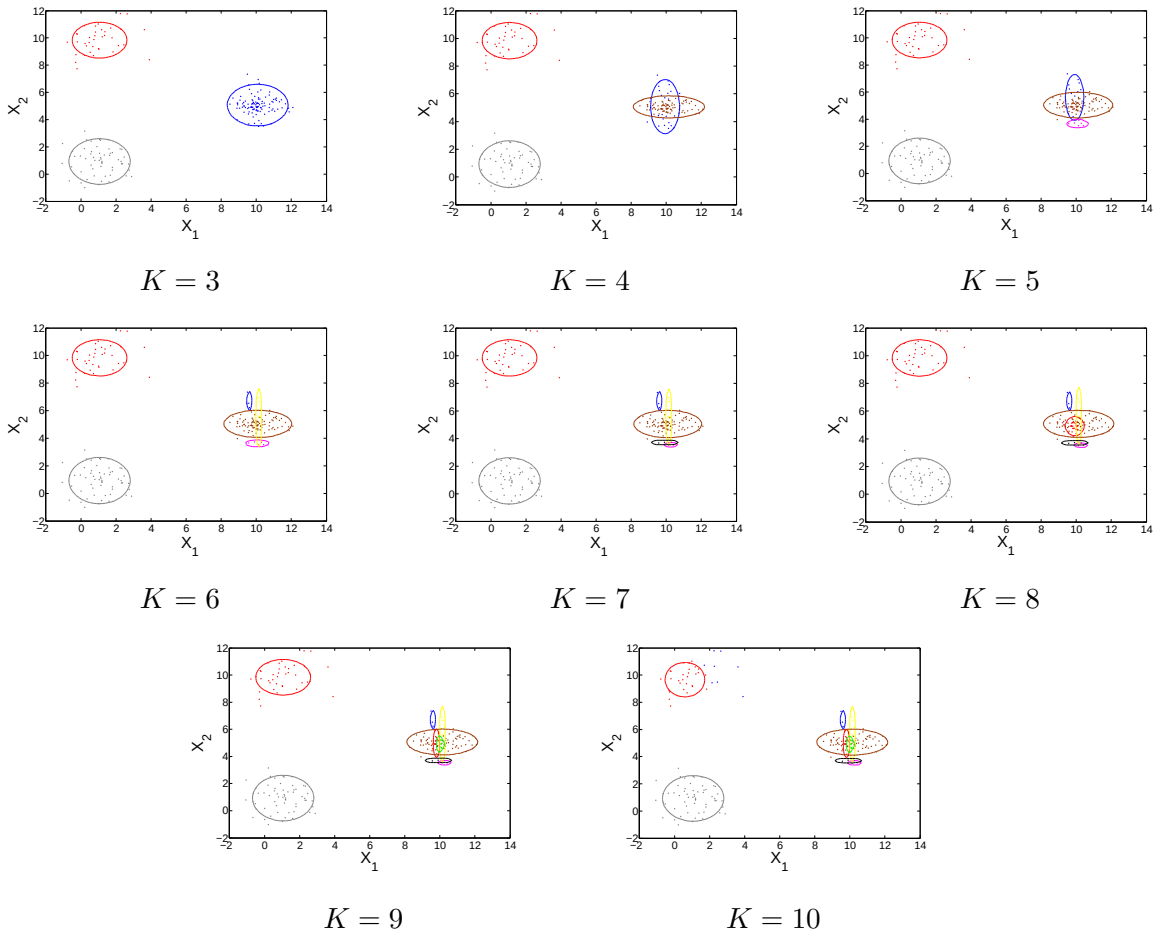


Figure 4.9: "Cross" Experiment.

$\widehat{\theta}_K^{\mathrm{MLE}}$ and the corresponding MAP classification for various values of K.

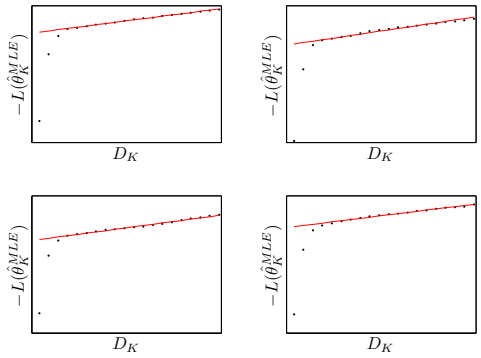


Figure 4.10: "Cross" Experiment.

A few examples of $D_K \mapsto \log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}})$ plots and of the linear regression (red).

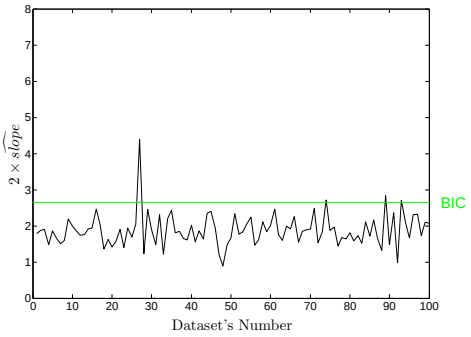


Figure 4.11: "Cross" Experiment.

Values of the penalties for SHL compared to the value of the BIC penalty.

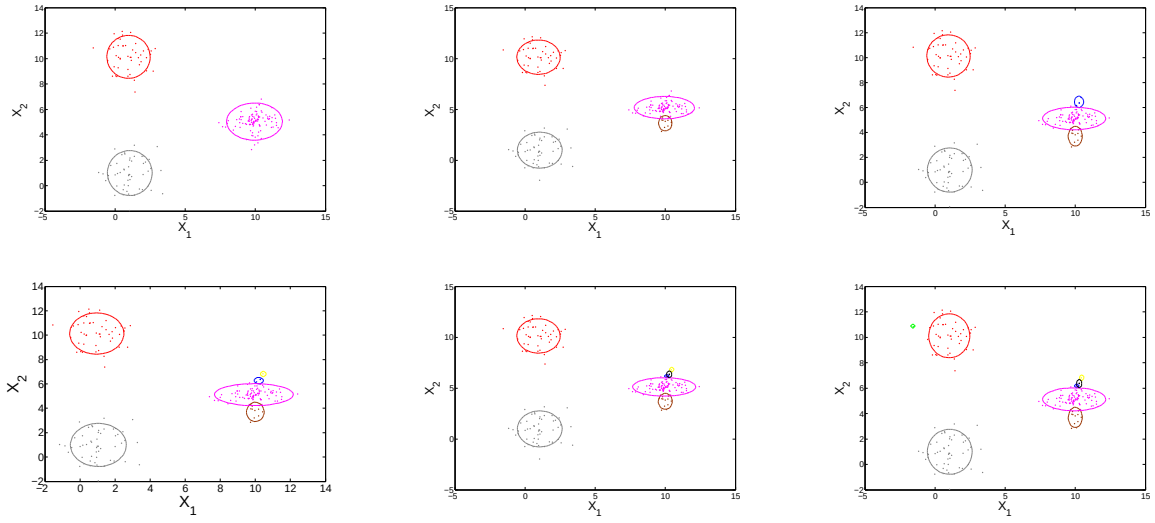


Figure 4.12: "Cross" Experiment.

$\widehat{\theta}_K^{\mathrm{MLccE}}$ and the corresponding MAP classification for various values of K.

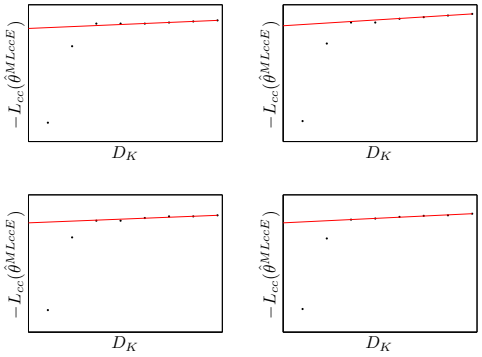


Figure 4.13: "Cross" Experiment.

A few examples of $D_K \mapsto \log \mathrm{L_{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}})$ plots and of the linear regression (red).

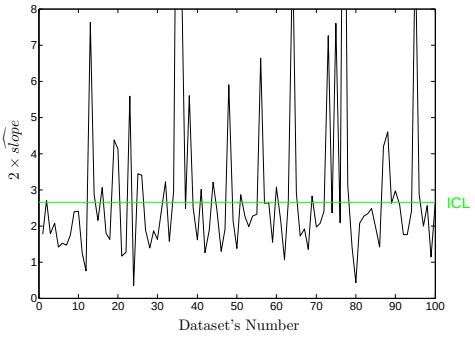


Figure 4.14: "Cross" Experiment.

Values of the penalties for $\mathrm{SHL_{cc}}$ compared to the value of the ICL penalty.

	Risk of the criterion $\times 10^3$	$\frac{\text{Risk of the criterion}}{\text{Risk of the oracle}}$
Oracle	59	1
AIC	506	8.03
BIC	65	1.10
(ICL)	156	2.62
SHL (estimation of the slope)	69	1.17
SHL (dimension jump)	68	1.14

Table 4.2: "Cross" Experiment Results.

Risk of each criterion in terms of Kullback-Leibler divergence to the true distribution, estimated by Monte Carlo simulations. The oracle results reported in the table correspond to the trajectory oracle

$$K_{\mathrm{oracle}} = \underset{1 \leq K \leq 20}{\operatorname{argmin}}\, d_{\mathrm{KL}}\big(f^{\wp}, f(\,.\,; \widehat{\theta}_K^{\mathrm{MLE}})\big)$$

for each dataset. The expected oracle number of components

$$K_{\mathrm{oracle}} = \underset{1 \leq K \leq 20}{\operatorname{argmin}}\, \mathbb{E}_{f^{\wp}}\left[d_{\mathrm{KL}}\big(f^{\wp}, f(\,.\,; \widehat{\theta}_K^{\mathrm{MLE}})\big)\right]$$

is four. The true number of components is four.

	"Risk" of the criterion $\times 10^3$
Oracle	3618
ICL	3622
$\mathrm{L_{cc}}$-ICL	3623
$\mathrm{SHL_{cc}}$ (estimation of the slope)	3632

Table 4.3: "Cross" Experiment Results.

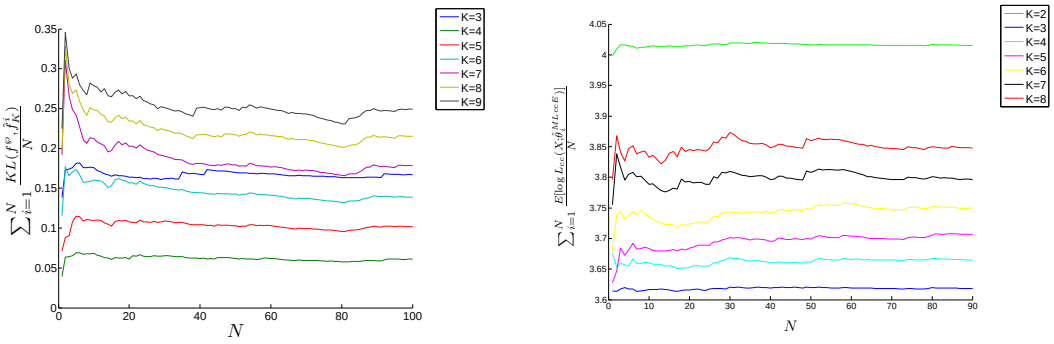
"Risk" of each criterion for the $\mathrm{L_{cc}}$ contrast, estimated by Monte Carlo simulations. The oracle results reported in the table correspond to the trajectory oracle:

$$K_{\text{oracle}} = \underset{1 \leq K \leq 8}{\operatorname{argmin}} \, \mathbb{E}_X \left[-\log \mathrm{L_{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}; X) \right]$$

for each dataset. The expected oracle number of components

$$K_{\text{oracle}} = \underset{1 \leq K \leq 8}{\operatorname{argmin}} \, \mathbb{E}_X \mathbb{E}_{X_1,\ldots,X_n} \left[-\log \mathrm{L_{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}; X) \right]$$

is three.



(a) (b)

Figure 4.15: "Cross" Experiment.

Convergence of the Monte Carlo simulations for the computation of

$$K_{\text{oracle}} = \underset{1 \leq K \leq 20}{\operatorname{argmin}} \, \mathbb{E}_{X_1,\ldots,X_n} \left[d_{\mathrm{KL}}\big(f^{\wp}, f(\,.\,; \widehat{\theta}_K^{\mathrm{MLE}})\big) \right] \text{ (a)}$$

$$K_{\text{oracle}} = \underset{1 \leq K \leq 8}{\operatorname{argmin}} \, \mathbb{E}_X \mathbb{E}_{X_1,\ldots,X_n} \left[-\log \mathrm{L_{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}; X) \right] \text{ (b).}$$

From Table 4.4:

AIC really has troubles in this situation, which is not surprising since we already noticed its tendency to select a number of components much higher than expected in the previous experiment. It is not adapted to this aim.

BIC tends to select a quite high number of components, in this experiment, too, which is not a problem from the point of view to which it corresponds. Indeed, the number of diagonal components needed to approximate $f^{\wp}$ is greater than four, because of the two non-diagonal components. See Figure 4.17 (a) for an example.

SHL, both applied with the data-driven slope estimation or with the dimension jump approach, yields the selections of $\hat{K}$ the closest to the oracle's (Table 4.4). However, it does not yield better risk results than BIC in this experiment (Table 4.5): the three criteria achieve good results. Giving a precise penalty value for the slope heuristics is delicate (see for example Section 5.2: our package yields a penalties interval rather than a single penalty value). However, remark that the comparison of Figure 4.19 and Figure 4.14 suggests the data-driven feature of the criterion: the slope heuristics yields lower penalties than BIC in both cases, but the values of the penalties are greater in the misspecified experiment.

ICL and $\mathrm{L_{cc}}$-ICL select the expected four classes half of the time. The number of observations (200, and the bottom-right component has proportion 0.2, the top-right, 0.3) does not always enable them to notice that some fitted components of the five- or six-component solution overlap. Remark that $\mathrm{L_{cc}}$-ICL, if it recovers four classes a little more often than ICL, also selects six components more often than ICL.

$\mathrm{SHL_{cc}}$ reaches the "best" results (from the clustering point of view) in this experiment, in the sense that it recovers the expected four classes most often. It yields heavier penalties than ICL (see Figure 4.21) in this experiment. It is more robust than ICL to a misspecified model and is less troubled by a non-asymptotic situation.

The efficiency point of view (Tables 4.5 and 4.6) confirms the difficulty of this setting for the criteria linked to $\log \mathrm{L}$: the oracle risk is quite large (it is for example much larger than in the "cross" experiment: compare with Table 4.2). Moreover, Figure 4.22(a) illustrates how difficult it is to select a number of components, even with Monte Carlo simulations. Things are easier for the $\mathrm{L_{cc}}$: see Figure 4.22(b).

Experiment with a Distorted Component

In this experiment, $f^{\wp}$ is a four-component Gaussian mixture in $\mathbb{R}^2$: see Figure 4.23. Three of the components are well separated but the fourth is smaller than the others (in size: $\pi = 0.1$ against 0.3 for the others, and in volume: $\det \Sigma_4 = 0.01$ against 1 or 0.5 for the others). Diagonal mixture models are fitted: this is a not a misspecified situation.

From Table 4.7:

BIC and SHL mostly recover the four Gaussian components. This is what they are expected to do, and SHL behaves a little better than BIC in this experiment. This is confirmed in an efficiency perspective: see Table 4.8.

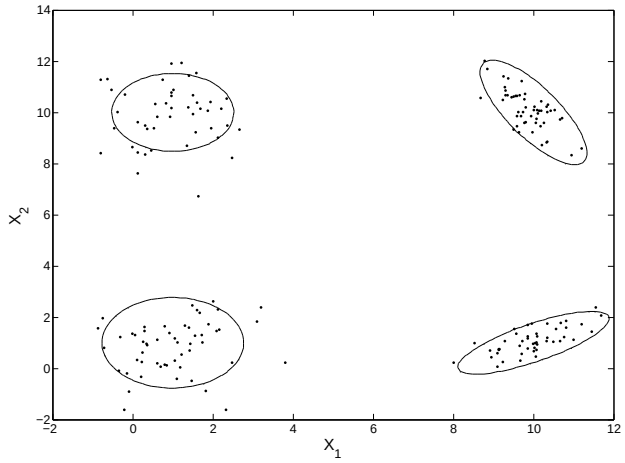


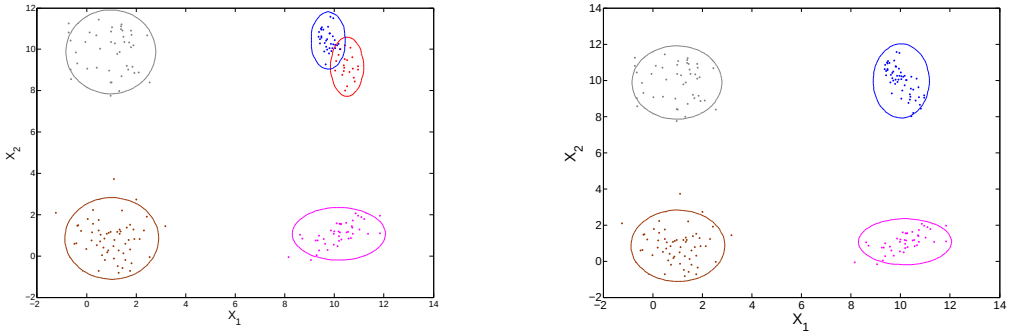
Figure 4.16: Misspecified Models Experiment.

Simulated observations example with isodensity contours of the true distribution. $n = 200$

Selected number of components	4	5	6	7	8	9–16	17	18	19	20
Oracle	4	10	30	43	12	1	0	0	0	0
AIC	0	0	0	0	0	20	14	12	26	28
BIC	3	43	38	13	3	0	0	0	0	0
SHL (ddes)	2	19	26	32	11	10	0	0	0	0
SHL (dj)	2	25	33	20	11	9	0	0	0	0
ICL	49	35	9	5	2	0	0	0	0	0
L_{cc}-ICL	54	29	13	4	0					
SHL_{cc}	81	17	2	0	0					

Table 4.4: Misspecified Models Experiment Results.

"ddes" indicates the data-driven slope estimation approach for the slope heuristics application and "dj" the dimension jump approach. Some boxes are left blank to recall the criteria related to L_{cc} have been computed with $K \in \{1, \ldots, 8\}$ only.



MLE (BIC)
(a)

MLccE (ICL)
(b)

Figure 4.17: Misspecified Models Experiment.

Typical MLE solution selected by BIC (a) and MLccE solution selected by ICL (b).

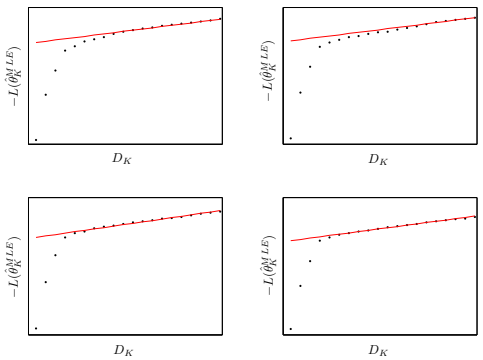


Figure 4.18: Misspecified Models Experiment.

A few examples of $D_K \mapsto \log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}})$ plots and of the linear regression (red)

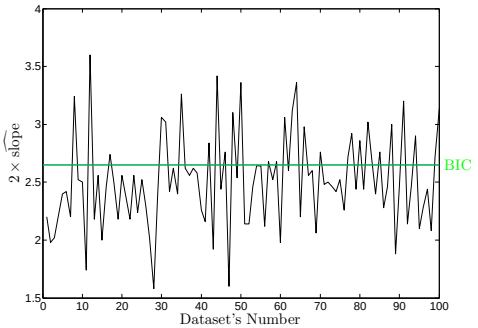


Figure 4.19: Misspecified Models Experiment.

Values of the penalties for SHL compared to the value of the BIC penalty

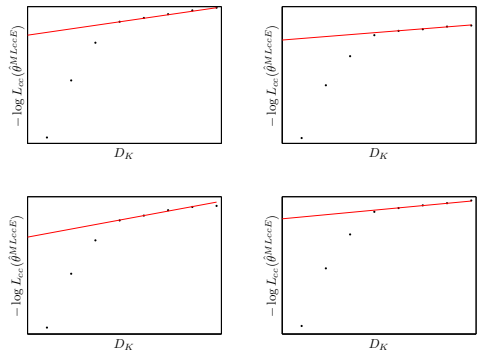


Figure 4.20: Misspecified Models Experiment.

A few examples of $D_K \mapsto \log \mathrm{L_{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}})$ plots and of the linear regression (red)

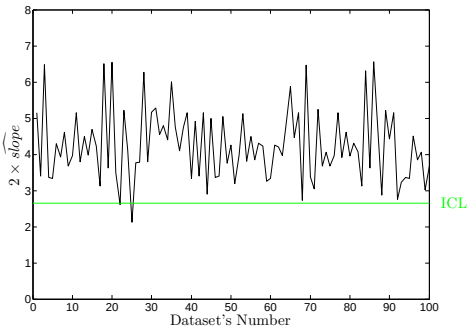


Figure 4.21: Misspecified Models Experiment.

Values of the penalties for $\mathrm{SHL_{cc}}$ compared to the value of the ICL penalty

	Risk of the criterion $\times 10^3$	$\frac{\text{Risk of the criterion}}{\text{Risk of the oracle}}$
Oracle	206	1
AIC	712	3.45
BIC	240	1.16
(ICL)	272	1.32
SHL (estimation of the slope)	249	1.21
SHL (dimension jump)	243	1.18

Table 4.5: Misspecified Models Experiment Results.

Risk of each criterion in terms of Kullback-Leibler divergence to the true distribution, estimated by Monte Carlo simulations. The oracle results reported in the table correspond to the trajectory oracle:

$$K_{\text{oracle}} = \underset{1 \leq K \leq 20}{\operatorname{argmin}}\, d_{\mathrm{KL}}\big(f^{\wp}, f(\,.\,; \widehat{\theta}_K^{\mathrm{MLE}})\big)$$

for each dataset. The expected oracle number of components:

$$K_{\text{oracle}} = \underset{1 \leq K \leq 20}{\operatorname{argmin}}\, \mathbb{E}_{f^{\wp}}\left[d_{\mathrm{KL}}\big(f^{\wp}, f(\,.\,; \widehat{\theta}_K^{\mathrm{MLE}})\big)\right]$$

is six or seven (see Figure 4.22). The "true" number of components is four, but the model is misspecified.

	"Risk" of the criterion $\times 10^3$
Oracle	3910
ICL	3926
L_{cc}-ICL	3928
$\mathrm{SHL}_{\mathrm{cc}}$ (estimation of the slope)	3915

Table 4.6: Misspecified Models Experiment Results.

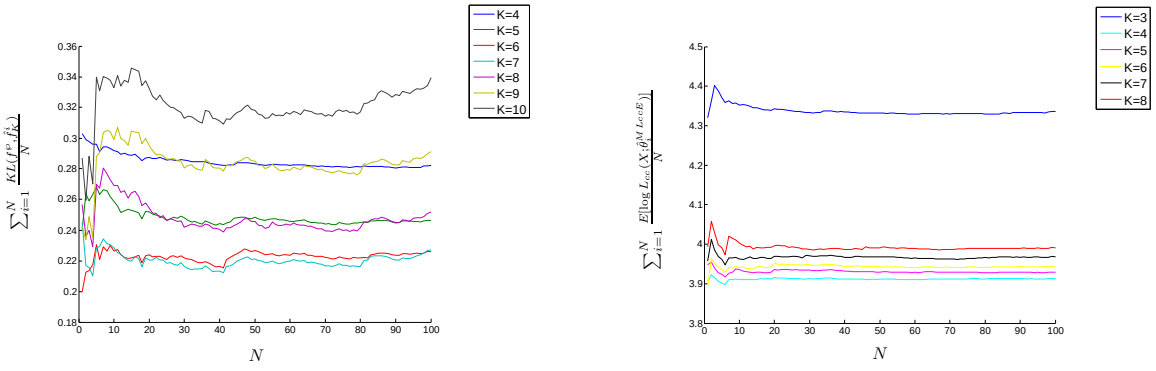
"Risk" of each criterion for the L_{cc} contrast, estimated by Monte Carlo simulations. The oracle results reported in the table correspond to the trajectory oracle:

$$K_{\text{oracle}} = \underset{1\leq K\leq 8}{\operatorname{argmin}}\, \mathbb{E}_X\left[-\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}; X)\right]$$

for each dataset. The expected oracle number of components:

$$K_{\text{oracle}} = \underset{1\leq K\leq 8}{\operatorname{argmin}}\, \mathbb{E}_X\mathbb{E}_{X_1,\ldots,X_n}\left[-\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}; X)\right]$$

is four.



(a) (b)

Figure 4.22: Misspecified Models Experiment.

Convergence of the Monte Carlo simulations for the computation of
$K_{\text{oracle}} = \underset{1\leq K\leq 20}{\operatorname{argmin}}\, \mathbb{E}_{X_1,\ldots,X_n}\left[d_{\mathrm{KL}}\left(f^{\wp}, f(\,.\,;\widehat{\theta}_K^{\mathrm{MLE}})\right)\right]$ (a)
$K_{\text{oracle}} = \underset{1\leq K\leq 8}{\operatorname{argmin}}\, \mathbb{E}_X\mathbb{E}_{X_1,\ldots,X_n}\left[-\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}; X)\right]$ (b).

ICL, L_{cc}-ICL and SHL_{cc} mostly select three classes, in what they conform to the notion of cluster they derive. Both ICL criteria behave even better from this point of view than SHL_{cc}: the optimization difficulties might be the cause.

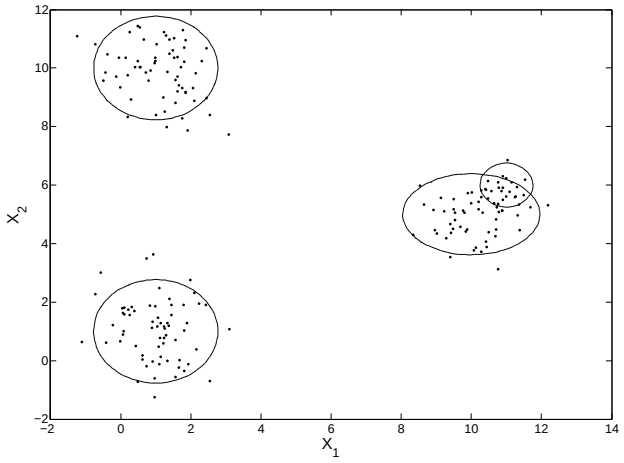


Figure 4.23: Distorted Component Experiment.

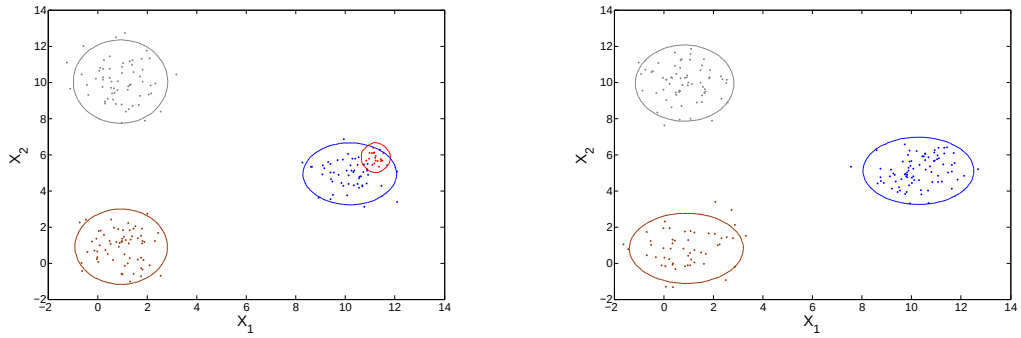
Simulated observations example with isodensity contours of the true distribution. $n = 200$

Selected number of components	3	4	5	6	7	8
AIC	0	24	30	23	3	20
BIC	42	57	0	0	0	1
SHL	22	67	10	1	0	0
ICL	93	7	0	0	0	0
L_{cc}-ICL	98	2	0	0	0	0
SHL_{cc}	78	17	4	1	0	0

Table 4.7: Distorted Component Experiment Results.

Conclusion

In conclusion, these experiments illustrate the different points of view of BIC on the one hand and ICL on the other hand. Beyond the choice of the penalty, they illustrate through the tabulated results that the most decisive — and then the first choice to be made — is the contrast to be involved. From this point of view, BIC and SHL belong to the same family and behave analogously, on the one hand. ICL, L_{cc}-ICL and SHLcc, on the other hand, behave differently from them and conform a clustering point of view. The choice of the penalty, and the comparison of the criteria based on the same contrast, comes after, and from this point of view, BIC and ICL in their respective families of criteria, perform pretty well, while being quite easy to perform, as compared notably to the slope heuristics methods. But those last methods enjoy their data-driven property, at least in some experiments.



MLE (BIC)
(a)

MLccE (ICL)
(b)

Figure 4.24: Distorted Component Experiment.

Typical MLE solution selected by BIC (a) and MLccE solution selected by ICL (b).

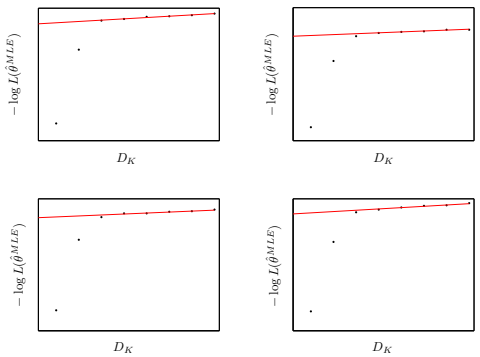


Figure 4.25: Distorted Component Experiment.

A few examples of $D_K \mapsto \log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}})$ plots and of the linear regression (red).

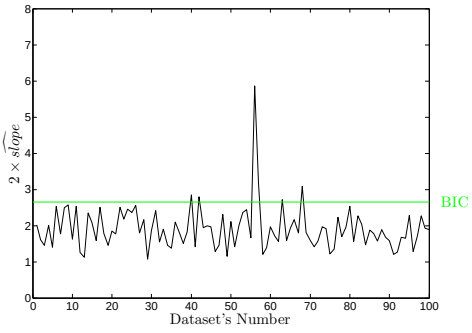


Figure 4.26: Distorted Component Experiment.

Values of the penalties for SHL compared to the value of the BIC penalty.

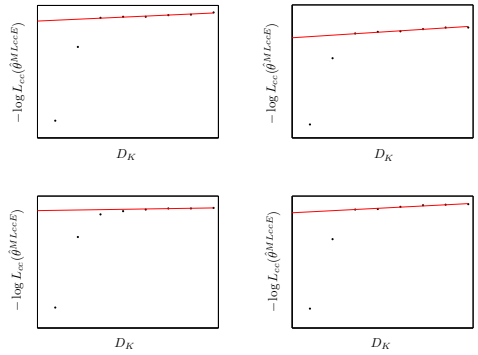


Figure 4.27: Distorted Component Experiment.

A few examples of $D_K \mapsto \log \mathrm{L_{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}})$ plots and of the linear regression (red).

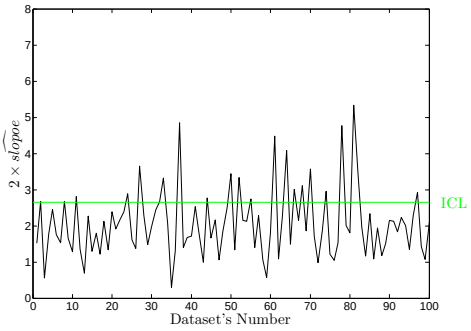


Figure 4.28: Distorted Component Experiment.

Values of the penalties for $\mathrm{SHL_{cc}}$ compared to the value of the ICL penalty.

	Risk of the criterion $\times 10^3$	$\frac{\text{Risk of the criterion}}{\text{Risk of the oracle}}$
Oracle	58.3	1
AIC	108.5	1.9
BIC	73.7	1.3
(ICL)	99.9	1.7
SHL (estimation of the slope)	68.0	1.2

Table 4.8: Distorted Component Experiment Results.

Risk of each criterion in terms of Kullback-Leibler divergence to the true distribution, estimated by Monte Carlo simulations. The oracle results reported in the table correspond to the trajectory oracle:

$$K_{\text{oracle}} = \underset{1 \leq K \leq 8}{\operatorname{argmin}}\, d_{\mathrm{KL}}\big(f^{\wp}, f(\,.\,; \widehat{\theta}_K^{\mathrm{MLE}})\big)$$

for each dataset. The expected oracle number of components:

$$K_{\text{oracle}} = \underset{1 \leq K \leq 8}{\operatorname{argmin}}\, \mathbb{E}_{f^{\wp}}\left[d_{\mathrm{KL}}\big(f^{\wp}, f(\,.\,; \widehat{\theta}_K^{\mathrm{MLE}})\big)\right]$$

is four. The "true" number of components is four.

	"Risk" of the criterion $\times 10^3$
Oracle	3857
ICL	3859
L_{cc}-ICL	3857
$\mathrm{SHL}_{\mathrm{cc}}$ (estimation of the slope)	3863

Table 4.9: Distorted Component Experiment Results.

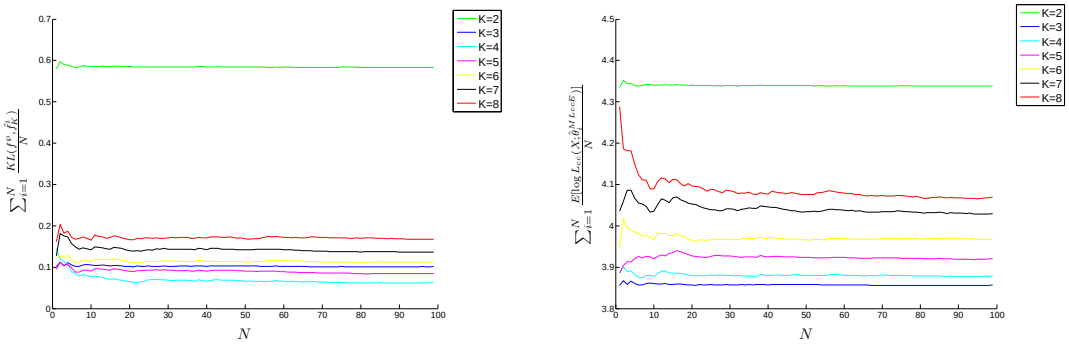
"Risk" of each criterion for the L_{cc} contrast, estimated by Monte Carlo simulations. The oracle results reported in the table correspond to the trajectory oracle:

$$K_{\text{oracle}} = \underset{1\leq K\leq 8}{\operatorname{argmin}}\, \mathbb{E}_X\left[-\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}; X)\right]$$

for each dataset. The expected oracle number of components:

$$K_{\text{oracle}} = \underset{1\leq K\leq 8}{\operatorname{argmin}}\, \mathbb{E}_X\mathbb{E}_{X_1,\ldots,X_n}\left[-\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}; X)\right]$$

is three.



(a) (b)

Figure 4.29: Distorted Component Experiment.

Convergence of the Monte Carlo simulations for the computation of

$$K_{\text{oracle}} = \underset{1\leq K\leq 8}{\operatorname{argmin}}\, \mathbb{E}_{f^{\wp}}\left[d_{\mathrm{KL}}\big(f^{\wp}, f(\,.\,;\widehat{\theta}_K^{\mathrm{MLE}})\big)\right] \text{ (a)}$$

$$K_{\text{oracle}} = \underset{1\leq K\leq 8}{\operatorname{argmin}}\, \mathbb{E}_X\mathbb{E}_{X_1,\ldots,X_n}\left[-\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}; X)\right] \text{ (b)}.$$

4.5 Discussion

Two families of criteria, in the clustering framework, are distinguished in this chapter: it is shown that ICL's purpose is of different nature than that of BIC or AIC. What theory enables to understand, is confirmed by the simulations. The identification theory for the contrasts based on the conditional classification likelihood is — not surprisingly — very similar to the classical one for the usual likelihood.

The main interest of the newly introduced estimator and criteria is to better understand the ICL criterion. The computing difficulties are rather a limit up to now, particularly for applying the slope heuristics to the new contrast: estimation in very complex mixture models is even tougher than in the usual likelihood framework (see Chapter 5). The new criterion L_{cc}-ICL yields choices of the number of classes quite similar to ICL.

ICL leads to discover classes matching a subtle combination of the notions of well separated, compact, clusters, and (Gaussian) mixture components. It then enjoys the flexibility and modeling possibilities of the model-based clustering approach, but does not break the expected notion of cluster.

The choice of the involved mixture components must be handled with care in this framework since it leads the cluster shape underlying the study. Several forms of Gaussian mixtures may be involved: for example, spherical (i.e. with covariance matrices proportional to the identity matrix) and general models may be compared, or models with free proportions may be compared with models with equal proportions. This definitely makes sense from a clustering point of view and depends on the application. But some settings do not enable the use of the slope heuristics: it is at least required that the bias be stationary for the most complex models, which means that there should not be a subfamily of models with lower bias than the other, when D_K is great. Moreover, it should be justified that the variance of both subfamilies behaves analogously for large values of D_K, which is not obvious. In this case, only criteria such as ICL or L_{cc}-ICL may be applied.

Besides it should be further studied how the complexity of the models should be measured when several model kinds are compared. The dimension of the model as a parametric space works for the reported theoretical results. But we are not completely convinced that it is the finest measure of the complexity of Gaussian mixture models. As a matter of fact this simple parametric point of view amounts to considering that all parameters play an analogous role, at least when measuring the complexity of the model. This is not really natural. Think for example about how different are the respective roles of a mean parameter on the one hand, and a non-diagonal covariance matrix coefficient on the other hand. Remark that a model with diagonal covariance matrices and equal proportions has dimension $2Kd$. A model with spherical covariance matrices with equal volumes has dimension $(Kd+1)$. But it does not seem really natural that a spherical model with $2K$ components is about as "complex" as a diagonal model with K components...

Further results would be necessary to fully justify that penalties proportional to the dimension are optimal. The situation in this mixture framework is known to be a little different, as compared to other frameworks as the regression, since the AIC penalty is

not heavy enough to yield efficient procedures.

The optimization algorithms need be improved to be more reliable, and above all to run much faster, which would obviously be a condition for a spread practical use of the new contrast, and then for a further understanding of its features.

A possibility to make this contrast more flexible would be to assign different weights to the log likelihood and the entropy:

$$\log \mathrm{L}_{\mathrm{cc}_\alpha} = \alpha \log \mathrm{L} + (1-\alpha)\,\mathrm{ENT},$$

with $\alpha \in [0;1]$. This would enable to tune how important the assignment confidence is with respect to the Gaussian fit...the practical interest of such a procedure is obvious, but the difficulty coming with it is also: how to choose α? The derivation of L_{cc} as the conditional classification likelihood would not hold anymore. A first insight which comes in mind is to choose α from simulations of situations in which the user knows what solution he expects.

Chapter 5

Practical

Contents

Presentation In this chapter are introduced and discussed practical solutions for the application of methods introduced in the preceding chapters. In Chapter 4 a new estimator has been proposed as a concurrent choice to the MLE in the clustering framework: MLccE. The MLE in the Gaussian mixture model framework is known to be difficult to be computed (see Section 1.2.2). The EM algorithm makes this computation tractable. The MLccE is seemingly even tougher to be computed: fortunately an algorithm can be derived, which makes this possible. This is an adapted EM algorithm and we shall see that it inherits its fundamental property — namely the monotonicity of the contrast along the iterations. Moreover, as for the MLE, the initialization step is crucial and must be sensible so as to get good results. Several initialization methods are discussed and a new one is introduced to overcome the limitations of the known ones. This method may be applied to the EM initialization and improve sensibly the results in this framework, too. Finally, it is introduced and discussed how to choose and practically impose bounds on the parameter space. Section 5.2 is about the slope heuristics, which has been introduced in Section 3.2, and applied in simulation studies in Sections 3.3 and 4.4.6. Its practical use involves difficulties, notably while applying the data-driven slope estimation approach (as compared to the dimension jump approach: see Sections 3.2.2 and 3.2.3). Yet the slope heuristics is a promising model selection approach and some first simulations show the data-driven slope estimation to be competitive with the dimension jump, and even to be more relevant than it in some cases at least. Hence the necessity of proposing practical solutions for its application. Jointly with C. Maugis and B. Michel a Matlab package is being developed to make those solutions available and easy to use for any interested user. The introduction of the embraced solutions and of the package, and a simulation study which highlights both the good behavior of the slope heuristics and the differences between the dimension jump and the data-driven slope estimation constitute the second section of this chapter.

5.1 Computing MLccE: L_{cc}-EM and Practical Considerations

A new contrast, $-\log \mathrm{L_{cc}}$, has been defined in Section 4.2[1], and applied to the definition of a new estimator. Computing this estimator requires to develop adapted tools. For a given Gaussian mixture model $\mathcal{M}_K$ with parameter space Θ_K, $\widehat{\theta}_K^{\mathrm{MLccE}}$ is

$$\widehat{\theta}_K^{\mathrm{MLccE}} = \operatorname*{argmax}_{\theta \in \Theta_K} \log \mathrm{L_{cc}}(\theta).$$

Recall, with the notation of Chapter 4, that

$$\begin{aligned} \log \mathrm{L_{cc}}(\theta) &= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}(\theta) \log \pi_k \phi(x_i; \omega_k) \\ &= \log \mathrm{L}(\theta) + \underbrace{\sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}(\theta) \log \tau_{ik}(\theta)}_{-\mathrm{ENT}(\theta;\mathbf{x})}. \end{aligned}$$

[1] Please refer to the introduction sections of Chapter 4 for the notation.

$\theta \in \Theta_K \mapsto \log \mathrm{L_{cc}}(\theta)$ is obviously not an easy function to optimize. The dimension of Θ_K may be quite high (for a general mixture model, this is $(K-1)+\frac{d(d+3)}{2} \times K$, which is for example with $d=2$ and $K=5$: $D_K=29$); the function is not convex and it may even have lots of local maxima. The usual likelihood is known in the mixture framework to be tough to optimize. The conditional classification likelihood analytic expression is even harder to handle, because of the entropy term. It is for example hopeless to try to solve the likelihood equations, even in the simplest situations. Actually, we did not achieve such a calculation by hand exactly in any situation, up to now.

Because the functions are nonconvex, and the dimension is high, it is hard to reach the likelihood or the conditional classification likelihood maximum by an usual, simple algorithm as a gradient or Newton method, too.

Fortunately, the EM algorithm has been developed, which enables to reasonably optimize the likelihood of mixtures within reasonable computation time. Dempster et al. (1977) proposed this now widespread algorithm (see Section 1.2.2). It was actually employed to maximize the likelihood for the computation of the criteria relying on it in the simulations section (Section 4.4.6). We did not have to implement it since several softwares are available, which run the EM algorithm in the Gaussian mixture framework. We employed the MIXMOD software of Biernacki et al. (2006). But the EM algorithm cannot directly be applied to our purpose with $\log \mathrm{L_{cc}}$. It may however be adapted to make this task tractable.

The EM algorithm has been presented in Section 1.2.2 already. Let us merely first recall the EM algorithm steps. Recall $\mathrm{L_c}(\theta;(\mathbf{x},\mathbf{z})) = \prod_{i=1}^{n}\prod_{k=1}^{K}(\pi_k\phi(x_i;\omega_k))^{z_{ik}}$. The steps to compute θ^{j+1}, the current parameter estimation being θ^j, are

E step Compute for any $\theta \in \Theta_K$, $Q(\theta,\theta^j) = \mathbb{E}_{\theta^j}\left[\log \mathrm{L_c}(\theta;\mathbf{X},\mathbf{Z})|\mathbf{X}\right]$. This amounts to computing each $\tau_{ik}^j = \tau_k(X_i;\theta^j)$.

M step Maximize $Q(\theta,\theta^j)$ with respect to $\theta \in \Theta_K$ to get θ^{j+1}.

Recall the fundamental property of EM (see Theorem 2, Chapter 1, page 32): $\log \mathrm{L}(\theta^j)$ is increased at each iteration of the algorithm. This property still holds if Q is increased — and not necessarily maximized — at the M step.

This algorithm has been adapted to several different situations. Our adaptation was notably inspired by the so-called BEM algorithm (Bayesian Expectation Maximization): see for example Lange (1999) for this algorithm which is an adapted-EM for the case the likelihood has to be maximized while tacking into account a prior on the parameter.

The solutions to practical problems suggested in this section have been applied to perform the simulations of the simulations sections (Section 4.3.4 and Section 4.4.6).

5.1.1 Definition and Fundamental Property of $\mathrm{L_{cc}}$-EM

Let us call $\mathrm{L_{cc}}$-EM the adapted algorithm. The steps of the j^{th} algorithm iteration $(\theta^{j-1} \to \theta^j)$ are:

E step Compute for any $\theta \in \Theta_K$,

$$\begin{aligned} Q(\theta, \theta^{j-1}) &= \mathbb{E}_{\theta^{j-1}}\left[\log \mathrm{L_c}(\theta)|\mathbf{X}\right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}(\theta^{j-1}) \log \pi_k \phi(X_i;\theta_k). \end{aligned}$$

M step Maximize $Q(\theta, \theta^{j-1}) - \mathrm{ENT}(\theta;\mathbf{X})$ with respect to $\theta \in \Theta_K$:

$$\theta^j \in \operatorname*{argmax}_{\theta\in\Theta_K} \Bigg\{ \underbrace{\mathbb{E}_{\theta^{j-1}}\left[\log \mathrm{L_c}(\theta)|\mathbf{X}\right] - \mathrm{ENT}(\theta;\mathbf{X})}_{\log \mathrm{L}(\theta) + \sum_{i=1}^{n}\sum_{k=1}^{K}(\tau_{ik}(\theta^{j-1}) + \tau_{ik}(\theta)) \log \tau_{ik}(\theta)} \Bigg\}$$

The maximization in the M-step may be replaced by an increase of $Q(\theta, \theta^j) - \mathrm{ENT}(\theta;\mathbf{X})$: the following still holds. The convergence of the algorithm is nevertheless expected to be even better that this increase is large.

Remark that the $\mathrm{L_{cc}}$-EM algorithm differs from the EM algorithm through the M step. Unfortunately, $\mathrm{L_{cc}}$-EM does not enjoy the nice property of EM that, in many situations, can be run with a closed-form M step (see for example Celeux and Govaert (1995) for many examples of models with closed-form M steps). M step in the $\mathrm{L_{cc}}$-EM therefore has to be performed through a maximization algorithm, and we employed in practice a Matlab function (*fminsearch*, which uses derivative-free method since caclulating the differential in this situation is particularly unpleasant).

The $\mathrm{L_{cc}}$-EM algorithm inherits the fundamental property of the EM algorithm:

Proposition 1 (Fundamental Property of the $\mathrm{L_{cc}}$-EM algorithm)

$$\begin{aligned} &\forall \theta, \theta' \in \Theta_K, \\ &\qquad Q(\theta', \theta) - \mathrm{ENT}(\theta';\mathbf{X}) > Q(\theta, \theta) - \mathrm{ENT}(\theta;\mathbf{X}) \Longrightarrow \log L_{cc}(\theta') > \log L_{cc}(\theta). \end{aligned}$$

The proof is straightforward.

Proof *If*

$$\log L(\theta') + \sum_{i=1}^{n}\sum_{k=1}^{K}\big(\tau_{ik}(\theta)+\tau_{ik}(\theta')\big)\log \tau_{ik}(\theta') > \log L(\theta) + \sum_{i=1}^{n}\sum_{k=1}^{K}\big(\tau_{ik}(\theta)+\tau_{ik}(\theta)\big)\log \tau_{ik}(\theta),$$

then

$$\log L(\theta') - \mathrm{ENT}(\theta';\mathbf{X}) > \log L(\theta) - \mathrm{ENT}(\theta;\mathbf{X}) + \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}(\theta) \log \frac{\tau_{ik}(\theta)}{\tau_{ik}(\theta')}.$$

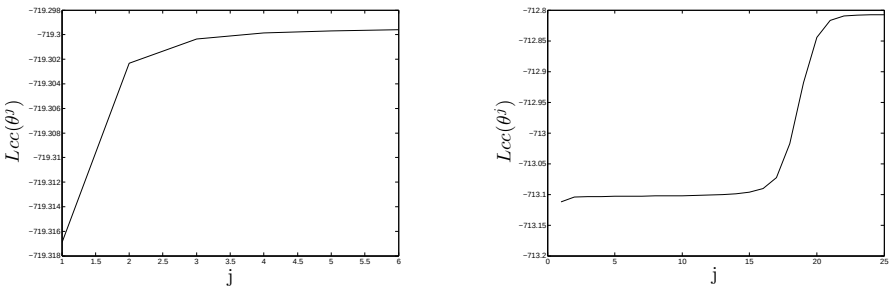
But $\sum_{k=1}^{K} \tau_{ik}(\theta) \log \frac{\tau_{ik}(\theta)}{\tau_{ik}(\theta')}$ *is the Kullback-Leibler divergence from the distribution over* $\{1,\ldots,K\}$ *with probabilities* $\{\tau_{i1}(\theta),\ldots,\tau_{iK}(\theta)\}$ *to that with probabilities* $\{\tau_{i1}(\theta'),\ldots,\tau_{iK}(\theta')\}$ *and hence is a nonnegative quantity.* ■

This property suggests that the algorithm should help optimizing $\log \mathrm{L_{cc}}(\theta)$. Of course, neither does it guarantee that θ^j converges to $\widehat{\theta}^{\mathrm{MLccE}}$, nor does it give any insight into the rate of the increase of $\log \mathrm{L_{cc}}(\theta^j)$ at each iteration...The interesting behavior of this algorithm is mainly assessed from simulations results.

A rule to stop the algorithm has to be specified. Two, at least, seem to be relevant:

- choose $\varepsilon > 0$ and stop the algorithm as soon as $\left| \frac{\log \mathrm{L_{cc}}(\theta^{j+1}) - \log \mathrm{L_{cc}}(\theta^j)}{\log \mathrm{L_{cc}}(\theta^j)} \right| < \varepsilon$;
- choose $\mathrm{Nb_{run}} \in \mathbb{N}$ and run at most $\mathrm{Nb_{run}}$ iterations of the algorithm.

Both rules can be combined: stop the algorithm as soon as one of those conditions holds. Limiting the number of iterations enables to limit the computation time. The choice of ε is more tricky since $\log \mathrm{L_{cc}}(\theta^j)$ might be almost constant for a while, before rising again. But it enables to stop the algorithm when the increase is so small that it seems hopeless that a real change could occur within a reasonable number of iterations. Remark that a good solution to get an idea about the convergence of the algorithm, is to plot $j \mapsto \log \mathrm{L_{cc}}(\theta^j)$: see Figure 5.1. In the experiments we performed, it seems that the most reliable rule is the one based on $\mathrm{Nb_{run}} \in \mathbb{N}$: involving ε is interesting to limit the computation time, but may be misleading, probably because of the phenomenon illustrated in Figure 5.1(b).



(a) (b)

Figure 5.1: Two Typical Examples of the $j \mapsto \log \mathrm{L_{cc}}(\theta^j)$ Graph Behavior

The involved constants have to be carefully chosen. We are not able to provide universal choices, which would provide reasonable results in reasonable computation time whatever the particular situation at hand, up to now. A compromise has to be found between the computation time spent for each M step and the number of iterations. Our experience is that it is worth it to have a good optimization — and thus to spend time — at each M step. Typical settings, when the M step is performed carefully, are $N_{\mathrm{run}} = 15$ or $N_{\mathrm{run}} = 25$. It may be necessary to choose N_{run} smaller because of the computation time, notably while applying the "small_$\mathrm{L_{cc}}$ _EM" initialization (see Section 5.1.3).

Exactly as the EM algorithm does, the sensibility of the solution of the $\mathrm{L_{cc}}$-EM is tightly linked to the initial parameter θ^0: if it is initialized near the attraction of a local maximum, there is little chance that it can leave it, since it would have to cross the valley which separates it from the relevant maximum to do so. Yet it can only cross the

valley in one single step since the $\log L_{cc}$ value cannot decrease, from the fundamental property...Hence the importance of the initialization step.

5.1.2 Initialization: Generalities and Known Methods

The choice of θ^0 is decisive for the L_{cc}-EM, as for the EM algorithm. Besides, all the approaches which use to be employed with the EM algorithm cannot be easily mimicked for the L_{cc}-EM because an iteration of the L_{cc}-EM may cost much more computation time than an iteration of the EM, particularly as the M step is closed-form for EM (see Section 5.1.1 above).

The main idea we advocate is, as Biernacki et al. (2003) claim in the EM context: *"For a good solution, do not skimp on the number of iterations"*. The idea is that several — and actually numerous — initial values of the parameter should be tried. Generate several initial parameters $\theta^{0,1}, \ldots, \theta^{0,N}$ and run the L_{cc}-EM from each one of those parameters, until a "small" stop rule is reached (ε or $\mathrm{Nb}_{\mathrm{run}}$: see Section 5.1.1). Get $\theta^{j_1,1}, \ldots, \theta^{j_N,N}$ (for example, if the stop rule is $\mathrm{Nb}_{\mathrm{run}}$, $j_1 = \cdots = j_N = \mathrm{Nb}_{\mathrm{run}}$). The aim is to explore lots of potentially sensible solutions and to bring the algorithm as close as possible to the highest relevant peak: it then just has to finish the climb. Keep $\theta_0 = \mathrm{argmax}_{\theta \in \{\theta^{j_1,1}, \ldots, \theta^{j_N,N}\}} \log L_{cc}(\theta)$ (which is called from now on the "best" solution) and run the L_{cc}-EM from this parameter until a finer stop rule is reached. This whole procedure may be repeated several times to strengthen the stability of the estimation. We would however moderate what *"do not skimp"* means. Of course, this is limited by the available computation time, but it should beside be taken care that the chance of starting the algorithm from a spurious solution seemingly increases as the number of initial parameters does. We would then restate the advice of Biernacki et al. (2003) as the trickier *"For a good solution, find the right compromise on the number of iterations"*. This holds both for the EM and the L_{cc}-EM.

There still remains to define procedures to design an initialization parameter. Several are employed at the same time, since none has shown to be always the best. Moreover, it is recommended to involve several starts from *each* procedure. Here are some known initialization procedures we tried and which have sometimes been helpful. They are more or less ranked by decreasing usefulness, according to our experience.

CEM The solution generated by the CEM algorithm (see Section 1.4.1) from random starts initialization, as generated by the MIXMOD software, often provided a sensible initialization parameter. This particularly holds as the number K of classes to be designed is reasonably high. This is no surprise, since CEM was designed to maximize the classification likelihood with respect both to the parameter and the labels. However, in the simulations we performed, this approach seldom yields sensible enough initialization parameter for large numbers of components (typically, as K is quite larger than the relevant number of classes).

EM The solutions provided by short runs of EM from random starts (or small-EM starts: run EM from the best solution among those obtained by random starts followed each by short runs of EM, see Biernacki et al., 2003) often yield sensible initialization parameters, too.

K-means Presumably the k-means algorithm should be able to provide interesting results in some situations, since its aim is not that different from the one of $\log \mathrm{L}_{\mathrm{cc}}$. But we quite seldom observed it to be competitive with the other procedures.

The "CEM" and "EM" initialization procedures can be easily applied, with quite low computation time, notably thanks to the efficiency of the EM algorithm and its implementation in the MIXMOD software. Those initialization procedures however do not suffice for our purpose, mainly as the aim is slope heuristics. This method comes with the drawback that the contrast has to be optimized for much higher dimensional models than the one which is finally selected. This means that sensible estimators have to be computed in situations where the number of components to be fitted is much higher than the sensible number of classes — and this is obviously the most difficult situation for the estimation. The procedures presented above may not reach good enough results to apply this method: the obtained values of $\log \mathrm{L}_{\mathrm{cc}}$ are not close enough to the actual maximum value. The obtained values are not even always increasing with K.

5.1.3 Initialization: Adapted and New Methods

Moreover, it is expected to design procedures which do not depend on EM when initializing L_{cc}-EM. Let us introduce two such initialization procedures. The first one is adapted from the Small_EM procedure (see Biernacki et al., 2003). Its usual version in the EM framework provides sensible results. The second one is new, up to our knowledge, and is particularly interesting when successive values of K are considered.

Small_$\mathbf{L_{cc}}$_EM It might be interesting for this method to consider a subsample in case the original sample has a great size: this would make it faster to run. Let us call it $\mathbf{x}$ in any case. We need to define what shall be called "initializing a component at random" in the following: let k be a component label within a K-component mixture. The proportions parameter π_k are always initialized at the uniform distribution: $p_k^0 = \frac{1}{K}$. Then the mean parameter μ_k is chosen at random among the observations: $\mu_k = x_{\tilde{i}}$ for $\tilde{i}$ random. But, to increase the variability of the initialization parameters and since it seemingly enables to find sensible solutions, μ_k is chosen with small probability — say 10% — at random uniformly in the range of the observations (with respect to the sup norm). Now, let us define $\mathbf{x}_s$ as the set of the n_s observations the closest to μ_k. n_s must obviously be chosen with respect to the sample size and to the number of components K. Choose Σ_k as the empirical covariance matrix of $\mathbf{x}_s$. Each time a component of a mixture is chosen at random this way, an iteration of L_{cc}-EM is applied to the whole obtained parameter (and to the whole dataset) to get a parameter which has the wanted form (according to the model form, etc.).

Now, let us describe the initialization procedure. For any $N \in \{1, \ldots, N_{\mathrm{small}}\}$, initialize all components of $\theta^{0,N}_{\mathrm{random}}$ at random, according to the procedure described above. Compute the value of $\log \mathrm{L}_{\mathrm{cc}}$. Choose at random a component label and replace the corresponding component by a component initialized at random. Compute the value of $\log \mathrm{L}_{\mathrm{cc}}$. Repeat this procedure several times. Choose as $\theta^{0,N}_{\mathrm{random}}$ the mixture which maximizes $\log \mathrm{L}_{\mathrm{cc}}$ among those obtained. Apply a few iterations of L_{cc}-EM to it and get $\theta^{1,N}_{\mathrm{random}}$. Now, choose among $\{\theta^{1,1}_{\mathrm{random}}, \ldots, \theta^{1,N_{\mathrm{random}}}_{\mathrm{random}}\}$

the parameter which maximizes $\log L_{cc}$. Apply L_{cc}-EM and choose θ^0_{random} as the obtained parameter.

This method is adapted from the Small_EM method of Biernacki et al. (2003). The differences, beside the replacement of EM by L_{cc}-EM, lie in the way components are chosen at random: Biernacki et al. (2003) choose the mean parameters among the observations and do not enable them to lie anywhere in the range of the data; they initialize the covariance matrices as the covariance of the whole data. We believe this can provide less sensible initialization than the procedure we propose, particularly when the data consists of several subgroups of data well separated from each other. Finally, they choose all components at random at each step, whereas we replace only one component chosen at random. It is expected that this procedure should better explore the parameter space. But the idea of the procedure is the same.

Km1 "Km1" stands for "K minus 1". Figure 5.2 illustrates this strategy. Suppose $K \geq 2$ and $\widehat{\theta}^{\text{MLccE}}_{K-1}$ is available. Then, choose one of the classes designed (through MAP) from $\widehat{\theta}^{\text{MLccE}}_{K-1}$ (say, k_0) and divide it into two classes. This can be done by applying to the corresponding observations the L_{cc}-EM algorithm with a two-component Gaussian mixture model of the same kind as the K-component model being fitted. The result of this L_{cc}-EM should not depend on the initialization procedure employed since a two-component model should mostly not be tough to fit. Now, build $\theta^{0k_0}_{\text{Km1}}$ by keeping the parameters of components with label different from k_0 as in $\widehat{\theta}^{\text{MLccE}}_{K-1}$, and use the parameters obtained for the two components corresponding to the two classes into which k_0 has been divided for the parameters of the k_0^{th} component and the K^{th}. Apply a few iteration of L_{cc}-EM to the obtained parameter and get $\theta^{1k_0}_{\text{Km1}}$. Apply the same procedure with any k_0. Then, apply L_{cc}-

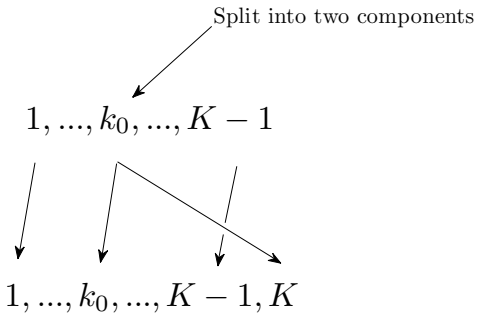


Figure 5.2: The Km1 Strategy

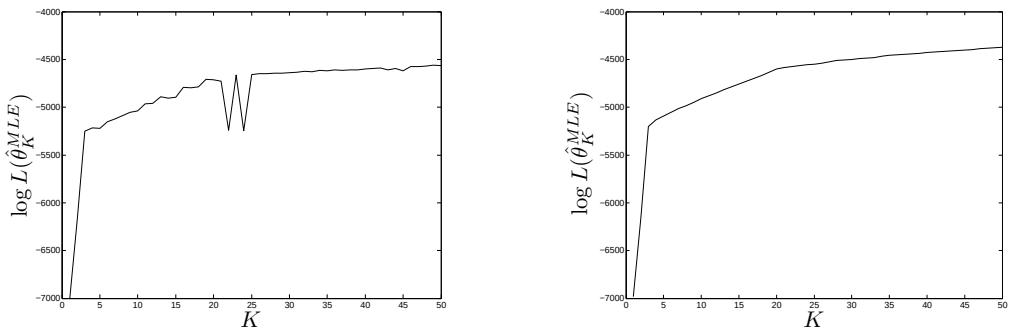
EM to the parameter which maximizes $\log L_{cc}$ among $\{\theta^{11}_{\text{Km1}}, \ldots, \theta^{1K-1}_{\text{Km1}}\}$ and get θ^0_{Km1}.

This method works quite well, particularly as the number of components is larger than the sensible number of classes. When it is smaller, a relevant solution with K components may be quite different from the solution with $K-1$ components and a good initialization algorithm has to be able to discover classes which have different shape and structure than those of the $K-1$ solution, and perhaps merge some of them while splitting others, etc. But when the number of classes is "over-estimated", it is rather expected that all the structure of the classes has been

discovered, and that supplementary classes may be designed by "artificially" dividing some. Remark that it is even expected that this situation favors maxima at the boundary of the parameter space and that small-size classes (i.e. corresponding to components with small proportions) are then presumably to be designed. Besides, it may then make little difference to divide one class or an other. Recall from the slope heuristics, for example, that it is expected that nothing is left to gain from the "bias" point of view and that the only rise of contrast that is observed follows an increased variance.

Initializing $\mathrm{L_{cc}}$-EM from the best solution among $\theta^0_{\mathrm{small}}$ and θ^0_{Km1} provides sensible enough results for the application of the slope heuristics. As expected, $\theta^0_{\mathrm{small}}$ is often more sensible than θ^0_{Km1} for small values of K and the situation is reversed for large values of K. This is a satisfying result to have procedures which do not depend on EM. But those are quite time consuming, and require a sensible choice of the allocation of the computation time among the steps of the algorithms .

To conclude, let us remark that both initialization procedures may be applied for the EM algorithm. When applying the Small_EM procedure in the simulations of this thesis, the procedure of Biernacki et al. (2003) is applied, which is not exactly the one implemented in MIXMOD. Km1 does not seem to correspond to any known procedure. We found that it helps improving the results notably when performing the slope heuristics, since models with high number of components with respect to the "sensible" number are involved. Km1 is much longer to run than Small_EM. Figure 5.3 illustrates the difference between the results obtained with the Small_EM procedure only on the one hand, and with both procedures, on the other hand. Remark that the results of the Small_EM procedures may probably be improved by better choosing the number of iterations of each step. But we already involved for this example quite a stronger setting than the one implemented in MIXMOD. Figure 5.3(a) highlights two difficulties: the first one is that the contrast is not really maximized (compare the values with Figure 5.3(b)) and some solutions may be far from being optimal (see the two "crevasses" between $K = 20$ and $K = 25$, which unfortunately occur in a critical area). The second one is — as a consequence — that the slope is seemingly under-estimated.



(a) (b)

Figure 5.3: "Bubbles" Experiment (see Section 5.2.4): Optimization of the Contrast for Each Model. (a) Initialization Without Km1. (b) Initialization With Km1.

5.1.4 Imposing Parameter Bounds in Practice

How to Impose Parameter Bounds While Applying L_{cc}-EM

The difficulty is to take the chosen bounds on the parameter space into account at the M step. As already mentioned in Chapter 4, it is mainly necessary to guarantee lower bounds on the covariance matrices (and in practice, it suffices to impose a lower bound on their determinant: this is discussed below) and on the proportions.

This can be done by adding terms to the quantity Q to be maximized at the M step (see Section 5.1.1). Those terms should go quickly to minus infinity as the conditions on the parameters are broken. The determinant of the covariance matrix may be imposed to be greater than (around) $\det_{\min}$ by adding to the Q quantity of each M step an exponential term

$$-\exp\left(-10\frac{\min_K \det \Sigma_K - \det_{\min}}{\det_{\min}}\right).$$

This works well in practice. Remark however that this does not provide a bound independent from the data, as it should. As a matter of fact, this term has an effect which is relative to the values of the likelihood. But it grows fast enough that it should always practically provide almost the same bound in reasonable settings.

How to Choose Parameter Bounds

Although most theoretical results have been derived under the compactness assumption on all parameters, it has been mentioned already that this assumption is not necessary in practice for the means parameters. The practice confirms this since we almost never observed any estimated mean larger than the most extremal observation. Moreover, the same reasoning holds for the covariance matrices upper bounds, and the practice confirms it as well.

Things are a little more involved for the proportions and the lower bound on the covariance matrices.

Of course, the theoretical results assume a compact has been fixed, independently from the data. But there is a convenient practical rule on the proportions, which, although it depends on the data, seems to be the least that can be expected: the user probably does not want a class with proportion smaller than $\frac{1}{n}$. This rule can be adopted to bound the proportions.

The bound on the covariance matrices we use is a lower bound on the determinant. It is not a complete lower bound on the covariance matrices since a matrix of any given determinant can be extremely thin in one direction, as soon as it is very large in an other direction — but this rule on the determinant works pretty well in practice. It may actually be a sufficient condition for the theoretical results: it is sufficient for the contrast $-\log L_{cc}$ to be bounded from above (with x fixed) and situations where such a very long and thin covariance matrix would be estimated should occur very seldom with $f^{\wp}$ having reasonably low tails. There still remains to choose a lower bound on the determinant.

In the most favorable cases, a bound can be fixed by the user, according to the

application. Otherwise, a bound can be fixed automatically. The method introduced here provides sensible bounds but can only give a rough order of the constant.

The order $\log \mathrm{L}_{\mathrm{cc}_{\mathrm{typ}}}$ of the $\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}})$ values and the order of the differences between successive values δ_{typ} $(\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}) - \log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_{K-1}^{\mathrm{MLccE}}))$ have to be roughly known. The idea is to avoid covariance matrices which would enable to win against a sensible solution by simply replacing a sensible component by a component centered at any observation, with a very small covariance matrix. Hence δ_{typ} should be chosen by considering numbers of components smaller than the number to be selected: as already mentioned, it is not a problem that the estimates for high values of K involve such components at the boundary, and this is even expected to occur. Since the optimization of the contrast is often easier for the small values of K (the sensible solutions are quite obvious in this situation and must be quite attractive for the algorithm), the procedure might be to first compute $\widehat{\theta}_K^{\mathrm{MLccE}}$ for a few small values of K, choose δ_{typ} and $\log \mathrm{L}_{\mathrm{cc}_{\mathrm{typ}}}$ based on those estimations, deduce a value for $\det_{\min}$ as explained below, and use it to compute $\widehat{\theta}_K^{\mathrm{MLccE}}$ for larger values of K. It may be verified first that the found value of $\det_{\min}$ is consistent with the values of $\det \Sigma_k$ for the first computed $\widehat{\theta}_K^{\mathrm{MLccE}}$.

Remark that only the log likelihood term in $\log \mathrm{L}_{\mathrm{cc}} = \log \mathrm{L} - \mathrm{ENT}$ has to be taken care of since the entropy is bounded. By the way, this method could then be applied as well when the contrast at hand is the usual likelihood. Now, assume the solution with $K-1$ components is $\sum_{k=1}^{K-1} \pi_k \phi(\,.\,;\omega_k)$, and the K-component solution is obtained from it by adding a component which is centered at an observation x_{i_0} and with the smallest covariance matrix determinant as possible, the other components being almost the same as in the $K-1$-component solution. This is the situation we want to avoid. Assuming that $\pi_K \approx \frac{1}{n}$, $\phi(x_i;\omega_K) \approx 0$ as soon as $i \neq i_0$, and $\log \sum_{k=1}^{K-1} \pi_k \phi(x_{i_0};\omega_k) \approx \frac{1}{n} \log \mathrm{L}_{\mathrm{cc}_{\mathrm{typ}}}$ (x_{i_0} is a "mean" point in the $K-1$-component solution...), and with crude approximations:

$$\begin{aligned}
\sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \phi(x_i;\omega_k) - \sum_{i=1}^{n} \log \sum_{k=1}^{K-1} \pi_k \phi(x_i;\omega_k) & \\
= \sum_{i \neq i_0} \underbrace{\log\left(1 + \frac{\pi_K \phi(x_i;\omega_K)}{\sum_{k=1}^{K-1} \pi_k \phi(x_i;\omega_k)}\right)}_{\approx 0} & \\
+ \log\left(1 + \frac{\pi_K \phi(x_{i_0};\omega_K)}{\sum_{k=1}^{K-1} \pi_k \phi(x_{i_0};\omega_k)}\right) & \\
\approx \log\left(1 + \frac{(\det \Sigma_K)^{-\frac{1}{2}}}{n(2\pi)^{\frac{d}{2}} \mathrm{e}^{\frac{1}{n}\log \mathrm{L}_{\mathrm{cc}_{\mathrm{typ}}}}}\right). &
\end{aligned}$$

Now, to avoid such a situation, guarantee that

$$\begin{aligned}
\log\left(1 + \frac{(\det \Sigma_K)^{-\frac{1}{2}}}{n(2\pi)^{\frac{d}{2}} \mathrm{e}^{\frac{1}{n}\log \mathrm{L}_{\mathrm{cc}_{\mathrm{typ}}}}}\right) << \delta_{\mathrm{typ}} &\iff \frac{(\det \Sigma_K)^{-\frac{1}{2}}}{n(2\pi)^{\frac{d}{2}} \mathrm{L}_{\mathrm{cc}_{\mathrm{typ}}}^{\frac{1}{n}}} << \mathrm{e}^{\delta_{\mathrm{typ}}} \\
&\iff \det \Sigma_K >> \frac{\mathrm{L}_{\mathrm{cc}_{\mathrm{typ}}}^{-\frac{2}{n}}}{\mathrm{e}^{2\delta_{\mathrm{typ}}} (2\pi)^d n^2}.
\end{aligned}$$

This should help the user to choose the order of the lower bound on the covariance matrices determinants.

5.2 A Matlab Package for the Slope Heuristics

A work in progress jointly with C. Maugis and B. Michel aims at developing an easy-to-use software for the application of the slope heuristics. Presumably, the data-driven and non-asymptotic properties of the slope heuristics may be valuable for some applications. This is a new method, which is theoretically validated in the regression with histograms framework (homoscedastic, fixed design and Gaussian framework, with possibly huge families of models in Birgé and Massart (2006) and heteroscedastic, random design and not necessarily Gaussian framework, but with polynomial complexity (with respect to n) of the family of models in Arlot and Massart, 2009). However, the authors of those articles expect the heuristics to hold in a much wider range of applications. This conjecture is supported by some practical applications, already (see Section 3.2). However, beside possible further theoretical results to come, much more examples of practical applications would be needed to better understand its practical behavior.

But, as mentioned in Section 3.2, its practical use involves some technical difficulties, which may prevent many statisticians to try it for the sake of curiosity. This Matlab package should make it easy to give the slope heuristics a try, and perhaps then elaborate further on it if it seems to yield interesting results in the considered framework. Therefore is this a graphical package: the user mainly has to click with the mouse and see...

Hopefully shall this package contribute to a widespread use of the slope heuristics!

5.2.1 Slope Heuristics Recalled

Recall from Section 3.2: in a contrast minimization framework with the contrast γ and with a models family $\left(\mathcal{M}_m\right)_{m\in M}$, the penalty $\mathrm{pen}_{\mathrm{opt}}$ of an optimal model selection criterion $\mathrm{crit}_{\mathrm{opt}}$ is assumed to be known from theory up to a multiplying factor:

$$\exists \kappa_{\mathrm{opt}} \text{ such that } \mathrm{pen}_{\mathrm{opt}}(m) = \kappa_{\mathrm{opt}}\, \mathrm{pen}_{\mathrm{shape}}(m).$$

$\mathrm{pen}_{\mathrm{shape}}(m)$ is denoted $f(m)$ from now on, to be consistent with the package notation. Moreover, the assumptions (SH1) and (SH2) underlying the slope heuristics are assumed and recalled:

SH1 there exists a minimal penalty $\mathrm{pen}_{\mathrm{min}}$ such that any lighter penalty selects models with clearly too high complexities and such that heavier penalties select models with reasonable complexity;

SH2 twice the minimal penalty is an optimal penalty.

$\mathrm{pen}_{\mathrm{min}}$ is assumed to be of the form $\kappa_{\mathrm{min}}\, \mathrm{pen}_{\mathrm{shape}}$, too. It is to be estimated: the optimal penalty is then chosen as twice the minimal penalty.

Let us introduce the Matlab package we propose to perform both the data-driven slope estimation (see Section 3.2.3) and the dimension jump (see Section 3.2.2) to take advantage of the slope heuristics to estimate κ_{opt}, and the solutions we embraced to overcome the practical difficulties when applying it.

Figure 5.4 is an example of the main window of the Matlab package.

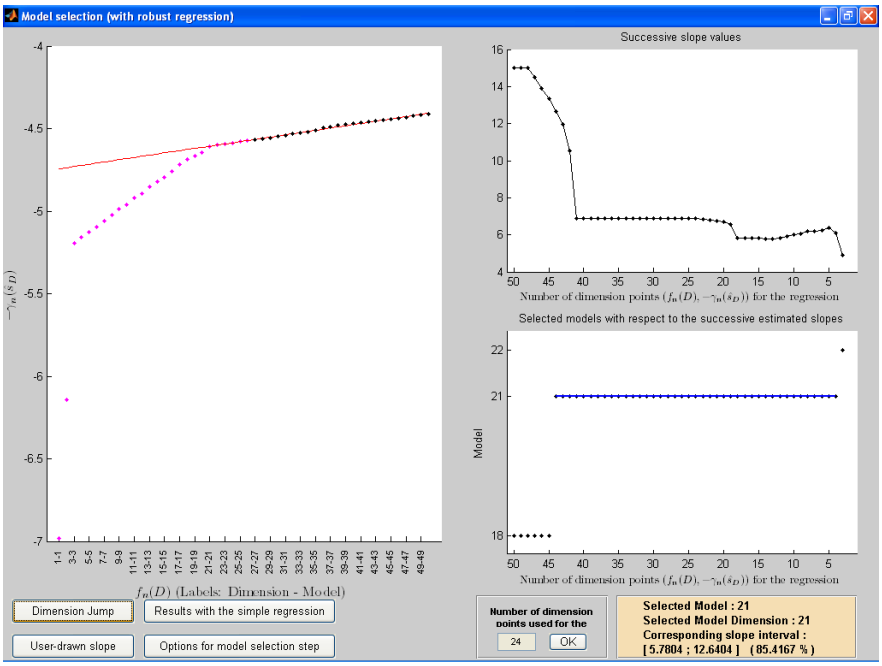


Figure 5.4: The Main Window

The user provides the package a simple file *file.txt* which contains the needed informations for each model m:

- the model name m;
- the model "dimension" D_m;
- the value of the penalty shape $f(m)$;
- the maximum value of the empirical contrast in the model $\gamma_n(\hat{s}_m)$.

Recall a "dimension" (a measure of the complexity of the model) is needed to apply the dimension jump method, but not necessarily to apply the data-driven slope estimation. In case the user is not interested in the dimension jump solution, any value can be provided for the dimension of the model. However, a penalty shape value is needed in any case, and could be actually regarded as a complexity measure itself. It is not required that $f(m)$ be a function of D_m in any case.

Remark that, in the current version of the package, models are gathered according to the "dimension" column. This is not consistent with the following presentation, where models are gathered according to their respective penalty shape value. Both approaches match when $f(m)$ is a function of D_m, but may not otherwise. Obviously, the user may then use the package by replacing the dimension values by the corresponding penalty shape values in the input file. But this requires to use two different files to perform the data-driven slope estimation and the dimension jump — for which the dimension column in the file must actually consist of the dimensions values.

5.2.2 Data-Driven Slope Estimation.

First, the models are gathered according to their penalty shape $f(m)$ value: for a given value of the penalty, only the model reaching the lowest value of the contrast is of interest.

$$\forall p \in \{f(m) : m \in M\},\ m(p) \in \operatorname{argmin} \left\{ \gamma_n(\hat{s}_m) \,\middle|\, m \in M : f(m) = p \right\}.$$

Then, from the slope heuristics (see Section 3.2.3), the function[2]

$$p \in \{f(m) : m \in M\} \longmapsto -\gamma_n(\hat{s}_{m(p)})$$

is expected to be linear for the largest values of p, with slope $\kappa_{\min}$. The algorithm is then:

1. choose $p_{\text{lin}} \in \{f(m) : m \in M\}$ such that $p \geq p_{\text{lin}} \longmapsto -\gamma_n(\hat{s}_{m(p)})$ can be considered as linear;
2. estimate the slope $\frac{\hat{\kappa}}{2}$ of this linear relation;
3. select $\hat{m} = \operatorname{argmin}_{m \in M} \{\gamma_n(\hat{s}_m) + \hat{\kappa} f(m)\}$.

Those tasks are summed up in Figure 5.4: $p \in \{f(m) : m \in M\} \mapsto -\gamma_n(\hat{s}_{m(p)})$ is plotted, together with the regressed linear part (in red) in the left (large) graph. For a given value of p_{lin}, the regression of $p \geq p_{\text{lin}} \mapsto -\gamma_n(\hat{s}_{m(p)})$ is performed through robust linear regression[3] . Robust regression yields more stable results than simple linear regression. This is notably interesting so as to get a method as robust as possible with respect to poor estimations of the values $\gamma_n(\hat{s}_m)$. A button is available to compare the results of the simple and robust regressions: see Figure 5.5. The plots corresponding to the values used for the regression: $\left\{\left(p_{\text{lin}}, -\gamma_n(\hat{s}_{m(p_{\text{lin}})})\right), \ldots, \left(p_{\max}, -\gamma_n(\hat{s}_{m(p_{\max})})\right)\right\}$ are colored in black, the others in violet.

The choice of the p_{lin} value is a crucial and quite hard step. The most reliable approach we tried consists in considering a stability criterion with respect to the selected model, which is actually the quantity of interest.

- Compute the estimated slope $\frac{\hat{\kappa}_{p_0}}{2}$ for the values $\left\{\left(p, -\gamma_n(\hat{s}_{m(p)})\right) : p \in \{f(m) : m \in M\} \text{ and } p \geq p_0\right\}$. p_0 may take any value in $\{f(m) : m \in M\}$, but the two largest ones since at least three points are needed to perform robust regression. Let us denote P_0 this set of possible values of p_0. The number of values of $p \in \{f(m) : m \in M\}$ such that $p \geq p_0$ (namely the number of values from which $\frac{\hat{\kappa}_{p_0}}{2}$ is estimated) is the "number of dimension points for the regression" in the main window Figure 5.4. The values of the corresponding estimated slopes are represented, in the top-right corner.

[2]We consider from now on $-\gamma$ instead of γ for the sake of consistence with the graphical representations.

[3]Iteratively reweighted least squares regression is applied, with a bisquare weighting function: $(1 - r^2)^2 I_{|r|<1}$ (where r is a function of the residuals which has to be tuned according to the expected robustness of the procedure), through the Matlab *Robustfit* function.

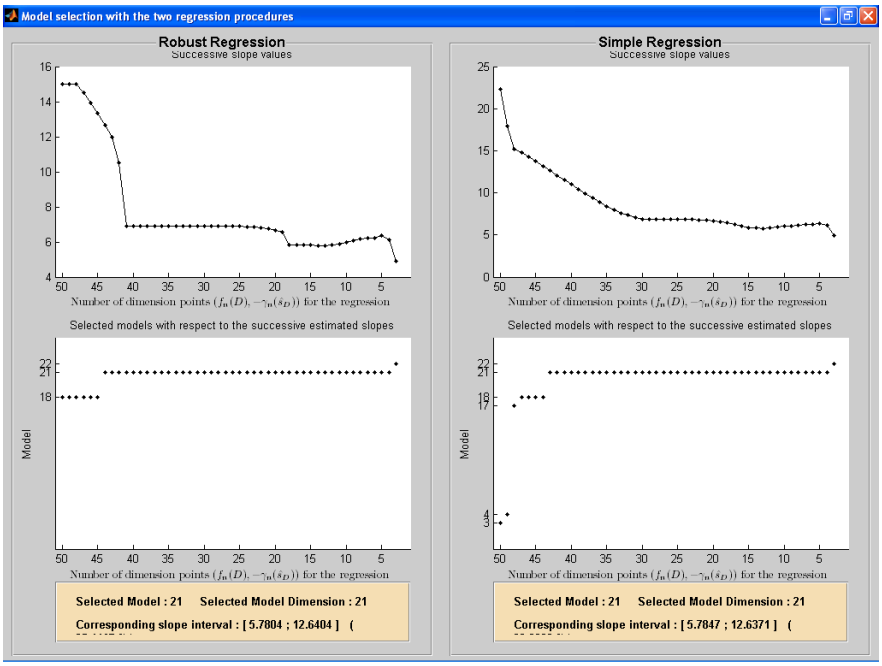


Figure 5.5: Compare the Robust and the Simple Regression

- For each value of p_0, compute the corresponding selected model $\hat{m}(p_0) = \text{argmin}_{m \in M} \{\gamma_n(\hat{s}_m) + \hat{\kappa}_{p_0} f(m)\}$. This is represented in the main window Figure 5.4, too, in the bottom-right corner.

- Find the most to the right "plateau". A *plateau* is a continuous sequence of values of p_0 for which $\hat{m}(p_0)$ is the same: see below for a more rigorous definition. Any value of p_0 corresponding to this plateau may be chosen as p_{lin}. We actually report the interesting result: the corresponding selected model. Moreover, the percentage of values of p_0 belonging to the selected plateau among all possible values and the range of the corresponding estimated slope values, are reported, too ("Corresponding slope interval"): there extent is a clue of the stability of the selection of the model.

The user may try an other value of p_{lin} by simply changing the value in the box "Number of dimension points used for the regression".

Now, of course, this method may be sensitive to the definition of what a *plateau* is. According to the definition above, it still remains to specify how large the sequence of values of p_0 yielding the same $\hat{m}(p_0)$ has to be to be considered a plateau. Two possibilities are left to the user:

- a plateau must be larger than *pct* percent of the total number of possible values of p_0;

- a plateau must contain a least *Nbr* different values of p_0.

By default, the first method is set, with $pct = 15$. This is rather an arbitrary choice, which should be reconsidered with respect to the application at hand. However, this choice may make sense rather generally, even in cases the user has no idea what to choose, and this is an appealing feature of this approach. Remark that whatever the choice at this step, and provided that it enables the method to be applied, the reported actual percent of values p_0 yielding the same model $\hat{m}(p_0)$ is a helpful clue about the stability of the method.

A formal writing of the selection of the model is: choose $\hat{m}$ as the least complex model in

$$\operatorname{argmax}\left\{ f(m) \,\middle|\, m \in M : \operatorname{card}\Big(\big\{p_0 \in P_0 | \hat{m}(p_0) = m\big\}\Big) \geq N \right\},$$

where N is either Nbr or $\frac{pct}{100} \times \operatorname{card}(P_0)$, according to the chosen plateau definition.

5.2.3 Options

Validation step. As already mentioned in Section 3.2.3, an important feature of the data-driven slope estimation method is that it comes with a natural and seemingly efficient method to check the validity of the assumptions underlying the slope heuristics. It consists of making use of one or several points $\big(p_v, -\gamma_n(\hat{s}_{m(p_v)})\big)$ with p_v large (according to the possibilities of the framework, it/they should be much larger than $p_{\max}$), and to check that it/they belong to the previously regressed line. This verification may be done with the package: the user merely has to specify the number of such values the provided file contains and the "Validation Step" button enables to check both by eye and by the mean of a test, whether the assessed selection may be validated or not: see Figure 5.6. If several validation points are available, a Bonferroni procedure is applied to build the test.

In case the validation step fails, two possibilities should be considered:

- It might be that not complex enough models have been involved in the study. More complex models should be added and the method applied again. See Example 8 for an illustration of such a situation.
- The values of $-\gamma_n(\hat{s}_m)$ might not be precise enough: they should perhaps be computed again.
- If this does not help, it might be that the slope heuristics is not justified in the considered framework, or that the penalty shape is not suitable.

Dimension Jump An option enables to compare the result of the data-driven slope estimation selection with the dimension jump method: see Figure 5.7. If both methods agree, the selection can be considered more confidently.

The dimension jump method is applied with the κ^{dj} definition of the estimated constant (see Section 3.2.2). Indeed, applying the κ^{thresh} definition would require to ask the user for a supplementary tuning parameter (the threshold dimension has to be chosen

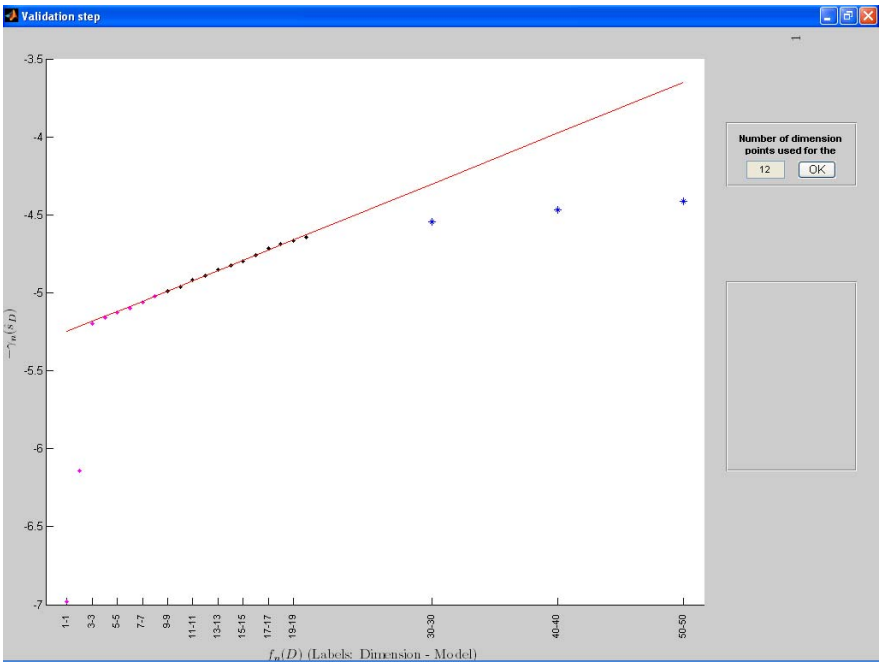


Figure 5.6: The Validation Step

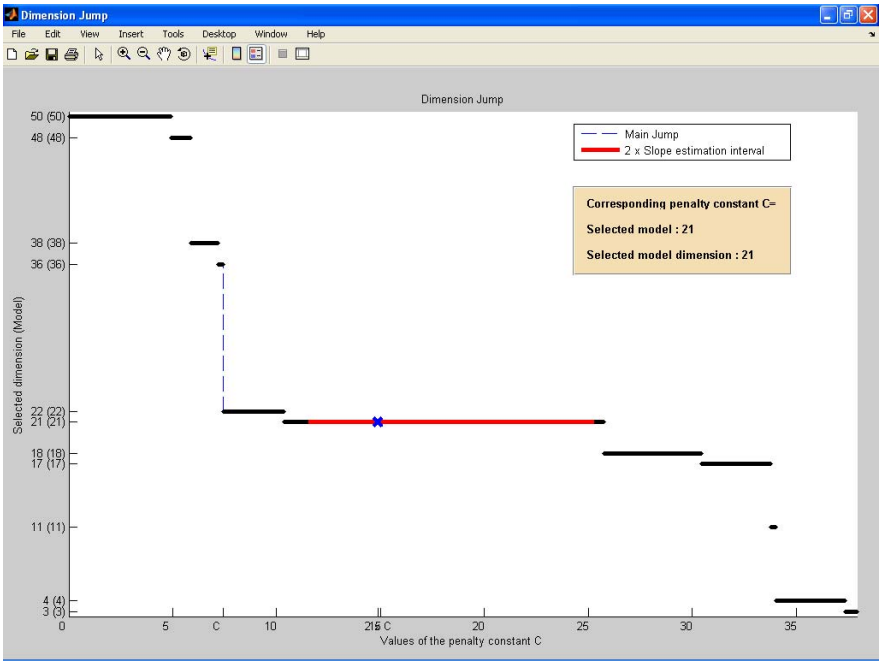


Figure 5.7: The Dimension Jump Method

with respect to the application at hand). The user should however be cautioned that the result of the dimension jump as is depends on the values of the models dimensions which have been involved in the study.

Draw your own slope! An option is offered to the user: a button enables to draw one's own linear part and to get the corresponding selected model. The interest of this option is that the user may not be confident in the regressed linear part. It may also be used as a tool to check the stability of the selection. But it should obviously be handled with care! See Figure 5.8.

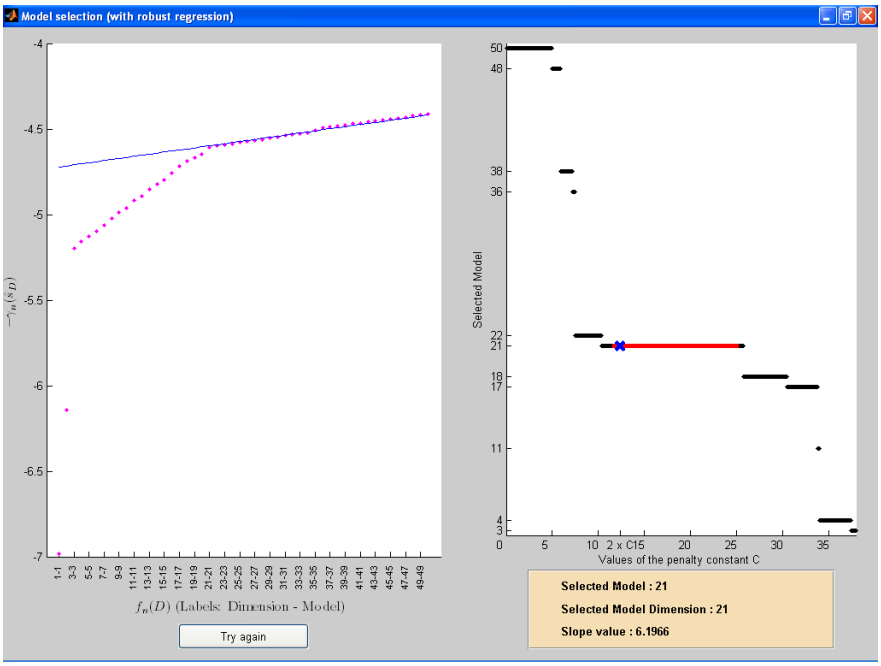


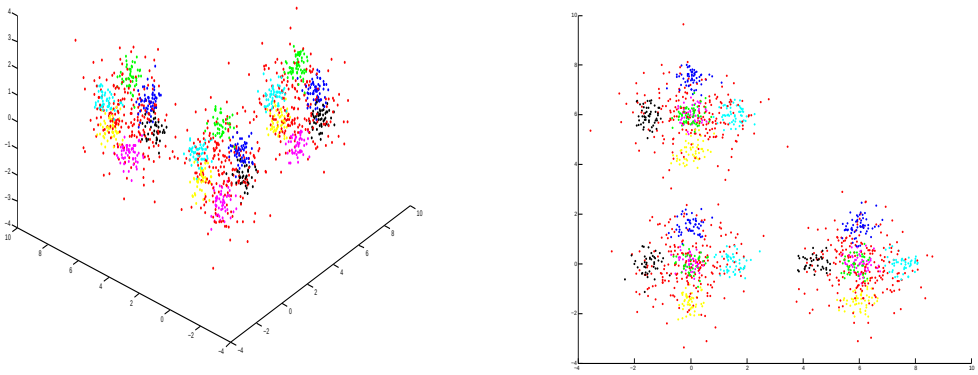
Figure 5.8: User-drawn Slope!

5.2.4 Example

Example 8 ("Bubbles") *A 21-component Gaussian mixture in* $\mathbb{R}^3$ *is simulated*[4]*. See Figure 5.9. The sample size is* 1000*. The data consists of three large "bubbles", quite far from each other. Each bubble consists of a mixture of a large component (with mixing proportion* $\frac{1}{3} \times 0.4$*) and six little bubbles around it (with mixing proportion* $\frac{1}{3} \times 0.1$ *each, and small covariance matrices). The model used for the simulation is spherical (the covariance matrices are proportional to the identity matrix).*

The fitted models $\left(\mathcal{M}_K\right)_{K\in\{1,\dots,50\}}$ *have the same form (i.e. spherical Gaussian mixture models). The problem is then quite easy and has been mainly chosen for the sake of*

[4]Details on the simulation settings and the applied algorithms may be found in Section A.2.



(a) (b)

Figure 5.9: "Bubbles" Experiment Dataset Example. (a) 3D View. (b) View From Above.

illustration. Remark however that the dimension of the largest models is not negligible with respect to the sample size (this is $(2+d)K-1$*, namely* 249 *for* $K = 50$*). Each model is fitted through the* MIXMOD *software of Biernacki et al. (2006) and with initializations as described in Sections 1.2.2 and 5.1.3 (Small_EM and Km1 are involved).*

To illustrate the interest of the validation step, let us first assume models with numbers of components smaller or equal than 20 *only are fitted. The obtained main window with the package is Figure 5.10. The estimated slope values do not seem stable, but the selected model is.*

Now, assume three models with much larger complexities are fitted and the obtained values used for validation: the package yields Figure 5.6. There is obviously a problem. Those values may be poorly estimated, but fitting all models with $21 \leq K \leq 50$ *actually confirms that the linear behavior for the most complex models was not reached yet with* $K = 20$*: see Figure 5.4. According to this figure, when* $K_M = 50$*, the slope heuristics applied with the data-driven slope estimation approach recovers the true number of components:* 21*.*

The slope heuristics applied with the dimension jump approach yields $\hat{K} = 21$*, too (see Figure 5.7).*

We conducted an experiment with 100 *such datasets. The results are summed up in Table 5.1 for comparison with classical criteria: AIC and BIC (see Section 2.1) and with the oracle. The risks of the criteria are compared in Table 5.2.*

The oracle is close to the true distribution, which is a consequence that the sample size is quite large. As a consequence, the identification and efficiency purposes almost match in this experiment. As usual in a mixture framework, AIC obviously underpenalizes the complexity of the models. BIC does a good job and mostly recovers the true number of

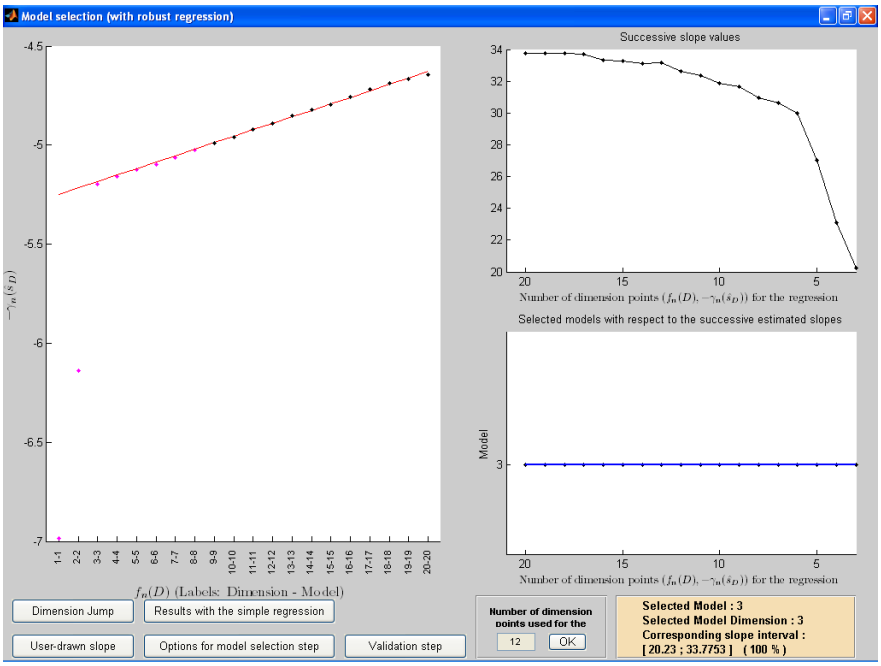


Figure 5.10: The Main Window for $K \leq 20$

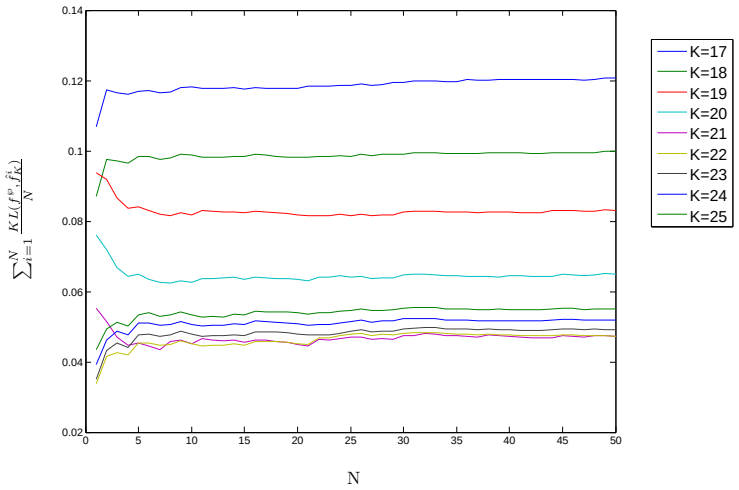


Figure 5.11: "Bubbles" Experiment. Convergence of the Monte Carlo Simulations for the Computation of $K_{\text{oracle}} = \text{argmin}_{1 \leq K \leq 50} \mathbb{E}_{f^{\wp}} \left[d_{\text{KL}}(f^{\wp}, f(\,.\,; \widehat{\theta}_K^{\text{MLE}})) \right]$.

Selected number of components	3	4	15–18	19	20	21	22	23	24	25	35–45	46	47	48	49	50
Oracle	0	0	0	0	1	76	15	3	3	2	0	0	0	0	0	0
AIC	0	0	0	0	0	0	0	0	0	0	7	3	7	9	15	59
BIC	0	0	3	6	23	57	9	1	1	0	0	0	0	0	0	0
SLH (estimation of the slope, $K_M = 50$)	0	0	0	3	7	59	20	6	3	2	0	0	0	0	0	0
SLH (estimation of the slope, $K_M = 40$)	0	0	1	3	7	61	18	4	4	2	0	0	0	0	0	0
SLH (dimension jump, $K_M = 50$)	4	2	0	3	7	59	18	2	3	2	0	0	0	0	0	0
SLH (dimension jump, $K_M = 40$)	28	2	0	2	4	51	10	2	1	0	0	0	0	0	0	0

Table 5.1: "Bubbles" Experiment Results.

	Risk of the criterion $\times 10^3$	$\frac{\text{Risk of the criterion}}{\text{Risk of the oracle}}$
Oracle	64	1
AIC	166	2.59
BIC	75	1.17
SLH (estimation of the slope, $K_M = 50$)	68	1.06
SLH (estimation of the slope, $K_M = 40$)	70	1.09
SLH (dimension jump, $K_M = 50$)	96	1.49
SLH (dimension jump, $K_M = 40$)	210	3.27

Table 5.2: "Bubbles" Experiment Results. Risk of Each Criterion in Terms of Kullback-Leibler Divergence to the True Distribution, Estimated by Monte-Carlo Simulations. The oracle results reported in the table correspond to the trajectory oracle ($K_{\text{oracle}} = \text{argmin}_{1 \leq K \leq 50} d_{\text{KL}}(f^{\wp}, \widehat{f}_K^{\text{MLE}})$ for each dataset). The expected oracle number of components ($K_{\text{oracle}} = \text{argmin}_{1 \leq K \leq 50} \mathbb{E}_{f^{\wp}} \left[d_{\text{KL}}(f^{\wp}, \widehat{f}_K^{\text{MLE}}) \right]$) is 21. The true number of components is 21.

components.

The best risk results (though close to the results of BIC), as compared to the oracle, are achieved by the slope heuristics, applied with the data-driven slope estimation approach. Those results are quite good, since the ratio of this method's risk to the oracle's is very close to 1. *The "plateaus" have been defined by the default method and value: a plateau must involve at least* 15% *of the total number of models involved.*

The dimension jump approach, applied with $K_M = 50$, yields the same selection as the data-driven slope estimation approach, but in 6% *of the datasets (Table 5.1). This is seldom but suffices to worsen sensibly its risk results because the result of those few cases are very poor (with $\hat{K} = 3$ or 4, $d_{KL}(f^{\wp}, \widehat{f}_{\hat{K}}^{MLE})$ is much worse than $d_{KL}(f^{\wp}, \widehat{f}_{K_{oracle}}^{MLE})$). The results achieved with this same approach when $K_M = 40$ are provided, too, since they illustrate a difficulty which can be encountered while applying the dimension jump. This approach leads to select $\hat{K} = 3$ for about* 30% *of the considered datasets, which is a poor result. The reason of this difficulty is illustrated in Figure 5.12: there seemingly occurs a dimension jump for the most complex models, but it occurs in several steps. Therefore the largest of those "sub-jumps" is still smaller than the jump leading to $\hat{K} = 3$, which is quite large because of the structure of the data. This illustrates the sensibility of the dimension jump approach to the choice of the most complex models involved in the study: with $K_M = 50$, this seldom occurs, but this is quite a huge number of components as compared to the interesting number of components.*

The data-driven slope estimation results are only worsened a little if $K_M = 40$ instead of $K_M = 50$ (see Table 5.1 and Table 5.2).

Conclusion

Some improvements are in progress, notably about the validation step tests. But the package is from now on functional.

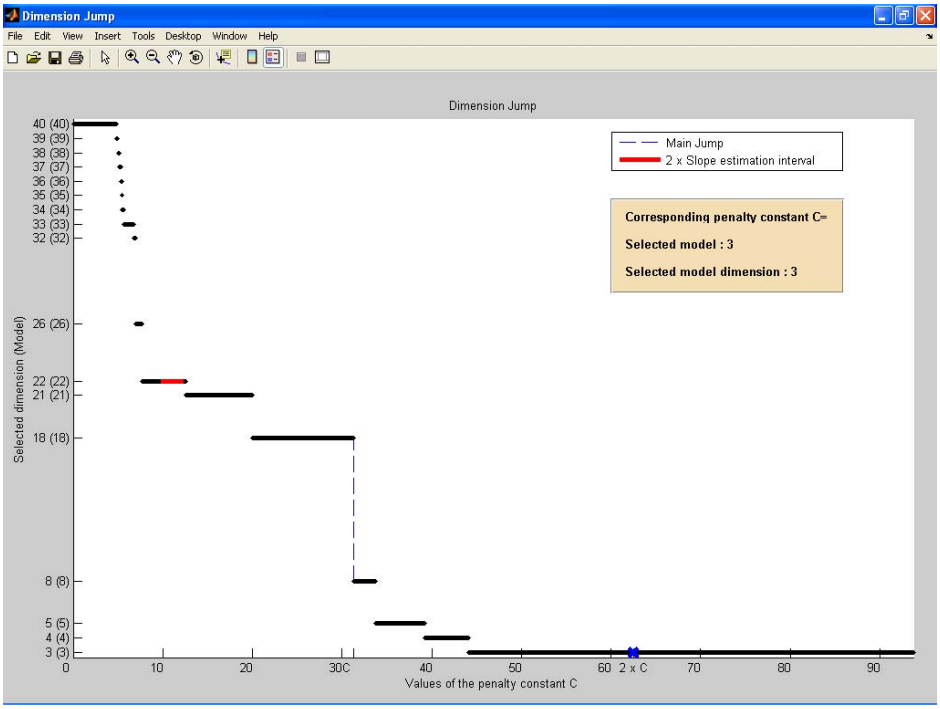


Figure 5.12: The Dimension Jump Method. An Example Where it Fails With $K_M = 40$.

It shall make the slope heuristics application easier and faster.

A matter that it might help to deal with is the comparison of the dimension jump and data-driven slope estimation approaches. A first step in this direction may be an extensive simulation study.

Chapter 6

Note on the Breakdown Point Properties of the L_{CC}-ICL Criterion

Contents

In this chapter are studied the robustness properties of the procedure defined in Chapter 4, based on the MLccE estimator and the L_{cc}-ICL model selection criterion. This notably illustrates that considering the usual ICL criterion (see for example Section 4.1.2) as an approximation of this procedure (as suggested in Section 4.4.4) helps studying its theoretical properties.

The considered robustness point of view is the *breakdown point* notion defined and studied in Hennig (2004) for a model-based clustering framework. It is a measure of how many supplementary observations have to be added to a dataset to break the solution down, in the sense that the original components cannot be recovered in the updated solution. This notion is adapted to a situation where the number of components is unknown and has to be chosen. Notably, there is not breakdown if some supplementary components are estimated but the original components can still be recovered. It is considered as a good behavior that a procedure isolates outliers by fitting supplementary components specifically to them. Hennig (2004) derives results for the usual procedure based on the MLE estimator and the BIC criterion. We derive results for the procedure based on the MLccE estimator and the L_{cc}-ICL criterion and compare those results.

The breakdown point notion is first recalled and adapted to the framework of our procedure, which is straightforward. The results of Hennig (2004) are also recalled. The results for our procedure are stated and proved in Section 6.2 and Section 6.3. They are compared to the usual procedure with examples in Section 6.4 and discussed in Section 6.5.

In this chapter, data in $\mathbb{R}$ only are considered.

6.1 Definitions, Reminder from Hennig (2004)

The subsequent breakdown definition is exactly the one proposed by Hennig (2004). We shall apply it to the estimator and model selection criterion we defined in Chapter 4.

Results are derived for the same mixture models as Hennig (2004), without having to restrict the study to Gaussian mixtures. The component densities, denoted by ϕ, are not necessarily Gaussian densities, but the same notation as introduced in Chapter 1 is used since we mainly have Gaussian mixtures in view. The following assumptions about ϕ are typically fulfilled by Gaussian densities:

$$\omega = (\mu, \sigma) \text{ and } \phi(x;\omega) = \frac{1}{\sigma}\phi\Big(\frac{x-\mu}{\sigma}\Big),$$

with $\phi : \mathbb{R} \to \mathbb{R}^{+*}$ such that

$$\begin{gathered}\phi \text{ symmetrical about zero;}\\ \phi \text{ decreasing on } [0,\infty];\\ \phi \text{ continuous.}\end{gathered}$$

Recall (from Section 1.1.1 for example) a mixture density is then written as $f(\,.\,;\theta)$, with $\theta = (\pi_1, \ldots, \pi_K, \omega_1, \ldots, \omega_k)$ $(\sum_{k=1}^{K} \pi_k = 1)$ and $\mathcal{M}_K = \{f(\,.\,;\theta) : \theta \in \Theta_K\}$. As already mentioned, there is no difficulty to define the conditional classification likelihood

if those models are considered from the parametric point of view: let $\mathbf{x}_n = (x_1, \ldots, x_n)$ be observations in $\mathbb{R}$ (which will be omitted in notation when not ambiguous or suggested by the superscript n otherwise) and define as in Chapter 4:

$$\forall \theta \in \Theta_K, \log \mathrm{L}_{\mathrm{cc}}(\theta; \mathbf{x}_n) = \log \mathrm{L}(\theta; \mathbf{x}_n) - \mathrm{ENT}(\theta; \mathbf{x}_n),$$

with

$$\log \mathrm{L}(\theta; \mathbf{x}_n) = \sum_{i=1}^{n} \log \underbrace{\sum_{k=1}^{K} \pi_k \phi(x_i; \omega_k)}_{f(x_i;\theta)}$$

and

$$\mathrm{ENT}(\theta; \mathbf{x}_n) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}(\theta) \log \tau_{ik}(\theta),$$

where

$$\tau_{ik}(\theta) = \frac{\pi_k \phi(x_i; \omega_k)}{\sum_{k'=1}^{K} \pi_{k'} \phi(x_i; \omega_{k'})}.$$

See Section 4.2 for discussion about those quantities and the choice of the conditional classification likelihood in the clustering framework.

Recall (we write $\widehat{\theta}$ instead of $\widehat{\theta}^{\mathrm{MLccE}}$ since only the MLccE will be involved in this chapter)

$$\widehat{\theta}_K^n \in \operatorname*{argmax}_{\theta \in \Theta_K} \log \mathrm{L}_{\mathrm{cc}}(\theta; \mathbf{x}_n).$$

This is well defined if the parameter space is assumed to be compact (this is discussed in Section 4.3). The only parameter bound which will be involved in the subsequent results, though, is the same as the one involved in Hennig (2004) in a usual likelihood framework:

$$\forall K, \forall \theta \in \Theta_K, \forall k \in \{1, \ldots, K\}, \sigma_k \geq \sigma_0 > 0.$$

Moreover, let us assume the set of considered models to have a bounded maximum number of components: $K \in \{1, \ldots, K_M\}$. Then, recall the L_{cc}-ICL criterion (see Section 4.4.4)

$$\mathrm{crit}_n(K) = \log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^n) - \frac{\log n}{2} D_K,$$

where D_K is the number of free parameters necessary to describe Θ_K. If the models are not constrained, this is $3K - 1$. Finally, write

$$\widehat{K}_n = \min \operatorname*{argmax}_{K \in \{1, \ldots, K_M\}} \mathrm{crit}_n(K),$$

$$\widehat{\theta}^n = \widehat{\theta}^n_{\widehat{K}_n},$$

and

$$\mathrm{crit}_n = \log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}^n) - \frac{\log n}{2} D_{\widehat{K}_n}.$$

Now, like Hennig (2004), let us define the *parameter breakdown point* of the procedure (see Hennig, 2004, Definition 3.2):

Definition 4 (Adapted from Definition 3.2 in Hennig, 2004)
Let $\mathbf{x}_n = (x_1 \ldots, x_n)$ *be a dataset. Let* $\widehat{K}_n$ *be defined as above. Then the* parameter breakdown point *of the procedure is defined as*

$$B_n(\mathbf{x}_n) = \min_g \left\{ \frac{g}{n+g} : \forall C \subset \Theta_{\widehat{K}_n} \textit{compact}, \exists x_{n+1}, \ldots, x_{n+g} \textit{ such that } \widehat{\theta}^{n+g} \textit{ fulfills: } \right.$$
$$\left. \textit{pairwise distinct } j_1, \ldots, j_{\widehat{K}_n} \textit{ do not exist such that } \left(\widehat{\omega}_{j_1}^{n+g}, \ldots, \widehat{\omega}_{j_{\widehat{K}_n}}^{n+g}\right) \in C \right\}$$

This breakdown notion is relevant when the number of components has to be chosen at the same time as the estimator. There is breakdown of the solution if g new observations can be added to the original dataset, such that, for at least one of the original components, none of the obtained components in the new solution "looks like" it. In particular, there is breakdown if $\widehat{K}_{n+g} < \widehat{K}_n$. But in case $\widehat{K}_{n+g} > \widehat{K}_n$, if $\widehat{K}_n$ of the $\widehat{\theta}^{n+g}$ solution stay close to the original components in $\widehat{\theta}^n$, the supplementary components may be anything without the original solution being "broken down". This allows the procedure to handle outliers by supplementary components designed for them specifically. The main situation of breakdown is then in this setting the addition of observations between the original ones, which would compel the procedure to select a single component where there were several in the original solution. As a matter of fact, this is the reason why considering breakdown with a compact parameter space makes sense. The breakdown point is the minimal proportion of supplementary observations necessary so that breakdown occurs. Remark that the situation where a mixing proportion goes to 0 is not considered as breakdown with this definition. Indeed this situation is not of concern since — like Hennig (2004) notices in the usual likelihood framework — our procedure shall not prefer a solution with K components, one of which has proportion very close to zero, to the corresponding $K-1$-component solution. Further discussion on this breakdown notion and several others, notably for situations where the number of components is fixed, can be found in Hennig (2004).

Let us recall the breakdown result obtained in Hennig (2004) for the usual maximum likelihood, when the number of components is selected with the BIC criterion (see Section 2.1.3 for the BIC criterion, and Section 1.2.1 for $\widehat{\theta}_K^{\mathrm{MLE}}$):

Theorem 9 (Theorem 4.13 in Hennig, 2004)
In this theorem,

$$\widehat{\theta}_K^n = \widehat{\theta}_K^{MLE}(\mathbf{x}_n)$$

and

$$\widehat{K}_n = \min \operatorname*{argmax}_{K \in \mathbb{N}^*} \ \mathrm{crit}_n^{BIC}(K)$$
$$= \min \operatorname*{argmax}_{K \in \mathbb{N}^*} \left\{ \log L(\widehat{\theta}_K^n) - \frac{\log n}{2} D_K \right\}.$$

If

$$\min_{K < \widehat{K}_n} \left[\log L(\widehat{\theta}_{\widehat{K}_n}^n) - \log L(\widehat{\theta}_K^n) - \frac{1}{2}(5g + 3\widehat{K}_n - 3K + 2n) \log(n+g) + n \log n \right] > 0,$$

then

$$B_n(\mathbf{x}_n) > \frac{g}{n+g}.$$

Remark that in this usual likelihood framework, it is not necessary to guarantee that $K \geq K_M$ since it can be proved (Lindsay, 1983) that the maximum of the likelihood itself — and *a fortiori* of the BIC criterion — is achieved for $K \leq n+1$. The number of components is however often expected to be quite smaller than the number of observations and should then be bounded anyway.

A sufficient condition to derive an upper bound on the parameter breakdown point in this usual framework can be found in Hennig (2004).

Let us now consider what can be derived in the L_{cc} framework.

6.2 Breakdown Point for Maximum Conditional Classification Likelihood

We shall prove the following theorem. The function h is defined in Lemma 10 below.

Theorem 10
$\mathbf{x}_n = (x_1, \ldots, x_n)$. *In the L_{cc} framework, with the number of components selected through L_{cc}-ICL (see Section 6.1), if*

$$\min_{K<\widehat{K}_n} \Big[\log L_{cc}(\widehat{\theta}^n_{\widehat{K}_n}) - \log L_{cc}(\widehat{\theta}^n_K) - \frac{1}{2}(5g + 3\widehat{K}_n - 3K + 2n)\log(n+g) + n\log n - g\log(\widehat{K}_n + g) - nh(g)\Big] > 0,$$

then

$$B_n(\mathbf{x}_n) > \frac{g}{n+g}.$$

This theorem provides a condition under which it is guaranteed that at least $g+1$ additional observations are necessary so that the solution of MLccE with the number of components selected through L_{cc}-ICL is broken down. The condition is seemingly stronger than the one involved in the result of Hennig (2004) for the usual likelihood with the number of components selected through BIC (Theorem 9). This is apparent from the form of the condition, and shall be highlighted through an example in Section 6.4. As a matter of fact

$$\big(\log \mathrm{L_{cc}}(\widehat{\theta}^{\mathrm{MLccE}}_{\widehat{K}_n^{\mathrm{Lcc\text{-}ICL}}}) - \log \mathrm{L_{cc}}(\widehat{\theta}^{\mathrm{MLccE}}_K)\big)$$

should mostly be of the same order as

$$\big(\log \mathrm{L}(\widehat{\theta}^{\mathrm{MLE}}_{\widehat{K}_n^{\mathrm{BIC}}}) - \log \mathrm{L}(\widehat{\theta}^{\mathrm{MLE}}_K)\big),$$

but perhaps if $\widehat{K}_n^{\mathrm{BIC}} \neq \widehat{K}_n^{\mathrm{Lcc\text{-}ICL}}$. This suggests that the solution in our framework may be less robust, from this breakdown point of view, than the usual MLE solution, but perhaps in case both procedures do not select the same number of components. This

particular situation is illustrated in Section 6.4, too. However, this result only gives insight into this since it only provides a necessary condition for breakdown, and may be pessimistic.

The following lemma will be useful to derive controls of the entropy terms along the proof of Theorem 10. There is defined the function h involved in the statement of Theorem 10.

Lemma 10
If $w_1, \cdots, w_g > 0$ are such that $\sum_{l=1}^{g} w_l \leq 1$, then:

$$\sum_{l=1}^{g} w_l \log w_l \geq -h(g) \tag{6.1}$$

with $h : g \in \mathbb{N}^* \longmapsto \begin{cases} \frac{g}{e} & \text{if } g \in \{1,2\} \\ \log g & \text{if } g \geq 3 \end{cases}$

Lemma 11 will be necessary to guarantee uniform continuity properties. The observations being fixed, it states that the density of a mixture evaluated at any of those observations is bounded away from zero, provided that it is guaranteed it is bounded away from zero for at least one observation. The important feature is that the constant does not depend on the mixture parameters.

Lemma 11
$\forall x_1, \ldots, x_n \in \mathbb{R},\ \forall \varepsilon > 0,\ \exists \eta > 0 / \forall K \leq K_M, \forall \theta \in \Theta_K$ such that $\sigma_k \geq \sigma_0$ for all k,

$$\sup_{i \in \{1,\ldots,n\}} \sum_{j=1}^{K} \pi_j \phi(x_i; \omega_j) > \varepsilon \Longrightarrow \inf_{i \in \{1,\ldots,n\}} \sum_{j=1}^{K} \pi_j \phi(x_i; \omega_j) > \eta.$$

η depends on σ_0 and K_M, but not on θ such that $\sigma_k \geq \sigma_0$ for all k.

The proofs of those lemmas are technical. Let us give the main ideas leading the proof of Theorem 10 before proving those results.

Sketch of Proof (Theorem 10)

As Hennig (2004), let us suppose there is breakdown with g supplementary observations (see Definition 4), and try to get a contradiction with the condition stated in Theorem 10. This can be done by comparing an upper and a lower bound on crit_{n+g}.

crit_{n+g} upper bound.

There is no difficulty, but technical ones. To upper-bound $\log L_{cc}(\widehat{\theta}^{n+g})$, *the contribution of the g supplementary observations can only be roughly upper-bounded by the density function maximum value. No better bound can be derived since they may be chosen such that they attract some components to them. Now, the contribution of $(x_1, \ldots, x_n)$ might not be really larger than the L_{cc} value achieved with the mixture made of the K components we are not sure whether they break down or not. The main difficulty is that*

we need uniform continuity properties. To guarantee them, we have to make sure that the K first components of the mixture do not explode as the $(\widehat{K}^{n+g} - K)$ last ones break down. As a matter of fact, this may happen, but we show that the result still holds in this case. Otherwise, Lemma 11 is the main tool in this task.

$\mathrm{crit}_{\mathbf{n+g}}$ **lower bound.**

The lower bound is easier. It is obtained by the construction of a solution with $\hat{K}_n + g$ components which L_{cc} value can be lower-bounded: starting from $\hat{\theta}^{n+g}$, g components are added, each one being centered at one of the supplementary observations and having the smallest possible variance. This is one of the two typical expected breakdown situations. The case where $\hat{K}_{n+g} < \hat{K}_n$ would be more difficult to control.

■

Let us now write this proof.

6.3 Proofs

Proof (Theorem 10)
An upper bound and a lower bound on $\mathrm{crit}_{\mathbf{n+g}}$ *are derived and compared to yield the conclusion.*

$\mathrm{crit}_{\mathbf{n+g}}$ **upper bound.**

Suppose there is breakdown for g, in the sense of Definition 4 (i.e. $B_n(\mathbf{x}_n) \leq \frac{g}{n+g}$). Then up to an index switching:

$$\forall C \subset \mathbb{R} \times [\sigma_0; \infty[\text{ compact},$$
$$\exists (x_{n+1}, \ldots, x_{n+g}), \exists K \in \left\{1, \ldots, (\widehat{K}_n - 1)\right\} \text{ s.t.}$$
$$\widehat{\omega}_j^{n+g} \notin C \text{ for } \widehat{K}_{n+g} \geq j > K.$$

This occurs in the particular case where $\widehat{K}_{n+g} < \widehat{K}_n$.

Let

$$a = 1 \wedge \frac{1}{2}\left(e^{\inf_{K < \widehat{K}_n} \log L_{cc}(\widehat{\theta}_K^n)}\right)^{\frac{1}{n}},$$

and $\tilde{a}$ such that

$$\inf_{i=1,\ldots,n} \sum_{j=1}^{K} \pi_j \phi(x_i; \omega_j) > \tilde{a} \text{ as soon as } \sup_{i=1,\ldots,n} \sum_{j=1}^{K} \pi_j \phi(x_i; \omega_j) > a,$$

whatever θ with $\sigma_k \geq \sigma_0$ and $K \leq K_M$, which is possible from Lemma 11, with $\tilde{a}$ which only depends on a, σ_0, $x_1, \ldots, x_n$ and K_M and not on θ, nor on K. Let $\varepsilon > 0$ and $b > 0$ such that

$$\forall x \leq b, \begin{cases} \log(\tilde{a} + x) & \leq \log \tilde{a} + \varepsilon, \\ \dfrac{1}{1 + \frac{x}{\tilde{a}}} & \geq 1 - \varepsilon. \end{cases}$$

Let $a = a \wedge b$. Choose C^a compact such that $\phi(x_i;\omega) < a$ for any $1 \le i \le n$, as soon as $\omega \notin C^a$. Let $x_{n+1}, \ldots, x_{n+g}$ and $K \in \{1, \ldots, (\widehat{K}_n - 1)\}$ such that $\widehat{\omega}_j^{n+g} \notin C^a$ for $\widehat{K}_{n+g} \ge j > K$. Remark that this is only possible since $\widetilde{a}$ — and then C^a — does not depend on $x_{n+1}, \ldots, x_{n+g}$.

We have to bound crit_{n+g} *from above:*

$$\mathrm{crit}_{n+g} = \sum_{i=1}^{n} \log \Big[\sum_{j=1}^{K} \widehat{\pi}_j^{n+g} \phi(x_i; \widehat{\omega}_j^{n+g}) + \sum_{j=K+1}^{\widehat{K}_{n+g}} \widehat{\pi}_j^{n+g} \phi(x_i; \widehat{\omega}_j^{n+g}) \Big] \tag{6.2a}$$

$$+ \sum_{i=n+1}^{n+g} \log \sum_{j=1}^{\widehat{K}_{n+g}} \widehat{\pi}_j^{n+g} \phi(x_i; \widehat{\omega}_j^{n+g}) \tag{6.2b}$$

$$+ \sum_{i=1}^{n+g} \sum_{j=1}^{\widehat{K}_{n+g}} \widehat{\tau}_{ij}^{n+g} \log \widehat{\tau}_{ij}^{n+g} \tag{6.2c}$$

$$- (3\widehat{K}_{n+g} - 1) \frac{\log(n+g)}{2}, \tag{6.2d}$$

with $\widehat{\tau}_{ij}^{n+g} = \tau_{ij}(\widehat{\theta}^{n+g})$.

There is no difficulty with (6.2b)*:*

$$(6.2\mathrm{b}) \le g \log f_{max} \qquad \left(f_{max} = \frac{f(0)}{\sigma_0} \right). \tag{6.3}$$

We cannot expect to derive a lower upper bound on (6.2b) *since it may definitely be that the $\widehat{\theta}^{n+g}$ solution involves g component, each one of which handles one of the supplementary observations $x_{n+1}, \ldots, x_{n+g}$ by being centered at this observation and with the smallest possible variance.*

Now, consider both situations, depending on whether $\sup_{i=1,\ldots,n} \sum_{j=1}^{K} \widehat{\pi}_j^{n+g} \phi(x_i; \widehat{\omega}_j^{n+g})$ *is smaller or larger than a, to bound* (6.2a) *and* (6.2c)*. We distinguish between both situations, so that we can use arguments relying on the uniform continuity of* $\log(1+x)$ *and* $\frac{1}{1+x}$*, in the latter case. The first case is the easiest.*

- $\sup\limits_{i=1,\ldots,n} \sum_{j=1}^{K} \widehat{\pi}_j^{n+g} \phi(x_i; \widehat{\omega}_j^{n+g}) < a.$

Then:

$$\begin{aligned}
\big((6.2\mathrm{a}) + (6.2\mathrm{c})\big) &= \sum_{i=1}^{n} \log \Bigg[\sum_{j=1}^{K} \widehat{\pi}_j^{n+g} \phi(x_i; \widehat{\omega}_j^{n+g}) \\
&\qquad + \sum_{j=K+1}^{\widehat{K}_{n+g}} \widehat{\pi}_j^{n+g} \phi(x_i; \widehat{\omega}_j^{n+g}) \Bigg] + \sum_{i=1}^{n+g} \sum_{j=1}^{\widehat{K}_{n+g}} \widehat{\tau}_{ij}^{n+g} \log \widehat{\tau}_{ij}^{n+g} \\
&\le \sum_{i=1}^{n} \log \Bigg[\sum_{j=1}^{K} \widehat{\pi}_j^{n+g} \phi(x_i; \widehat{\omega}_j^{n+g}) + \sum_{j=K+1}^{\widehat{K}_{n+g}} \widehat{\pi}_j^{n+g} \phi(x_i; \widehat{\omega}_j^{n+g}) \Bigg] \\
&< n \log(2a) \qquad (\textit{since } \widehat{\omega}_j^{n+g} \notin C^a \textit{ for } j > K) \\
&\le \log L_{cc}(\widehat{\theta}_K^n) \qquad (\textit{by definition of } a).
\end{aligned} \tag{6.4}$$

- $\sup_{i=1,\dots,n} \sum_{j=1}^{K} \widehat{\pi}_j^{n+g} \phi(x_i; \widehat{\omega}_j^{n+g}) \geq a$.

We then have, by definition of $\tilde{a}$*:*

$$\inf_{i=1,\dots,n} \sum_{j=1}^{K} \widehat{\pi}_j^{n+g} \phi(x_i; \widehat{\omega}_j^{n+g}) > \tilde{a}.$$

Let us define θ^* *as the* K*-component mixture whose components are the* K *first ones of* $\widehat{\theta}^{n+g}$ *(those we are not sure whether they are or not in* C^a*):*

$$\begin{aligned} \forall j \in \{1, \dots, K\}, \qquad \omega_j^* &= \widehat{\omega}_j^{n+g} \\ \pi_j^* &= \frac{\widehat{\pi}_j^{n+g}}{\sum_{k=1}^{K} \widehat{\pi}_k^{n+g}}. \end{aligned}$$

Then:

$$\begin{aligned} (6.2a) &\leq \sum_{i=1}^{n} \log \sum_{j=1}^{K} \widehat{\pi}_j^{n+g} \phi(x_i; \widehat{\omega}_j^{n+g}) + n\varepsilon \\ &\leq \sum_{i=1}^{n} \log \sum_{j=1}^{K} \pi_j^* \phi(x_i; \omega_j^*) + n\varepsilon \\ &= \log L(\theta^*; \mathbf{x}_n) + n\varepsilon. \end{aligned} \tag{6.5}$$

$$(6.2c) = \sum_{i=1}^{n} \sum_{j=1}^{K} \widehat{\tau}_{ij}^{n+g} \log \widehat{\tau}_{ij}^{n+g} \tag{6.2c.1}$$

$$+ \sum_{i=1}^{n} \sum_{j=K+1}^{\widehat{K}_{n+g}} \widehat{\tau}_{ij}^{n+g} \log \widehat{\tau}_{ij}^{n+g} \tag{6.2c.2}$$

$$+ \sum_{i=n+1}^{n+g} \sum_{j=1}^{\widehat{K}_{n+g}} \widehat{\tau}_{ij}^{n+g} \log \widehat{\tau}_{ij}^{n+g}. \tag{6.2c.3}$$

For $1 \leq i \leq n$ *and* $1 \leq j \leq K$*:*

$$\begin{aligned} \widehat{\tau}_{ij}^{n+g} &= \frac{\widehat{\pi}_j^{n+g} \phi(x_i; \widehat{\omega}_j^{n+g})}{\sum_{l=1}^{K} \widehat{\pi}_l^{n+g} \phi(x_i; \widehat{\omega}_l^{n+g}) + \sum_{l=K+1}^{\widehat{K}_{n+g}} \widehat{\pi}_l^{n+g} \phi(x_i; \widehat{\omega}_l^{n+g})} \\ &\leq \frac{\frac{\widehat{\pi}_j^{n+g}}{\sum_{m=1}^{K} \widehat{\pi}_m^{n+g}} \phi(x_i; \widehat{\omega}_j^{n+g})}{\sum_{l=1}^{K} \frac{\widehat{\pi}_l^{n+g}}{\sum_{m=1}^{K} \widehat{\pi}_m^{n+g}} \phi(x_i; \widehat{\omega}_l^{n+g})} = \tau_{ij}^* \\ \widehat{\tau}_{ij}^{n+g} &= \frac{\widehat{\pi}_j^{n+g} \phi(x_i; \widehat{\omega}_j^{n+g})}{\sum_{l=1}^{K} \widehat{\pi}_l^{n+g} \phi(x_i; \widehat{\omega}_l^{n+g})} \times \frac{1}{1 + \frac{\sum_{l=K+1}^{\widehat{K}_{n+g}} \widehat{\pi}_l^{n+g} \phi(x_i; \widehat{\omega}_l^{n+g})}{\sum_{l=1}^{K} \widehat{\pi}_l^{n+g} \phi(x_i; \widehat{\omega}_l^{n+g})}} \\ &\geq \frac{\widehat{\pi}_j^{n+g} \phi(x_i; \widehat{\omega}_j^{n+g})}{\sum_{l=1}^{K} \widehat{\pi}_l^{n+g} \phi(x_i; \widehat{\omega}_l^{n+g})} \times \frac{1}{1 + \frac{b}{\tilde{a}}} \\ &\geq \tau_{ij}^* \times (1 - \varepsilon) \end{aligned}$$

Then

$$\begin{aligned}\widehat{\tau}_{ij}^{n+g}\log\widehat{\tau}_{ij}^{n+g} &\leq \widehat{\tau}_{ij}^{n+g}\log\tau_{ij}^{*}\\ &\leq \tau_{ij}^{*}\log\tau_{ij}^{*}-\varepsilon\tau_{ij}^{*}\log\tau_{ij}^{*}\end{aligned}$$

And, since $-\sum_{j=1}^{K}\tau_{ij}^{*}\log\tau_{ij}^{*}\leq\log K$ *for any* i*:*

$$(6.2\text{c}.1)\leq\sum_{i=1}^{n}\sum_{j=1}^{K}\tau_{ij}^{*}\log\tau_{ij}^{*}+\varepsilon\, n\log K$$

Moreover, $(6.2\text{c}.2)\leq 0$ *and* $(6.2\text{c}.3)\leq 0$*.*

Then, together with (6.5)*:*

$$\begin{aligned}(6.2\text{a})+(6.2\text{c}) &\leq \underbrace{\log L(\theta^{*};\mathbf{x}_n)+\sum_{i=1}^{n}\sum_{j=1}^{K}\tau_{ij}^{*}\log\tau_{ij}^{*}}_{\log L_{cc}(\theta^{*};\mathbf{x}_n)}+(n+n\log K)\varepsilon\\ &\leq \log L_{cc}(\widehat{\theta}_K^n)+n\big(1+\log K_M\big)\varepsilon.\end{aligned}\tag{6.7}$$

And finally, from (6.2), (6.3), (6.4) *and* (6.7)*:*

$$\text{crit}_{n+g}\leq\log L_{cc}(\widehat{\theta}_K^n)+g\log f_{max}-(3\widehat{K}_{n+g}-1)\frac{\log(n+g)}{2}+\kappa\varepsilon\tag{6.8}$$

holds in any case, with $\kappa>0$ *depending only on* n *and* K_M*.*

$\text{crit}_{\mathbf{n+g}}$ **lower bound.**

Define now $\widetilde{\theta}$ *as the following* $\widetilde{K}$*-component* ($\widetilde{K}=\widehat{K}_n+g$) *mixture:*

$$\begin{aligned}\forall j\in\{1,\dots,\widehat{K}_n\},\quad \widetilde{\pi}_j &= \frac{n}{n+g}\widehat{\pi}_j^n\\ \widetilde{\theta}_j &= \widehat{\theta}_j^n\\ \forall j\in\{\widehat{K}_n+1,\dots,\widehat{K}_n+g\},\quad \widetilde{\pi}_j &= \frac{1}{n+g}\\ \widetilde{\theta}_j &= (x_{n+j-\widehat{K}_n},\sigma_0).\end{aligned}$$

$$\begin{aligned}\text{crit}_{n+g} &\geq \text{crit}_{n+g}(\widehat{K}_n+g)\\ &\geq \log L(\widetilde{\theta};\mathbf{x}_{n+g}) &(6.9\text{a})\\ &\quad -\text{ENT}(\widetilde{\theta};\mathbf{x}_{n+g}) &(6.9\text{b})\\ &\quad -\big(3(\widehat{K}_n+g)-1\big)\frac{\log(n+g)}{2}. &(6.9\text{c})\end{aligned}$$

$$(6.9\text{a})=\sum_{i=1}^{n}\log\sum_{j=1}^{\widehat{K}_n+g}\widetilde{\pi}_j\phi(x_i;\widetilde{\omega}_j)\tag{6.10a}$$

$$+\sum_{i=n+1}^{n+g}\log\sum_{j=1}^{\widehat{K}_n+g}\widetilde{\pi}_j\phi(x_i;\widetilde{\omega}_j)\tag{6.10b}$$

$$(6.9\text{b}) = \sum_{i=1}^{n}\sum_{j=1}^{\widehat{K}_n} \widetilde{\tau}_{ij}\log\widetilde{\tau}_{ij} \tag{6.11a}$$

$$+ \sum_{i=1}^{n}\sum_{j=\widehat{K}_n+1}^{\widehat{K}_n+g} \widetilde{\tau}_{ij}\log\widetilde{\tau}_{ij} \tag{6.11b}$$

$$+ \sum_{i=n+1}^{n+g}\sum_{j=1}^{\widehat{K}_n+g} \widetilde{\tau}_{ij}\log\widetilde{\tau}_{ij}. \tag{6.11c}$$

For $1 \leq i \leq n$ *and* $1 \leq j \leq \widehat{K}_n$*:*

$$\begin{aligned}\widetilde{\tau}_{ij} &= \frac{\frac{n}{n+g}\widehat{\pi}_j^n\phi(x_i;\widehat{\omega}_j^n)}{\frac{n}{n+g}\sum_{l=1}^{\widehat{K}_n}\widehat{\pi}_l^n\phi(x_i;\widehat{\omega}_l^n) + \frac{1}{n+g}\sum_{l=n+1}^{n+g}\phi\big(x_i;(x_l,\sigma_0)\big)} \\ &= \widehat{\tau}_{ij}^n\frac{\sum_{l=1}^{\widehat{K}_n}\widehat{\pi}_l^n\phi(x_i;\widehat{\omega}_l^n)}{\sum_{l=1}^{\widehat{K}_n}\widehat{\pi}_l^n\phi(x_i;\widehat{\omega}_l^n) + \frac{1}{n}\sum_{l=n+1}^{n+g}\phi\big(x_i;(x_l,\sigma_0)\big)} \\ &\leq \widehat{\tau}_{ij}^n.\end{aligned}$$

For $1 \leq i \leq n$ *and* $\widehat{K}_n + 1 \leq j \leq \widehat{K}_n + g$*:*

$$\widetilde{\tau}_{ij} = \frac{\frac{1}{n}\phi\big(x_i;(x_{n+j-\widehat{K}_n},\sigma_0)\big)}{\sum_{l=1}^{\widehat{K}_n+g}\widetilde{\pi}_l\phi(x_i;\widetilde{\omega}_l)}.$$

For $n+1 \leq i \leq n+g$ *and* $1 \leq j \leq \widehat{K}_n + g$*:*

$$\widetilde{\tau}_{ij} = \frac{\widetilde{\pi}_j\phi(x_i;\widetilde{\omega}_j)}{\sum_{l=1}^{\widehat{K}_n+g}\widetilde{\pi}_l\phi(x_i;\widetilde{\omega}_l)}.$$

And then:

$$\begin{aligned}(6.11\text{a}) &\geq \sum_{i=1}^{n}\sum_{j=1}^{\widehat{K}_n}\widehat{\tau}_{ij}^n\log\widehat{\tau}_{ij}^n + \sum_{i=1}^{n}\overbrace{\sum_{j=1}^{\widehat{K}_n}\widehat{\tau}_{ij}^n}^{\leq 1}\log\frac{\overbrace{\sum_{l=1}^{\widehat{K}_n}\widehat{\pi}_l^n\phi(x_i;\widehat{\omega}_l^n)}^{\leq 1}}{\frac{n+g}{n}\sum_{l=1}^{\widehat{K}_n+g}\widetilde{\pi}_l\phi(x_i;\widetilde{\omega}_l)} \\ &\geq -\operatorname{ENT}(\widehat{\theta}^n) + \sum_{i=1}^{n}\log\frac{\sum_{l=1}^{\widehat{K}_n}\widehat{\pi}_l^n\phi(x_i;\widehat{\omega}_l^n)}{\frac{n+g}{n}\sum_{l=1}^{\widehat{K}_n+g}\widetilde{\pi}_l\phi(x_i;\widetilde{\omega}_l)}.\end{aligned}$$

From lemma 10:

$$(6.11\text{b}) \geq -nh(g)$$

Now,

$$(6.10\text{a}) + (6.11\text{a}) \geq -\operatorname{ENT}(\widehat{\theta}^n) + \log L(\widehat{\theta}^n) - n\log\frac{n+g}{n}.$$

And

$$(6.10\text{b}) + (6.11\text{c}) \geq g \log \frac{f_{max}}{n+g} + g \log \frac{1}{\widehat{K}_n + g}\cdot$$

Finally:

$$\text{crit}_{n+g} \geq \log L_{cc}(\widehat{\theta}^n) + n \log \frac{n}{n+g} + g \log \frac{1}{\widehat{K}_n + g} + g \log \frac{f_{max}}{n+g} - nh(g) - \frac{1}{2}\big(3(\widehat{K}_n + g) - 1\big) \log(n+g). \quad (6.12)$$

Conclusion.

From (6.8) *and* (6.12)*:*

$$\log L_{cc}(\widehat{\theta}^n) - \log L_{cc}(\widehat{\theta}^n_K) - \big(n + \frac{5}{2}g + \frac{3}{2}\widehat{K}_n - \frac{3}{2}\widehat{K}_{n+g}\big) \log(n+g) + n \log n - g \log(\widehat{K}_n + g) - nh(g) - \kappa\varepsilon < 0,$$

and

$$\log L_{cc}(\widehat{\theta}^n) - \log L_{cc}(\widehat{\theta}^n_K) - \frac{1}{2}\big(5g + 3\widehat{K}_n - 3K + 2n\big) \log(n+g) + n \log n - g \log(\widehat{K}_n + g) - nh(g) - \kappa\varepsilon < 0.$$

Since this holds for every $\varepsilon > 0$ for a good choice of $x_{n+1}, \cdots, x_{n+g}$, we get a contradiction with the condition of the theorem and there cannot be breakdown with g additional observations. ■

Proof (Lemma 10) *Let, for every $l \in \{1, \cdots, g\}$, $\tilde{w}_l = \frac{w_l}{\sum_{j=1}^g w_j}$. Then*

$$\sum_{l=1}^g \tilde{w}_l = 1 \Rightarrow \sum_{l=1}^g \tilde{w}_l \log \tilde{w}_l \geq -\log g \qquad \textit{(see for example Section 4.2.2)}$$
$$\Rightarrow \sum_{l=1}^g w_l \log w_l - \sum_{l=1}^g w_l \log \sum_{j=1}^g w_j \geq -\sum_{l=1}^g w_l \log g$$
$$\Rightarrow \sum_{l=1}^g w_l \log w_l \geq \sum_{l=1}^g w_l \log \frac{\sum_{l=1}^g w_l}{g}$$

Now, if $x \leq 1$ and $g \geq 3$,

$$\frac{x}{g} \log \frac{x}{g} \geq \frac{1}{g} \log \frac{1}{g},$$

since $\frac{x}{g} \leq \frac{1}{g} \leq \frac{1}{e}$ and $x \mapsto x \log x$ is decreasing on $]0; \frac{1}{e}]$. Applied to $x = \sum_{l=1}^g \omega_l$ and if $g \geq 3$, this yields $\sum_{l=1}^g w_l \log w_l \geq -\log g$. This inequality is tight (the equality is achieved in the case $w_l = \frac{1}{g}$ for all l).

If $g \in \{1, 2\}$, let us write: $\forall x \in [0, 1],\ x \log x \geq -\frac{1}{e}$ and then, $\sum_{l=1}^g w_l \log w_l \geq -\frac{g}{e}$, which is the best inequality we can get in those cases, since the equality can be achieved, with $w_l = \frac{1}{e}$ for all l (which is only possible if $g \leq 2$). ■

Proof (Lemma 11) *Let $\varepsilon > 0$. Let K, θ be such that the corresponding mixture verifies the lemma condition. Then*

$$\begin{aligned}
\varepsilon &< \sup_i \sum_{j=1}^{K} \pi_j \phi(x_i; \omega_j) \\
&= \sum_{j=1}^{K} \pi_j \phi(x_{i^M}; \omega_j) \\
&\leq K \sup_j \pi_j \phi(x_{i^M}; \omega_j) \\
&= K \pi_{j^M} \phi(x_{i^M}; \omega_{j^M}).
\end{aligned}$$

Write $\inf_i \sum_{j=1}^{K} \pi_j \phi(x_i; \omega_j) = \sum_{j=1}^{K} \pi_j \phi(x_{i^m}; \omega_j)$. We have

$$\begin{aligned}
\forall j, \ \frac{\phi(x_{i^m}; \omega_j)}{\phi(x_{i^M}; \omega_j)} &= e^{-\frac{1}{\sigma_j^2}\left(x_{i^m}^2 - x_{i^M}^2 - 2a_j(x_{i^m} - x_{i^M})\right)} \\
&\geq e^{-\frac{1}{\sigma_j^2}(d_2 + 2a_j d)},
\end{aligned} \tag{6.13}$$

with $d = \max_i x_i - \min_i x_i$ and $d_2 = \max_{i,j}(x_i^2 - x_j^2)$.

Moreover,

$$\phi(x_{i^M}; \omega_{j^M}) \geq \pi_{j^M} \phi(x_{i^M}; \omega_{j^M}) > \frac{\varepsilon}{K} \Longrightarrow \frac{1}{\sqrt{2\pi\sigma_{j^M}^2}} e^{-\frac{(x_{i^M} - a_{j^M})^2}{\sigma_{j^M}^2}} > \frac{\varepsilon}{K},$$

from which on the one hand $\sigma_{j^M}^2 < \frac{K^2}{2\pi\varepsilon^2}$, and then on the other hand, since moreover, $\sigma_{j^M}^2 > \sigma_0^2$,

$$\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{2\pi}{K^2}\varepsilon^2 (x_{i^M} - a_{j^M})^2} > \frac{\varepsilon}{K}.$$

This implies that $\exists A(x_1 \ldots, x_n, \varepsilon, \sigma_0, K_M)/|a_{j^M}| \leq A$.

And then, from (6.13),

$$\begin{aligned}
\sum_{j=1}^{K} \pi_j \phi(x_{i^m}; \omega_j) &\geq \pi_{j^M} \phi(x_{i^m}; \omega_{j^M}) \\
&\geq \pi_{j^M} \phi(x_{i^M}; \omega_{j^M}) e^{-\frac{1}{\sigma_0^2}(d_2 + 2a_{j^M} d)} \\
&> \frac{\varepsilon}{K} e^{-\frac{1}{\sigma_0^2}(d_2 + 2a_{j^M} d)} \\
&\geq \frac{\varepsilon}{K_M} e^{-\frac{1}{\sigma_0^2}(d_2 + 2Ad)} = \eta(x_1, \ldots, x_n, \varepsilon, \sigma_0, K_M)
\end{aligned}$$

which is the expected conclusion. ■

6.4 Examples

Let us consider the same type of datasets as Hennig (2004). The base of those datasets is not random so that it can be repeated in order to compare several procedures. Hennig (2004) denote by (μ, σ^2)-*Normal standard dataset* (NSD) the set of quantiles of respective levels $\frac{1}{n+1}, \ldots, \frac{n}{n+1}$ of a Gaussian distribution with parameters (μ, σ^2). Then, the considered dataset is of size 50 and consists of the combination of a $(0, 1)$-NSD with 25 points and a $(\mu^*, 1)$-NSD with 25 points. In Hennig (2004) $\mu^* = 5$. See Figure 6.1(a) for an example of such a dataset.

In Hennig (2004, Example 4.14), such a dataset with $\mu^* = 5$ is considered: for the MLE and when the number of components is chosen through BIC ($\widehat{K}_n^{\text{BIC}} = 2$), the breakdown point is proved to be larger than $\frac{1}{51}$ since the condition of Theorem 9 is fulfilled for $g = 1$. But it is not fulfilled for $g = 2$. However, Hennig (2004) found empirically that thirteen points are necessary to actually break the solution down, which suggests that the condition is too conservative.

In our framework, when fitting models by MLccE and selecting the number of components through L_{cc}-ICL, the condition in Theorem 10 is not fulfilled even for $g = 1$. As expected, $\hat{K}^{\text{L}_{\text{cc}}-\text{ICL}} = 2$ with reasonable σ_0 (see for example Section 5.1.4 for the choice of σ_0). The typical value of σ_0 we involved is smaller than the one chosen by Hennig (2004) so that those results still hold for this value. The minimal value of μ^* for which the condition is fulfilled for $g = 1$ is around 7.5 (see Figure 6.1(b)). For such a value, the condition in Theorem 9 is fulfilled up to $g = 3$. See Figure 6.2 for the common solution of BIC and L_{cc}-ICL.

This experiment illustrates that the condition in Theorem 10 is actually quite stronger than the condition in Theorem 9. This suggests, as already mentioned, that the procedure relying on the MLccE and on L_{cc}-ICL might be less robust than the procedure based on the MLE and BIC. However, it has been mentioned also that this might fail to be true in case $\widehat{K}^{\text{BIC}} \neq \widehat{K}^{\text{L}_{\text{cc}}-\text{ICL}}$. Let us consider such a situation.

If the components are made closer to each other, L_{cc}-ICL is the first of both criteria to select a single component instead of two, because it favors well separated components. To make the example richer, let us consider a situation where three components are selected by BIC: this is the dataset (c) in Figure 6.1. BIC based on the MLE selects three components (see Figure 6.3(a)) and the condition of Theorem 9 fails for $K = 2$. But L_{cc}-ICL based on MLccE selects only two components (see Figure 6.3(b)) and the condition of Theorem 10 is fulfilled. Actually, no theorem is necessary in this case to guess that the solution of L_{cc}-ICL is more robust than that of BIC. This is apparent from the solutions (see Figures 6.3(a) and 6.3(b)): it is necessary to roughly "fill in" — in the sense of MLccE — the space between the left (centered at -1.75) and the right (centered at 9) classes (Figure 6.3(b)) to break the solution of L_{cc}-ICL down, while it suffices to "fill in" — in the sense of MLE — the space between the two left components (namely respectively centered at -3.5 and 0: see Figure 6.3(a)) to break the solution of BIC down, which presumably requires less supplementary observations. This is confirmed by Figure 6.4: recall that the conditions in Theorems 9 and 10 essentially depend on the difference of the values of each contrast between the selected model and models with lower numbers of components. From this point of view, Figure 6.4 illustrates that the L_{cc}-ICL solution is expected to be more robust than that of BIC in this setting.

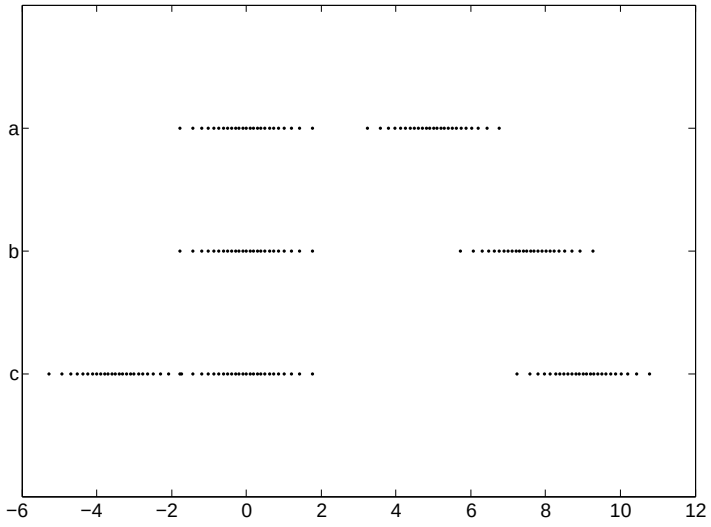


Figure 6.1: NSD-based datasets.

(a) $(0, 1)$-NSD $(n = 25)$ + $(5, 1)$-NSD $(n = 25)$,
(b) $(0, 1)$-NSD $(n = 25)$ + $(7.5, 1)$-NSD $(n = 25)$,
(c) $(-3.5, 1)$-NSD $(n = 25)$ + $(0, 1)$-NSD $(n = 25)$ + $(9, 1)$-NSD $(n = 25)$.

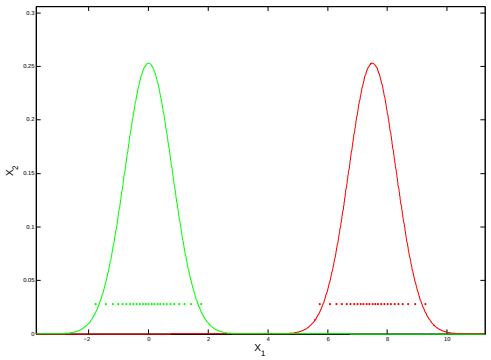
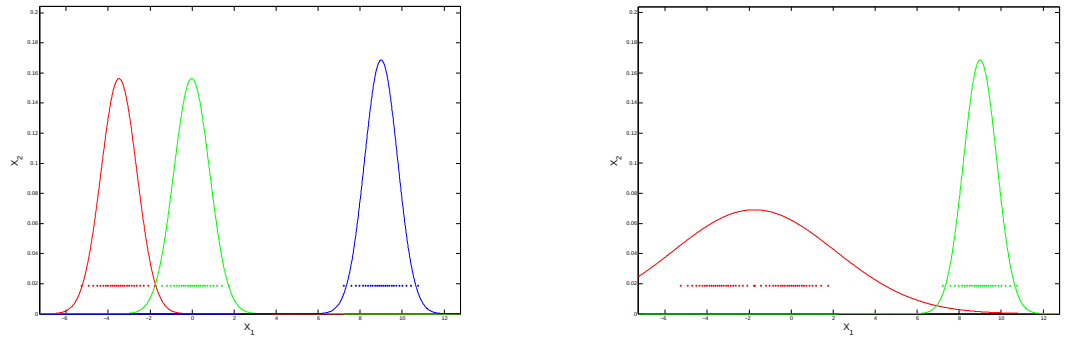


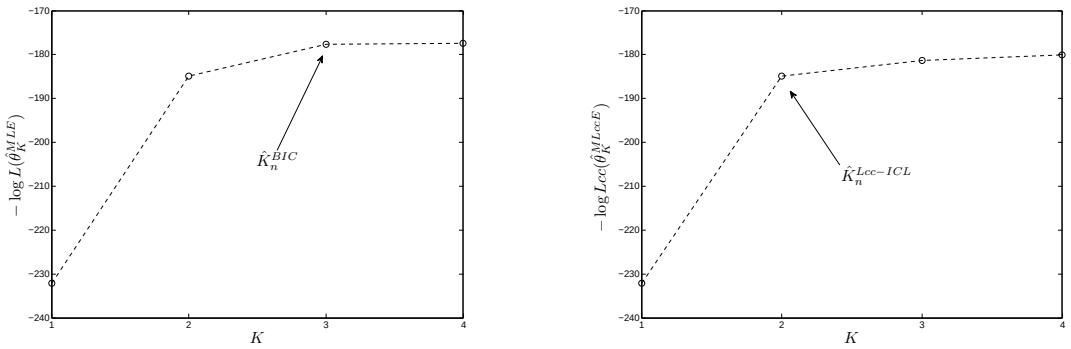
Figure 6.2: Common solution of BIC and L_{cc}-ICL for the $(0, 1)$-NSD $(n = 25)$ + $(7.5, 1)$-NSD $(n = 25)$ Dataset.



(a) (b)

Figure 6.3: Solutions for the $(-3.5, 1)$-NSD $(n = 25)$ + $(0, 1)$-NSD $(n = 25)$ + $(9, 1)$-NSD $(n = 25)$ Dataset.

(a) MLE solution selected through BIC,
(b) MLccE solution selected through $\mathrm{L_{cc}}$-ICL.



(a) (b)

Figure 6.4: Maximum Contrast Values vs K for the $(-3.5, 1)$-NSD $(n = 25)$ + $(0, 1)$-NSD $(n = 25)$ + $(9, 1)$-NSD $(n = 25)$ Dataset.

(a) $K \mapsto \log \mathrm{L}(\hat{\theta}_K^{\mathrm{MLE}})$ graph,
(b) $K \mapsto \log \mathrm{L_{cc}}(\hat{\theta}_K^{\mathrm{MLccE}})$ graph.

6.5 Discussion

The theoretical results reported in Hennig (2004) for BIC and the analogous results reported here for L_{cc}-ICL seem to be quite conservative, in the sense that the conditions they respectively impose seemingly suffice to guarantee more robustness than they claim. This might be a consequence of the difficulty to take into account in the calculations the most realistic situation of breakdown when the number of components is unknown, i.e. that where components vanish because of observations added between the original ones.

However, those results provide an interesting basis to compare the robustness properties of both procedures in different situations. They notably enable to expect that the BIC procedure is presumably more robust than the L_{cc}-ICL procedure in case both procedures select the same number of components. A dataset modelled by two components through L_{cc}-ICL and BIC can be made to be modelled by a single component through L_{cc}-ICL easier than through BIC because MLccE and L_{cc}-ICL favor well separated components. There is however a situation where L_{cc}-ICL may be more robust than BIC: when this is not clear with the original dataset whether one or two components should be selected, BIC might choose two while L_{cc}-ICL chooses one. This is a situation where breaking the BIC solution down is quite easy and where the L_{cc}-ICL solution is much more robust.

The theoretical results do not seem to be of direct practical interest but, as in Hennig (2004) for the comparison of several solutions to make the mixture model estimation more robust, they enable to improve the study and understanding of the procedures' respective and relative behaviors. Moreover they involve theoretical studies which are interesting for their own, such as Lemma 11.

Chapter 7

BIC/ICL: Combining Mixture Components to Achieve the Best of Both Worlds

Contents

Presentation. Chapter 4 largely illustrates to what extent the density estimation and clustering purposes are different, and may be contradictory. Therefore, performing model-based clustering (Chapter 1) involves a severe dilemma, notably about the choice of the number of components: should the study be based on a mixture which provides a precise estimation of the data distribution? The use of Gaussian mixture models is justified by their nice approximation properties. The BIC criterion (Section 2.1.3) is helpful to this aim. Or should the number of classes be considered as the main relevant quantity, hence the number of components be chosen with regards to this purpose? This is the point of view underlying ICL (Section 2.1.4), which is a good criterion for this purpose. This dilemma may be overcome by breaking the "one component=one class" rule. See Section 2.2 for a first introduction of this idea and further discussion and references. A methodology to do so is introduced in this chapter, which consists of an article (to appear) corresponding to a joint work with A. Raftery, G. Celeux, K. Lo and R. Gottardo Baudry et al., 2008b. The notation has been changed to be consistent with the preceding chapters, and a little more material has been included. But notably the presentation of ICL has not been changed so that a comparison of this chapter with the preceding ones — particularly with Chapter 4 — illustrates the two different points of view about ICL. In this work, we propose to first estimate the data density with the BIC solution and to hierarchically merge some of the obtained components to design classes which match a notion of cluster related to the entropy. Presumably the user may mostly be interested in the whole obtained hierarchy, or at least in several solutions, which may be compared on substantive ground. Graphical tools are provided to help analyzing this hierarchy and identifying the interesting numbers of classes. An automatic way of selecting the number of classes when it is necessary is derived from these tools.

7.1 Introduction

Model-based clustering is based on a finite mixture of distributions, in which each mixture component is taken to correspond to a different group, cluster or subpopulation. For continuous data, the most common component distribution is a multivariate Gaussian (or normal) distribution. A standard methodology for model-based clustering consists of using the EM algorithm to estimate the finite mixture models corresponding to each number of clusters considered and using BIC to select the number of mixture components, taken to be equal to the number of clusters (Fraley and Raftery, 1998). The clustering is then done by assigning each observation to the cluster to which it is most likely to belong *a posteriori*, conditionally on the selected model and its estimated parameters. For reviews of model-based clustering, see McLachlan and Peel (2000) and Fraley and Raftery (2002).

Biernacki et al. (2000) argued that the goal of clustering is not the same as that of estimating the best approximating mixture model, and so BIC may not be the best way of determining the number of clusters, even though it does perform well in selecting the number of components in a mixture model. Instead they proposed the ICL criterion, whose purpose is to assess the number of mixture components that leads to the best clustering. This turns out to be equivalent to BIC penalized by the entropy of the corresponding clustering.

We argue here that the goal of selecting the number of mixture components for

estimating the underlying probability density is well met by BIC, but that the goal of selecting the number of *clusters* may not be. Even when a multivariate Gaussian mixture model is used for clustering, the number of mixture components is not necessarily the same as the number of clusters. This is because a cluster may be better represented by a mixture of normals than by a single normal distribution.

We propose a method for combining the points of view underlying BIC and ICL to achieve the best of both worlds. BIC is used to select the number of components in the mixture model. We then propose a sequence of possible solutions by hierarchical combination of the components identified by BIC. The decision about which components to combine is based on the same entropy criterion that ICL implicitly uses. In this way, we propose a way of interpreting the mixture model in terms of clustering by identifying a subset of the mixture components with each cluster. We suggest assessing all the resulting clusterings substantively. We also describe an automatic method for choosing the number of clusters based on a piecewise linear regression fit to the rescaled entropy plot. The number of clusters selected, either substantively or automatically, can be different from the number of components chosen with BIC.

Often the number of clusters identified by ICL is smaller than the number of components selected by BIC, raising the question of whether BIC tends to overestimate the number of groups. On the other hand, in almost all simulations based on assumed true mixture models, the number of components selected by BIC does not overestimate the true number of components (Biernacki et al., 2000; McLachlan and Peel, 2000; Steele, 2002). Our approach resolves this apparent paradox.

In Section 7.2 we provide background on model-based clustering, BIC and ICL, and in Section 7.3 we describe our proposed methodology. In Section 7.4 we give results for simulated data, and in Section 7.5 we give results from the analysis of a flow cytometry dataset. There, one of the sequence of solutions from our method is clearly indicated substantively, and seems better than either the original BIC or ICL solutions. In Section 7.6 we discuss issues relevant to our method and other methods that have been proposed.

7.2 Model Selection in Model-Based Clustering

Model-based clustering assumes that observations $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ in $\mathbf{R}^{nd}$ are a sample from a finite mixture density

$$p(\mathbf{x}_i \mid K, \theta_K) = \sum_{k=1}^{K} \pi_k \phi(\mathbf{x}_i \mid \omega_k), \tag{7.1}$$

where the π_k's are the mixing proportions ($0 < \pi_k < 1$ for all $k = 1, \ldots, K$ and $\sum_k \pi_k = 1$), $\phi(. \mid \omega_k)$ denotes a parameterized density, and $\theta_K = (\pi_1, \ldots, \pi_{K-1}, \omega_1, \ldots, \omega_K)$. When the data are multivariate continuous observations, the component density is usually the d-dimensional Gaussian density with parameter $\omega_k = (\mu_k, \Sigma_k)$, μ_k being the mean and Σ_k the variance matrix of component k.

For estimation purposes, the mixture model is often expressed in terms of complete data, including the groups to which the data points belong. The complete data are

$$\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) = ((\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n)),$$

where the missing data are $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, with $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ being binary vectors such that $z_{ik} = 1$ if $\mathbf{x}_i$ arises from group k. The $\mathbf{z}_i$'s define a partition $P = (P_1, \ldots, P_K)$ of the observed data $\mathbf{x}$ with $P_k = \{\mathbf{x}_i \text{ such that } z_{ik} = 1\}$.

From a Bayesian perspective, the selection of a mixture model can be based on the integrated likelihood of the mixture model with K components Kass and Raftery (1995), namely

$$p(\mathbf{x}|K) = \int p(\mathbf{x}|K, \theta_K)\pi(\theta_K)d\theta_K, \tag{7.2}$$

where $\pi(\theta_K)$ is the prior distribution of the parameter θ_K. Here we use the BIC approximation of Schwarz (1978) to the log integrated likelihood, namely

$$\text{BIC}(K) = \log p(\mathbf{x}|K, \hat{\theta}_K) - \frac{D_K}{2}\log n, \tag{7.3}$$

where $\hat{\theta}_K$ is the maximum likelihood estimate of θ_K and D_K is the number of free parameters of the model with K components. This was first applied to model-based clustering by Dasgupta and Raftery (1998). Keribin (2000) has shown that under certain regularity conditions the BIC consistently estimates the number of mixture components, and numerical experiments show that the BIC works well at a practical level (Fraley and Raftery, 1998; Biernacki et al., 2000; Steele, 2002). See Section 2.1.3 for further discussion and references about BIC.

There is one problem with using this solution directly for clustering. Doing so is reasonable if each mixture component corresponds to a separate cluster, but this may not be the case. In particular, a cluster may be both cohesive and well separated from the other data (the usual intuitive notion of a cluster), without its distribution being Gaussian. This cluster may be represented by two or more mixture components, if its distribution is better approximated by a mixture of Gaussians than by a single Gaussian component. Thus the number of clusters in the data may be different from the number of components in the best approximating Gaussian mixture model.

To overcome this problem, Biernacki et al. (2000) proposed estimating the number of *clusters* (as distinct from the number of mixture components) in model-based clustering using the integrated complete likelihood (ICL), defined as the integrated likelihood of the complete data $(\mathbf{x}, \mathbf{z})$. ICL is defined as

$$p(\mathbf{x}, \mathbf{z} \mid K) = \int_{\Theta_K} p(\mathbf{x}, \mathbf{z} \mid K, \theta)\pi(\theta \mid K)d\theta, \tag{7.4}$$

where

$$p(\mathbf{x}, \mathbf{z} \mid K, \theta) = \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{z}_i \mid K, \theta)$$

with

$$p(\mathbf{x}_i, \mathbf{z}_i \mid K, \theta) = \prod_{k=1}^{K} \pi_k^{z_{ik}} \left[\phi(\mathbf{x}_i \mid \omega_k)\right]^{z_{ik}}.$$

To approximate this integrated complete likelihood, Biernacki et al. (2000) proposed using a BIC-like approximation, leading to the criterion

$$\text{ICL}(K) = \log \mathbf{p}(\mathbf{x}, \hat{\mathbf{z}} \mid K, \hat{\theta}_K) - \frac{D_K}{2}\log n, \tag{7.5}$$

where the missing data have been replaced by their most probable values, given the parameter estimate $\hat{\theta}_K$.

Roughly speaking, ICL is equal to BIC penalized by the mean entropy

$$\text{Ent}(K) = -\sum_{k=1}^{K}\sum_{i=1}^{n} \tau_{ik}(\hat{\theta}_K) \log \tau_{ik}(\hat{\theta}_K) \geq 0, \tag{7.6}$$

where τ_{ik} denotes the conditional probability that $\mathbf{x}_i$ arises from the kth mixture component ($1 \leq i \leq n$ and $1 \leq k \leq K$), namely

$$\tau_{ik}(\hat{\theta}_K) = \frac{\hat{\pi}_k \phi(\mathbf{x}_i|\hat{\omega}_k)}{\sum_{j=1}^{K} \hat{\pi}_j \phi(\mathbf{x}_i|\hat{\omega}_j)}.$$

Thus the number of clusters, K', favored by ICL tends to be smaller than the number K favored by BIC because of the additional entropy term. ICL aims to find the number of clusters rather than the number of mixture components. However, if it is used to estimate the number of mixture components it can underestimate it, particularly in data arising from mixtures with poorly separated components. In that case, the fit is worsened.

See Chapter 4 for a different point of view about ICL.

Thus the user of model-based clustering faces a dilemma: do the mixture components really all represent clusters, or do some subsets of them represent clusters with non-Gaussian distributions? In the next section, we propose a methodology to help resolve this dilemma.

7.3 Methodology

The idea is to build a sequence of clusterings, starting from a mixture model that fits the data well. Its number of components is chosen using BIC. We design a sequence of candidate soft clusterings with $\hat{K}^{\text{BIC}}$, $\hat{K}^{\text{BIC}} - 1, \ldots, 1$ clusters by successively merging the components in the BIC solution.

At each stage, we choose the two mixture components to be merged so as to minimize the entropy of the resulting clustering. Let us denote by $\tau_{i1}^K, \ldots, \tau_{iK}^K$ the conditional probabilities that $\mathbf{x}_i$ arises from cluster $1, \ldots, K$ with respect to the K-cluster solution. If clusters k and k' from the K-cluster solution are combined, the τ_{ij}'s remain the same for every j except for k and k'. The new cluster $k \cup k'$ then has the following conditional probability:

$$\tau_{i\,k\cup k'}^K = \tau_{ik}^K + \tau_{ik'}^K.$$

Then the resulting entropy is:

$$-\sum_{i=1}^{n}\left(\sum_{j\neq k,k'} \tau_{ij}^K \log \tau_{ij}^K + (\tau_{ik}^K + \tau_{ik'}^K) \log (\tau_{ik}^K + \tau_{ik'}^K)\right). \tag{7.7}$$

Thus, the two clusters k and k' to be combined are those that maximize the criterion

$$-\sum_{i=1}^{n}\{\tau_{ik}^K \log(\tau_{ik}^K) + \tau_{ik'}^K \log(\tau_{ik'}^K)\} + \sum_{i=1}^{n} \tau_{i\,k\cup k'}^K \log \tau_{i\,k\cup k'}^K$$

among all possible pairs of clusters (k, k'). Then $\tau_{ik}^{K-1},\ i = 1, \ldots, n,\ k = 1, \ldots, K-1$ can be updated.

At the first step of the combining procedure, $K = \hat{K}^{\text{BIC}}$ and τ_{ik}^{K} is the conditional probability that $\mathbf{x}_i$ arises from the kth mixture component ($1 \leq i \leq n$ and $1 \leq k \leq K$). But as soon as at least two components are combined in a cluster k (hence $K < \hat{K}^{\text{BIC}}$), τ_{ik}^{K} is the conditional probability that observation $\mathbf{x}_i$ belongs to one of the combined components in cluster k.

Our method is a soft clustering one that yields probabilities of cluster membership rather than cluster assignments. However, it can be used as the basis for a hard clustering method, simply by assigning the maximum a posteriori cluster memberships. Note that this will not necessarily be a strictly hierarchical clustering method. For example, an observation that was not assigned to either cluster k or k' by the K-cluster solution might be assigned to cluster $k \cup k'$ by the $(K-1)$-cluster solution.

Any combined solution fits the data as well as the BIC solution, since it is based on the same Gaussian mixture; the likelihood does not change. Only the number and definition of clusters are different. Our method yields just one suggested set of clusters for each K, and the user can choose between them on substantive grounds. Our flow cytometry data example in Section 7.5 provides one instance of this.

If a more automated procedure is desired for choosing a single solution, one possibility is to select, among the possible solutions, the solution providing the same number of clusters as ICL. An alternative is to use an elbow rule on the graphic displaying the entropy variation against the number of clusters. Both these strategies are illustrated in our examples.

The algorithm implementing the suggested procedure is given in Section 7.7.

7.4 Simulated Examples

We first present some simulations to highlight the possibilities of our methodology. They have been chosen to illustrate cases where BIC and ICL do not select the same number of components[1].

7.4.1 Simulated Example with Overlapping Components

The data, shown in Figure 7.1(a), were simulated from a two-dimensional Gaussian mixture. There are six components, four of which are axis-aligned with diagonal variance matrices (the four components of the two "crosses"), and two of which are not axis-aligned, and so do not have diagonal variance matrices. There were 600 points, with mixing proportions 1/5 for each non axis-aligned component, 1/5 for each of the upper left cross components, and 1/10 for each of the lower right cross components.

We fitted Gaussian mixture models to this simulated dataset. This experiment was repeated with 100 different such datasets, but we first present a single one of them to

[1]Details on the simulation settings may be found in Section A.2.

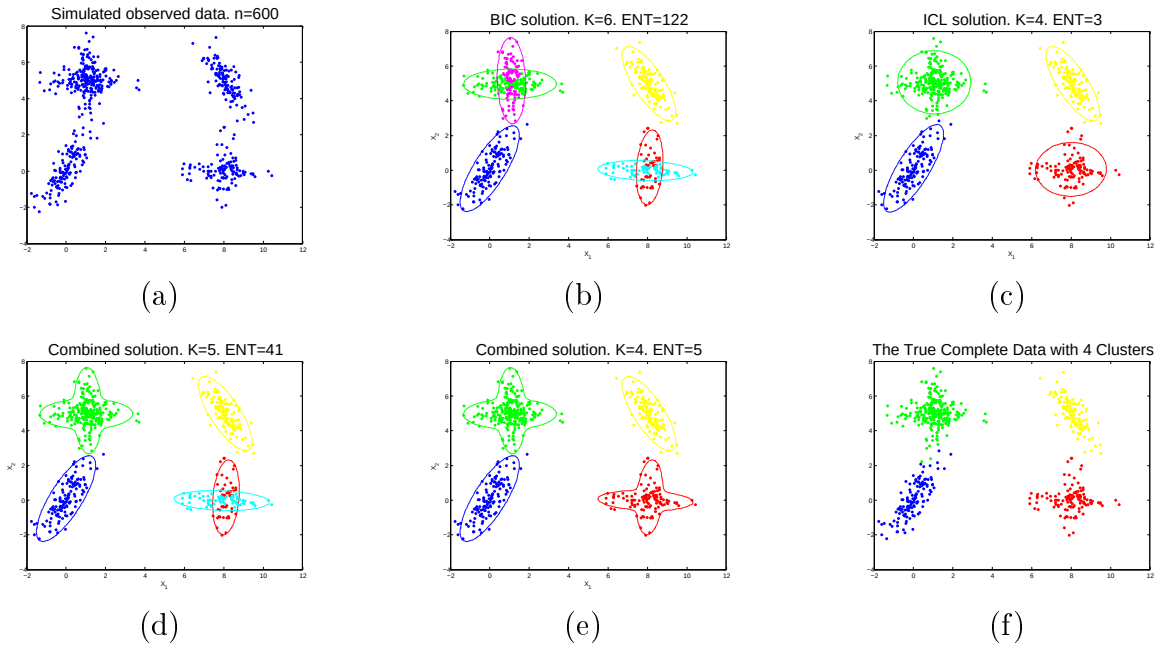


Figure 7.1: Simulated Example 1. (a) Simulated data from a six-component two-dimensional Gaussian mixture. (b) BIC solution with six components. (c) ICL solution with four clusters. (d) Combined solution with five clusters. (e) Combined solution with four clusters. (f) The true labels for a four-cluster solution. In (b) and (c) the entropy, ENT, is defined by equation (7.6) with respect to the the maximum likelihood solution, and in (d) and (e) ENT is defined by equation (7.7).

illustrate the method. Although all the features of our approach cannot be tabulated, results illustrating the stability of the method are reported and discussed at the end of this subsection.

For the dataset at hand, the BIC selected a six-component mixture model, which was the correct model; this is shown in Figure 7.1(b). ICL selected a four-cluster model, as shown in Figure 7.1(c). The four clusters found by ICL are well separated.

Starting from the BIC six-component solution, we combined two components to get the five-cluster solution shown in Figure 7.1(d). To decide which two components to merge, each pair of components was considered, and the entropy after combining these components into one cluster was computed. The two components for which the resulting entropy was the smallest were combined.

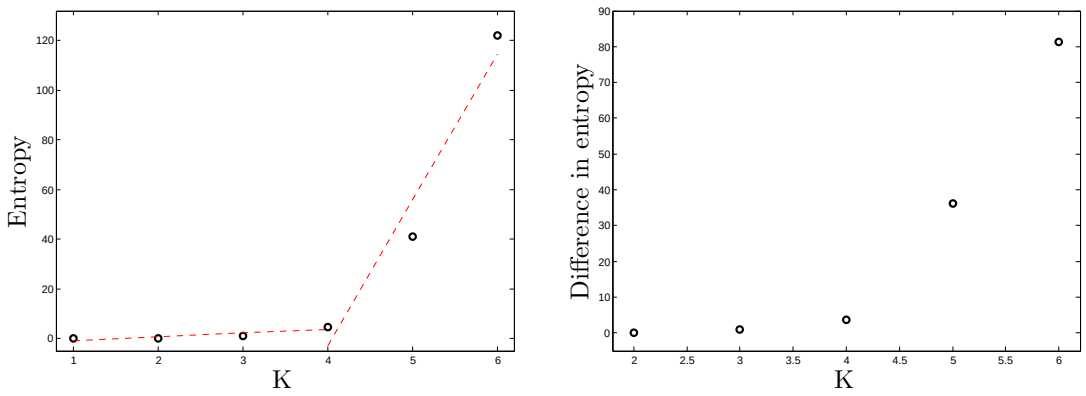
The same thing was done again to find a four-cluster solution, shown in Figure 7.1(e). This is the number of clusters identified by ICL. Note that there is no conventional formal statistical inferential basis for choosing between different numbers of clusters, as the likelihood and the distribution of the observations are the same for all the numbers of clusters considered.

However, the decrease of the entropy at each step of the procedure may help guide the choice of the number of clusters, or of a small number of solutions to be considered. The entropies of the combined solutions are shown in Figure 7.2, together with the differences between successive entropy values. There seems to be an elbow in the plot at $K = 4$, and together with the choice of ICL, this leads us to focus on this solution.

A finer examination of those graphics gives more information about the merging process. The first merging (from six to five clusters) is clearly necessary, since the decrease in entropy is large (with respect for example to the minimal decreases, when merging from two to one clusters, say). The second merging (from five to four clusters) also seems to be necessary for the same reason, although it results in a smaller decrease of the entropy (about half of the first one). This is far from zero, but indicates either that the components involved in this merging overlap less than the first two to be merged, or that this merging involves only about half as many observations as the first merging.

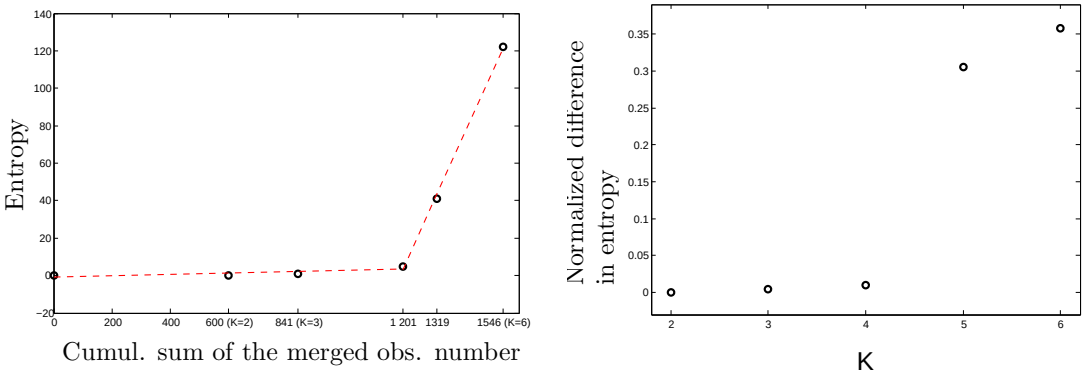
To further analyze the situation, we suggest changing the scale of the first of those graphics so that the difference between the abscissas of two successive points is proportional to the number of observations involved in the corresponding merging step: see Figure 7.3(a). This plot leads to the conclusion that the reason why the second merging step gives rise to a smaller entropy decrease than the first one is that it involves fewer observations. The mean decrease in entropy for each observation involved in the corresponding merging step is about the same in both cases, since the last three points of this graphic are almost collinear. The same result can be seen in a slightly different way by plotting the differences of entropies divided by the number of observations involved at each step, as shown in Figure 7.3(b). These new graphical representations accentuate the elbow at $K = 4$.

In the four-cluster solution, the clusters are no longer all Gaussian; now two of them are modeled as mixtures of two Gaussian components each. Note that this four-cluster solution is not the same as the four-cluster solution identified by ICL: ICL identifies a mixture of four Gaussians, while our method identifies four clusters of which two are not Gaussian. Figure 7.1(f) shows the true classification. Only three of the 600 points



(a) (b)

Figure 7.2: (a) Entropy values for the K-cluster Combined Solution, as defined by equation (7.7), for Simulated Example 1. The dashed line shows the best piecewise linear fit, with a breakpoint at $K = 4$ clusters. (b) Differences between successive entropy values.



(a) (b)

Figure 7.3: Simulated Example 1: (a) Entropy values for the K-cluster Combined Solution, as defined by equation (7.7), plotted against the cumulative sum of the number of observations merged at each step. The dashed line shows the best piecewise linear fit, with a breakpoint at $K = 4$ clusters. (b) Rescaled differences between successive entropy values: $\frac{\text{ENT}(K+1)-\text{ENT}(K)}{\text{Number of merged obs.}}$.

were misclassified.

It will often be scientifically useful to examine the full sequence of clusterings our method yields and assess them on substantive grounds, as well as by inspection of the entropy plots. However, an automatic way of choosing the number of clusters may be desired. A simple approach to this was proposed by Byers and Raftery (1998), in a different context. Consider successively each possible breakpoint for a two-part piecewise linear regression in the entropy plot; fit a linear regression model to the values to the left (resp. to the right) of this breakpoint (included); finally, use the breakpoint leading to the smallest total least-square value as the selected number of clusters.

For simulated example 1, this is shown as the dashed line in Figure 7.2(a) for the raw entropy plot and in Figure 7.5(a) for the rescaled entropy plot. The method chooses $K = 4$ using both the raw and rescaled entropy plots, but the fit of the piecewise linear regression model is better for the rescaled entropy plot, as expected.

We repeated this experiment 100 times to assess the stability of the method, simulating new data from the same model each time. The piecewise linear regression model fit to the rescaled entropy plot selected $K = 4$, 95 times out of 100.

We carried out an analogous experiment in dimension 6. The "crosses" involved two components each, with four discriminant directions between them and two noisy directions. The proportions of the components were equal. Our piecewise linear regression model method almost always selected 4 clusters.

7.4.2 Simulated Example with Overlapping Components and Restrictive Models

We now consider the same data again, but this time with more restrictive models. Only Gaussian mixture models with diagonal variance matrices are considered. This illustrates what happens when the mixture model generating the data is not in the set of models considered.

BIC selects more components than before, namely 10 (Figure 7.4(a)). This is because the true generating model is not considered, and so more components are needed to approximate the true distribution. For example, the top right non-axis-aligned component cannot be represented correctly by a single Gaussian with a diagonal variance matrix, and BIC selects three diagonal Gaussians to represent it. ICL still selects four clusters (Figure 7.4(b)).

In the hierarchical merging process, the two components of one of the "crosses" were combined first (Figure 7.4(c)), followed by the components of the other cross (Figure 7.4(d)). The nondiagonal cluster on the lower left was optimally represented by three diagonal mixture components in the BIC solution. In the next step, two of these three components were combined (Figure 7.4(e)). Next, two of the three mixture components representing the upper right cluster were combined (Figure 7.4(f)). After the next step there were five clusters, and all three mixture components representing the lower left cluster had been combined (Figure 7.4(g)).

The next step got us to four clusters, the number identified by ICL (Figure 7.4(h)). After this last combination, all three mixture components representing the upper right

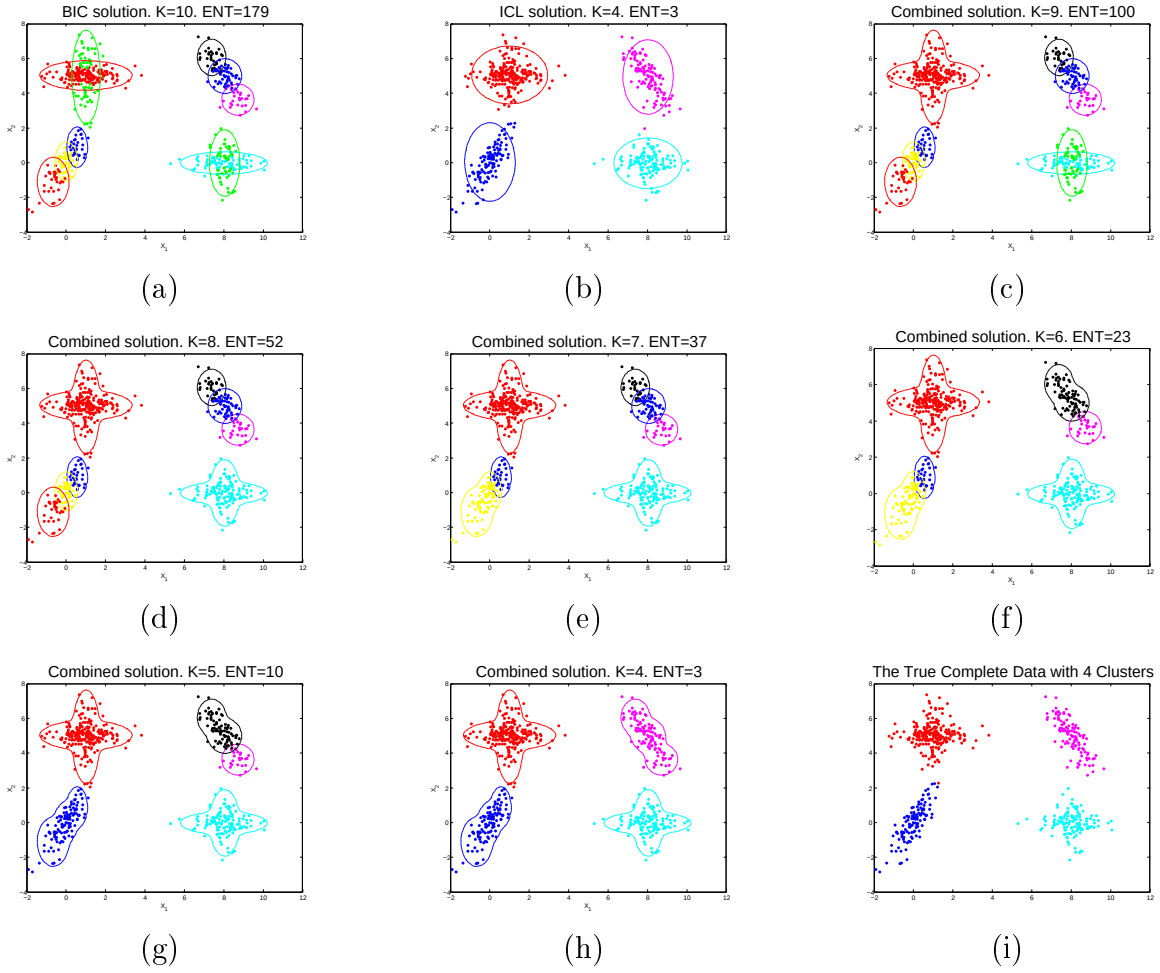


Figure 7.4: Simulated Example 2. The data are the same as in Simulated Example 1, but the model space is more restrictive, as only Gaussian mixture models with diagonal covariance matrices are considered. See Fig.7.1 legends for explanations about ENT. (a) BIC solution with ten mixture components. (b) ICL solution with four clusters. (c) Combined solution with nine clusters. (d) Combined solution with eight clusters. (e) Combined solution with seven clusters. (f) Combined solution with six clusters. (g) Combined solution with five clusters. (h) Combined solution with four clusters. (i) True labels with four clusters.

cluster had been combined. Note that this four-cluster solution is not the same as the four-cluster solution got by optimizing ICL directly. Strikingly, this solution is almost identical to that obtained with the less restrictive set of models considered in Section 7.4.1.

The plot of the combined solution entropies against the number of components in Figure 7.5 suggests an elbow at $K = 8$, with a possible second, less apparent one at $K = 4$. In the $K = 8$ solution the two crosses have been merged, and in the $K = 4$ solution all four visually apparent clusters have been merged. Recall that the choice of the number of clusters is not based on formal statistical inference, unlike the choice of the number of mixture components. Our method generates a small set of possible solutions that can be compared on substantive grounds. The entropy plot is an exploratory device that can help to assess separation between clusters, rather than a formal inference tool.

In this example, the elbow graphics (Figures 7.5(a) and 7.5(c)) exhibit three different stages in the merging process (a two-change-point piecewise line is necessary to fit to fit them well):

- The two first merging steps (from ten to eight clusters) correspond to a large decrease in entropy (Figure 7.5(a)). They are clearly necessary. The mean entropy is equivalent in each one of those two steps (Figure 7.5(c)). Indeed, Figure 7.4 shows that they correspond to the formation of the two crosses.

- The four following merging steps (from eight to four clusters) correspond to smaller decreases in entropy (Figure 7.5(a)). They have a comparable common mean decrease of entropy, but it is smaller than that of the first stage: a piece of the line would be fitted for them only (as appears in Figure 7.5(c)). They correspond to the merging of components which overlap in a different way than those merged at the first stage (Figure 7.4).

- The four last merging steps should not be applied.

In this case the user can consider the solutions with four and eight clusters, and take a final decision according to the needs of the application. The automatic rule in Section 7.4.1 (see Figure 7.5(d)) selects $K = 6$ clusters, which splits the difference between the two solutions we identified by inspection of the plot. This seems reasonable if a single automatic choice is desired, but either four or eight clusters might be better in specific contexts.

7.4.3 Circle/Square Example

This example was presented by Biernacki et al. (2000). The data shown in Figure 7.6(a) were simulated from a mixture of a uniform distribution on a square and a spherical Gaussian distribution. Here, for illustrative purposes, we restricted the models considered to Gaussian mixtures with spherical variance matrices with the same determinant. Note that the true generating model does not belong to this model class.

In the simulation results of Biernacki et al. (2000), BIC chose two components in only 60% of the simulated cases. Here we show one simulated dataset in which BIC

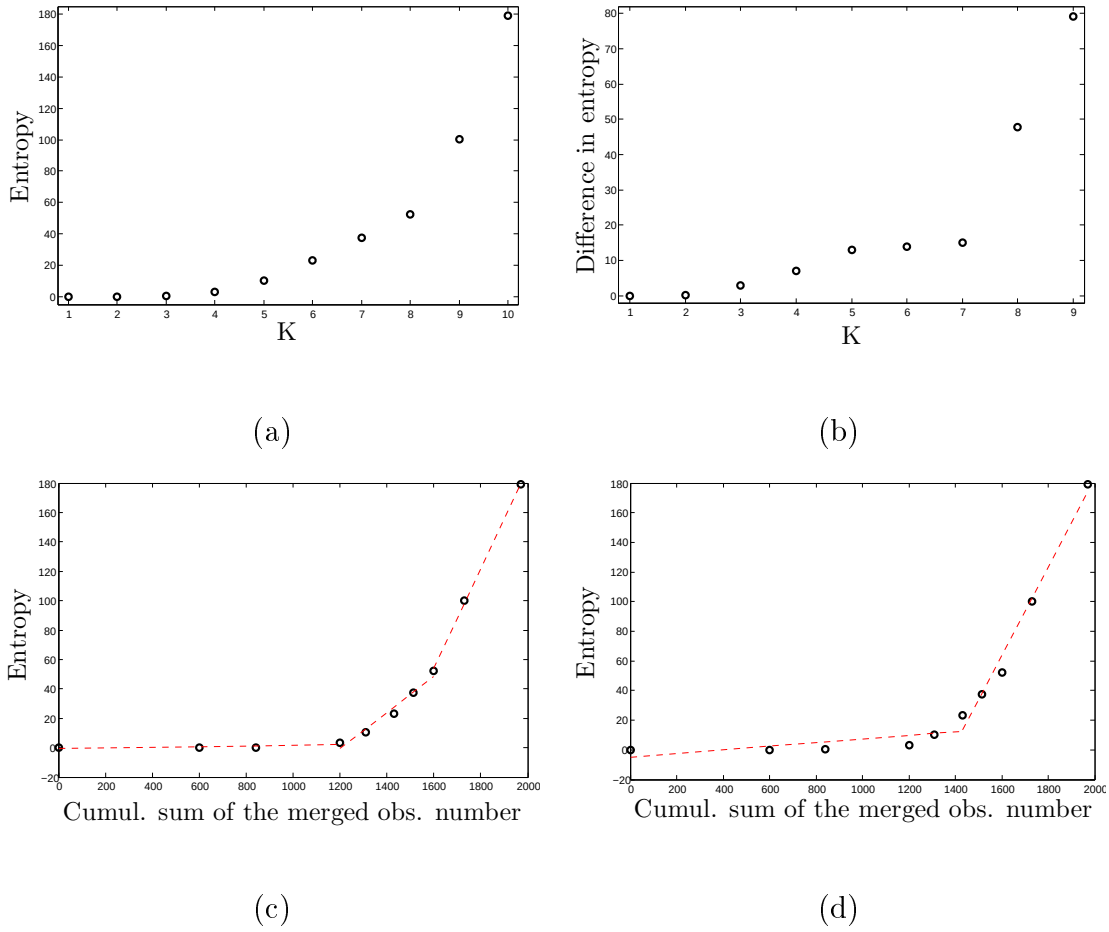


Figure 7.5: (a) Entropy values for the K-cluster Combined Solution, as defined by equation (7.7), for Simulated Example 2. (b) Differences between successive entropy values. (c) Entropy values with respect to the cumulative sum of the number of observations merged at each step $K+1 \rightarrow K$. Two change-points piecewise linear regression. (d) Entropy values with respect to the cumulative sum of the number of observations merged at each step $K+1 \rightarrow K$. Single change-point piecewise linear regression with minimum least-squares choice of the change-point.

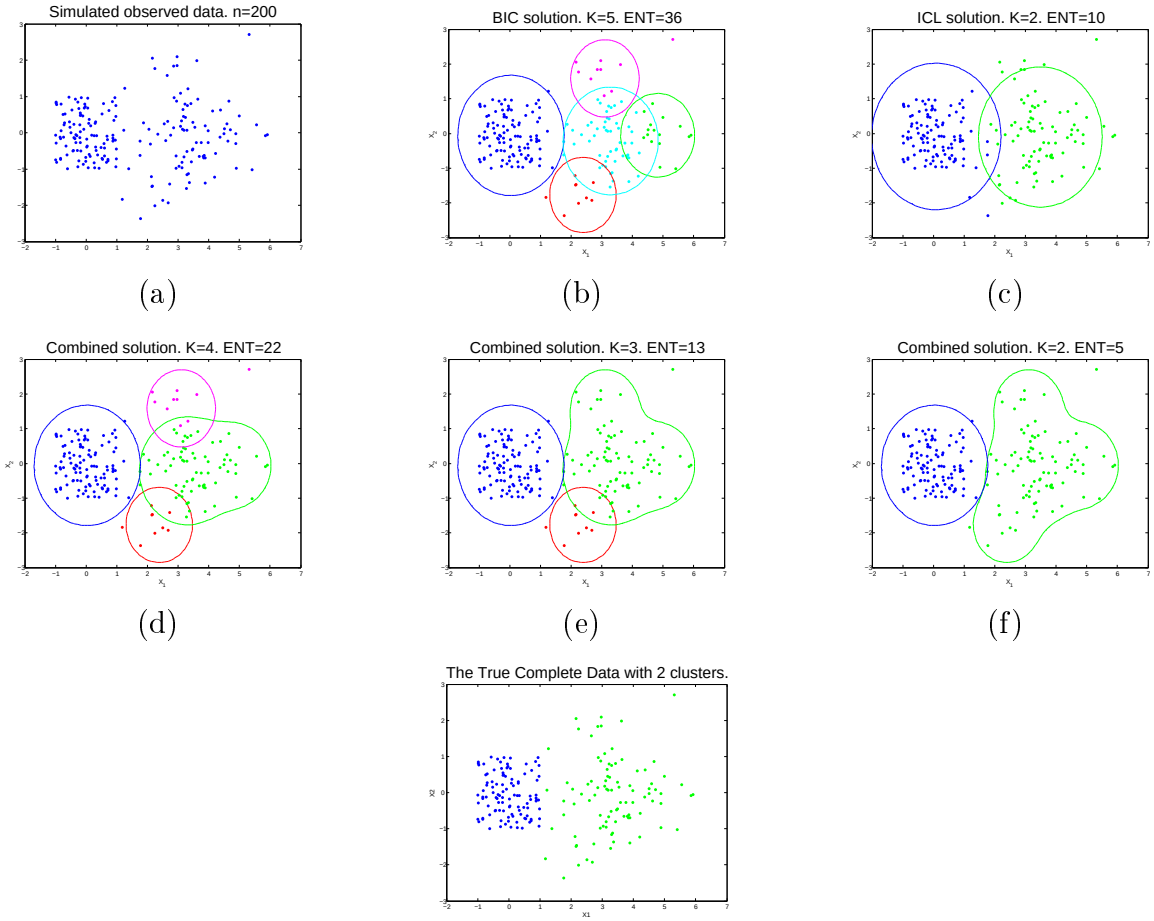
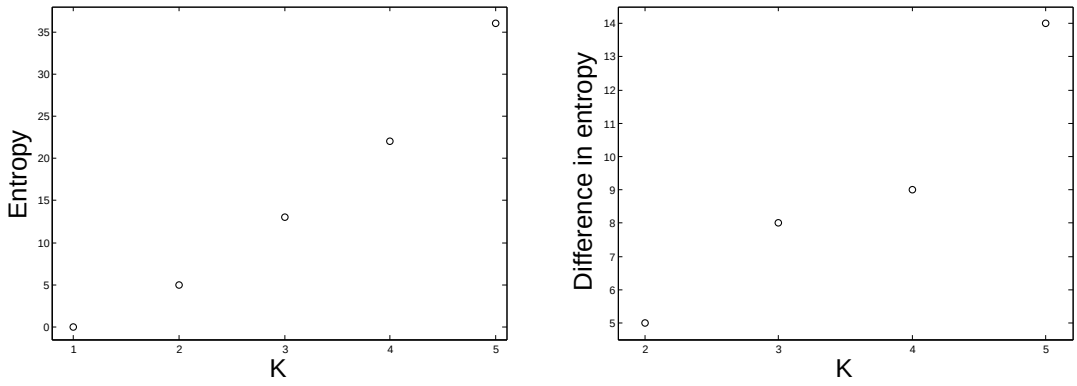


Figure 7.6: Circle-Square Example. See Figure 7.1 legends for explanations about ENT. (a) Observed data simulated from a mixture of a uniform distribution on a square and a spherical Gaussian distribution. (b) The BIC solution, with five components. (c) The ICL solution with two clusters. (d) The combined solution with four clusters. (e) The combined solution with three clusters. (f) The final combined solution, with two clusters. (g) The true labels.



(a) (b)

Figure 7.7: (a) Entropy values for the K-cluster Combined Solution, as defined by equation (7.7), for the Circle-Square Example. (b) Differences between successive entropy values.

approximated the underlying non-Gaussian density using a mixture of five normals (Figure 7.6(b)). ICL always selected two clusters (Figure 7.6(c)).

The progress of the combining algorithm is shown in Figure 7.6(d-f). The final two-cluster solution, obtained by hierarchical merging starting from the BIC solution, is slightly different from the clustering obtained by optimizing ICL directly. It also seems slightly better: ICL classifies seven observations into the uniform cluster that clearly do not belong to it, while the solution shown misclassifies only three observations in the same way. The true labels are shown in Figure 7.6(g). The entropy plot in Figure 7.7 does not have a clear elbow.

7.4.4 Comparison With Li's Method

In this section, our methodology is compared with the related method of Li (2005). Similarly to our approach, Li proposed modeling clusters as Gaussian mixtures, starting with the BIC solution, and then merging mixture components. However, unlike us, Li assumed that the true number of clusters is known in advance. The author also used k-means clustering to merge components; this works well when the mixture components are spherical but may have problems when they are not.

In the framework of the so-called multilayer mixture model, Li (2005) proposed two methods for partitioning the components of a mixture model into a fixed number of clusters. They are both initialized with the same double-layer k-means procedure. Then the first method consists of computing the maximum likelihood estimator of a Gaussian mixture model with a greater number of components than the desired number of clusters. The components are then merged by minimizing a within-cluster inertia criterion (sum of squares) on the mean vectors of the mixture components. The second method consists of fitting the Gaussian mixture model through a CEM-like algorithm (Celeux and Govaert, 1992), to maximize the classification likelihood, where the clusters are taken as mixtures of components. The total number of components (for each method)

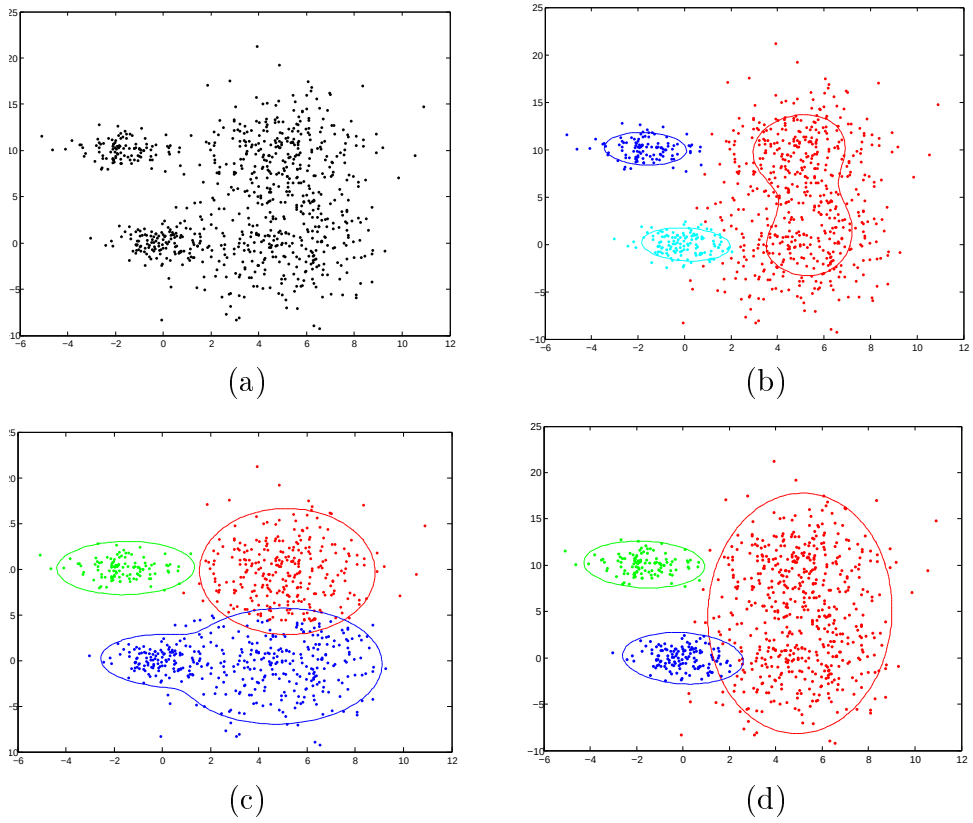


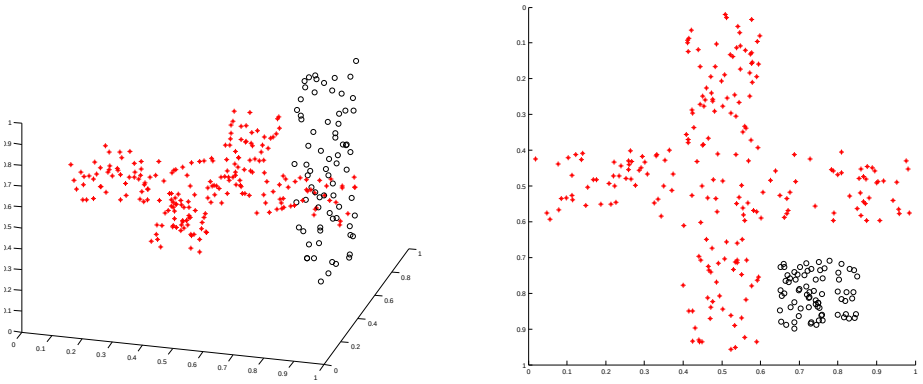
Figure 7.8: Comparison with Li's method. (a) A simulated data set to compare Li's method with ours (b) The three cluster solution with our method (c) The three cluster solution with most of Li's methods (d) The typical three cluster solution with Li's "k-means on the means + ICL" method

and the number of components per cluster (for the CEM method) are selected either through BIC or ICL.

First Experiment: Gaussian Mixture

We simulated 100 samples of size $n = 800$ of a four component Gaussian mixture in $\mathbf{R}^2$. An example of such a sample is shown in Figure 7.8(a).

Since Li's method imposes a fixed number of clusters, we fixed it to three and stopped our algorithm as soon as it yielded three clusters. For each simulated sample we always obtained the same kind of result for both methods. They are depicted in Figure 7.8(b) for our method, which always gave the same result. Figure 7.8(c) shows the results for the four variants of Li's method. Li's method with the CEM-like algorithm always gave rise to the solution in Figure 7.8(c). Li's method with the k-means on the means and the selection through BIC found the same solution in 93 of the 100 cases. The method with the k-means on the means and the selection through ICL found such a solution in 27 cases, but in most other cases found a different solution whose fit was poorer (Figure 7.8(d)).



3D View View from above

Figure 7.9: Simulated Example 2: There are two clusters, the 3D cross (red) and the uniform pillar (black). The true cluster memberships are shown here.

Second Experiment: 3D Uniform Cross

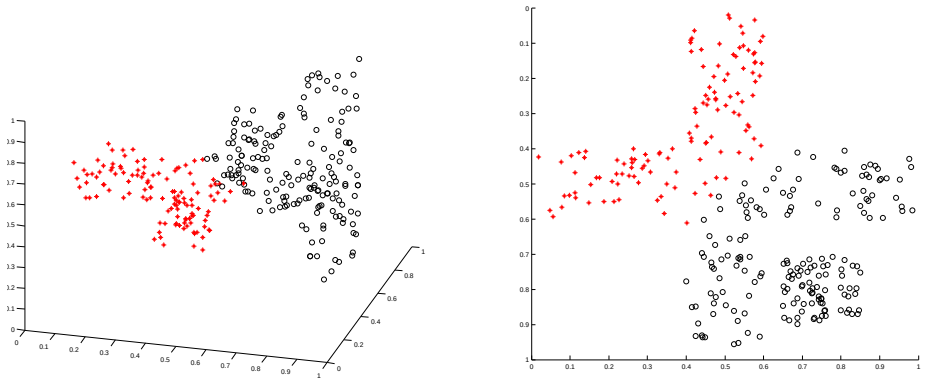
We simulated data in $\mathbf{R}^3$ from a mixture of two uniform components: see Figure 7.9. One is a horizontal thick cross (red in Figure 7.9) and has proportion 0.75 in the mixture, while the other is a vertical pillar (black in Figure 7.9) and has proportion 0.25. We simulated 100 datasets of size 300, and we applied Li's procedures, Ward's sum of squares method, and ours. We fixed the number of clusters to be designed at its true value (two), and we then fitted general Gaussian mixture models.

BIC selected 4 components for 69 of the 100 datasets, and 3 components for 18 of them. ICL selected 4 components for 60 of the datasets, and 3 components for 29 of them.

As in the preceding example, Li's approach did not recover the true clusters. Li's CEM-like methods always yielded a bad solution: sometimes one of the arms of the cross merged to the pillar, and sometimes two, as in Figure 7.10. Li's BIC + k-means method recovered the true clusters in 19 cases out of 100, and Li's ICL + k-means method did so in 33 cases out of 100. This occurred almost every time the number of Gaussian components was 3 (two for the cross, which then have almost the same mean, and one for the pillar). When the number of fitted components is higher, the distance between the means of the components is no longer a relevant criterion, and those methods yielded clusterings such as Figure 7.10.

Our merging procedure almost always (95 times out of 100) recovered the true clusters.

The same experiment performed with 50 datasets of size 1000 (instead of size 300) strengthened those conclusions in favor of our method in this setting.



3D View View from above

Figure 7.10: Example Solution with Li's Procedures

Conclusions on the Comparisons with Other Methods

Here are some comments on the comparison between Li's methods and ours based on these simulations. Our method takes into account the overlap between components to choose which ones to merge, whereas Li's method is based on the distances between the component means, through the initialization step in each method, and also through the merging procedure in the first method. This sometimes leads to mergings that are not relevant from a clustering point of view.

Our method is appreciably faster since only one EM estimation has to be run for each considered number of components, whereas numerous runs are needed with Li's method. Our procedure can also be applied when the number of clusters is unknown, unlike Li's method.

We also compared our results with those of a non-model-based clustering method: Ward's hierarchical method (Ward, 1963). We used Matlab's *datacluster* function to apply this procedure in each of the experiments described in this section. Ward's method always found irrelevant solutions, close to Li's ones, for each of the 200($= 2 \times 100$) datasets.

7.5 Flow Cytometry Example

We now apply our method to the GvHD data of Brinkman et al. (2007, page 4201). Two samples of this flow cytometry data have been used, one from a patient with the graft-versus-host disease (GvHD), and the other from a control patient. GvHD occurs in allogeneic hematopoietic stem cell transplant recipients when donor-immune cells in the graft attack the skin, gut, liver, and other tissues of the recipient. GvHD is one of the most significant clinical problems in the field of allogeneic blood and marrow transplantation.

The GvHD positive and control samples consist of 9,083 and 6,809 observations respectively. Both samples include four biomarker variables, namely, CD4, CD8β, CD3

and CD8. The objective of the analysis is to identify $CD3^+$ $CD4^+$ $CD8\beta^+$ cell sub-populations present in the GvHD positive sample. In order to identify all cell sub-populations in the data, we use a Gaussian mixture model with unrestricted covariance matrix. Adopting a similar strategy to that described by Lo et al. (2008, page 4577), for a given number of components, we locate the $CD3^+$ sub-populations by labeling components with means in the CD3 dimension above 280 $CD3^+$. This threshold was based on a comparison with a negative control sample, as explained by Brinkman et al. (2007, page 4201).

We analyze the positive sample first. A previous manual analysis of the positive sample suggested that the $CD3^+$ cells could be divided into six $CD3^+$ cell sub-populations Brinkman et al. (2007, page 4201). ICL selected nine clusters, five of which correspond to the $CD3^+$ population (Figure 7.11(b)). Compared with the result shown in Lo et al. (2008, page 4577), the $CD4^+$ $CD8\beta^-$ region located at the bottom right of the graph is missing.

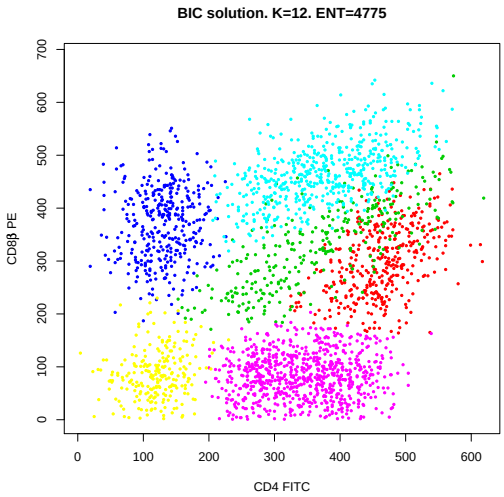
BIC selected 12 components to provide a good fit to the positive sample, six of which are labeled $CD3^+$ (Figure 7.11(a)). The $CD4^+$ $CD8\beta^+$ region seems to be encapsulated by the cyan, green, and red components. Starting from this BIC solution, we repeatedly combined two components causing maximal reduction in the entropy. The first three combinations all occurred within those components originally labeled $CD3^-$, and the CD4 vs $CD8\beta$ projection of the $CD3^+$ sub-populations remains unchanged.

The combined solution with nine clusters, in which six are labeled $CD3^+$, provides the most parsimonious view of the positive sample while retaining the six important $CD3^+$ cell sub-populations. However, when the number of clusters is reduced to eight, the magenta cluster representing the $CD3^+$ $CD4^+$ $CD8\beta^-$ population is combined with the big $CD3^-$ cluster, resulting in an incomplete representation of the $CD3^+$ population (Figure 7.11 (c)). Note that the entropy of the combined solution with nine clusters (1474) was smaller (i.e. better) than that of the ICL solution (3231). The entropy plot along with the piecewise regression analysis (Figure 7.12) suggests an elbow at $K = 9$ clusters, agreeing with the number of clusters returned by the ICL as well as our more substantively-based conclusion.

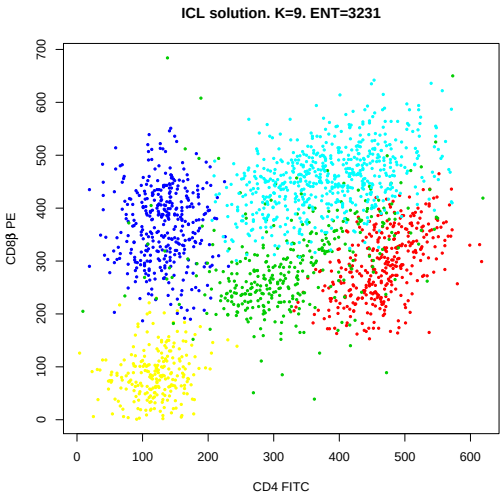
Next we analyze the control sample. A satisfactory analysis would show an absence of the $CD3^+$ $CD4^+$ $CD8\beta^+$ cell sub-populations. ICL chose seven clusters, three of which correspond to the $CD3^+$ population (Figure 7.13 (b)). The red cluster on the left of the graph represents the $CD4^-$ region. The blue cluster at the bottom right of the graph represents the $CD4^+CD8\beta^-$ region. It seems that it misses a part of this cluster near the red cluster. In addition, contrary to previous findings in which $CD4^+$ $CD8\beta^+$ cell sub-populations were found only in positive samples but not in control samples, a cyan cluster is used to represent the observations in the $CD4^+$ $CD8\beta^+$ region. These suggest that the ICL solution could be improved.

BIC selected 10 components, four of which are labeled $CD3^+$ (Figure 7.13(a)). A green component is found next to the blue component, filling in the missing part in the ICL solution and resulting in a more complete representation of the $CD4^+CD8\beta^-$ region. Meanwhile, similarly to the ICL solution, a cyan component is used to represent the observations scattered within the $CD4^+$ $CD8\beta^+$ region.

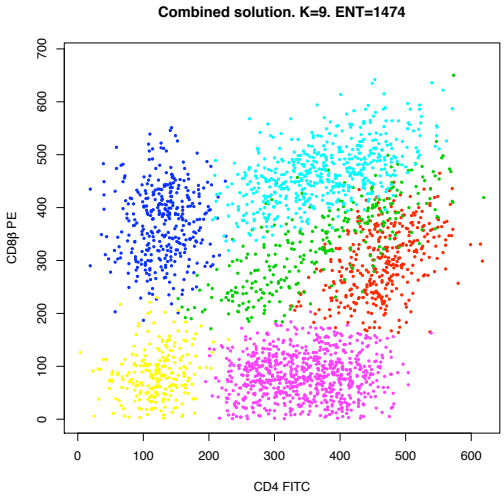
When we combined the components in the BIC solution, the first few combinations



(a)

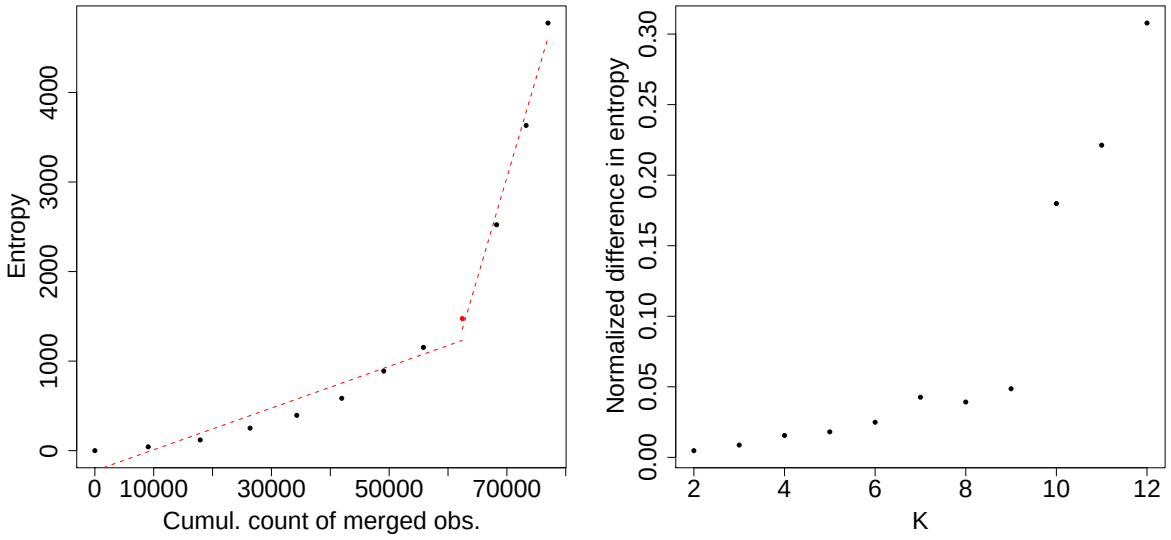


(b)



(c)

Figure 7.11: GvHD positive sample. Only components labeled CD3$^+$ are shown. (a) BIC Solution ($K = 12$). The combined solutions for $K = 11$ and $K = 10$ are almost identical for these CD3$^+$ components. (b) ICL Solution ($K = 9$). (c) Combined Solution ($K = 9$).



(a) (b)

Figure 7.12: (a) Entropy values for the GvHD positive sample. The piecewise regression analysis suggests that $K = 9$ is the optimal number of clusters. (b) Differences between successive entropy values.

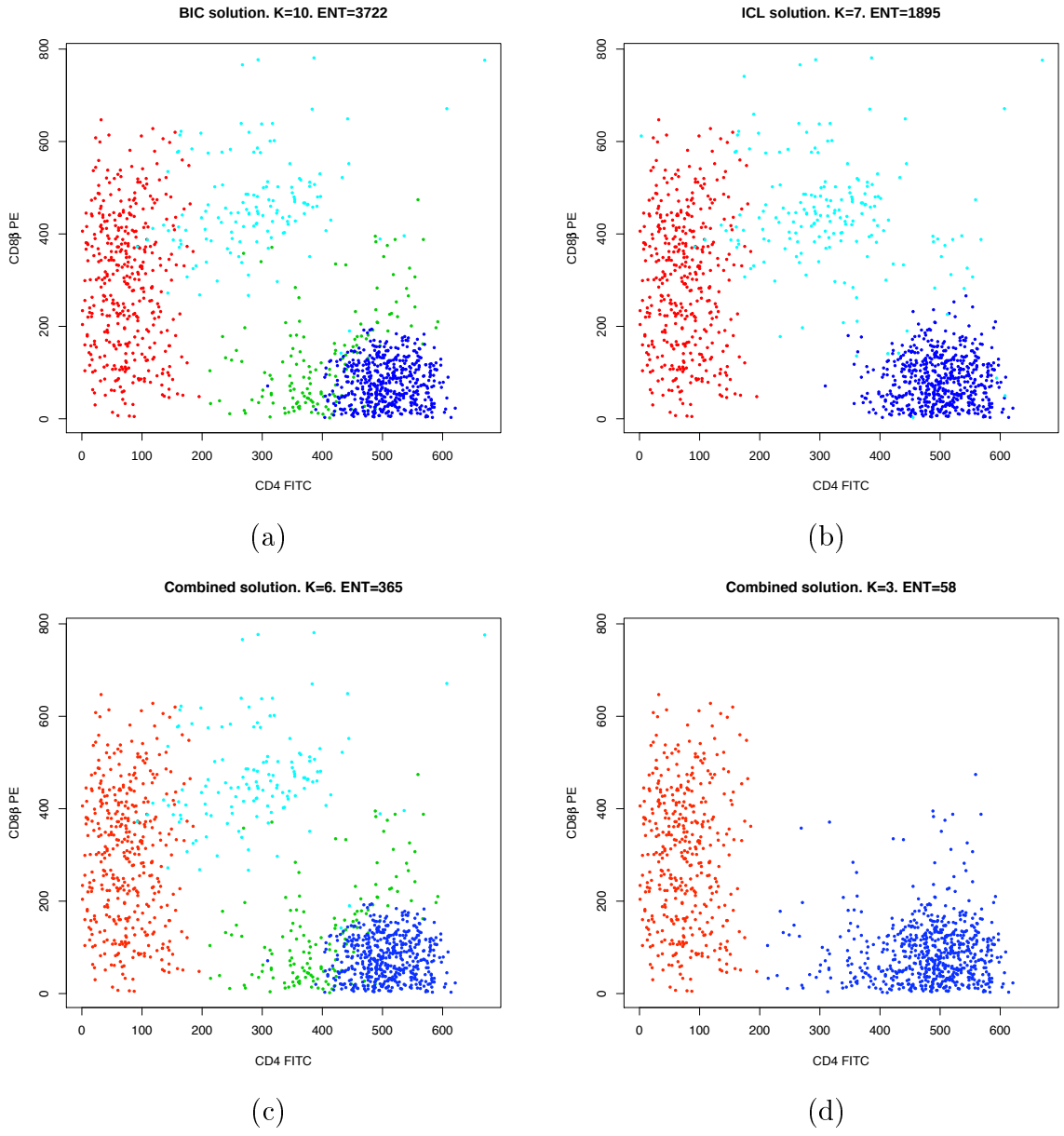
took place within those components initially labeled $CD3^-$, similarly to the result for the positive sample. Going from $K = 5$ to $K = 4$, the blue and green components in Figure 7.13(a) were combined, leaving the $CD3^+$ sub-populations to be represented by three clusters.

After one more combination ($K = 3$), the cyan component merged with a big $CD3^-$ cluster. Finally we had a "clean" representation of the $CD3^+$ population with no observations from the $CD3^+$ $CD4^+$ $CD8\beta^+$ region, consistent with the results of Brinkman et al. (2007, page 4201) and Lo et al. (2008, page 4577). This solution results in the most parsimonious view of the control sample with only three clusters but showing all the relevant features (Figure 7.13(d)). Once again, the entropy of the combined solution (58) was much smaller than that of the ICL solution (1895). Note that in this case we ended up with a combined solution that has fewer clusters than the ICL solution. The entropy plot along with the piecewise regression analysis (Figure 7.14) suggests an elbow at $K = 6$, but substantive considerations suggest that we can continue merging past this number.

7.6 Discussion

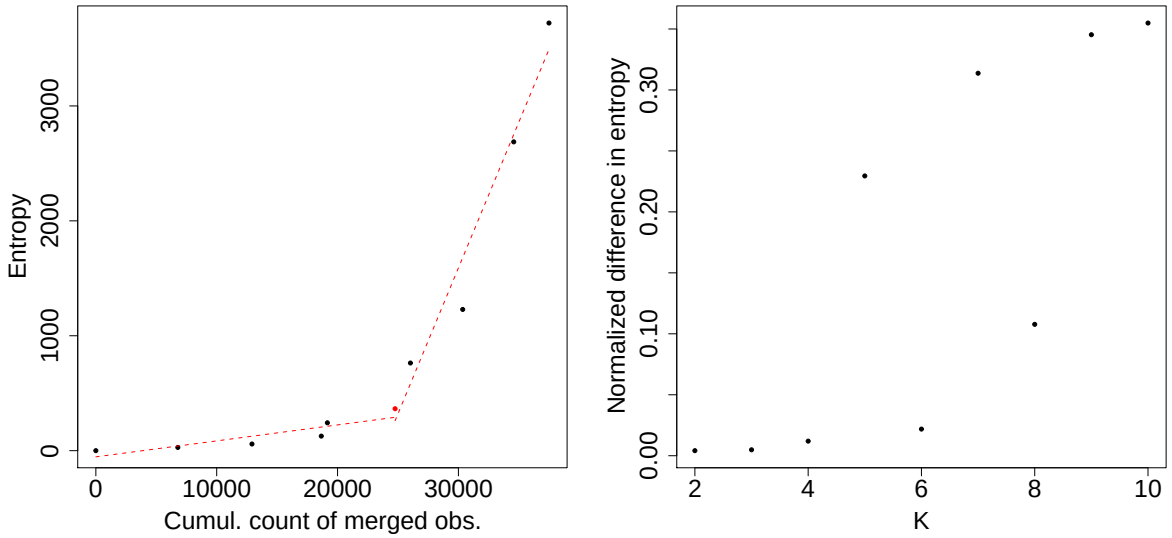
We have proposed a way of addressing the dilemma of model-based clustering based on Gaussian mixture models, namely that the number of mixture components selected is not necessarily equal to the number of clusters. This arises when one or more of the clusters has a non-Gaussian distribution, which is approximated by a mixture of several Gaussians.

Our strategy is as follows. We first fit a Gaussian mixture model to the data by max-



(a) (b) (c) (d)

Figure 7.13: GvHD control sample. Only components labeled $CD3^+$ are shown. (a) BIC Solution ($K = 10$). (b) ICL Solution ($K = 7$). (c) Combined Solution ($K = 6$). (d) Combined Solution ($K = 3$).



(a) (b)

Figure 7.14: (a) Entropy values for the K-cluster Combined Solution for the GvHD Control Sample. The piecewise regression analysis suggests that $K = 6$ is the optimal number of clusters. (b) Differences between successive entropy values.

imum likelihood estimation, using BIC to select the number of Gaussian components. Then we successively combine mixture components, using the entropy of the conditional membership distribution to decide which components to merge at each stage. This yields a sequence of possible solutions, one for each number of clusters, and in general we expect that users would consider these solutions from a substantive point of view.

The underlying statistical model is the same for each member of this sequence of solutions, in the sense that the likelihood and the modeled probability distribution of the data remain unchanged. What changes is the interpretation of this model. Thus standard statistical testing or model selection methods cannot be used to choose the preferred solution.

If a data-driven choice is required, however, we also describe two automatic ways of selecting the number of clusters, one based on a piecewise linear regression fit to the rescaled entropy plot, the other choosing the number of clusters selected by ICL. An inferential choice could be made, for example using the gap statistic (Tibshirani et al., 2001). However, the null distribution underlying the resulting test does not belong to the class of models being tested, so that it does not have a conventional statistical interpretation in the present context. It could still possibly be used in a less formal sense to help guide the choice of number of clusters.

Our method preserves the advantages of Gaussian model-based clustering, notably a good fit to the data, but it allows us to avoid the overestimation of the number of clusters that can occur when some clusters are non-Gaussian. The mixture distribution selected by BIC allows us to start the hierarchical procedure from a good summary of the data set. The resulting hierarchy is easily interpreted in relation to the mixture components. We stress that the whole hierarchy from K to 1 clusters might be informative.

Our merging procedure generally improves the entropy over the ICL solution. This highlights the better fit of the clusters that result from the merging procedure. Note that our method can also be used when the number of clusters K^* is known, provided that the number of mixture components in the BIC solution is at least as large as K^*.

One attractive feature of our method is that it is computationally efficient, as it uses only the conditional membership probabilities. Thus it could be applied to any mixture model, and not just to a Gaussian mixture model, effectively without modification. This includes latent class analysis Lazarsfeld (1950); Hagenaars and McCutcheon (2002), which is essentially model-based clustering for discrete data.

Several other methods for joining Gaussian mixture components to form clusters have been proposed. Walther (2002) considered the problem of deciding whether a univariate distribution is better modeled by a mixture of normals or by a single, possibly non-Gaussian and asymmetric distribution. To our knowledge, this idea has not yet been extended to more than one dimension, and it seems difficult to do so. Our method seems to provide a simple alternative approach to the problem addressed by Walther (2002), in arbitrary dimensions.

Wang and Raftery (2002, Section 4.5) considered the estimation of elongated features in a spatial point pattern with noise, motivated by a minefield detection problem. They suggested first clustering the points using Gaussian model-based clustering with equal spherical covariance matrices for the components. This leads to the feature being covered by a set of "balls" (spherical components), and these are then merged if their centers are close enough that the components are likely to overlap. This works well for joining spherical components, but may not work well if the components are not spherical, as it takes account of the component means but not their shapes.

Tantrum et al. (2003) proposed a different method based on the hierarchical model-based clustering method of Banfield and Raftery (1993). Hierarchical model-based clustering is a "hard" clustering method, in which each data point is assigned to one group. At each stage, two clusters are merged, with the likelihood used as the criterion for deciding which clusters to merge. Tantrum et al. (2003) proposed using the dip test of Hartigan and Hartigan (1985) to decide on the number of clusters. This method differs from ours in two main ways. Ours is a probabilistic ("soft") clustering method that merges mixture components (distributions), while that of Tantrum et al. (2003) is a hard clustering method that merges groups of data points. Secondly, the merging criterion is different.

As discussed earlier, Li (2005) assumed that the number of clusters K is known in advance, used BIC to estimate the number of mixture components, and joined them using k-means clustering applied to their means. This works well if the clusters are spherical, but may not work as well if they are elongated, as the method is based on the means of the clusters but does not take account of their shape. The underlying assumption that the number of clusters is known may also be questionable in some applications. Jörnsten and Keleş (2008) extended Li's method so as to apply it to multifactor gene expression data, allowing clusters to share mixture components, and relating the levels of the mixture to the experimental factors.

7.7 Merging Algorithm

Choose a family of mixture models: $\{\mathcal{M}_{K_{\min}}, \ldots, \mathcal{M}_{K_{\max}}\}$. General Gaussian mixture models are recommended, since the purpose is to take advantage of the nice approximation properties of the Gaussian mixture models to get a good estimate of the distribution of the data, through the BIC criterion.

Recall the method consists of first estimating the data distribution through BIC, and then merging hierarchically some of the obtained classes to get mixtures of Gaussian mixtures. The criterion to choose which classes to merge at each step is the decrease of entropy, which is to be maximized.

1. Compute MLE(K) for each model using the EM algorithm:

$$\forall K \in \{K_{\min}, \ldots, K_{\max}\}, \quad \hat{\theta}_K \in \arg\max_{\theta_K \in \Theta_K} \log p(\mathbf{x} \mid K, \theta_K)$$

2. Compute the BIC solution:

$$\hat{K}^{\mathrm{BIC}} = \underset{K \in \{K_{\min}, \ldots, K_{\max}\}}{\operatorname{argmin}} \left\{ -\log p(\mathbf{x} \mid K, \hat{\theta}_K) + \frac{D_K}{2} \log n \right\}$$

3. Compute the density f_k^K of each combined cluster k for each K from $\hat{K}^{\mathrm{BIC}}$ to $K_{\min}$:

$$\forall k \in \{1, \ldots, \hat{K}^{\mathrm{BIC}}\}, \quad f_k^{\hat{K}^{\mathrm{BIC}}}(\cdot) = \hat{\pi}_k^{\hat{K}^{\mathrm{BIC}}} \phi\left(\cdot \mid \hat{\omega}_k^{\hat{K}^{\mathrm{BIC}}}\right).$$

   For $K = \hat{K}^{\mathrm{BIC}}, \ldots, (K_{\min} + 1)$:

   - Choose the clusters l and l' to be combined at step $K \to K - 1$:

$$(l, l') = \underset{(k,k') \in \{1,\ldots,K\}^2,\ k \neq k'}{\operatorname{argmax}} \left\{ \begin{array}{l} -\sum_{i=1}^n \left\{ \tau_{ik}^K \log(\tau_{ik}^K) + \tau_{ik'}^K \log(\tau_{ik'}^K) \right\} \\ \qquad + \sum_{i=1}^n (\tau_{ik}^K + \tau_{ik'}^K) \log(\tau_{ik}^K + \tau_{ik'}^K) \end{array} \right\},$$

     where $\tau_{ik}^K = \frac{f_k^K(x_i)}{\sum_{j=1}^K f_j^K(x_i)}$ is the conditional probability of component k given the K-cluster combined solution.

   - Define the densities of the combined clusters for the (K-1) cluster solution by combining l and l':

$$\begin{array}{ll} f_k^{K-1} = f_k^K & \text{for } k = 1, \ldots, (l \wedge l' - 1), (l \wedge l' + 1), \ldots, (l \vee l' - 1); \\ f_{l \wedge l'}^{K-1} = f_l^K + f_{l'}^K; & \\ f_k^{K-1} = f_{k+1}^K & \text{for } k = l \vee l', \ldots, (K - 1). \end{array}$$

4. To select the number of clusters through ICL:

$$\hat{K}^{\text{ICL}} = \underset{K \in \{K_{\min}, \ldots, K_{\max}\}}{\operatorname{argmin}} \left\{ -\log p(\mathbf{x} \mid K, \hat{\theta}_K) - \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}(\hat{\theta}_K) \log \tau_{ik}(\hat{\theta}_K) + \frac{D_K}{2} \log n \right\},$$

where $\tau_{ik}(\hat{\theta}_K) = \frac{\hat{\pi}_k^K \phi(x_i|\hat{\omega}_k^K)}{\sum_{j=1}^{K} \hat{\pi}_j^K \phi(x_i|\hat{\omega}_j^K)}$ is the conditional probability of component k given the MLE for the model with K Gaussian components.

Chapter 8

Selecting a Clustering Model in View of an External Classification

Contents

Presentation. This chapter deals with a work in progress. This is a collaboration with Gilles Celeux and Ana Sousa Ferreira. The notations have been updated to be consistent with the thesis. Thus the reader who already read the chapters introducing mixture models, model-based clustering and the BIC and ICL penalized likelihood criteria, may skip Section 8.2. It has been left in the chapter to make it self-contained. This work takes advantage of the idea that a model selection criterion should take into account the user's purpose. This is the idea underlying the derivation of ICL (see Section 2.1.4 and Chapter 4), when the purpose is clustering. Now, assume that beside the observations, an external classification is available, which is expected to help enlighten the classification derived from the observations. The idea is that a class structure in the data distribution may be revealed by the model-based approach and that the corresponding classification may be — or not — related to the external classification. A penalized criterion is derived, based on a heuristics analogous to that which originally led to ICL, which purpose is to select a number of classes to this aim. It involves the study of contingency tables between the considered external classification and the classification derived from the model-based study. This is notably illustrated by a real dataset, for which several external classifications are available but only one influences the selection of the number of components.

8.1 Introduction

In model selection, assuming that the data arose from one of the models in competition is often somewhat unrealistic and could be misleading. However this assumption is implicitly present in standard model selection criteria such as AIC or BIC. This "true model" assumption could lead to overestimate the model complexity in practical situations. On the other hand, a common feature of standard penalized likelihood criteria such as AIC and BIC is that they do not take into account the modeling purpose. Our opinion is that taking account of the modeling purpose when selecting a model would lead to more flexible penalties favoring useful and parsimonious models. This point of view could be exploited in many statistical learning situations. Here, it is developed in a model-based clustering context to choose a sensible partition of the data favoring eventually partitions leading to a relevant interpretation with respect to external qualitative variables. The chapter is organized as follows. In Section 8.2, the framework of model-based clustering is described. Our new penalized likelihood criterion is presented in Section 8.3. Numerical experiments on simulated and real datasets are presented in Section 8.4 to illustrate the behavior of this criterion and highlight its possible interest. A short discussion section ends the chapter.

8.2 Model-Based Clustering

Model-based clustering consists of assuming that the dataset to be classified arises from a mixture model and to associate each class with one of the mixture components. Embedding cluster analysis in this precise framework is useful in many aspects. In particular, it allows to choose the number K of classes in a proper way: Choosing K is choosing the number of mixture components.

8.2.1 Finite Mixture Model

Data to be classified $\mathbf{x}$ in $\mathbb{R}^{nd}$ are assumed to arise from a mixture

$$f(x_i;\theta_K) = \sum_{k=1}^{K} \pi_k \phi(x_i;\omega_k)$$

where the π_k's are the mixing proportions and $\phi(\,.\,;\omega_k)$ denotes density (as the d-dimensional Gaussian density) with parameter ω_k, and $\theta_K = (\pi_1,\ldots,\pi_{K-1},\omega_1,\ldots,\omega_K)$. A mixture model is a latent structure model involving unknown label data $\mathbf{z} = (z_1,\ldots,z_n)$ which are binary vectors with $z_{ik} = 1$ if and only if x_i arises from component k. Those indicator vectors define a partition $P = (P_1,\ldots,P_K)$ of data $\mathbf{x}$ with $P_k = \{x_i : z_{ik} = 1\}$. The maximum likelihood estimator $\hat{\theta}_K$ is generally derived with the EM algorithm (see Section 1.2; Dempster et al., 1977; McLachlan and Krishnan, 1997). From a density estimation perspective, a classical way for choosing a mixture model is to select the model maximizing the integrated likelihood,

$$f(\mathbf{x};K) = \int f(\mathbf{x};\theta_K)\eta(\theta_K)d\theta_K,$$

$$f(\mathbf{x};\theta_K) = \prod_{i=1}^{n} f(x_i;\theta_K),$$

$\eta(\theta_K)$ being a weakly informative prior distribution on θ_K. It can be approximated with the BIC criterion (see Section 2.1.3)

$$\log f(\mathbf{x};K) \approx \log f(\mathbf{x};\hat{\theta}_K) - \frac{D_K}{2}\log n,$$

with $\hat{\theta}_K$ the maximum likelihood estimator and D_K the number of free parameters in the mixture model with K components. Numerical experiments (see for instance Roeder and Wasserman, 1995) show that BIC works well at a practical level for mixture models.

8.2.2 Choosing K From the Clustering Viewpoint

In the model-based clustering context, an alternative to the BIC criterion is the ICL criterion (see Section 2.1.4; Biernacki et al., 2000) which aims at maximizing the integrated likelihood of the complete data $(\mathbf{x},\mathbf{z})$

$$f(\mathbf{x},\mathbf{z};K) = \int_{\Theta_K} f(\mathbf{x},\mathbf{z};\theta_K)\eta(\theta_K)d\theta_K,$$

It can be approximated with a BIC-like approximation:

$$\log f(\mathbf{x},\mathbf{z};K) \approx \log f(\mathbf{x},\mathbf{z};\hat{\theta}^*_K) - \frac{D_K}{2}\log n$$

$$\hat{\theta}^*_K = \arg\max_{\theta_K} f(\mathbf{x},\mathbf{z};\theta_K).$$

But $\mathbf{z}$ and $\hat{\theta}_K^*$ are unknown. Arguing that $\hat{\theta}_K \approx \hat{\theta}_K^*$ if the mixture components are well separated for n large enough, Biernacki et al. (2000) replace $\hat{\theta}_K^*$ by $\hat{\theta}_K$ and the missing data $\mathbf{z}$ with $\hat{\mathbf{z}} = \text{MAP}(\hat{\theta}_K)$ defined by

$$\hat{z}_{ik} = \begin{cases} 1 & \text{if } \text{argmax}_\ell \, \tau_{i\ell}(\hat{\theta}_K) = k \\ 0 & \text{otherwise,} \end{cases}$$

$\tau_{ik}(\hat{\theta}_K)$ denoting the conditional probability that x_i arises from the kth mixture component ($1 \leq i \leq n$ and $1 \leq k \leq K$):

$$\tau_{ik} = \frac{\pi_k \phi(x_i; \omega_k)}{\sum_{\ell=1}^{K} \pi_\ell \phi(x_i; \omega_\ell)}.$$

Finally the ICL criterion is

$$\text{ICL}(K) = \log f(\mathbf{x}, \hat{\mathbf{z}}; \hat{\theta}_K) - \frac{D_K}{2} \log n.$$

Roughly speaking ICL is the criterion BIC penalized by the estimated mean entropy

$$\text{ENT}(K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} \tau_{ik}(\hat{\theta}_K) \log \tau_{ik}(\hat{\theta}_K) \geq 0.$$

Because of this additional entropy term, ICL favors values of K giving rise to partitioning the data with the greatest evidence. The derivation and approximations leading to ICL are questioned in Chapter 4. However, in practice, ICL appears to provide a stable and reliable estimate of K for real datasets and also for simulated datasets arising from mixtures with well separated components. ICL, which is not aiming at discovering the true number of mixture components, can underestimate the number of components for simulated data arising from mixture with poorly separated components (see Biernacki et al., 2000).

Remark that, for a given number of components K and a parameter θ_K, the class of each observation x_i is assigned according to the MAP rule defined above.

8.3 A Particular Clustering Selection Criterion

Suppose that the problem is to classify observations described with vectors $\mathbf{x}$'s. But, in addition, a known classification $\mathbf{u}$, associated to a qualitative variable not directly related to the variables defining the vectors $\mathbf{x}$'s, is defined on the population. Relating the classification $\mathbf{z}$ and the classification $\mathbf{u}$ could be of interest to get a suggestive and simple interpretation of the classification $\mathbf{z}$. With this purpose in mind, it is possible to define a penalized likelihood criterion which selects a model providing a good compromise between the mixture model fit and its ability to lead to a clear classification of the observations well related to the external classification $\mathbf{u}$. Ideally, it is wished that $\mathbf{x}$ and $\mathbf{u}$ are conditionally independent knowing $\mathbf{z}$, as holds if $\mathbf{u}$ can be written as a function of $\mathbf{z}$. Let us consider the following heuristics. The problem is to find the mixture model m maximizing the integrated completed likelihood

$$f(\mathbf{x}, \mathbf{u}, \mathbf{z}; m) = \int f(\mathbf{x}, \mathbf{u}, \mathbf{z}; \theta_m) \eta(\theta_m) d\theta_m.$$

Using a BIC-like approximation as in Biernacki et al. (2000),

$$\log f(\mathbf{x}, \mathbf{u}, \mathbf{z}; m) \approx \log f(\mathbf{x}, \mathbf{u}, \mathbf{z}; \hat{\theta}_m^*) - \frac{\nu_m}{2} \log n,$$

with

$$\hat{\theta}_m^* = \arg\max_{\theta_m} f(\mathbf{x}, \mathbf{u}, \mathbf{z}; \theta_m).$$

An approximation analogous to that leading to ICL is done: $\hat{\theta}_m^*$ is replaced by $\hat{\theta}_m$, the maximum likelihood estimator. $\mathbf{z}$ is then deduced from the MAP rule under this estimator. Assuming moreover that $\mathbf{x}$ and $\mathbf{u}$ are conditionally independent knowing $\mathbf{z}$, which should hold at least for mixtures with enough components, this yields

$$\log f(\mathbf{x}, \mathbf{u}, \mathbf{z}; m) \approx \log f(\mathbf{x}, \mathbf{z}; \hat{\theta}_m) + \log f(\mathbf{u} \mid \mathbf{z}; \hat{\theta}_m) - \frac{D_m}{2} \log n,$$

and the estimation of $\log f(\mathbf{u} \mid \mathbf{z}; \hat{\theta}_m)$ is derived from the contingency table $(n_{k\ell})$ relating the qualitative variables $\mathbf{u}$ and $\mathbf{z}$: for any $k \in \{1, \ldots, K\}$ and $\ell \in \{1, \ldots, U_{\max}\}$, $U_{\max}$ being the number of levels of the variable $\mathbf{u}$,

$$n_{k\ell} = \text{card}\big\{i : z_{ik} = 1 \text{ and } u_i = \ell\big\}.$$

Finally, this leads to the Supervised Integrated Completed Likelihood (SICL) criterion

$$SICL(m) = \text{ICL}(m) + \sum_{\ell=1}^{U_{max}} \sum_{k=1}^{K} n_{k\ell} \log \frac{n_{k\ell}}{n_{k\cdot}},$$

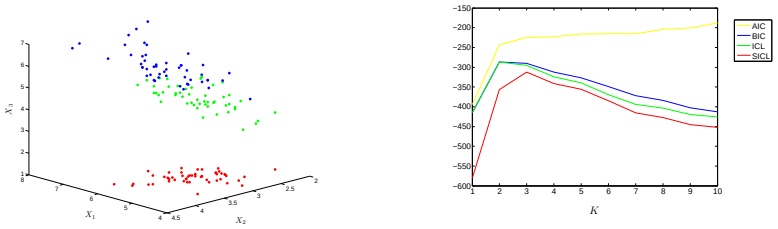
where $n_{k\cdot} = \sum_{\ell=1}^{K} n_{k\ell}$. The last additional term $\sum_{\ell=1}^{U_{max}} \sum_{k=1}^{K} n_{k\ell} \log \frac{n_{k\ell}}{n_{k\cdot}}$ is all the smaller that the link between the qualitative variables $\mathbf{u}$ and $\mathbf{z}$ is stronger.

8.4 Numerical Experiments

We first present simple applications to show that the SICL criterion is doing the job it is expected to do[1]. The first example is an application to the Iris dataset (Fisher, 1936) which consists of 150 observations of four measurements ($\mathbf{x}$) for three species of Iris ($\mathbf{u}$). Those data are depicted in Figure 8.1(a) and the variations of criteria BIC, ICL and SICL in function of K are provided in Figure 8.1(b). While BIC and ICL choose two classes, SICL selects the three-component mixture solution which is closely related to the species of Iris, as attested by the contingency table between the two partitions (Table 8.1).

For the second experiment, we simulated 200 observations from a Gaussian mixture in $\mathbb{R}^2$ depicted in Figure 8.2(a) and the variable $\mathbf{u}$ corresponds exactly to the mixture component from which each observation arises. Diagonal mixture models (i.e. with diagonal variance matrices) are fitted. The variations of the criteria BIC, ICL and SICL in function of K are provided in Figure 8.2(b). We repeated this experiment with 100 different simulated datasets. BIC almost always recovers the four Gaussian components, while ICL almost always selects three because of the two very overlapping

[1]Details on the simulation settings and the applied algorithms may be found in Section A.2.



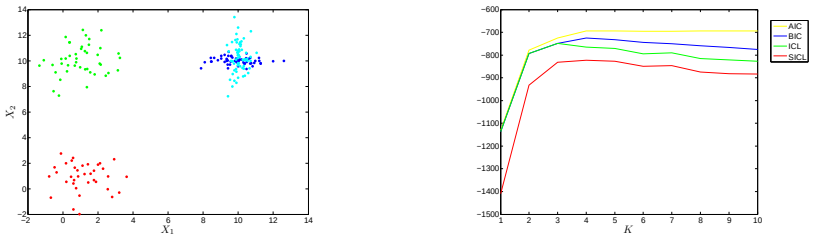
(a) Iris dataset (b) AIC, BIC, ICL, SICL vs. K

Figure 8.1: Comparison of AIC, BIC, ICL and SICL to choose K for the Iris dataset

Species \ k	1	2	3
Setosa	0	50	0
Versicolor	45	0	5
Virginica	0	0	50

Table 8.1: Iris data. Contingency table between the "species" variable and the classes derived from the three-component mixture.

ones (the "cross"). Since the solution obtained through MLE with the four-component mixture model yields classes nicely related to the considered $\mathbf{u}$ classes, SICL favors the four-component solution more than ICL does. But since it also takes the overlapping into account, it still selects the three-component model about half of the times (56 times out of 100 in our experiments), and selects the four-component model in almost all the remaining cases (40 out of 100). Figure 8.2(b) illustrates that the choice of SICL is not clear.

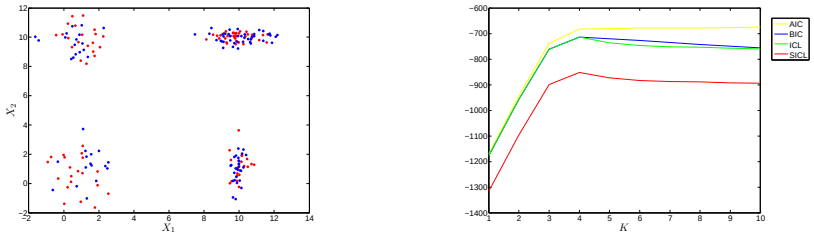


(a) "Cross" dataset (b) BIC, ICL, SICL in function of K

Figure 8.2: Comparison of BIC, ICL and SICL to choose K for the "Cross" dataset

In the two next experiments, we illustrate that SICL is not harmful when $\mathbf{u}$ cannot be related with any classification yielded by one of the mixture distributions. At first, we consider a situation where $\mathbf{u}$ is a two-class partition which has no link at all with a four-component mixture data. In Figure 8.3(a) the classes of $\mathbf{u}$ are in red and in blue. As is apparent from Figure 8.3(b), SICL does not change the solution $K = 4$ provided by BIC and ICL.

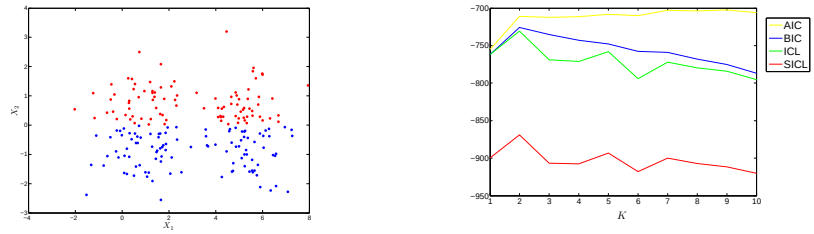
Then we consider a two-component mixture and a two-class $\mathbf{u}$ partition "orthogonal"



(a) Simulated dataset (b) BIC, ICL, SICL in function of K

Figure 8.3: Comparison of BIC, ICL and SICL to choose K for this simulated dataset

to this mixture. In Figure 8.4(a) the classes of $\mathbf{u}$ are in red and in blue. As is apparent from Figure 8.4(b), SICL does not change the solution $K = 2$ provided by BIC and ICL despite this solution has no link at all with the $\mathbf{u}$ classes.



(a) Simulated dataset (b) BIC, ICL, SICL in function of K

Figure 8.4: Comparison of BIC, ICL and SICL to choose K for this Simulated dataset

8.4.1 Real Dataset: Teachers Professional Development

The data arise from an international survey about teachers' views on the opportunities for their learning and professional development at workplace in Finland, Portugal and Serbia. Only data for teachers in Portugal are considered.

The dataset consists of discrete data background informations about the teachers and their working conditions on the one hand, and quantitative data which are their answers to a questionnaire. The original dataset involves 252 teachers, but some data are missing. Missing data among the background information variables do not cause any trouble, as there are not too many. But we handle missing data among the quantitative variables by removing the corresponding individuals from the study. To keep an interesting sample size, we first removed some variables with many missing data, and then the individuals with at least one missing value among the remaining quantitative variables. A Principal Component Analysis highlights that one of the individuals is an outlier: it has been removed from the study. The actual dataset is then composed of 190 individuals for which are available:

- Nine background discrete variables, with some missing data:

  - gender;
  - age (by classes);
  - academic qualification;
  - years of experience (by classes);
  - years of experience at current school (by classes);
  - level of teaching;
  - discipline of teaching;
  - school type (rural, suburban, urban);
  - number of inhabitants in the municipality (by classes).

- Twelve quantitative data corresponding to the questionnaire answers, with no missing data:
  - two variables about "opportunities for learning";
  - five variables about "professional development";
  - five variables about "motives".

The quantitative variables are scales with common range and spherical Gaussian mixture models (i.e. with variance matrices proportional to the identity matrix) with equal variance matrices[2] have been fitted through the EM algorithm with the MIXMOD software (Biernacki et al., 2006) to the obtained data for numbers of components K between one and ten.

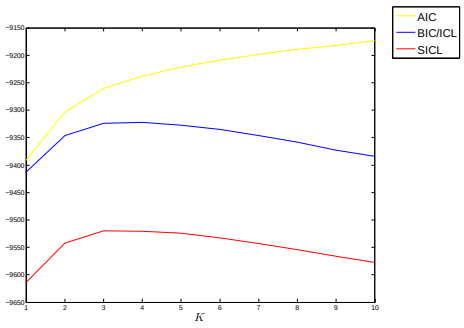


Figure 8.5: BIC, ICL, SICL (for the "school type" quantitative variable) in function of K.

AIC selects ten components; BIC and ICL select four components. Let us however remark that the selection of the two last criteria is not very clear: from Figure 8.5, the three and four-component solutions have values very close from each other.

SICL does not select the same number of components, depending on the involved qualitative variable. It selects four components, as BIC or ICL, for all qualitative variables but the eighth: "school type", in which case it selects three components. This is an

[2]Those models are denoted by $[p_k\_L\_I]$ in Celeux and Govaert (1995).

interesting result: this illustrates that the classes yielded by the Gaussian components of the three-component solution must be related to the "school type" classes. This is no surprise that this is a relevant variable. The study of the contingency tables (see Table 8.2, Table 8.3 and Table 8.4) confirms that the three-component classes are better related than the four-component classes to the three classes deduced from the "school type" variable. One of the Gaussian components can be regarded as reflecting the urban school type, whereas an other can be regarded as a "suburban" component. The proportion of rural schools, not surprisingly, is almost the same in any component (see Table 8.4).

School Type \ k	1	2	3
Rural	14	9	18
Suburban	32	20	16
Urban	25	17	38

Table 8.2: Teachers dataset. Contingency table between the "school type" external variable and the classes derived from the three-component mixture.

School Type \ k	1	2	3	4
Rural	10	9	10	12
Suburban	22	18	15	13
Urban	19	16	18	27

Table 8.3: Teachers dataset. Contingency table for the education data between the "school type" external variable and the classes derived from the four-component mixture.

Interestingly, SICL helps choosing between the three and four-component solutions by relating the corresponding classifications to the external variable as it is relevant.

8.5 Discussion

The criterion SICL has been conceived in the model-based clustering context to choose a sensible number of classes possibly well related to a qualitative variable of interest which is not entering into the variables used to design the classification. This criterion can be useful to draw attention to a well-grounded classification related to this external qualitative variable. It is an example of model selection criterion taking account of the modeler purpose to choose a useful and stable model. We think that this SICL criterion could have many fruitful applications as illustrated in the real data example for an education dataset.

School Type \ k	1	2	3	Any
Rural	0.20	0.20	0.25	0.22
Suburban	0.45	0.43	0.22	0.36
Urban	0.35	0.37	0.53	0.42
Component Size	71	46	72	189

(a)

School Type \ k	1	2	3	Class Size
Rural	0.34	0.22	0.44	41
Suburban	0.47	0.29	0.24	68
Urban	0.31	0.21	0.48	80
Any	0.38	0.24	0.38	189

(b)

School Type \ k	1	2	3	4	Any
Rural	0.20	0.21	0.23	0.23	0.22
Suburban	0.43	0.42	0.35	0.25	0.36
Urban	0.37	0.37	0.42	0.52	0.42
Component Size	51	43	43	52	189

(c)

School Type \ k	1	2	3	4	Class Size
Rural	0.24	0.22	0.24	0.30	41
Suburban	0.32	0.27	0.22	0.19	68
Urban	0.24	0.20	0.22	0.34	80
Any	0.27	0.23	0.23	0.28	189

(d)

Table 8.4: Teachers dataset. Frequency contingency tables with the "school type" external qualitative variable, for the three-component (a and b) and four-component (c and d) mixture classes. Proportion of each class within each Gaussian component (a and c) and distribution of each class among the Gaussian components (b and d).

Conclusion and Perspectives

In conclusion, let us stress that clustering is a challenging problem which definitely deserves the interest it generates among the statistical — and others — community: it answers a real need of many applications in various areas and is quite a stimulating research field. Probably the main unanswered question is that of a precise definition — in statistical terms — of what is expected to be done: as a matter of fact there is no clear consensus yet about the objective, even in the case where the data distribution would be known.

The theoretical study of ICL is an interesting and promising task from this point of view: since it seems to meet a notion of what people expect to be a class, the idea is that understanding what it does is a first step to understand what a class should be and to define an explicit aim for clustering. The usual model-based clustering approach, based on the maximum likelihood estimator and criteria such as BIC, provides an interesting methodology but the statistical aim is not easy to identify. This is illustrated by the discussion about BIC, ICL (with MLE), and mixtures of mixtures: it is difficult to choose between a good fit, which seems necessary to get a solid ground for the study, and a relevant number of classes. Defining the new L_{cc} contrast enables to recast the clustering question in a fully defined statistical framework. This is driving the logic underlying ICL to its conclusion. The choice of the contrast may still be discussed, but it provides an objective to get components both reasonable with respect to the data distribution and which yield an interesting, firm classification. Then the statistical study may be driven with usual tools: Chapter 4 illustrates how the study of this contrast is analogous to the study of the usual likelihood. The same ideas employed fruitfully with the usual likelihood contrast for density estimation may be applied to the new contrast: for example, we saw in Chapter 4 that the consistent penalized criteria correspond to the same range of penalties than in the likelihood framework, and that the slope heuristics may be straightforwardly adapted — though not theoretically validated with this new contrast yet. Another possibility which should probably be considered is cross-validation, as Smyth (2000) does with the observed likelihood. We learnt about the works of Perry (2009) about cross-validation for unsupervised learning, and notably for choosing the number of axes to keep when driving a Principal Component Analysis, while writing this conclusion. Actually cross-validation, instead of penalized criteria, may straightforwardly be adapted to the problem of choosing the number of components for clustering with the L_{cc} contrast considered. This would probably be worth further studying this possibility. Finally ICL is better understood as an approximation of L_{cc}-ICL. But the link and differences between both criteria should be further studied, at least practically. It has been mentioned that ICL offers the advantage of being more easily computed. This may be decisive in some settings, but the price of this needs be better understood. The provided simulations show that the MLE and MLccE estimators may be quite different at times, but do not really highlight situations where both model selection criteria would sensibly differ. Such situations should be studied.

This raises the question of the computation of the MLccE estimator. Solutions are provided, which enable to compute it in the considered situations. Let us highlight that the Km1 initialization method (Section 5.1.3) highly improves the solutions, and that we also advocate its use for the usual MLE computation. But those practical methods are quite greedy: the computation time for the reported simulations is counted in hours, and even in days for an experiment with a hundred datasets, run on a modern computer. This computation cost may get prohibitive for some applications, notably as the number

of considered classes is high. New and more efficient algorithms may be to appear in the forthcoming years, and perhaps the solutions considered in this thesis may still be markedly improved. Anyhow improvements of the possibilities of such methods follow mechanically from the progress of computers capabilities.

The question of how to impose bounds on the parameter space seems to be satisfyingly answered from a practical point of view (Section 5.1.4). Let us however highlight that the choice of the variances lower bound (typically $\sigma_0^2 = \det_{\min}$ in a multivariate setting) is not merely a technical assumption. This is essential in a mixture model framework: we mentioned that the contrast (the observed likelihood and the conditional classification likelihood as well) is not upper-bounded if the component variances are not lower-bounded. A sensible choice of the variance bound is then essential for a relevant clustering analysis. This choice may be decisive: a small value may favor spurious solutions and prevent relevant solutions while a too large value may hide the structure of interesting classes. Let us illustrate to what extent it may affect the study by the following informal remarks about the slope heuristics applied to the $\mathrm{L_{cc}}$ contrast. They apply likewise to the usual likelihood contrast. First remark that, for a small enough lower-bound on the variances and a great number of components $K-1$, it may be that any supplementary component is fitted such that its mean equals an observation value x_{i_0}, and its covariance matrix has minimal determinant value σ_0^2. Let us write (recall definitions from Chapter 4.2.1)

$$\begin{aligned}\log \mathrm{L_{cc}}(\hat{\theta}_K) &= \sum_{i\neq i_0}\sum_{k=1}^{K} \hat{\tau}_{ik} \log \hat{\pi}_k \hat{\phi}_k(x_i) + \sum_{k=1}^{K} \hat{\tau}_{i_0 k} \log \hat{\pi}_k \hat{\phi}_k(x_{i_0}) \\ &\approx \sum_{i\neq i_0}\sum_{k=1}^{K-1} \hat{\tau}_{ik} \log \hat{\pi}_k \hat{\phi}_k(x_i) + \underbrace{\hat{\tau}_{i_0 K} \log \hat{\pi}_K \hat{\phi}_K(x_{i_0})}_{\approx \log \frac{1}{n}(\frac{1}{2\pi})^{\frac{d}{2}}\frac{1}{\sigma_0}}.\end{aligned}$$

Should this occur each time a supplementary component is added, then the relation between $\mathrm{L}(\hat{\theta}_K)$ and K is mechanically roughly linear, with a slope depending on σ_0. We do not know yet whether this may be the basis of a validation of the slope heuristics in this framework, which would then be quite simple, or whether on the contrary, this reveals a phenomenon which may mask the slope heuristics, what we are afraid of. Remark that the method deduced from the slope heuristics should anyway provide a relevant number of components in such a situation with a sensible value of σ_0 since the phenomenon is only expected to occur for values K larger than the relevant number of classes. This should be further studied, with assumptions being precisely specified and approximations made rigorous. But anyhow, this illustrates how decisive the choice of σ_0^2 can be. Insights for this choice are given in Section 5.1.4 and could presumably be improved.

The results of Chapter 4 could be extended to get an oracle inequality. This might notably help to better understand what the "good" measure of the complexity of a mixture model is. As a matter of fact this would involve a finer control of the bracketing entropies, from which might emerge a complexity measure which we are not convinced it would be the number of free parameters of the model, as already discussed in the conclusion of Chapter 4 (see page 146).

The slope heuristics provides a promising method to calibrate the penalties of efficient criteria, based on the data, whatever the contrast (L or $\mathrm{L_{cc}}$). We presented two

approaches for its application: the dimension jump and the data-driven slope estimation. Hopefully the Matlab package introduced in Chapter 5 shall contribute to its use. This would notably enable to get more material to compare both practical approaches. As a matter of fact, the reported simulations show that the data-driven slope estimation may be more reliable and less computationally expensive than the dimension jump — at least in some situations.

Finally, the methodology reported in Chapter 7 seems to be an interesting approach for model-based clustering. As we advocate in this chapter, users of such a method should consider the whole hierarchy rather than a single solution. However it would presumably be an interesting further work to link this work with those dedicated to the choice of the number of classes.

Appendix A

Appendix

A.1 Theorem 7.11 in Massart (2007)

The following Theorem of Massart (2007) is referred to several time in this thesis. Please refer to Massart (2007) for discussion and proofs. Only the notation is changed slightly.

Recall the definition of the Hellinger distance between two densities t and u:

$$d_{\text{hel}}(u,t) = \frac{1}{\sqrt{2}}\|\sqrt{t} - \sqrt{u}\|_2.$$

Recall the definition of the L_2 bracketing entropy of $\sqrt{S} \subset L_2$ from Section 4.3.2, denoted by $H(\varepsilon, \sqrt{S}) = \mathcal{E}_{[]}(\varepsilon, \sqrt{S}, \|\cdot\|_2)$.

Let $(S_m)_{m\in\mathcal{M}}$ be a given collection of models.

Now, a function ϕ_m on $\mathbb{R}^+$ is considered such that ϕ_m is nondecreasing, $x \mapsto \frac{\phi_m(x)}{x}$ is nonincreasing on $]0, +\infty[$ and for every $\sigma \in \mathbb{R}^+$ and every $u \in S_m$

$$\int_0^\sigma \sqrt{H\Big(x, \sqrt{S_m(u,\sigma)}\Big)}dx \leq \phi_m(x),$$

where $S_m(u,\sigma) = \{t \in S_m : \|\sqrt{t} - \sqrt{u}\|_2 \leq \sigma\}$.

The following separability assumption is made: There exists some countable subset S'_m of S_m and a set $\Omega \subset \mathbb{R}^d$ with $\lambda(\mathbb{R}^d \backslash \Omega) = 0$ such that for every $t \in S_m$, there exists some sequence $(t_k)_{k\geq 1}$ of elements of S'_m such that for every $x \in \Omega$, $\log t_k(x)$ tends to $\log t(x)$ as k tends to infinity.

Then, the following Theorem holds.

Theorem 11 (Theorem 7.11 in Massart, 2007)
Let $X_1, \ldots, X_n$ be i.i.d. random variables with unknown density s with respect to some positive measure λ. Let $\{S_m\}_{m\in\mathcal{M}}$ be some at most countable collection of models, where for each $m \in \mathcal{M}$, the elements of S_m are assumed to be probability densities with respect to λ and S_m fulfills the separability assumption above. We consider a corresponding collection of ρ-MLEs $(\hat{s}_m)_{m\in\mathcal{M}}$ which means that for every $m \in \mathcal{M}$[1]

$$\mathbb{P}_n\big[-\log \hat{s}_m\big] \leq \inf_{t\in S_m} \mathbb{P}_n\big[-\log t\big] + \rho.$$

Let $\{x_m\}_{m\in\mathcal{M}}$ be some family of nonnegative numbers such that

$$\sum_{m\in\mathcal{M}} e^{-x_m} = \Sigma < \infty,$$

and for every $m \in \mathcal{M}$ considering ϕ_m which fulfills the property stated above, define σ_m as the unique solution of the equation

$$\phi_m(\sigma) = \sqrt{n}\sigma^2.$$

Let pen $: \mathcal{M} \to \mathbb{R}^+$ *and consider the penalized log likelihood criterion*

$$\text{crit}(m) = \mathbb{P}_n\big[-\log \hat{s}_m\big] + \text{pen}(m).$$

[1] With: $\forall g$ measurable, $\mathbb{P}_n g = \frac{1}{n}\sum_{i=1}^n g(X_i)$.

Then there exists some absolute constants κ *and* C *such that whenever*

$$\text{pen}(m) \geq \kappa\Big(\sigma_m^2 + \frac{x_m}{n}\Big) \text{ for every } m \in \mathcal{M},$$

some random variable $\hat{m}$ *minimizing* crit *over* $\mathcal{M}$ *does exist and moreover, whatever the density* s

$$\mathbb{E}_s\left[d_{hel}^2\big(s, \hat{s}_{\hat{m}}\big)\right] \leq C\Bigg(\inf_{m\in\mathcal{M}}\big(d_{KL}(s, S_m) + \text{pen}(m)\big) + \rho + \frac{\Sigma}{n}\Bigg),$$

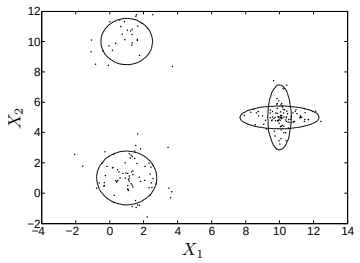
where, for every $m \in \mathcal{M}$, $d_{KL}(s, S_m) = \inf_{t\in S_m} d_{KL}(s, t)$.

A.2 Simulation Settings

Let us give in this section the precise settings of the simulation studies reported in the thesis: in Section A.2.1 are given the parameters of the simulated datasets and the interested reader shall find the precise methodologies and tuning constants employed to run the algorithms in Section A.3.

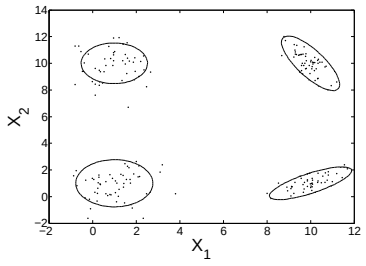
A.2.1 Simulated Datasets Settings

All simulated datasets are mixtures, and mostly Gaussian mixtures, which parameters are given now. A Gaussian distribution with parameters μ and Σ is denoted by $\mathcal{N}(\mu, \Sigma)$, while a uniform distribution over the set $C \subset \mathbb{R}^d$ is denoted by $\mathcal{U}(C)$. The rotation matrix with angle θ is denoted by R_θ : $R_\theta = \left(\begin{smallmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{smallmatrix}\right)$.



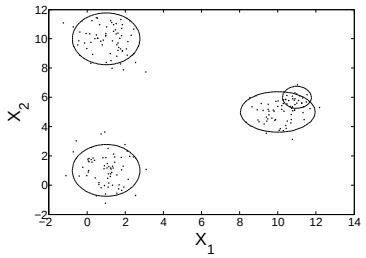
k	Mixing Proportion	Component Distribution
1	0.3	$\mathcal{N}\Big(\left(\begin{smallmatrix}1\\1\end{smallmatrix}\right), \left(\begin{smallmatrix}1 & 0\\0 & 1\end{smallmatrix}\right)\Big)$
2	0.2	$\mathcal{N}\Big(\left(\begin{smallmatrix}1\\10\end{smallmatrix}\right), \left(\begin{smallmatrix}1 & 0\\0 & 1\end{smallmatrix}\right)\Big)$
3	0.3	$\mathcal{N}\Big(\left(\begin{smallmatrix}10\\5\end{smallmatrix}\right), \left(\begin{smallmatrix}1 & 0\\0 & 0.1\end{smallmatrix}\right)\Big)$
4	0.2	$\mathcal{N}\Big(\left(\begin{smallmatrix}10\\5\end{smallmatrix}\right), \left(\begin{smallmatrix}0.1 & 0\\0 & 1\end{smallmatrix}\right)\Big)$

"Cross" Dataset (Section 3.3.3 and Section 4.4.6).
$n = 200$. This is a "diagonal" four-component Gaussian mixture (i.e. in the family denoted by $[pk\,\lambda k\,Bk]$ in Celeux and Govaert, 1995).



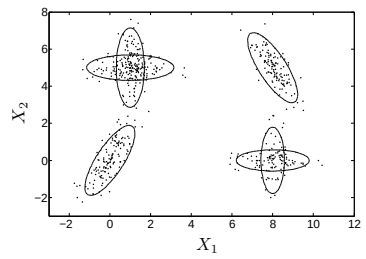
k	Mixing Proportion	Component Distribution
1	0.3	$\mathcal{N}\left(\binom{1}{1}, \left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)\right)$
2	0.2	$\mathcal{N}\left(\binom{1}{10}, \left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)\right)$
3	0.3	$\mathcal{N}\left(\binom{10}{10}, R_{-\frac{\pi}{3}}\left(\begin{smallmatrix}1 & 0\\ 0 & 0.1\end{smallmatrix}\right)R_{\frac{\pi}{3}}\right)$
4	0.2	$\mathcal{N}\left(\binom{10}{1}, R_{\frac{\pi}{6}}\left(\begin{smallmatrix}1 & 0\\ 0 & 0.1\end{smallmatrix}\right)R_{-\frac{\pi}{6}}\right)$

"Misspecified" Experiments Dataset (Section 3.3.3 and Section 4.4.6). $n = 200$. This is a general four-component Gaussian mixture (i.e. in the family denoted by $[pk\,\lambda k\,Ck]$ in Celeux and Govaert, 1995).



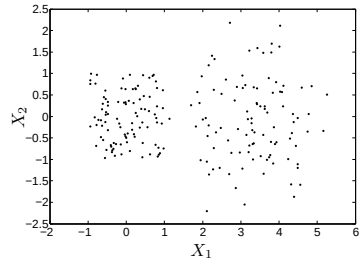
k	Mixing Proportion	Component Distribution
1	0.3	$\mathcal{N}\left(\binom{1}{1}, \left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)\right)$
2	0.3	$\mathcal{N}\left(\binom{1}{10}, \left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)\right)$
3	0.3	$\mathcal{N}\left(\binom{10}{5}, \left(\begin{smallmatrix}1 & 0\\ 0 & 0.5\end{smallmatrix}\right)\right)$
4	0.1	$\mathcal{N}\left(\binom{11}{6}, \left(\begin{smallmatrix}0.1 & 0\\ 0 & 0.1\end{smallmatrix}\right)\right)$

"Distorted Component" Experiment Dataset (Section 4.4.6). $n = 200$. This is a "diagonal" four-component Gaussian mixture (i.e. in the family denoted by $[pk\,\lambda k\,Bk]$ in Celeux and Govaert, 1995).



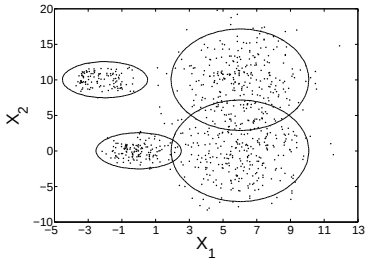
k	Mixing Proportion	Component Distribution
1	0.2	$\mathcal{N}\left(\binom{0}{0}, R_{\frac{\pi}{3}}\left(\begin{smallmatrix}1 & 0\\ 0 & 0.1\end{smallmatrix}\right)R_{-\frac{\pi}{3}}\right)$
2	0.2	$\mathcal{N}\left(\binom{8}{5}, R_{-\frac{\pi}{3}}\left(\begin{smallmatrix}1 & 0\\ 0 & 0.1\end{smallmatrix}\right)R_{\frac{\pi}{3}}\right)$
3	0.2	$\mathcal{N}\left(\binom{1}{5}, \left(\begin{smallmatrix}0.1 & 0\\ 0 & 1\end{smallmatrix}\right)\right)$
4	0.2	$\mathcal{N}\left(\binom{1}{5}, \left(\begin{smallmatrix}1 & 0\\ 0 & 0.1\end{smallmatrix}\right)\right)$
5	0.1	$\mathcal{N}\left(\binom{8}{0}, \left(\begin{smallmatrix}0.1 & 0\\ 0 & 1\end{smallmatrix}\right)\right)$
6	0.1	$\mathcal{N}\left(\binom{8}{0}, \left(\begin{smallmatrix}1 & 0\\ 0 & 0.1\end{smallmatrix}\right)\right)$

"Overlapping Components" Dataset (Section 7.4.1 and Section 7.4.2). $n = 600$. This is a general six-component Gaussian mixture (i.e. in the family denoted by $[pk\,\lambda k\,Ck]$ in Celeux and Govaert, 1995).



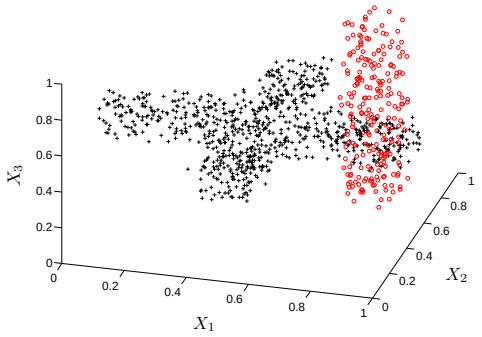
k	Mixing Proportion	Component Distribution
1	0.5	$\mathcal{N}\left(\begin{pmatrix}3.3\\0\end{pmatrix}, \begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$
2	0.5	$\mathcal{U}\left([-1;1]\times[-1;1]\right)$

"Circle-Square" Dataset (Section 7.4.3). $n = 200$.

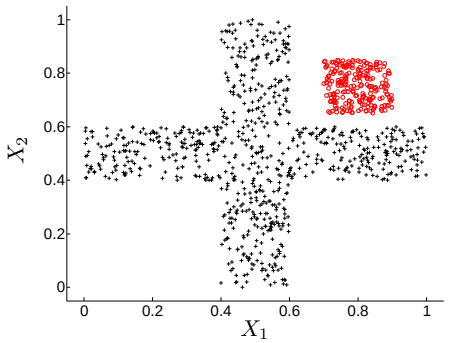


k	Mixing Proportion	Component Distribution
1	0.15	$\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$
2	0.35	$\mathcal{N}\left(\begin{pmatrix}6\\0\end{pmatrix}, \begin{pmatrix}4&0\\0&12.25\end{pmatrix}\right)$
3	0.35	$\mathcal{N}\left(\begin{pmatrix}6\\10\end{pmatrix}, \begin{pmatrix}4&0\\0&12.25\end{pmatrix}\right)$
4	0.15	$\mathcal{N}\left(\begin{pmatrix}-2\\10\end{pmatrix}, \begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$

"Gaussian Mixture" Dataset (Section 7.4.4). $n = 800$.



3D View

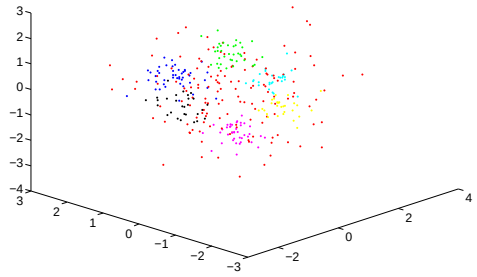


View from above

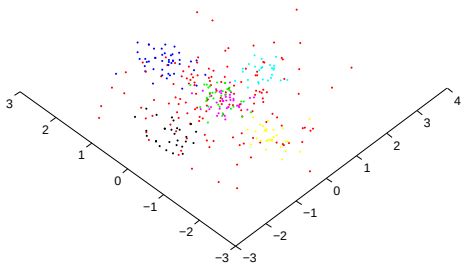
k	Mixing Proportion	Component Distribution
1	0.75	$\mathcal{U}\left([0;1]\times[0.4;0.6]\times[0.4;0.6]\cup[0.4;0.6]\times[0;1]\times[0.4;0.6]\right)$
2	0.25	$\mathcal{U}\left([0.7;0.9]\times[0.65;0.85]\times[0;1]\right)$

"3D Uniform Cross" Dataset (Section 7.4.4). $n = 300$ or $n = 1000$.

"Bubbles" Dataset (Section 5.2.4) For the sake of readability, let us first define the "bubble" distribution , defined as a spherical Gaussian mixture distribution in $\mathbb{R}^3$ and denoted by $\mathcal{B}$:



3D View

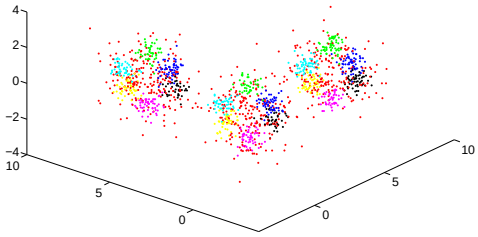


View From Above

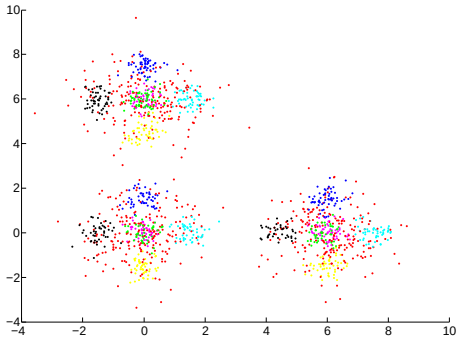
k	Mixing Proportion	Component Distribution
1	0.4	$\mathcal{N}\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}\right)$
2	0.1	$\mathcal{N}\left(\begin{pmatrix}0\\0\\1.5\end{pmatrix}, \begin{pmatrix}0.1&0&0\\0&0.1&0\\0&0&0.1\end{pmatrix}\right)$
3	0.1	$\mathcal{N}\left(\begin{pmatrix}0\\1.5\\0\end{pmatrix}, \begin{pmatrix}0.1&0&0\\0&0.1&0\\0&0&0.1\end{pmatrix}\right)$
4	0.1	$\mathcal{N}\left(\begin{pmatrix}1.5\\0\\0\end{pmatrix}, \begin{pmatrix}0.1&0&0\\0&0.1&0\\0&0&0.1\end{pmatrix}\right)$
5	0.1	$\mathcal{N}\left(\begin{pmatrix}0\\0\\-1.5\end{pmatrix}, \begin{pmatrix}0.1&0&0\\0&0.1&0\\0&0&0.1\end{pmatrix}\right)$
6	0.1	$\mathcal{N}\left(\begin{pmatrix}0\\-1.5\\0\end{pmatrix}, \begin{pmatrix}0.1&0&0\\0&0.1&0\\0&0&0.1\end{pmatrix}\right)$
7	0.1	$\mathcal{N}\left(\begin{pmatrix}-1.5\\0\\0\end{pmatrix}, \begin{pmatrix}0.1&0&0\\0&0.1&0\\0&0&0.1\end{pmatrix}\right)$

"Bubble" Distribution, denoted by $\mathcal{B}$ $n = 1000$. This is a "spherical" seven-component Gaussian mixture (i.e. in the family denoted by $[pk\,\lambda k\,I]$ in Celeux and Govaert, 1995).

Now, the "bubbles" dataset is a mixture of three $\mathcal{B}$ distributions:



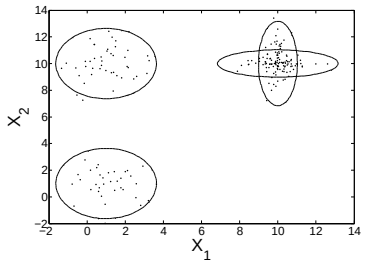
3D View



View from Above

k	Mixing Proportion	Component Distribution
1	0.3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} + \mathcal{B}$
2	0.3	$\begin{pmatrix} 6 \\ 0 \\ 0 \end{pmatrix} + \mathcal{B}$
3	0.3	$\begin{pmatrix} 0 \\ 6 \\ 0 \end{pmatrix} + \mathcal{B}$

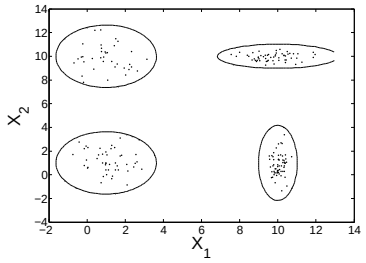
"Bubbles" Dataset.



k	Mixing Proportion	Component Distribution
1	0.2	$\mathcal{N}\left(\binom{1}{1}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$
2	0.2	$\mathcal{N}\left(\binom{1}{10}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$
3	0.3	$\mathcal{N}\left(\binom{10}{10}, \begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix}\right)$
4	0.3	$\mathcal{N}\left(\binom{10}{10}, \begin{pmatrix} 0.1 & 0 \\ 0 & 1 \end{pmatrix}\right)$

"Cross" Dataset (Section 8.4)

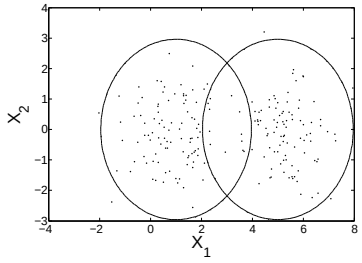
$n = 200$. This is a "diagonal" four-component Gaussian mixture (i.e. in the family denoted by $[pk\,\lambda k\,Bk]$ in Celeux and Govaert, 1995).



k	Mixing Proportion	Component Distribution
1	0.2	$\mathcal{N}\left(\binom{1}{1}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$
2	0.2	$\mathcal{N}\left(\binom{1}{10}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$
3	0.3	$\mathcal{N}\left(\binom{10}{10}, \begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix}\right)$
4	0.3	$\mathcal{N}\left(\binom{10}{1}, \begin{pmatrix} 0.1 & 0 \\ 0 & 1 \end{pmatrix}\right)$

Simulated Dataset of Section 8.4

$n = 200$. This is a "diagonal" four-component Gaussian mixture (i.e. in the family denoted by $[pk\,\lambda k\,Bk]$ in Celeux and Govaert, 1995).



k	Mixing Proportion	Component Distribution
1	0.5	$\mathcal{N}\left(\binom{1}{0}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$
2	0.5	$\mathcal{N}\left(\binom{5}{0}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$

Simulated Dataset of Section 8.4

$n = 200$. This is a "diagonal" two-component Gaussian mixture (i.e. in the family denoted by $[pk\,\lambda k\,Bk]$ in Celeux and Govaert, 1995).

A.3 Algorithms Settings

A.3.1 EM

The EM algorithm has been applied in this thesis mainly with Matlab and the MIXMOD software (Biernacki et al., 2006). The initialization procedures we involved differed — both by the involved methods and by the choice of the tuning constants — from those implemented in MIXMOD and have been introduced in Section 1.2.2 (Small_EM) and Section 5.1.3 (Km1). Let us give typical tuning constants that may be involved for these. They may differ a little from a simulation study to another, though.

Small_EM

- Number of random starts before each "short run" of EM : 25.
- Length of each short run of EM (for each of the 25 tries): 50 iterations.
- Length of the final long run of EM : 1000 iterations.

Km1

- Length of each short run of EM (for each one of the $K-1$ tries): 50 iterations.
- Length of the final long run of EM : 1000 iterations.

A.3.2 L_{cc}-EM

The notation below refers to the corresponding sections. This algorithm has been implemented and run with Matlab.

Let us first describe how each M step of the L_{cc}-EM algorithm (Section 5.1.1) is performed. Since the quantity $Q(\theta, \theta^j) - \mathrm{ENT}(\theta; X)$ has to be numerically optimized, the Matlab *fminsearch*[2] function is applied. This function must be provided an initialization parameter and the number of iterations of the function algorithm can be specified. It is denoted by iter_fmin in the sequel. Our L_{cc}-EM algorithm is then parameterized by the number of iterations iter_L_{cc}_EM, the number of iterations iter_fmin and the initialization parameter. Moreover, minimum accepted values for the covariance matrices determinants and the proportions may be specified and are imposed in this case through the method introduced in Section 5.1.4[3].

Let us now give typical tuning constants for the L_{cc}-EM algorithm initialization steps, which have been introduced in Section 5.1.2 (CEM) and Section 5.1.3 (Small_L_{cc}_EM, Km1). This may have to be adapted according to each particular situation.

[2]According to the Matlab Software documentation, "*fminsearch* uses the simplex search method of Lagarias et al. (1998). This is a direct search method that does not use numerical or analytic gradients." This is interesting since the gradient here is tough to be computed.

[3]In practice, it did not seem necessary to impose bounds on the proportions, and a typical value for the minimum accepted covariance determinant is 10^{-4} for the considered datasets when the dimension is 2.

CEM

- A 10-times run of CEM through MIXMOD (with 100 iterations of the algorithm and a standard Small_EM initialization) is performed.
- This first step is repeated 50 times, and the best solution (according to the L_{cc} values) is chosen.
- A long run of L_{cc}-EM with iter_fmin=100 is performed until one of the following stopping criteria is reached: either 100 iterations, or the increase of L_{cc} along an algorithm iteration is less than 10^{-2}.

Small_L_{cc}_EM

- When initializing a component at random, the probability to choose the mean uniformly at random in the range of the observations (instead of choosing it at random among the observations values) is set to be 10%.
- n_s has been chosen as $max(floor(\frac{nr}{K}), 5)$ (*floor* is a Matlab function which rounds its argument to the nearest integer less than or equal to it). nr is the size of the sample, or of the subsample in case a subsample is considered at this step.
- Number of random updates of a component chosen at random (followed by a single shot of L_{cc}-EM, with iter_fmin=100 each), before each "short run" of L_{cc}-EM (5 iterations with iter_fmin=1000) : 5.
- All this procedure is repeated 25 times and followed by a long run of L_{cc}-EM (10 iterations with iter_fmin=1000) .

Km1

- Length of each short run of L_{cc}-EM (for each one of the $K-1$ tries): 25 iterations, with iter_fmin=50.
- Length of the final long run of L_{cc}-EM : 10 iterations, with iter_fmin=1000.

Final L_{cc}-EM run Finally, a long run of 15 iterations (with iter_fmin=1000) initialized at the best solution obtained along the initialization steps, is performed.

Bibliography

Aitkin, M. and Rubin, D. (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society B*, 47(1):67–75.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings, 2nd Internat. Symp. on Information Theory*, pages 267–281.

Arlot, S. (2007). *Rééchantillonnage et sélection de modèles*. PhD thesis, University Paris XI.

Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research (JMLR)*, 10:245–279 (electronic).

Banfield, J. and Raftery, A. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803 – 821.

Baudry, J.-P., Celeux, G., and Marin, J.-M. (2008a). Selecting models focussing on the modeller's purpose. In *COMPSTAT 2008: Proceedings in Computational Statistics*, pages 337–348, Heidelberg. Physica-Verlag.

Baudry, J.-P., Raftery, A., Celeux, G., Lo, K., and Gottardo, R. (2008b). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*. To Appear.

Bickel, P. and Doksum, K. (2001). *Mathematical Statistics. Vol. I*. Prentice Hall, second edition.

Biernacki, C., Celeux, G., and Govaert, G. (1998). Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood. Research Report RR-3521, INRIA.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transaction on PAMI*, 22:719–725.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):567 – 575.

Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). Model-based cluster and discriminant analysis with the mixmod software. *Computational Statistics and Data Analysis*, 51(2):587–600.

Biernacki, C. and Govaert, G. (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics*, 29:451–457.

Biernacki, C. and Govaert, G. (1999). Choosing models in model-based clustering and discriminant analysis. *Journal of Statistical Computation and Simulation*, 64:49–71.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268.

Birgé, L. and Massart, P. (2006). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1-2):33–73.

Brinkman, R., Gasparetto, M., Lee, S.-J., Ribickas, A., Perkins, J., Janssen, W., Smiley, R., and Smith, C. (2007). High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biology of Blood and Marrow Transplantation*, 13:691–700.

Byers, S. D. and Raftery, A. E. (1998). Nearest neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association*, 93:577–584.

Caillerie, C. and Michel, B. (2009). Model selection for simplicial approximation. Research Report RR-6981, INRIA.

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2(1):73–82.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3):315–332.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781 – 793.

Chanda, K. (1954). A note on the consistency and maxima of the roots of likelihood equations. *Biometrika*, 41(1/2):56 – 61.

Chapelle, O., Schölkopf, B., and Zien, A. (2006). *Semi-supervised learning*. MIT press.

Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: Population mixtures and stationary arma processes. *Annals of Statistics*, 27(4):1178–1209.

Dasgupta, A. and Raftery, A. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93:294–302.

De Granville, C., Southerland, J., and Fagg, A. (2006). Learning grasp affordances through human demonstration. In *Proceedings of the International Conference on Development and Learning, electronically published*.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38.

Denis, M. and Molinari, N. (2009). Choix du nombre de noeuds en régression spline par l'heuristique des pentes. In *41èmes Journées de Statistique, SFdS, Bordeaux*, Bordeaux, France France.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

Fraley, C. and Raftery, A. (1998). How many clusters? Answers via model-based cluster analysis. *The Computer Journal*, 41:578–588.

Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the american statistical association*, 97:611–631.

Goutte, C., Hansen, L., Liptrot, M., and Rostrup, E. (2001). Feature-space clustering for fMRI meta-analysis. *Human Brain Mapping*, 13(3):165–183.

Hagenaars, J. A. and McCutcheon, A. L. (2002). *Applied Latent Class Analysis*. Cambridge University Press, Cambridge, U.K.

Hamelryck, T., Kent, J. T., and Krogh, A. (2006). Sampling realistic protein conformations using local structural bias. *PLoS Comput Biol*, 2:e131.

Hartigan, J. A. and Hartigan, P. M. (1985). The dip test of unimodality. *Annals of Statistics*, 13:78–84.

Hastie, T. and Tibshirani, R. (1996). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society, Series B*, 58:155–176.

Hathaway, R. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics*, 13(2):795–800.

Hathaway, R. (1986). A constrained EM algorithm for univariate normal mixtures. *Journal of Statistical Computation and Simulation*, 23(3):211–230.

Hennig, C. (2004). Breakdown points for maximum likelihood estimators of location-scale mixtures. *The Annals of Statistics*, 32(4):1313 – 1340.

Hennig, C. (2009). Methods for merging Gaussian mixture components. Research Report 302, Department of Statistical Science, UCL.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium in mathematical statistics*, volume 1, pages 221–233.

Jörnsten, R. and Keleş, S. (2008). Mixture models with multiple levels, with application to the analysis of multifactor gene expression data. *Biostatistics*, 9:540–554.

Kass, R. and Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Keribin (2000). Consistent estimation of the order of mixture models. *Sankhya Series A*, 62(1):49–66.

Lagarias, J., Reeds, J., Wright, M., and Wright, P. (1998). Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9(1):112 – 147.

Lange, K. (1999). *Numerical Analysis for statisticians*. Springer-Verlag, New-York.

Lazarsfeld, P. F. (1950). The logical and mathematical foundations of latent structure analysis. In Stouffer, S. A., editor, *Measurement and Prediction, Volume IV of The American Soldier: Studies in Social Psychology in World War II*, page Chapter 10. Princeton University Press.

Le Cam, L. (1991). Maximum likelihood, an introduction. Research Report 168.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85(4):717–736.

Lepez, V. (2002). *Some estimation problems related to oil reserves*. PhD thesis, University Paris XI.

Lerasle, M. (2009). *Rééchantillonnage et sélection de modèles optimale pour l'estimation de la densité de variables indépendantes ou mélangeantes*. PhD thesis, Toulouse.

Leroux, M. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20:1350–1360.

Li, J. (2005). Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics*, 14:547–568.

Lindsay, B. G. (1983). The geometry of mixing likelihoods: A general theory. *The Annals of Statistics*, 11(1):86–94.

Lo, K., Brinkman, R., and Gottardo, R. (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry A*, 73:321–32.

Mallows, C. L. (1973). Some comments on CP. *Technometrics*, 15(4):661–675.

Massart, P. (2007). *Concentration Inequalities and Model Selection. École d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics*. Springer.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009). Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics and Data Analysis*, 53:3872 – 3882.

Maugis, C. and Michel, B. (2008). Slope heuristics for variable selection and clustering via Gaussian mixtures. Research Report 6550, INRIA.

Maugis, C. and Michel, B. (2009). A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM*. To appear.

McLachlan, G. and Krishnan, T. (1997). *The EM-algorithm and Extensions*. New York : Wiley.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. New York : Wiley.

McNicholas, P. and Murphy, T. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, 18(3):285–296.

McQuarrie, A. and Tsai, C. (1998). *Regression and time series model selection*. World Scientific.

Nishii, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate Analysis*, 27:392–403.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A*, 185:71–110.

Perry, P. O. (2009). *Cross-Validation for Unsupervised Learning*. PhD thesis, Stanford University.

Peters, B. and Walker, H. (1978). An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM Journal on Applied Mathematics*, 35(2):362–378.

Pigeau, A. and Gelgon, M. (2005). Building and tracking hierarchical geographical & temporal partitions for image collection management on mobile devices. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 141–150. ACM New York, NY, USA.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics*, 9(1):225–228.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195 – 239.

Ripley, B. (1995). *Pattern Recognition and Neural Network*. Cambridge University Press.

Roeder, K. and Wasserman, L. (1995). Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92:894–902.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6:461–464.

Scott, A. and Symons, M. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, 27:387–397.

Shlezinger, M. (1968). An algorithm for solving the selforganization problem. *Cybernetics*, 2.

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10:63–72.

Steele, R. J. (2002). *Importance sampling methods for inference in mixture models and missing data*. PhD thesis, Department of Statistics, University of Washington, Seattle, Wash.

Tantrum, J., Murua, A., and Stuetzle, W. (2003). Assessment and pruning of hierarchical model based clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 197–205, New York, NY. Association for Computing Machinery.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B*, 63:411–423.

Titterington, D. (1990). Some recent research in the analysis of mixture distributions. *Statistics*, 21(4):619–641.

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite mixture Distributions*. New York : Wiley.

Ueda, N. and Nakano, R. (1998). Deterministic annealing EM algorithm. *Neural Networks*, 11(2):271–282.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

Verzelen, N. (2009). Data-driven neighborhood selection of a Gaussian field. Research Report 6798, INRIA.

Villers, F. (2007). *Tests et sélection de modèles pour l'analyse de données protéomiques et transcriptomiques*. PhD thesis, University Paris XI.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20:595–601.

Walther, G. (2002). Detecting the presence of mixing with multiscale maximum likelihood. *Journal of the American Statistical Association*, 97:508–513.

Wang, N. and Raftery, A. (2002). Nearest neighbor variance estimation (NNVE): Robust covariance estimation via nearest neighbor cleaning (with discussion). *Journal of the American Statistical Association*, 97:994–1019.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Yakowitz, S. and Spragins, J. (1968). On the identifiability of finite mixtures. *Annals of Mathematical Statistics*, 39:209–214.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 92(4):937–950.

Sélection de modèle pour la classification non supervisée.
Choix du nombre de classes.

Résumé. Le cadre principal de cette thèse est la classification non supervisée, traitée par une approche statistique dans le cadre des modèles de mélange. Plus particulièrement, nous nous intéressons au choix du nombre de classes et au critère de sélection de modèle ICL. Une approche fructueuse de son étude théorique consiste à considérer un contraste adapté à la classification non supervisée : ce faisant, un nouvel estimateur ainsi que de nouveaux critères de sélection de modèle sont proposés et étudiés. Des solutions pratiques pour leur calcul s'accompagnent de retombées positives pour le calcul du maximum de vraisemblance dans les modèles de mélange. La méthode de l'heuristique de pente est appliquée pour la calibration des critères pénalisés considérés. Aussi les bases théoriques en sont-elles rappelées en détails, et deux approches pour son application sont étudiées.
Une autre approche de la classification non supervisée est considérée : chaque classe peut être modélisée elle-même par un mélange. Une méthode est proposée pour répondre notamment à la question du choix des composantes à regrouper.
Enfin, un critère est proposé pour permettre de lier le choix du nombre de composantes, lorsqu'il est identifié au nombre de classes, à une éventuelle classification externe connue a priori.

Mots-clés : Classification non supervisée, Sélection de modèle, Modèles de mélange, Vraisemblance classifiante, Critères pénalisés, BIC, ICL, Minimisation de contraste, Sélection de modèle data-driven, Heuristique de pente, EM, Point d'effondrement, Mélanges de mélanges, SICL.

Model Selection for Clustering.
Choosing the Number of Classes.

Abstract. The reported works take place in the statistical framework of model-based clustering. We particularly focus on choosing the number of classes and on the ICL model selection criterion. A fruitful approach for theoretically studying it consists of considering a contrast related to the clustering purpose. This entails the definition and study of a new estimator and new model selection criteria. Practical solutions are provided to compute them, which can also be applied to the computation of the usual maximum likelihood estimator within mixture models. The slope heuristics is applied to the calibration of the considered penalized criteria. Thus its theoretical bases are recalled in details and two approaches for its application are studied.
Another approach for model-based clustering is considered: each class itself may be modeled by a Gaussian mixture. A methodology is proposed, notably to tackle the question of which components have to be merged.
Finally a criterion is proposed, which enables to choose a number of components — when identified to the number of classes — related to a known external classification.

Keywords: Model-based clustering, Model selection, Mixture models, Classification likelihood, Penalized criteria, BIC, ICL, Contrast minimization, Data-driven model selection, Slope heuristics, EM, Breakdown point, Mixtures of mixtures, SICL.