Sélection de variables pour la classification non supervisée par mélanges gaussiens. Application à l'étude de données transcriptomes.

Cathy Maugis

N° d'ordre : 9233

UNIVERSITE PARIS-SUD 11
FACULTE DES SCIENCES D'ORSAY

THÈSE

Présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS-SUD 11

Spécialité : Mathématiques

par

Cathy MAUGIS

Sélection de variables pour la classification non supervisée par mélanges gaussiens. Application à l'étude de données transcriptomes.

Soutenue le 21 novembre 2008 après avis des rapporteurs

M. Yannick BARAUD
M. Mark van der LAAN

devant la Commission d'examen composée de :

M. Christophe AMBROISE	(Examinateur)
M. Sébastien AUBOURG	(Examinateur)
M. Yannick BARAUD	(Rapporteur)
M. Gilles CELEUX	(Directeur de thèse)
Mme. Marie-Laure MARTIN-MAGNIETTE	(Co-directrice de thèse)
M. Pascal MASSART	(Président du jury)

Remerciements

Je souhaite tout d'abord exprimer toute ma reconnaissance envers Gilles et Marie-Laure pour m'avoir encadrée durant mon stage de DEA et ma thèse. Je vous remercie pour votre confiance, votre soutien permanent, votre disponibilité et vos conseils qui m'ont permis de progresser durant ces années et de dépasser les moments de doute. Complémentaires, vous avez su me faire profiter de vos expériences et de votre goût pour la recherche. Je n'oublierai pas les nombreuses discussions mathématiques (ou non) très souvent accompagnées d'un bon café (j'ai trouvé le sucre !). Un écho du bâtiment 440 vous a qualifiés de « chefs quatre étoiles » ce qui résume parfaitement ces années de thèse auprès de vous. J'ai pris beaucoup de plaisir à travailler avec vous et j'espère que cette fin de thèse n'est pas un point final à notre collaboration.

J'adresse mes sincères remerciements à Yannick Baraud et Mark van der Laan pour l'intérêt qu'ils ont porté à mon travail en acceptant de rapporter ma thèse. Je remercie également Christophe Ambroise et Sébastien Aubourg de m'avoir fait l'honneur de leur présence dans le jury et Pascal Massart pour avoir accepté de le présider.

Cette seconde partie de thèse ne serait rien sans ma collaboration avec Bertrand. Je te remercie, « sympathique » collaborateur, d'avoir accepté de faire ce bout de chemin avec moi. Merci pour ta gentillesse, ta bonne humeur, ta persévérance et d'avoir accepté de te convertir aux mélanges gaussiens. On a su se soutenir et progresser pour finalement dompter ces calculs d'entropie rebels et ces pentes récalcitrantes. Mais ce n'est pas encore le moment de se déchausser, le col n'est pas encore atteint. Heureusement, nous avons pris notre « JPS » (Jean-Patrick) pour poursuivre une partie de cette ascension !

Je tiens également à remercier Sébastien Aubourg, Alain Lecharny, Sandra Pelletier et Jean-Pierre Renou, membres de l'URGV, pour leur patience face à mes ignorances en biologie et nos discussions pluridisciplinaires qui ont permis de donner un véritable aspect appliqué à cette thèse.

Effectuer ma thèse au sein de l'équipe de probabilités et de statistiques fut un réel plaisir. Je tiens en particulier à remercier Pascal Massart pour m'avoir permis de satisfaire mes envies d'applications en m'orientant vers Gilles et Marie-Laure pour cette thèse et d'avoir provoqué ma rencontre scientifique avec Bertrand. Je remercie également Vincent Rivoirard pour sa bonne humeur communicative et son soutien, Yves Misiti et Patrick Jakubowicz pour leur aide en informatique et Christine Keribin pour ses conseils en C++. Un merci tout particulier à Jean-Michel Marin (notre bayésien préféré reparti dans le sud) pour m'avoir supportée lors de mes nocturnes informatiques et pour tous les moments inoubliables de délires non maîtrisés au 440 ! Merci également à Valérie Lavigne, Marie-Carol Lopes et Katia Evrat pour leur aide, leur efficacité et leur gentillesse.

Que d'années inoubliables passées sur le campus d'Orsay. J'ai eu la chance de suivre les cours de professeurs exceptionnels, qui ont su me faire partager leur passion des mathématiques. Depuis

la Licence, ce fut également l'occasion de rencontres, d'amitiés et d'échanges. Merci à Juliette, Séverine, Stéphanie, Antoine, Maxime, Frédéric et tant d'autres qui me pardonneront (je l'espère) de ne pas les citer.

Pour m'avoir supportée au jour le jour, je tiens à remercier chaleureusement mes co-bureaux du 227, Sourour, Annalisa, Jérôme, Ayman, Antoine, Pierre, Nicolas et l'autruche Bernadette. Merci pour cette excellente ambiance, ces fous rires inoubliables, cette solidarité, ces dégustations culinaires, ces repas animés au CESFO, ... Merci également à tous les doctorants que j'ai côtoyés durant ces années. Je vous souhaite à tous bon vent !

Bien loin du monde de la recherche, je tiens à remercier mes parents, mes grands-parents, Marie-Odile et Bernard pour leur soutien constant. Je remercie également mes amis de longue date, Gaëlle, Sandrine et Grégory pour leur amitié tourangelle.

Enfin, merci Flavien pour ta patience, ton soutien inconditionnel et ton amour infini. Merci de prêter une oreille attentive à tout mon charabia de statisticienne alors que tu préfères parler de corps creux, de déchets ultimes et de CSD. J'espère que cette thèse est un grand pas vers le dépôt des valises pour enfin avancer au quotidien à tes côtés.

Table des matières

Présentation générale

Les progrès informatiques et le développement de technologies de pointe performantes, comme les puces à ADN, participent activement à la création de données de plus en plus complexes, décrites par un nombre croissant de variables. On rencontre ce phénomène dans de nombreux domaines comme l'informatique, les sciences sociales, la finance ou encore la biologie. Cette abondance de variables descriptives peut sembler un atout pour déterminer une bonne classification des données. Néanmoins, seul un sous-ensemble de ces variables descriptives peut contenir la structure d'intérêt pour la classification, les autres variables pouvant être redondantes, non significatives ou même néfastes pour classifier les données. Dans le but de maîtriser les informations nécessaires pour l'obtention d'une bonne classification des observations, la sélection des variables pertinentes parmi l'ensemble des variables disponibles doit être envisagée. On espère ainsi améliorer le processus de classification et faciliter l'interprétation de la classification obtenue.

En classification supervisée, une profusion de méthodes de sélection de variables sont disponibles. Une vue d'ensemble de ces méthodes est proposée par Guyon et Elisseeff (2003) dans un numéro spécial du Journal of Machine Learning Research consacré à la sélection de variables. Ces méthodes sont basées sur un critère mesurant l'adéquation à la vraie partition connue des données. Dans le cadre de la classification non supervisée, tout l'enjeu est de déterminer les labels inconnus des observations. Les publications sur les méthodes de sélection de variables dans ce contexte sont beaucoup moins nombreuses. La principale difficulté réside dans la construction d'un critère permettant de guider la sélection des variables mais ne pouvant pas être basé sur les labels. Les méthodes proposées sont généralement classées en *filter* ou *wrapper* selon la terminologie empruntée au cas supervisé, introduite par Kohavi and John (1997). Les méthodes dites *filter* traitent le problème de la sélection de variables indépendamment du processus de classification. Parmi ces méthodes, on peut citer les travaux de Dash et al. (2002) et Jouve et Nicoloyannis (2005). À l'opposé, les méthodes *wrapper* sont des procédures de sélection de variables incluses dans le processus de classification. Les premières méthodes *wrapper* ont été proposées pour des processus de classification basés sur des distances comme par exemple les procédures de Fowlkes et al. (1988) et de Brusco et Cradit (2001) pour la classification hiérarchique et l'algorithme des k plus proches voisins respectivement. Pour la classification basée sur des modèles, des méthodes de sélection de variables ont été développées plus récemment,

en particulier dans le cadre de la classification par mélanges gaussiens. Law et al. (2004) proposent d'évaluer l'importance des variables pour le processus de classification en introduisant la notion de *feature saliency* et utilisent un critère *Minimum Message Length*. Leur méthode est basée sur l'hypothèse que les variables non pertinentes pour la classification sont totalement indépendantes des variables significatives. Pour remédier à cette hypothèse restrictive, Raftery et Dean (2006b) proposent une méthode basée sur une modélisation plus réaliste du rôle des variables par rapport à la classification. Ils supposent que les variables non pertinentes sont expliquées par toutes les variables significatives pour la classification selon une régression linéaire. Leur méthode est fondée sur la comparaison de modèles via un facteur de Bayes. Dans ces deux articles, le problème de sélection de variables pour la classification est ramené à un problème de sélection de modèles. Les auteurs ont alors recours à des critères asymptotiques de sélection de variables.

Dans cette thèse, nous nous plaçons dans le cadre de la classification non supervisée par mélanges gaussiens comme Law et al. (2004) et Raftery et Dean (2006b). L'objectif des travaux exposés dans ce manuscrit consiste à proposer de nouvelles méthodes de sélection de variables dans ce contexte de classification non supervisée. Dans les deux parties de cette thèse, ce problème de sélection de variables par la classification est abordé comme un problème de sélection de modèles. Ces deux parties se distinguent en particulier par la nature du critère de sélection proposé.

La première partie est consacrée à la construction d'une procédure de sélection de variables généralisant celle proposée par Raftery et Dean (2006b). Elle est basée sur une modélisation plus réaliste du rôle des variables par rapport au processus de classification, affinant en particulier celle de Raftery et Dean (2006b). La sélection de modèles sous-jacente est traitée grâce un critère asymptotique de type BIC. D'un point de vue théorique, l'identifiabilité des modèles et la consistance du critère sont établies. Pour la mise en pratique de la procédure, un algorithme différent de celui de Raftery et Dean est développé.

La seconde partie est consacrée à la construction d'un critère pénalisé avec une approche non asymptotique pour la sélection de modèles de mélanges gaussiens multidimensionnels. La principale difficulté théorique résolue réside dans le contrôle des entropies à crochets des familles de mélanges gaussiens multidimensionnels considérés. Aboutissant à un critère pénalisé dépendant de constantes inconnues, une méthode heuristique est mise en œuvre pour la calibration des constantes.

Les travaux de cette thèse sont en particulier motivés par l'étude de données transcriptomes. Après le séquençage du génome de différentes espèces, les biologistes s'attachent maintenant à découvrir la ou les fonctions des gènes. Partant de l'hypothèse que des gènes co-exprimés sont liés fonctionnellement, le but est d'extraire des groupes de gènes co-exprimés à partir de données transcriptomes. Par la création de bases de données, ces données transcriptomes sont des ressources génomiques pour étudier le profil d'expression des gènes. Face au nombre croissant des expériences (variables) pour décrire les gènes, la sélection de variables est envisagée pour améliorer la classification et accroître son interprétation biologique.

Cette thèse est divisée en deux parties distinctes qui peuvent être lues indépendamment. Dans ces deux parties, les preuves et les calculs techniques sont reportés dès que possible en annexe du chapitre correspondant pour faciliter la lecture de ce manuscrit. Ces deux parties, exposées en anglais, sont précédées de deux chapitres introductifs en français consacrés respectivement aux mélanges gaussiens et à l'étude de l'expression des gènes.

Chapitre 1 : Classification non supervisée par mélanges gaussiens
Ce chapitre est consacré aux prérequis sur les mélanges gaussiens. Nous présentons la famille des 28 modèles de mélanges gaussiens que nous considérons par la suite. Puis nous décrivons l'algorithme EM utilisé pour estimer les paramètres d'un mélange. Les critères asymptotiques de sélection de modèles BIC et ICL sont ensuite rappelés. Les choix d'utilisation du logiciel MIXMOD pour les applications numériques sont finalement précisés.

Chapitre 2 : Etude de l'expression des gènes
Ce chapitre décrit les enjeux de l'étude des données transcriptomes. Après avoir rappelé le principe des puces à ADN, les méthodes de normalisation et d'analyse différentielle des données transcriptomes mises en place à l'URGV (Unité de Recherche en Génomique Végétale) sont explicitées. Enfin, des précisions sur les données transcriptomes étudiées dans cette thèse issue de l'URGV sont exposées ainsi que les enjeux qui ont en particulier motivé ce travail.

Partie I : Variable role modelling for Gaussian mixture clustering

Ce travail a été réalisé en collaboration avec mes deux directeurs de thèse, Gilles Celeux et Marie-Laure Martin-Magniette. Cette partie est motivée par la méthode de sélection de variables proposée par Raftery et Dean (2006b). Ces auteurs supposent que les variables non significatives pour la classification sont toutes liées aux variables pertinentes par une régression linéaire pour répondre à l'hypothèse trop restrictive de totale indépendance de Law et al. (2004). Néanmoins, leur méthode a tendance à surpénaliser des modèles, forçant les liens entre variables significatives et variables non significatives pour la classification. Nous proposons une amélioration de leur modélisation dans le chapitre 3 puis deux extensions, l'une affinant et rendant plus réaliste la modélisation du rôle des variables pour la classification, et l'autre pour l'étude de données avec valeurs manquantes au hasard. Toutes ces méthodes sont basées sur un problème de sélection de modèles où le critère de sélection est asymptotique de type BIC.

Chapitre 3 : Variable selection for clustering with Gaussian mixture models
Dans ce chapitre, nous autorisons les variables non significatives à ne dépendre que d'un sous-groupe de variables pertinentes pour la classification, selon une régression linéaire multivariée multidimensionnelle. Ce sous-groupe peut contenir toutes les variables significatives, revenant alors à la modélisation proposée par Raftery et Dean (2006b), mais peut également être vide, rendant les variables non significatives indépendantes comme supposé

par Law et al. (2004). Notre procédure de sélection de variables permet d'envisager le cas de variables blocs, motivé par l'étude des données transcriptomes. D'un point de vue théorique, l'identifiabilité des modèles est établie, prouvant ainsi celle de la modélisation de Raftery et Dean (2006b). La consistance du critère de sélection de variables, sous un modèle de mélange gaussien fixé est également démontrée. L'algorithme proposé, résultant de cette modélisation, imbrique deux algorithmes descendants de sélection de variables, l'un pour la régression et l'autre pour la classification. L'intérêt de cette méthode est mis en évidence par son application sur données simulées mais aussi sur un exemple de données transcriptomes.
Ce chapitre a fait l'objet d'une publication dans Biometrics :
Maugis, C., Celeux, G. and Martin-Magniette, M.-L. (2008) Variable Selection for Clustering with Gaussian Mixture Models. *Biometrics*, à paraître.

Chapitre 4 : Improving the variable roles in variable selection for clustering
La procédure de sélection de variables exposée dans le chapitre 3 ne résout pas tous les problèmes de surpénalisation des modèles. En particulier, si certaines variables non pertinentes pour la classification sont indépendantes alors que d'autres ont un lien linéaire avec certaines variables significatives, des coefficients de régression, déclarés libres dans le modèle, sont en réalité nuls. Cette surpénalisation touche principalement les modèles de mélanges gaussiens les plus parcimonieux. Pour résoudre ce problème, une amélioration de la modélisation du rôle des variables est proposée. Elle consiste à répartir les variables déclarées non significatives pour la classification selon deux catégories. Certaines de ces variables peuvent être dépendantes d'une partie, voire de la totalité, des variables significatives pour la classification alors que d'autres sont totalement indépendantes. Les propriétés théoriques d'identifiabilité et de consistance sont prolongées pour cette nouvelle modélisation et l'algorithme de sélection de variables est adapté. Enfin, les améliorations apportées par cette nouvelle procédure de sélection de variables sont illustrées sur des données simulées.

Chapitre 5 : Extension of the variable selection procedure for missing at random data
La procédure de sélection de variables proposée au chapitre 3 ne permet pas l'étude de données avec valeurs manquantes. Or ce problème de données manquantes est classiquement rencontré dans l'étude de données d'expression de gènes. Aussi, ce chapitre est consacré à l'extension de notre procédure de sélection de variables dans le cas de données avec valeurs supposées manquantes au hasard. Cette extension, évitant l'estimation des données manquantes, nécessite un calcul différent du critère de sélection de modèles et une nouvelle stratégie pour l'estimation des paramètres. L'algorithme de sélection de variables est modifié en conséquence. Cette nouvelle procédure de sélection de variables est comparée à la procédure initiale du chapitre 3 appliquée sur les données dont les valeurs manquantes sont préalablement estimées par une méthode d'imputation. Cette étape préliminaire d'affectation est classiquement employée pour l'étude des données d'expression de gènes, de nombreuses méthodes étant proposées depuis une dizaine d'années.

Cette partie s'achève par un chapitre de conclusion et perspectives (chapitre 6). Les algorithmes associés aux différentes versions de la procédure de sélection de variables étant gourmands en temps calculs, leur programmation a été effectuée en langage orienté objet C^{++}. Chacun de ces programmes est constitué d'environ 1000 lignes de code.

Partie II : Construction of a penalized likelihood criterion for variable selection in Gaussian mixture clustering with a non asymptotic point of view

Les travaux exposés dans cette partie sont le fruit d'une collaboration avec Bertrand Michel (Université Paris-Sud 11).

Nous considérons des mélanges gaussiens multivariés multidimensionnels pour reformuler notre problème de sélection de variables pour la classification en un problème de sélection de modèles. L'objectif est de construire un critère pénalisé selon un point de vue non asymptotique pour résoudre ce problème. Ce travail est novateur dans l'étude de la sélection de modèles par pénalisation pour des mélanges gaussiens multidimensionnels avec une approche non asymptotique, apparue il y a environ une dizaine d'années avec les travaux de Birgé et Massart (1997) et de Barron et al. (1999).

Chapitre 7 : A non asymptotic penalized likelihood criterion for specific Gaussian mixture model selection

Ce chapitre est consacré à la construction théorique d'un critère pénalisé par une approche non asymptotique dont les principes sont rappelés. Classiquement, la construction d'un tel critère revient au calcul explicite de l'espérance du processus empirique et à l'utilisation d'une inégalité de concentration. Dans notre contexte, une telle stratégie ne peut pas être envisagée, le contraste de Kullback-Leibler n'ayant pas un comportement linéaire sur les mélanges gaussiens. Nous avons donc recours à un théorème général de sélection de modèles pour l'estimation par maximum de vraisemblance proposé par Massart (2007). L'application de ce théorème nécessite de contrôler les entropies à crochets des familles de mélanges gaussiens multidimensionnels étudiés. Ghosal et van der Vaart (2001) et Genovese et Wasserman (2000) ont proposé des méthodes pour calculer de telles entropies pour des mélanges gaussiens unidimensionnels afin d'obtenir des vitesses de convergence en distance d'Hellinger pour l'estimation de densités. Nous nous sommes appuyés ici sur les travaux de Genovese et Wasserman (2000) pour ramener ce problème du contrôle des entropies à crochets de nos mélanges à celui des entropies à crochets des familles de densités gaussiennes qui composent les mélanges étudiés. La méthode proposée par Genovese et Wasserman (2000) a été étendue pour le cas de densités gaussiennes multidimensionnelles avec matrices de variance diagonales. Par contre, une nouvelle stratégie est présentée pour le cas de densités gaussiennes multidimensionnelles avec matrices de variance générales.

Chapitre 8 : Slope heuristics for a practical use of our penalized criterion
Le travail théorique proposé au chapitre 7 permet de construire un critère pénalisé non asymptotique et d'obtenir une inégalité oracle justifiant le comportement du modèle sélectionné par rapport à l'objectif donné. Néanmoins, ce critère, dépendant de constantes inconnues, ne peut être utilisé directement en pratique. La méthode dite « de la pente » proposée par Birgé et Massart (2006) est mise en œuvre pour calibrer la pénalité. Ce critère pénalisé est utilisé pour étudier des données simulées, un exemple de données transcriptomes ainsi qu'un ensemble de courbes. Ces différentes applications permettent en particulier de valider les hypothèses faites sur la forme de la pénalité et de comparer ce critère aux critères asymptotiques AIC, BIC et ICL.

Le chapitre 9 de conclusion et perspectives évoque en particulier le cas de la sélection du nombre de composantes d'un mélange gaussien, sans sélection de variables. C'est un problème fondamental en classification non supervisée par mélanges gaussiens.

Les chapitres 7 et 8 correspondent à deux articles actuellement soumis.

Chapitre 1

Classification non supervisée par mélanges gaussiens

1.1 Mélanges finis de distributions de probabilité

Depuis l'article de Newcomb (1886) pour la détection de points aberrants et l'article de Pearson (1894) sur l'estimation de cinq paramètres d'un mélange de deux lois normales, les mélanges finis de distribution de probabilité, « Finite Mixtures », ont fait l'objet de nombreux travaux. Ces dernières années, on constate un regain de popularité à l'égard des mélanges finis. Cette attention est due au fait que ces mélanges reflètent l'idée intuitive qu'une population est composée de plusieurs classes, caractérisées chacune par une distribution de probabilité. De plus, leur flexibilité permet de modéliser une large variété de phénomènes aléatoires. L'intérêt porté à ces mélanges finis aussi bien d'un point de vue théorique que pratique, est visible à travers le livre de Everitt et Hand (1981), de Titterington et al. (1985), McLachlan et Basford (1988) ou encore le livre plus récent de McLachlan et Peel (2000).

Une variable aléatoire Y prenant ses valeurs dans un espace $\mathcal{Y}$ suit une loi de mélange fini si sa densité est une combinaison convexe d'un nombre fini K de densités :

$$y \in \mathcal{Y} \mapsto \sum_{k=1}^{K} p_k f_k(y)$$

où les $f_k(\cdot)$ sont les densités de probabilité de chacun des composants du mélange et les p_k sont les proportions du mélange ($\forall k$, $p_k \in]0,1[$ et $\sum_{k=1}^{K} p_k = 1$). Généralement, on suppose que les densités $f_k(\cdot)$ appartiennent à une famille de densités paramétriques $f(.|\alpha)$. La densité du mélange fini s'écrit alors

$$y \in \mathcal{Y} \mapsto \sum_{k=1}^{K} p_k f(y|\alpha_k)$$

où $\theta = (p_1, \ldots, p_K, \alpha_1, \ldots, \alpha_K)$ est le vecteur des paramètres.

Dans cette thèse, nous considérons l'étude d'observations décrites par Q variables quantitatives. On suppose que chaque composant du mélange est modélisé par une densité gaussienne Q-dimensionnelle. Ainsi, la loi des observations est un mélange gaussien de densité

$$\forall \mathbf{x} \in \mathbb{R}^Q,\ f(\mathbf{x}|\theta) = \sum_{k=1}^{K} p_k \Phi(\mathbf{x}|\mu_k, \Sigma_k).$$

La fonction $\Phi(.|\mu, \Sigma)$ est la densité d'une loi gaussienne Q-dimensionnelle de moyenne μ et de matrice de variance Σ. Son expression est donnée par

$$\Phi(\mathbf{x}|\mu, \Sigma) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right].$$

Le vecteur des paramètres est alors $\theta = (p_1, \ldots, p_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$.

D'un point de vue théorique, ces modèles de mélanges gaussiens sont identifiables à une permutation près des composants, contrairement à d'autres mélanges finis comme par exemple les mélanges de lois uniformes, de Bernoulli ou binomiales. En effet, deux mélanges gaussiens ayant la même densité ont exactement les mêmes paramètres à une permutation près des composants (on peut se référer par exemple à McLachlan et Peel, 2000). Cette propriété d'identifiabilité est importante pour que l'estimation des paramètres du mélange soit bien définie.

1.2 28 formes de mélanges gaussiens

Géométriquement, les densités gaussiennes associées à chacun des composants du mélange correspondent à des ellipsoïdes d'inertie centrés en les moyennes μ_k. Les caractéristiques géométriques de ces ellipsoïdes sont liées à la décomposition spectrale des matrices de variance Σ_k. Selon Banfield et Raftery (1993) et Celeux et Govaert (1995), chaque matrice de variance Σ_k est décomposable de la façon suivante

$$\Sigma_k = L_k D_k A_k D_k'$$

où $L_k = |\Sigma_k|^{\frac{1}{Q}}$, D_k est la matrice orthogonale des vecteurs propres de Σ_k et A_k est une matrice diagonale constituée des valeurs propres normalisées de Σ_k, rangées par ordre décroissant et telle que $|A_k| = 1$. L'intérêt est d'avoir une interprétation géométrique de ces paramètres : L_k caractérise le volume du composant k, D_k précise son orientation et A_k indique sa forme.

Une collection de modèles peut alors être obtenue en faisant varier ou non les volumes, formes et orientations entre les composants. Ces modèles sont regroupés selon les 3 familles suivantes :

- **La famille générale :**
  Si l'on permet aux volumes, formes et orientations de varier ou d'être identiques entre les composants, on obtient 8 modèles dits « généraux ». Ces modèles sont par

convention notés sous la forme $[L_k D_k A_k D'_k]$, l'indexation en k étant abandonnée si le paramètre est identique pour tous les composants du mélange. Ainsi par exemple, le modèle noté $[LD_k AD'_k]$ indique un mélange dont les ellipsoïdes associés ont même volume et même forme mais des orientations différentes. Les modèles en $D_k A_k D'_k$ sont résumés par la notation C_k et les DAD' par C.

- **La famille diagonale :**
  Les matrices de variance Σ_k sont supposées diagonales, les matrices D_k étant alors des matrices de permutation. Dans ce cas, les matrices de variance s'écrivent sous la forme $\Sigma_k = L_k B_k$ où B_k est une matrice diagonale de déterminant 1. On obtient ainsi 4 modèles notés $[LB], [L_k B], [LB_k]$ et $[L_k B_k]$.
- **La famille sphérique :**
  La dernière famille suppose des formes sphériques à savoir $A_k = I$, I étant la matrice identité. Deux modèles sont alors obtenus : $[LI]$ et $[L_k I]$.

Finalement, les différentes combinaisons de contraintes sur les matrice de variance permettent d'obtenir 14 modèles. Ces modèles sont représentés en Figure 1.1 dans le cas d'un mélange de $K = 2$ composants en dimension $Q = 2$.

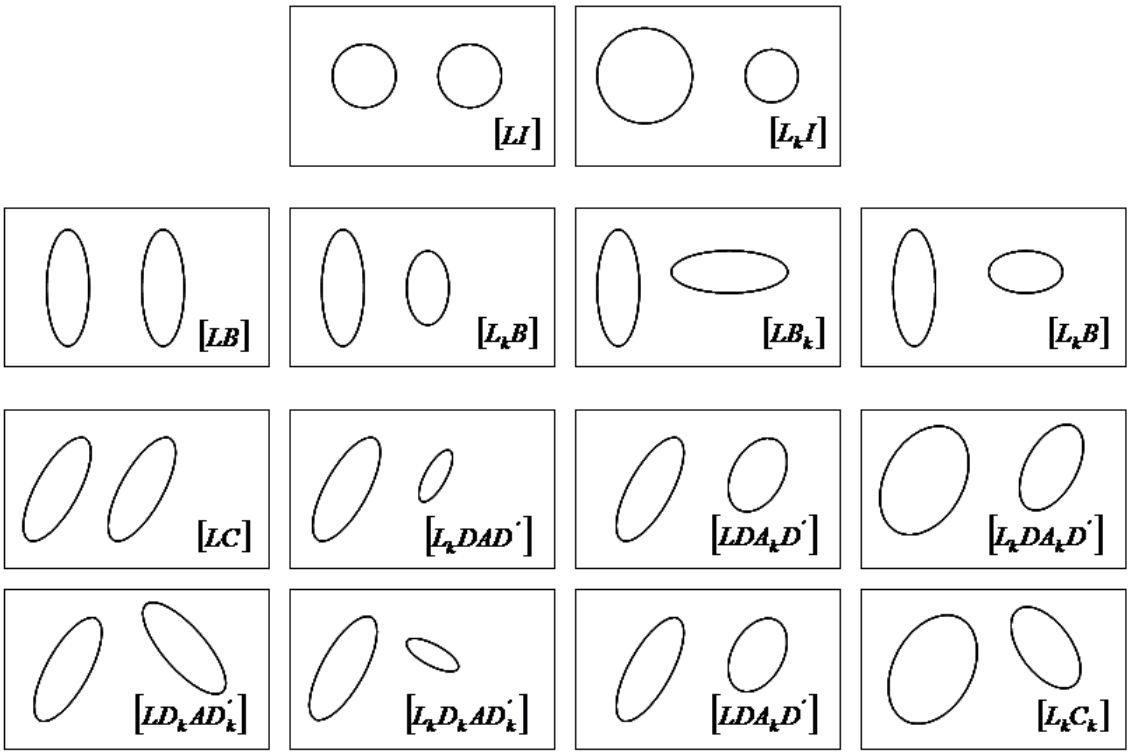


FIG. 1.1 – Représentation graphique des ellipses d'isodensité pour chacun des 14 modèles dans le cas de $K = 2$ groupes en dimension $Q = 2$.

En plus de cette flexibilité sur les matrices de variance, on peut supposer les proportions du mélange égales ou dépendantes des composants, disposant ainsi d'une collection de 28 modèles de mélanges gaussiens répertoriés dans la Table 1.1. Ces 28 modèles, faciles

à interpréter géométriquement, permettent de modéliser un grand nombre de situations de natures diverses. Ces modèles sont plus ou moins parcimonieux comme l'indique leur nombre de paramètres libres reportés en Table 1.1.

Famille	Modèle	Proportion	Volume	Orientation	Forme	nombre de paramètres libres
Sphérique	$[pLI]$	égal	égal	égal	NA	$a+1$
	$[pL_kI]$	égal	variable	égal	NA	$a+K$
Diagonale	$[pLB]$	égal	égal	axes	égal	$a+Q$
	$[pL_kB]$	égal	variable	axes	égal	$a+Q-1+K$
	$[pLB_k]$	égal	égal	axes	variable	$a+KQ-K+1$
	$[pL_kB_k]$	égal	variable	axes	variable	$a+KQ$
Générale	$[pLC]$	égal	égal	égal	égal	$a+b$
	$[pL_kC]$	égal	variable	égal	égal	$a+b+K-1$
	$[pLDA_kD']$	égal	égal	égal	variable	$a+b+(K-1)(Q-1)$
	$[pL_kDA_kD']$	égal	variable	égal	variable	$a+b+(K-1)Q$
	$[pLD_kAD'_k]$	égal	égal	variable	égal	$a+Kb-(K-1)Q$
	$[pL_kD_kAD'_k]$	égal	variable	variable	égal	$a+Kb-(K-1)(Q-1)$
	$[pLC_k]$	égal	égal	variable	variable	$a+Kb-(K-1)$
	$[pL_kC_k]$	égal	variable	variable	variable	$a+Kb$
Sphérique	$[p_kLI]$	variable	égal	égal	NA	$c+1$
	$[p_kL_kI]$	variable	variable	égal	NA	$c+K$
Diagonale	$[p_kLB]$	variable	égal	axes	égal	$c+Q$
	$[p_kL_kB]$	variable	variable	axes	égal	$c+Q-1+K$
	$[p_kLB_k]$	variable	égal	axes	variable	$c+KQ-K+1$
	$[p_kL_kB_k]$	variable	variable	axes	variable	$c+KQ$
Générale	$[p_kLC]$	variable	égal	égal	égal	$c+b$
	$[p_kL_kC]$	variable	variable	égal	égal	$c+b+K-1$
	$[p_kLDA_kD']$	variable	égal	égal	variable	$c+b+(K-1)(Q-1)$
	$[p_kL_kDA_kD']$	variable	variable	égal	variable	$c+b+(K-1)Q$
	$[p_kLD_kAD'_k]$	variable	égal	variable	égal	$c+Kb-(K-1)Q$
	$[p_kL_kD_kAD'_k]$	variable	variable	variable	égal	$c+Kb-(K-1)(Q-1)$
	$[p_kLC_k]$	variable	variable	variable	variable	$c+Kb-(K-1)$
	$[p_kL_kC_k]$	variable	variable	variable	variable	$c+Kb$

TAB. 1.1 – Liste des 28 différentes formes de mélanges gaussiens. Pour le nombre de paramètres libres, $a = KQ$, $c = a + (K-1)$ et $b = \frac{Q(Q+1)}{2}$ où Q est le nombre de variables et K le nombre de composants du mélange gaussien.

1.3 Classification et modèles de mélanges gaussiens

Considérons un échantillon $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ où les $\mathbf{y}_i = (y_{i1}, \ldots, y_{iQ})$ sont décrits par Q variables quantitatives. On désire obtenir une classification de ces données en un nombre fini K de groupes, K étant inconnu, soit une partition $(P_1, \ldots, P_K)$ des données. Cette partition inconnue des données est formalisée par $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ où ces n vecteurs

$\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ sont tels que

$$z_{ik} = \begin{cases} 1 & \text{si } \mathbf{y}_i \text{ appartient à la classe } k \\ 0 & \text{sinon.} \end{cases}$$

La résolution de ce problème se fait en deux temps. Tout d'abord on cherche la meilleure classification des données sous un modèle de mélange (K, m) fixé, K étant le nombre de composants du mélange et m sa forme. Puis nous devons résoudre un problème de sélection de modèles grâce à un critère de sélection. Cette deuxième phase est abordée dans la section suivante, nous nous concentrons ici sur la première étape, (K, m) étant fixé. La classification des données peut être obtenue de deux manières différentes. La première, appelée «approche mélange» consiste à estimer les paramètres du mélange puis à en déduire une classification en affectant chaque observation à la classe dont la probabilité d'appartenance est la plus élevée. La seconde, appelée « approche classifiante » consiste à considérer $\mathbf{z}$ comme un paramètre et donc estimer $\mathbf{z}$ et θ en même temps.

1.3.1 Algorithme EM

Dans une approche mélange, on désire estimer tout d'abord les paramètres du mélange, à savoir déterminer le vecteur des paramètres donnant le mélange gaussien à K composants de forme m le plus proche de la densité inconnue des données au sens de la divergence de Kullback-Leibler. On est donc ramené à un problème d'estimation de densité en cherchant à estimer le vecteur des paramètres $\hat{\theta}$ maximisant la logvraisemblance observée

$$L(\theta|\mathbf{y}) = \sum_{i=1}^{n} \ln \left\{ \sum_{k=1}^{K} p_k \Phi(\mathbf{y}_i|\mu_k, \Sigma_k) \right\}.$$

Pour déterminer $\hat{\theta}$, l'algorithme EM (Expectation Maximization algorithm) proposé par Dempster et al. (1977) est le plus couramment utilisé. Il est basé sur la maximisation par itérations successives de l'espérance de la logvraisemblance complétée conditionnellement aux observations $\mathbf{y}$ et à une valeur courante $\theta^{(r)}$ du vecteur des paramètres

$$\mathcal{Q}(\theta|\theta^{(r)}) = \mathbb{E}\left[\ln\{f(\mathbf{y}, \mathbf{z}|\theta)\}|\mathbf{y}, \theta^{(r)}\right]$$

où la vraisemblance complétée s'exprime par

$$f(\mathbf{y}, \mathbf{z}|\theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \{p_k \Phi(\mathbf{y}_i|\mu_k, \Sigma_k)\}^{z_{ik}}.$$

Après initialisation du vecteur des paramètres $\theta^{(1)}$, cet algorithme alterne les deux étapes suivantes. À la $r^{\text{ième}}$ itération,

- Étape E : Cette étape consiste à calculer l'espérance $\mathcal{Q}(\theta|\theta^{(r)})$, revenant à exprimer les probabilités conditionnelles notées $t_{ik}^{(r)}$ que $\mathbf{y}_i$ appartienne au composant k :

$$t_{ik}^{(r)} = P\left(z_{ik} = 1|\mathbf{y}, \theta^{(r)}\right) = \frac{p_k^{(r)} \Phi\left(\mathbf{y}_i|\mu_k^{(r)}, \Sigma_k^{(r)}\right)}{\sum_{l=1}^{K} p_l^{(r)} \Phi\left(\mathbf{y}_i|\mu_l^{(r)}, \Sigma_l^{(r)}\right)}.$$

- Étape M : Cette étape de maximisation consiste à déterminer le vecteur des paramètres $\theta^{(r+1)}$ maximisant $\mathcal{Q}(\theta|\theta^{(r)})$. Ceci est équivalent à déterminer le vecteur des proportions maximisant

$$(p_1, \ldots, p_K) \mapsto \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik}^{(r)} \ln(p_k)$$

sachant que $\sum_{k=1}^{K} p_k = 1$, si les proportions sont laissées libres dans la forme m du mélange, et à minimiser

$$(\mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K) \mapsto \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik}^{(r)} (\mathbf{y}_i - \mu_k)' \Sigma_k^{-1} (\mathbf{y}_i - \mu_k).$$

Dans tous les cas, les proportions sont données par

$$p_k^{(r+1)} = \frac{\sum_{i=1}^{n} t_{ik}^{(r)}}{n}$$

et les vecteurs moyenne par

$$\mu_k^{(r+1)} = \frac{\sum_{i=1}^{n} t_{ik}^{(r)} \mathbf{y}_i}{\sum_{i=1}^{n} t_{ik}^{(r)}}.$$

Pour les matrices de variance, le calcul dépend des conditions sur ces matrices imposées par la forme m du mélange. Les calculs selon les différents modèles sont développés dans Celeux et Govaert (1995). Dans le cas de la forme la plus générale $[L_k C_k]$, les matrices de variance sont données par

$$\Sigma_k^{(r+1)} = \frac{\sum_{i=1}^{n} t_{ik}^{(r)} (\mathbf{y}_i - \mu_k^{(r+1)})(\mathbf{y}_i - \mu_k^{(r+1)})'}{\sum_{i=1}^{n} t_{ik}^{(r)}}.$$

La croissance de $Q(\theta|\theta^{(r)})$ à chaque étape de l'algorithme implique celle de la logvraisemblance observée $L(\theta|\mathbf{y})$ puisque

$$\mathcal{Q}(\theta|\theta^{(r)}) = L(\theta|\mathbf{y}) + H(\theta|\theta^{(r)})$$

où $H(\theta|\theta^{(r)}) := \mathbb{E}\left[\ln\{f(\mathbf{z}|\mathbf{y}, \theta)\}|\mathbf{y}, \theta^{(r)}\right]$ satisfait $H(\theta|\theta^{(r)}) \leq H(\theta^{(r)}|\theta^{(r)})$ d'après l'inégalité de Jensen. Sous certaines conditions de régularité, l'estimateur obtenu par cet algorithme EM converge vers un maximum local de la logvraisemblance. En pratique, cet algorithme peut parfois converger lentement et surtout, il est influencé par la valeur initiale $\theta^{(1)}$ du vecteur des paramètres. Nous reviendrons sur des stratégies d'initialisation en Section 1.5. Les propriétés théoriques de cet algorithme sont détaillées dans Dempster et al. (1977). Une vue d'ensemble des travaux dont l'algorithme EM a fait l'objet est disponible dans McLachlan et Krishnan (1997).

1.3.2 Règle de classification

Une fois l'estimation du vecteur des paramètres $\hat{\theta}$ effectuée, on détermine la meilleure partition des observations en attribuant à chaque individu la classe pour laquelle il a la plus forte probabilité d'appartenance. Pour cela, les probabilités conditionnelles $P(\mathbf{y}_i \in P_k|\mathbf{y})$ que l'observation $\mathbf{y}_i$ appartienne à la classe k sachant l'ensemble des observations sont calculées. Par le théorème de Bayes,

$$\begin{aligned} t_{ik}(\hat{\theta}) &:= P(\mathbf{y}_i \in P_k|\mathbf{y}) \\ &= \frac{P(\mathbf{y}|\mathbf{y}_i \in P_k)P(\mathbf{y}_i \in P_k)}{P(\mathbf{y})} \\ &= \frac{\hat{p}_k \Phi(\mathbf{y}_i|\hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{l=1}^{K} \hat{p}_l \Phi(\mathbf{y}_i|\hat{\mu}_l, \hat{\Sigma}_l)}. \end{aligned}$$

Chaque observation est finalement attribuée à la classe pour laquelle la probabilité conditionnelle est la plus grande

$$\hat{z}_{ik} = \begin{cases} 1 & \text{si } t_{ik}(\hat{\theta}) > t_{il}(\hat{\theta}),\ \forall l \neq k \\ 0 & \text{sinon} \end{cases}.$$

Cette règle de classification est appelée règle du maximum a posteriori (MAP).

1.3.3 Autres algorithmes d'estimation

Toujours dans une approche mélange, Celeux et Diebolt (1985) proposent un algorithme SEM (Stochastic EM) consistant à intercaler une étape stochastique de classification afin de limiter les risques d'obtenir un maximum local de la logvraisemblance observée. Après une étape E identique à celle de l'algorithme EM, l'étape S consiste à simuler des labels inconnus $z_{ik}^{(r)}$ selon une loi multinomiale $\mathcal{M}(t_{i1}^{(r)}, \ldots, t_{iK}^{(r)})$. Puis à l'étape M, les paramètres sont évalués en maximisant la logvraisemblance complétée associée à la restauration des labels. Au final, on obtient un estimateur $\hat{\theta}$ de la logvraisemblance observée et la règle du MAP permet d'en déduire une classification des observations.

Dans une approche classifiante, Celeux et Govaert (1992) proposent une autre variante de l'algorithme EM. Cet algorithme, appelé CEM, considérant les labels inconnus comme des paramètres, vise à maximiser la logvraisemblance complétée. L'étape de classification, intercalée entre les étapes E et M, affecte chaque observation à un des composants selon la règle du MAP. L'étape M consiste ensuite à actualiser les paramètres en maximisant la logvraisemblance complétée associée à la restauration des données manquantes. La classification des données est actualisée à chaque itération de l'algorithme.

1.4 Critères de sélection de modèles

Lors de l'utilisation des mélanges gaussiens en classification non supervisée, il faut après l'étape d'estimation déterminer la forme du mélange m et son nombre de composants K

grâce à un critère de sélection de modèles. Le critère asymptotique le plus couramment utilisé est le critère BIC (Bayesian Information Criterion) de Schwarz (1978). Ce critère est basé sur la maximisation de la vraisemblance intégrée

$$f(\mathbf{y}|K,m) = \int f(\mathbf{y}|K,m,\theta)\pi(\theta|K,m)d\theta$$

où $\pi(\theta|K,m)$ est la distribution a priori du vecteur des paramètres. Cette vraisemblance intégrée étant difficilement calculable, une approximation de Laplace est utilisée pour approximer cette intégrale. Finalement, le modèle sélectionné minimise le critère BIC défini par

$$\text{BIC}(K,m) = -\ln[f(\mathbf{y}|K,m,\hat{\theta})] + \frac{\lambda_{(K,m)}}{2}\ln(n)$$

où $\lambda_{(K,m)}$ dénote le nombre de paramètres libres des mélanges gaussiens de la collection (K,m). Ce critère BIC prend donc la forme d'un critère de vraisemblance pénalisé. Bien que les conditions de régularité classiques ne sont pas satisfaites par les mélanges pour justifier BIC (Schwarz, 1978), BIC converge pour de nombreux modèles (Keribin, 2000) et est efficace en pratique.

Le critère BIC, basé sur la vraisemblance intégrée se place dans le cadre de l'approche d'estimation de densité. Biernacki et al. (2000) proposent le critère ICL, Integrated Completed Likelihood, construit pour une approche classifiante. Ce critère est basé sur la maximisation de la vraisemblance intégrée complétée

$$f(\mathbf{y},\mathbf{z}|K,m) = \int f(\mathbf{y},\mathbf{z}|K,m,\theta)\pi(\theta|K,m)d\theta.$$

Par une approximation de type Laplace, la logvraisemblance intégrée complétée est approximée par

$$\ln[f(\mathbf{y},\mathbf{z}|K,m)] \approx -\ln[f(\mathbf{y},\mathbf{z}|K,m,\widehat{\theta^\star})] + \frac{\lambda_{(K,m)}}{2}\ln(n)$$

où $\widehat{\theta^\star}$ est le vecteur des paramètres maximisant la vraisemblance complétée. Néanmoins, les labels étant inconnus, cet estimateur est approximé par celui du maximum de vraisemblance $\hat{\theta}$ et le vecteur des labels $\mathbf{z}$ est remplacé par $\hat{\mathbf{z}} = \text{MAP}(\hat{\theta})$. Finalement, le modèle sélectionné minimise le critère ICL défini par

$$\text{ICL}(K,m) = -\ln[f(\mathbf{y},\hat{\mathbf{z}}|K,m,\hat{\theta})] + \frac{\lambda_{(K,m)}}{2}\ln(n).$$

Ce critère ICL peut être décomposé en

$$\begin{aligned}
\text{ICL}(K,m) &= -\ln[f(\mathbf{y},\hat{\mathbf{z}}|K,m,\hat{\theta})] + \frac{\lambda_{(K,m)}}{2}\ln(n) \\
&= -\ln[f(\mathbf{y}|K,m,\hat{\theta})] + \frac{\lambda_{(K,m)}}{2}\ln(n) \\
&\quad -\ln[f(\mathbf{y},\hat{\mathbf{z}}|K,m,\hat{\theta})] + \ln[f(\mathbf{y}|K,m,\hat{\theta})].
\end{aligned}$$

Et puisque le second terme dans la partie droite de l'égalité précédente s'écrit comme

$$\begin{aligned} \ln\left[\frac{f(\mathbf{y},\hat{\mathbf{z}}|K,m,\hat{\theta})}{f(\mathbf{y}|K,m,\hat{\theta})}\right] &= \ln[f(\mathbf{z}|\mathbf{y},K,m,\hat{\theta})] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{z}_{ik}\ln[t_{ik}(\hat{\theta})], \end{aligned}$$

on obtient que $\mathrm{ICL}(K,m) = \mathrm{BIC}(K,m) + \mathrm{ENT}(K,m)$ où le terme d'entropie est donné par

$$\mathrm{ENT}(K,m) = -\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{z}_{ik}\ln[t_{ik}(\hat{\theta})].$$

Ainsi ICL peut être vu comme un critère BIC auquel on a ajouté une pénalité sous forme d'un terme d'entropie. Ce terme d'entropie mesure la capacité du mélange gaussien à fournir une bonne classification des données. Si les classes obtenues sont bien distinctes, ce terme d'entropie est proche de zéro alors qu'il est grand lorsque les classes sont peu séparées.

1.5 Mise en pratique

Les modèles de mélanges étant utilisés dans de nombreuses disciplines, plusieurs logiciels dédiés à ces modèles ont été développés. Sont disponibles par exemple le logiciel Emmix de McLachlan et al. (1999), le logiciel Snob de Wallace et Dowe (1994) ou le logiciel Mclust de Fraley et Raftery (2003). Dans cette thèse, nous avons utilisé le logiciel Mixmod[1] (MIXture MODelling) développé par Biernacki et al. (2006). Ce logiciel est dévolu à l'analyse de mélanges de lois de probabilité sur des données multidimensionnelles dans un but d'estimation de densité, de classification ou d'analyse discriminante. Les 28 modèles de mélanges gaussiens, les critères de sélection BIC et ICL ainsi que les algorithmes d'estimation EM, CEM et SEM sont disponibles dans ce logiciel. Nous avons utilisé Mixmod directement en C^{++} pour les algorithmes de la première partie de cette thèse et via l'interface Matlab pour les programmes de la seconde partie.

Dans cette thèse, nous nous sommes placés dans une approche mélange. L'algorithme EM a été utilisé pour l'estimation des paramètres de tous les mélanges gaussiens considérés. Cet algorithme étant sensible à la valeur initialisée du vecteur des paramètres, différentes stratégies sont proposées dans Mixmod pour cette étape d'initialisation. Nous avons opté pour l'initialisation dite «SMALL_EM»(Biernacki et al., 2003) : elle consiste à se placer en la position produisant la plus grande vraisemblance obtenue après le lancement aléatoire de courtes et nombreuses exécutions de l'algorithme EM lui-même. Nous avons pu également exploiter la famille complète des 28 modèles de mélanges gaussiens grâce à ce logiciel.

[1]Le logiciel Mixmod est disponible à l'adresse suivante : http ://www-math.univ-fcomte.fr/mixmod/

Chapitre 2

Etude de l'expression des gènes

Les problématiques abordées dans cette thèse sont en particulier motivées par l'étude de données transcriptomes. Après le séquençage du génome de plusieurs organismes comme la levure *S.cerevisiae*, la drosophile *D.melanogaster*, la plante *Arabidopsis thaliana* ou l'homme, l'enjeu consiste à déterminer la (les) fonction(s) des gènes. Pour parvenir à ce type d'informations, la première étape consiste à s'intéresser au transcriptome, à savoir l'étude de la population des ARNm exprimés par un organisme à un instant donné. Le transcriptome reflète la dynamique de la cellule et des processus biologiques en cours. Techniquement, son étude est rendue possible grâce aux puces à ADN qui permettent l'étude de milliers de gènes simultanément. Leur utilisation permet d'acquérir une mesure relative du niveau d'expression des gènes dans un échantillon cellulaire par rapport à un témoin de référence, par exemple une souche mutée comparée à une souche sauvage, ou des cellules cultivées dans deux conditions différentes. Le transcriptome peut être abordé comme un outil de biologie moléculaire pour comprendre quels sont les gènes impliqués dans une différence phénotypique observée. Ainsi après la normalisation des données pour éliminer les différents biais techniques, on peut utiliser une analyse différentielle consistant à réaliser des tests d'hypothèses pour comparer deux transcriptomes, gène à gène. Elle permet de mettre en évidence des gènes différentiellement exprimés et d'émettre des hypothèses sur leur implication dans la différence phénotypique. Mais grâce à la création de ressources transcriptomes, comme GEO[1] (Gene Expression Omnibus), CATdb[2] (Gagnot et al., 2008) ou ArrayExpress[3] (Parkinson et al., 2007), il est possible d'avoir accès à l'expression des gènes dans un grand nombre d'expériences. Le transcriptome peut alors être abordé comme une ressource génomique à partir de laquelle on va chercher à déterminer la fonction des gènes. Le transcriptome fournit dans ce cas un « profil d'expression » pour chaque gène, traduisant la variation de son niveau d'expression dans un ensemble d'expériences. Puisque les biologistes supposent que des gènes co-exprimés participent à une même fonction biologique (Eisen et al., 1998), on cherche par des méthodes de classification non supervisée à

[1] GEO : `http://www.ncbi.nlm.nih.gov/geo/`
[2] CATdb : `http://urgv.evry.inra.fr/CATdb`
[3] ArrayExpress : `http://www.ebi.ac.uk/microarray-as/ae/`

mettre en lumière des groupes de gènes co-exprimés. Le travail présenté dans ce manuscrit s'inscrit dans ce deuxième contexte d'étude des données transcriptomes.

Dans ce chapitre, nous allons tout d'abord décrire le principe des puces à ADN (Section 2.1). Puis nous préciserons les méthodes de normalisation des données (Section 2.2.1) et d'analyse différentielle (Section 2.2.2) mises en place sur la plate-forme de l'Unité de Recherche en Génomique Végétale (URGV[4]). Une vue d'ensemble des méthodes de classification non supervisée les plus couramment utilisées pour l'étude de données d'expression sera ensuite proposée en Section 2.3. La dernière section sera consacrée aux données transcriptomes produites par la plate-forme de l'URGV et étudiées durant ma thèse.

2.1 Principe des puces à ADN

Au départ élaborées sur des membranes de nylon, les puces à ADN ont progressivement été conçues sur des lames de verre à partir de la fin des années 90. Cette technologie connaît actuellement un essor exceptionnel et suscite un formidable intérêt dans la communauté scientifique. La miniaturisation donne lieu à la fabrication de puces (dites « microarray ») comportant une très forte densité de spots, permettant l'étude de l'expression de milliers de gènes simultanément dans deux conditions sur une simple lame de microscope. Le principe de la technologie des puces à ADN, résumé par la Figure 2.1, est constitué de quatre étapes présentées ci-dessous.

Fabrication des puces à ADN (microarrays) : La fabrication des puces à ADN consiste à déposer sur une lame de verre, préalablement recouverte de polylysine (pour la fixation), des milliers de fragments d'ADN amplifiés par PCR (Polymerase Chain Reaction) de façon organisée avec une micropipette robotisée. Chacun de ces fragments, appelés sondes, correspond à une tâche dite un « spot » sur la puce et est spécifique d'un gène que l'on veut étudier. En dehors des puces à ADN fabriquées à partir de produits PCR, il existe aussi des puces à oligonucléotides, dites « puces à oligos », dont les sondes sont synthétisées *in situ* (Lipshutz et al., 1999).

Préparation des cibles : On extrait des deux conditions les ARN messagers (ARNm) qui seront co-hybridés. Une transcription inverse est alors réalisée à partir de ces ARNm pour obtenir des fragments complémentaires, dits ADNc, consistant en des ADN simple brins artificiellement synthétisés. Lors de cette étape, l'ADNc de la première condition est marqué par le traceur fluorescent Cy3 (lecture scanner à 532nm) et le second par Cy5 (lecture scanner à 635nm). L'ensemble des fragments d'ADNc ainsi marqués par fluorescence forme les cibles.

Hybridation : Cette étape consiste à déposer en excès les cibles marquées sur la puce à ADN. La mise en contact permet la reconstitution de la double hélice d'ADN. Les ADNc non appariés sont ensuite éliminés par lavage de la puce. Ceux fixés aux sondes sont alors

[4]Page Web de l'URGV : http ://www.versailles.inra.fr/urgv/index.htm

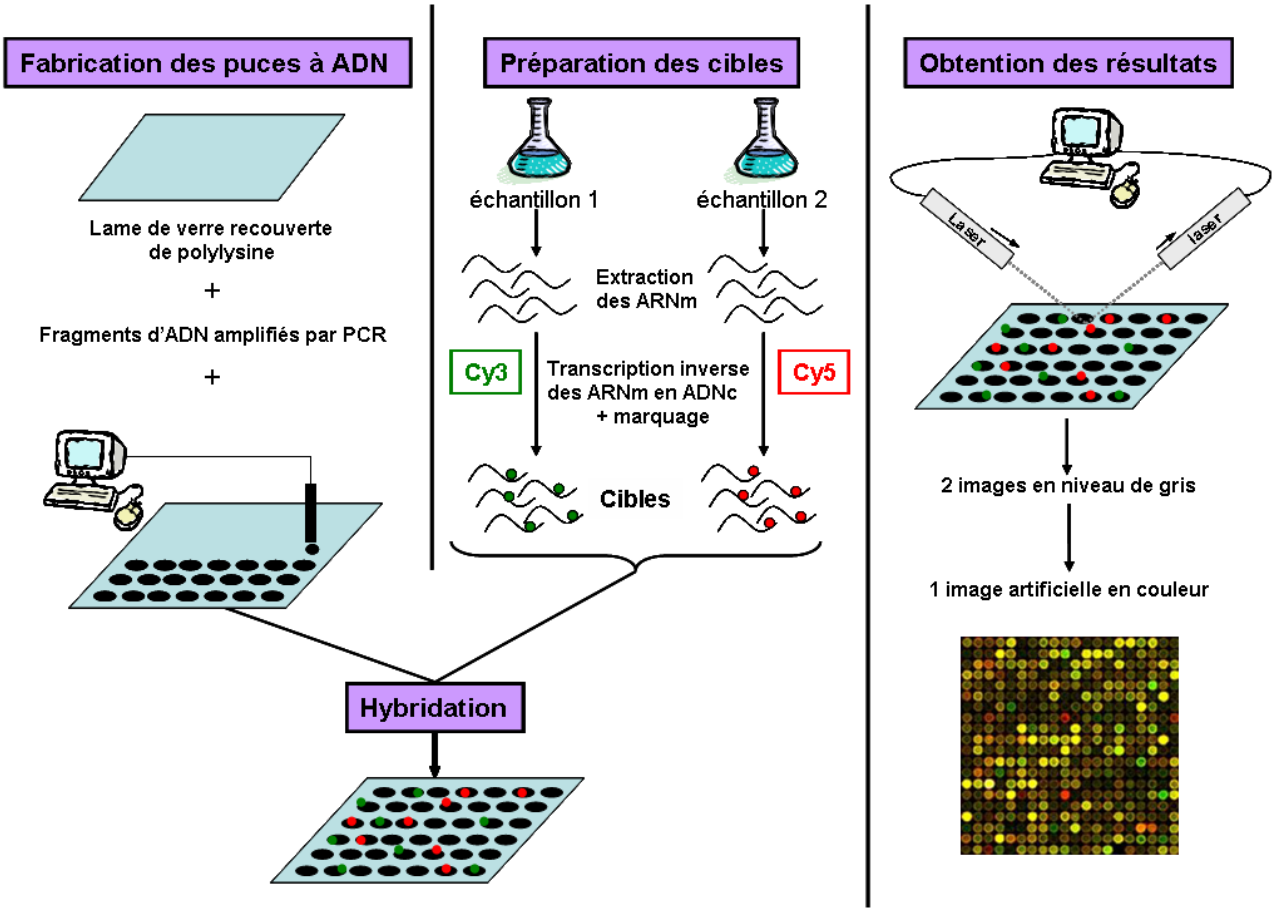


FIG. 2.1 – Principe de la technologie des puces à ADN.

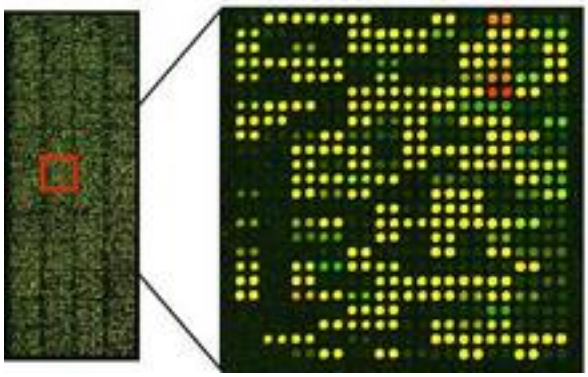


FIG. 2.2 – Exemple d'une image couleur artificielle obtenue après hybridation.

visibles grâce à leur fluorescence.

Lecture des résultats : Chaque spot est excité par un laser et la fluorescence est mesurée grâce à un scanner pour déterminer la quantité de cibles hybridées. Le scanner génère deux images en niveau de gris représentant l'intensité du signal fluorescent pour chaque fluorochrome. Une image couleur artificielle (voir Figure 2.2) est alors obtenue allant du vert pour caractériser la condition marquée par Cy3 jusqu'au rouge pour l'autre condition marquée par Cy5. Les cibles hybridées en proportions égales sont jaune.

2.2 Normalisation et analyse différentielle : exemple de la plate-forme de l'URGV

La normalisation des données et l'analyse différentielle sont deux étapes fondamentales pour l'étude des données transcriptomes. Nous n'allons pas ici faire une revue exhaustive des nombreuses méthodes proposées mais nous focaliser sur la présentation des procédures mises en place sur la plate-forme transcriptome de l'URGV. Le lecteur pourra consulter Martin-Magniette et Robin (2004), le livre de Parmigiani et al. (2003) ou encore la page web de Bioconductor, `http://www.bioconductor.org/`, pour plus de détails sur les différentes méthodes existantes.

2.2.1 Normalisation des données transcriptomes

Les signaux mesurés sont affectés par de nombreux biais techniques, non contrôlables par les biologistes. Le rôle de la normalisation est de les identifier, de les quantifier et de les soustraire du signal avant d'analyser l'expression des gènes. Nous présentons dans cette section la procédure de normalisation mise en place à l'URGV.

Nous supposons tout d'abord que l'effet biologique n'est pas confondu avec les biais techniques que l'on souhaite corriger et que la majorité des sondes s'hybrident de la même manière dans les deux conditions pour effectuer la normalisation à partir de toutes les sondes présentes sur la puce. Cette dernière hypothèse doit être prise en compte dès la construction du plan d'expérience. La normalisation n'est effectuée que sur les sondes jugées correctes par l'expérimentateur lors de la quantification de l'hybridation. Dans toute la suite de l'analyse, les signaux sont log-transformés dans le but de stabiliser la variance et de travailler avec un modèle additif. Le bruit de fond n'est pas soustrait de ces données pour ne pas accroître artificiellement la variabilité du signal.

Le logarithme en base 2 de la mesure sur la lame i, du gène g sous la condition k marquée par le fluorochrome j est noté Y_{ijkg} et est supposé suivre le modèle ANOVA suivant :

$$Y_{ijkg} = \mu + I_i + D_j + V_k + G_g + (VG)_{kg} + (DG)_{jg} + E_{ijkg}$$

où les E_{ijkg} sont des erreurs indépendantes de moyennes nulles, I_i représente l'effet lame, D_j l'effet du fluorochrome, V_k celui de la condition et G_g l'effet gène. L'effet de l'interaction

gène-condition est représenté par $(VG)_{gk}$ et l'interaction gène-fluorochrome par $(DG)_{jg}$. Sur une lame i où l'on s'intéresse à l'expression des gènes dans une expérience e dont les deux conditions e_1 et e_2 sont marquées respectivement par les fluorochromes j et j' (l'un rouge, l'autre vert), l'intensité moyenne du gène g est donnée par

$$A_{ieg} = \frac{1}{2}\{Y_{ije_1g} + Y_{ij'e_2g}\}$$

et la différence d'expression par

$$\begin{aligned} M_{ieg} &= Y_{ije_1g} - Y_{ij'e_2g} \\ &= (V_{e_1} - V_{e_2}) + (D_j - D_{j'}) + [(VG)_{e_1g} - (VG)_{e_2g}] + [(DG)_{jg} - (DG)_{j'g}] + \tilde{E}_{ieg} \end{aligned}$$

où les erreurs $\tilde{E}_{ieg}$ sont indépendantes et de moyennes nulles. La quantité d'intérêt est $(VG)_{e_1g} - (VG)_{e_2g}$ correspondant à la différence d'expression du gène g selon les deux conditions de l'expérience e. Les autres termes sont la différence d'effet entre les deux traitements et des biais de marquage à corriger. Ces biais sont principalement dûs à l'utilisation de fluorochromes et à l'autofluorescence verte de l'ADN. Ils sont observables sur le graphe MA qui consiste à représenter la différence d'expression (log ratio, M) contre l'intensité moyenne du spot (A) (voir Figure 2.3). Sous l'hypothèse que peu de gènes s'expriment différentiellement entre les deux conditions et que la quantité de fluorochrome incorporée n'a pas d'influence sur le rapport, le nuage de points devrait se situer autour de l'axe des abscisses. Or une déformation est observée, principalement due à la différence d'efficacité des fluorochromes et qui varie en plus d'une sonde à l'autre. Pour corriger cet effet, Yang

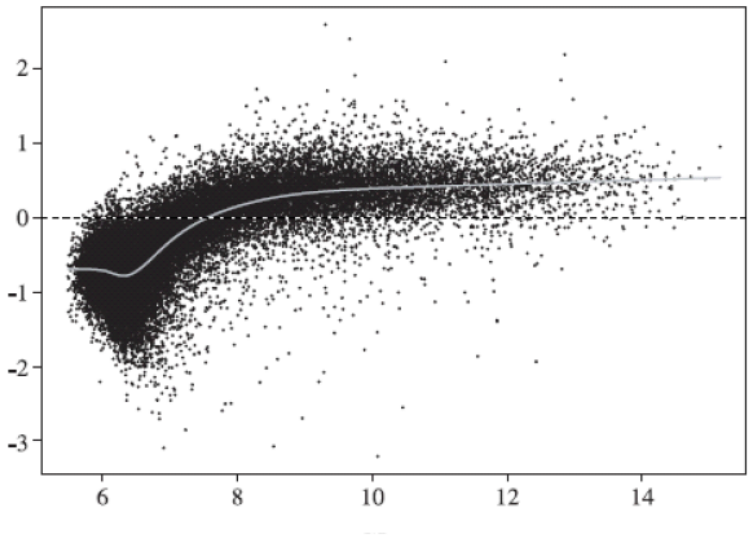


FIG. 2.3 – Graphe MA : représentation des points de coordonnées $(A_{ieg}, M_{ieg})_{1 \leq g \leq N}$.

et al. (2002) proposent d'utiliser une correction *lowess* pour normaliser les données sur

chaque lame. Cette méthode de régression locale pondérée est valable dans le cas de puces à ADN où deux conditions sont hybridées sur la même lame. Elle consiste à déterminer la fonction c, polynomiale de degré 1 à partir des données telle que $M_{ieg} = c(A_{ieg}) + \varepsilon_g$ où les ε_g sont des variables aléatoires de distribution symétrique d'espérance nulle et de variance constante. Kerr et al. (2002) ont montré que la normalisation *lowess* modifie le terme d'interaction gène-fluorochrome mais pas celui d'interaction gène-condition. Aussi la différence d'expression normalisée du gène g sur la lame i dans l'expérience e s'exprime par

$$\Delta_{ieg} = (VG)_{e_1 g} - (VG)_{e_2 g} + (-1)^{i+1}\{(DG)'_{jg} - (DG)'_{j'g}\} + F_{ieg}$$

où les F_{ieg} sont des erreurs de moyennes nulles. Cette normalisation *lowess* impose que $\sum_{g=1}^{N} \Delta_{ieg} = 0$ et $\sum_{g=1}^{N} F_{ieg} = 0$ d'où une dépendance faible de l'ordre de $1/N$ entre les F_{ieg} qui sont supposées indépendantes par la suite. Après cette étape, la variabilité de la différence des signaux par bloc, due aux problèmes de lavage et de séchage ainsi qu'à l'utilisation de plusieurs aiguilles pour déposer les sondes sur la lame, reste à maîtriser. La médiane du bloc est donc retranchée à la différence d'expression corrigée par *lowess*.

Dans les différences d'expression normalisées, le terme $(DG)'_{jg} - (DG)'_{j'g}$ représente le biais de marquage spécifique du gène. Si ce terme est non nul, la sonde correspondant au gène g a une préférence vis-à-vis d'un des marqueurs. En particulier pour corriger implicitement ce biais, au moins un dye-swap, représenté schématiquement en Figure 2.4, est mis en place pour toutes les expériences étudiées sur la plate-forme. Le principe est de faire une répétition technique en inversant les marquages fluorescents pour que chaque condition soit marquée par les deux fluorochromes. En considérant que la condition e_1 est marquée par le fluorochrome rouge (Cy5) sur une lame impaire, le carré latin représenté Table 2.1 précise le marquage des échantillons selon les conditions de l'expérience.

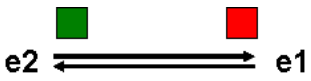


FIG. 2.4 – Représentation d'un dye-swap.

		Condition			
		e_1		e_2	
lame	impaire	rouge	(Cy5)	vert	(Cy3)
	paire	vert	(Cy3)	rouge	(Cy5)

TAB. 2.1 – Carré latin précisant la répartition des marquages.

Supposons maintenant que p dye-swaps soit $2p$ lames sont utilisés pour une expérience e. Pour chaque gène g, n_{eg} correspond au nombre de dye-swaps parmi p pour lesquels l'expérimentateur a conservé les données relatives au gène g pour les deux lames. On peut alors exprimer la différence moyenne d'expression normalisée du gène g dans l'expérience e par

$$\bar{\Delta}_{eg} = \frac{1}{2n_{eg}} \sum_{i=1}^{2n_{eg}} \Delta_{ieg} = (VG)_{e_1 g} - (VG)_{e_2 g} + \frac{1}{2n_{eg}} \sum_{i=1}^{2n_{eg}} F_{ieg}.$$

Elle dépend de la différence d'expression $(VG)_{e_1 g} - (VG)_{e_2 g}$ du gène g selon les deux

conditions de l'expérience e. L'analyse différentielle présentée ci-après est basée sur cette quantité d'intérêt.

2.2.2 Analyse différentielle

L'objectif de l'analyse différentielle est de déterminer quels sont les gènes différentiellement exprimés dans une expérience. Des tests d'hypothèses sont donc définis pour répondre statistiquement à ce problème. Pour décider si un gène g a une différence d'expression significative entre les deux conditions d'une expérience e, le test d'hypothèses suivant est étudié :

$$\begin{aligned} H_{0,(e,g)} &= \{\text{la différence d'expression du gène } g \text{ dans l'expérience } e \text{ est nulle}\} \\ &= \{\mathbb{E}[\bar{\Delta}_{eg}] = 0\} \end{aligned}$$

contre l'hypothèse alternative

$$\begin{aligned} H_{1,(e,g)} &= \{\text{la différence d'expression du gène } g \text{ dans l'expérience } e \text{ est non nulle}\} \\ &= \{\mathbb{E}[\bar{\Delta}_{eg}] \neq 0\}. \end{aligned}$$

Dans cette section, on suppose que tous les gènes sont observés pour le même nombre de dye-swaps, $n_{eg} = n_e$, et les différences d'expression Δ_{ieg} suivent une loi normale $\mathcal{N}(\mu_{eg}, \sigma^2_{eg})$ où les paramètres dépendent de l'expérience et du gène. Pour ce test, on voudrait utiliser la statistique de test suivante, qui sous l'hypothèse nulle suit une loi de Student à $2n_e - 1$ degrés de liberté,

$$\frac{\sqrt{2n_e}}{\hat{\sigma}_{eg}} \bar{\Delta}_{eg}$$

où $\hat{\sigma}^2_{eg} = \frac{1}{2n_e - 1} \sum_{i=1}^{2n_e} (\Delta_{ieg} - \bar{\Delta}_{eg})^2$ est un estimateur de la variance. Néanmoins, le nombre d'observations par gène étant généralement faible, il est inadéquat d'estimer la variance par gène. Pour cette raison, la variance est considérée comme identique pour tous les gènes. Elle est estimée par la moyenne des $\hat{\sigma}^2_{eg}$ et notée $\hat{\sigma}^2_e$. Cette hypothèse étant peu réaliste, on essaie de s'en approcher en retirant les gènes dont la variance de différence d'expression est extrême. Pour cela, on utilise le théorème de Cochran et les gènes ne vérifiant pas la condition suivante

$$\chi^2\left(\frac{\alpha'}{2}, 2n_e - 1\right) \leq (2n_e - 1)\frac{\hat{\sigma}^2_{eg}}{\hat{\sigma}^2_e} \leq \chi^2\left(1 - \frac{\alpha'}{2}, 2n_e - 1\right)$$

où $\chi^2(1 - \alpha', ddl)$ est le quantile d'ordre α' d'une loi du chi-deux à ddl degrés de liberté, sont exclus. La variance est alors réestimée par la moyenne des variances des $\tilde{N}_e$ gènes conservés pour l'expérience e, toujours notée $\hat{\sigma}^2_e$, et la statistique de test considérée pour ces gènes est

$$T_{eg} := \frac{\sqrt{2n_e}}{\hat{\sigma}_e} \bar{\Delta}_{eg}. \tag{2.1}$$

Finalement, l'hypothèse nulle $H_{0,(e,g)}$ est rejetée avec une erreur de première espèce α si la statistique de test calculée à partir des observations T_{eg}^{obs} est dans la zone de rejet $\{|T_{eg}^{\text{obs}}| > t(\alpha, \tilde{N}_e(2n_e - 1))\}$ où $t(\alpha, ddl)$ est le quantile d'ordre α de la loi de Student à ddl degrés de liberté. Comme cette règle dépend du seuil α, il est préférable de la reformuler à partir de la probabilité critique P_{eg} qui est la probabilité d'observer les données si l'hypothèse nulle est vraie :

$$P_{eg} = P_{H_{0,(e,g)}}(|T| > |T_{eg}^{\text{obs}}|)$$

où T est une variable aléatoire qui suit la même loi que la statistique de test sous l'hypothèse nulle. Ainsi, de manière équivalente, l'hypothèse nulle $H_{0,(e,g)}$ sera rejetée si $P_{eg} < \alpha$.

Il existe une seconde difficulté. En effet, la technologie des puces à ADN permet de comparer simultanément l'expression de milliers de gènes entre deux conditions. Aussi, si le test pour chaque gène est réalisé avec un risque de première espèce égal à α alors en réalisant $\tilde{N}_e$ tests indépendants, le nombre moyen de faux positifs (gènes déclarés différentiellement exprimés à tort) est égal à $\tilde{N}_e\alpha$. Il est donc important pour le test multiple $H_{0,e} = \{$tous les gènes ne sont pas différentiellement exprimés dans l'expérience $e\}$ contre $H_{1,e} = \{$au moins un gène est différentiellement exprimé dans l'expérience $e\}$ de chercher à contrôler une fonction du nombre de faux-positifs. Ce contrôle a fait l'objet de nombreux travaux dont une vue d'ensemble est disponible entre autre dans Dudoit et al. (2003) et Dudoit et van der Laan (2008). Nous ne présentons ici que les deux méthodes employées à l'URGV. Si la fonction des faux-positifs est la probabilité d'avoir au moins un faux-positif, on contrôle alors le FWER (Family-wise error rate) par la procédure de Bonferroni. Ce FWER est majoré par

$$FWER \leq \sum_{g=1}^{\tilde{N}_e} P_{H_{0,(e,g)}}\left(\text{rejeter } H_{0,(e,g)}\right).$$

Ce contrôle implique que le FWER est inférieur au seuil γ si tous les tests simples sont réalisés au seuil $\alpha = \gamma/\tilde{N}_e$. Les probabilités critiques ajustées sont donc définies par $\tilde{P}_{eg} = \min(\tilde{N}_e P_{eg}, 1)$. Une autre possibilité est de contrôler l'espérance de la proportion de faux positifs parmi les gènes déclarés différentiellement exprimés. Cette fonction des faux-positifs, appellée le FDR (False Discovery Rate), est contrôlée par la procédure de Benjamini et Hochberg (1995). Les probabilités critiques ajustées sont alors définies par

$$\tilde{P}_{e(g)} = \min_{j \geq g} \left\{ \min\left(1, P_{e(j)} \frac{\tilde{N}_e}{j}\right) \right\}$$

où $P_{e(1)} \leq P_{e(2)} \leq \ldots \leq P_{e(\tilde{N}_e)}$ sont les probabilités critiques ordonnées.

2.3 Méthodes de classification non supervisée des gènes

Les biologistes faisant l'hypothèse que des gènes ayant des profils d'expression similaires ont des liens fonctionnels, l'objectif est de déterminer des classes de gènes co-exprimés.

Usuellement, les méthodes de classification non supervisée utilisées par les biologistes sont la classification hiérarchique, les K-means et le Self-Organizing Map.

La classification hiérarchique : elle consiste à générer une suite de classes emboîtées représentée graphiquement par un dendogramme (voir Figure 2.5). Elle se base sur la matrice de similarité obtenue à partir des données d'expression et du choix d'une distance de similarité, les plus couramment employées étant la distance euclidienne et la distance de corrélation de Pearson. L'algorithme peut être ascendant ou descendant et nécessite le choix d'une distance inter-groupe pour obtenir une règle d'agglomération des classes. Les trois principales règles utilisées sont le lien moyen (« average-linkage »), le lien complet (« complete-linkage ») et le lien simple (« single-linkage ») pour lesquelles la distance entre deux groupes est donnée respectivement par la moyenne des distances entre toutes les paires d'objets, la distance entre les deux points les plus éloignés ou la distance entre les deux points les plus proches. D'autres distances comme la distance aux centroïdes ou le lien de Ward peuvent être également considérées. Cette technique de classification est très populaire dans la communauté pour l'étude des profils d'expression des gènes car elle est facile à utiliser et implémentée. Eisen et al. (1998) ont développé le logiciel CLUSTER basé sur une classification hiérarchique ascendante avec le critère « average-linkage ». Ce logiciel s'accompagne d'un programme de visualisation TreeView [5]. Il faut tout de même noter que cette procédure a une complexité algorithmique en $O(N^2 \ln N)$ si N est le nombre total de gènes à classer (Jain et al., 1999) et est très influencée par le bruit et l'ordre des données. Elle souffre d'un manque de robustesse (Tamayo et al., 1999) car une petite perturbation des données peut nettement changer la structure du dendogramme. De plus, un inconvénient de cette méthode réside dans le choix par l'utilisateur du nombre de groupes à partir du dendrogramme.

Méthode des K-means : Cette méthode (McQueen, 1967) partitionne les données en K classes, K étant fixé préalablement par l'utilisateur. Elle est basée sur la minimisation de l'inertie intra-classe (distance de chaque gène par rapport au centre du groupe auquel il appartient). Initialement, chaque point est assigné aléatoirement à l'une des classes. Puis le centre d'inertie de chaque classe est calculé et chaque point est alors affecté à la classe dont il est le plus proche du centre d'inertie. Le processus est alors itéré jusqu'à convergence de l'algorithme. Un exemple d'application de cette méthode sur des données d'expression est présenté dans Tavazoie et al. (1999). La difficulté d'utiliser une telle méthode pour la classification de données d'expression réside dans le choix initial du nombre de classes. De plus, cette méthode est sensible aux valeurs extrêmes. Pour remédier à ce dernier point, Kaufman et Rousseeuw (1987) proposent la méthode PAM (Partitioning Around Medoids) où la médoïde d'un groupe est le point possédant la distance médiane la plus faible avec les autres individus du groupe.

[5] Les deux logiciels CLUSTER et TREEVIEW sont disponibles sur `http://rana.lbl.gov/EisenSoftware.htm`

Self-Organizing Map (SOM) : Cette méthode proposée par Kohonen (1997) est basée sur un réseau de neurones artificiel, appelé aussi carte de Kohonen. L'utilisateur doit commencer par spécifier la topologie du réseau c'est-à-dire l'ensemble des nœuds et leur disposition. Le réseau peut être rectangulaire ou hexagonal et dans un espace de dimension 1, 2 ou 3. Ce réseau permet de visualiser l'ensemble V des données multidimensionnelles dans un espace de faible dimension, chaque nœud étant lié à un vecteur référent de V. Les nœuds sont liés entre eux ainsi qu'aux vecteurs de données par une fonction de voisinage. Initialement, des vecteurs référents sont choisis au hasard et chaque gène est associé au nœud dont il est le plus proche du vecteur référent. À chaque étape de cet algorithme itératif, un vecteur (un gène) de V est choisi au hasard. Le vecteur référent le plus proche du vecteur choisi est déterminé et la position du nœud associé et des nœuds voisins est ajustée. Après itération de ce processus, chaque gène est assigné à un nœud correspondant à une classe. Ce processus est résumé par la Figure 2.6. Tamayo et al. (1999) ont proposé un logiciel GeneCluster pour la classification de profils d'expression par SOM. La difficulté d'utilisation de cet algorithme réside dans le choix initial des paramètres. De plus, si de nombreuses données sont non significatives ou ont le même profil, elles vont peupler une grande majorité des classes rendant difficile la distinction de profils intéressants. Dans ce contexte, SOTA (Self-Organizing Tree Algorithm, Dopazo et Carozo, 1997) basé sur un réseau de neurones avec une topologie binaire, peut donner une meilleure classification car il combine les avantages de la classification hiérarchique et de SOM. Il retourne une classification hiérarchique avec la précision et la robustesse du réseau de neurone. Un exemple d'application de cette méthode sur données d'expression est proposé par Herrero et al. (2001).

Au delà de ces méthodes classiquement utilisées, différentes procédures de classification pour les données d'expression ont été proposées depuis une dizaine d'années. Certaines de ces méthodes sont basées sur la théorie des graphes. Sharan et Shamir (2000) proposent la méthode CLICK (CLuster Identification via Connectivity Kernels). Se donnant une similarité S_{ij} entre chaque paire de gènes (i, j), l'ensemble des gènes est représenté par un graphe complet valué selon le degré de similarité. Ce graphe est alors découpé de façon itérative en sous-graphes selon une coupe qui minimise la somme des valeurs associées aux arêtes supprimées. Au final, chaque sous-graphe représente une classe. Le nombre optimal de classes est estimé à partir des données. Notons que l'hypothèse de normalité sur la distribution de similarité est assez forte. Un schéma résumant la procédure est donné par la Figure 2.7. Ben-Dor et al. (1999) proposent un algorithme théorique et une heuristique d'utilisation appelé CAST (Cluster Affinity Search Technique). L'algorithme prend en entrée la matrice de similarité des gènes et un seuil $t \in [0, 1]$ et construit les classes une à une. Pour la classe en construction C, on mesure l'affinité d'un gène comme la somme des valeurs de similarité entre ce gène et tous les gènes de la classe considérée. L'algorithme alterne entre ajouter les gènes de forte affinité (affinité $\geq t|C|$) et enlever les gènes de faible affinité dans C jusqu'à stabilisation de la classe. Une fois une classe construite, elle n'intervient pas dans la construction des autres. L'algorithme itère le processus jusqu'à attribuer une classe à chaque gène. Il n'est pas nécessaire de définir le nombre de classes et l'algorithme est assez

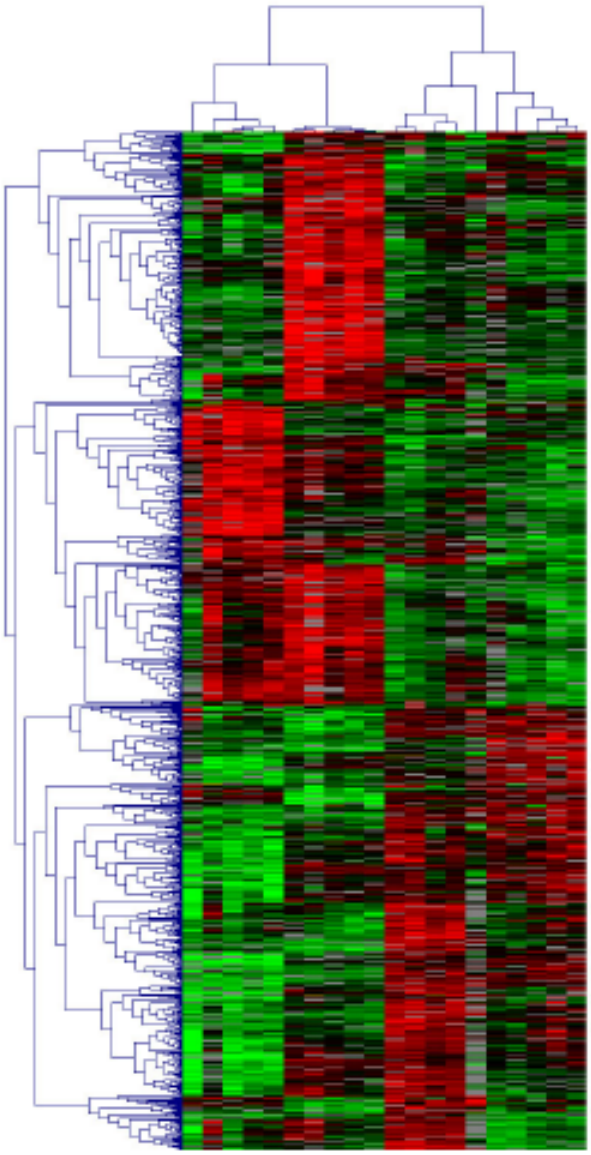


FIG. 2.5 – Exemple d'un dendogramme obtenu par classification hiérarchique sur données d'expression.

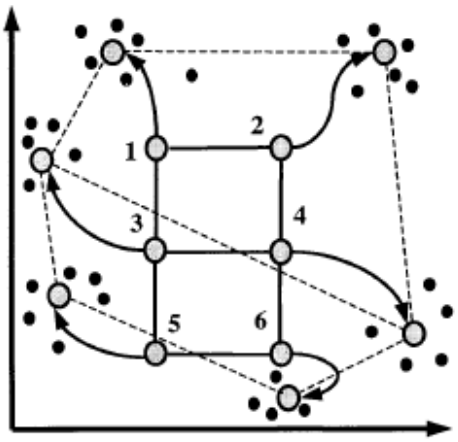


FIG. 2.6 – Schéma de l'algorithme SOM : le réseau initial est un rectangle 3×2 matérialisé par les lignes continues. Les trajectoires des nœuds (cercles) lors d'itérations successives de l'algorithme pour classer les données, représentées par des points noirs, sont indiquées par des flèches. Le réseau résultant est représenté en pointillés.

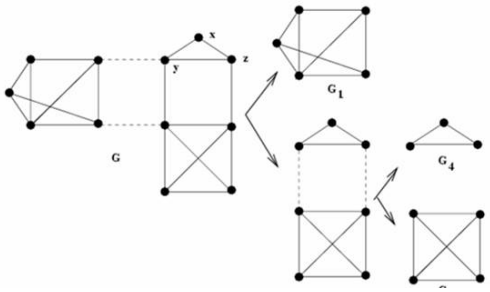


FIG. 2.7 – Schéma résumant le principe de l'algorithme CLICK.

robuste face aux données aberrantes. Par contre, il est difficile d'ajuster le seuil d'affinité t. Des approches de classification basées sur des modèles sont également peu à peu proposées. Elles fournissent un cadre statistique solide pour modéliser la structure de telles données. Par exemple, l'ensemble des données peut être supposé provenir d'un mélange de densités où chaque composant représente une classe. Le but est d'estimer le vecteur des paramètres maximisant la logvraisemblance puis de calculer les probabilités a posteriori qu'un gène appartienne à chacune des classes. Classiquement, la règle du maximum a posteriori est alors utilisée pour affecter un gène à la classe pour laquelle il a la plus forte probabilité d'appartenance. Yeung et al. (2001) ont étudié plusieurs possibilités de transformations des données et une modélisation par mélange gaussiens. On peut également citer les travaux de Ghosh et Chinnaiyan (2002). Cette liste de méthodes de classification non supervisée n'est pas exhaustive car on peut encore citer la méthode « Gene Shaving » de Hastie et al. (2000), la classification floue de Gasch et Eisen (2002), l'algorithme de Xu et al. (2001) basé sur le Minimum Spanning Tree, l'approche DHC (Density-based Hierarchical Clustering method) proposée par Jiang et al. (2003) ou encore la méthode de classification basée des mélanges infinis bayesien de Medvedovic et Sivaganesan (2002). Pour une vue d'ensemble de ces méthodes de classification non supervisée pour des données d'expression, on peut se référer aux articles de Sharan et al. (2002) et Jiang et al. (2004).

2.4 Données transcriptomes étudiées dans cette thèse

Les données transcriptomes étudiées durant ma thèse proviennent de l'Unité de Recherche en Génomique Végétale. Ce laboratoire, spécialisé dans la génomique des plantes, a participé en 2000 au séquençage du génome d'*Arabidopsis thaliana*. Cette plante, de nom commun arabette de Thalius, mesure 10 à 15 cm de haut à l'état adulte. Elle est formée d'une rosette de feuilles de 2 à 5 cm de diamètre située au ras du sol dont se détache une courte racine et un pédoncule floral portant une inflorescence blanche typique de quelques millimètres (voir Figure 2.8). Cette plante est constituée de 5 paires de chromosomes dont la séquence du génome se compose de 125 millions de nucléotides (25498 gènes). La communauté scientifique a fait de cette plante le représentant des végétaux chlorophylliens vasculaires parmi les organismes modèles utilisés en génétique.

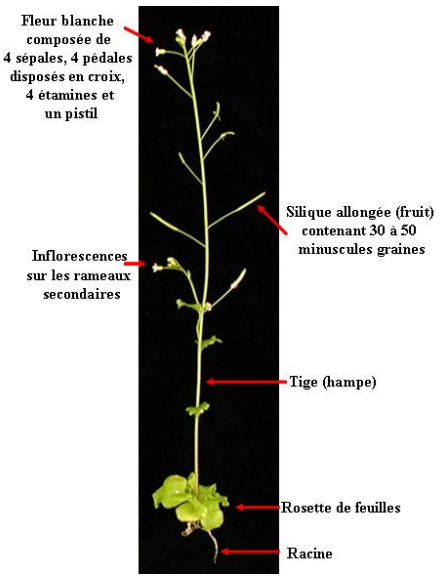


FIG. 2.8 – *Arabidopsis thaliana*.

Après le séquençage de cette plante, l'initiative européenne CATMA (Complete Arabidopsis Transcriptome MicroArray) a conçu une puce à ADN permettant d'étudier son

génome complet (Crowe et al., 2003; Hilson et al., 2004). Les sondes spécifiques employées sont des « Genes Sequence Tags »(GSTs) dessinées à partir de l'annotation structurale du génome prédite par le logiciel EuGene (Foissac et al., 2003). La version produite à l'URGV depuis 2003 comprend 24576 GSTs. La constitution d'une telle puce couvrant le génome entier, utilisable pour toutes les expériences, permet la mise en évidence de réseaux d'expression spécifiques, en vue de contribuer à élucider la fonction des gènes d'*Arabidopsis thaliana*. Précisons qu'à partir de l'analyse différentielle on dira par la suite qu'un gène est « différentiellement exprimé » dans une expérience e lorsque sa probabilité critique ajustée par Bonferroni, $\tilde{P}_{eg}$, est inférieure au seuil $\alpha = 0,05$.

La procédure de normalisation présentée en Section 2.2.1 et la procédure d'analyse différentielle décrite en Section 2.2.2 sont mises en places à l'URGV et sont disponibles dans le package R « Anapuce »[6]. Un exemple d'application de ces procédures sur des données transcriptomes issues de la plateforme de l'URGV est proposé dans Lurin et al. (2004).

L'URGV réalise des analyses du transcriptome en collaboration avec de nombreux laboratoires. Ces collaborations ont conduit à la production de données d'hybridation d'un grand nombre de puces concernant 13 types d'organes (cellule, racine, hypocotyle, protoplaste, fleur, feuille, tige, pollen, graine, ...). Ces projets concernent différentes conditions et différents stades de développement, des mutants et différents stress biotiques et abiotiques. Ces données, une fois publique, sont regroupées dans la base de données CATdb[7], Complete Arabidopsis Transcriptome database (Gagnot et al., 2008). Pour chaque projet disponible dans CATdb, on trouve le plan d'expérience mis en place par la plateforme pour répondre à la question biologique considérée (voir exemple Figure 2.9) ainsi que les valeurs des intensités et des différences d'expression de chaque gène dans chacune des expériences composant le projet.

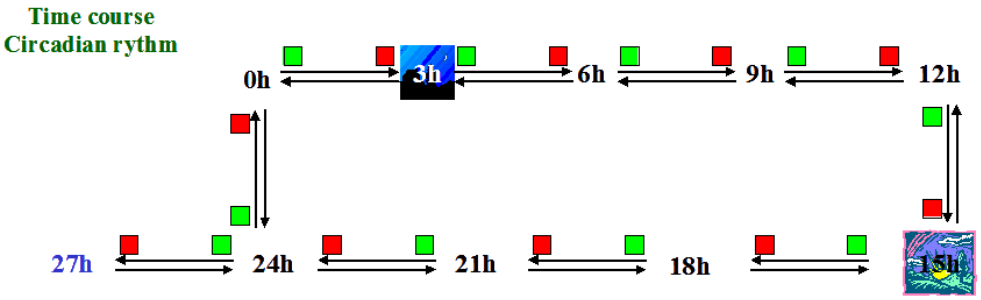


FIG. 2.9 – Plan d'expérience pour le projet « cycle circadien » disponible dans la base de données CATdb.

La classification hiérarchique sur les intensités moyennes corrigées ou les différences d'expression moyennes corrigées pour former des classes de gènes co-exprimés est la plus couramment utilisée à l'URGV. Néanmoins, les résultats obtenus sont difficilement exploitables et parfois même, les biologistes nous ont montré que la classification obtenue

[6]Anapuce : `http://www.agroparistech.fr/mmip/maths/outil.html`
[7]CATdb : `http://urgv.evry.inra.fr/CATdb`

n'est pas en adéquation avec les connaissances biologiques. Aussi dans le cadre de cette thèse, une nouvelle orientation a été choisie. Nous considérons la classification des gènes par mélanges gaussiens, nous plaçant ainsi dans un cadre statistique plus robuste. De plus, une particularité de ce travail est de décrire les gènes par leurs statistiques de test T_{eg} et non par leurs différences d'expression corrigées. Ainsi, un gène g est décrit par un vecteur dont la $e^{\text{ième}}$ coordonnée est la statistique de test du gène g dans l'expérience e donnée par l'expression (2.1). Rappelons que la statistique de test comprend le terme d'intérêt, à savoir la différence d'expression d'un gène selon les deux conditions d'une expérience, ainsi qu'une maîtrise de la variabilité. Ainsi la matrice des données reflète l'activité de chaque gène dans chaque expérience tout en bénéficiant d'une variabilité plus faible que celle de la matrice des différences d'expression. Ces données comprennent des valeurs manquantes provenant de deux sources différentes. Les premières sont dues à des problèmes rencontrés lors de la fabrication et l'hybridation des puces à ADN. Les secondes concernent les gènes avec une trop forte ou trop faible variance, leur statistique de test n'étant pas alors évaluée lors de l'analyse différentielle. Néanmoins, il est possible dans ce second cas de définir une « pseudo » statistique de test en prenant la différence d'expression disponible et en la normalisant par l'estimateur de l'écart-type évalué avec seulement les gènes satisfaisant l'hypothèse d'homoscédasticité (voir Section 2.2.2).

Début août 2008, 70 projets soient 2256 expériences sont rendus publiques dans la base de données CATdb. Intuitivement, on pourrait penser que ce flot d'informations est bénéfique pour déterminer des sous-groupes de gènes co-exprimés. Pourtant, la majorité des gènes étudiés sont non différentiellement exprimés dans une expérience. Cette masse de points autour de zero rend alors difficile la détection de profils similaires intéressants. Aussi, il semble judicieux d'inclure dans le processus de classification une procédure de sélection de variables. On espère ainsi améliorer la classification des gènes et faciliter son interprétation par les biologistes. Il faut également savoir prendre en compte les valeurs manquantes dans la matrice d'expression étudiée. Enfin d'après la nature des données, le découpage en variables peut se faire à différents niveaux : chaque variable peut correspondre à une expérience ou représenter un projet puisque les expériences le composant sont liées à une même question biologique. Dans ce second cas, la taille d'une variable dite « variable bloc » est égale au nombre d'expériences dans le projet correspondant. Toutes ces questions détectées dans l'analyse des données transcriptomes ont motivé les problématiques statistiques abordées dans cette thèse.

Part I

Variable role modelling for Gaussian mixture clustering

Chapter 3

Variable selection for clustering with Gaussian mixture models

Résumé: Ce chapitre s'intéresse à la sélection de variables en classification non supervisée. Le problème est ramené à un problème de sélection de modèles de mélanges. Un modèle global, généralisant celui de Raftery et Dean (2006b) est proposé pour spécifier le rôle des variables pour le processus de classification. Ce modèle ne nécessite aucune hypothèse a priori sur le lien entre les variables sélectionnées et les variables écartées pour la classification. Les modèles sont comparés avec un critère de type BIC. Le statut des variables est obtenu en pratique grâce à un algorithme imbriquant deux algorithmes descendants de sélection de variables pour la classification et pour la régression linéaire. L'identifiabilité des modèles est établie et la consistance du critère de sélection est démontrée sous des conditions de régularité. Des exemples numériques sur données simulées et sur une application génomique mettent en évidence l'intérêt de cette procédure de sélection de variables.

3.1 Introduction

The goal of clustering methods is to discover structures (clusters) among individuals described by several variables. Many clustering methods exist and roughly fall into two categories. The first one is based on similarity or dissimilarity distances. It gathers hierarchical clusterings, which build trees and also methods like K-means algorithm which classify data through a number of clusters fixed a priori. The second category is model-based methods which consist of using a model for clusters and optimizing the fit between the data and the model. In practice, each cluster is represented by a parametric distribution, like a Gaussian one and the entire dataset is modelled by a mixture of these distributions. An advantage of model-based clustering is to provide a rigourous framework to assess the number of clusters and the role of each variable in the clustering process.

In principle, the more information we have about each individual, the better a clustering

method is expected to perform. However the structure of interest may often be contained into a subset of the available variables and a lot of variables may be useless or even harmful to detect a reasonable clustering structure. It is thus important to select the relevant variables from the cluster analysis view point. It is a recent research topic in contrast to variable selection in regression and classification models (Kohavi and John, 1997; Guyon and Elisseeff, 2003; Miller, 1990). This new interest for variable selection in clustering comes from the increasingly frequent use of these methods on high dimensional datasets, such as transcriptome datasets. It is usually considered that coexpressed genes are often implicated in the same biological function and consequently are potential candidates to be co-regulated genes (see for instance Sharan et al., 2002, or Jiang et al., 2004, and references therein). Since the number of transcriptome experiments always increases, an experiment selection in the clustering procedure is desirable to reveal important biological phenomena.

Three types of approach dealing with variable selection in clustering have been proposed. The first one includes clustering methods with weighted variables (see for instance Friedman and Meulman, 2004) and dimension reduction methods. For this later, McLachlan et al. (2002) use a mixture of factor analyzers to reduce the extremely high dimensionality of a gene expression problem. A suitable Gaussian mixture family is considered in Bouveyron et al. (2007) to take the dimension reduction and the data clustering simultaneously into account. In contrast to this first method type, the two last approaches select explicitly relevant variables. The so-called "filter" approaches select the variables before a clustering analysis (see for instance Dash et al., 2002; Jouve and Nicoloyannis, 2005). Their main weakness is the influence of independent selection step on the clustering results. In contrast, the so-called "wrapper" approaches combine variable selection and clustering. For distance-based methods, one can cite Fowlkes et al. (1988) for a forward selection approach with complete linkage hierarchical clustering, Devaney and Ram (1997) who propose a stepwise algorithm where the quality of the feature subsets is measured with the COBWEB algorithm or the method of Brusco and Cradit (2001) based on the adjusted Rand index for K-means clustering. There exist also wrapper methods in the model-based clustering setting. When the number of variables is greater than the number of individuals, Tadesse et al. (2005) propose a fully Bayesian method using a reversible jump algorithm to simultaneously choose the number of mixture components and select variables. Kim et al. (2006) use a similar approach by formulating clustering in terms of Dirichlet process mixtures. In Gaussian mixture model clustering, Law et al. (2004) propose to evaluate the importance of the variables in the clustering process via "feature saliencies" and use the *Minimum Message Length* criterion. Raftery and Dean (2006b) recast the problem of comparing two nested variable subsets as a model comparison problem and address it using Bayes factor. An interesting aspect of their model formulation is that irrelevant variables are not required to be independent of the clustering variables. They avoid thus the unrealistic independence assumption between the relevant and irrelevant variables for the clustering, considered in Tadesse et al. (2005), Kim et al. (2006) and Law et al. (2004). In their model, the whole irrelevant variable subset depends on the whole relevant variables through a linear regression equation. However, some relevant variables are not necessarily required to explain all irrelevant variables in the linear regression and their introduction

involves additional parameters without a significant increase of the loglikelihood.

In this chapter, we improve their method by considering another type of relation between the irrelevant variables for clustering and the relevant ones. We consider that the irrelevant variables can be independent of some relevant variables. This modelling allows us to improve the clustering and its interpretation. Our variable selection implementation is based on a backward stepwise algorithm. Moreover, we look at a more general situation where the variables are partitioned into blocks which cannot be splitted.

The chapter is organized as follows: Gaussian mixture models for clustering are reviewed in Section 3.2. Our variable selection approach is presented in Section 3.3. The associated search algorithm is described in Section 3.4. The model identifiability and the consistency of the variable selection criterion are stated in Section 3.5 and proved in Appendix 3.C and Appendix 3.D respectively. Simulated experiments are presented to validate the method and to compare it with Raftery and Dean's approach in Section 3.6. An example of transcriptome data clustering is addressed in Section 3.7. Finally, a discussion on the overall method is given in Section 3.8.

3.2 Multivariate Gaussian models and clustering

Model-based clustering consists of assuming that data come from several subpopulations modelled separately and the overall population is a mixture of these subpopulations. The resulting model is a finite mixture model. When data are multivariate continuous observations, the parameterized component density is usually a multidimensional Gaussian density. We consider n individuals $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_n')'$ described by Q variables ($\mathbf{y}_i'$ in $\mathbb{R}^Q$). Observations are assumed to be a sample from a probability distribution with density

$$f(\mathbf{y}_i \mid K, \alpha) = \sum_{k=1}^{K} p_k \Phi(\mathbf{y}_i \mid \mu_k, \Sigma_k),$$

where the p_k's are the mixing proportions ($p_k \in]0,1[$ for all $k = 1, \ldots, K$ and $\sum_{k=1}^{K} p_k = 1$), and $\Phi(. \mid \mu_k, \Sigma_k)$ denotes the Q-dimensional Gaussian density with mean μ_k and variance matrix Σ_k. The vector parameter is denoted $\alpha = (p_1, \ldots, p_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$. The mixture model is an incomplete data structure model: The complete data consist of $((\mathbf{y}_1, \mathbf{z}_1)', \ldots, (\mathbf{y}_n, \mathbf{z}_n)')'$ where the missing data are $\mathbf{z} = (\mathbf{z}_1', \ldots, \mathbf{z}_n')'$, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ being binary vectors such that $z_{ik} = 1$ iff $\mathbf{y}_i$ arises from the k^{th} subpopulation. The $\mathbf{z}$'s define an ideal clustering of the data $\mathbf{y}$, associated to the mixture model.

As in Banfield and Raftery (1993) and Celeux and Govaert (1995), the mixture component variance matrix can be decomposed into $\Sigma_k = L_k D_k A_k D_k'$ where $L_k = |\Sigma_k|^{1/Q}$, D_k is the Σ_k's eigenvector matrix and A_k is the diagonal matrix of normalized eigenvalues of Σ_k. They control respectively the volume, the orientation and the shape of the k^{th} cluster. According to constraints required on the different elements of this decomposition, a collection of parsimonious and interpretable models is available. Moreover, the proportions can

be assumed to be equal or free. Finally, the considered model family is

$$\mathcal{T} = \{(K, m) \in \{2, \ldots, K_{\max}\} \times \mathcal{M}\}$$

where $\mathcal{M}$ is a collection of 28 models, described in Chapter 1 (see Table 1.1), and $K_{\max}$ is the maximum number of clusters specified by the user. Those 28 models are available in the MIXMOD software (Biernacki et al., 2006) and, for most of them, in the MCLUST software (Fraley and Raftery, 2003).

In this inferential framework, it is possible to choose one of the models $(K, m) \in \mathcal{T}$, by using model selection methods or criteria (see McLachlan and Peel, 2000). In a Bayesian perspective, the model maximizing the posterior probability $P[(K, m)|\mathbf{y}]$ is to be chosen. By Bayes theorem

$$P[(K, m)|\mathbf{y}] = \frac{f(\mathbf{y}|K, m)P[(K, m)]}{f(\mathbf{y})},$$

and supposing a non informative uniform prior distribution $P[(K, m)]$ on the models, it leads to $P[(K, m)|\mathbf{y}] \propto f(\mathbf{y}|K, m)$. Thus the chosen model satisfies

$$(\tilde{K}, \tilde{m}) = \underset{(K,m)\in\mathcal{T}}{\operatorname{argmax}}\ f(\mathbf{y}|K, m),$$

where the integrated likelihood $f(\mathbf{y}|K, m)$ is defined by

$$f(\mathbf{y}|K, m) = \int f(\mathbf{y}|K, m, \alpha)\pi(\alpha|K, m)d\alpha,$$

$\pi(\alpha|K, m)$ being the prior distribution of the vector parameter α of the (K, m) model (Kass and Raftery, 1995). Since this integrated likelihood is typically difficult to calculate, an asymptotic approximation of $2\ln\{f(\mathbf{y}|K, m)\}$ is generally used. This approximation is the Bayesian Information Criterion (BIC) defined by

$$\mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}|K, m) = 2\ln\{f(\mathbf{y}|K, m, \hat{\alpha})\} - \lambda_{(K,m)}\ln(n) \tag{3.1}$$

where $\lambda_{(K,m)}$ is the number of free parameters for the (K, m) model and $f(\mathbf{y}|K, m, \hat{\alpha})$ is the maximum likelihood under this model (Schwarz, 1978). In this perspective, the selected model is

$$(\hat{K}, \hat{m}) = \underset{(K,m)\in\mathcal{T}}{\operatorname{argmax}}\ \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}|K, m).$$

For deriving $(\hat{K}, \hat{m})$, the maximum likelihood estimate (mle) $\hat{\alpha}$ is computed using generally the EM algorithm (Dempster et al., 1977). Finally, the clustering is performed using the Maximum a Posteriori (MAP) rule defined by

$$\hat{z}_{ik} = \begin{cases} 1 & \text{if } \hat{p}_k\Phi(\mathbf{y}_i|\hat{\mu}_k, \hat{\Sigma}_k) > \hat{p}_l\Phi(\mathbf{y}_i|\hat{\mu}_l, \hat{\Sigma}_l), \forall l \neq k, \\ 0 & \text{otherwise.} \end{cases}$$

Here all the Q variables are supposed to enter in the mixture models. When there are numerous variables, it can be sensible to choose which variables are actually required in the mixture models. This can be regarded as a model selection problem as well.

3.3 Selecting variables

The approach we propose for selecting relevant variables for clustering is related to the Raftery and Dean (2006b) approach that is sketched first. Their idea is to divide the variable set into a subset of relevant clustering variables and its complement which does not provide information for the clustering but which depends on the relevant variables through a linear regression. This is an interesting aspect since they avoid the unrealistic and usual independence assumption between the relevant and irrelevant variables for the clustering. As they stressed, the independence assumption would often lead to wrongly declare a variable as relevant for the clustering because this variable is related to some relevant variables, but not necessarily to the clustering itself. Although relevant variables are not all required to explain the whole irrelevant variable subset, Raftery and Dean (2006b) force them to enter in the regression model. This involves additional parameters in the model without necessarily leading to a significant increase of its loglikelihood. One consequence is that models assigning some irrelevant variables as relevant could be wrongly preferred when model comparison is performed with Bayes factor or penalized likelihood criteria. In this chapter, we opt for a more realistic model where irrelevant variables are explained by a subset of relevant variables. Moreover, we consider a more general framework where the Q variables are partitioned into T blocks. That is there exists a function Ψ such that each variable $j \in \{1, \ldots, Q\}$ belongs to a unique variable block $\Psi(j) \in \{1, \ldots, T\}$. This common situation appears for instance in the genomic application considered in Section 3.7. Obviously in the standard situation where each block reduces to a single variable, we have $T = Q$ and all the following formula can be straightforwardly particularized to this simple case. Throughout the chapter, $\mathbf{y}^j = (y_1^j, \ldots, y_n^j)'$ and for a subset of variable blocks A, $\mathbf{y}^A$ denotes the set $\{\mathbf{y}^j \in \mathbb{R}^n;\ \Psi(j) \in A\}$ and $\text{card}(A) = \text{card}\{j; \Psi(j) \in A\}$.

Let $\mathcal{F}$ be the family of variable block index subset, $S \in \mathcal{F}$ the set of relevant clustering variable block indexes and, S^c its complement in $\{1, \ldots, T\}$, denoting the irrelevant variables. In order to distinguish the role of each clustering variable block in the regression, those entering in the regression equation of the irrelevant variables constitute the subset R. The information is thus summarized into a couple (S, R) belonging to

$$\mathcal{V} = \{(S, R); (S, R) \in \mathcal{F}^2,\ S \neq \emptyset,\ R \subseteq S\}.$$

This division of the variable block roles is illustrated in Figure 3.1. Finally, the considered model set is defined by $\mathcal{N} = \{(K, m, S, R); (K, m) \in \mathcal{T};\ (S, R) \in \mathcal{V}\}$.

The models in competition are compared with their integrated likelihoods decomposed into two multiplicative parts

$$f(\mathbf{y}|K, m, S, R) = f_{\text{clust}}(\mathbf{y}^S|K, m) f_{\text{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^R).$$

The function $f_{\text{clust}}(\mathbf{y}^S|K, m) = \int f_{\text{clust}}(\mathbf{y}^S|K, m, \alpha)\pi(\alpha|K, m, S)d\alpha$ is the integrated likelihood of the (K, m) mixture model on the relevant clustering variables. The function $f_{\text{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^R) = \int f_{\text{reg}}(\mathbf{y}^{S^c}|a + \mathbf{y}^R\beta, \Omega)\pi(a, \beta, \Omega|S^c, R)da d\beta d\Omega$ is the integrated likelihood of

multidimensional regression of the irrelevant variables on a subset of the clustering variables, a , β and Ω denoting the intercept vector, the regression coefficient matrix and the variance matrix respectively (see Appendix 3.A). In practice, the integrated likelihoods are approximated using the BIC approximation as in (3.1), and the chosen model is

$$(\hat{K}, \hat{m}, \hat{S}, \hat{R}) = \underset{(K,m,S,R)\in\mathcal{N}}{\operatorname{argmax}} \{\mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S|K, m) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^R)\} \tag{3.2}$$

where

$$\mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S|K, m) = 2\ln\{f_{\mathrm{clust}}(\mathbf{y}^S|K, m, \hat{\alpha})\} - \lambda_{(K,m,S)}\ln(n)$$

and

$$\mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^R) = 2\ln\{f_{\mathrm{reg}}(\mathbf{y}^{S^c}|\hat{a} + \mathbf{y}^R\hat{\beta}, \hat{\Omega})\} - \nu_{(S^c,R)}\ln(n), \tag{3.3}$$

$\lambda_{(K,m,S)}$ being the number of free parameters of the (K, m) mixture model with $\mathrm{card}(S)$ variables, $(\hat{a}, \hat{\beta}, \hat{\Omega})$ the maximum likelihood estimate of the regression parameters, and $\nu_{(S^c,R)} = \{\mathrm{card}(R) + 1\}\,\mathrm{card}(S^c) + \frac{\mathrm{card}(S^c)\{\mathrm{card}(S^c)+1\}}{2}$. The computation of BIC for multidimensional multivariate regression is detailed in Appendix 3.A.

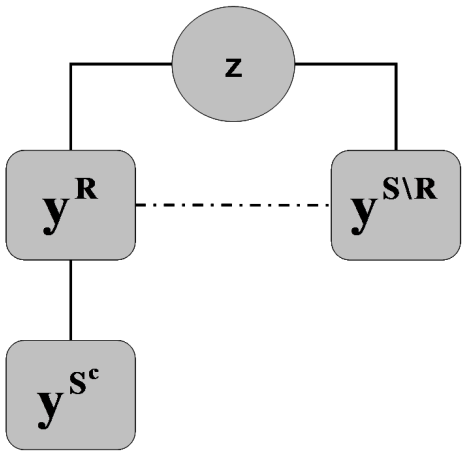


Figure 3.1: Graphical representation of the variable repartition in the model.

3.4 The variable selection procedure

The number of models in $\mathcal{N}$ is $28(K_{\max} - 1)\sum_{t=1}^{T}\binom{T}{t}\sum_{l=0}^{t}\binom{t}{l}$, where $K_{\max}$ is the maximum number of clusters. Thus an exhaustive search of the optimal model is impossible in most situations. The algorithm we propose is a two-nested-step algorithm.

1. For all (K, m), we search

$$(\hat{S}(K, m), \hat{R}(K, m)) = \underset{(S,R)\in\mathcal{V}}{\operatorname{argmax}} \left\{\mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S|K, m) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^R)\right\}$$

   by a backward stepwise procedure detailed hereafter.

2. We determine

$$(\hat{K}, \hat{m}) = \underset{(K,m)\in\mathcal{T}}{\operatorname{argmax}} \left\{ \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^{\hat{S}(K,m)}|K,m) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^{\hat{S}^c(K,m)}|\mathbf{y}^{\hat{R}(K,m)}) \right\}.$$

Finally, the selected model is $(\hat{K}, \hat{m}, \hat{S}(\hat{K}, \hat{m}), \hat{R}(\hat{K}, \hat{m}))$.
We opt for a backward stepwise selection algorithm. It means that all the variables are selected at the beginning and at each step, a variable block is excluded or included.

3.4.1 The models in competition

At each step of this algorithm, the variable set $\{1, \ldots, T\}$ is divided into three subgroups: S the set of selected clustering variable blocks, j the candidate variable block being considered for inclusion into or exclusion from the set of clustering variables and U the irrelevant variable set. The integrated likelihood can be thus decomposed into

$$f(\mathbf{y}|K,m) = f(\mathbf{y}^U|\mathbf{y}^j, \mathbf{y}^S) f(\mathbf{y}^j, \mathbf{y}^S|K,m).$$

The decision of exclusion (resp. inclusion) of variable block j from (resp. in) the set of clustering variables is made by the comparison of the following two models:

1. $M_1(K,m)$ specifies that given $\mathbf{y}^S$, $\mathbf{y}^j$ does not provide additional information for the clustering and is explained by a subset $\mathbf{y}^{R[j]}$ of $\mathbf{y}^S$,

$$\begin{aligned} f_1(\mathbf{y}^j, \mathbf{y}^S|K,m) &= \sum_{\mathbf{z}} f(\mathbf{y}^j, \mathbf{y}^S|\mathbf{z}, K, m) f(\mathbf{z}|K,m) \\ &= \sum_{\mathbf{z}} f_{\mathrm{reg}}(\mathbf{y}^j|\mathbf{y}^{R[j]}) f(\mathbf{y}^S|\mathbf{z}, K, m) f(\mathbf{z}|K,m) \\ &= f_{\mathrm{reg}}(\mathbf{y}^j|\mathbf{y}^{R[j]}) f_{\mathrm{clust}}(\mathbf{y}^S|K,m). \end{aligned}$$

2. $M_2(K,m)$ specifies that given $\mathbf{y}^S$, $\mathbf{y}^j$ provides additional information for the clustering,

$$f_2(\mathbf{y}^j, \mathbf{y}^S|K,m) = f_{\mathrm{clust}}(\mathbf{y}^j, \mathbf{y}^S|K,m).$$

The two models are compared with the Bayes factor, $B_{12}(K,m)$ for $M_1(K,M)$ against $M_2(K,M)$:

$$B_{12}(K,m) = \frac{f_{\mathrm{reg}}(\mathbf{y}^j|\mathbf{y}^{R[j]}) f_{\mathrm{clust}}(\mathbf{y}^S|K,m)}{f_{\mathrm{clust}}(\mathbf{y}^j, \mathbf{y}^S|K,m)}.$$

Since the integrated likelihoods are difficult to evaluate, $-2\ln\{B_{12}(K,m)\}$ is approximated with

$$\mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^j|K,m) = \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S, \mathbf{y}^j|K,m) - \left\{ \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S|K,m) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^j|\mathbf{y}^{R[j]}) \right\}. \quad (3.4)$$

If $\mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^j|K,m)$ is positive, Model M_2 is chosen, otherwise Model M_1 is chosen.

3.4.2 The backward stepwise selection algorithm

Let (K, m) be fixed, this algorithm is making use of an exclusion and an inclusion procedures now described. The decision of excluding (resp. including) a variable block from (resp. into) the set of clustering variables is based on the comparison of the two models with the BIC approximation of the Bayes factor.
Initialisation: $S = \{1, \ldots, T\}$, $j_E = \emptyset$ and $j_I = \emptyset$.
Exclusion step: In this step, the proposed variable block for removal from the set of currently selected clustering variables is chosen to be the variable block which gives the smallest value of BIC_{diff} defined in (3.4). It is as follows:

For all j in S, use the backward stepwise selection algorithm, described in Appendix 3.B, to choose the subset $R[j]$ of dependent variables for the regression of $\mathbf{y}^j$ on $\mathbf{y}^{S\setminus j}$, and compute $\text{BIC}_{\text{diff}}(\mathbf{y}^j|K, m)$. Then, compute

$$j_E = \underset{j\in S}{\text{argmin}}\, \text{BIC}_{\text{diff}}(\mathbf{y}^j|K, m).$$

- If $\text{BIC}_{\text{diff}}(\mathbf{y}^{j_E}|K, m) \leq 0$, $S = S\setminus j_E$ and: Stop if $j_E = j_I$, otherwise go to the inclusion step;
- otherwise: Stop if $j_I = \emptyset$ or go to the inclusion step otherwise.

Inclusion step: In this step, the proposed new variable block is chosen to be the variable block which gives the greatest value of BIC_{diff}. It is as follows:

For all j in S^c, use the backward stepwise selection algorithm, described in Appendix 3.B, to choose the subset $R[j]$ of dependent variables for the regression of $\mathbf{y}^j$ on $\mathbf{y}^S$, and compute $\text{BIC}_{\text{diff}}(\mathbf{y}^j|K, m)$. Then, compute

$$j_I = \underset{j\in S^c}{\text{argmax}}\, \text{BIC}_{\text{diff}}(\mathbf{y}^j|K, m).$$

- If $\text{BIC}_{\text{diff}}(\mathbf{y}^{j_I}|K, m) > 0$, stop if $j_I = j_E$, otherwise $S = S \cup j_I$ and go to the exclusion step;
- otherwise go to the exclusion step.

Starting from the exclusion step, the backward variable selection algorithm consists of alternating exclusion and inclusion steps. It returns the relevant clustering variable subset $\hat{S}(K, m)$. Next, $\hat{R}(K, m)$ is obtained using the backward stepwise algorithm for the regression of $\mathbf{y}^{S^c}$ on $\mathbf{y}^S$ (see Appendix 3.B).

3.5 Theoretical properties

In this section, necessary and sufficient conditions are given to ensure the model identifiability and a consistency theorem of the criterion is stated. In model (K, m, S, R), the parameterized densities are denoted $f(.|\theta)$ where $\theta = (\alpha, a, \beta, \Omega) \in \Upsilon_{(K,m,S,R)}$ in the sequel.

3.5.1 Identifiability

The model identifiability is based on the following remark: Let s be a nonempty subset included strictly into S and $\bar{s}$ be its complement in S, then the density $f(.|\theta)$ under the model (K, m, S, R) can be decomposed as

$$\begin{aligned} f(x|\theta) &= \sum_{k=1}^{K} p_k \Phi(x^S|\mu_k, \Sigma_k)\Phi(x^{S^c}|a + x^R\beta, \Omega) \\ &= \sum_{k=1}^{K} p_k \Phi(x^s|\mu_{ks}, \Sigma_{k,ss})\Phi(x^{\bar{s}}|\mu_{k,\bar{s}|s} + x^s\Sigma_{k,\bar{s}|s}, \Sigma_{k,\bar{s}\bar{s}|s})\Phi(x^{S^c}|a + x^R\beta, \Omega) \end{aligned}$$

where mixture parameters are decomposed into $\mu_k = (\mu_{ks}, \mu_{k\bar{s}})$ and Σ_k into submatrices $\Sigma_{k,ss}$, $\Sigma_{k,s\bar{s}}$ and $\Sigma_{k,\bar{s}\bar{s}}$ (according to Theorem 2.5.1 of Anderson, 2003, page 35). The conditional parameters are defined by $\mu_{k,\bar{s}|s} = \mu_{k\bar{s}} - \mu_{ks}\Sigma_{k,ss}^{-1}\Sigma_{k,s\bar{s}}$, $\Sigma_{k,\bar{s}|s} = \Sigma_{k,ss}^{-1}\Sigma_{k,s\bar{s}}$ and $\Sigma_{k,\bar{s}\bar{s}|s} = \Sigma_{k,\bar{s}\bar{s}} - \Sigma_{k,\bar{s}s}\Sigma_{k,ss}^{-1}\Sigma_{k,s\bar{s}}$. If these parameters $\mu_{k,\bar{s}|s}$, $\Sigma_{k,\bar{s}|s}$ and $\Sigma_{k,\bar{s}\bar{s}|s}$ are identical for all clusters, the identifiability cannot be ensured because the regression density of $\bar{s}$ on s can be factorized from the Gaussian mixture and regrouped with the regression density of S^c on R. This remark leads to the following identifiability theorem.

Theorem 3.5.1. *Let $\Theta_{(K,m,S,R)}$ be a subset of $\Upsilon_{(K,m,S,R)}$ whose elements $\theta = (\alpha, a, \beta, \Omega)$*

- *contain distinct couples (μ_k, Σ_k) fulfilling*

$$\forall s \subsetneq S, \exists (k, k'), 1 \leq k < k' \leq K; \mu_{k,\bar{s}|s} \neq \mu_{k',\bar{s}|s} \text{ or } \Sigma_{k,\bar{s}|s} \neq \Sigma_{k',\bar{s}|s} \text{ or } \Sigma_{k,\bar{s}\bar{s}|s} \neq \Sigma_{k',\bar{s}\bar{s}|s} \tag{3.5}$$

- *and for all variable blocks r of R, there exists a variable block l of S^c such that the restriction of the regression coefficient matrix β associated to r and l is not equal to zero.*

Let (K, m, S, R) and $(K^\star, m^\star, S^\star, R^\star)$ be two models. If there exist $\theta \in \Theta_{(K,m,S,R)}$ and $\theta^\star \in \Theta_{(K^\star,m^\star,S^\star,R^\star)}$ such that $f(.|\theta) = f(.|\theta^\star)$ then $(K, m, S, R) = (K^\star, m^\star, S^\star, R^\star)$ and $\theta = \theta^\star$ (up to a permutation of mixture components).

The proof of this theorem can be summarized as follows. First, each density $f(.|\theta)$ is written as a Gaussian mixture allowing us to use the identifiability of Gaussian mixture models. Thus, $K = K^\star$ and the parameters of the Gaussian mixture are equal, up to a permutation of mixture components. Second, it is proved by contradiction that $S \cap S^\star \neq \emptyset$ since the couples (μ_k, Σ_k) are not identical. Finally, it is deduced from (3.5) that the only possible case is $S = S^\star$ leading to $m = m^\star, R = R^\star$ and $\theta = \theta^\star$. The complete proof can be found in Appendix 3.C.

3.5.2 Consistency of our criterion

In this section, it is proved that the probability of selecting the true couple (S_0, R_0) by maximizing criterion (3.2) approaches 1 as $n \to \infty$ when the sampling distribution is one of the densities in competition and the true mixture model (K_0, m_0) is known. Denoting h the density function of the sample $\mathbf{y}$ and $\Theta_{(K,m,S,R)}$ being the subset defined in Theorem 3.5.1,

$$\begin{aligned}\theta^{\star}_{(K,m,S,R)} &= \underset{\theta_{(K,m,S,R)} \in \Theta_{(K,m,S,R)}}{\operatorname{argmin}} \mathrm{KL}[h, f(.|\theta_{(K,m,S,R)})] \\ &= \underset{\theta_{(K,m,S,R)} \in \Theta_{(K,m,S,R)}}{\operatorname{argmax}} \mathbb{E}_X\{\ln f(X|\theta_{(K,m,S,R)})\},\end{aligned}$$

where $\mathrm{KL}[h, f] = \int \ln\left\{\frac{h(x)}{f(x)}\right\} h(x)dx$ is the Kullback-Leibler divergence between the densities h and f, and

$$\hat{\theta}_{(K,m,S,R)} = \underset{\theta_{(K,m,S,R)} \in \Theta_{(K,m,S,R)}}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \ln\{f(\mathbf{y}_i|\theta_{(K,m,S,R)})\}.$$

The following assumption is considered:

(H1) The density h is assumed to be one of the densities in competition. By identifiability, there exists a unique model (K_0, m_0, S_0, R_0) and an associated parameter $\theta^{\star}_{(K_0,\, m_0,\, S_0,\, R_0)}$ such that $h = f(.|\theta^{\star}_{(K_0,\, m_0,\, S_0,\, R_0)})$. The couple (K_0, m_0) is supposed to be known.

To simplify the notation, all the dependencies over this couple (K_0, m_0) is omitted in the following. Moreover, an additional technical assumption is considered:

(H2) The vectors $\theta^{\star}_{(S,R)}$ and $\hat{\theta}_{(S,R)}$ are supposed to belong to a compact subspace $\Theta'_{(S,R)}$ of the following subset (included into $\Theta_{(S,R)}$),

$$\left(\begin{array}{l}\mathcal{P}_{K-1} \times \mathcal{B}(\eta, \mathrm{card}(S))^{K_0} \times \mathcal{D}^{K_0}_{\mathrm{card}(S)} \\ \times \mathcal{B}(\rho, 1, \mathrm{card}(S^c)) \times \mathcal{B}(\rho, \mathrm{card}(R), \mathrm{card}(S^c)) \times \mathcal{D}_{\mathrm{card}(S^c)}\end{array}\right) \cap\ \Theta_{(S,R)}$$

with

- $\mathcal{P}_{K-1} = \left\{(p_1, \ldots, p_K) \in [0,1]^K; \sum_{k=1}^{K} p_k = 1\right\}$ denotes the $K-1$ dimensional simplex containing the considered proportion vectors,
- $\mathcal{B}(\eta, r)$ is the closed ball in $\mathbb{R}^r$ of radius η centered at zero for the l^2-norm defined by $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{r} x_i^2}, \forall \mathbf{x} \in \mathbb{R}^r$,
- $\mathcal{B}(\rho, r, q)$ is the closed ball in $\mathcal{M}_{r\times q}(\mathbb{R})$ of radius ρ centered at zero for the matricial norm $|||.|||$ defined by

$$\forall A \in \mathcal{M}_{r\times q}(\mathbb{R}), |||A||| = \sup_{\|\mathbf{x}\|=1} \|\mathbf{x}A\|,$$

- $\mathcal{D}_r$ is the set of the $r \times r$ positive definite matrices with eigenvalues in $[s_{\mathrm{m}}, s_{\mathrm{M}}]$ with $0 < s_{\mathrm{m}} < s_{\mathrm{M}}$.

Theorem 3.5.2. *Under assumptions (H1),(H2), the couple of variable sets $(\hat{S}, \hat{R})$ maximizing Criterion (3.2) with fixed (K_0, m_0) is such that $P((\hat{S}, \hat{R}) = (S_0, R_0)) \underset{n\to\infty}{\to} 1$.*

The complete proof of Theorem 3.5.2 is given in Appendix 3.D.

3.6 Method validation

Through the presentation of simulated experiments, we first want to highlight the difference with Raftery and Dean's method. Second, we illustrate the variable selection gain for clustering and its interpretation. Third, we discuss the robustness of our methodology.

3.6.1 Comparison with Raftery and Dean's method

The dataset consists of 2000 data points from a mixture of four Gaussian distributions $\mathcal{N}(\mu_k, I_2)$ with $\mu_1 = (-2, -2)$, $\mu_2 = (-2, 2)$, $\mu_3 = (2, -2)$, $\mu_4 = (2, 2)$ and with a proportion vector $\mathbf{p} = (0.3, 0.2, 0.3, 0.2)$ (see Figure 3.2). Eight irrelevant variables are appended, simulated according to $\mathbf{y}_i^{\{3,\dots,10\}} = \mathbf{y}_i^{\{1,2\}}\beta + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \Omega)$. Different scenarii for the eight variables are proposed, ranging from all are independent of the relevant variables to all depend on the relevant variables (see Table 3.1). In all scenarii, the true mixture is $(K_0 = 4, m_0 = [p_k LI])$ and the true relevant variable subset is $S_0 = \{1, 2\}$. Only the subset R_0 changes according to the scenario. Raftery and Dean's algorithm, called CLUSTVARSEL (Raftery and Dean, 2006a) and our algorithm are compared on these different scenarii.

Both algorithms choose the true number of components $\hat{K} = 4$ in all scenarii. Raftery and Dean's procedure selects all variables when the irrelevant ones are simulated from $\mathcal{N}(0, 1)$ but selects the true relevant variables when the noise is lower (Scenarii 1 and 2). For Scenarii 3-5 where the number of independent variables is larger than the regressed variables, their procedure selects the independent variables. Nevertheless their method has a good behaviour when the number of regressed variables is larger (Scenarii 6 and 7). Our procedure selects the true variable partition for all scenarii, but for the two last scenarii, it selects a too complex mixture form. This result seems to be related to numerical variability due to the two nested backward stepwise algorithms making up our variable selection algorithm.

3.6.2 Variable selection interest

Each dataset consists of 800 data points from a mixture of four equiprobable Gaussian distributions $\mathcal{N}(\mu_k, \Sigma_k)$ with $\mu_1 = (-c, -c)$, $\mu_2 = (-c, c)$, $\mu_3 = (c, -c)$, $\mu_4 = (c, c)$ where $c \in \{2, 3, 5\}$ and $\Sigma_1 = \mathrm{diag}(1, 2)$, $\Sigma_2 = \mathrm{diag}(3, 0.5)$, $\Sigma_3 = I_2$ and $\Sigma_4 = \Sigma_1$. The third variable is defined by $\mathbf{y}^3 = 3\mathbf{y}^1 + \varepsilon$, ε being sampled from a $\mathcal{N}(0, 0.5I_{800})$ density. Five noisy independent standard centered Gaussian variables are also appended. The true model

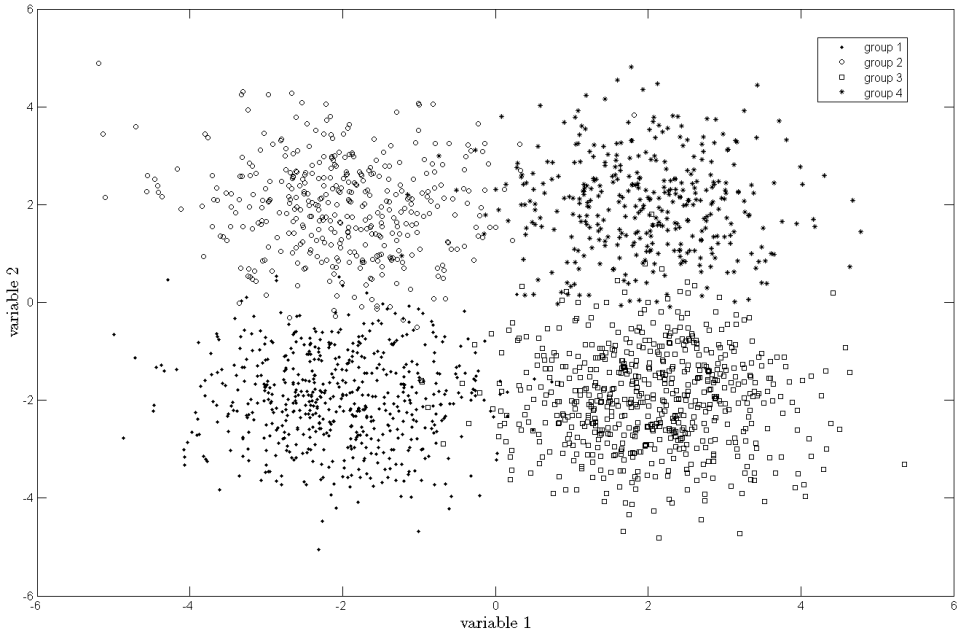


Figure 3.2: Simulated dataset representation according to the two relevant variables.

	Raftery & Dean		Our algorithm		
Scenario	$\hat{m}$	$\hat{S}$	$\hat{m}$	$\hat{S}$	$\hat{R}$
1: $\beta = 0_8$, $\Omega = I_8$	$[p_k LI]$	$\{1-10\}$	$[p_k LI]$	$\{1,2\}$	$\emptyset$
2: $\beta = 0_8$, $\Omega = 0.5I_8$	$[p_k LI]$	$\{1,2\}$	$[p_k LI]$	$\{1,2\}$	$\emptyset$
3: $\beta = ((2,0)', 0_7)$, $\Omega = I_8$	$[p_k LI]$	$\{1,2,4-10\}$	$[p_k LI]$	$\{1,2\}$	$\{1\}$
4: $\beta = ((0.5,0)', (0,1)', 0_6)$, $\Omega = I_8$	$[p_k LI]$	$\{1,2,5-10\}$	$[p_k LI]$	$\{1,2\}$	$\{1,2\}$
5: $\beta = (\beta_1, 0_4)$,$\Omega = \mathrm{diag}(I_2, 0.5I_2, I_4)$	$[p_k LI]$	$\{1,2,7-10\}$	$[p_k LI]$	$\{1,2\}$	$\{1,2\}$
6: $\beta = (\beta_1, \beta_2, 0_2)$, $\Omega = \mathrm{diag}(I_2, 0.5I_4, I_2)$	$[p_k LI]$	$\{1,2\}$	$[p_k LC]$	$\{1,2\}$	$\{1,2\}$
7: $\beta = (\beta_1, \beta_2, \beta_3)$, $\Omega = \mathrm{diag}(I_2, 0.5I_4, I_2)$	$[p_k LI]$	$\{1,2\}$	$[p_k LC]$	$\{1,2\}$	$\{1,2\}$

Table 3.1: Model selection by Raftery and Dean's and our algorithms in seven scenarii where $\beta_1 = ((0.5,0)', (0,1)', (2,0)', (0,3)')$, $\beta_2 = ((2,0.5)', (0.5,1)')$, $\beta_3 = ((2,0)', (0,3)')$, 0_p is the $2 \times p$ zero matrix. In all cases, both methods select a mixture with $\hat{K} = 4$ components.

is $(K_0 = 4, m_0 = [pL_kB_k], S_0 = \{1, 2\}, R_0 = \{1\})$. Results, summarized in Table 3.2, show that our procedure selects the true variable partition, the true number of clusters and a diagonal model, for the three values of c. Compared to MIXMOD (Biernacki et al., 2006), which does not proceed to a variable selection, our method provides the smallest clustering error rates, especially when c decreases. Raftery and Dean's method finds also the same relevant variable subset in this example but our method gives additional information about the variable role.

	our algorithm					MIXMOD		
c	$\hat{K}$	$\hat{m}$	$\hat{S}$	$\hat{R}$	error rate	$\hat{K}$	$\hat{m}$	error rate
5	4	$[pL_kB_k]$	{1,2}	{1}	0%	4	$[pL_kC_k]$	0%
3	4	$[pL_kB_k]$	{1,2}	{1}	0.875%	5	$[p_kLC]$	2.38%
2	4	$[pLB_k]$	{1,2}	{1}	9.25%	5	$[pLC]$	13.88%

Table 3.2: Results with our algorithm and MIXMOD for this simulated example.

3.6.3 Waveform dataset

This experiment allows us to assess the method behaviour on a non Gaussian mixture dataset. It is extracted at random from the waveform dataset, available at the UCI repository (Blake et al., 1999). It consists of 900 observations divided into three equiprobable groups, based on a random convex combination of two of three waveforms sampled at integers $\{1, \ldots, 21\}$ with noise added. Nineteen noisy standard centered Gaussian variables are appended. A detailed description is available in Breiman et al. (1984). We compare our method with the one of Raftery and Dean for $K \in \{3, 4, 5, 6\}$ with twenty models (spherical models, diagonal models and models with the following form $[p\_L\_C\_]$, see Table 1.1). The selected model with our method is

$$(\hat{K} = 6, \hat{m} = [pLI], \hat{S} = \{3, 4, 6 - 15, 18, 19\}, \hat{R} = \{8, 11, 15\}).$$

All noisy variables and variables $\{1, 2, 5, 16, 17, 20, 21\}$ are declared irrelevant. The final clustering is coherent with the construction of the sample: Three clusters correspond to the three wave functions and the three others are convex combinations of two wave functions (see Figure 3.3). Raftery and Dean's method selects a mixture $(K, m) = (3, [pL_kB])$ and declares that all variables are relevant except Variables 5 and 16. Their resulting clustering reveals the three wave functions but not their convex combinations.

We now study the complete waveform dataset (5000 observations) in the same conditions. The following selected model with our method is similar to the previous selected model:

$$(\hat{K} = 6, \hat{m} = [p_kLC], \hat{S} = \{4 - 18\}, \hat{R} = \{7, 11, 15\})$$

and the associated clustering is always coherent with the sample construction (see Table 3.3). In this case where the sample size is larger, Raftery and Dean's method has a better behaviour since it selects a mixture $(K, m) = (6, [p_kLC])$ and declares that the

relevant variable subset is $\hat{S} = \{4 - 19\}$. In our opinion, our more realistic model is able to detect easier the variable roles for smaller datasets although our algorithm is composed of two backward stepwise algorithms which can imply more variability.

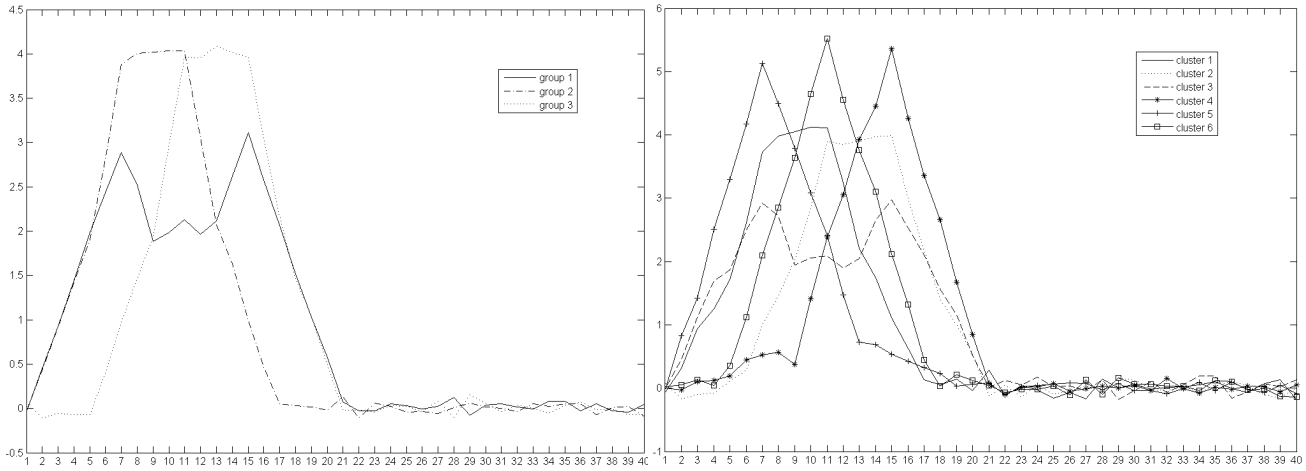


Figure 3.3: On the left, average profiles of the three real groups and on the right, average profiles of the six clusters found with our variable selection procedure.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Total
Group 1	590	589	6	6	0	501	1692
Group 2	573	1	0	598	481	0	1653
Group 3	0	2	575	2	512	564	1655
Total	1163	592	581	606	993	1065	5000

Table 3.3: Clustering of the 5000 observations into the six clusters.

3.7 Analysis of transcriptome data

As explained in the introduction, variable selection is desirable in cluster analysis of transcriptome data. We illustrate it by studying a transcriptome dataset of *Arabidopsis thaliana*, extracted from the database CATdb (Gagnot et al., 2008). In this database, data are organized by project, each project being composed of a set of experiments dedicated to a specific biological question. We focus on 1020 genes of *Arabidopsis thaliana* declared differentially expressed at least once in a time course of the hypocotyl growth switch (Project 6 in Table 3.4) and we study their behaviour in other projects (see Table 3.4): We study $n = 1020$ genes described by $Q = 27$ experiments partitioned into $T = 7$ variable blocks. Gene i is described with a vector $\mathbf{y}_i' \in \mathbb{R}^{27}$, the component y_i^j corresponding to the test statistic calculated in the experiment j for the differential analysis (see Chapter 2 for details on the normalization and the differential analysis steps).

Project	exper. num.	aim of the project
1	8	transcriptome of the circadian cycle
2	4	transcriptome response to iron signaling
3	4	transcriptome profiling from a protoplast culture
4	3	transcriptome profiling from cell division to differentiation
5	3	transcriptome response to nematode infection
6	3	transcriptome time course of the hypocotyl growth switch
7	2	transcriptome of the hypocotyl growth switch to isoxaben treatment

Table 3.4: Description of the transcriptome projects used to define the seven variable blocks. The number of experiments and the aim of each project are given in Columns 2 and 3 respectively.

cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
gene number	7	120	25	11	42	424	14	29	43	94	11	149	19	4	6	2	20

Table 3.5: Size of the 17 estimated clusters with variable selection on the transcriptome dataset.

clustering without variable selection

	Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Cluster	nb of genes	21	3	27	18	144	10	35	46	15	8	397	5	97	13	40	141
1	7		1		6												
2	120					107			1		6			1	5		
3	25			22			1					2					
4	11		2		6				2			1					
5	42				1	2	5	2	29			3					
6	424			2	3	2	3	26	1			373		11	1	2	
7	14								12						1	1	
8	29					11		3				11				1	3
9	43	1				6	1					1				33	1
10	94				1	9						1		83			
11	11							2				4	1		2	2	
12	149					5		1			2			2	2		137
13	19					1			1	15					2		
14	4			2	1			1									
15	6					1						1	4				
16	2	1		1													
17	20	19														1	

clustering with variable selection

Figure 3.4: Comparison of the two transcriptome dataset clusterings with and without variable selection.

In this example, Gaussian mixtures with equal volumes and a maximal number of mixture components fixed to $K_{\max} = 20$ (see Table 1.1) are considered. Using all variables, the selected model is $[p_k LC]$ with 16 clusters. When our variable selection procedure is performed, the selected mixture model is $[p_k LC]$ with $\hat{K} = 17$ clusters with different sizes (see Table 3.5). The relevant variable blocks are Projects 1, 3, 4, 6 and 7 and the four last ones enter in the regression model. Some gene subsets are common to the two clusterings (see Figure 3.4). Nevertheless with our variable selection procedure, clusters seem to be more homogeneous than without variable selection. Some clusters interesting from a biological point of view are highlighted (see for example Figure 3.7).

Concerning the variable selection, Project 6, used to define the gene subset, has been declared relevant for the clustering, as Project 7, also related to the hypocotyl growth switch. Regression parameters are given in Table 3.6. The two irrelevant projects are related to stress conditions: Project 5 investigates root-knot nematode infection, known to induce the redifferentiation of root cells and formation of giant cells (Jammes et al., 2005). The cell redifferentiation and the cell elongation are two phenomena also studied in Project 3 and Projects 6, 7 respectively. Project 2 investigates iron stress of cells. It is mainly explains by Projects 3 and 6 but no biological interpretation is yet available. Since the Project 5 consists of a time-course project, it is logical to observe large correlations between experiments of this project. Project 2 is also a time-course project but the first and the fourth experiment, corresponding to a precocious and a tardive response to iron stress respectively, are few related to the two other experiments as illustrated in the correlation matrix. Moreover, the fact that these two irrelevant projects study different tissus can explained that they are little correlated. Finally we explore whether the irrelevant projects offer a different gene clustering and we find six clusters poorly related to the 17 clusters previously discussed.

Sandra Pelletier (URGV) studied 1359 genes which are differentially expressed at least once in the time course of the hypocotyl growth switch (Project 6). In order to understand the biological mechanisms occurring in each phase of the growth switch, she divided these 1359 genes into nine expression profiles of interest in Project 6 (see Figure 3.5). Then she studied the expression profiles of these genes in 227 other experiments, available in CATdb, with an exploratory method. She underlines 11 subgroups of coexpressed genes, characterized by dissimilar expression profiles in this time course. Note that Subgroups 5a2 and 5a3 having the same profile 5a in Project 6 are distinguished since they have different characteristic behaviours in other experiments. The genes of Subgroup 5b2 have the profile 5b in Project 6 but are characterized by a specific behaviour in Project 7 (their profiles increase in the second experience). Some of these 11 gene subgroups are recovered among the 17 clusters obtained with our procedure (see Figure 3.6). For instance, the genes of Subgroup 2a are clustered with our procedure in Cluster 17 whose the interesting expression profile is given in Figure 3.7(a). Genes of Subgroups 5a2 and 5a3 are globally separate in our clustering. Subgroup 5b2 is totally contained in Cluster 13 whose the profile highlights the specific behaviour of these genes in Project 7 (see Figure 3.7(c)). These genes are clustered with 15 other genes which are not discovered by the exploratory method. Thus the result of our procedure allows biologists to formulate new assumptions.

	P2-1	P2-2	P2-3	P2-4	P5-1	P5-2	P5-3
a	-0.06	0.12	-0.02	0,26	0.38	-0.10	-0,45
P3-1	-0.05	0.00	-0.03	-0.02	0.10	0.12	0.16
P3-2	0.10	0.21	0.19	-0.05	-0.06	0.28	0.19
P3-3	-0.06	-0.12	-0.20	-0.03	0.01	-0.10	-0.06
P3-4	0.00	0.07	0.13	-0.07	-0.05	0.24	0.22
P4-1	-0.05	0.05	0.08	0.00	-0.15	-0.02	-0.01
P4-2	0.02	-0.03	-0.08	0.02	0.09	-0.01	-0.02
P4-3	-0.05	0.07	0.12	0.00	-0.17	-0.09	-0.11
P6-1	0.08	0.14	0.12	-0.02	0.01	0.14	0.17
P6-2	-0.10	-0.12	-0.10	0.07	-0.16	-0.11	-0.05
P6-3	0.04	0.07	0.03	-0.03	0.03	-0.03	-0.11
P7-1	-0.06	0.03	0.03	0.10	-0.29	-0.25	-0.20
P7-2	-0.05	-0.09	-0.23	-0.02	0.14	-0.05	0.12

	P2-1	P2-2	P2-3	P2-4	P5-1	P5-2	P5-3
P2-1	2.18	0.29	-0.59	-1.05	0.46	0.48	0.25
P2-2	0.29	2.77	3.07	0.28	-0.56	0.03	-0.27
P2-3	-0.59	3.07	9.95	2.03	-0.31	0.20	-0.23
P2-4	-1.05	0.28	2.03	3.34	-0.34	-0.18	-0.29
P5-1	0.46	-0.56	-0.31	-0.34	9.51	6.65	7.32
P5-2	0.48	0.03	0.20	-0.18	6.65	13.59	12.62
P5-3	0.25	-0.27	-0.23	-0.29	7.32	12.62	17.26

	P2-1	P2-2	P2-3	P2-4	P5-1	P5-2	P5-3
P2-1	1	0.12	-0.13	-0.39	0.10	0.09	0.04
P2-2	0.12	1	0.59	0.09	-0.11	0	-0.04
P2-3	-0.13	0.59	1	0.35	-0.03	0.02	-0.02
P2-4	-0.39	0.09	0.35	1	-0.06	-0.03	-0.04
P5-1	0.10	-0.11	-0.03	-0.06	1	0.59	0.57
P5-2	0.09	0	0.02	-0.03	0.59	1	0.82
P5-3	0.04	-0.04	-0.02	-0.04	0.57	0.82	1

Table 3.6: Estimated regression parameters for the transcriptome dataset: on the left, the regression coefficient $(\hat{a}', \hat{\beta}')'$ and on the right, the variance matrix $\hat{\Omega}$ on the top and the correlation matrix on the bottom. $Pi - j$ denotes Experiment j in Project i.

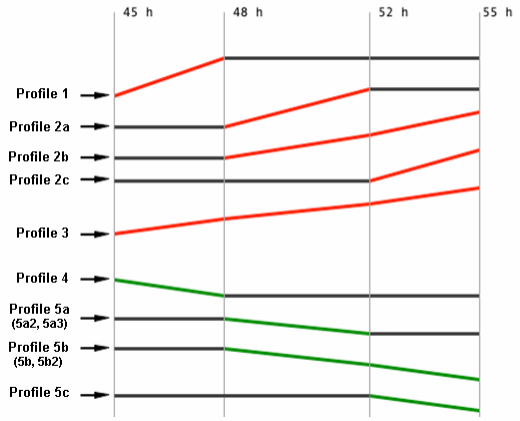
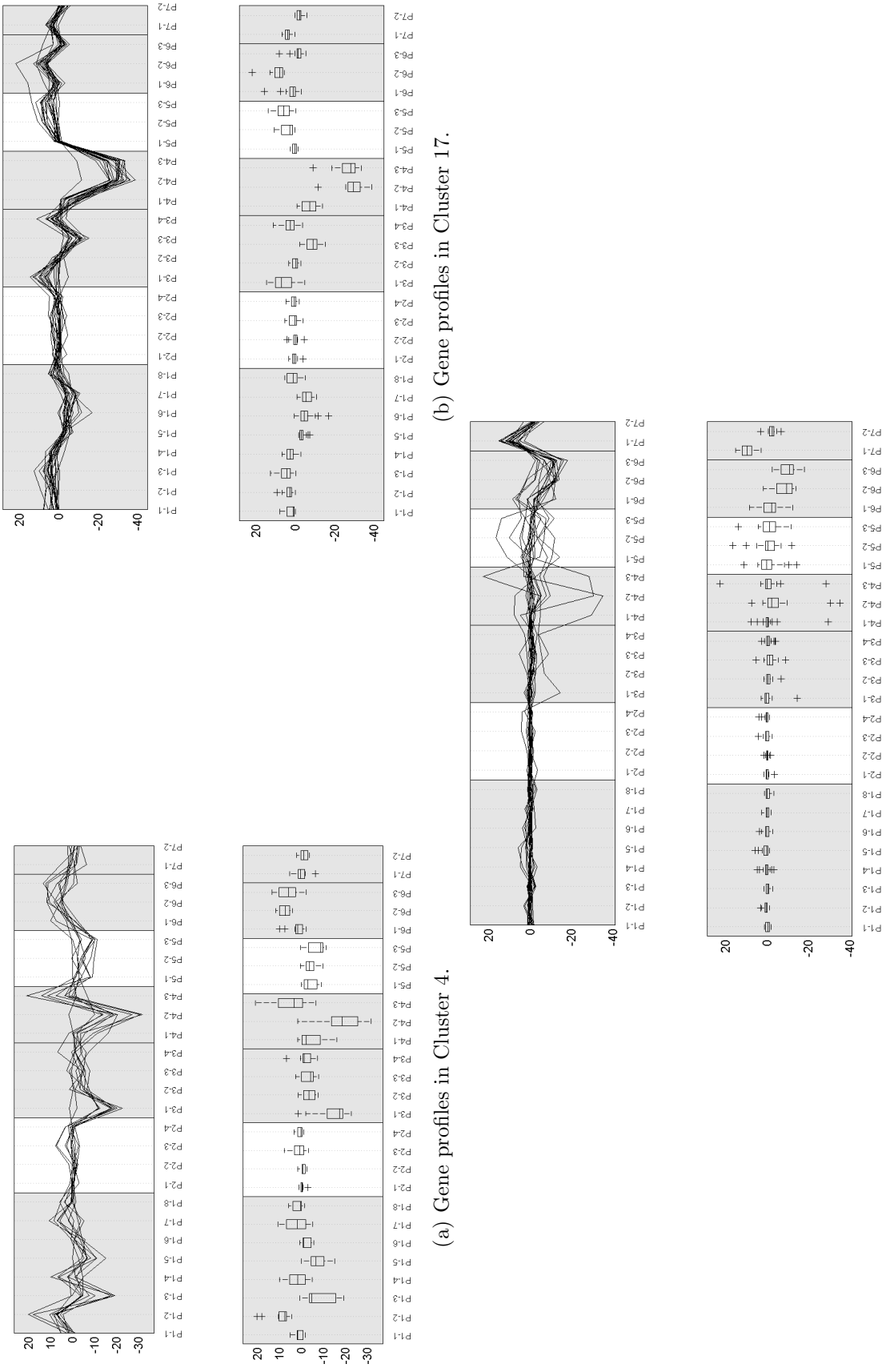


Figure 3.5: Specific profiles in Project 6 studied by S. Pelletier.

	clusters	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
subgroup	nb of genes	7	120	25	11	42	424	14	29	43	94	11	149	19	4	6	2	20
1	3										3							
2a	17									3								14
2b	11			2	1		8											
2c	3						2				1							
3	9						6		1			2						
4	8					4	1	3										
5a2	15												15					
5a3	8										6		2					
5b	15												15					
5b2	4													4				
5c	22		2							15			4		1			

Figure 3.6: Comparison between the clustering obtained by our procedure and the known gene subgroups.



(a) Gene profiles in Cluster 4.

(b) Gene profiles in Cluster 17.

(c) Gene profiles in Cluster 13.

Figure 3.7: Graphical representation of gene profiles in three clusters. Relevant projects are colored in grey and, on the x-axis, $Pi - j$ denotes Experiment j in Project i.

3.8 Discussion

We have presented a general variable selection methodology for cluster analysis when the individual number is greater than the variable number. Following Raftery and Dean (2006b), this methodology considers the problem in the model-based cluster analysis context. In our approach, variables could be partitioned into blocks and the role of the clustering variables with respect to the other variables is more versatile and is expected to be more realistic. On the theoretical side, we have established the model identifiability and proved the consistency of our variable selection criterion under reasonable assumptions. As a by product, we have proved the identifiability of Raftery and Dean's models and the consistency of their variable selection criterion could be proved easily following the same line of proof. Compared to Raftery and Dean (2006b), our more general definition of variable role could avoid to overpenalize models with independent variables as illustrated in the numerical experiments. Nevertheless, the variable partition modelling can be again improved by distinguish between the irrelevant variables which are totally independent of the relevant variables and the ones which are dependent of a subset of the relevant variables. This improvement is addressed in Chapter 4.

One of the interests of our model is to allow for a better and, sometimes subtle, interpretation of the variable role. Thus, this method can be regarded as promising in the field of microarray gene expression dataset analysis where the behaviour of several thousand genes are described by an increasing number of experiments. Nevertheless, these gene expression datasets have often missing values due to various reasons in the laboratory process. Thus an extension of our method for such datasets is proposed in Chapter 5 for a more realistic analysis of gene expression datasets.

Finally, we want to stress that the defined procedure can work with alternative models linking the clustering and remaining variables, provided that a BIC-like criterion analogous with our $\mathrm{BIC}_{\mathrm{reg}}$ criterion can be computed. Moreover, it is possible to base the criterion construction on the integrated complete likelihood instead of the integrated likelihood, in order to take the clustering aim into account. Briefly in this case, the chosen model $(\tilde{K}, \tilde{m}, \tilde{S}, \tilde{R})$ maximizes the integrated complete likelihood

$$\begin{aligned} f(\mathbf{y}, \mathbf{z}|K, m, S, R) &= \int f(\mathbf{y}, \mathbf{z}|K, m, S, R, \theta)\pi(\theta|K, m, S, R)d\theta \\ &= \int f_{\mathrm{clust}}(\mathbf{y}^S, \mathbf{z}|K, m, \alpha)\pi(\alpha|K, m)d\alpha \\ &\quad \times \int f_{\mathrm{reg}}(\mathbf{y}^{S^c}|a + \mathbf{y}^R\beta, \Omega)\pi(a, \beta, \Omega)dad\beta d\Omega \end{aligned}$$

where $\mathbf{z}$ is the label vector (see Section 3.2). The second term of the right-hand side can be still approximated by to $\mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^R)$. The first term, corresponding to the integrated complete likelihood of a sample distributed from a Gaussian mixture, can be approximated

by an ICL-type criterion (see Chapter 1) defined by

$$\mathrm{ICL}_{\mathrm{clust}}(\mathbf{y}^S|K,m) = \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S|K,m) + 2\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{z}_{ik}\ln[t_{ik}(\hat{\alpha})]$$

where the conditional probabilities

$$t_{ik}(\hat{\alpha}) = \frac{\hat{p}_k\Phi(\mathbf{y}_i^S|\hat{\mu}_k,\hat{\Sigma}_k)}{\sum_{l=1}^{K}\hat{p}_l\Phi(\mathbf{y}_i^S|\hat{\mu}_l,\hat{\Sigma}_l)}.$$

The selected model $(\hat{K},\hat{m},\hat{S},\hat{R})$ thus maximizes the criterion $\mathrm{ICL}_{\mathrm{clust}}(\mathbf{y}^S|K,m)+\mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^R)$ and the variable selection algorithm can be easily adapted.

Appendices

3.A Multidimensional Multivariate Regression

Let H be a $n\times V$ observed matrix of V response variables on n individuals. Let M be a $n\times A$ known matrix which represents a matrix of A observed variables on the n individuals. The regression model is defined by

$$\begin{aligned} H &= (1,\ldots,1)'a + M\beta + E \\ &= XB + E, \end{aligned}$$

with

$$H = \begin{pmatrix} H_1^1 & \ldots & H_1^V \\ \vdots & & \vdots \\ H_n^1 & \ldots & H_n^V \end{pmatrix}, X = \left(\begin{array}{c|ccc} 1 & M_1^1 & \ldots & M_1^A \\ \vdots & \vdots & & \vdots \\ 1 & M_n^1 & \ldots & M_n^A \end{array}\right) \text{ and } B = \left(\begin{array}{ccc} a_1 & \ldots & a_V \\ \hline \beta_1^1 & \ldots & \beta_1^V \\ \vdots & & \vdots \\ \beta_A^1 & \ldots & \beta_A^V \end{array}\right)$$

and E such that $E_i \sim \mathcal{N}_V(0,\Omega)$. It is proved in Mardia et al. (1979, Theorem 6.2.1) or Anderson (2003) that if $(X'X)^{-1}$ exists then defining $P = I - X(X'X)^{-1}X'$, the maximum likelihood estimates of B and Ω are

$$\hat{B} = (X'X)^{-1}X'H \qquad \hat{\Omega} = \frac{1}{n}H'PH. \tag{3.6}$$

BIC criterion for multidimensional multivariate regression is now derived. The model likelihood for data H is

$$f(H|X,B,\Omega) = |2\pi\Omega|^{-n/2}\exp\left[-\frac{1}{2}\,\mathrm{tr}\{(H-XB)\Omega^{-1}(H-XB)'\}\right].$$

The integrated likelihood is defined by

$$f(H|X) = \int f(H|X, B, \Omega)\pi(B, \Omega)d(B, \Omega)$$

where π is the prior distribution of the parameters. It can be written

$$f(H|X) = \int e^{nL_n(B,\Omega)}d(B, \Omega)$$

where

$$nL_n(B, \Omega) = -\frac{n}{2}\ln\{|2\pi\Omega|\} - \frac{1}{2}\operatorname{tr}\left\{(H - XB)\Omega^{-1}(H - XB)'\right\} + \ln\{\pi(B, \Omega)\}.$$

Using Laplace approximation along a line detailed in Burnham and Anderson (2002) or in Lebarbier and Mary-Huard (2006), we get

$$f(H|X) = e^{nL_n(B^\star,\Omega^\star)}\left(\frac{2\pi}{n}\right)^{\nu/2} | - L_n''(B^\star, \Omega^\star)|^{-1/2}\{1 + \mathcal{O}(n^{-1/2})\},$$

where $(B^\star, \Omega^\star)$ maximize $L_n(B, \Omega)$ and $\nu = (A + 1)V + \frac{V(V+1)}{2}$ is the number of free parameters in the regression model. Replacing $(B^\star, \Omega^\star)$ with $(\hat{B}, \hat{\Omega})$ maximizing $f(H|X, B, \Omega)$ defined by (3.6) and $| - L_n''(B^\star, \Omega^\star)|$ with the Fisher information $I_{(\hat{B},\hat{\Omega})}$ which must be bounded, we get

$$2\ln\{f(H|X)\} = 2nL_n(\hat{B}, \hat{\Omega}) + \nu\ln\left(\frac{2\pi}{n}\right) - \ln\{I_{(\hat{B},\hat{\Omega})}\} + \mathcal{O}(n^{-1/2}).$$

From

$$2nL_n(\hat{B}, \hat{\Omega}) = -n\ln(|2\pi\hat{\Omega}|) - \operatorname{tr}\left\{(H - X\hat{B})\hat{\Omega}^{-1}(H - X\hat{B})'\right\} + 2\ln\{\pi(\hat{B}, \hat{\Omega})\}$$

and noticing that

$$\operatorname{tr}\left\{(H - X\hat{B})\hat{\Omega}^{-1}(H - X\hat{B})'\right\} = nV,$$

we get

$$\begin{aligned} 2\ln\{f(H|X)\} &= -n\ln(|2\pi\hat{\Omega}|) - nV + \nu\ln\left(\frac{2\pi}{n}\right) - \ln\{I_{(\hat{B},\hat{\Omega})}\} + 2\ln\{\pi(\hat{B}, \hat{\Omega})\} + \mathcal{O}(n^{-1/2}) \\ &\approx -n\ln[(|2\pi\hat{\Omega}|) - nV - \nu\ln(n). \end{aligned}$$

We conclude that the BIC criterion for multivariate regression is

$$\mathrm{BIC}_{\mathrm{reg}}(H|X) = -n\ln(|2\pi\hat{\Omega}|) - nV - \nu\ln(n)$$

and in the simple regression context, ($V = 1$ and $\Omega = \sigma^2 > 0$), it becomes

$$\mathrm{BIC}_{\mathrm{reg}}(H|X) = -n\ln(2\pi\hat{\sigma}^2) - n - (A + 2)\ln(n).$$

3.B The backward variable selection in regression

The following algorithm allows to determine the subset $R[\ell]$ of variables among S required to explain a variable $\mathbf{y}^\ell$ with a linear regression. The model comparison is performed with criterion $\mathrm{BIC}_{\mathrm{reg}}$ defined in (3.3). The algorithm is making use of exclusion and inclusion steps now described.
Initialisation: $R[\ell] = S$, $j_E = \emptyset$ and $j_I = \emptyset$.
Exclusion step: For all j in $R[\ell]$, compute $\mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^j) = \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^\ell|\mathbf{y}^{R[\ell]}) - \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^\ell|\mathbf{y}^{R[\ell]\setminus j})$. Then, compute $j_E = \underset{j\in R[\ell]}{\operatorname{argmin}}\ \mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^j)$.

- If $\mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^{j_E}) \leq 0$, set $R[\ell] = R[\ell]\setminus j_E$ and go to the inclusion step if $j_E \neq j_I$ or stop otherwise.
- otherwise go to the inclusion step if $j_I \neq \emptyset$ or stop otherwise.

Inclusion step: For all j in $S\setminus R[\ell]$, compute $\mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^j) = \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^\ell|\mathbf{y}^{R[\ell]\cup j}) - \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^\ell|\mathbf{y}^{R[\ell]})$. Then, compute $j_I = \underset{j\in S\setminus R[\ell]}{\operatorname{argmax}}\ \mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^j)$.

- If $\mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^{j_I}) > 0$, $R[\ell] = R[\ell] \cup j_I$ and go to the exclusion step if $j_I \neq j_E$ or stop otherwise.
- otherwise go to the exclusion step.

Starting from the exclusion step, the backward variable selection algorithm consists of alternating the exclusion and the inclusion steps.

3.C Proof by contradiction of the model identifiability theorem

Recall that $\Phi(.|\mu_k, \Sigma_k)$ denotes the Q-dimensional Gaussian density with mean μ_k and variance matrix Σ_k. The mixture parameters can be decomposed into $\mu_k = (\mu_{ks}, \mu_{k\bar{s}})$ and Σ_k into submatrices $\Sigma_{k,ss}$, $\Sigma_{k,s\bar{s}}$ and $\Sigma_{k,\bar{s}\bar{s}}$, where s is a nonempty subset of S and $\bar{s}$ its complement in S. Moreover, conditional parameters are defined by $\mu_{k,\bar{s}|s} = \mu_{k\bar{s}} - \mu_{ks}\Sigma_{k,ss}^{-1}\Sigma_{k,s\bar{s}}$, $\Sigma_{k,\bar{s}|s} = \Sigma_{k,ss}^{-1}\Sigma_{k,s\bar{s}}$ and $\Sigma_{k,\bar{s}\bar{s}|s} = \Sigma_{k,\bar{s}\bar{s}} - \Sigma_{k,\bar{s}s}\Sigma_{k,ss}^{-1}\Sigma_{k,s\bar{s}}$. For the regression part, we assume that for every variable block $r \in R$, there exists a variable block $j \in S^c$ such that the restriction of β associated to r and j, denoted β_{rj} is a nonzero matrix. It can be also rewritten $a + x^R\beta = a + x^S\Lambda$, where the matrix Λ is defined by

$$\forall j \in S^c, \Lambda_{lj} = \begin{cases} \beta_{lj} & \text{for } l \in R, \\ 0 & \text{for } l \in S\setminus R. \end{cases}$$

For $s \subseteq S$ and $t \subseteq S^c$, those restricted matrices are defined by $\Lambda_{st} = (\Lambda_{pq})_{p\in s, q\in t}$, $\Lambda_{.t} = \Lambda_{St}$ and $\Lambda_{s.} = \Lambda_{sS^c}$. Finally, recall that in the parameter vector $\theta = (\alpha, a, \beta, \Omega)$ where

$\alpha = (p_1, \ldots, p_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$, the couples (μ_k, Σ_k) are distinct and fulfill the condition

$$\forall s \subsetneq S, \exists (k, k'), 1 \leq k < k' \leq K; \mu_{k,\bar{s}|s} \neq \mu_{k',\bar{s}|s} \text{ or } \Sigma_{k,\bar{s}|s} \neq \Sigma_{k',\bar{s}|s} \text{ or } \Sigma_{k,\bar{s}\bar{s}|s} \neq \Sigma_{k',\bar{s}\bar{s}|s}. \tag{3.7}$$

Proof. First, the density functions $f(.|\theta)$ and $f(.|\theta^\star)$ are written as standard Gaussian mixtures:

$$\begin{aligned} f(x|\theta) &= f(x|\theta, K, m, S, R) \\ &= f_{\text{clust}}(x^S|K, m, \alpha) f_{\text{reg}}(x^{S^c}|a + x^R\beta, \Omega) \\ &= \sum_{k=1}^{K} p_k \Phi(x^S|\mu_k, \Sigma_k) \Phi(x^{S^c}|a + x^S\Lambda, \Omega) \\ &= \sum_{k=1}^{K} p_k \Phi(x^S, x^{S^c}|\nu_k, \Delta_k) \end{aligned}$$

where $\nu_k = (\mu_k, a + \mu_k\Lambda)$ and $\Delta_k = \begin{pmatrix} \Sigma_k & \Sigma_k\Lambda \\ \Lambda'\Sigma_k & \Omega + \Lambda'\Sigma_k\Lambda \end{pmatrix}$ according to Lemma 3.C.1 stated hereafter. In the same way, $f(x|\theta^\star) = \sum_{k=1}^{K^\star} p_k^\star \Phi(x^{S^\star}, x^{S^{\star\,c}}|\nu_k^\star, \Delta_k^\star)$. Then, since the couples (ν_k, Δ_k) are distinct as well as $(\nu_k^\star, \Delta_k^\star)$, the identifiability of Gaussian mixture models gives that $K = K^\star$ and up to a permutation of mixture components, $p_k = p_k^\star$, $\nu_k = \nu_k^\star$ and $\Delta_k = \Delta_k^\star$ (see for instance McLachlan and Peel, 2000).

Second, assume that $S \cap S^\star = \emptyset$ and consider the subsets $s = S^\star \cap S^c$ and $t = S^{\star\,c} \cap S$. The equality of covariance matrices Δ_k and $\Delta_k^\star$ on s and between s and t gives respectively for all k,

$$\begin{cases} \Omega_{ss} + \Lambda'_{.s}\Sigma_k\Lambda_{.s} = \Sigma_k^\star \\ \Sigma_k^\star\Lambda_{.t}^\star = \Lambda'_{.s}\Sigma_k \end{cases}.$$

According to these two equalities, it is deduced that $\Omega_{ss} = \Sigma_k^\star(I - \Lambda_{.t}^\star\Lambda_{.s})$ for all k. Since Ω_{ss} and $\Sigma_k^\star$ are positive definite matrices, $I - \Lambda_{.t}^\star\Lambda_{.s}$ is a nonsingular matrix and consequently all covariance matrices $\Sigma_k^\star = \Omega_{ss}(I - \Lambda_{.t}^\star\Lambda_{.s})^{-1}$ are equal. Moreover, the equality of mean vectors on s and t gives

$$\begin{cases} \mu_k^\star = a_s + \mu_k\Lambda_{.s} \\ \mu_k = a_t^\star + \mu_k^\star\Lambda_{.t}^\star \end{cases},$$

implying that $\mu_k^\star(I - \Lambda_{.t}^\star\Lambda_{.s}) = a_s + a_t^\star\Lambda_{.s}$, for all k. Since $I - \Lambda_{.t}^\star\Lambda_{.s}$ is nonsingular, all $\mu_k^\star$ are equal. This is in contradiction with the assumption that the couples $(\mu_k^\star, \Sigma_k^\star)$ are distinct and thus $S \cap S^\star \neq \emptyset$.

Third, assume that $S \cap S^\star \neq \emptyset$ and $S^c \cap S^{\star\,c} \neq \emptyset$, and consider the nonempty subsets $t = S^c \cap S^\star$, $\bar{s} = S \cap S^{\star\,c}$ and $s = S \cap S^\star$. The equality of covariance matrices on $\bar{s}$, on s,

between t and s, between s and $\bar{s}$, and between t and $\bar{s}$ gives respectively for all k,

$$\begin{aligned}
\Sigma_{k,\bar{s}\bar{s}} &= \Omega^\star_{\bar{s}\bar{s}} + \Lambda^{\star\prime}_{s\bar{s}}(\Sigma^\star_{k,ss}\Lambda^\star_{s\bar{s}} + \Sigma^\star_{k,st}\Lambda^\star_{t\bar{s}}) + \Lambda^{\star\prime}_{t\bar{s}}(\Sigma^\star_{k,ts}\Lambda^\star_{s\bar{s}} + \Sigma^\star_{k,tt}\Lambda^\star_{t\bar{s}}) && (3.8)\\
\Sigma_{k,ss} &= \Sigma^\star_{k,ss} && (3.9)\\
\Lambda'_{\bar{s}t}\Sigma_{k,\bar{s}s} + \Lambda'_{st}\Sigma_{k,ss} &= \Sigma^\star_{k,ts} && (3.10)\\
\Sigma_{k,\bar{s}s} &= \Lambda^{\star\prime}_{s\bar{s}}\Sigma^\star_{k,ss} + \Lambda^{\star\prime}_{t\bar{s}}\Sigma^\star_{k,ts} && (3.11)\\
\Sigma_{k,\bar{s}\bar{s}}\Lambda_{\bar{s}t} + \Sigma_{k,\bar{s}s}\Lambda_{st} &= \Lambda^{\star,\prime}_{s\bar{s}}\Sigma^\star_{k,st} + \Lambda^{\star,\prime}_{t\bar{s}}\Sigma^\star_{k,tt}\ . && (3.12)
\end{aligned}$$

From (3.8), (3.11), (3.12), we get

$$\Sigma_{k,\bar{s}\bar{s}}(I - \Lambda_{\bar{s}t}\Lambda^\star_{t\bar{s}}) = \Omega^\star_{\bar{s}\bar{s}} + \Sigma_{k,\bar{s}s}(\Lambda^\star_{s\bar{s}} + \Lambda^\star_{st}\Lambda_{t\bar{s}}) \tag{3.13}$$

and Equations (3.9), (3.10) and (3.11) allow to deduce

$$\Lambda^\star_{s\bar{s}} + \Lambda^\star_{st}\Lambda_{t\bar{s}} = \Sigma^{-1}_{k,ss}\Sigma_{k,s\bar{s}}(I - \Lambda_{\bar{s}t}\Lambda^\star_{t\bar{s}}). \tag{3.14}$$

Finally, Equations (3.13) and (3.14) imply $\Sigma_{k,\bar{s}\bar{s}|s}(I - \Lambda_{\bar{s}t}\Lambda^\star_{t\bar{s}}) = \Omega^\star_{\bar{s}\bar{s}}$. Since $\Omega^\star_{\bar{s}\bar{s}}$ and $\Sigma_{k,\bar{s}\bar{s}|s}$ are positive definite matrices, the matrix $I - \Lambda_{\bar{s}t}\Lambda^\star_{t\bar{s}}$ is nonsingular and all the matrices $\Sigma_{k,\bar{s}\bar{s}|s}$ are equal. Similarly, according to (3.14), all matrices $\Sigma_{k,\bar{s}|s}$ are equal. Then the equality of mean vectors on $\bar{s}$, s and t gives the following equations: For all k,

$$\left\{\begin{array}{l}
\mu_{k\bar{s}} = a^\star_{\bar{s}} + \mu^\star_{ks}\Lambda^\star_{s\bar{s}} + \mu^\star_{k,t}\Lambda^\star_{t\bar{s}}\\
\mu_{ks} = \mu^\star_{ks}\\
a_t + \mu_{k\bar{s}}\Lambda_{\bar{s}t} + \mu_{ks}\Lambda_{st} = \mu^\star_{k,t}
\end{array}\right.$$

implying

$$\mu_{k\bar{s}}(I - \Lambda_{\bar{s}t}\Lambda^\star_{t\bar{s}}) = (a^\star_{\bar{s}} + a_t\Lambda^\star_{t\bar{s}}) + \mu_{ks}(\Lambda^\star_{s\bar{s}} + \Lambda_{st}\Lambda^\star_{t\bar{s}})$$

and Equation (3.14) leads to

$$(\mu_{k\bar{s}} - \mu_{ks}\Sigma^{-1}_{k,ss}\Sigma_{k,s\bar{s}})(I - \Lambda_{\bar{s}t}\Lambda^\star_{t\bar{s}}) = a_{\bar{s}} + a^\star_t\Lambda^\star_{t\bar{s}}\ .$$

Since $I - \Lambda_{\bar{s}t}\Lambda^\star_{t\bar{s}}$ is non singular, the mean vectors $\mu_{k,\bar{s}|s}$ are also equal, and thus the constraint (3.7) is contradicted. In the same way, assuming that $\bar{s}$ or t is empty, we prove that $S \subsetneq S^\star$ and $S^\star \subsetneq S$ are impossible.

Finally, it leads to $S = S^\star$ and, by the equality of covariance matrices and mean vectors, we easily obtain that $\mu_k = \mu^\star_k$, $\Sigma_k = \Sigma^\star_k$ (and then $m = m^\star$), $a = a^\star$, $\Omega = \Omega^\star$ and $\Lambda = \Lambda^\star$. Then, the relation between R, β and Λ allows to deduce that $R = R^\star$ and $\beta = \beta^\star$. □

Lemma 3.C.1.
If the distribution of X_1 is normal with mean μ and variance matrix Σ, and if the conditional distribution of X_2 given $X_1 = x_1$ is normal with mean $a + x_1\Lambda$ and variance matrix Ω then the distribution of (X_1, X_2) is normal with mean $\nu = (\mu, a + \mu\Lambda)$ and variance matrix $\Delta = \begin{pmatrix} \Sigma & \Sigma\Lambda \\ \Lambda'\Sigma & \Omega + \Lambda'\Sigma\Lambda \end{pmatrix}$.

3.D Proof of the criterion consistency theorem

Proof. By definition $(\hat{S},\hat{R}) = \underset{(S,R)\in\mathcal{V}}{\operatorname{argmax}}\ \mathbf{BIC}(S,R)$ with

$$\begin{aligned}\mathbf{BIC}(S,R) &= \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^R)\\ &= 2\ln[f_{\mathrm{clust}}(\mathbf{y}^S|\hat{\alpha})] - \lambda_{(S)}\ln(n) + 2\ln[f_{\mathrm{reg}}(\mathbf{y}^{S^c}|\hat{a}+\mathbf{y}^R\hat{\beta},\hat{\Omega})] - \nu_{(S^c,R)}\ln(n)\\ &= 2\sum_{i=1}^{n}\ln[f(\mathbf{y}_i|\hat{\theta}_{(S,R)})] - \Xi_{(S,R)}\ln(n),\end{aligned}$$

where $\Xi_{(S,R)} = \lambda_{(S)} + \nu_{(S^c,R)}$ is the number of model parameters for variable set (S,R). Thus

$$\begin{aligned}P((\hat{S},\hat{R}) = (S_0,R_0)) &= P(\mathbf{BIC}(S_0,R_0) \geq \mathbf{BIC}(S,R), \forall (S,R)\in\mathcal{V})\\ &= P(\mathbf{BIC}(S_0,R_0) - \mathbf{BIC}(S,R) \geq 0, \forall (S,R)\in\mathcal{V}). \quad (3.15)\end{aligned}$$

Denoting $\gamma_{(S,R)} = \Xi_{(S,R)} - \Xi_{(S_0,R_0)}$ and $\Delta\mathbf{BIC}(S,R) = \mathbf{BIC}(S_0,R_0) - \mathbf{BIC}(S,R)$, we get

$$\Delta\mathbf{BIC}(S,R) = 2n\left[\frac{1}{n}\sum_{i=1}^{n}\ln\left\{\frac{f(\mathbf{y}_i|\hat{\theta}_{(S_0,R_0)})}{h(\mathbf{y}_i)}\right\} - \frac{1}{n}\sum_{i=1}^{n}\ln\left\{\frac{f(\mathbf{y}_i|\hat{\theta}_{(S,R)})}{h(\mathbf{y}_i)}\right\}\right] + \gamma_{(S,R)}\ln(n). \quad (3.16)$$

Under regularity conditions and from (H1), $\Delta\mathbf{BIC}(S,R)$ converges to $-\mathrm{KL}[h, f(.|\theta^\star_{(S,R)})]$ when n tends to infinity. If $\mathrm{KL}[h, f(.|\theta^\star_{(S,R)})] \neq 0$ the first term at the right-hand side of (3.16) dominates and tends to infinity with n. Otherwise, by the unicity assumption in (H1) and the model identifiability, $\mathrm{KL}[h, f(.|\theta^\star_{(S,R)})] = 0$ immediately implies that $S = S_0$ and $R = R_0$. It leads to consider that $\mathcal{V}$ can be decomposed as follows

$$\mathcal{V} = \{(S_0,R_0)\} \cup \mathcal{V}_1$$

where $\mathcal{V}_1 = \{(S,R)\in\mathcal{V};\ \mathrm{KL}[h, f(.|\theta^\star_{(S,R)})] \neq 0\}$.

From (3.15), the theorem is demonstrated if it is proved that

$$\forall (S,R)\in\mathcal{V}_1,\ P\left(\Delta\mathbf{BIC}(S,R) < 0\right) \underset{n\to\infty}{\to} 0.$$

Let $(S,R)\in\mathcal{V}_1$. Denoting $\mathbb{M}_n(S,R) = \frac{1}{n}\sum_{i=1}^{n}\ln\left\{\frac{f(\mathbf{y}_i|\hat{\theta}_{(S,R)})}{h(\mathbf{y}_i)}\right\}$ and $M(S,R) = -\mathrm{KL}[h, f(.|\theta^\star_{(S,R)})]$, from (3.16) we have

$$P(\Delta\mathbf{BIC}(S,R) < 0) = P(2n\{\mathbb{M}_n(S_0,R_0) - \mathbb{M}_n(S,R)\} + \gamma_{(S,R)}\ln(n) < 0) =$$
$$P\left(\mathbb{M}_n(S_0,R_0) - M(S_0,R_0) + M(S_0,R_0) - M(S,R) + M(S,R) - \mathbb{M}_n(S,R) + \frac{\gamma_{(S,R)}\ln(n)}{2n} < 0\right).$$

Thus, for all $\epsilon > 0$, according to Lemma 3.D.5,

$$\begin{aligned}P(\Delta\mathbf{BIC}(S,R) < 0) &\leq P(M(S_0,R_0) - \mathbb{M}_n(S_0,R_0) > \epsilon) + P(\mathbb{M}_n(S,R) - M(S,R) > \epsilon)\\ &\quad + P\left(M(S_0,R_0) - M(S,R) + \frac{\gamma_{(S,R)}\ln(n)}{2n} < 2\epsilon\right).\end{aligned}$$

From Proposition 3.D.1, stated hereafter, $\forall (S,R), \mathbb{M}_n(S,R) \underset{n\to\infty}{\overset{P}{\to}} M(S,R)$. Thus,

$$\forall \epsilon > 0,\ P(\mathbb{M}_n(S,R) - M(S,R) > \epsilon) \leq P(|\mathbb{M}_n(S,R) - M(S,R)| > \epsilon) \underset{n\to\infty}{\to} 0.$$

For the third term, note

$$P\left(M(S_0,R_0) - M(S,R) + \frac{\gamma_{(S,R)}\ln(n)}{2n} < 2\epsilon\right) \leq P\left(M(S_0,R_0) - M(S,R) - 2\epsilon < \left|\frac{\gamma_{(S,R)}\ln(n)}{2n}\right|\right).$$

Since $\gamma_{(S,R)}\ln(n)/2n \underset{n\to\infty}{\to} 0$ and $M(S_0,R_0) - M(S,R) > 0$ because $(S,R) \in \mathcal{V}_1$, taking $\epsilon = \{M(S_0,R_0) - M(S,R)\}/4 > 0$, we get

$$P\left(M(S_0,R_0) - M(S,R) + \frac{\gamma_{(S,R)}\ln(n)}{2n} < 2\epsilon\right) \leq P\left(\frac{M(S_0,R_0) - M(S,R)}{2} < \left|\frac{\gamma_{(S,R)}\ln(n)}{2n}\right|\right) \underset{n\to\infty}{\to} 0.$$

Finally, $P(\Delta\mathbf{BIC}(S,R) < 0) \underset{n\to\infty}{\to} 0$. □

The following proposition implies that $\forall (S,R),\ \mathbb{M}_n(S,R) \underset{n\to\infty}{\overset{P}{\to}} M(S,R)$.

Proposition 3.D.1.
Under assumptions (H1) and (H2), $\forall (S,R) \in \mathcal{V},\ \frac{1}{n}\sum_{i=1}^{n} \ln\left[\frac{h(\mathbf{y}_i)}{f(\mathbf{y}_i|\hat{\theta}_{(S,R)})}\right] \underset{n\to\infty}{\overset{P}{\to}} KL[h, f(.|\theta^\star_{(S,R)})]$.

Proof. For making easier the reading of this proof, the notation Card(S) is replaced with $\sharp S$ and all the vectors are implicitly row vectors. Let $(S,R) \in \mathcal{V}$. By the law of large numbers, if $\mathbb{E}[|\ln(h(X))|] < \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} \ln\left[h(\mathbf{y}_i)\right] \underset{n\to\infty}{\overset{P}{\to}} \mathbb{E}_X[\ln(h(X))]. \tag{3.17}$$

And, if the Proposition 3.D.2 can be applied with the family

$$\mathcal{F}_{(S,R)} := \{\ln[f(.|\theta)];\ \theta \in \Theta'_{(S,R)}\}$$

thus

$$\frac{1}{n}\sum_{i=1}^{n} \ln\left[f(\mathbf{y}_i|\hat{\theta}_{(S,R)})\right] \underset{n\to\infty}{\overset{P}{\to}} \mathbb{E}_X[\ln f(X|\theta^\star_{(S,R)})]. \tag{3.18}$$

Then (3.17) and (3.18) give the result. Thus we have to prove that (H2) allows to verify the hypotheses of the Proposition 3.D.2 and $\mathbb{E}_X[|\ln h(X)|] < \infty$.
Firstly, according to (H2), $\Theta'_{(S,R)}$ is a compact metric space. Moreover, for all $\mathbf{x}$ in $\mathbb{R}^Q$, $\theta_{(S,R)} \in \Theta'_{(S,R)} \mapsto \ln[f(\mathbf{x}|\theta_{(S,R)})]$ is continuous. Let us verify now that there is an envelope function F of $\mathcal{F}_{(S,R)}$ being h-integrable. Recalling that

$$\ln[f(\mathbf{x}|\theta_{(S,R)})] = \ln[f_{\text{clust}}(\mathbf{x}^S|\alpha)] + \ln[f_{\text{reg}}(\mathbf{x}^{S^c}|a + \mathbf{x}^R\beta, \Omega)],$$

these two terms are bounded separately.

Study of the first term:

Due to $\|\mathbf{x}^S - \mu_k\|^2_{\Sigma_k^{-1}} \geq 0$, $|\Sigma_k|^{-\frac{1}{2}} \leq s_{\mathrm{m}}^{-\frac{\sharp S}{2}}$ according to Lemma 3.D.3 and $\sum_{k=1}^{K} p_k = 1$, the upper bound of this first term is given by

$$\begin{aligned}
\ln[f_{\mathrm{clust}}(\mathbf{x}^S|\alpha)] &= \ln\left[\sum_{k=1}^{K} p_k |2\pi\Sigma_k|^{-\frac{1}{2}} \exp\left(-\frac{\|\mathbf{x}^S - \mu_k\|^2_{\Sigma_k^{-1}}}{2}\right)\right] \\
&\leq \ln\left[\sum_{k=1}^{K} p_k (2\pi s_{\mathrm{m}})^{-\frac{\sharp S}{2}}\right] \\
&\leq -\frac{\sharp S}{2}\ln\left[2\pi s_{\mathrm{m}}\right]
\end{aligned}$$

where $\|\mathbf{x}^S - \mu_k\|^2_{\Sigma_k^{-1}} = (\mathbf{x}^S - \mu_k)\Sigma_k^{-1}(\mathbf{x}^S - \mu_k)'$.

For obtaining a lower bound, the concavity of the logarithm function is used thus

$$\begin{aligned}
\ln[f_{\mathrm{clust}}(\mathbf{x}^S|\alpha)] &\geq \sum_{k=1}^{K} p_k \ln\left[|2\pi\Sigma_k|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|\mathbf{x}^S - \mu_k\|^2_{\Sigma_k^{-1}}\right)\right] \\
&= -\frac{\sharp S}{2}\ln[2\pi] - \frac{1}{2}\sum_{k=1}^{K} p_k \left\{\ln\left[|\Sigma_k|\right] + \left[\|\mathbf{x}^S - \mu_k\|^2_{\Sigma_k^{-1}}\right]\right\}
\end{aligned}$$

since $\forall k,\ |\Sigma_k| \leq s_{\mathrm{M}}^{\sharp S}$ according to Lemma 3.D.3 and

$$\begin{aligned}
\|\mathbf{x}^S - \mu_k\|^2_{\Sigma_k^{-1}} &\leq \frac{\|\mathbf{x}^S - \mu_k\|^2}{s_{\mathrm{m}}} \\
&\leq \frac{2(\|\mathbf{x}^S\|^2 + \|\mu_k\|^2)}{s_{\mathrm{m}}} \\
&\leq \frac{2(\|\mathbf{x}^S\|^2 + \eta^2)}{s_{\mathrm{m}}}
\end{aligned}$$

because $\mu_k \in \mathcal{B}(\eta, \sharp S)$. Thus,

$$\begin{aligned}
\ln[f_{\mathrm{clust}}(\mathbf{x}^S|\alpha)] &\geq -\frac{\sharp S}{2}\ln[2\pi] - \frac{1}{2}\sum_{k=1}^{K} p_k \left\{\ln[s_{\mathrm{M}}^{\sharp S}] + \frac{2}{s_{\mathrm{m}}}(\|\mathbf{x}\|^2 + \eta^2)\right\} \\
&= -\frac{\sharp S}{2}\ln[2\pi s_{\mathrm{M}}] - \frac{\|\mathbf{x}\|^2 + \eta^2}{s_{\mathrm{m}}}.
\end{aligned}$$

Finally the first term is bounded by

$$-\frac{\sharp S}{2}\ln[2\pi s_{\mathrm{M}}] - \frac{\|\mathbf{x}\|^2 + \eta^2}{s_{\mathrm{m}}} \leq \ln[f_{\mathrm{clust}}(\mathbf{x}^S|\alpha)] \leq -\frac{\sharp S}{2}\ln\left[2\pi s_{\mathrm{m}}\right]. \tag{3.19}$$

Study of the second term:

The second term is expressed as follows:

$$\begin{aligned}\ln\left[f_{\text{reg}}(\mathbf{x}^{S^c}|a+\mathbf{x}^R\beta,\Omega)\right] &= \ln\left[|2\pi\Omega|^{-1/2}\exp\left(-\frac{1}{2}\|\mathbf{x}^{S^c}-a-\mathbf{x}^R\beta\|^2_{\Omega^{-1}}\right)\right]\\ &= -\frac{\sharp S^c}{2}\ln[2\pi]-\frac{1}{2}\ln[|\Omega|]-\frac{1}{2}\|\mathbf{x}^{S^c}-a-\mathbf{x}^R\beta\|^2_{\Omega^{-1}}.\end{aligned}$$

Using Lemma 3.D.3, the following upper bound is found

$$\ln\left[f_{\text{reg}}(\mathbf{x}^{S^c}|a+\mathbf{x}^R\beta,\Omega)\right]\leq-\frac{\sharp S^c}{2}\ln[2\pi s_{\text{m}}].$$

According to Lemma 3.D.3, $|\Omega|\leq s_{\text{M}}^{\sharp S^c}$ and $\|\mathbf{x}^{S^c}-a-\mathbf{x}^R\beta\|^2_{\Omega^{-1}}\leq s_{\text{m}}^{-1}\|\mathbf{x}^{S^c}-a-\mathbf{x}^R\beta\|^2$. In addition,

$$\begin{aligned}\|\mathbf{x}^{S^c}-a-\mathbf{x}^R\beta\|^2 &\leq 2(\|\mathbf{x}^{S^c}\|^2+\|a+\mathbf{x}^R\beta\|^2)\\ &\leq 2(\|\mathbf{x}^{S^c}\|^2+\|a\|^2+|||\beta|||^2\|\mathbf{x}^R\|^2)\\ &\leq 2(\|\mathbf{x}^{S^c}\|^2+\rho^2[1+\|\mathbf{x}^R\|^2])\end{aligned}$$

because $a\in\mathcal{B}(\rho,1,\sharp S^c)$ and $\beta\in\mathcal{B}(\rho,\sharp R,\sharp S^c)$. Moreover, $\|\mathbf{x}^{S^c}\|^2\leq\|\mathbf{x}\|^2$ and $\|\mathbf{x}^R\|^2\leq\|\mathbf{x}\|^2$ hence

$$\|\mathbf{x}^{S^c}-a-\mathbf{x}^R\beta\|^2\leq 2([1+\rho^2]\|\mathbf{x}\|^2+\rho^2).$$

Then a lower bound of $\ln[f_{\text{reg}}(\mathbf{x}^{S^c}|a+\mathbf{x}^R\beta,\Omega)]$ is

$$\ln\left[f_{\text{reg}}(\mathbf{x}^{S^c}|a+\mathbf{x}^R\beta,\Omega)\right]\geq-\frac{\sharp S^c}{2}\ln[2\pi s_{\text{M}}]-\frac{\rho^2}{s_{\text{m}}}-\frac{1+\rho^2}{s_{\text{m}}}\|\mathbf{x}\|^2.$$

Finally the second term is bounded by

$$-\frac{\sharp S^c}{2}\ln[2\pi s_{\text{M}}]-\frac{\rho^2}{s_{\text{m}}}-\frac{1+\rho^2}{s_{\text{m}}}\|\mathbf{x}\|^2\leq\ln\left[f_{\text{reg}}(\mathbf{x}^{S^c}|a+\mathbf{x}^R\beta,\Omega)\right]\leq-\frac{\sharp S^c}{2}\ln[2\pi s_{\text{m}}]. \quad (3.20)$$

Using (3.19), (3.20) and $\sharp S+\sharp S^c=Q$, each function of the family $\mathcal{F}_{(S,R)}$ is bounded by

$$-\frac{Q}{2}\ln[2\pi s_{\text{M}}]-\frac{\rho^2+\eta^2}{s_{\text{m}}}-\frac{2+\rho^2}{s_{\text{m}}}\|\mathbf{x}\|^2\leq\ln\left[f(\mathbf{x}|\theta_{(S,R)})\right]\leq-\frac{Q}{2}\ln\left[2\pi s_{\text{m}}\right].$$

Thus, for all $\theta_{(S,R)}\in\Theta'_{(S,R)}$ and all $\mathbf{x}\in\mathbb{R}^Q$,

$$|\ln[f(\mathbf{x}|\theta_{(S,R)})]|\leq C_1(s_{\text{m}},s_{\text{M}},Q,\eta,\rho)+C_2(\rho,s_{\text{m}})\|\mathbf{x}\|^2$$

defining the envelope function F, where $C_1(s_{\text{m}},s_{\text{M}},Q,\eta,\rho)$ and $C_2(\rho,s_{\text{m}})$ are two positive constantes.

To verify that F is h-integrable, we have to show that $\int \|\mathbf{x}\|^2 h(\mathbf{x})dx < \infty$.

$$\begin{aligned}
\int \|\mathbf{x}\|^2 h(\mathbf{x})d\mathbf{x} &= \int \|\mathbf{x}\|^2 f(\mathbf{x}|\theta^\star_{(S_0,R_0)})dx \\
&= \int (\|\mathbf{x}^{S_0}\|^2 + \|\mathbf{x}^{S_0^c}\|^2) f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) f_{\text{reg}}(\mathbf{x}^{S_0^c}|a^\star + \mathbf{x}^{R_0}\beta^\star, \Omega^\star) d\mathbf{x}^{S_0^c} d\mathbf{x}^{S_0} \\
&\leq \int \|\mathbf{x}^{S_0}\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&+ \int 2\|\mathbf{x}^{S_0^c} - a^\star - \mathbf{x}^{R_0}\beta^\star\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) f_{\text{reg}}(\mathbf{x}^{S_0^c}|a^\star + \mathbf{x}^{R_0}\beta^\star, \Omega^\star) d\mathbf{x}^{S_0^c} d\mathbf{x}^{S_0} \\
&+ \int 2\|a^\star + \mathbf{x}^{R_0}\beta^\star\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) f_{\text{reg}}(\mathbf{x}^{S_0^c}|a^\star + \mathbf{x}^{R_0}\beta^\star, \Omega^\star) d\mathbf{x}^{S_0^c} d\mathbf{x}^{S_0} \\
&= \int \|\mathbf{x}^{S_0}\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} + 2\int \|a + \mathbf{x}^{R_0}\beta^\star\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&+ \int 2\|\mathbf{x}^{S_0^c} - a^\star - \mathbf{x}^{R_0}\beta^\star\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) f_{\text{reg}}(\mathbf{x}^{S_0^c}|a^\star + \mathbf{x}^{R_0}\beta^\star, \Omega^\star) d\mathbf{x}^{S_0^c} d\mathbf{x}^{S_0} \\
&= A_1 + A_2 + A_3. \qquad (3.21)
\end{aligned}$$

The behaviour of the three integrals A_1, A_2 and A_3 is studied separately. The first integral

$$\begin{aligned}
A_1 &= \int \|\mathbf{x}^{S_0}\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&= \sum_{k=1}^{K} p_k \int \|\mathbf{x}^{S_0}\|^2 \Phi(\mathbf{x}^{S_0}|\mu_k, \Sigma_k) d\mathbf{x}^{S_0} \\
&\leq \sum_{k=1}^{K} p_k [2\|\mu_k\|^2 + 2\,\text{tr}(\Sigma_k)]
\end{aligned}$$

according to Lemma 3.D.4. Thus, from Lemma 3.D.3 and since $\sum_{k=1}^{K} p_k = 1$,

$$A_1 \leq 2\eta^2 + 2s_{\text{M}}\sharp S_0.$$

The second integral is upper bounded by

$$\begin{aligned}
A_2 &= \int \|a^\star + \mathbf{x}^{R_0}\beta^\star\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&\leq \int \rho^2(1 + \|\mathrm{x}^{S_0}\|^2) f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&\leq \rho^2 \int f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} + \rho^2 A_1 \\
&\leq \rho^2 + \rho^2[2\eta^2 + 2s_{\text{M}}\sharp S_0].
\end{aligned}$$

Finally, the third integral can be written

$$\begin{aligned}
A_3 &= \int \|\mathbf{x}^{S_0^c} - a^\star - \mathbf{x}^{R_0}\beta^\star\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) f_{\text{reg}}(\mathbf{x}^{S_0^c}|a^\star + \mathbf{x}^{R_0}\beta^\star, \Omega^\star) d\mathbf{x}^{S_0^c} d\mathbf{x}^{S_0} \\
&= \int f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) \int \|\mathbf{x}^{S_0^c} - a^\star - \mathbf{x}^{R_0}\beta^\star\|^2 |2\pi\Omega^\star|^{-\frac{1}{2}} \exp\left[-\frac{\|\mathbf{x}^{S_0^c} - a^\star - \mathbf{x}^{R_0}\beta^\star\|^2_{\Omega^\star}}{2}\right] d\mathbf{x}^{S_0^c} d\mathbf{x}^{S_0} \\
&\leq \int f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) \int \|\mathbf{x}^{S_0^c} - a^\star - \mathbf{x}^{R_0}\beta^\star\|^2 (2\pi s_{\text{M}})^{-\frac{\sharp S_0^c}{2}} \exp\left[-\frac{\|\mathbf{x}^{S_0^c} - a^\star - \mathbf{x}^{R_0}\beta^\star\|^2}{2s_{\text{M}}}\right] d\mathbf{x}^{S_0^c} d\mathbf{x}^{S_0}
\end{aligned}$$

because $|\Omega^\star|^{-1/2} \leq s_{\text{M}}^{-\sharp S_0^c/2}$ and $\|\mathbf{x}^{S_0^c} - a^\star - \mathbf{x}^{R_0}\beta^\star\|^2_{\Omega^\star} \geq s_{\text{M}}^{-1}\|\mathbf{x}^{S_0^c} - a^\star - \mathbf{x}^{R_0}\beta^\star\|^2$ according to Lemma 3.D.3. Thus, from Lemma 3.D.4,

$$\begin{aligned}
A_3 &\leq \int f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \times \int \|u\|^2 \Phi(u|0, bI_{\sharp S_0^c}) du \\
&= s_{\text{M}}\sharp S_0^c.
\end{aligned}$$

So turning back (3.21), $\int \|\mathbf{x}\|^2 h(\mathbf{x}) d\mathbf{x} \leq 2(s_{\text{M}}\sharp S_0^c + \rho^2) + (1 + 2\rho^2)(2\eta^2 + 2s_{\text{M}}\sharp S_0)$ and finally F is h-integrable. Since $\ln(h) \in \mathcal{F}_{(S_0,\, R_0)}$, it implies that $\mathbb{E}[|\ln h(X)|] \leq \mathbb{E}[F(X)] < \infty$ and the law of large numbers can be applied to obtain (3.17). □

Proposition 3.D.2.
Assume that

1. $(X_1, \ldots, X_n)$ *is a n-sample with unknown density* h.
2. Θ *is a compact metric space.*
3. $\theta \in \Theta \mapsto \ln[f(\boldsymbol{x}|\theta)]$ *is continuous for every* $\boldsymbol{x} \in \mathbb{R}^Q$.
4. F *is an envelope function of* $\mathcal{F} := \{\ln[f(.|\theta)];\ \theta \in \Theta\}$ *which is h-integrable.*
5. $\theta^\star = \underset{\theta\in\Theta}{\text{argmax}}\, KL[h, f(.|\theta)]$
6. $\hat{\theta} = \underset{\theta\in\Theta}{\text{argmax}} \sum_{i=1}^n f(X_i|\theta)$.

Then $\frac{1}{n}\sum_{i=1}^n \ln\left[f(X_i|\hat{\theta})\right] \underset{n\to\infty}{\overset{P}{\to}} \mathbb{E}_X[\ln f(X|\theta^\star)]$.

Proof. We consider the following inequality

$$\begin{aligned}
\left|\mathbb{E}_X[\ln f(X|\theta^\star)] - \frac{1}{n}\sum_{i=1}^n \ln[f(X_i|\hat{\theta})]\right| &\leq \left|\mathbb{E}_X[\ln f(X|\theta^\star)] - \mathbb{E}_X[\ln f(X|\hat{\theta})]\right| \\
&\quad + \sup_{\theta\in\Theta}\left|\mathbb{E}_X[\ln f(X|\theta)] - \frac{1}{n}\sum_{i=1}^n \ln[f(X_i|\theta)]\right|.
\end{aligned}$$

According to the definition of $\theta^\star$, $\mathbb{E}_X[\ln(f(X|\theta^\star))] - \mathbb{E}_X[\ln(f(X|\hat{\theta}_n))] \geq 0$, thus

$$\begin{aligned}\left|\mathbb{E}_X[\ln f(X|\theta^\star)] - \mathbb{E}_X[\ln f(X|\hat{\theta})]\right| &= \mathbb{E}_X[\ln f(X|\theta^\star)] - \frac{1}{n}\sum_{i=1}^{n}\ln[f(X_i|\theta^\star)] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\ln[f(X_i|\theta^\star)] - \frac{1}{n}\sum_{i=1}^{n}\ln[f(X_i|\hat{\theta})] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\ln[f(X_i|\hat{\theta})] - \mathbb{E}_X[\ln f(X|\hat{\theta})] \\ &\leq 2\sup_{\theta\in\Theta}\left|\mathbb{E}_X[\ln f(X|\theta)] - \frac{1}{n}\sum_{i=1}^{n}\ln[f(X_i|\theta)]\right|.\end{aligned}$$

According to Example 19.8 in van der Vaart (1998), the bracketing numbers of $\mathcal{F}$ are finite under the assumptions. Hence, using Theorem 19.4 in van der Vaart (1998), $\mathcal{F}$ is P-Glivenko-Cantelli. Thus $\sup_{\theta\in\Theta}\left|\mathbb{E}_X[\ln f(X|\theta)] - \frac{1}{n}\sum_{i=1}^{n}\ln[f(X_i|\theta)]\right| \xrightarrow[n\to\infty]{P} 0$, which concludes the proof. □

Lemma 3.D.3. *Let $\Sigma \in \mathcal{D}_r$ where $\mathcal{D}_r$ is defined in (H2). Then*

1. $s_m^r \leq |\Sigma| \leq s_M^r$ *and* $\mathrm{tr}(\Sigma) \leq s_M r$
2. $\forall \mathbf{x} \in \mathbb{R}^r, s_M^{-1}\|\mathbf{x}\|^2 \leq \|\mathbf{x}\|_{\Sigma^{-1}}^2 \leq s_m^{-1}\|\mathbf{x}\|^2$

Proof. The proof is based on the eigenvalue decomposition of the variance matrix Σ and the bounded constraint on the eigenvalues because $\Sigma \in \mathcal{D}_r$. □

Lemma 3.D.4.
Let $\Phi(.|\mu, \Sigma)$ be the density of the multivariate Gaussian distribution $\mathcal{N}_r(\mu, \Sigma)$. Then

1. $\int \|\mathbf{x}\|^2\Phi(\mathbf{x}|0, \Sigma)d\mathbf{x} = \mathrm{tr}(\Sigma)$
2. $\int \|\mathbf{x}\|^2\Phi(\mathbf{x}|\mu, \Sigma)d\mathbf{x} \leq 2\left[\|\mu\|^2 + \mathrm{tr}(\Sigma)\right]$

Proof. The first result is a classical property of multivariate Gaussian densities. The second result is deduced from the first one using the triangle inequality. □

Lemma 3.D.5.
Let A and B be two real random variables,

$$\forall \epsilon \in \mathbb{R},\ P(A + B \leq 0) \leq P(A \leq \epsilon) + P(-B > \epsilon).$$

Chapter 4

Improving the variable roles in variable selection for clustering

Résumé: Nous proposons dans ce chapitre une amélioration de la modélisation du rôle des variables dans le processus de classification non supervisée. Ce raffinement de modélisation consiste à répartir les variables déclarées non significatives pour la classification selon deux catégories: ces variables peuvent être dépendantes d'une partie, voire de la totalité, des variables significatives pour la classification ou être totalement indépendantes. Le critère de sélection de modèles et l'algorithme associé sont modifiés selon cette nouvelle partition des variables. Des propriétés d'identifiabilité et de consistance sont également établies. L'intérêt de cette nouvelle procédure de sélection de variables est mise en évidence sur des données simulées.

4.1 Introduction

Among the variable selection procedure available for the clustering with Gaussian mixture models, the procedure of Law et al. (2004) assumes the irrelevant variables to be independent of the relevant clustering variables. Raftery and Dean (2006b) propose a first answer of this limitation by assuming that the irrelevant variables are regressed on the whole relevant variables. Their modelling enforce the dependency link between the two types of variables. In Chapter 3, we suggest an improvement of Raftery and Dean's approach. Models in competition are composed of the number of mixture components, the Gaussian mixture form, the relevant clustering variables S and the subset R of S required to explain the irrelevant variables according to a linear regression. The selected model is the maximizer of the BIC approximation of the integrated likelihood. These models cover in particular Raftery and Dean's modelling since R can be equal to S and also the approach of Law et al. (2004) since R can be an empty subset. Nevertheless, our previous variable selection model did not allow some irrelevant variables to be independent and others to be dependent of the relevant variables as the same time. Moreover, it implies an overpenaliza-

tion of some models, in particular with the more parcimonious Gaussian mixture models. It could have a pernicious effect as illustrated with the following example.

Consider a dataset consisting of 2000 points from a mixture of four equiprobable Gaussian distribution $\mathcal{N}(\mu_k, I_2)$ where $\mu_1 = (0,0)$, $\mu_2 = (4,0)$, $\mu_3 = (0,2)$ and $\mu_4 = (4,2)$ is considered. The third variable is defined by $\mathbf{y}^3 = 0.5\mathbf{y}^1 + \mathbf{y}^2 + \varepsilon$, ε being sampled from a $\mathcal{N}(0, I_{2000})$ density and eleven noisy independent are also appended. For each individual i, $\mathbf{y}_i^{4,5,\ldots,14}$ is simulated according to the Gaussian density $\mathcal{N}((0, 0.4, \ldots, 3.6, 4), I_{11})$. Using our variable selection procedure, the selected model is

$$(\hat{K} = 2, \hat{m} = [pLI], \hat{S} = \{1, 5, 7, 10 - 12\}, \hat{R} = \{1\})$$

and not the true model $(K_0 = 4, m_0 = [pLI], S_0 = \{1, 2\}, R_0 = \{1, 2\})$. In this example, there is a dilemma to choose a clustering with two or four clusters as it can be seen in Figure 4.1. For these two models, the difference between the two associated loglikelihoods is smaller than the difference between the two penalties which differ by the number of free parameters. For the selected model, this number of free parameters is equal to 46 whereas the true model yields to 123 free parameters. For the true model, several regression coefficients are assumed to be free whereas they are equal to zero for the studied dataset, only the third variable is explained by the two first variables. Thus this large complexity of the true model favours the selected model with two clusters. Note that this phenomenon of overpenalization also occurs in the local steps of the variable selection algorithm.

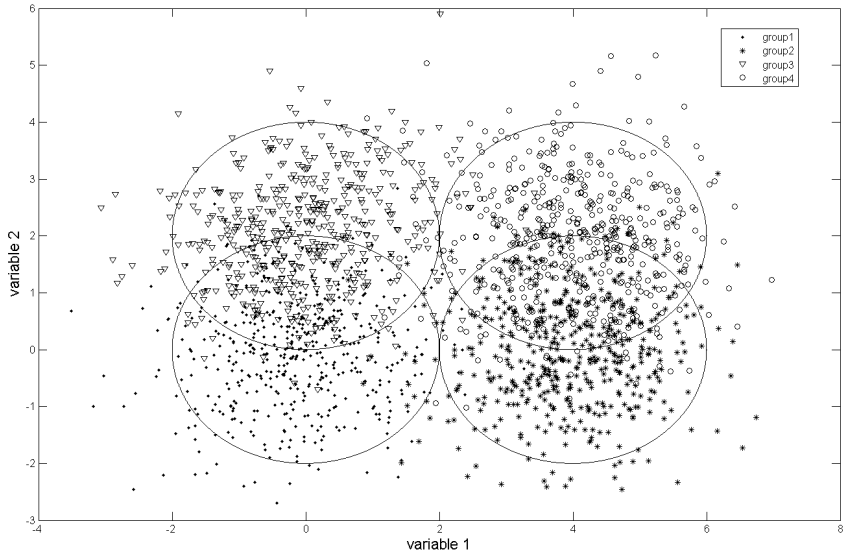


Figure 4.1: Representation of the dataset according to the two first variables.

In order to remedy to such situations, we propose to refine the variable role modelling and to take into account the possibility that some irrelevant clustering variables are independent of all the relevant clustering variables and others are linked to some relevant

variables at the same time. Acting in such a way, we hope to improve the dataset clustering and the variable role analysis. Nevertheless, since the algorithm described in Chapter 3 is already greedy, we have to propose a new algorithm taking the new variable role modelling into account while keeping a reasonable computing time. Note that the improvement proposed in this chapter is only described for single variables, the case of variable blocks is discussed in Section 4.7.

This chapter is organized as follows. A new possible role of the variables is presented in Section 4.2. Section 4.3 is devoted to the presentation of the related model selection criterion. The model identifiability and the consistency of the variable selection criterion are analyzed in Section 4.4. The new associated algorithm is described in Section 4.5 and experimented on simulated datasets in Section 4.6. Finally, a discussion on the overall method is given in Section 4.7.

4.2 A new variable role modelling

A sample of n individuals $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_n')'$ described by Q single variables (a coordinate corresponds to a variable) is considered. In order to improve the variable role for the clustering of these individuals using a model-based clustering method, a new possible variable role is proposed. The nonempty set of relevant clustering variables is again denoted S. Its complement S^c containing the irrelevant clustering variables is now divided into two variable subsets U and W. The variables belonging to U are explained by a variable subset R of S according to a linear regression while the variables of W are assumed to be independent of all the relevant variables. Note that if U is empty, R is empty too and otherwise R is assumed to be no empty. Denoting $\mathcal{F}$ the family of variable index subsets of $\{1, \ldots, Q\}$, the new variable partition set can be formalized as follows:

$$\mathcal{V} = \left\{ (S, R, U, W) \in \mathcal{F}^4; \begin{array}{l} S \cup U \cup W = \{1, \ldots, Q\} \\ S \cap U = \emptyset, S \cap W = \emptyset, U \cap W = \emptyset \\ S \neq \emptyset, R \subseteq S \\ R = \emptyset \text{ if } U = \emptyset \text{ and } R \neq \emptyset \text{ otherwise} \end{array} \right\}.$$

Throughout this chapter, a quadruplet (S, R, U, W) of $\mathcal{V}$ is denoted $\mathbf{V} = (S, R, U, W)$ for simplicity. The new variable partition is summarized in Figure 4.2.

Since the variable partition is changed, a new density family associated to a variable partition $\mathbf{V}$ is now specified to model the unknown density h of the sample $\mathbf{y}$. On the variable subset S, a Gaussian mixture characterized by its number of clusters K and its form m, is still considered. The set of such models (K, m) is denoted $\mathcal{T}$ and the likelihood on S for (K, m) is

$$f_{\text{clust}}(\mathbf{y}^S | K, m, \alpha) = \sum_{k=1}^{K} p_k \Phi(\mathbf{y}^S | \mu_k, \Sigma_k)$$

where the parameter vector is $\alpha = (p_1, \ldots, p_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$, the proportion vector and the variance matrices satisfying the form m.

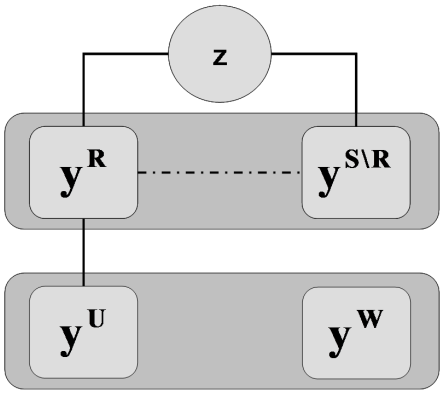


Figure 4.2: Graphical representation of a variable partition $\mathbf{V} = (S, R, U, W)$.

The variables of the subset U are explained by the variables of the subset R according to a multidimensional linear regression where the variance matrix can be assumed to have a spherical, diagonal or general form. These forms are denoted $[LI]$, $[LB]$ and $[LC]$ respectively by analogy with the notation of Gaussian mixture models (see Table 1.1). The variance matrix form is specified by $r \in \mathcal{T}_{\text{reg}} := \{[LI], [LB], [LC]\}$. The likelihood associated to the linear regression of $\mathbf{y}^U$ on $\mathbf{y}^R$ is then

$$f_{\text{reg}}(\mathbf{y}^U | r, a + \mathbf{y}^R\beta, \Omega) = \prod_{i=1}^{n} \Phi(\mathbf{y}_i^U | a + \mathbf{y}_i^R\beta, \Omega)$$

where a is the $1 \times \text{card}(U)$ intercept vector, β is the $\text{card}(R) \times \text{card}(U)$ coefficient regression matrix and Ω is the $\text{card}(U) \times \text{card}(U)$ variance matrix.

Individuals restriction on the variable subset W, which contains the variables independent of all relevant variables, are assumed to be a sample from a Gaussian density with mean vector γ and variance matrix τ. The form of the variance matrix can be spherical or diagonal and is specified by $l \in \mathcal{T}_{\text{indep}} := \{[LI], [LB]\}$. The associated likelihood on W is then

$$f_{\text{indep}}(\mathbf{y}^W | l, \gamma, \tau) = \prod_{i=1}^{n} \Phi(\mathbf{y}_i^W | \gamma, \tau).$$

Finally, the model family is

$$\mathcal{N} = \{(K, m, r, l, \mathbf{V}); (K, m) \in \mathcal{T}, r \in \mathcal{T}_{\text{reg}}, l \in \mathcal{T}_{\text{indep}}, \mathbf{V} \in \mathcal{V}\} \tag{4.1}$$

and the likelihood for a model $(K, m, r, l, \mathbf{V})$ is given by

$$f(\mathbf{y} | K, m, r, l, \mathbf{V}, \theta) = f_{\text{clust}}(\mathbf{y}^S | K, m, \alpha) f_{\text{reg}}(\mathbf{y}^U | r, a + \mathbf{y}^R\beta, \Omega) f_{\text{indep}}(\mathbf{y}^W | l, \gamma, \tau)$$

where the global parameter vector $\theta = (\alpha, a, \beta, \Omega, \gamma, \tau)$ belongs to the parameter vector set $\Upsilon_{(K,m,r,l,\mathbf{V})}$.

4.3 Model selection criterion

The new model collection (4.1) allows to recast the variable selection problem for clustering into a model selection problem. The aim of this section is to adapt the model selection criterion for this new model family. Ideally, we search the model maximizing the integrated loglikelihood

$$(\tilde{K}, \tilde{m}, \tilde{r}, \tilde{l}, \tilde{\mathbf{V}}) = \underset{(K,m,r,l,\mathbf{V})\in\mathcal{N}}{\operatorname{argmax}} \ln\{f(\mathbf{y}|K, m, r, l, \mathbf{V})\}$$

where the integrated likelihood can be decomposed into

$$f(\mathbf{y}|K, m, r, l, \mathbf{V}) = f_{\text{clust}}(\mathbf{y}^S|K, m) f_{\text{reg}}(\mathbf{y}^U|r, \mathbf{y}^R) f_{\text{indep}}(\mathbf{y}^W|l) \tag{4.2}$$

with

$$f_{\text{clust}}(\mathbf{y}^S|K, m) = \int f_{\text{clust}}(\mathbf{y}^S|K, m, \alpha)\pi(\alpha|K, m)d\alpha,$$

$$f_{\text{reg}}(\mathbf{y}^U|r, \mathbf{y}^R) = \int f_{\text{reg}}(\mathbf{y}^U|r, a + \mathbf{y}^R\beta, \Omega)\pi(a, \beta, \Omega|r)d(a, \beta, \Omega)$$

and

$$f_{\text{indep}}(\mathbf{y}^W|l) = \int f_{\text{indep}}(\mathbf{y}^W|l, \gamma, \tau)\pi(\gamma, \tau|l)d(\gamma, \tau).$$

The three functions π are the prior distributions of the different vector parameters. Since these integrated likelihoods are difficult to evaluate, they are approximated by their associated BIC criterion in practice:

- **Bayesian Information Criterion for Gaussian mixture:**
  The BIC criterion associated to the Gaussian mixture on the relevant variable subset S is given by

  $$\text{BIC}_{\text{clust}}(\mathbf{y}^S|K, m) = 2\ln[f_{\text{clust}}(\mathbf{y}^S|K, m, \hat{\alpha})] - \lambda_{(K,m,S)}\ln(n) \tag{4.3}$$

  where $\hat{\alpha}$ is the maximum likelihood estimator obtained using the EM algorithm (Dempster et al., 1977) and $\lambda_{(K,m,S)}$ is the number of free parameters of this Gaussian mixture model (K, m) on the variable subset S.

- **Bayesian Information Criterion for linear regression:**
  For the linear regression of the variable subset U on R, the associated BIC criterion is defined by

  $$\text{BIC}_{\text{reg}}(\mathbf{y}^U|r, \mathbf{y}^R) = 2\ln[f_{\text{reg}}(\mathbf{y}^U|r, \hat{a} + \mathbf{y}^R\hat{\beta}, \hat{\Omega})] - \nu_{(r,U,R)}\ln(n) \tag{4.4}$$

  where $\hat{a}$, $\hat{\beta}$ and $\hat{\Omega}$ are the maximum likelihood estimators. The estimated intercept vector and the regression coefficient matrix are given by $(\hat{a}', \hat{\beta}')' = (X'X)^{-1}X'\mathbf{y}^U$ where $X = (1_n, \mathbf{y}^R)$, 1_n being a n-vector of ones. The estimated variance matrix $\hat{\Omega}$ and the number of free parameters in this linear regression denoted $\nu_{(r,U,R)}$ depend

on the form index r. If r assigns the general form ($r = [LC]$), the estimated variance matrix is given by

$$\hat{\Omega} = \frac{1}{n}\mathbf{y}^{U\prime}\{I_n - X(X'X)^{-1}X'\}\mathbf{y}^U$$

and the number of free parameters is equal to

$$\nu_{(r,U,R)} = \text{card}(U) \times \{\text{card}(R) + 1\} + \frac{\text{card}(U)\{\text{card}(U) + 1\}}{2}.$$

If r is the diagonal form ($r = [LB]$), the estimated variance matrix is written $\hat{\Omega} = \text{diag}(\hat{\omega}_1^2, \ldots, \hat{\omega}_{\text{card}(U)}^2)$ where the diagonal elements are defined by

$$\hat{\omega}_j^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i^U - \hat{a} - \mathbf{y}_i^R\hat{\beta})_j^2, \ \forall j \in \{1, \ldots, \text{card}(U)\}$$

and the number of free parameters is $\nu_{(r,U,R)} = \text{card}(U) \times \{\text{card}(R) + 1\} + \text{card}(U)$. When r assigns the spherical form ($r = [LI]$), the estimated variance matrix is equal to $\hat{\Omega} = \hat{\omega}^2 I_{\text{card}(U)}$ where

$$\hat{\omega}^2 = \frac{1}{n\,\text{card}(U)}\sum_{i=1}^{n}\|\mathbf{y}_i^U - \hat{a} - \mathbf{y}_i^R\hat{\beta}\|_2^2$$

and the number of free parameters is $\nu_{(r,U,R)} = \text{card}(U) \times \{\text{card}(R) + 1\} + 1$.

- **Bayesian Information Criterion for a Gaussian density:**
The BIC criterion associated to the Gaussian density on the variable subset W is given by

$$\text{BIC}_{\text{indep}}(\mathbf{y}^W|l) = 2\ln[f_{\text{indep}}(\mathbf{y}^W|l, \hat{\gamma}, \hat{\tau})] - \rho_{(l,W)}\ln(n). \tag{4.5}$$

The parameters $\hat{\gamma}$ and $\hat{\tau}$ denote the maximum likelihood estimators and $\rho_{(l,W)}$ is the number of free parameters. Whatever the form of the variance matrices, the estimated mean vector is given by

$$\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i^W.$$

If l assigns the diagonal form ($l = [LB]$), the estimated variance matrix is expressed as $\hat{\tau} = \text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_{\text{card}(W)}^2)$ where the diagonal elements are given by

$$\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i^W - \hat{\gamma})_j^2, \ \forall j \in \{1, \ldots, \text{card}(W)\}$$

and the number of free parameters is equal to $\rho_{(l,W)} = 2\,\text{card}(W)$. Otherwise, l indicating the spherical form ($l = [LI]$), the estimated variance matrix is $\hat{\tau} = \hat{\sigma}^2 I_{\text{card}(W)}$ where

$$\hat{\sigma}^2 = \frac{1}{n\,\text{card}(W)}\sum_{i=1}^{n}\|\mathbf{y}_i^W - \hat{\gamma}\|^2$$

and the number of free parameters is equal to $\rho_{(l,W)} = \text{card}(W) + 1$.

Finally, the three terms of the likelihood (4.2) are replaced with their BIC approximation (4.3), (4.4) and (4.5) respectively. Then the selected model satisfies

$$(\hat{K}, \hat{m}, \hat{r}, \hat{l}, \hat{\mathbf{V}}) = \underset{(K,m,r,l,\mathbf{V})\in\mathcal{N}}{\operatorname{argmax}} \operatorname{crit}(K, m, r, l, \mathbf{V}) \tag{4.6}$$

where the model selection criterion is the sum of the three BIC criteria

$$\operatorname{crit}(K, m, r, l, \mathbf{V}) = \operatorname{BIC}_{\text{clust}}(\mathbf{y}^S|K, m) + \operatorname{BIC}_{\text{reg}}(\mathbf{y}^U|r, \mathbf{y}^R) + \operatorname{BIC}_{\text{indep}}(\mathbf{y}^W|l).$$

This criterion can be also written

$$\operatorname{crit}(K, m, r, l, \mathbf{V}) = 2\ln[f(\mathbf{y}|K, m, r, l, \mathbf{V}, \hat{\theta})] - \Xi_{(K,m,r,l,\mathbf{V})}\ln(n) \tag{4.7}$$

where the maximum likelihood estimator is $\hat{\theta} = (\hat{\alpha}, \hat{a}, \hat{\beta}, \hat{\Omega}, \hat{\gamma}, \hat{\tau})$ and the overall number of free parameters is $\Xi_{(K,m,r,l,\mathbf{V})} = \lambda_{(K,m,S)} + \nu_{(r,U,R)} + \rho_{(l,W)}$.

4.4 Theoretical properties

The theoretical properties established in Chapter 3 for the previous modelling can be generalized to the new modelling. First, necessary and sufficient conditions are given to ensure the identifiability of the new variable selection model collections. Second, a consistency theorem of our new variable selection criterion is stated.

4.4.1 Identifiability

The characterization of identifiability is based on the identifiability property of the previous model collection and the difference between the variables in U and W.

Theorem 4.4.1. *Let $\Theta_{(K,m,r,l,\mathbf{V})}$ be a subset of the parameter set $\Upsilon_{(K,m,r,l,\mathbf{V})}$ whose elements $\theta = (\alpha, a, \beta, \Omega, \gamma, \tau)$*

- *contain distinct couples (μ_k, Σ_k) fulfilling*

$$\forall s \subsetneq S, \exists(k, k'), 1 \leq k < k' \leq K; \mu_{k,\bar{s}|s} \neq \mu_{k',\bar{s}|s} \text{ or } \Sigma_{k,\bar{s}|s} \neq \Sigma_{k',\bar{s}|s} \text{ or } \Sigma_{k,\bar{s}\bar{s}|s} \neq \Sigma_{k',\bar{s}\bar{s}|s}, \tag{4.8}$$

- *if $U \neq \emptyset$,*
  - *for all variables j of R, there exists a variable u of U such that the restriction β_{uj} of the regression coefficient matrix β associated to j and u is not the zero matrix.*
  - *for all variables u of U, there exists a variable j of R such that $\beta_{uj} \neq 0$.*
- *parameters Ω and τ exactly respect the forms r and l respectively: they are both diagonal matrices with at least two different eigenvalues if $r = [LB]$ and $l = [LB]$ and Ω has at least a non-zero entry outside the main diagonal if $r = [LC]$.*

Let $(K, m, r, l, \mathbf{V})$ *and* $(K^\star, m^\star, r^\star, l^\star, \mathbf{V}^\star)$ *be two models. If there exist* $\theta \in \Theta_{(K,m,r,l,\mathbf{V})}$ *and* $\theta^\star \in \Theta_{(K^\star,m^\star,r^\star,l^\star,\mathbf{V}^\star)}$ *such that*

$$f(.|K, m, r, l, \mathbf{V}, \theta) = f(.|K^\star, m^\star, r^\star, l^\star, \mathbf{V}^\star, \theta^\star)$$

then $(K, m, r, l, \mathbf{V}) = (K^\star, m^\star, r^\star, l^\star, \mathbf{V}^\star)$ *and* $\theta = \theta^\star$ *(up to a permutation of mixture components).*

Proof. First, we remark that for all row vector x of size Q,

$$\begin{aligned} f_{\text{reg}}(x^U|r, a + x^R\beta, \Omega) f_{\text{indep}}(x^W|l, \gamma, \tau) &= \Phi(x^U|a + x^R\beta, \Omega)\Phi(x^W|\gamma, \tau) \\ &= \Phi(x^{U\cup W}|\tilde{a} + x^R\tilde{\beta}, \tilde{\Omega}) \end{aligned}$$

where $\tilde{a} = (a, \gamma)$, $\tilde{\beta} = (\beta, 0)$ and $\tilde{\Omega}$ is the block diagonal matrix with diagonal elements Ω and τ. This remark allows us to consider parameter vectors $\tilde{\theta} = (\alpha, \tilde{a}, \tilde{\beta}, \tilde{\Omega})$ in the previous model (K, m, S, R) in order to rewrite the densities in the following way

$$f(x|K, m, r, l, \mathbf{V}, \theta) = \tilde{f}_{\text{clust}}(x^S|K, m, \alpha)\tilde{f}_{\text{reg}}(x^{S^c}|\tilde{a} + x^R\tilde{\beta}, \tilde{\Omega}) = \tilde{f}(x|K, m, S, R, \tilde{\theta})$$

where $\tilde{f}_{\text{clust}}$, $\tilde{f}_{\text{reg}}$ and $\tilde{f}$ denote the density functions used in the previous modelling (see Chapter 3). In the same way, $f(.|K^\star, m^\star, r^\star, l^\star, \mathbf{V}^\star, \theta^\star) = \tilde{f}(.|K^\star, m^\star, S^\star, R^\star, \tilde{\theta}^\star)$. According to Hypothesis (4.8) and the identifiability property for the previous modelling (see Theorem 3.5.1), the equality

$$\tilde{f}_{\text{clust}}(x^S|K, m, \alpha)\tilde{f}_{\text{reg}}(x^{S^c}|\tilde{a} + x^R\tilde{\beta}, \tilde{\Omega}) = \tilde{f}_{\text{clust}}(x^{S^\star}|K^\star, m^\star, \alpha^\star)\tilde{f}_{\text{reg}}(x^{S^{\star\,c}}|\tilde{a}^\star + x^{R^\star}\tilde{\beta}^\star, \tilde{\Omega}^\star)$$

implies that $K = K^\star$, $m = m^\star$, $\alpha = \alpha^\star$, $S = S^\star$, $R = R^\star$, $\tilde{a} = \tilde{a}^\star$, $\tilde{\beta} = \tilde{\beta}^\star$ and $\tilde{\Omega} = \tilde{\Omega}^\star$. Then we consider the decompositions $S^c = U \cup W$ and $S^{\star\,c} = U^\star \cup W^\star$ knowing that $S^c = S^{\star\,c}$. If there exists a variable j belonging to $U^\star \cap W$ then for all $q \in R$, $(\beta, 0)_{qj} = 0 = (\beta^\star, 0)_{qj}$ and there exists $q \in R^\star = R$ such that $\beta^\star_{qj} \neq 0$. Thus by contradiction, we obtain that $U^\star \cap W$ is empty and in the same way, $U \cap W^\star$ is an empty set. Finally, it leads to $W = W^\star$, $U = U^\star$ and, identifying each parameter term $\tilde{a}$, $\tilde{\beta}$ and $\tilde{\Omega}$, we obtain that $a = a^\star$, $\beta = \beta^\star$, $\gamma = \gamma^\star$, $\tau = \tau^\star$, $\Omega = \Omega^\star$ and then $r = r^\star$ and $l = l^\star$. □

4.4.2 Consistency of our criterion

A consistency property of our criterion restricted to the variable partition selection can still be checked. In this section, it is proved that the probability of selecting the true variable partition $\mathbf{V}_0 = (S_0, R_0, U_0, W_0)$ by maximizing criterion (4.7) approaches 1 as $n \to \infty$ when the sampling distribution is one of the densities in competition and the true model (K_0, m_0, r_0, l_0) is known. Denoting h the density function of the sample $\mathbf{y}$, the two following vectors are considered

$$\begin{aligned} \theta^\star_{(K,m,r,l,\mathbf{V})} &= \underset{\theta_{(K,m,r,l,\mathbf{V})} \in \Theta_{(K,m,r,l,\mathbf{V})}}{\text{argmin}} \text{KL}[h, f(.|\theta_{(K,m,r,l,\mathbf{V})})] \\ &= \underset{\theta_{(K,m,r,l,\mathbf{V})} \in \Theta_{(K,m,r,l,\mathbf{V})}}{\text{argmax}} \mathbb{E}_X\{\ln f(X|\theta_{(K,m,r,l,\mathbf{V})})\}, \end{aligned}$$

where $\mathrm{KL}[h,f] = \int \ln\left\{\frac{h(x)}{f(x)}\right\} h(x)dx$ is the Kullback-Leibler divergence between the densities h and f and

$$\hat{\theta}_{(K,m,r,l,\mathbf{V})} = \underset{\theta_{(K,m,r,l,\mathbf{V})}\in\Theta_{(K,m,r,l,\mathbf{V})}}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} \ln\{f(\mathbf{y}_i|\theta_{(K,m,r,l,\mathbf{V})})\}.$$

Recall that $\Theta_{(K,m,r,l,\mathbf{V})}$ is the subset defined in Theorem 4.4.1 where the model identifiability is ensured.

The following assumption is considered:

(H1) The density h is assumed to be one of the densities in competition. By identifiability, there exists a unique model $(K_0, m_0, r_0, l_0, \mathbf{V}_0)$ and an associated parameter $\theta^{\star}_{(K_0,\,m_0,\,r_0,\,l_0,\,\mathbf{V}_0)}$ such that $h = f(.|\theta^{\star}_{(K_0,\,m_0,\,r_0,\,l_0,\,\mathbf{V}_0)})$. The model (K_0, m_0, r_0, l_0) is supposed to be known.

To simplify the notation, all the dependencies over this model (K_0, m_0, r_0, l_0) is omitted in the following. Moreover, an additional technical assumption is considered:

(H2) The vectors $\theta^{\star}_{\mathbf{V}}$ and $\hat{\theta}_{\mathbf{V}}$ are supposed to belong to a compact subspace $\Theta'_{\mathbf{V}}$ of the following subset

$$\left(\begin{array}{c} \mathcal{P}_{K-1}\times\mathcal{B}(\eta,\mathrm{card}(S))^{K_0}\times\mathcal{D}^{K_0}_{\mathrm{card}(S)}\times\mathcal{B}(\rho,\mathrm{card}(U)) \\ \times\mathcal{B}(\rho,\mathrm{card}(R),\mathrm{card}(U))\times\mathcal{D}_{\mathrm{card}(U)}\times\mathcal{B}(\eta_1,\mathrm{card}(W))\times\mathcal{D}_{\mathrm{card}(W)} \end{array}\right)\cap\ \Theta_{\mathbf{V}}$$

where

- $\mathcal{P}_{K-1} = \left\{(p_1,\ldots,p_K)\in[0,1]^K; \sum_{k=1}^{K} p_k = 1\right\}$ denotes the $K-1$ dimensional simplex containing the considered proportion vectors,
- $\mathcal{B}(\eta, r)$ is the closed ball in $\mathbb{R}^r$ of radius η centered at zero for the l^2-norm defined by $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{r} x_i^2}, \forall \mathbf{x}\in\mathbb{R}^r$,
- $\mathcal{B}(\rho, r, q)$ is the closed ball in $\mathcal{M}_{r\times q}(\mathbb{R})$ of radius ρ centered at zero for the matricial norm $|||.|||$ defined by

$$\forall A\in\mathcal{M}_{r\times q}(\mathbb{R}), |||A||| = \sup_{\|\mathbf{x}\|=1}\|\mathbf{x}A\|,$$

- $\mathcal{D}_r$ is the set of the $r\times r$ positive definite matrices with eigenvalues in $[s_{\mathrm{m}}, s_{\mathrm{M}}]$ with $0 < s_{\mathrm{m}} < s_{\mathrm{M}}$.

Theorem 4.4.2. *Under assumptions (H1) and (H2), the variable partition* $\hat{\mathbf{V}} = (\hat{S},\hat{R},\hat{U},\hat{W})$ *maximizing Criterion (4.7) with fixed* (K_0, m_0, r_0, l_0) *is such that*

$$P(\hat{\mathbf{V}} = \mathbf{V}_0) = P((\hat{S},\hat{R},\hat{U},\hat{W}) = (S_0,R_0,U_0,W_0)) \underset{n\to\infty}{\to} 1.$$

The proof of this theorem is given in Appendix 4.A.

4.5 The new variable selection procedure

An exhaustive research of the model which maximizes the criterion (4.7) is impossible since the number of models is very large and in particular, larger than with the previous variable selection model. Thus we design a procedure, embedding backward stepwise algorithms to determine the best model, which is now described.

4.5.1 The models in competition

At a fixed step of the algorithm, the variable set $\{1, \ldots, Q\}$ is divided into the set of selected clustering variables S, the set U of irrelevant variables which are linked to some relevant variables, the set W of independent irrelevant variables and j the candidate variable for inclusion into or exclusion from the clustering variable set. Under the model (K, m, r, l), the integrated likelihood can be decomposed into

$$\begin{aligned} f(\mathbf{y}^S, \mathbf{y}^j, \mathbf{y}^U, \mathbf{y}^W | K, m, r, l) &= f(\mathbf{y}^U, \mathbf{y}^W | \mathbf{y}^S, \mathbf{y}^j, K, m, r, l) f(\mathbf{y}^S, \mathbf{y}^j | K, m, r, l) \\ &= f_{\text{indep}}(\mathbf{y}^W | l) f_{\text{reg}}(\mathbf{y}^U | r, \mathbf{y}^S, \mathbf{y}^j) f(\mathbf{y}^S, \mathbf{y}^j | K, m, r, l). \end{aligned}$$

Three situations can then occur for the candidate variable j:

- M1: Given $\mathbf{y}^S$, $\mathbf{y}^j$ provides additional information for the clustering,

$$f(\mathbf{y}^S, \mathbf{y}^j | K, m, r, l) = f_{\text{clust}}(\mathbf{y}^S, \mathbf{y}^j | K, m).$$

- M2: Given $\mathbf{y}^S$, $\mathbf{y}^j$ does not provide additional information for the clustering but has a linear link with the variables of $R[j]$ (the nonempty subset of S containing the relevant variables for the regression of $\mathbf{y}^j$ on $\mathbf{y}^S$),

$$f(\mathbf{y}^S, \mathbf{y}^j | K, m, r, l) = f_{\text{clust}}(\mathbf{y}^S | K, m) f_{\text{reg}}(\mathbf{y}^j | [LI], \mathbf{y}^{R[j]}).$$

- M3: Given $\mathbf{y}^S$, $\mathbf{y}^j$ is independent of all the variables of S,

$$f(\mathbf{y}^S, \mathbf{y}^j | K, m, r, l) = f_{\text{clust}}(\mathbf{y}^S | K, m) f_{\text{indep}}(\mathbf{y}^j | [LI]).$$

In models M2 and M3, the form of the variance matrices in the regression and in the Gaussian density are $r = [LI]$ and $l = [LI]$ respectively since j is a single variable. In order to compare those three situations, we remark that $f_{\text{indep}}(\mathbf{y}^j | [LI])$ can be written $f_{\text{reg}}(\mathbf{y}^j | [LI], \mathbf{y}^{\emptyset})$. Thus, an explicative variable subset, denoted $\tilde{R}[j]$ and defined by $\tilde{R}[j] = \emptyset$ if j follows model M3 and $\tilde{R}[j] = R[j]$ if j follows model M2, is considered. This subset allows us to recast the comparison of the three models into the comparison of two models by the following Bayes factor

$$\frac{f_{\text{clust}}(\mathbf{y}^S, \mathbf{y}^j | K, m)}{f_{\text{clust}}(\mathbf{y}^S | K, m) f_{\text{reg}}(\mathbf{y}^j | [LI], \mathbf{y}^{\tilde{R}[j]})}.$$

This Bayes factor being difficult to evaluate, it is approximated by

$$\mathrm{BIC}_{\mathrm{diff}}(j) = \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S, \mathbf{y}^j | K, m) - \left\{ \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S | K, m) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^j | [LI], \mathbf{y}^{\tilde{R}[j]}) \right\}.$$

That is the criterion $\mathrm{BIC}_{\mathrm{diff}}(.)$ which allows us to construct the variable selection backward algorithm associated to the previous modelling (see Equation 3.4).

4.5.2 The general steps of our algorithm

This algorithm makes use of the clustering variable selection backward algorithm proposed in the previous chapter (see Section 3.4.2) and the regression backward variable selection algorithm presented in Appendix 3.B:

- For each mixture model (K, m):
  - The variable partition into $\hat{S}(K, m)$ and $\hat{S}^c(K, m)$ is determined by the backward stepwise selection algorithm described in Chapter 3.
  - The variable subset $\hat{S}^c(K, m)$ is divided into $\hat{U}(K, m)$ and $\hat{W}(K, m)$: for each variable j belonging to $\hat{S}^c(K, m)$, the variable subset $\tilde{R}[j]$ of $\hat{S}(K, m)$ allowing to explain j by a linear regression is determined with the backward stepwise regression algorithm. If $\tilde{R}[j] = \emptyset$, $j \in \hat{W}(K, m)$ and otherwise, $j \in \hat{U}(K, m)$.
  - For each form r:
    - The variable subset $\hat{R}(K, m, r)$, included into $\hat{S}(K, m)$ and explaining the variables of $\hat{U}(K, m)$, is determined using a backward stepwise regression algorithm taken the fixed form r into account.
    - For each form l: $\hat{\theta}$ is estimated and the following criterion value is computed

$$\widetilde{\mathrm{crit}}(K, m, r, l) := \mathrm{crit}(K, m, r, l, \hat{S}(K, m), \hat{R}(K, m, r), \hat{U}(K, m), \hat{W}(K, m)).$$

- The model satisfying the following condition is then selected

$$(\hat{K}, \hat{m}, \hat{r}, \hat{l}\,) = \underset{(K,m,r,l) \in \mathcal{T} \times \mathcal{T}_{\mathrm{reg}} \times \mathcal{T}_{\mathrm{indep}}}{\mathrm{argmax}} \widetilde{\mathrm{crit}}(K, m, r, l).$$

- Finally, the complete selected model is

$$\left(\hat{K}, \hat{m}, \hat{r}, \hat{l}, \hat{S}(\hat{K}, \hat{m}), \hat{R}(\hat{K}, \hat{m}, \hat{r}), \hat{U}(\hat{K}, \hat{m}), \hat{W}(\hat{K}, \hat{m}) \right).$$

Remark: It is worth noticing that despite there are now four possible roles for the variables rather than three, the complexity of the algorithm is not increased.

4.6 Method validation

This section is devoted to illustrate the behaviour of our new variable selection algorithm and compare it to the previous selection method. First, we study a simulated example where different scenarii for the irrelevant clustering variables are considered. In particular, this example contains the dataset considered in the introduction. Second, the study of the waveform dataset is performed.

4.6.1 Variable selection improvement with this new modelling

The dataset consists of 2000 data points from a mixture of four equiprobable Gaussian distributions $\mathcal{N}(\mu_k, I_2)$ with $\mu_1 = (0,0)$, $\mu_2 = (4,0)$, $\mu_3 = (0,2)$ and $\mu_4 = (4,2)$. The dataset representation given in Figure 4.1 shows the difficulty to choose between 2 or 4 clusters for the dataset clustering. Twelve variables are appended, simulated according to $\mathbf{y}_i^{\{3,\ldots,14\}} = \tilde{a} + \mathbf{y}_i^{\{1,2\}}\tilde{\beta} + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \tilde{\Omega})$ and $\tilde{a} = (0, 0, 0.4, 0.8, \ldots, 3.6, 4)$. Different scenarii are studied, ranging from all variables are independent of the relevant variables to all irrelevant variables depend on relevant variables and with different forms for the variance matrices in the regression and the independent Gaussian density. These different scenarii are described in Table 4.1.

The algorithms associated to the previous and the new variable selection models are compared on these seven scenarii (see Table 4.2 and Table 4.3). The previous variable selection procedure has difficulties with the six first scenarii. It selects a spherical Gaussian mixture with two components. Although Variable 1 is the more significant and seems to be only required to obtain such a clustering in two groups, the procedure selects besides some noise variables (see Table 4.2). It only succeeds in finding the true variable partition for Scenario 7 where irrelevant variables are all dependent to the relevant variables. Moreover it chooses the true number of clusters for this dataset but a more complex Gaussian mixture form $[pLC]$. With the new variable selection procedure, these difficulties of selection disappear. This new method selects the true variable partition and chooses a clustering in four groups (see Table 4.3). The form of variance matrices for the regression and for the independent Gaussian density are correctly identified. Only the true mixture form in Scenario 7 is not selected as with the previous method. This variable selection improvement is due to the use of a larger and more realistic model family and leading to a fairer penalization of the models. For example in Scenario 3, the chosen model is $(\hat{K} = 2, \hat{m} = [pLI], \hat{S} = \{1, 5, 7, 10 - 12\}, \hat{R} = \{1\})$ using the algorithm associated to the previous model whereas our new procedure selects the true model

$$(\hat{K} = 4, \hat{m} = [pLI], \hat{r} = [LI], \hat{l} = [LI], \hat{S} = \{1, 2\}, \hat{R} = \{1, 2\}, \hat{U} = \{3\}, \hat{W} = \{4 - 14\}).$$

With the previous modelling, the free number of parameters for the selected model is 46 whereas it is equal to 123 for the model $(K = 4, m = [pLI], S = \{1, 2\}, R = \{1, 2\})$. In the last case, several regression coefficients are equal to zero since eleven variables are totally independent of the relevant clustering variables but however are counted as free parameters. Thus the difference between the two penalty functions can be larger than

Scenario	$\tilde{\beta}$	$\tilde{\Omega}$
n° 1	0_{12}	I_{12}
n° 2	$((3,0)', 0_{11})$	$\text{diag}(0.5, I_{11})$
n° 3	$((0.5,1)', 0_{11})$	I_{12}
n° 4	$(\beta_1, 0_{10})$	I_{12}
n° 5	$(\beta_1, \beta_2, 0_7)$	$\text{diag}(I_3, 0.5I_5, I_4)$
n° 6	$(\beta_1, \beta_2, \beta_3, 0_3)$	$\text{diag}(I_3, 0.5I_2, \Omega_1, \Omega_2, I_3)$
n° 7	$(\beta_1, \beta_2, \beta_3, (-1,-2)', (0,0.5)', (1,1)')$	$\text{diag}(I_3, 0.5I_2, \Omega_1, \Omega_2, I_3)$

Table 4.1: Description of the seven scenarii where 0_p is the $2 \times p$ zero matrix, $\beta_1 = ((0.5,1)', (2,0)')$, $\beta_2 = ((0,3)', (-1,2)', (2,-4)')$, $\beta_3 = ((0.5,0)', (4,0.5)', (3,0)', (2,1)')$, $\Omega_1 = \text{Rot}(\pi/3)' \,\text{diag}(1,3) * \text{Rot}(\pi/3)$, $\Omega_2 = \text{Rot}(\pi/6)' \,\text{diag}(2,6)\text{Rot}(\pi/6)$, $\text{Rot}(\theta)$ denoting the plane rotation matrix with angle θ.

Scenario	$\hat{K}$	$\hat{m}$	$\hat{S}$	$\hat{R}$
n° 1	2	$[pLI]$	$\{1, 6, 8, 9, 12-14\}$	$\emptyset$
n° 2	2	$[pLI]$	$\{1, 4, 6\}$	$\{1\}$
n° 3	2	$[pLI]$	$\{1, 5, 7, 10-12\}$	$\{1\}$
n° 4	2	$[pLI]$	$\{1, 5-8, 11, 13\}$	$\{1\}$
n° 5	2	$[pLI]$	$\{1, 4\}$	$\{1\}$
n° 6	2	$[pLI]$	$\{1, 13, 14\}$	$\{1\}$
n° 7	4	$[pLC]$	$\{1, 2\}$	$\{1, 2\}$

Table 4.2: Model selection results obtained with our previous variable selection algorithm. The true model is composed of $K_0 = 4$, $m_0 = [pLI]$, $S_0 = \{1, 2\}$ and $R_0 = \emptyset$ for Scenario 1, $R_0 = \{1\}$ for Scenario 2 and $R_0 = \{1, 2\}$ for the other scenarii, with the previous modelling.

Scenario	$\hat{K}$	$\hat{m}$	$\hat{r}$	$\hat{l}$	$\hat{S}$	$\hat{R}$	$\hat{U}$	$\hat{W}$
n° 1	4	$[pLI]$	-	$[LI]$	$\{1, 2\}$	$\emptyset$	$\emptyset$	$\{3-14\}$
n° 2	4	$[pLI]$	$[LI]$	$[LI]$	$\{1, 2\}$	$\{1\}$	$\{3\}$	$\{4-14\}$
n° 3	4	$[pLI]$	$[LI]$	$[LI]$	$\{1, 2\}$	$\{1, 2\}$	$\{3\}$	$\{4-14\}$
n° 4	4	$[pLI]$	$[LI]$	$[LI]$	$\{1, 2\}$	$\{1, 2\}$	$\{3, 4\}$	$\{5-14\}$
n° 5	4	$[pLI]$	$[LB]$	$[LB]$	$\{1, 2\}$	$\{1, 2\}$	$\{3-7\}$	$\{8-14\}$
n° 6	4	$[pLI]$	$[LC]$	$[LI]$	$\{1, 2\}$	$\{1, 2\}$	$\{3-11\}$	$\{12-14\}$
n° 7	4	$[pLC]$	$[LC]$	-	$\{1, 2\}$	$\{1, 2\}$	$\{3-14\}$	$\emptyset$

Table 4.3: Model selection results obtained with our new variable selection algorithm. For all scenarii, the three first elements of the true model are $K_0 = 4$, $m_0 = [pLI]$ and $S_0 = \{1, 2\}$. The selected $\hat{r}$, $\hat{l}$, $\hat{R}$, $\hat{U}$ and $\hat{W}$ correspond to the true model elements for all scenarii.

the difference between the two loglikelihoods, implying an inaccurate selection model. Moreover, this remark for the global criterion value comparison remains valid in all the local criterion versions evaluated in the backward stepwise algorithm. With the new more realistic variable role modelling, this problem is removed, in particular the number of free parameters is now equal to 25 for the true model.

4.6.2 Waveform dataset

The waveform dataset example presented in Section 3.6.3 is now considered. Recall that this dataset is composed of 5000 points based on a random convex combination of two of three waveforms (see Fig 4.3) sampled at integers $\{1, \ldots, 21\}$ with noise added and nineteen noisy standard centered Gaussian variables are appended. Our new algorithm is performed in the same conditions namely the number of components K belongs to $\{3, 4, 5, 6\}$ and twenty mixture forms are used (spherical forms, diagonal forms and the general forms assigned by $[p\_L\_C\_]$). Our algorithm selects the Gaussian mixture model ($\hat{K} = 6, \hat{m} = [p_k LC]$) and a spherical form for the variance matrix in the regression and in the independent Gaussian density ($\hat{r} = [LI]$ and $\hat{l} = [LI]$). It selects also the following variable partition

$$(\hat{S} = \{4-18\}, \hat{R} = \{5-7, 9-12, 14, 15, 17\}, \hat{U} = \{2, 3, 19, 20, 38\}, \hat{W} = \{1, 21-37, 39, 40\}).$$

This new procedure allows to highlight that several variables are independent of the relevant variables. Except Variable 38, the standard centered Gaussian variables are declared independent. Moreover, it reveals that the link between the variables of $\hat{U}$ with the relevant variables is more complex. In fact, recall that the previous algorithm selects the model ($\hat{K} = 6, \hat{m} = [p_k LC], \hat{S} = \{4 - 18\}, \hat{R} = \{7, 11, 15\}$). Only the maximum of each wave $\{7, 11, 15\}$ are selected to explain irrelevant variables because all the noise variables are regressed. With the new modelling, the independent variables being identified, the dependence of the irrelevant variables of $\hat{U}$ requires several relevant variables. It is more realistic since the dataset is based on a random convex combination of two of three waveforms (see Fig 4.3).

4.7 Discussion

A new modelling of the variable partition is proposed to improve the clustering, its interpretation and the determination of the variable role for this clustering. A larger model family is considered to improve the behaviour of our procedure, in particular when the clustering is difficult to determine or when the clustering is supported by spherical or diagonal Gaussian mixtures for which the variable selection is sensitive. Theoretically, the model identifiability and the criterion consistency are extended to this more versatile collection of models. In practice, the algorithmic complexity is preserved: Although models in competition are larger than under the previous modelling, the complexity of the proposed algorithm is of the same order than the one of the previous algorithm.

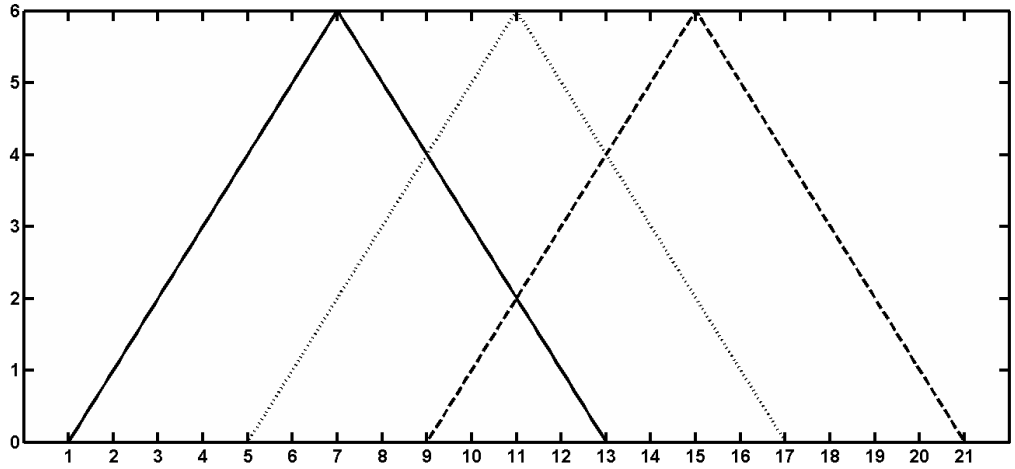


Figure 4.3: Representation of the three wave functions used to construct the waveform dataset.

In this chapter, only the case of single variables is addressed. It is easy to provide a model selection criterion in the same way if variable blocks are considered. But it is more difficult to construct the algorithm in this case. The comparison of the three situations (see Section 4.5.1) which is the main ingredient of the algorithm construction cannot come down to the one of the previous algorithm because of the constraints on variance matrix forms in the regression and the Gaussian independent density. For similar reasons, we do not allow the variance matrix of the Gaussian density to have a general form ($l = [LC]$) in our modelling. The comparison of the three situations becomes more difficult since the variable of interest j can then be correlated to variables of W.

Assume that alternative models are considered, where the linear regression is replaced with an other link and/or an other law is chosen for the independent variables. If BIC criteria associated to these changes are available, an analogous BIC-like criterion can be obtained. Under these conditions, our work can be easily extended theoretically, after taking care to check the model identifiability. On the contrary, the construction of the associated algorithm can require deeper changes.

Appendix

4.A Proof of the criterion consistency theorem

This appendix is devoted to the proof of Theorem 4.4.2 given the criterion consistency. This proof is based on the one of the previous criterion consistency given in Appendix 3.D.

Proof. According to the expressions (4.6) and (4.7), the selected variable partition satisfies $\hat{\mathbf{V}} = \underset{\mathbf{V}\in\mathcal{V}}{\operatorname{argmax}}\ \mathbf{BIC}(\mathbf{V})$ with

$$\mathbf{BIC}(\mathbf{V}) = 2\sum_{i=1}^{n} \ln[f(\mathbf{y}_i|\hat{\theta}_{\mathbf{V}})] - \Xi_{(\mathbf{V})}\ln(n).$$

Thus

$$P(\hat{\mathbf{V}} = \mathbf{V}_0) = P(\mathbf{BIC}(\mathbf{V}_0) - \mathbf{BIC}(\mathbf{V}) \geq 0, \forall\, \mathbf{V} \in \mathcal{V}). \tag{4.9}$$

Denoting $\Delta\mathbf{BIC}(\mathbf{V}) = \mathbf{BIC}(\mathbf{V}_0) - \mathbf{BIC}(\mathbf{V})$, we get

$$\Delta\mathbf{BIC}(\mathbf{V}) = 2n\left[\frac{1}{n}\sum_{i=1}^{n}\ln\left\{\frac{f(\mathbf{y}_i|\hat{\theta}_{\mathbf{V}_0})}{h(\mathbf{y}_i)}\right\} - \frac{1}{n}\sum_{i=1}^{n}\ln\left\{\frac{f(\mathbf{y}_i|\hat{\theta}_{\mathbf{V}})}{h(\mathbf{y}_i)}\right\}\right] + \left[\Xi_{(\mathbf{V})} - \Xi_{(\mathbf{V}_0)}\right]\ln(n). \tag{4.10}$$

Note that for a variable partition $\mathbf{V} \in \mathcal{V}\backslash\{\mathbf{V}_0\}$, $\mathrm{KL}[h, f(.|\theta^\star_{\mathbf{V}})] \neq 0$ since $\theta^\star_{\mathbf{V}} \in \Theta'_{\mathbf{V}} \subset \Theta_{\mathbf{V}}$ and according to the model identifiability. Thus, the variable partition set $\mathcal{V}$ can be decomposed into $\mathcal{V} = \{\mathbf{V}_0\} \cup \mathcal{V}_1$ where $\mathcal{V}_1 = \{\mathbf{V} \in \mathcal{V};\ \mathrm{KL}[h, f(.|\theta^\star_{\mathbf{V}})] \neq 0\}$. From (4.9), the theorem is then established if it is proved that

$$\forall\, \mathbf{V} \in \mathcal{V}_1,\ P\left(\Delta\mathbf{BIC}(\mathbf{V}) < 0\right) \underset{n\to\infty}{\to} 0. \tag{4.11}$$

Let $\mathbf{V} \in \mathcal{V}_1$. Denoting $\mathbb{M}_n(\mathbf{V}) = \frac{1}{n}\sum\limits_{i=1}^{n}\ln\left\{\frac{f(\mathbf{y}_i|\hat{\theta}_{\mathbf{V}})}{h(\mathbf{y}_i)}\right\}$ and $M(\mathbf{V}) = -\mathrm{KL}[h, f(.|\theta^\star_{\mathbf{V}})]$, from (4.10) we have

$$\begin{aligned}
&P(\Delta\mathbf{BIC}(\mathbf{V}) < 0) = P(2n\{\mathbb{M}_n(\mathbf{V}_0) - \mathbb{M}_n(\mathbf{V})\} + \left[\Xi_{(\mathbf{V})} - \Xi_{(\mathbf{V}_0)}\right]\ln(n) < 0)\\
&= P\left(\mathbb{M}_n(\mathbf{V}_0) - M(\mathbf{V}_0) + M(\mathbf{V}_0) - M(\mathbf{V}) + M(\mathbf{V}) - \mathbb{M}_n(\mathbf{V}) + \frac{\left[\Xi_{(\mathbf{V})} - \Xi_{(\mathbf{V}_0)}\right]\ln(n)}{2n} < 0\right).
\end{aligned}$$

Thus, for all $\epsilon > 0$, according to Lemma 3.D.5,

$$\begin{aligned}
P(\Delta\mathbf{BIC}(\mathbf{V}) < 0) \quad &\leq \quad P(M(\mathbf{V}_0) - \mathbb{M}_n(\mathbf{V}_0) > \epsilon) + P(\mathbb{M}_n(\mathbf{V}) - M(\mathbf{V}) > \epsilon)\\
&+ \quad P\left(M(\mathbf{V}_0) - M(\mathbf{V}) + \frac{\left[\Xi_{(\mathbf{V})} - \Xi_{(\mathbf{V}_0)}\right]\ln(n)}{2n} < 2\epsilon\right).
\end{aligned}$$

As in Appendix 3.D, it only requires to show that $\forall\, \mathbf{V} \in \mathcal{V}, \mathbb{M}_n(\mathbf{V}) \underset{n\to\infty}{\overset{P}{\to}} M(\mathbf{V})$ in order to prove (4.11). Thus the proof is finished using the result of the following Lemma 4.A.1. □

Lemma 4.A.1. *Under assumptions (H1) and (H2),*

$$\forall\, \mathbf{V} \in \mathcal{V},\ \frac{1}{n}\sum_{i=1}^{n}\ln\left[\frac{h(\mathbf{y}_i)}{f(\mathbf{y}_i|\hat{\theta}_{\mathbf{V}})}\right] \underset{n\to\infty}{\overset{P}{\to}} KL[h, f(.|\theta^\star_{\mathbf{V}})].$$

Proof. For making easier the reading of this proof, the notation Card(S) is replaced with $\sharp S$ and we recall that all the vectors are implicitly row vectors. Let $\mathbf{V} = (S, R, U, W) \in \mathcal{V}$. As in the proof of Proposition 3.D.1, we want to apply Proposition 3.D.2 with the family

$$\mathcal{F}_{(\mathbf{V})} := \{\ln[f(.|\theta)];\ \theta \in \Theta'_{\mathbf{V}}\}$$

in order to obtain

$$\frac{1}{n}\sum_{i=1}^{n} \ln\left[f(\mathbf{y}_i|\hat{\theta}_{\mathbf{V}})\right] \underset{n\to\infty}{\overset{P}{\to}} \mathbb{E}_X[\ln f(X|\theta^\star_{\mathbf{V}})].$$

Thus we have to prove that (H2) allows to verify the hypotheses of the Proposition 3.D.2 and $\mathbb{E}_X[|\ln h(X)|] < \infty$.

Firstly, according to (H2), $\Theta'_{\mathbf{V}}$ is a compact metric space. Moreover, for all $\mathbf{x}$ in $\mathbb{R}^Q$, $\theta_{\mathbf{V}} \in \Theta'_{\mathbf{V}} \mapsto \ln[f(\mathbf{x}|\theta_{\mathbf{V}})]$ is continuous. Let us verify now that there is an envelope function F of $\mathcal{F}_{(\mathbf{V})}$ being h-integrable. Recalling that

$$\ln[f(\mathbf{x}|\theta_{\mathbf{V}})] = \ln[f(\mathbf{x}|\theta_{(S,R,U,W)})] = \ln[f_{\text{clust}}(\mathbf{x}^S|\alpha)]+\ln[f_{\text{reg}}(\mathbf{x}^U|a+\mathbf{x}^R\beta,\Omega)]+\ln[f_{\text{indep}}(\mathbf{x}^W|\gamma,\tau)],$$

these three terms on the right-hand side are bounded separately. Using the calculus of the proof of Proposition 3.D.1, the two first terms are bounded by

$$-\frac{\sharp S}{2}\ln[2\pi s_{\text{M}}] - \frac{\|\mathbf{x}\|^2+\eta^2}{s_{\text{m}}} \le \ln[f_{\text{clust}}(\mathbf{x}^S|\alpha)] \le -\frac{\sharp S}{2}\ln[2\pi s_{\text{m}}] \tag{4.12}$$

and

$$-\frac{\sharp U}{2}\ln[2\pi s_{\text{M}}] - \frac{\rho^2}{s_{\text{m}}} - \frac{1+\rho^2}{s_{\text{m}}}\|\mathbf{x}\|^2 \le \ln\left[f_{\text{reg}}(\mathbf{x}^U|a+\mathbf{x}^R\beta,\Omega)\right] \le -\frac{\sharp U}{2}\ln[2\pi s_{\text{m}}]. \tag{4.13}$$

For the third term,

$$\begin{aligned}\ln\left[f_{\text{indep}}(\mathbf{x}^W|\gamma,\tau)\right] &= \ln\left[|2\pi\tau|^{-1/2}\exp\left(-\frac{1}{2}\|\mathbf{x}^W-\gamma\|^2_{\tau^{-1}}\right)\right]\\ &= -\frac{\sharp W}{2}\ln[2\pi] - \frac{1}{2}\ln[|\tau|] - \frac{1}{2}\|\mathbf{x}^W-\gamma\|^2_{\tau^{-1}}.\end{aligned}$$

Using Lemma 3.D.3, the third term can be upper bounded by

$$\ln\left[f_{\text{indep}}(\mathbf{x}^W|\gamma,\tau)\right] \le -\frac{\sharp W}{2}\ln[2\pi s_{\text{m}}].$$

According to Lemma 3.D.3, $|\tau| \le s_{\text{M}}^{\sharp W}$ and

$$\begin{aligned}\|\mathbf{x}^W-\gamma\|^2_{\tau^{-1}} &\le s_{\text{m}}^{-1}\|\mathbf{x}^W-\gamma\|^2\\ &\le \frac{2}{s_{\text{m}}}\left(\|\mathbf{x}^W\|^2+\|\gamma\|^2\right)\\ &\le \frac{2}{s_{\text{m}}}\left(\|\mathbf{x}\|^2+\eta^2\right)\end{aligned}$$

because $\gamma \in \mathcal{B}(\eta, \sharp W)$. Then a lower bound of $\ln[f_{\text{indep}}(\mathbf{x}^W|\gamma,\tau)]$ is

$$\ln[f_{\text{indep}}(\mathbf{x}^W|\gamma,\tau)] \geq -\frac{\sharp W}{2}\ln[2\pi s_{\text{M}}] - \frac{2(\|\mathbf{x}\|^2+\eta^2)}{s_{\text{m}}}.$$

Finally the third term is bounded by

$$-\frac{\sharp W}{2}\ln[2\pi s_{\text{M}}] - \frac{2(\|\mathbf{x}\|^2+\eta^2)}{s_{\text{m}}} \leq \ln\left[f_{\text{indep}}(\mathbf{x}^W|\gamma,\tau)\right] \leq -\frac{\sharp W}{2}\ln[2\pi s_{\text{m}}]. \tag{4.14}$$

Using (4.12), (4.13), (4.14) and $\sharp S + \sharp U + \sharp W = Q$, each function of the family $\mathcal{F}_{(\mathbf{V})}$ is bounded by

$$-\frac{Q}{2}\ln[2\pi s_{\text{M}}] - \frac{2(\|\mathbf{x}\|^2+\eta^2)}{s_{\text{m}}} - \frac{\rho^2}{s_{\text{m}}} - \frac{(1+\rho^2)\|\mathbf{x}\|^2}{s_{\text{m}}} \leq \ln\left[f(\mathbf{x}|\theta_{\mathbf{V}})\right] \leq -\frac{Q}{2}\ln\left[2\pi s_{\text{m}}\right].$$

Thus, for all $\theta_{\mathbf{V}} \in \Theta'_{\mathbf{V}}$ and all $\mathbf{x} \in \mathbb{R}^Q$, $|\ln[f(\mathbf{x}|\theta_{\mathbf{V}})]| \leq C_1(s_{\text{m}}, s_{\text{M}}, Q, \eta, \rho) + C_2(\rho, s_{\text{m}})\|\mathbf{x}\|^2$ defining the envelope function F, where $C_1(s_{\text{m}}, s_{\text{M}}, Q, \eta, \rho)$ and $C_2(\rho, s_{\text{m}})$ are two positive constants. To verify that F is h-integrable, we have to show that $\int \|\mathbf{x}\|^2 h(\mathbf{x})dx < \infty$:

$$\begin{aligned}
\int \|\mathbf{x}\|^2 h(\mathbf{x})d\mathbf{x} &= \int \|\mathbf{x}\|^2 f(\mathbf{x}|\theta^\star_{(S_0,R_0,U_0,W_0)})dx \\
&= \int \|\mathbf{x}\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) f_{\text{reg}}(\mathbf{x}^{U_0}|a^\star + \mathbf{x}^{R_0}\beta^\star, \Omega^\star) f_{\text{indep}}(\mathbf{x}^{W_0}|\gamma^\star,\tau^\star) d\mathbf{x}^{W_0} d\mathbf{x}^{U_0} d\mathbf{x}^{S_0} \\
&= \int \|\mathbf{x}^{S_0}\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&+ \int \|\mathbf{x}^{U_0}\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) f_{\text{reg}}(\mathbf{x}^{U_0}|a^\star + \mathbf{x}^{R_0}\beta^\star, \Omega^\star) d\mathbf{x}^{U_0} d\mathbf{x}^{S_0} \\
&+ \int \|\mathbf{x}^{W_0}\|^2 f_{\text{indep}}(\mathbf{x}^{W_0}|\gamma^\star,\tau^\star) d\mathbf{x}^{W_0} \\
&\leq \int \|\mathbf{x}^{S_0}\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&+ \int 2\|a^\star + \mathbf{x}^{R_0}\beta^\star\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&+ \int 2\|\mathbf{x}^{U_0} - a^\star - \mathbf{x}^{R_0}\beta^\star\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) f_{\text{reg}}(\mathbf{x}^{U_0}|a^\star + \mathbf{x}^{R_0}\beta^\star, \Omega^\star) d\mathbf{x}^{U_0} d\mathbf{x}^{S_0} \\
&+ \int \|\mathbf{x}^{W_0}\|^2 f_{\text{indep}}(\mathbf{x}^{W_0}|\gamma^\star,\tau^\star) d\mathbf{x}^{W_0}.
\end{aligned} \tag{4.15}$$

By a similar study as in Appendix 3.D, the three first terms on the right-hand side of Inequality (4.15) are upper bounded respectively by $2\eta^2 + 2s_{\text{M}}\sharp S_0$, $\rho^2 + \rho^2[2\eta^2 + 2s_{\text{M}}\sharp S_0]$ and $s_{\text{M}}\sharp U_0$. For the fourth term

$$\begin{aligned}
\int \|\mathbf{x}^{W_0}\|^2 f_{\text{indep}}(\mathbf{x}^{W_0}|\gamma^\star,\tau^\star) d\mathbf{x}^{W_0} &= \int \|\mathbf{x}^{W_0}\|^2 \Phi(\mathbf{x}^{W_0}|\gamma^\star,\tau^\star) d\mathbf{x}^{W_0} \\
&\leq 2[\|\gamma^\star\|^2 + \text{tr}(\tau^\star)] \\
&\leq 2(\eta^2 + \sharp W_0 s_{\text{M}})
\end{aligned}$$

according to Lemma 3.D.4. So turning back Inequality (4.15), the integral $\int \|\mathbf{x}\|^2 h(\mathbf{x}) d\mathbf{x} \leq 4\eta^2 + 2s_{\mathrm{M}}(\sharp S_0 + \sharp W_0) + s_{\mathrm{M}}\sharp U_0 + \rho^2(1 + 2\eta^2 + 2s_{\mathrm{M}}\sharp S_0)$ and finally F is h-integrable. Since $\ln(h) \in \mathcal{F}_{(S_0, R_0, U_0, W_0)}$, it implies that $\mathbb{E}[|\ln h(X)|] \leq \mathbb{E}[F(X)] < \infty$ and the law of large numbers can be applied to end the proof. □

Chapter 5

Extension of the variable selection procedure for missing at random data

Résumé: La procédure de sélection de variables proposée au chapitre 3 ne permet pas l'étude de jeux de données avec valeurs manquantes. C'est une faiblesse en particulier pour l'étude de données transcriptomes car l'utilisation de la technologie des puces à ADN engendre des données manquantes de natures différentes. Classiquement, ces valeurs manquantes sont préalablement estimées grâce à une méthode d'imputation avant le processus de classification. Nous proposons dans ce chapitre une extension de notre procédure de sélection de variables, tenant compte de l'existence des données manquantes. Cette nouvelle procédure est ensuite comparée à la procédure sur données préalablement complétées grâce à une méthode d'imputation choisie parmi les méthodes proposées couramment pour l'étude de données d'expression.

5.1 Introduction

A variable selection procedure included into the Gaussian mixture clustering process is proposed in Chapter 3. The variable set is decomposed into the subset S of relevant clustering variables and the subset S^c of irrelevant variables. On the relevant variables, the sample density is modelled by a Gaussian mixture with a form m and K clusters. The irrelevant variables are not assumed to be independent of the relevant variables and are explained by a linear regression on a subset R of the relevant variable subset S. A BIC-based criterion is then used to select the best model and in practice, a backward stepwise algorithm is proposed. A weakness of our variable selection procedure is that it cannot be apply to datasets with missing values. The transcriptome dataset studied in Chapter 3 was preliminary restricted to the subset of totally observed genes, removing thus potential interesting genes. As soon as a gene has a missing value, it cannot be studied and then cannot be clustered in order to deduce some functional conclusions. For biologists, a method more sophisticated which does not allow for missing data is useless.

Missing values of transcriptome datasets are often encountered in a clustering context. These missing values are due to various reasons in the laboratory process. First, the spots on the slides are minuscule and are packed very tightly. A tiny imperfection, a smudge, a dust or a scratch on the slide corrupt the signals of some spots. These missing values are also due to hybridization failures. Imputation methods are widely used in a preprocessing step to solve this problem. The imputations by the value zero or the average of row or of column are more widespread but are more and more criticized (see for instance Troyanskaya et al., 2001; Wang et al., 2006). New imputation methods have been proposed to estimate the missing values of gene expression matrices in the ten last years. A review of such methods is presented in Section 5.3.1. Nevertheless, the imputed values are not modified during the clustering process and we think that this preprocessing imputation step may biased the final clustering and also the variable selection.

In this chapter, a sample $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_n')'$ is considered where an individual i is described by a vector $\mathbf{y}_i' \in \mathbb{R}^Q$. This sample is decomposed into $\mathbf{y} = (\mathbf{y}^{\mathbf{o}}, \mathbf{y}^{\mathbf{m}})$ where $\mathbf{y}^{\mathbf{o}}$ is the subset of observed values and $\mathbf{y}^{\mathbf{m}}$ are the missing entries. The general variable block framework is considered: The Q coordinates are grouped into T variable blocks and for a coordinate $j \in \{1, \ldots, Q\}$, $\psi(j) \in \{1, \ldots, T\}$ assigns the corresponding variable block. If A is a subset of variable blocks, $\mathbf{y}_i^{A \cap \mathbf{o}}$ denotes the restriction of $\mathbf{y}_i$ on the coordinates j which are observed for this individual i and such that $\psi(j) \in A$. An analogous notation $\mathbf{y}_i^{A \cap \mathbf{m}}$ is used for the missing data. The aim is to find under which conditions our variable selection procedure can be extended for taking missing values $\mathbf{y}^{\mathbf{m}}$ into account. We are also interested in the influence of the $\mathbf{y}^{\mathbf{m}}$ imputation in preprocessing on our variable selection procedure and in the competitiveness of the suggested extension with respect to imputation method.

The chapter is organized as follows: Section 5.2 is devoted to the description of an extension of our variable selection procedure, taking the missing entries into account. After specifying the nature of missing values in Section 5.2.1, the model selection criterion is adapted in Section 5.2.2. The calculation of this new criterion requires to specify the maximum likelihood parameter estimates, described in Section 5.2.3. The description of the resulting variable selection algorithm and some applications of this algorithm are addressed in Section 5.2.4. Section 5.3 is devoted to the study of imputation methods. A review of imputation methods proposed to complete gene expression datasets is given in Section 5.3.1. Six of these imputation methods are compared on different datasets in Section 5.3.2. Finally a comparison between our method and methods based on imputation strategy is performed in Section 5.3.3.

5.2 Extension of our variable selection procedure

We are interested in the extension of our variable selection procedure for the study of datasets with missing values. The aim is to take the existence of these missing values into account by avoiding the preliminary use of a imputation method. Thus we have to specify under which assumptions we can extend our model selection criterion, provide a parameter

estimation strategy and adapt our variable selection algorithm.

5.2.1 Nature of missing values

It is possible to distinguish three types of missing data according to the missing-data mechanism (Rubin, 1976). If the probability that a value is missing is independent of the observed data, the missing entries are called "missing completely at random" (MCAR). In this case, the completely observed data is a random subsample of the original sample. When the probability that a value is missing is related to the observed data but not to the missing data, the missing values are called "missing at random" (MAR). The MAR assumption is less restrictive than MCAR because it only requires that the missing values behave like a random sample of all values within subclasses defined by observed data. In these two cases, the missing-data mechanism is called ignorable (Little and Rubin, 1986; Schafer, 1997) and likelihood-based inferences can be obtained by ignoring the missing-data mechanism. On the contrary, when the probability depends on the observed data and external phenomena, the missing-data mechanism is not ignorable and missing data are called "not missing at random" (MNAR). In this case, a modelling of the missing-data mechanism is required but is difficult to be suggested since this mechanism is most often unknown.

In this chapter, the missing data are assumed to be MAR. The precise definition of MAR proposed by Rubin (1976) in terms of a probability model is now stated. The missing-data indicator matrix M defined by

$$M_{ij} = \begin{cases} 1 & \text{if} \quad y_i^j \text{ is observed} \\ 0 & \text{if} \quad y_i^j \text{ is missing} \end{cases} \tag{5.1}$$

is considered as a random variable. The density of observed data $\mathbf{y^o}$ ignoring the missing-data mechanism is denoted $f(\mathbf{y^o}|\theta) = \int f(\mathbf{y^o}, \mathbf{y^m}|\theta) d\mathbf{y^m}$ and the joint distribution of M and $\mathbf{y}$ is $f(\mathbf{y}, M|\theta, \psi)$. This last density can be specified as the product of the densities of the distribution of $\mathbf{y}$ and the distribution of M conditionally to $\mathbf{y}$

$$f(\mathbf{y}, M|\theta, \psi) = f(\mathbf{y}|\theta) f(M|\mathbf{y}, \psi).$$

If the distribution of the missing-data mechanism is independent of missing values $\mathbf{y^m}$ then

$$f(M|\mathbf{y}, \psi) = f(M|\mathbf{y^o}, \psi). \tag{5.2}$$

According to Rubin (1976), the missing data are called missing at random when Equation (5.2) is satisfied.

5.2.2 Model selection criterion

In order to extend our variable selection procedure for the clustering of dataset with missing at random values, the model selection criterion given by (3.2) has to be adapted.

5.2.2.1 Theoretical principle

The model family consists of $\mathcal{N} = \{(K, m, S, R);\ (K, m) \in \mathcal{T},\ (S, R) \in \mathcal{V}\}$. The mixture family $\mathcal{T}$ is composed of couples (K, m) where K is the number of mixture components and m is the form of the Gaussian mixture. The set of variable partitions $\mathcal{V}$ contains couples (S, R) where S is the nonempty set of relevant clustering variables and R is a subset of S containing the relevant variables required to explain irrelevant variables according to a linear regression. For the model (K, m, S, R), the data distribution is modelled by

$$f(\mathbf{y}|K, m, S, R, \theta) = f_{\text{clust}}(\mathbf{y}^S|K, m, \alpha) f_{\text{reg}}(\mathbf{y}^{S^c}|a + \mathbf{y}^R\beta, \Omega)$$

where $f_{\text{clust}}(\mathbf{y}^S|K, m, \alpha)$ corresponds to the Gaussian mixture density on variables S with the parameter vector α and $f_{\text{reg}}(\mathbf{y}^{S^c}|a + \mathbf{y}^R\beta, \Omega)$ corresponds to the multidimensional multivariate linear regression density of $\mathbf{y}^{S^c}$ on $\mathbf{y}^R$ with the intercept vector a, the regression coefficient matrix β and the covariance matrix Ω. The parameter vector is denoted $\theta = (\alpha, a, \beta, \Omega)$.

When a dataset with missing values is considered, the selected model $(\tilde{K}, \tilde{m}, \tilde{S}, \tilde{R})$ maximizes the posterior probability

$$P(K, m, S, R|\mathbf{y}^o, M) = \frac{f(\mathbf{y}^o, M|K, m, S, R) P(K, m, S, R)}{f(\mathbf{y}^o, M)}.$$

And since a non informative uniform prior distribution $P(K, m, S, R)$ is assumed on the models, the selected model fulfills

$$(\tilde{K}, \tilde{m}, \tilde{S}, \tilde{R}) = \underset{(K,m,S,R)\in\mathcal{N}}{\operatorname{argmax}} f(\mathbf{y}^o, M|K, m, S, R).$$

This integrated observed likelihood $f(\mathbf{y}^o, M|K, m, S, R)$ is defined by

$$\begin{aligned} f(\mathbf{y}^o, M|K, m, S, R) &= \int f(\mathbf{y}^o, \mathbf{y}^m, M|K, m, S, R) d\mathbf{y}^m \\ &= \int\int f(\mathbf{y}^o, \mathbf{y}^m, M|K, m, S, R, \theta, \psi) \pi(\theta, \psi|K, m, S, R) d(\theta, \psi) d\mathbf{y}^m \end{aligned}$$

where $\pi(\theta, \psi|K, m, S, R)$ is the prior distribution of the complete parameter vector (θ, ψ). Using the MAR assumption given by Equation (5.2) and assuming that the parameters θ and ψ are distinct, in the sense that the joint parameter space of (θ, ψ) is the product of the parameter spaces of θ and ψ, it leads to

$$\begin{aligned} f(\mathbf{y}^o, M|K, m, S, R) &= \int\int f(M|\mathbf{y}^o, \psi) f(\mathbf{y}^o, \mathbf{y}^m|K, m, S, R, \theta) \pi(\theta|K, m, S, R) \pi(\psi) d\theta d\psi d\mathbf{y}^m \\ &= \int f(M|\mathbf{y}^o, \psi) \pi(\psi) d\psi \\ &\quad \times \int \left\{ \int f(\mathbf{y}^o, \mathbf{y}^m|K, m, S, R, \theta) d\mathbf{y}^m \right\} \pi(\theta|K, m, S, R) d\theta \\ &= f(\mathbf{y}^o|K, m, S, R) \int f(M|\mathbf{y}^o, \psi) \pi(\psi) d\psi \end{aligned}$$

where $f(\mathbf{y}^{\mathbf{o}}|K,m,S,R)$ is the integrated observed likelihood ignoring the missing-data mechanism. Consequently, the selected model fulfills

$$(\tilde{K},\tilde{m},\tilde{S},\tilde{R}) = \underset{(K,m,S,R)\in\mathcal{N}}{\operatorname{argmax}} f(\mathbf{y}^{\mathbf{o}}|K,m,S,R).$$

But since this integrated observed likelihood ignoring the missing-data mechanism is difficult to calculate, a BIC approximation is used. Thus the chosen model is

$$(\hat{K},\hat{m},\hat{S},\hat{R}) = \underset{(K,m,S,R)\in\mathcal{N}}{\operatorname{argmax}} \operatorname{crit}(K,m,S,R)$$

with

$$\operatorname{crit}(K,m,S,R) = 2\ln\left\{f(\mathbf{y}^{\mathbf{o}}|K,m,S,R,\hat{\theta}\,)\right\} - \Xi_{(K,m,S,R)}\ln(n), \tag{5.3}$$

where $\hat{\theta}$ is the parameter vector maximizing the observed likelihood $f(\mathbf{y}^{\mathbf{o}}|K,m,S,R,\theta)$. The total number of free parameters $\Xi_{(K,m,S,R)}$ is the sum of the free parameter number $\lambda_{(K,m,S)}$ for the Gaussian mixture on the variable subset S and the free parameter number $\nu_{(S^c,R)}$ in the regression of the subset S^c on R (see Section 3.3).

5.2.2.2 Explicit observed likelihood expression

In order to be able to use the model selection criterion defined by (5.3), the parameter vector $\hat{\theta}$ maximizing the observed likelihood $f(\mathbf{y}^{\mathbf{o}}|K,m,S,R,\theta)$ has to be evaluated and the observed likelihood has to be made explicit. This section is devoted to the second point, the estimation of $\hat{\theta}$ being addressed in the next Section 5.2.3.

The observed likelihood ignoring the missing-data mechanism is defined by

$$\begin{aligned} f(\mathbf{y}^{\mathbf{o}}|K,m,S,R,\theta) &= \int f(\mathbf{y}^{\mathbf{o}},\mathbf{y}^{\mathbf{m}}|K,m,S,R,\theta)d\mathbf{y}^{\mathbf{m}} \\ &= \int f_{\text{clust}}(\mathbf{y}^{S\cap\mathbf{o}},\mathbf{y}^{S\cap\mathbf{m}}|K,m,\alpha) f_{\text{reg}}(\mathbf{y}^{S^c\cap\mathbf{o}},\mathbf{y}^{S^c\cap\mathbf{m}}|a+(\mathbf{y}^{R\cap\mathbf{o}},\mathbf{y}^{R\cap\mathbf{m}})\beta,\Omega)d\mathbf{y}^{\mathbf{m}}. \end{aligned}$$

The first idea to explicitly calculate this observed likelihood would consist of studying separately the two terms in the integral, setting apart the conditional law of missing data according to observed data and integrating on the missing values. For the first term related to the Gaussian mixture, it is possible to isolate the law of $\mathbf{y}^{S\cap\mathbf{m}}$ conditionally to $\mathbf{y}^{S\cap\mathbf{o}}$ according to Proposition 5.C.1. On the contrary, for the second term associated to the regression, such a strategy is made impossible because of the missing values $\mathbf{y}_i^{R\cap\mathbf{m}}$ in the mean vector of the Gaussian density.

Here, an other strategy is considered which consists of considering the distribution of

the sample as a global Gaussian mixture. Thus the likelihood can be expressed as

$$\begin{aligned} f(\mathbf{y}^{\mathbf{o}}, \mathbf{y}^{\mathbf{m}}|K, m, S, R, \theta) &= f_{\text{clust}}(\mathbf{y}^{S\cap\mathbf{o}}, \mathbf{y}^{S\cap\mathbf{m}}|K, m, \alpha) f_{\text{reg}}(\mathbf{y}^{S^c\cap\mathbf{o}}, \mathbf{y}^{S^c\cap\mathbf{m}}|a + (\mathbf{y}^{R\cap\mathbf{o}}, \mathbf{y}^{R\cap\mathbf{m}})\beta, \Omega) \\ &= \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} p_k \Phi(\mathbf{y}_i^{S\cap\mathbf{o}}, \mathbf{y}_i^{S\cap\mathbf{m}}|\mu_k, \Sigma_k) \right\} \Phi(\mathbf{y}_i^{S^c\cap\mathbf{o}}, \mathbf{y}_i^{S^c\cap\mathbf{m}}|a + (\mathbf{y}_i^{R\cap\mathbf{o}}, \mathbf{y}_i^{R\cap\mathbf{m}})\beta, \Omega) \\ &= \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} p_k \Phi(\mathbf{y}_i^{S\cap\mathbf{o}}, \mathbf{y}_i^{S\cap\mathbf{m}}|\mu_k, \Sigma_k) \right\} \Phi(\mathbf{y}_i^{S^c\cap\mathbf{o}}, \mathbf{y}_i^{S^c\cap\mathbf{m}}|a + (\mathbf{y}_i^{S\cap\mathbf{o}}, \mathbf{y}_i^{S\cap\mathbf{m}})\tilde{\beta}, \Omega) \end{aligned}$$

where $\tilde{\beta}$ is deduced from β and (S, R): For all couples $(j, l) \in \{1, \ldots, Q\}^2$ such that $\psi(j) \in S$ and $\psi(l) \in S^c$,

$$\tilde{\beta}_{jl} = \begin{cases} \beta_{jl} & \text{if } \psi(j) \in R \\ 0 & \text{if } \psi(j) \in S \backslash R. \end{cases} \tag{5.4}$$

Then, according to Lemma 3.C.1, the likelihood can be written as the global Gaussian mixture

$$f(\mathbf{y}^{\mathbf{o}}, \mathbf{y}^{\mathbf{m}}|K, m, S, R, \theta) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} p_k \Phi(\mathbf{y}_i^{\mathbf{o}}, \mathbf{y}_i^{\mathbf{m}}|\nu_k, \Delta_k) \right\}$$

where the parameters are defined, for all variables $j \in \{1, \ldots, Q\}$, by

$$\nu_{kj} = \begin{cases} \mu_{kj} & \text{if } \psi(j) \in S \\ (a + \mu_k \tilde{\beta})_j & \text{if } \psi(j) \in S^c \end{cases} \tag{5.5}$$

and, for all variables l and j, by

$$\Delta_{k,jl} = \begin{cases} \Sigma_{k,jl} & \text{if } \psi(j) \in S, \psi(l) \in S \\ (\Sigma_k \tilde{\beta})_{jl} & \text{if } \psi(j) \in S, \psi(l) \in S^c \\ (\tilde{\beta}' \Sigma_k)_{jl} & \text{if } \psi(j) \in S^c, \psi(l) \in S \\ (\Omega + \tilde{\beta}' \Sigma_k \tilde{\beta})_{jl} & \text{if } \psi(j) \in S^c, \psi(l) \in S^c. \end{cases} \tag{5.6}$$

In order to set apart the conditional law of the missing values according to the observed values, the decompositions of the mean vectors and the variance matrices into

$$\nu_k = (\nu_{k,\mathbf{o}}^{(i)}, \nu_{k,\mathbf{m}}^{(i)}) \text{ and } \Delta_k = \begin{pmatrix} \Delta_{k,\mathbf{oo}}^{(i)} & \Delta_{k,\mathbf{om}}^{(i)} \\ \Delta_{k,\mathbf{mo}}^{(i)} & \Delta_{k,\mathbf{mm}}^{(i)} \end{pmatrix}$$

according to the position of missing values for $\mathbf{y}_i$ are considered. Moreover the conditional parameters are denoted by $\nu_{k,\mathbf{m}|\mathbf{o}}^{(i)} = \nu_{k,\mathbf{m}}^{(i)} - \nu_{k,\mathbf{o}}^{(i)} \Delta_{k,\mathbf{m}|\mathbf{o}}^{(i)}$, $\Delta_{k,\mathbf{m}|\mathbf{o}}^{(i)} = (\Delta_{k,\mathbf{oo}}^{(i)})^{-1} \Delta_{k,\mathbf{om}}^{(i)}$ and $\Delta_{k,\mathbf{mm}|\mathbf{o}}^{(i)} = \Delta_{k,\mathbf{mm}}^{(i)} - \Delta_{k,\mathbf{mo}}^{(i)} (\Delta_{k,\mathbf{oo}}^{(i)})^{-1} \Delta_{k,\mathbf{om}}^{(i)}$. According to Proposition 5.C.1, the Gaussian mixture density can be decomposed as follows

$$\sum_{k=1}^{K} p_k \Phi(\mathbf{y}_i^{\mathbf{o}}, \mathbf{y}_i^{\mathbf{m}}|\nu_k, \Delta_k) = \sum_{k=1}^{K} p_k \Phi\left(\mathbf{y}_i^{\mathbf{o}}|\nu_{k,\mathbf{o}}^{(i)}, \Delta_{k,\mathbf{oo}}^{(i)}\right) \Phi\left(\mathbf{y}_i^{\mathbf{m}}|\nu_{k,\mathbf{m}|\mathbf{o}}^{(i)} + \mathbf{y}_i^{\mathbf{o}} \Delta_{k,\mathbf{m}|\mathbf{o}}^{(i)}, \Delta_{k,\mathbf{mm}|\mathbf{o}}^{(i)}\right).$$

Considering the label vector $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ where $\mathbf{z}_i \in \{0,1\}^K$ such that z_{ik} is equal to 1 if the individual i belongs to the k^{th} component, the observed likelihood can be written

$$\begin{aligned} f(\mathbf{y}^{\mathbf{o}}|K,m,S,R,\theta) &= \int f(\mathbf{y}^{\mathbf{o}},\mathbf{y}^{\mathbf{m}},\mathbf{z}|K,m,S,R,\theta)d\mathbf{y}^{\mathbf{m}}d\mathbf{z} \\ &= \prod_{i=1}^{n}\int\prod_{k=1}^{K}\left[p_k\Phi\left(\mathbf{y}_i^{\mathbf{o}}|\nu_{k,\mathbf{o}}^{(i)},\Delta_{k,\mathbf{oo}}^{(i)}\right)\Phi\left(\mathbf{y}_i^{\mathbf{m}}|\nu_{k,\mathbf{m}|\mathbf{o}}^{(i)}+\mathbf{y}_i^{\mathbf{o}}\Delta_{k,\mathbf{m}|\mathbf{o}}^{(i)},\Delta_{k,\mathbf{mm}|\mathbf{o}}^{(i)}\right)\right]^{z_{ik}}d\mathbf{y}_i^{\mathbf{m}}d\mathbf{z}_i \\ &= \prod_{i=1}^{n}\int\prod_{k=1}^{K}p_k^{z_{ik}}\Phi\left(\mathbf{y}_i^{\mathbf{o}}|\nu_{k,\mathbf{o}}^{(i)},\Delta_{k,\mathbf{oo}}^{(i)}\right)^{z_{ik}}d\mathbf{z}_i \\ &= \prod_{i=1}^{n}\sum_{k=1}^{K}p_k\Phi\left(\mathbf{y}_i^{\mathbf{o}}|\nu_{k,\mathbf{o}}^{(i)},\Delta_{k,\mathbf{oo}}^{(i)}\right). \end{aligned} \tag{5.7}$$

Consequently, it is possible to calculate explicitly the observed likelihood using the global Gaussian mixture with parameters $(p_k,\nu_k,\Delta_k)_{1\leq k\leq K}$ and using the expression (5.7). It remains to estimate the parameter vector $\hat{\theta}$ maximizing this observed likelihood in order to use the criterion (5.3).

5.2.3 Maximum observed likelihood estimator

This section is devoted to the determination of the parameter vector $\hat{\theta}$ maximizing the observed likelihood $f(\mathbf{y}^{\mathbf{o}}|K,m,S,R,\theta)$. Since the sample density can be formulated as a global Gaussian mixture, an EM algorithm could be used in order to estimate the parameters $(\hat{p}_k,\hat{\nu}_k,\hat{\Delta}_k)_{1\leq k\leq K}$ and to deduce $\hat{\theta}=(\hat{\alpha},\hat{a},\hat{\beta},\hat{\Omega})$. But these parameters are defined under the constraints (5.4), (5.5) and (5.6) allowing not to make explicit the potential EM algorithm. Thus the parameter vector α of the Gaussian mixture on S and the parameter vector (a,β,Ω) of the regression of S^c on R are separately estimated. These distinct parameter estimations also allow us to adapt our algorithm using a backward stepwise algorithm for the variable selection in regression.

5.2.3.1 Estimation of Gaussian mixture parameters

In this section, we are interested in the computation of the Gaussian mixture parameter vector $\hat{\alpha}$ maximizing $f_{\text{clust}}(\mathbf{y}^{S\cap\mathbf{o}}|K,m,\alpha)$. This estimation is only made explicit for the mixture form $m=[p_kLC]$, which is the only mixture form programmed and applied in this chapter but it can be made explicit for other mixture forms.

We begin to restrict the sample to individuals having at least one observed value for the variables of S, $\mathbf{I}_{S\cap\mathbf{o}} = \{i\in\{1,\ldots,n\};\ \mathbf{y}_i^{S\cap\mathbf{m}}\neq\mathbf{y}_i^{S}\}$, since individuals being not observed on S are non informative to estimate α. An EM algorithm is used to determine the parameter vector $\hat{\alpha}=(\hat{p}_1,\ldots,\hat{p}_K,\hat{\mu}_1,\ldots,\hat{\mu}_K,\hat{\Sigma})$ maximizing the observed likelihood $f_{\text{clust}}(\mathbf{y}_{\mathbf{I}_{S\cap\mathbf{o}}}^{S\cap\mathbf{o}}|K,m,\alpha)$ (see for instance Little and Rubin, 1986). The missing data being composed of the label vector $\mathbf{z}$ and the missing values $\mathbf{y}_{\mathbf{I}_{S\cap\mathbf{o}}}^{S\cap\mathbf{m}}$, the aim is to maximize the

following expectation of the complete loglikelihood conditionally to the observed data and a current parameter value $\alpha^{(r)}$ at the r^{th} iteration:

$$\mathcal{Q}(\alpha|\alpha^{(r)}) = \mathbb{E}\left[\ln f_{\text{clust}}(\mathbf{y}^{S\cap\mathbf{o}}_{\mathbf{I}_{S\cap\mathbf{o}}}, \mathbf{y}^{S\cap\mathbf{m}}_{\mathbf{I}_{S\cap\mathbf{o}}}, \mathbf{z}|K, m, \alpha)|\mathbf{y}^{S\cap\mathbf{o}}_{\mathbf{I}_{S\cap\mathbf{o}}}, \alpha^{(r)}\right]$$

where the complete likelihood is

$$f_{\text{clust}}(\mathbf{y}^{S\cap\mathbf{o}}_{\mathbf{I}_{S\cap\mathbf{o}}}, \mathbf{y}^{S\cap\mathbf{m}}_{\mathbf{I}_{S\cap\mathbf{o}}}, \mathbf{z}|K, m, \alpha) = \prod_{i\in\mathbf{I}_{S\cap\mathbf{o}}}\prod_{k=1}^{K} p_k^{z_{ik}}\Phi(\mathbf{y}_i^{S\cap\mathbf{o}}, \mathbf{y}_i^{S\cap\mathbf{m}}|\mu_k, \Sigma)^{z_{ik}}.$$

The initial parameter value $\alpha^{(1)}$ is obtained by applying the EM algorithm for a Gaussian mixture model $(K, [p_k LC])$ on the totally observed individuals $\{i \in \{1, \ldots, n\};\ \mathbf{y}_i^{S\cap\mathbf{o}} = \mathbf{y}_i^S\}$. Then the algorithm alternates the following E-step and M-step, detailed in Appendix 5.A, until convergence.

E-Step: This step consists of calculating $\mathcal{Q}(\alpha|\alpha^{(r)})$ that is equivalent to making explicit the conditional probabilities that the individuals arise from one of the mixture components, the conditional expectations of y_i^j and the conditional covariances knowing the observed data and a current parameter value $\alpha^{(r)}$. First, the mean vectors and the variance matrix are decomposed into

$$\mu_k = (\mu^{(i)}_{k,\mathbf{o}}, \mu^{(i)}_{k,\mathbf{m}}) \text{ and } \Sigma = \begin{pmatrix} \Sigma^{(i)}_{\mathbf{oo}} & \Sigma^{(i)}_{\mathbf{om}} \\ \Sigma^{(i)}_{\mathbf{mo}} & \Sigma^{(i)}_{\mathbf{mm}} \end{pmatrix},$$

according to the position of the missing values of $\mathbf{y}_i$, and the conditional parameters defined by $\mu^{(i)}_{k,\mathbf{m}|\mathbf{o}} = \mu^{(i)}_{k,\mathbf{m}} - \mu^{(i)}_{k,\mathbf{o}}\Sigma^{(i)}_{\mathbf{m}|\mathbf{o}}$, $\Sigma^{(i)}_{\mathbf{m}|\mathbf{o}} = (\Sigma^{(i)}_{\mathbf{oo}})^{-1}\Sigma^{(i)}_{\mathbf{om}}$ and $\Sigma^{(i)}_{\mathbf{mm}|\mathbf{o}} = \Sigma^{(i)}_{\mathbf{mm}} - \Sigma^{(i)}_{\mathbf{mo}}(\Sigma^{(i)}_{\mathbf{oo}})^{-1}\Sigma^{(i)}_{\mathbf{om}}$ are considered. For all $i \in \mathbf{I}_{S\cap\mathbf{o}}$, the conditional probability that i belongs to the component k is

$$t^{(r)}_{ik} = \frac{p^{(r)}_k\Phi(\mathbf{y}_i^{S\cap\mathbf{o}}|\mu^{(r,\,i)}_{k,\mathbf{o}}, \Sigma^{(r,\,i)}_{\mathbf{oo}})}{\sum_{v=1}^{K} p^{(r)}_v\Phi(\mathbf{y}_i^{S\cap\mathbf{o}}|\mu^{(r,\,i)}_{v,\mathbf{o}}, \Sigma^{(r,\,i)}_{\mathbf{oo}})}.$$

For all $i \in \mathbf{I}_{S\cap\mathbf{o}}$ and for all $(j, l) \in \{1, \ldots, Q\}^2$ such that $\psi(j) \in S, \psi(l) \in S$, the expressions of the conditional expectations and the conditional covariance matrix are given by

$$\tilde{y}^{(k,\,r)}_{ij} := \mathbb{E}[y_i^j|\mathbf{y}_i^{S\cap\mathbf{o}}, \alpha^{(r)}, z_{ik} = 1] = \begin{cases} y_i^j & \text{if } y_i^j \text{ is observed} \\ \left[\mu^{(r,\,i)}_{k,\mathbf{m}|\mathbf{o}} + \mathbf{y}_i^{S\cap\mathbf{o}}\Sigma^{(r,\,i)}_{\mathbf{m}|\mathbf{o}}\right]_j & \text{otherwise} \end{cases}$$

and

$$C^{(r)}_{i,jl} = \begin{cases} \left[\Sigma^{(r,\,i)}_{\mathbf{mm}|\mathbf{o}}\right]_{jl} & \text{if } y_i^j \text{ and } y_i^l \text{ are missing} \\ 0 & \text{otherwise.} \end{cases}$$

M-Step: This step consists of finding the parameter vector

$$\alpha^{(r+1)} = \left(p_1^{(r+1)}, \ldots, p_K^{(r+1)}, \mu_1^{(r+1)}, \ldots, \mu_K^{(r+1)}, \Sigma^{(r+1)}\right)$$

which maximizes $\alpha \mapsto \mathcal{Q}(\alpha|\alpha^{(r)})$. This parameter vector $\alpha^{(r+1)}$ is given by, $\forall k \in \{1, \ldots, K\}$,

$$p_k^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n} t_{ik}^{(r)},$$

$$\mu_k^{(r+1)} = \frac{\sum_{i=1}^{n} t_{ik}^{(r)} \tilde{\mathbf{y}}_i^{(k,r)}}{\sum_{i=1}^{n} t_{ik}^{(r)}}$$

and

$$\Sigma^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n}\left[\left\{\sum_{k=1}^{K} t_{ik}^{(r)}\left(\tilde{\mathbf{y}}_i^{(k,r)} - \mu_k^{(r+1)}\right)'\left(\tilde{\mathbf{y}}_i^{(k,r)} - \mu_k^{(r+1)}\right)\right\} + C_i^{(r)}\right].$$

5.2.3.2 Estimation of regression parameters

This section is devoted to the maximum likelihood parameter estimation for a multidimensional multivariate linear regression where the response matrix and the predictor matrix can have some missing values. The following linear regression of the $n \times \text{card}(S^c)$ response matrix $\mathbf{y}^{S^c}$ on the $n \times \text{card}(R)$ predictor matrix $\mathbf{y}^R$ is considered: For all $i \in \{1, \ldots, n\}$,

$$\mathbf{y}_i^{S^c} = a + \mathbf{y}_i^R \beta + \epsilon_i$$

where the ϵ_i's are i.i.d $\mathcal{N}(0, \Omega)$. Both matrices $\mathbf{y}^{S^c}$ and $\mathbf{y}^R$ may contain missing values which are assumed to be missing at random. Each individual i is assumed to have at least an observed value in $\mathbf{y}_i^{S^c}$ and $\mathbf{y}_i^R$ and each variable is observed at least one time. A possible estimation method (see Little and Rubin, 1986) consists of considering vectors $(\mathbf{y}_i^{S^c}, \mathbf{y}_i^R)$ and assuming that these vectors have a normal distribution $\mathcal{N}(u, \Upsilon)$. The parameter vector $(\hat{u}, \hat{\Upsilon})$ maximizing the observed likelihood is then estimated using an algorithm EM. Then the parameters $\hat{a}$, $\hat{\beta}$ and $\hat{\Omega}$ are deduced from Proposition 5.C.1 and are given by

$$\begin{cases} \hat{a} = \hat{u}_1 - \hat{u}_2\hat{\Upsilon}_{22}^{-1}\hat{\Upsilon}_{21} \\ \hat{\beta} = \hat{\Upsilon}_{22}^{-1}\hat{\Upsilon}_{21} \\ \hat{\Omega} = \hat{\Upsilon}_{11} - \hat{\Upsilon}_{12}\hat{\Upsilon}_{22}^{-1}\hat{\Upsilon}_{21} \end{cases}$$

where $\hat{u} = (\hat{u}_1, \hat{u}_2)$ and $\hat{\Upsilon} = \begin{pmatrix} \hat{\Upsilon}_{11} & \hat{\Upsilon}_{12} \\ \hat{\Upsilon}_{21} & \hat{\Upsilon}_{22} \end{pmatrix}$ are decomposed as $(\mathbf{y}^{S^c}, \mathbf{y}^R)$. This strategy can be easily carried out but it is based on an unrealistic assumption. Indeed vectors $(\mathbf{y}_i^{S^c}, \mathbf{y}_i^R)$ cannot fulfill this normal law assumption since the vectors $\mathbf{y}_i^R$ are modelled by a Gaussian mixture.

Thereby, we suggest an other strategy requiring the predictor matrix $\mathbf{y}^R$ to be totally observed. Preliminary, the matrices $\mathbf{y}^{S^c}$ and $\mathbf{y}^R$ are restricted to the individual subset $\mathbf{I}_{R\cap\mathbf{o}} = \{i \in \{1,\ldots,n\};\ \mathbf{y}_i^{R\cap\mathbf{o}} = \mathbf{y}_i^R\}$ and thus, the missing values only belong to the response matrix $\mathbf{y}^{S^c}_{\mathbf{I}_{R\cap\mathbf{o}}}$. An EM algorithm is proposed to estimate the regression parameters $(\hat{a},\hat{\beta},\hat{\Omega})$ maximizing the observed likelihood $f_{\text{reg}}(\mathbf{y}^{S^c\cap\mathbf{o}}_{\mathbf{I}_{R\cap\mathbf{o}}}|a+\mathbf{y}^R_{\mathbf{I}_{R\cap\mathbf{o}}}\beta,\Omega)$. This algorithm consists of maximizing the expectation of the complete likelihood $f_{\text{reg}}(\mathbf{y}^{S^c}_{\mathbf{I}_{R\cap\mathbf{o}}}|a+\mathbf{y}^R_{\mathbf{I}_{R\cap\mathbf{o}}}\beta,\Omega)$ conditionally to the observed values and a current parameter value. A detailed description of the E-step and the M-step of this algorithm is given in Appendix 5.B.

- **Initialisation:** The required initial parameter values are determined by

$$(a^{(1)\prime},\beta^{(1)\prime})' = (X^{(R)\prime}_{\text{init}}X^{(R)}_{\text{init}})^{-1}X^{(R)\prime}_{\text{init}}\mathbf{y}^{S^c}_{\text{init}}$$

  and

$$\Omega^{(1)} = \frac{1}{n_{\text{init}}}\mathbf{y}^{S^c\prime}_{\text{init}}(I - X^{(R)}_{\text{init}}(X^{(R)\prime}_{\text{init}}X^{(R)}_{\text{init}})^{-1}X^{(R)\prime}_{\text{init}})\mathbf{y}^{S^c}_{\text{init}}$$

  where $\mathbf{y}^{S^c}_{\text{init}}$ denotes the restriction of $\mathbf{y}^{S^c}$ to the totally observed individual subset $\{i \in \{1,\ldots,n\};\ \mathbf{y}_i^{S^c\cap\mathbf{o}} = \mathbf{y}_i^{S^c}\}$ with cardinal n_{init} and $X^{(R)}_{\text{init}}$ is the restriction of $X^{(R)} = ((1,\ldots,1)',\mathbf{y}^{R\prime})'$ on the same individual subset.

- **E-Step:** At the r^{th} iteration, for all couples (j,l) such that $\psi(j),\psi(l) \in S^c$ and all $i \in \mathbf{I}_{R\cap\mathbf{o}}$, we compute

$$\tilde{y}^{(r)}_{i,j} = \begin{cases} y_i^j & \text{if } y_i^j \text{ is observed} \\ (a^{(r)} + \mathbf{y}_i^R\beta^{(r)})_j & \text{otherwise} \end{cases}$$

  and

$$C^{(r)}_{i,jl} = \begin{cases} \Omega^{(r)}_{jl} & \text{if } y_i^j \text{ and } y_i^l \text{ are missing} \\ 0 & \text{otherwise.} \end{cases}$$

- **M-Step:** The updated parameters are given by

$$(a^{(r+1)\prime},\beta^{(r+1)\prime})' = (X^{(R)\prime}X^{(R)})^{-1}X^{(R)\prime}\tilde{\mathbf{y}}^{(r)}$$

  and

$$\Omega^{(r+1)} = \frac{1}{n}\left[\tilde{\mathbf{y}}^{(r)\prime}\{I - X^{(R)}(X^{(R)\prime}X^{(R)})^{-1}X^{(R)\prime}\}\tilde{\mathbf{y}}^{(r)} + \sum_{i=1}^{n} C_i^{(r)}\right].$$

5.2.4 Modified variable selection algorithm and applications

5.2.4.1 Changes for the variable selection algorithm

From the description of the variable selection algorithm given in Chapter 3, we describe now the main changes required to analyse datasets with missing values, with a fixed mixture form $m = [p_k LC]$. The general principle of the new algorithm is as follows:

1. For all (K, m), we search

$$(\hat{S}(K,m), \hat{R}(K,m)) = \underset{(S,R)\in\mathcal{V}}{\operatorname{argmax}} \ \operatorname{crit}(K, m, S, R)$$

by a backward stepwise algorithm detailed hereafter.

2. We determine

$$\hat{K} = \underset{K\in\{1,\ldots,K_{\max}\}}{\operatorname{argmax}} \ 2\ln\left\{f(\mathbf{y}^{\mathbf{o}}|K, m, \hat{S}(K,m), \hat{R}(K,m), \hat{\theta}\,)\right\} - \Xi_{(K,m,\hat{S}(K,m),\hat{R}(K,m))}\ln(n)$$

where the parameter vector $\hat{\theta}$ is determined in two parts as explained in Section 5.2.3.

Finally, the selected model is $(\hat{K}, m, \hat{S}(\hat{K}, m), \hat{R}(\hat{K}, m))$.

The first step consists of a backward stepwise variable selection algorithm as presented in Section 3.4.2. It means that, all variables being selected at the beginning, at each step, a variable block is excluded or included. It is based on the comparison of the following criterion value differences for each variable block j:

$$\begin{aligned}\mathrm{BIC}_{\mathrm{diff}}(j) &= 2\ln\left[f_{\mathrm{clust}}(\mathbf{y}^{(S\cup j)\cap\mathbf{o}}|K, m, \hat{\alpha}_1)\right] - \lambda_{(K,m,S\cup j)}\ln(n) \\ &- \left\{2\ln\left[\prod_{i=1}^{n}\sum_{k=1}^{K}\hat{p}_k\Phi(\mathbf{y}^{(S\cup j)\cap\mathbf{o}}|\hat{\nu}_{k,\mathbf{o}}^{(i)}, \hat{\Delta}_{\mathbf{oo}}^{(i)})\right] - (\lambda_{(K,m,S)} + \nu_{(j,R[j])})\ln(n)\right\}\end{aligned}$$

where $\hat{\alpha}_1$ is the parameter estimator of the Gaussian mixture on $S\cup j$ (see Section 5.2.3.1), and $(\hat{p}_k, \hat{\nu}_k)_{1\leq k\leq K}$ and $\hat{\Delta}$ are computed according to the parameter vector $\hat{\alpha}_2$ maximizing $f_{\mathrm{clust}}(\mathbf{y}^{S\cap\mathbf{o}}|K, m, \alpha_2)$ (see Section 5.2.3.1) and the parameter estimates $(\hat{a}, \hat{\beta}, \hat{\Omega})$ of the linear regression of j on $R[j]$ (see Section 5.2.3.2) in order to evaluate the second observed likelihood as explained in Section 5.2.2.2. The subset $R[j]$ included into S and composed of the variable blocks required to explain the variable j, is determined using a backward variable selection algorithm. This second backward stepwise algorithm is analogous to the algorithm presented in Appendix 3.B. It is initialized with $R[j] = S$ and is making use of exclusion and inclusion steps. It is based on the comparison of the following quantities for each variable block ℓ of S:

$$\begin{aligned}\mathrm{BIC}_{\mathrm{diffreg}}(\ell) &= 2\ln\left[f_{\mathrm{reg}}(\mathbf{y}_{I_{S\cap\mathbf{o}}}^{j\cap\mathbf{o}}|\hat{a}_1 + \mathbf{y}_{I_{S\cap\mathbf{o}}}^{R[j]}\hat{\beta}_1, \hat{\Omega}_1)\right] - \nu_{(j,R[j])}\ln[\mathrm{card}(I_{S\cap\mathbf{o}})] \\ &- \left\{2\ln\left[f_{\mathrm{reg}}(\mathbf{y}_{I_{S\cap\mathbf{o}}}^{j\cap\mathbf{o}}|\hat{a}_2 + \mathbf{y}_{I_{S\cap\mathbf{o}}}^{R[j]\setminus\ell}\hat{\beta}_2, \hat{\Omega}_2)\right] - \nu_{(j,R[j]\setminus\ell)}\ln[\mathrm{card}(I_{S\cap\mathbf{o}})]\right\}\end{aligned}$$

where $I_{S\cap\mathbf{o}}$ is the subset of the individuals which are totally observed on S. The observed likelihoods for the regression are calculated on $I_{S\cap\mathbf{o}}$ since $R[j]$ and $R[j]\setminus\ell$ are subsets of S, allowing the predictor matrix to be totally observed at each step of this backward algorithm. The parameters $(\hat{a}_q, \hat{\beta}_q, \hat{\Omega}_q)_{q\in\{1,2\}}$ are estimated using the EM algorithm presented in Section 5.2.3.2.

5.2.4.2 Simulated example

A simulated dataset consisting of 2000 data points from a mixture of four Gaussian distributions $\mathcal{N}(\mu_k, \Sigma)$ is considered. The mean vectors are $\mu_1 = (0,0,0)$, $\mu_2 = (-6,6,0)$, $\mu_3 = (0,0,6)$, $\mu_4 = (-6,6,6)$ and the variance matrix is $\Sigma = A' \times \text{diag}(6\sqrt{2}, 1, 2) \times A$ where the matrix A is the product of two 3×3 rotation matrices around the Oz and Ox axis with angle $\pi/6$ and $\pi/3$ respectively. The proportion vector of this mixture is $\mathbf{p} = (0.25, 0.25, 0.2, 0.3)$. This dataset is plotted according to these three variables in Figure 5.1. The fourth and fifth variables are defined for all $i \in \{1, \ldots, n\}$ by

$$(y_i^4, y_i^5) = (-1, 2) + (y_i^1, y_i^2)((0.5, 2)', (1, 0)') + \varepsilon_i,$$

ε_i being sampled from a $\mathcal{N}(0, \text{rot}(\pi/6)' \,\text{diag}(1,3)\, \text{rot}(\pi/6))$ density where $\text{rot}(\pi/6)$ is the 2×2 plane rotation matrix with angle $\pi/6$. Two noisy independent standard centered Gaussian variables are also appended. The true model is thus

$$(K_0 = 4, m_0 = [p_k LC], S_0 = \{1,2,3\}, R_0 = \{1,2\}).$$

In order to evaluate the performance of our variable selection procedure with missing values, $c \in \{1, 5, 10, 15, 20\}$ percent of values, chosen at random, are marked as missing. Our procedure is applied with a maximum number of components $K_{\max} = 8$ and the true mixture form $m_0 = [p_k LC]$ is fixed. The results obtained by applying our new variable selection procedure in these different scenarii are presented in Table 5.1. In all scenarii, our procedure selects the true variable partition and the true number of clusters. But the clustering error rate, calculated on the individuals having at least one observed value on the three first variables, increases with the number of missing values.

The 264 misclassified individuals for the scenario with $c = 20\%$ of missing values are shared out according to their initial group and the position of their missing values on the three relevant clustering variables in Table 5.2, in order to explain the increase of the clustering error rates. We point out that few individuals without missing values on the three relevant variables are misclassified. Moreover, the individuals having at least a missing value for the third variable belong to Groups 2 and 3 most often. According to the model used to simulate the data, the third variable allows to distinguish between Groups 1 and 3, and between Groups 2 and 4 (see Figure 5.1). Hence the 52 individuals of Group 2 which are not observed on the third variable are consequently all clustered in Cluster 4. This reason explains the increase of the clustering error rate with the increase of the number of missing values since the applied MAP rule (see Chapter 1) is based on the conditional probabilities which are only evaluated on the observed values.

Note that some individuals are not clustered (see Table 5.1) because they are not observed on the relevant clustering variables. An intuitive solution would consist of attributing such a no classified individual to the cluster for which it is the closest to the average profile on all the variables. For each cluster k, the average profile $\bar{\mathbf{y}}_{[k]}$ of its individuals is determined on all the variables:

$$\forall j \in \{1, \ldots, Q\},\ \bar{y}_{[k]}^j = \frac{1}{\text{card}(\{i; M_{ij} = 1\})} \sum_{i=1}^{n} y_i^j \mathbb{1}_{M_{ij}=1}$$

where M is the missing-data indicator matrix defined by (5.1). Then for each no classified individual i, its distance to the average profile of each cluster is computed

$$d(k) = \sum_{j=1}^{Q} (y_i^j - \bar{y}_{[k]}^j)^2 \mathbb{1}_{M_{ij}=1}.$$

Finally, the individual i is assigned to the $\tilde{k}^{th}$ cluster such that

$$\tilde{k} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmin}} \; d(k).$$

Nevertheless, this strategy is unsuitable if there are few redundant variables among the irrelevant variables and if the redundant variables are not linked to the whole relevant variables. For instance, 80%, 67% and 50% of the no classified individuals are correctly clustered with this strategy for Scenarii $c = 10\%$, $c = 15\%$ and $c = 20\%$ respectively. For the simulated dataset, Variables 4 and 5 are redundant and regressed on Variables 1 and 2 which are not sufficient to distinguish the four clusters. Thus, we advise to keep these individuals as no classified as much as possible.

percentage of missing values	$\hat{K}$	$\hat{S}$	$\hat{R}$	Error rate	Number of no classified individuals
$c = 0\%$	4	$\{1,2,3\}$	$\{1,2\}$	1.15%	0
$c = 1\%$	4	$\{1,2,3\}$	$\{1,2\}$	1.55%	0
$c = 5\%$	4	$\{1,2,3\}$	$\{1,2\}$	3.85%	0
$c = 10\%$	4	$\{1,2,3\}$	$\{1,2\}$	7.17%	5
$c = 15\%$	4	$\{1,2,3\}$	$\{1,2\}$	10.17%	3
$c = 20\%$	4	$\{1,2,3\}$	$\{1,2\}$	13.27%	10

Table 5.1: Models selected by our procedure for different percentages of missing values.

Positions of the missing values	Group 1	Group 2	Group 3	Group 4
Variable 1	9	7	5	10
Variable 2	0	1	8	0
Variable 3	1	52	53	10
Variables 1 and 2	0	13	14	0
Variables 1 and 3	1	15	12	3
Variables 2 and 3	0	16	21	1
No missing values	4	3	2	3
Total	15	107	115	27

Table 5.2: Number of misclassified individuals, among the 264 ones for the scenario with $c = 20\%$ of missing values, according to their initial group and the position of their missing values on the three relevant clustering variables.

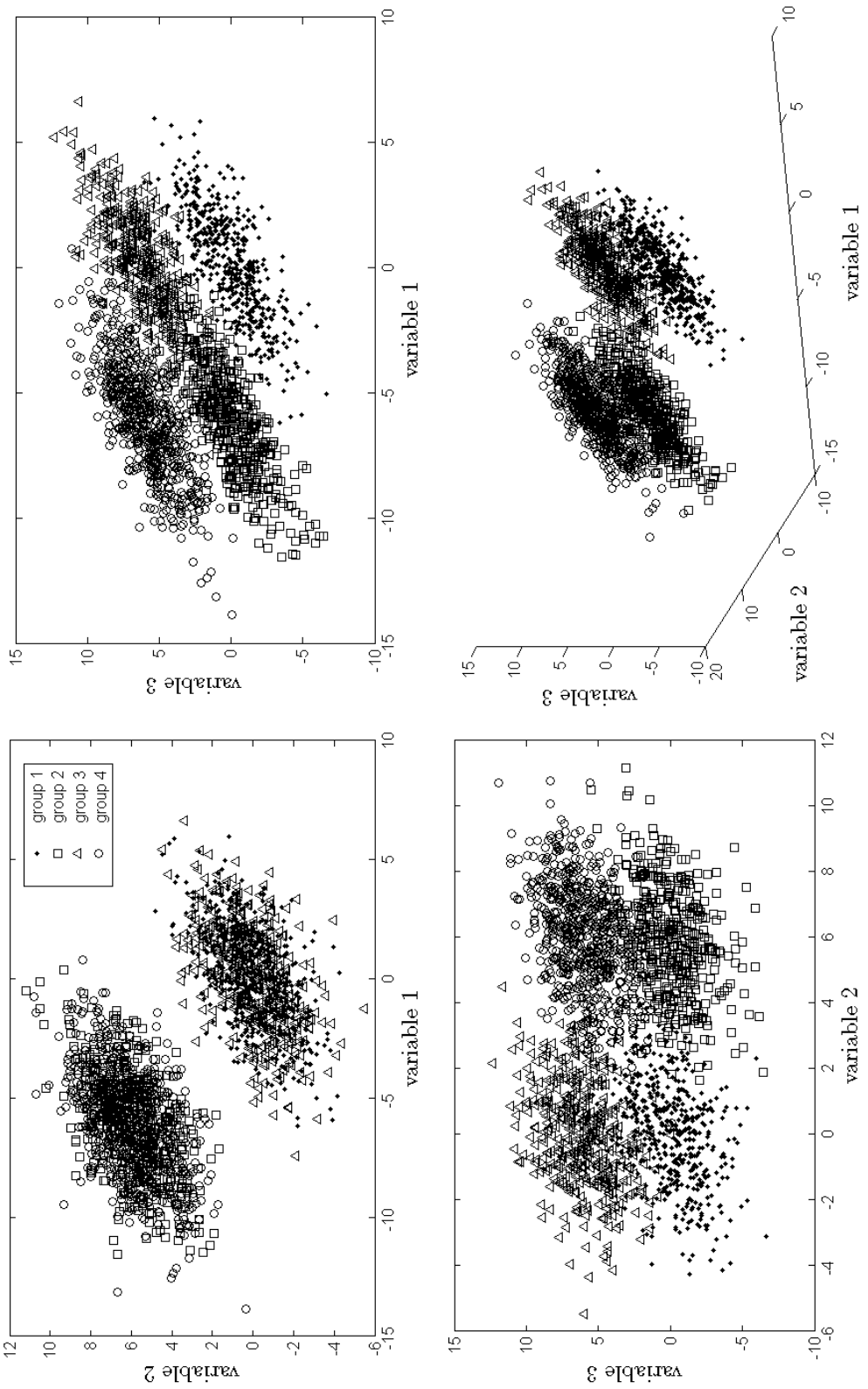


Figure 5.1: Representation of the simulated dataset on the three relevant variables.

5.2.4.3 Transcriptome dataset

This section is devoted to the clustering of a transcriptome dataset. We go back to the analysis of genes declared differentially expressed at least once in a time course of the hypocotyl growth switch addressed in Chapter 3. Recall that the behaviour of these genes is studied on $T = 7$ projects with $Q = 27$ experiments (see Table 3.4). The dataset is now composed of $n = 1267$ genes sharing out among the 1020 previous studied genes, 118 genes with missing values at random and 129 genes which were removed of the previous analysis because they do not satisfy the homoscedastic assumption in the differentially analysis (see Chapter 2 for more details). Nevertheless, for these 129 genes, a test statistic can be calculated in an experiment as the difference of expression normalized by the estimated standard deviation obtained with all the genes satisfying the homoscedastic assumption. If these test statistics were declared missing, the missing values of the dataset would not be missing at random. On the contrary, the 118 genes have at least one missing values due to various reasons in the laboratory process. These missing values occur when there is an imperfection on the array or hybridation failures for instance. Since the problematic spots are manually flagged as missing, the missing values are missing at random. Consequently, 9.3% of genes have at least one missing value in this dataset and $c = 0.38\%$ of the values are missing.

Our variable selection procedure is performed on this transcriptome dataset with a maximal number of mixture components equal to 20. The mixture form is fixed to $m = [p_k LC]$, this form being selected in the analysis of Chapter 3. Our method selects a clustering with 17 clusters, which have different sizes and different repartitions of gene types (see Figure 5.2). The variable blocks declared relevant for clustering are Projects 1, 2, 3, 4, 6 and 7 and the four last ones are required to explain Project 5. Several gene subsets of this clustering are close to the ones of the previous clustering obtained for the dataset restricted to the 1020 totally observed genes. For instance, Clusters 1, 3, 4, 6, 9, 12, 13, 15 and 17 of the previous clustering are similar to Clusters 6, 11, 17, 8, 3, 1, 7, 4 and 16 of the new clustering respectively.

Clusters 5 and 9, which contain 3 and 12 new genes respectively, have no link with the previous clustering according to Figure 5.2. Their genes have characteristic expressions in Project 2 contrary to the genes of the other new clusters. This can explain that Project 2 is selected in addition compared with the previous relevant clustering projects. The same projects are selected to explain the irrelevant Project 5. The explanations based on the redifferentiation of cells and the formation of giant cells given in Section 3.7 are always valid and large correlations between experiments of this project are still present (see Table 5.3).

Our new clustering is compared to the 11 gene subgroups underlined by S. Pelletier with an exploratory method (see its description in Section 3.7). Some of these 11 gene subgroups are recovered among the 17 clusters (see Figure 3.6). For instance, the genes of Subgroup 2a belong to Cluster 16 whose the expression profiles are very homogeneous and have a specific behaviour in Projects 3, 4, 6 and 7. Subgroup 5b2 is still totally contained in Cluster 7 which highlights the specific gene behaviour in Project 7. The genes of this subgroup are clustered with 22 other genes which are not discovered by the exploratory

method. These promising results obtained with our procedure are coherent with the prior knowledge of biologists but require to be more precisely studied by biologists in order to validate the use of our procedure for gene expression data analysis.

	P5-1	P5-2	P5-3
a	0.42	-0.04	-0,40
P3-1	0.10	0.12	0.15
P3-2	-0.02	0.31	0.25
P3-3	0.01	-0.11	-0.06
P3-4	-0.03	0.24	0.25
P4-1	-0.14	-0.02	-0.01
P4-2	0.08	-0.01	-0.02
P4-3	-0.15	-0.07	-0.09
P6-1	0.01	0.13	0.17
P6-2	-0.15	-0.08	-0.04
P6-3	0.00	-0.06	-0.13
P7-1	-0.28	-0.26	-0.22
P7-2	0.12	-0.09	0.11

	P5-1	P5-2	P5-3
P5-1	9.71	6.96	7.68
P5-2	6.96	14.56	13.36
P5-3	7.68	13.36	18.06

	P5-1	P5-2	P5-3
P5-1	1.00	0.59	0.58
P5-2	0.59	1.00	0.82
P5-3	0.58	0.82	1.00

Table 5.3: The regression coefficient matrix $(\hat{a}', \hat{\beta}')'$ is given on the left and on the right, the variance matrix $\hat{\Omega}$ on the top and the correlation matrix on the bottom. Pi-j denotes Experiment j in Project i.

			Without missing values																
		cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	cluster	nb of genes	7	120	25	11	42	424	14	29	43	94	11	149	19	4	6	2	20
With missing values	1	143 [132, 10, 1]								1		1		130					
	2	6 [6, 0,0]	1			5													
	3	67 [50, 8, 9]		5		1		1	1		38			3					1
	4	8 [6, 1,1]											1				5		
	5	22 [10, 3, 9]					6	3			1								
	6	8 [7, 0, 1]	6				1												
	7	28 [21,2,5]							1				1		19				
	8	510 [420, 56, 34]		1	1	4	6	383	2	10	3	6	4						
	9	12 [10, 2, 0]		6							1		1	2					
	10	33 [30, 1, 2]					7	20		1			2						
	11	40 [32, 0, 4]			24			5	1									2	
	12	36 [32, 7, 1]		1		1	22		5				1	2					
	13	16 [13, 0, 4]								11			1				1		
	14	167 [133, 20, 14]		106				1	4	6		11		5					
	15	143 [95, 8, 40]		1				11				76		7					
	16	24 [19, 0, 5]																	19
	17	4 [4, 0, 0]														4			

Figure 5.2: Comparison of the two clusterings obtained by applying our variable selection procedure with and without missing values. The number of genes per cluster for the dataset with missing values given in the second column, is followed by: [the number of genes totally observed and previously studied, number of genes with at least one missing value, the number of genes totally observed but not previously studied].

	clusters	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
n° of group	nb of genes	143	6	67	8	22	8	28	510	12	33	40	36	16	167	143	24	4
1	10 (12)															10		
2a	19			4													15	
2b	14								9		3	2						
2c	8															8		
3	11								9		1			1				
4	10			1		1					3	1	3		1			
5a2	17	17																
5a3	12															12		
5b	15 (17)	15																
5b2	6							6										
5c	28	2		21								1			3			1

Figure 5.3: Comparison with the subgroups underlined by S.Pelletier (see Section 3.7).

5.3 Imputation methods of missing data

5.3.1 Imputation methods for gene expression matrix study

In the gene expression clustering framework, imputation methods are usually used to complete the gene expression matrix before the clustering process. During the last ten years, several imputation methods have been proposed. The widely usual methods replace missing values with zero (Alizadeh et al., 2000), least often with the row (gene) average (ROW-imputation) or sometimes with the column (array) average (COL-imputation). These methods do not take the correlation structure of data into account. Troyanskaya et al. (2001) propose two correlation-based imputation methods. First, the KNN-imputation method consists of finding the k genes which are the closest to the gene of interest with missing values according to a distance metric, most frequently the Euclidian distance or Pearson correlation. The missing value is then estimated by the weighted average of these k-nearest neighbour genes for the same array where the weights are calculated from their similarity measurement. Second, the SVD-imputation method initially sets missing entries to the row average and then uses the singular value decomposition of the gene expression matrix to obtain orthogonal principal components called "eigengenes" (Alter et al., 2000). The k most significant eigengenes are selected and the linear regression coefficients of a gene with missing entries on these eigengenes are used to estimate its missing values. Oba et al. (2003) show that their BPCA-imputation method using Bayesian estimation to fit a probabilistic PCA model outperform these two previous methods. The GMC-imputation approach proposed by Ouyang et al. (2004) consists of modelling data by Gaussian mixtures with k components varying between 1 to K and of estimating missing values by an EM algorithm. Finally, the missing values are imputed by the average of these estimators indexed by k. Jornsten et al. (2005) search to take advantage of these previous methods. Their LinCmb method considers a convex combination of the estimates obtained by ROW-, KNN-, SVD-, BPCA- and GMC-imputation methods, where weights are chosen

according to the data structure. Others methods based on the least squares principle are also developed. Bø et al. (2004) propose the LS-imputation method which minimizes the sum of squared errors of a regression model and uses the correlation between genes and arrays. The Local Least Squares imputation (LLS-imputation) of Kim et al. (2005) finds neighbour genes by KNN method based on the Pearson correlation and estimates missing values according to a multiple regression on the k-nearest neighbour genes. Contrary to methods estimating independently missing entries, Friedland et al. (2006) suggest a fixed rank approximation algorithm (FRAA) in order to take the influence of a missing value estimation on the others into account. All these previous methods use only the gene expression matrix to estimate missing values. Other methods are proposed using external information in the imputation process. Tuikkala et al. (2006) use the information of gene ontology annotation to improve the missing value imputation in their GOKNN-imputation method. Gan et al. (2006) propose a set theoretic framework based on projection onto convex sets (POCS) in order to consider different biological knowledge types into the estimation process. The integrative Missing Value Estimation (iMISS) of Hu et al. (2006) allows them to derive neighbour genes by taking reference data sets into consideration. This imputation method list is not exhaustive, we can also quote the CMVE-imputation method of Sehgal et al. (2005), the CIAO-imputation proposed by Kim et al. (2007) or the method of Zhou et al. (2003). Brock et al. (2008) proposed the entropy-based selection (EBS) scheme and the simulation-based self-training selection (STS) scheme in order to select an appropriate method among an imputation method collection for a studied dataset. They conclude after a comparison of eight methods that the three top-performing methods are the LS-, LLS- and BPCA-imputation methods.

5.3.2 Comparison of some imputation methods

The behaviour of six of the imputation methods quoted in Section 5.3.1 on different datasets is now studied. The ZERO-, ROW- and COL-imputation methods which are the more used while they are simplistic, the KNN-imputation method and the BPCA- and LLS-imputation methods which seem to be the two best methods according to different analyses, are considered. These six imputation methods are now specified:

- **ZERO-imputation:** This imputation method consists of replacing all missing values by zero. It is the usual imputation method used to gene expression matrix.

- **ROW-imputation:** This imputation method consists of replacing all missing values of an individual i (a gene) by the average of its observed values.

- **COL-imputation:** This imputation method consists of replacing all missing values for a variable j (an experiment) by the average of its observed values.

- **KNN-imputation:** The k-nearest neighbours are found using the Euclidean metric for each individual with missing values, confined to the variables for which the individual is observed. If a candidate neighbour is missing some of the coordinates used

to calculate the distance, the distance is computed on the no missing coordinates. When the k-nearest neighbours are found for an individual, the missing elements are imputed by averaging those (no missing) elements of its neighbours. We use the R package "impute.knn"[1] proposed by Hastie et al. with the choice per default $k = 10$ number of neighbours.

- **LLS-imputation:** This method estimates all missing values of an individual simultaneously. First, it selects the k-nearest neighbours of an individual with missing values based on the Pearson correlation. Second, this method performs a multiple regression using all k neighbours. Then the missing values are imputed, based on the least square estimates. Kim et al. (2005) propose an heuristic method to select the number of neighbours k. The Matlab software "impute_llsq_l2.m"[2] proposed by Kim et al. is used in this chapter.

- **BPCA-imputation:** This method uses Bayesian estimation to fit a probabilistic PCA model, which is based on the assumption that the factor scores and the residuals obey normal distributions. The Bayesian estimation calculates the posterior distribution of model parameter θ and factor scores X according to the Bayes theorem $p(\theta, X|Y) \propto p(Y, X|\theta)p(\theta)$ where $p(\theta)$ is a prior distribution which denotes a priori preference for θ. An EM-like repetitive algorithm is used to iteratively estimate the posterior distribution $q(\theta)$ of the model parameter and the posterior distribution of the missing values $q(\mathbf{y^m}) = \int p(\mathbf{y^m}|\mathbf{y^o}, \theta)q(\theta)d\theta$. Finally, the missing entries are imputed by $\int \mathbf{y^m} q(\mathbf{y^m}) d\mathbf{y^m}$. The Matlab software "BPCAfill.m"[3] proposed by Oba et al. is used.

The accuracy measure for an imputation method is usually based on the root mean squared error (RMSE) between the original matrix $\mathbf{y}$ and the imputed matrix $\hat{\mathbf{y}}$. Different normalizations of the RMSE are used: Troyanskaya et al. (2001) and Friedland et al. (2006) normalize the RMSE by the mean of the complete true matrix values. Wang et al. (2006) divides the RMSE by the standard deviation of the complete matrix while Oba et al. (2003) and Kim et al. (2005) use the standard deviation of the true values of the missing entries. The normalization by the root mean squared true values of the missing entries is used by Ouyang et al. (2004), Hu et al. (2006), Tuikkala et al. (2006) and Jornsten et al. (2005). We prefer in this chapter to consider this last normalization defined by

$$NRMSE = \sqrt{\frac{\text{mean}(\mathbf{y^m} - \hat{\mathbf{y}}^\mathbf{m})^2}{\text{mean}(\mathbf{y^m})^2}} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{Q}\left(y_i^j - \hat{y}_i^j\right)^2 \mathbb{1}_{M_{ij}=0}}{\sum_{i=1}^{n}\sum_{j=1}^{Q}(y_i^j)^2 \mathbb{1}_{M_{ij}=0}}}$$

[1]The R package *impute.knn* is available in `http://rss.acs.unt.edu/Rdoc/library/impute/html/impute.knn.html`

[2]The Matlab software *impute_llsq_l2.m* is available in `http://www.cs.umn.edu/~hskim/tools.html`

[3]The Matlab software *BPCAfill.m* is available in `http://hawaii.aist-nara.ac.jp/~shige-o/tools/BPCAFill.html`

since it is equal to 1 for the ZERO-imputation, allowing us to consider this imputation method as reference. When the estimated values are accurate, the NRMSE reaches its minimum value 0 and when the missing value estimation is poor, the NRMSE becomes large.

The six imputation methods are compared on the following four datasets, composed of two simulated examples and two transcriptome datasets:

- **The simulated example:** It is the simulated dataset presented in Section 5.2.4.2. Recall that this dataset is composed of $n = 2000$ observations described by $Q = 7$ variables. On the three first variables, the data are sampled from a Gaussian mixture. The fourth and fifth variables are a linear combination of the two first variables and the two last variables are noisy variables.

- **Waveform dataset:** This dataset consists of $n = 5000$ individuals described by $Q = 40$ variables. It is studied in the two previous chapters (see Section 3.6.3 for a detailed description). Recall that the nineteen last variables are sampled from $\mathcal{N}(0, 1)$ and the twenty first are based on a linear combination of two of three wave functions.

- **First transcriptome dataset:** This transcriptome dataset is composed of $n = 5966$ genes studied on $Q = 10$ experiments. Its distinctiveness is that whole genes are declared non-differentially expressed in five experiments.

- **Second transcriptome dataset:** This transcriptome dataset consists of $n = 18417$ genes described by seven projects (see Table 3.4) with $Q = 27$ experiments. These genes are the whole genes of the CATMA microarray totally observed in these seven projects.

The NRMSE values for the different imputation methods applied to estimate c percent of missing entries for the four datasets are given in Table 5.4 and are graphically represented in Figure 5.4. The percentage of missing values belongs to $c \in \{1, 5, 10, 15, 20\}$ and the entries are marked as missing at random. On the whole, the KNN-imputation method has an intermediate behaviour between the simple ZERO-, ROW- and COL-imputation methods and the best methods are LLS- and BPCA-imputation methods. The ROW-imputation method behaves often worse. For transcriptome datasets, this method does not seem to be appropriate since a gene may have a different behaviour from an experiment to an other. For the waveform dataset, its better behaviour is due to the majority of noisy variables sampled from a standard centered Gaussian distribution. The COL- and ZERO-imputation methods have a similar behaviour for transcriptome datasets because much genes are non-differentially expressed in an experiment thus the average of values for an experiment is close to zero. On the contrary, they have different behaviours on the two simulated examples where the value average per variable is not necessarily null. The KNN-imputation method has a better behaviour that the three previous methods except for the waveform dataset. This improvement is certainly due to the local imputation since this

method is based on only the k-nearest neighbours of an individual having missing values to estimate them. In the four examples, the more accurate imputation values are estimated by the LLS- and the BPCA-imputation methods. For the transcriptome datasets, the LLS-imputation method has a better behaviour than BPCA for low percents of missing values while it is the opposite for large missing entry percents. As illustrated in Figure 5.5, for the imputation of the 20% of missing values in the second transcriptome dataset, the LLS- and BPCA-imputation methods give close impute values and underestimate extreme values (where genes are differentially expressed) in absolute value mainly. The estimated values given by the KNN-imputation method are more scattered along the main diagonal.

Studied dataset	Imputation method	$c = 1\%$	$c = 5\%$	$c = 10\%$	$c = 15\%$	$c = 20\%$
Simulated dataset	ZERO	1	1	1	1	1
	ROW	1.1008	1.0835	1.0737	1.0671	1.0925
	COL	0.8013	0.8255	0.8035	0.8284	0.8291
	KNN	0.3617	0.4883	0.5137	0.5180	0.5833
	LLS	**0.3337**	0.4175	0.4143	0.4467	0.5180
	BPCA	0.3344	**0.3784**	**0.4020**	**0.4082**	**0.4450**
Waveform dataset	ZERO	1	1	1	1	1
	ROW	0.3552	0.3872	0.3855	0.3814	0.3789
	COL	0.3217	0.3264	0.3177	0.3204	0.3171
	KNN	0.5652	0.5635	0.5565	0.5524	0.5624
	LLS	0.5450	0.5477	0.0313	0.0305	**0.0340**
	BPCA	**0.0070**	**0.0182**	**0.0306**	**0.0300**	0.0348
First transcriptome dataset	ZERO	1	1	1	1	1
	ROW	1.0618	1.0537	1.0443	1.0612	1.0619
	COL	0.9992	1.0004	0.9992	0.9995	0.9997
	KNN	0.7743	0.8481	0.8768	0.8717	0.9308
	LLS	**0.6572**	**0.6872**	**0.6883**	0.7418	0.7698
	BPCA	0.6788	0.7034	0.6961	**0.7181**	**0.7600**
Second transcriptome dataset	ZERO	1	1	1	1	1
	ROW	1.0180	1.0165	1.0176	1.0195	1.0214
	COL	0.9999	0.9995	0.9997	0.9996	0.9997
	KNN	0.8008	0.7957	0.8369	0.8328	0.8818
	LLS	**0.6089**	**0.6009**	**0.6241**	0.6640	0.7054
	BPCA	0.6211	0.6241	0.6342	**0.6618**	**0.6975**

Table 5.4: NRMSE values of the six imputation methods on the four datasets. The best value for each situation is in bold.

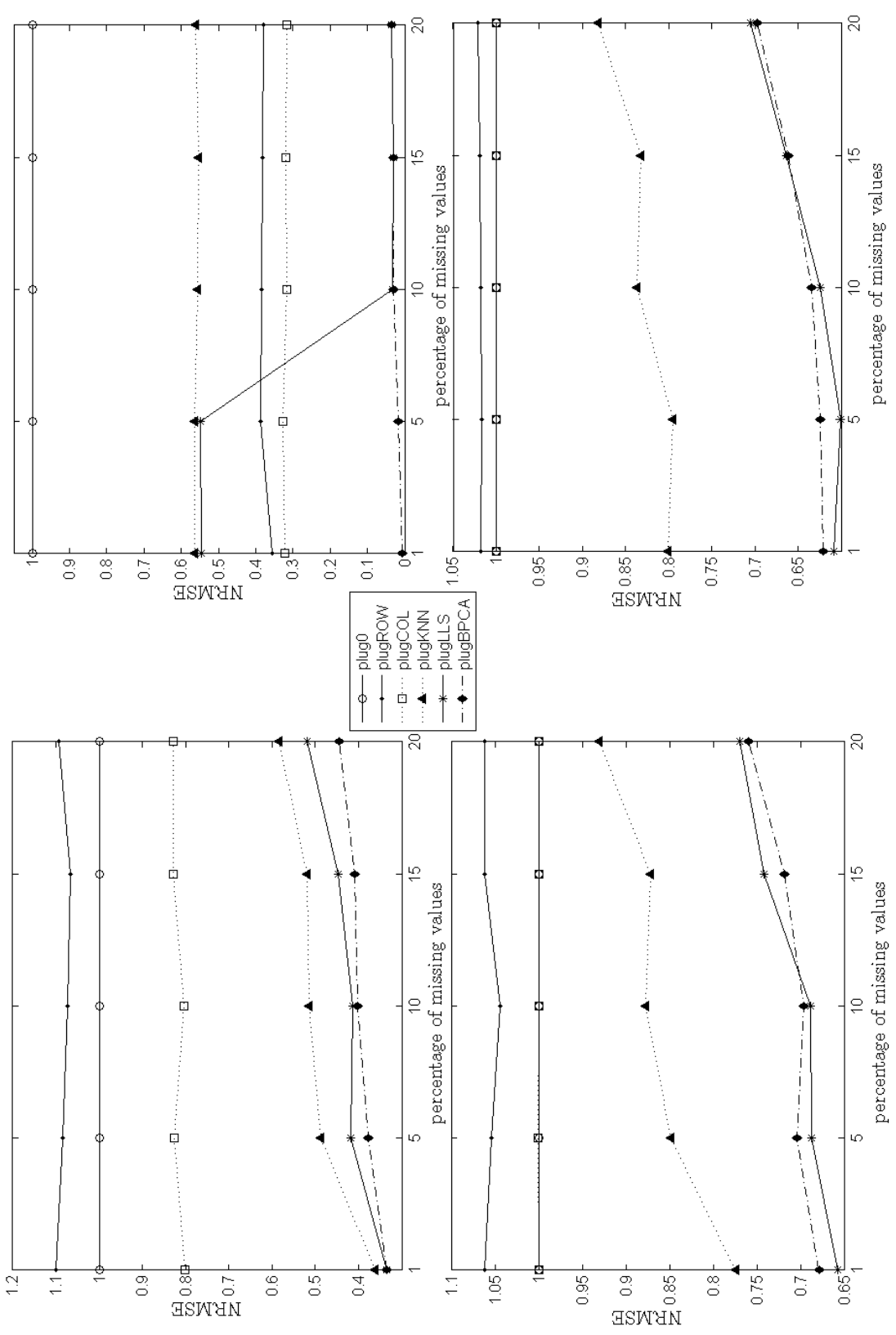


Figure 5.4: Estimation ability (NRMSE) by ZERO-, ROW-, COL-, KNN-, LLS- and BPCA-imputation on the simulated example (top, left), the waveform dataset (top, right), the first transcriptome dataset (bottom, left) and the second transcriptome dataset (bottom, right).

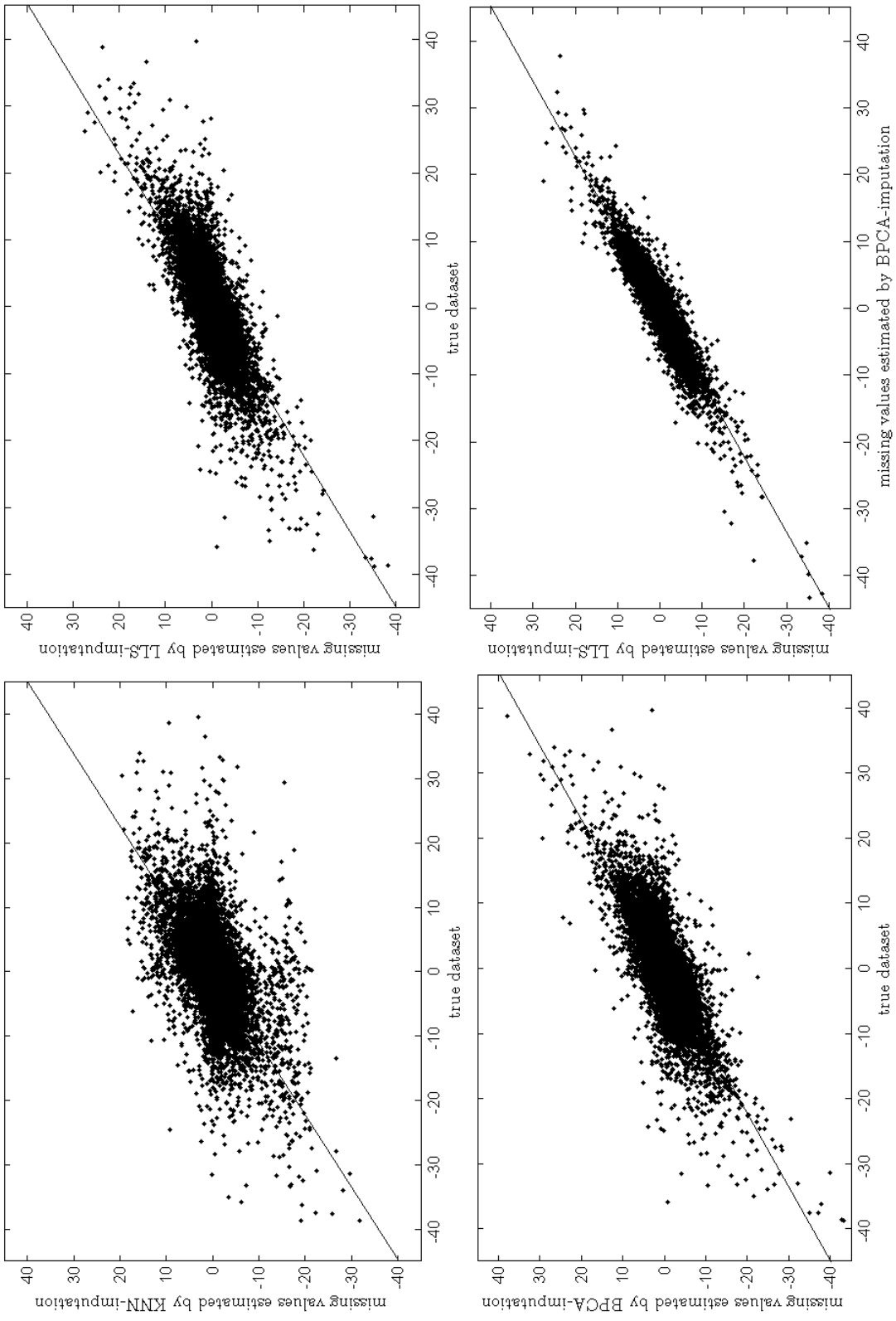


Figure 5.5: Comparison of the estimated values by the KNN-, LLS- and BPCA-imputation methods to the true values of the second transcriptome dataset with $c = 20\%$ of missing entries.

5.3.3 Behaviour of our variable selection procedure with and without a preprocessing imputation method

Our new variable selection procedure which avoids imputing missing values is compared to the variable selection procedure applied on completed datasets. The simulated dataset described in Section 5.2.4.2 is considered. In a preprocessing step, the missing entries of this dataset are estimated by one of the six studied imputation methods and then, the variable selection procedure without missing values is performed. The results of these different methods, for different percentages of missing values are summarized in Table 5.5. For each strategy, the selected model and the clustering error rate are given, the error rate being computed by assigning each cluster to one class according to the majority rule. In the case where the true model is selected, the clustering error rate for the totally observed individuals is given between brackets in the last column.

First, recall that our procedure taking the missing value into account selects the true model in all the cases. The variable selection procedure with a ZERO-, ROW- or COL-imputation in preprocessing does not select the true model in all the cases. With the KNN-imputation method, the procedure finds the true model only for $c = 1\%$ of missing values. With LLS- or BPCA-imputation method, the variable selection procedure has a better behaviour since it selects the true model, except for $c = 20\%$ when the LLS-imputation is applied. The clustering error rate is better than with our procedure taking the missing values into account. But this difference is due to the misclassified individuals with at least one missing value on the relevant clustering variables, as explained in Section 5.2.4.2, since the error rates are similar for the totally observed individuals.

5.4 Discussion

In this chapter, we were interested in the adaptability of our variable selection procedure (see Chapter 3) to study datasets with missing at random values. The extension of our procedure requires a new strategy to calculate the model selection criterion via the explicit expression of the observed loglikelihood and a new parameter estimation method. These different changes involve making alterations to the backward stepwise algorithm. The first results obtained by the use of our new variable selection procedure show a good behaviour for the variable selection but improvements are required for the final clustering rule. A new method seems to be needed to cluster the individuals having at least one missing value in the relevant clustering variables. Moreover, in order to take advantage of the versatility of Gaussian mixture forms, a long programming work remains to carry out for the generalization of this new procedure for the whole mixture forms. Note that an analogous extension for analyzing datasets with missing values can easily be obtained for the new modelling of the variable partition addressed in Chapter 4. Indeed, the major difficulties encountered to extend the procedure in this chapter are due to the regression part. Moreover, the parameters of the additional Gaussian density in the modelling of Chapter 4 can be estimated by an usual EM-algorithm and this Gaussian density can be

percentage of missing values	imputation method	$\hat{K}$	$\hat{S}$	$\hat{R}$	clustering error rate	
$c = 0\%$	-	**4**	**1,2,3**	**1,2**	1.15%	
$c = 1\%$	-	**4**	**1,2,3**	**1,2**	1.55%	[1.1%]
	ZERO	6	1,2,3,4,5,6,7	$\emptyset$	2.00%	
	ROW	5	1,2,3,4,5	$\emptyset$	2.10%	
	COL	6	1,2,3,4,5	$\emptyset$	1.70%	
	KNN	**4**	**1,2,3**	**1,2**	**1.45**%	[1.1%]
	LLS	**4**	**1,2,3**	**1,2**	**1.45**%	[1.1%]
	BPCA	**4**	**1,2,3**	**1,2**	1.55%	[1.1%]
$c = 5\%$	-	**4**	**1,2,3**	**1,2**	3.85%	[0.95%]
	ZERO	6	1,2,3,4,5,6,7	$\emptyset$	6.55%	
	ROW	6	1,2,3,4,5,6,7	$\emptyset$	44.45%	
	COL	6	1,2,3,4,5	$\emptyset$	6.35%	
	KNN	6	1,2,3,4,5	$\emptyset$	6.65%	
	LLS	**4**	**1,2,3**	**1,2**	**3.35**%	[0.9%]
	BPCA	**4**	**1,2,3**	**1,2**	3.60%	[0.95%]
$c = 10\%$	-	**4**	**1,2,3**	**1,2**	7.17%	[0.85%]
	ZERO	6	1,2,3,4,5,6,7	$\emptyset$	44.90%	
	ROW	6	1,2,4,5	4,5	45.25%	
	COL	6	1,2,3,4,5,7	$\emptyset$	12.45%	
	KNN	5	1,2,3,4,5,6	$\emptyset$	6.35%	
	LLS	**4**	**1,2,3**	**1,2**	**6.45**%	[0.8%]
	BPCA	**4**	**1,2,3**	**1,2**	6.85%	[0.85%]
$c = 15\%$	-	**4**	**1,2,3**	**1,2**	10.17%	[0.35%]
	ZERO	6	1,2,4,6	1,2,4	45.80%	
	ROW	5	1,2,3,4,5	1,3,4	45.65%	
	COL	6	1,2,3,4,5,6	$\emptyset$	11.40%	
	KNN	6	1,2,3,4,5	$\emptyset$	9.35%	
	LLS	**4**	**1,2,3**	**1,2**	**8.90**%	[0.6%]
	BPCA	**4**	**1,2,3**	**1,2**	9.15%	[0.6%]
$c = 20\%$	-	**4**	**1,2,3**	**1,2**	13.27%	[0.6%]
	ZERO	6	1,2,4,5	1,4	45.60%	
	ROW	6	1,2,3,4,5,6,7	$\emptyset$	45.85%	
	COL	6	1,2,3,4,5,6,7	$\emptyset$	45.05%	
	KNN	6	1,2,3,4,5	$\emptyset$	12.40%	
	LLS	5	1,2,3,4,5	4,5	12.45%	
	BPCA	**4**	**1,2,3**	**1,2**	**11.90**%	[0.7%]

Table 5.5: Results given by our variable selection procedure according to the percentage of missing values and a preprocessing imputation method. The "-" in the second column indicates the use of the variable selection procedure taking the missing values into account. The percentage into brackets in the last column is the clustering error rate calculated only for the totally observed individuals. When the true model is selected, the solution is in bold.

included into a global Gaussian mixture for the observed likelihood calculation.

The new procedure taking missing values into account is compared to the variable selection method on datasets, completed by an imputation method in preprocessing. This estimation of the missing entries are carried out by one of the six imputation methods tested in this chapter. To study data with missing at random values, we advise to test the LLS- and BPCA-imputation methods on the totally observed individual subset beforehand, marking at random some entries as missing. In particular, for the analysis of a gene subset, the imputation step has to be carried out on the whole genes of the DNA microarray and then the dataset is restricted to the interested gene subset. If one of these two imputation methods gives suitable results, we recommend to apply the variable selection procedure with a preprocessing imputation step because of a computational time gain for similar results. Otherwise, the variable selection procedure taking missing values into account has to be used.

Appendices

5.A EM algorithm for the Gaussian mixture form $[p_k LC]$

In this section, the EM algorithm to estimate the parameters of a Gaussian mixture with the $[p_k LC]$ form is described. We consider a sample $\mathbf{x} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$, where for all $i \in \{1, \ldots, n\}$, $\mathbf{x}_i' = (x_{i1}, \ldots, x_{id})' \in \mathbb{R}^d$ which is modelled by a Gaussian mixture with a general form $m = [p_k LC]$ and K clusters. This sample is decomposed into $(\mathbf{x}^{\mathbf{o}}, \mathbf{x}^{\mathbf{m}})$ where $\mathbf{x}^{\mathbf{o}}$ is the observed values and $\mathbf{x}^{\mathbf{m}}$ is the missing values. Each individual i is assumed to have at least an observed value and each variable j is observed for at least an individual. The aim is to determine the parameter $\hat{\alpha}$ maximizing the observed likelihood $f_{\text{clust}}(\mathbf{x}^{\mathbf{o}}|K, m, \alpha)$ using an Expectation-Maximization algorithm. This algorithm is based on the conditional expectation of the complete likelihood. In this context, the missing data are composed of the label $\mathbf{z}$ and the missing values $\mathbf{x}^{\mathbf{m}}$, and the complete likelihood is

$$f_{\text{clust}}(\mathbf{x}^{\mathbf{o}}, \mathbf{x}^{\mathbf{m}}, \mathbf{z}|K, m, \alpha) = \prod_{i=1}^{n} \prod_{k=1}^{K} p_k^{z_{ik}} \Phi(\mathbf{x}_i^{\mathbf{o}}, \mathbf{x}_i^{\mathbf{m}}|\mu_k, \Sigma)^{z_{ik}}.$$

Given an initial parameter vector value $\alpha^{(1)}$, the principle of the EM algorithm is to alternate an E-Step and a M-Step until it converges. These two steps are now made explicit at the r^{th} iteration.

E-Step: This Expectation step consists of calculating the conditional expectation of the complete loglikelihood given the observed data and a current parameter value $\alpha^{(r)}$:

$$\begin{aligned}
\mathcal{Q}(\alpha|\alpha^{(r)}) &= \mathbb{E}\left[\ln f_{\text{clust}}(\mathbf{x}^{\mathbf{o}}, \mathbf{x}^{\mathbf{m}}, z|K, m, \alpha)|\mathbf{x}^{\mathbf{o}}, \alpha^{(r)}\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}[z_{ik}|\mathbf{x}^{\mathbf{o}}, \alpha^{(r)}]\ln(p_k) - \frac{nd}{2}\ln(2\pi) - \frac{n}{2}\ln(|\Sigma|) \\
&- \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}\left[\frac{z_{ik}}{2}(\mathbf{x}_i - \mu_k)\Sigma^{-1}(\mathbf{x}_i - \mu_k)'|\mathbf{x}^{\mathbf{o}}, \alpha^{(r)}\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}[z_{ik}|\mathbf{x}^{\mathbf{o}}, \alpha^{(r)}]\ln(p_k) - \frac{nd}{2}\ln(2\pi) - \frac{n}{2}\ln(|\Sigma|) \\
&- \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j,l=1}^{d}(\Sigma^{-1})_{jl}\left\{\mathbb{E}[z_{ik}x_{ij}x_{il}|\mathbf{x}^{\mathbf{o}}, \alpha^{(r)}] + \mathbb{E}[z_{ik}|\mathbf{x}^{\mathbf{o}}, \alpha^{(r)}]\mu_{kj}\mu_{kl}\right\} \\
&+ \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j,l=1}^{d}(\Sigma^{-1})_{jl}\left\{\mu_{kj}\mathbb{E}[z_{ik}x_{il}|\mathbf{x}^{\mathbf{o}}, \alpha^{(r)}] + \mu_{kl}\mathbb{E}[z_{ik}x_{ij}|\mathbf{x}^{\mathbf{o}}, \alpha^{(r)}]\right\}. \quad (5.8)
\end{aligned}$$

Thus according to Equation (5.8), this step is equivalent to calculating for all $i \in \{1, \ldots, n\}$, $k \in \{1, \ldots, K\}$ and $j, l \in \{1, \ldots, d\}$, the conditional expectations $\mathbb{E}[z_{ik}|\mathbf{x}^{\mathbf{o}}, \alpha^{(r)}]$, $\mathbb{E}[z_{ik}x_{ij}|\mathbf{x}^{\mathbf{o}}, \alpha^{(r)}]$ and $\mathbb{E}[z_{ik}x_{ij}x_{il}|\mathbf{x}^{\mathbf{o}}, \alpha^{(r)}]$. In order to make explicit these conditional expectations, we introduce the following decomposition of parameters μ_k and Σ according to the positions of the observed and the missing values of $\mathbf{x}_i$:

$$\mu_k^{(i)} = (\mu_{k,\mathbf{o}}^{(i)}, \mu_{k,\mathbf{m}}^{(i)}) \text{ and } \Sigma^{(i)} = \begin{pmatrix} \Sigma_{\mathbf{oo}}^{(i)} & \Sigma_{\mathbf{om}}^{(i)} \\ \Sigma_{\mathbf{mo}}^{(i)} & \Sigma_{\mathbf{mm}}^{(i)} \end{pmatrix}.$$

Since $\mathbf{x}_i$ has a normal distribution $\mathcal{N}(\mu_k, \Sigma)$ conditionally to $z_{ik} = 1$, $\mathbf{x}_i^{\mathbf{o}}$ given $z_{ik} = 1$ is distributed from a normal density $\mathcal{N}\left(\mu_{k,\mathbf{o}}^{(i)}, \Sigma_{\mathbf{oo}}^{(i)}\right)$ and the conditional distribution of $\mathbf{x}_i^{\mathbf{m}}$ given $\mathbf{x}_i^{\mathbf{o}}$ and $z_{ik} = 1$ is $\mathcal{N}\left(\mu_{k,\mathbf{m}|\mathbf{o}}^{(i)} + \mathbf{x}_i^{\mathbf{o}}\Sigma_{\mathbf{m}|\mathbf{o}}^{(i)}, \Sigma_{\mathbf{mm}|\mathbf{o}}^{(i)}\right)$ where $\mu_{k,\mathbf{m}|\mathbf{o}}^{(i)} = \mu_{k,\mathbf{m}}^{(i)} - \mu_{k,\mathbf{o}}^{(i)}\Sigma_{\mathbf{m}|\mathbf{o}}^{(i)}$, $\Sigma_{\mathbf{m}|\mathbf{o}}^{(i)} = (\Sigma_{\mathbf{oo}}^{(i)})^{-1}\Sigma_{\mathbf{om}}^{(i)}$ and $\Sigma_{\mathbf{mm}|\mathbf{o}}^{(i)} = \Sigma_{\mathbf{mm}}^{(i)} - \Sigma_{\mathbf{mo}}^{(i)}(\Sigma_{\mathbf{oo}}^{(i)})^{-1}\Sigma_{\mathbf{om}}^{(i)}$ according to Proposition 5.C.1.

First, the conditional probability denoted $t_{ik}^{(r)}$ that $\mathbf{x}_i$ arises from the component k is given by

$$t_{ik}^{(r)} = \mathbb{E}[z_{ik}|\mathbf{x}^{\mathbf{o}}, \alpha^{(r)}] = \frac{p_k^{(r)}\Phi(\mathbf{x}_i^{\mathbf{o}}|\mu_{k,\mathbf{o}}^{(r,\,i)}, \Sigma_{\mathbf{oo}}^{(r,\,i)})}{\sum_{v=1}^{K} p_v^{(r)}\Phi(\mathbf{x}_i^{\mathbf{o}}|\mu_{v,\mathbf{o}}^{(r,\,i)}, \Sigma_{\mathbf{oo}}^{(r,\,i)})}. \quad (5.9)$$

The second-type expectations are formulated as

$$\mathbb{E}[z_{ik}x_{ij}|\mathbf{x}^{\mathbf{o}}, \alpha^{(r)}] = \mathbb{E}[z_{ik}|\mathbf{x}_i^{\mathbf{o}}, \alpha^{(r)}]\mathbb{E}[x_{ij}|\mathbf{x}_i^{\mathbf{o}}, \alpha^{(r)}, z_{ik} = 1] = t_{ik}^{(r)}\tilde{x}_{ij}^{(k,\,r)} \quad (5.10)$$

where $\tilde{x}_{ij}^{(k,\,r)} = \mathbb{E}[x_{ij}|\mathbf{x}_i^{\mathbf{o}}, \alpha^{(r)}, z_{ik} = 1]$ is equal to x_{ij} if x_{ij} is observed and otherwise, corresponds to the coordinate j of the vector $\mu_{k,\mathbf{m}|\mathbf{o}}^{(r,\,i)} + \mathbf{x}_i^{\mathbf{o}}\Sigma_{\mathbf{mm}|\mathbf{o}}^{(r,\,i)}$.

The third-type expectations are calculated using the conditional covariances:

$$\begin{aligned}\mathbb{E}[z_{ik}x_{ij}x_{il}|\mathbf{x}^{\mathbf{o}},\alpha^{(r)}] &= \mathbb{E}[z_{ik}|\mathbf{x}_i^{\mathbf{o}},\alpha^{(r)}]\mathbb{E}[x_{ij}x_{il}|\mathbf{x}_i^{\mathbf{o}},\alpha^{(r)},z_{ik}=1]\\ &= t_{ik}^{(r)}\mathbb{E}[x_{ij}|\mathbf{x}_i^{\mathbf{o}},\alpha^{(r)},z_{ik}=1]\mathbb{E}[x_{il}|\mathbf{x}_i^{\mathbf{o}},\alpha^{(r)},z_{ik}=1]\\ &\quad +t_{ik}^{(r)}\operatorname{cov}(x_{ij},x_{il}|\mathbf{x}_i^{\mathbf{o}},\alpha^{(r)},z_{ik}=1)\\ &= t_{ik}^{(r)}\left\{\tilde{x}_{ij}^{(k,\,r)}\tilde{x}_{il}^{(k,\,r)}+C_{i,jl}^{(r)}\right\} \qquad (5.11)\end{aligned}$$

where $C_{i,jl}^{(r)}=\operatorname{cov}(x_{ij},x_{il}|\mathbf{x}_i^{\mathbf{o}},\alpha^{(r)},z_{ik}=1)$. If x_{ij} and/or x_{il} are observed then $C_{jl}^{(i,r)}=0$. Otherwise, $C_{i,jl}^{(r)}$ is the (j,l) term of the matrix $\Sigma_{\mathbf{mm|o}}^{(r,\,i)}$.

M-Step: This step consists of determining the parameter vector $\alpha^{(r+1)}$ maximizing the function $\mathcal{Q}(\alpha|\alpha^{(r)})$. Replacing the conditional expectations by their expressions (5.9), (5.10) and (5.11) respectively in Equation (5.8) and defining $n_k^{(r)}=\sum_{i=1}^{n}t_{ik}^{(r)}$, it leads to

$$\begin{aligned}\mathcal{Q}(\alpha|\alpha^{(r)}) &= \sum_{k=1}^{K}n_k^{(r)}\ln(p_k)-\frac{nd}{2}\ln(2\pi)-\frac{n}{2}\ln(|\Sigma|)\\ &\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}t_{ik}^{(r)}\left\{(\tilde{\mathbf{x}}_i^{(k,\,r)}-\mu_k)\Sigma^{-1}(\tilde{\mathbf{x}}_i^{(k,\,r)}-\mu_k)'+\sum_{j,l=1}^{d}(\Sigma^{-1})_{jl}C_{i,jl}^{(r)}\right\}.\end{aligned}$$

First the proportion vector maximizing the function $(p_1,\ldots,p_K)\mapsto\sum_{k=1}^{K}n_k^{(r)}\ln(p_k)$ such that $\sum_{k=1}^{K}p_k=1$ is determined using a Lagrange multiplicator. We obtain easily that for all $k\in\{1,\ldots,K\}$, $p_k^{(r+1)}=\frac{1}{n}\sum_{i=1}^{n}t_{ik}^{(r)}=n_k^{(r)}/n$.

Second, we search the Gaussian density parameters maximizing the function

$$(\mu_1,\ldots,\mu_K,\Sigma)\mapsto -n\ln(|\Sigma|)-\sum_{k=1}^{K}\sum_{i=1}^{n}t_{ik}^{(r)}\left\{(\tilde{\mathbf{x}}_i^{(k,\,r)}-\mu_k)\Sigma^{-1}(\tilde{\mathbf{x}}_i^{(k,\,r)}-\mu_k)'+\sum_{j,l=1}^{d}(\Sigma^{-1})_{jl}C_{i,jl}^{(r)}\right\}.$$

Noting that

$$\sum_{k=1}^{K}\sum_{i=1}^{n}t_{ik}^{(r)}(\tilde{\mathbf{x}}_i^{(k,\,r)}-\mu_k)\Sigma^{-1}(\tilde{\mathbf{x}}_i^{(k,\,r)}-\mu_k)'=\operatorname{tr}\left(\Sigma^{-1}\left\{\sum_{i=1}^{n}\sum_{k=1}^{K}t_{ik}^{(r)}(\tilde{\mathbf{x}}_i^{(k,\,r)}-\mu_k)'(\tilde{\mathbf{x}}_i^{(k,\,r)}-\mu_k)\right\}\right)$$

and that

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j,l=1}^{d}t_{ik}^{(r)}\,(\Sigma^{-1})_{jl}\,C_{i,jl}^{(r)}=\operatorname{tr}\left(\Sigma^{-1}\sum_{i=1}^{n}C_i^{(r)}\right),$$

it is equivalent to maximizing

$$(\mu_1\ldots,\mu_K,\Sigma)\mapsto -n\ln(|\Sigma|)-\operatorname{tr}\left(\Sigma^{-1}\sum_{i=1}^{n}C_i^{(r)}\right)-\operatorname{tr}\left(\Sigma^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}t_{ik}^{(r)}(\tilde{\mathbf{x}}_i^{(k,\,r)}-\mu_k)'(\tilde{\mathbf{x}}_i^{(k,\,r)}-\mu_k)\right).$$

Denoting $A^{(r)} = \sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}^{(r)}(\tilde{\mathbf{x}}_i^{(k,r)} - \bar{\mu}_k^{(r)})'(\tilde{\mathbf{x}}_i^{(k,r)} - \bar{\mu}_k^{(r)})$ where $\bar{\mu}_k^{(r)} = \frac{1}{n_k^{(r)}} \sum_{i=1}^{n} t_{ik}^{(r)}\tilde{\mathbf{x}}_i^{(k,r)}$ and $B^{(r)} = \sum_{i=1}^{n} C_i^{(r)}$, the expression in the second trace term is decomposed into

$$\sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}^{(r)}(\tilde{\mathbf{x}}_i^{(k,r)} - \mu_k)'(\tilde{\mathbf{x}}_i^{(k,r)} - \mu_k) = A^{(r)} + \sum_{k=1}^{K} n_k^{(r)}(\bar{\mu}_k^{(r)} - \mu_k)'(\bar{\mu}_k^{(r)} - \mu_k)$$

and thus, we want to maximize the function

$$(\mu_1, \ldots, \mu_K, \Sigma) \mapsto -n\ln(|\Sigma|) - \mathrm{tr}\left\{\Sigma^{-1}(B^{(r)} + A^{(r)})\right\} - \mathrm{tr}\left\{\Sigma^{-1}\sum_{k=1}^{K} n_k^{(r)}(\bar{\mu}_k^{(r)} - \mu_k)'(\bar{\mu}_k^{(r)} - \mu_k)\right\}.$$

Since Σ is a positive definite matrix,

$$\mathrm{tr}\left\{\Sigma^{-1}\sum_{k=1}^{K} n_k^{(r)}(\bar{\mu}_k^{(r)} - \mu_k)'(\bar{\mu}_k^{(r)} - \mu_k)\right\} = \sum_{k=1}^{K} n_k^{(r)}(\bar{\mu}_k^{(r)} - \mu_k)\Sigma^{-1}(\bar{\mu}_k^{(r)} - \mu_k)'$$

is a non negative quantity and is equal to zero only when $\mu_k = \bar{\mu}_k^{(r)}$. Thus, we obtain

$$\mu_k^{(r+1)} = \frac{1}{n_k^{(r)}} \sum_{i=1}^{n} t_{ik}^{(r)}\tilde{\mathbf{x}}_i^{(k,r)}.$$

Then, we search the positive definite matrix maximizing the function

$$\Sigma \mapsto -n\ln(|\Sigma|) - \mathrm{tr}\left(\Sigma^{-1}\tilde{W}^{(r+1)}\right)$$

where $\tilde{W}^{(r+1)} = A^{(r)} + B^{(r)}$. If we prove that this matrix $\tilde{W}^{(r+1)}$ is positive definite almost surely then, according to Lemma 5.C.2, we get

$$\Sigma^{(r+1)} = \frac{1}{n}\left\{\sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}^{(r)}(\tilde{\mathbf{x}}_i^{(k,r)} - \mu_k^{(r+1)})'(\tilde{\mathbf{x}}_i^{(k,r)} - \mu_k^{(r+1)}) + \sum_{i=1}^{n} C_i^{(r)}\right\}.$$

It remains to prove that $\tilde{W}^{(r+1)}$ is a positive definite matrix almost surely. According to the definitions of $C_i^{(r)}$ and $t_{ik}^{(r)}$, the matrices $A^{(r)}$, $B^{(r)}$ and thus $\tilde{W}^{(r+1)}$ are positive matrices. The subset of variables for which the whole individuals are observed is denoted J_1 and J_2 is its complement. We are interested in the following probability

$$\begin{aligned} P(\forall u \neq 0;\ u\tilde{W}^{(r+1)}u' \neq 0) &= 1 - P(\exists u \neq 0;\ u\tilde{W}^{(r+1)}u' = 0) \\ &\geq 1 - \sum_{u\neq 0, u_{J_2}=0} P(uA^{(r)}u' = 0 \ \cap\ uB^{(r)}u' = 0) \\ &\quad - \sum_{u\neq 0, u_{J_2}\neq 0} P(uA^{(r)}u' = 0 \ \cap\ uB^{(r)}u' = 0). \end{aligned}$$

If $u_{J_2} \neq 0$, there exists a variable $j \in J_2$ such that $u_j \neq 0$ and an individual i such that x_{ij} is non observed. Then, according to the definition of the conditional variance matrix $C_i^{(r)}$, $uC_i^{(r)}u' = u_{J_2}C_{J_2J_2}^{(i,r)}u'_{J_2} > 0$. Otherwise, $u_{J_1} \neq 0$ since u is a nonzero vector. But

$$\begin{aligned} u\tilde{W}^{(r+1)}u' &= u_{J_1}A^{(r)}u'_{J_1} \\ &= \sum_{k=1}^{K} u_{J_1}\left\{\sum_{i=1}^{n} t_{ik}^{(r)}(\tilde{\mathbf{x}}_i^{(k,r)} - \mu_k^{(r+1)})'_{J_1}(\tilde{\mathbf{x}}_i^{(k,r)} - \mu_k^{(r+1)})_{J_1}\right\}u'_{J_1} \\ &= \sum_{k=1}^{K} u_{J_1}D_k^{(r)}u'_{J_1} \end{aligned}$$

where $D_k^{(r)} = \sum_{i=1}^{n} t_{ik}^{(r)}(\mathbf{x}_i - \mu_k^{(r+1)})'_{J_1}(\mathbf{x}_i - \mu_k^{(r+1)})_{J_1}$. It leads to

$$\begin{aligned} P\left(\sum_{k=1}^{K} u_{J_1}D_k^{(r)}u'_{J_1} = 0\right) &= \sum_{\mathbf{z}_1,\ldots,\mathbf{z}_n} P\left(\sum_{k=1}^{K} u_{J_1}D_k^{(r)}u'_{J_1} = 0\middle|\mathbf{z}_1,\ldots,\mathbf{z}_n\right)P(\mathbf{z}_1,\ldots,\mathbf{z}_n) \\ &= \sum_{\mathbf{z}_1,\ldots,\mathbf{z}_n} P\left(\forall 1 \leq k \leq K,\ u_{J_1}D_k^{(r)}u'_{J_1} = 0\middle|\mathbf{z}_1,\ldots,\mathbf{z}_n\right)P(\mathbf{z}_1,\ldots,\mathbf{z}_n) \\ &= \sum_{\mathbf{z}_1,\ldots,\mathbf{z}_n} P\left(\forall 1 \leq k \leq K,\ u_{J_1}F_k^{(r)}u'_{J_1} = 0 \cap u_{J_1}G_k^{(r)}u'_{J_1} = 0|\mathbf{z}_1,\ldots,\mathbf{z}_n\right) \\ &\qquad \times P(\mathbf{z}_1,\ldots,\mathbf{z}_n) \end{aligned}$$

with

$$F_k^{(r)} = \sum_{i;z_{ik}=1} t_{ik}^{(r)}(\mathbf{x}_i - \mu_k^{(r+1)})'_{J_1}(\mathbf{x}_i - \mu_k^{(r+1)})_{J_1}$$

and

$$G_k^{(r)} = \sum_{i;z_{ik}=0} t_{ik}^{(r)}(\mathbf{x}_i - \mu_k^{(r+1)})'_{J_1}(\mathbf{x}_i - \mu_k^{(r+1)})_{J_1}.$$

There exists $\tilde{k} \in \{1,\ldots,K\}$ such that $\sum_{i;z_{i\tilde{k}}=1} t_{i\tilde{k}}^{(r)} \neq 0$ and, according to Proposition 5.C.3, the matrix $F_{\tilde{k}}^{(r)}$ is positive definite almost surely since conditionally to $z_{i\tilde{k}} = 1$, $\mathbf{x}_i$ restricted to J_1 is sampled from $\mathcal{N}(\mu_{\tilde{k},J_1}, \Sigma_{J_1J_1})$. Consequently, $P\left(\sum_{k=1}^{K} u_{J_1}D_k^{(r)}u'_{J_1} = 0\right) = 0$ and then $\tilde{W}^{(r+1)}$ is a positive definite matrix almost surely.

5.B EM algorithm for multidimensional multivariate regression

This section is devoted to determine the maximum likelihood estimators of the following multidimensional multivariate linear regression

$$\forall i \in \{1,\ldots,n\},\ H_i = X_iB + E_i,\ E_i \sim \mathcal{N}(0,\Omega)$$

where the predictor matrix X is assumed to be totally observed and the $n \times V$ response matrix H can have missing values. The matrix H is decomposed into its observed values and missing values $H = (H^{\circ}, H^{\mathbf{m}})$. The parameter estimator $(\hat{B}, \hat{\Omega})$ are obtained using an iterative EM algorithm based on a conditional expectation of the complete loglikelihood given by

$$L(B, \Omega | H, X) = -\frac{nV}{2} \ln(2\pi) - \frac{n}{2} \ln(|\Omega|) - \frac{1}{2} \sum_{i=1}^{n} (H_i - X_i B)\Omega^{-1}(H_i - X_i B)'.$$

Given an initial parameter vector value $(B^{(1)}, \Omega^{(1)})$, the two alternate steps of the EM algorithm are now made explicit at the r^{th} iteration.

E-Step: This step consists of calculating the expectation of the complete loglikelihood conditionally to the observed values H° and X, and current parameter values $B^{(r)}$ and $\Omega^{(r)}$, $\mathcal{Q}(B, \Omega | B^{(r)}, \Omega^{(r)}) := \mathbb{E}[L(B, \Omega | H, X) | H^{\circ}, X, B^{(r)}, \Omega^{(r)}]$. It is equivalent to evaluating the two conditional expectations $\mathbb{E}\left[\sum_{i=1}^{n} H_{ij}(X_i B)_l \middle| H^{\circ}, X, B^{(r)}, \Omega^{(r)}\right]$ and $\mathbb{E}\left[\sum_{i=1}^{n} H_{ij}H_{il} \middle| H^{\circ}, X, B^{(r)}, \Omega^{(r)}\right]$ for all $j, l \in \{1, \ldots, V\}$. The first-type expectations are given by

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{n} H_{ij}(X_i B)_l \middle| H^{\circ}, X, B^{(r)}, \Omega^{(r)}\right] &= \sum_{i=1}^{n} \mathbb{E}\left[H_{ij} | H_i^{\circ}, X, B^{(r)}, \Omega^{(r)}\right] (X_i B)_l \\ &= \sum_{i=1}^{n} \tilde{H}_{i,j}^{(r)} (X_i B)_l \end{aligned}$$

where $\tilde{H}_{i,j}^{(r)} := \mathbb{E}\left[H_{ij} | H_i^{\circ}, X, B^{(r)}, \Omega^{(r)}\right]$ is equal to H_{ij} if H_{ij} is observed and $(X_i B^{(r)})_j$ otherwise. The second-type expectations are evaluated using the conditional covariance matrix as follows

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{n} H_{ij}H_{il} \middle| H^{\circ}, X, B^{(r)}, \Omega^{(r)}\right] &= \sum_{i=1}^{n} \operatorname{cov}(H_{ij}, H_{il} | H_i^{\circ}, X, B^{(r)}, \Omega^{(r)}) \\ &+ \sum_{i=1}^{n} \mathbb{E}\left[H_{ij} | H_i^{\circ}, X, B^{(r)}, \Omega^{(r)}\right] \mathbb{E}\left[H_{il} | H_i^{\circ}, X, B^{(r)}, \Omega^{(r)}\right] \\ &= \sum_{i=1}^{n} \tilde{H}_{i,j}^{(r)} \tilde{H}_{i,l}^{(r)} + C_{i,jl}^{(r)} \end{aligned}$$

where $C_{i,jl}^{(r)} = \operatorname{cov}(H_{ij}, H_{il} | H_i^{\circ}, X, B^{(r)}, \Omega^{(r)})$ is the coefficient of $\Omega^{(r)}$ associated to the couple (j, l) if H_{ij} and H_{il} are missing and is null otherwise. Finally, the conditional expectation of the complete likelihood is given by

$$\begin{aligned} \mathcal{Q}(B, \Omega | B^{(r)}, \Omega^{(r)}) &= -\frac{nV}{2} \ln(2\pi) - \frac{n}{2} \ln(|\Omega|) \\ &\quad -\frac{1}{2} \operatorname{tr}\left[\Omega^{-1} \left\{ \sum_{i=1}^{n} (\tilde{H}_i^{(r)} - X_i B)'(\tilde{H}_i^{(r)} - X_i B) + C_i^{(r)} \right\}\right]. \quad (5.12) \end{aligned}$$

M-Step: This step consists of determining the parameter vector $(B^{(r+1)}, \Omega^{(r+1)})$ maximizing $\mathcal{Q}(B, \Omega | B^{(r)}, \Omega^{(r)})$. According to (5.12), it is equivalent to maximizing

$$(B, \Omega) \mapsto -n \ln(|\Omega|) - \operatorname{tr}\left[\Omega^{-1}\left\{\sum_{i=1}^{n} (\tilde{H}_i^{(r)} - X_i B)'(\tilde{H}_i^{(r)} - X_i B) + C_i^{(r)}\right\}\right].$$

Denoting $T^{(r)} = \left(\sum_{i=1}^{n} X_i' X_i\right)^{-1} \sum_{i=1}^{n} X_i' \tilde{H}_i^{(r)}$, it leads to

$$\begin{aligned}\sum_{i=1}^{n} (\tilde{H}_i^{(r)} - X_i B)'(\tilde{H}_i^{(r)} - X_i B) + C_i^{(r)} &= \sum_{i=1}^{n} (\tilde{H}_i^{(r)} - X_i T^{(r)})'(\tilde{H}_i^{(r)} - X_i T^{(r)}) + C_i^{(r)} \\ &\quad + (T^{(r)} - B)' \left(\sum_{i=1}^{n} X_i' X_i\right) (T^{(r)} - B).\end{aligned}$$

Using the property that given two positive definite matrices F and G, $\operatorname{tr}(S'FSG) > 0$ for all $S \neq 0$ (see for instance Anderson, 2003, Lemma 8.2.2), the following trace

$$\operatorname{tr}\left(\left\{\Omega^{-1}(T^{(r)} - B)' \left(\sum_{i=1}^{n} X_i' X_i\right) (T^{(r)} - B)\right\}\right)$$

is non negative and is equal to zero only if $B = T^{(r)}$. Thus

$$B^{(r+1)} = \left(\sum_{i=1}^{n} X_i' X_i\right)^{-1} \sum_{i=1}^{n} X_i' \tilde{H}_i^{(r)} = (X'X)^{-1} X' \tilde{H}^{(r)}.$$

Secondly, we maximize the function $\Omega \mapsto -n \ln(|\Omega|) - \operatorname{tr}\left(\Omega^{-1} W^{(r+1)}\right)$ where

$$W^{(r+1)} = \sum_{i=1}^{n} (\tilde{H}_i^{(r)} - X_i B^{(r+1)})'(\tilde{H}_i^{(r)} - X_i B^{(r+1)}) + C_i^{(r)}.$$

According to Lemma 5.C.2, if $W^{(r+1)}$ is a positive definite matrix then

$$\begin{aligned}\Omega^{(r+1)} &= \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{H}_i^{(r)} - X_i B^{(r+1)}\right)' \left(\tilde{H}_i^{(r)} - X_i B^{(r+1)}\right) + \frac{1}{n} \sum_{i=1}^{n} C_i^{(r)} \\ &= \frac{1}{n} \tilde{H}^{(r)\prime} \left\{I - X(X'X)^{-1} X'\right\} \tilde{H}^{(r)} + \frac{1}{n} \sum_{i=1}^{n} C_i^{(r)}.\end{aligned}$$

It remains to prove by contradiction that $W^{(r+1)}$ is a positive definite matrix. First, $W^{(r+1)}$ is a positive matrix as the sum of positive matrices. The proof is analogous to this of Appendix 5.A, the distinction between a vector $u \neq 0$ such that $u_{J_2} \neq 0$ or not and the Corollary 5.C.4 provide the result: If $u_{J_2} \neq 0$, there exists $j \in J_2$ and $i \in \{1, \ldots, n\}$ such that H_{ij} is non observed and then, $u C_i^{(r)} u' \neq 0$. Otherwise, the vector $u_{J_1} \neq 0$ hence

$$u W^{(r+1)} u' = u_{J_1} \left\{\sum_{i=1}^{n} (H_i - X_i B^{(r+1)})'_{J_1} (H_i - X_i B^{(r+1)})_{J_1}\right\} u'_{J_1} > 0$$

according to Corollary 5.C.4.

5.C Technical results

The following proposition stated for instance in Anderson (2003, Theorem 2.5.1) gives the distribution of a part of a Gaussian vector conditionally to the other part.

Proposition 5.C.1.
If a vector $(\mathbf{x}_1, \mathbf{x}_2)$ *has a normal density with mean vector* $\mu = (\mu_1, \mu_2)$ *and variance matrix* $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ *then the conditional law of* $\mathbf{x}_2$ *given* $\mathbf{x}_1$ *is a normal distribution with mean* $\mu_{2|1} + \mathbf{x}_1\Sigma_{2|1}$ *and variance matrix* $\Sigma_{22|1}$ *where*

$$\begin{cases} \Sigma_{2|1} = \Sigma_{11}^{-1}\Sigma_{12} \\ \mu_{2|1} = \mu_2 - \mu_1\Sigma_{2|1} \\ \Sigma_{22|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \end{cases}$$

The following Lemma (see for instance Anderson, 2003, Lemma 3.2.2) allows to find the positive definite matrix maximizing an usual function for the maximum likelihood estimation of a Gaussian distribution.

Lemma 5.C.2.
If D *is a positive definite matrix of order* p*, the maximum of the function*

$$f(\Sigma) = -N\ln(|\Sigma|) - \mathrm{tr}(\Sigma^{-1}D)$$

with respect to positive definite matrices Σ *exists, occurs at* $\Sigma = \frac{D}{N}$*, and has the value*

$$f\left(\frac{D}{N}\right) = pN\ln(N) - pN\ln(|D|) - pN.$$

Proposition 5.C.3.
Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *are* n *independent row vectors distributed according to* $\mathcal{N}(\mu, \Sigma)$ *and* $p_1, \ldots, p_n$ *are* n *real of* $(0, 1)$ *such that* $\sum_{i=1}^n p_i = 1$*. Let be the vector* $\hat{\mu} = \sum_{i=1}^n p_i\mathbf{x}_i$*. Then the matrix* $A := \sum_{i=1}^n p_i(\mathbf{x}_i - \hat{\mu})'(\mathbf{x}_i - \hat{\mu})$ *is positive definite almost surely.*

Proof. First, we note that the matrix A is a positive matrix. For all $i \in \{1, \ldots, n\}$, let be the vector $\mathbf{y}_i = \mathbf{x}_i - \mu$ distributed from $\mathcal{N}(0, \Sigma)$. Then the interested matrix

$$\begin{aligned} A &= \sum_{i=1}^n p_i(\mathbf{x}_i - \hat{\mu})'(\mathbf{x}_i - \hat{\mu}) \\ &= \sum_{i=1}^n p_i\mathbf{y}_i'\mathbf{y}_i - (\hat{\mu} - \mu)'(\hat{\mu} - \mu). \end{aligned}$$

Denoting $\hat{\nu} = \hat{\mu} - \mu$ and $\forall i, X_i = \sqrt{p_i}\mathbf{y}_i \sim \mathcal{N}(0, p_i\Sigma)$, it leads to $A = \sum_{i=1}^n X_i'X_i - \hat{\nu}'\hat{\nu}$. An orthogonal matrix $B = (b_{ij})$ is considered such that $Z_i := \sum_{\delta=1}^n b_{i\delta}X_\delta$ and $Z_n := \hat{\nu} =$

$\sum_{i=1}^{n} \sqrt{p_i} X_i$. Since B is an orthogonal matrix,

$$\sum_{i=1}^{n} Z_i' Z_i = \sum_{i=1}^{n} \left(\sum_{\delta=1}^{n} b_{i\delta} X_\delta \right)' \left(\sum_{\delta'=1}^{n} b_{i\delta'} X_{\delta'} \right) = \sum_{\delta=1}^{n} X_\delta' X_\delta$$

and then $A = \sum_{i=1}^{n-1} Z_i' Z_i$. The vector Z_i is a Gaussian vector as a linear combination of Gaussian vectors with a null expectation and a variance $\mathrm{Var}(Z_i) = \gamma_i \Sigma$ with $\gamma_i = \sum_{\delta=1}^{n} b_{i\delta}^2 p_\delta > 0$. Defining $F_i = Z_i / \sqrt{\gamma_i} \sim \mathcal{N}(0, \Sigma)$, it leads to $A = \sum_{i=1}^{n-1} \gamma_i F_i' F_i$. Finally, since for all $u \neq 0$, $\gamma_i u F_i' F_i u' \geq 0$,

$$\begin{aligned} P(\forall u \neq 0, uAu' > 0) &= P\left(\forall u \neq 0, \sum_{i=1}^{n-1} \gamma_i u F_i' F_i u' > 0 \right) \\ &= P\left(\forall u \neq 0, \exists i; \gamma_i u F_i' F_i u' > 0 \right) \\ &= 1 - P\left(\exists u \neq 0, \forall i, \gamma_i \frac{u F_i' F_i u'}{u \Sigma u'} = 0 \right). \end{aligned}$$

The $F_i' F_i$ matrices have a Wishart distribution (see Definition 3.4.1 Mardia et al., 1979) and according to Theorem 3.4.2 in Mardia et al. (1979), for all $u \Sigma u' \neq 0$, the ratio $u F_i' F_i u' / u \Sigma u'$ has a χ^2 distribution. Thus $P(\forall u \neq 0, uAu' > 0) = 1$. □

Corollary 5.C.4.
Suppose $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are a set of n observations, $\mathbf{y}_i$ being sampled from $\mathcal{N}(\mathbf{x}_i B, \Omega)$. Let $\hat{B}$ be the estimated regression coefficient matrix. Then the matrix $\sum_{i=1}^{n} \left(\mathbf{y}_i - \mathbf{x}_i \hat{B} \right)' \left(\mathbf{y}_i - \mathbf{x}_i \hat{B} \right)$ is positive definite almost surely.

Chapter 6

Conclusion and perspectives

A variable selection procedure for clustering with Gaussian mixture models has been proposed. It is based on a new variable role modelling for clustering, generalizing the model of Raftery and Dean (2006b) and being more realistic. The model identifiability and the consistency of the model selection criterion are proved, establishing at the same time the ones of Raftery and Dean's modelling, not proved until now. The algorithm, associated to the model selection criterion, consists of embedding two backward stepwise algorithms. It visits the model collection differently from Raftery and Dean's algorithm CLUSTVARSEL. Two extensions of our variable selection procedure are also suggested. The first extension improves the variable roles in variable selection for clustering. It specifies the link between irrelevant variables and relevant variables and allows for a possible independence of some irrelevant variables. This extension allows us to avoid an overpenalization of some models and to improve the data clustering and its interpretation. The second extension consists of taking the missing values of datasets into account in the variable selection procedure. It allows for a more realistic analysis of transcriptome data since it avoids the imputation of missing entries or the restriction to the totally observed gene subset. These two extensions being proposed separately, a global variable selection procedure, including the qualities of these extensions remains to be achieved.

Programming of the different variable selection procedures

Algorithms presented in this part are greedy in computation time. Consequently, they are implemented in C^{++} with the object-oriented programming. According to the description of algorithms given in the three chapters, an implementation on a grid or a cluster is conceivable to analysis larger datasets. Note that, the implementation of the EM algorithm, used to estimate Gaussian missing parameters to cluster data with missing values, is not available in the MIXMOD software. The programming of such an EM algorithm was only realized for the mixture form $[p_k LC]$. The implementation associated to the other mixture forms has to be planned in order to allow for the choice of the mixture form in the second procedure extension.

Possible extensions of the variable selection procedure

Different extensions could be considered for our variable selection procedure.

First, it is possible to change the structure of dependency between the irrelevant variables and the relevant clustering variables. If the linear regression is replaced with an other dependency link for which a BIC-like criterion is available, the procedure presented in Chapter 3 can be adapted straightforwardly. On the other hand, one can have difficulty in adapting one of the two extensions proposed in Chapters 4 and 5, especially for the change of the variable selection algorithm. In every case, the identifiability of the new model collection has to be studied.

Second, the model selection criterion proposed in these chapters is a BIC-type criterion, based on the maximization of the integrated likelihood. We can plan to use a criterion based on the integrated complete likelihood in order to more take the clustering aim as the ICL criterion into account.

It also could be interesting to consider other variable types. For instance, we could envisage adapting our procedure for the clustering of data described by qualitative variables or a mixture of qualitative and quantitative variables. An other interesting point could be to select variables for a semi-supervised clustering process, allowing to take biological knowledge for the transcriptome dataset analysis into account for instance.

Final clustering rule

The maximum a posteriori (MAP) rule used in the different procedures to obtain the final data clustering is a strict rule which seems to be unsuitable for gene clustering. For instance, some of the 1020 genes clustered in Chapter 3 have not a strong conditional probability for any of the 17 clusters. To illustrate this remark, these 1020 genes are grouped according to the greatest value taken by their conditional probabilities in Table 6.1. Only 902 genes are assigned to a cluster with a conditional probability greater than 0.9. Consequently, a clustering rule graduating the belonging of a gene for a cluster and perhaps allowing a gene to belong to several clusters, remains to be defined.

max c.p.	$]0.9, 1]$	$]0.8, 0.9]$	$]0.7, 0.8]$	$]0.6, 0.7]$	$]0.5, 0.6]$	$[0, 0.5]$	Total
number of genes	902	42	35	17	21	3	1020

Table 6.1: Repartition of the 1020 genes studied in Section 3.7 according to the value of their greatest conditional probability (max c.p.).

Moreover, a clustering problem is highlighted in Chapter 5. The individuals having missing values on relevant clustering variables are often misclassified. Indeed, the MAP rule is based on the conditional probabilities evaluated on the relevant variables where such an individual is observed. This problem remains to be solved.

Possible future uses of our variable selection procedure on a platform as URGV

The programming improvements of the variable selection algorithm, pointed out previ-

ously, should allow to consider the analysis of largest transcriptome datasets. Moreover, the transcriptome data analysis with our procedure can be carried out in different situations:

- Clusters of coexpressed genes for a fixed project can be searched, the variable selection concerning experiments of this project. This situation occurs for instance for the analysis of a time-course gene expression dataset.
- For a subset of interested genes, the procedure can be used to determine stable gene subgroups and biological situations where these genes are transcripted.
- The final aim consists of the global analysis of a database as CATdb. This analysis can be performed by using our method with all genes and a large experiment subset (for example study of all physiological conditions to find patterns of expressions) or a large gene subset on all available experiments (for example study of genes involved in the seed development).

Part II

Construction of a penalized likelihood criterion for variable selection in Gaussian mixture clustering with a non asymptotic point of view

In collaboration with Bertrand MICHEL

Chapter 7

A non asymptotic penalized likelihood criterion for specific Gaussian mixture model selection

Résumé: Dans ce chapitre, des mélanges gaussiens de formes spécifiques sont considérés pour résoudre un problème de sélection de variables en classification non supervisée. Un critère pénalisé non asymptotique est proposé pour sélectionner le nombre de composantes du mélange et l'ensemble des variables pertinentes pour cette classification. Le contraste de Kullback-Leibler ayant un comportement non linéaire sur les mélanges gaussiens, un théorème général de sélection de modèles pour l'estimation de densités par maximum de vraisemblance dû à Massart (2007) est utilisé pour déterminer la forme de la pénalité. Ce théorème nécessite le contrôle des entropies à crochets des familles de mélanges gaussiens étudiées. Le cas des variables ordonnées et celui des variables non ordonnées sont tous deux considérés dans ce chapitre.

7.1 Introduction

Model-based clustering methods consist of modelling clusters with parametric distributions and considering the mixture of these distributions to describe the whole dataset. They provide a rigorous framework to assess the number of mixture components and to take the variable roles into account. Currently, cluster analysis is more and more concerned with large datasets where observations are described by many variables. This large number of predictor variables could be beneficial to data clustering. Nevertheless, the useful information for clustering can be contained into only a variable subset and some of the variables can be useless or even harmful to choose a reasonable clustering structure. Several authors have suggested variable selection methods for Gaussian mixture clustering which is the most widely used mixture model for clustering multivariate continuous datasets. These methods are called "wrapper" since they are included into the clustering process. Law

et al. (2004) have introduced the feature saliency concept. Regardless of cluster membership, relevant variables are assumed to be independent of the irrelevant variables which are supposed to have the same distribution. Raftery and Dean (2006b) recast variable selection for clustering into a global model selection problem. Irrelevant variables are explained by all the relevant clustering variables according to a linear regression. The comparison between two nested variable subsets is performed using Bayes factor. A variation of this method is proposed in Maugis et al. (2008) where irrelevant variables can only depend on a relevant clustering variable subset and variables can have different sizes (variable blocks). Since all these methods are based on a variable selection procedure included into the clustering process, they do not impose specific constraints on Gaussian mixture forms. On the contrary, Bouveyron et al. (2007) consider a suitable Gaussian mixture family to take into account that data are in low-dimensional subspaces hidden in the original space. However, since this dimension reduction is based on principal components, it is difficult to deduce from this approach an interpretation of the variable roles. In all these methods, an asymptotic criterion is used to solve the underlying model selection problem.

In this chapter, a modelling taken the variable role for clustering process into account recasts variable selection and clustering problems into a model selection problem in a density estimation framework. Suppose that we observe a sample from an unknown probability distribution with density s. A specific collection of models is defined: A model $\mathcal{S}_{(K,\mathbf{v})}$ corresponds to a particular clustering situation with K clusters and a clustering "relevant" variable subset $\mathbf{v}$. A density t in $\mathcal{S}_{(K,\mathbf{v})}$ has the following form: Its projection on the relevant variable space is a Gaussian mixture density with K components and its projection on the space of the other variables is a multidimensional Gaussian density. Definitions of models $\mathcal{S}_{(K,\mathbf{v})}$ are precised in Section 7.2.1. The problem can be recast into the selection of a model among the model collection since this choice automatically leads to a data clustering and a variable selection. We propose a penalized criterion to solve this model selection problem with a non asymptotic point of view. In this approach, the "best" model is the one whose the associated maximum likelihood estimator of s gives the lowest estimation error.

In the density estimation framework, the principle of selecting a model by penalizing a loglikelihood type criterion has emerged during the seventies. Akaike (1973) proposed the AIC criterion (Akaike's information criterion) and Schwarz (1978) suggested the BIC (Bayesian Information Criterion). These two classical criteria assume implicitly that the true distribution belongs to the model collection (see for instance Burnham and Anderson, 2002). With a different point of view, the criterion ICL (Integrated Completed Likelihood) proposed by Biernacki et al. (2000) takes the clustering aim into account. Although the behaviours of these asymptotic criteria were tested in practice, there are little proved theoretical properties. For instance, the BIC consistency is only stated for the assessing cluster number under restrictive regularity assumptions and assuming that the true density belongs to the considered Gaussian mixture family (Keribin, 2000).

A non asymptotic approach for model selection via penalization has emerged during the last ten years, mainly with works of Birgé and Massart (1997) and Barron et al. (1999). An overview is available in Massart (2007). The aim of this approach is to define penalized

data-driven criteria which lead to oracle inequalities. The belonging of the true density to the model collection is not required. The penalty function depends on the number of parameters of each model and also on the complexity of the whole model collection. This approach has been carried out in several frameworks where penalty functions are explicitly assessed. In our context, a general model selection theorem for maximum likelihood estimation (MLE) is used to obtain a penalized criterion and an associated oracle inequality. This theorem proposed by Massart (2007) is a version of Theorem 2 in Barron et al. (1999). Its application requires to control the bracketing entropy of the considered Gaussian mixture models.

The chapter is organized as follows: Section 7.2 gives the model selection principles. The Gaussian mixture models, considered in this chapter, are described in Section 7.2.1 and principles of non asymptotic theory for density estimation based on Kullback-Leibler contrast are reviewed in Section 7.2.2. This section is completed by the statement of the general model selection theorem. The main results are stated in Section 7.3 and a discussion is given in Section 7.4. The proof of the main results is given in Appendix 7.A. It requires to control the bracketing entropy of mixture families. This problem is recast into a control of the bracketing entropy of Gaussian density families. This bracketing entropy is upper bounded for different Gaussian mixture forms in Appendix 7.B.

7.2 Model selection principles

7.2.1 Framework

Centered observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, with $\mathbf{y}_i \in \mathbb{R}^Q$ are assumed to be a sample from a probability distribution with unknown density s. This target s is proposed to be estimated by a finite mixture model in a clustering purpose. Note that s itself is not assumed to be a Gaussian mixture density. Model-based clustering consists of assuming that the data come from a source with several subpopulations, modelled separately and the overall population is a mixture of them. The resulting model is a finite mixture model. When the data are multivariate continuous observations, the parameterized component density is usually a multidimensional Gaussian density. Thus, a Gaussian mixture density with K components is written

$$\sum_{k=1}^{K} p_k \Phi(.|\eta_k, \Lambda_k)$$

where the p_k's are the mixing proportions ($\forall k = 1, \ldots, K, 0 < p_k < 1$ and $\sum_{k=1}^{K} p_k = 1$) and $\Phi(.|\eta_k, \Lambda_k)$ denotes the Q-dimensional Gaussian density with mean η_k and variance matrix Λ_k. The parameter vector is $(p_1, \ldots, p_K, \eta_1, \ldots, \eta_K, \Lambda_1, \ldots, \Lambda_K)$.

The mixture model is an incomplete data structure model: the complete data are $((\mathbf{y}_1, \mathbf{z}_1), \ldots, (\mathbf{y}_n, \mathbf{z}_n))$ where the missing data are $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ with $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ such that $z_{ik} = 1$ if and only if $\mathbf{y}_i$ arises from the component k. The vector $\mathbf{z}$ defines an ideal clustering of the data $\mathbf{y}$ associated to the mixture model. After an estimation of the

parameter vector thanks to the EM algorithm (Dempster et al., 1977), a data clustering is deduced from the maximum a posteriori principle:

$$\hat{z}_{ik} = \begin{cases} 1 & \text{if } \hat{p}_k \Phi(\mathbf{y}_i|\hat{\eta}_k, \hat{\Lambda}_k) > \hat{p}_l \Phi(\mathbf{y}_i|\hat{\eta}_l, \hat{\Lambda}_l),\ \forall l \neq k \\ 0 & \text{otherwise.} \end{cases} \tag{7.1}$$

Currently, statistics deals with problems where data are explained by many variables. In principle, the more information we have about each individual, the better a clustering method is expected to perform. Nevertheless, some variables can be useless or even harmful to obtain a good data clustering. Thus, it is important to take the variable role in the clustering process into account. To this aim, Gaussian mixtures with a specific form are considered. On irrelevant variables, data are assumed to have an homogeneous behavior around the null mean (centered data) allowing not to distinguish a possible clustering. Hence the data density is modelled by a spherical Gaussian joint law with null mean vector on these variables. On the contrary, the different component mean vectors are free on relevant variables. Moreover, the variance matrices restricted on relevant variables are either taken completely free or are chosen in a specified set of positive definite matrices. Note that the terms "relevant" and "irrelevant" have not the same meaning than Part I. A variable is called irrelevant if it is not informative for the clustering and relevant otherwise. Thus the redundant informative clustering variables are relevant in this part.

This modelling idea is now formalized. Let $\mathcal{V}$ be the collection of the nonempty subsets of $\{1, \ldots, Q\}$. A Gaussian mixture family is characterized by its number of mixture components $K \in \mathbb{N}^*$ and its relevant variable index subset $\mathbf{v} \in \mathcal{V}$ whose cardinal is denoted α. In the sequel, the set of index couples $(K, \mathbf{v})$ is $\mathcal{M} = \mathbb{N}^* \times \mathcal{V}$. Consider the decomposition of a vector $x \in \mathbb{R}^Q$ into its restriction on relevant variables $x_{[\mathbf{v}]} = (x_{j_1}, \ldots, x_{j_\alpha})'$ and its restriction on irrelevant variables $x_{[\mathbf{v}^c]} = (x_{l_1}, \ldots, x_{l_{Q-\alpha}})'$ where $\mathbf{v} = \{j_1, \ldots, j_\alpha\}$ and $\mathbf{v}^c = \{l_1, \ldots, l_{Q-\alpha}\} = \{1, \ldots, Q\} \backslash \mathbf{v}$. On relevant variables, a Gaussian mixture f is chosen among the following mixture family

$$\mathcal{L}_{(K,\alpha)} = \left\{ \sum_{k=1}^{K} p_k \Phi(.|\mu_k, \Sigma_k);\ \begin{array}{l} \forall k,\ \mu_k \in [-a, a]^\alpha,\ (\Sigma_1, \ldots, \Sigma_K) \in \mathcal{D}^+_{(K,\alpha)} \\ 0 < p_k < 1, \sum_{k=1}^{K} p_k = 1 \end{array} \right\}$$

where $a \in \mathbb{R}^*_+$ and $\mathcal{D}^+_{(K,\alpha)}$ denotes a family of K-tuples of $\alpha \times \alpha$ symmetric positive definite matrices whose eigenvalues are bounded. The family $\mathcal{D}^+_{(K,\alpha)}$ is related to the Gaussian mixture shape and the associated set of K-tuples Gaussian densities composing mixtures of $\mathcal{L}_{(K,\alpha)}$ is denoted $\mathcal{F}_{(K,\alpha)}$. These notations are specified hereafter. On irrelevant variables, a spherical Gaussian density g is considered, belonging to the following family

$$\mathcal{G}_{(\alpha)} = \left\{ \Phi(.|0, \omega^2 I_{Q-\alpha});\ \omega^2 \in [\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}] \right\} \tag{7.2}$$

where $0 < \lambda_{\mathrm{m}} < \lambda_{\mathrm{M}}$. Finally, the family of Gaussian mixtures associated to $(K, \mathbf{v}) \in \mathcal{M}$ is defined by

$$\mathcal{S}_{(K,\mathbf{v})} = \left\{ x \in \mathbb{R}^Q \mapsto f(x_{[\mathbf{v}]})\, g(x_{[\mathbf{v}^c]});\ f \in \mathcal{L}_{(K,\alpha)},\ g \in \mathcal{G}_{(\alpha)} \right\}. \tag{7.3}$$

The dimension of the model $\mathcal{S}_{(K,\mathbf{v})}$ is denoted $D(K,\alpha)$ and corresponds to the number of free parameters common to all Gaussian mixtures in this model. It only depends on the number of components K and the number of relevant variables α. Note that a density of $\mathcal{S}_{(K,\mathbf{v})}$ can be written as a global Gaussian mixture with mean vectors $\eta_k = (\mu_k, 0, \ldots, 0)$ and block-diagonal variance matrices Λ_k with diagonal-blocks Σ_k and $\omega^2 I_{Q-\alpha}$. A data clustering can be deduced from such a Gaussian mixture using the MAP rule (see (7.1)).

In this chapter, four collections of Gaussian mixtures are considered. For each collection, constraints are imposed on the variance matrices of the K Gaussian densities constituting a K-tuple of $\mathcal{F}_{(K,\alpha)}$. This implies a specific shape for mixtures of the associated family $\mathcal{L}_{(K,\alpha)}$. The Gaussian mixture notation for those four collections is taken from Biernacki et al. (2006) (see Chapter 1).

- For the $[L_k B_k]$ collection, the variance matrices on relevant variables are assumed to be diagonal and free, and their eigenvalues belong to the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. Thus the relevant variables are independent conditionally to mixture component belonging. In this context, an element of $\mathcal{F}_{(K,\alpha)}$ is composed of K Gaussian densities belonging to the following set

$$\mathcal{F}_{(\alpha)} = \left\{\Phi(.|\mu,\Sigma);\ \mu \in [-a,a]^\alpha,\ \Sigma = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_\alpha^2),\ \sigma_1^2,\ldots,\sigma_\alpha^2 \in [\lambda_{\mathrm{m}},\lambda_{\mathrm{M}}]\right\} \quad (7.4)$$

  and the dimension $D(K,\alpha)$ of model $\mathcal{S}_{(K,\mathbf{v})}$ is equal to $K(2\alpha+1)$.

- For the $[L_k C_k]$ collection, the variance matrices are assumed to be totally free. They belong to the set $\mathcal{D}^+_{(\alpha)}(\lambda_{\mathrm{m}},\lambda_{\mathrm{M}})$ of $\alpha \times \alpha$ positive definite matrices with eigenvalues in the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. The relevant variables are thus admitted to be correlated conditionally to mixture component belonging. The set $\mathcal{F}_{(K,\alpha)}$ composing mixtures can be assimilated to the Gaussian density family

$$\mathcal{F}_{(\alpha)} = \left\{w \in \mathbb{R}^\alpha \mapsto \Phi(w|\mu,\Sigma),\ \mu \in [-a,a]^\alpha,\ \Sigma \in \mathcal{D}^+_{(\alpha)}(\lambda_{\mathrm{m}},\lambda_{\mathrm{M}})\right\} \quad (7.5)$$

  and the dimension of the family $\mathcal{S}_{(K,\mathbf{v})}$ is equal to $D(K,\alpha) = K\left\{1+\alpha+\frac{\alpha(1+\alpha)}{2}\right\}$.

- For the $[LB_k]$ collection, the variance matrices are assumed to be diagonal and to have the same volume i.e. $\forall k \neq k', |\Sigma_k|^{\frac{1}{\alpha}} = |\Sigma_{k'}|^{\frac{1}{\alpha}}$. The variance matrices are decomposed into $\Sigma_k = \beta B_k$ where the common volume β belongs to $[\beta_{\mathrm{m}}, \beta_{\mathrm{M}}]$ and B_k is a diagonal matrix with a determinant 1 and with diagonal coefficients in the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. Thus the family of K-tuples of Gaussian densities composing mixtures of $\mathcal{L}_{(K,\alpha)}$ is

$$\mathcal{F}_{(K,\alpha)} = \left\{(\Phi(.|\mu_1,\beta B_1),\ldots,\Phi(.|\mu_K,\beta B_K));\ \begin{array}{l}\forall 1 \leq k \leq K,\ \mu_k \in [-a,a]^\alpha \\ B_k \in \Delta^1_{(\alpha)}(\lambda_{\mathrm{m}},\lambda_{\mathrm{M}}), \beta \in [\beta_{\mathrm{m}},\beta_{\mathrm{M}}]\end{array}\right\} \quad (7.6)$$

  where $\Delta^1_{(\alpha)}(\lambda_{\mathrm{m}},\lambda_{\mathrm{M}})$ is the set of $\alpha \times \alpha$ diagonal matrices with determinant 1 and whose eigenvalues are in the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$ where $0 < \lambda_{\mathrm{m}} < \lambda_{\mathrm{M}}$. Here, the model dimension is equal to $D(K,\alpha) = 2K\alpha + 1$.

- For the $[LC]$ collection, the variance matrices are all equal to a free positive definite matrix Σ whose eigenvalues are assumed to be in the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. The set of such variance matrices is denoted $\mathcal{D}^{+}_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}})$. The family $\mathcal{F}_{(K,\alpha)}$ is thus defined by

$$\mathcal{F}_{(K,\alpha)} = \{(\Phi(.|\mu_1, \Sigma), \ldots, \Phi(.|\mu_K, \Sigma));\ \mu_k \in [-a, a]^{\alpha},\ \Sigma \in \mathcal{D}^{+}_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}})\} \qquad (7.7)$$

  and the model dimension is $D(K, \alpha) = K(1 + \alpha) + \frac{\alpha(\alpha+1)}{2}$.

Note that the family $\mathcal{F}_{(K,\alpha)}$ cannot be assimilated to a Gaussian density set $\mathcal{F}_{(\alpha)}$ for the $[LB_k]$ and $[LC]$ collections since the variance matrices have a common characteristic: The variance matrices have a common volume for the $[LB_k]$ collection and are equal for the $[LC]$ collection. This distinction with the two other collections will be important to obtain the penalized criterion. It is interesting to consider these different collections since the results obtained further are stated in function of the model dimension. Whereas a mixture can belong to different collections, its number of free parameters is different according to the mixture shape. Furthermore, the consideration of different Gaussian mixture collections will allow for a large practical use of our results. To make easier the reading of this chapter, the same notation $\mathcal{S}_{(K,\mathbf{v})}$ is used for the four model collections. Finally in order to extend the application field, the cases of ordered and non-ordered variables are both addressed in this chapter. If variables are assumed to be ordered, the relevant variable subset is $\mathbf{v} = \{1, \ldots, \alpha\}$ and can be assimilated to its cardinal α. Thus, in order to distinguish between these two cases, Gaussian mixture families are denoted $\mathcal{S}_{(K,\alpha)}$ when variables are assumed to be ordered.

These Gaussian mixture families allow us to recast clustering and variable selection problems into a global model selection problem. A criterion is now required to select the best model according to the dataset. We propose a penalized criterion using a non asymptotic approach whose principles are given in the following section.

7.2.2 Non asymptotic model selection

Density estimation deals with the problem of estimating an unknown distribution corresponding to the observation of a sample $\mathbf{y}$. In many cases, it is not easy to choose a model of adequate dimension. For instance, a model with few parameters tends to be efficiently estimated whereas it could be far from the true distribution. In the opposite situation, a more complex model easily fits data but estimates have larger variances. The aim of model selection is to construct a data-driven criterion to select a model of proper dimension among a model collection. A general theory on this topic, with a non asymptotic approach is proposed in the works of Birgé and Massart (see for instance Birgé and Massart, 2001a,b). This model selection principle is now described in our density estimation framework.

Let $\mathcal{S}$ be the set of all densities with respect to the Lebesgue measure on $\mathbb{R}^Q$. The contrast $\gamma(t, .) = -\ln\{t(.)\}$ is considered, leading to the maximum likelihood criterion.

The corresponding loss function is the Kullback-Leibler information. It is defined for two densities s and t in $\mathcal{S}$ by

$$\mathrm{KL}(s,t) = \int \ln\left\{\frac{s(x)}{t(x)}\right\} s(x)\,dx$$

if sdx is absolutely continuous with respect to tdx and $+\infty$ otherwise. The density s being the unique minimizer of the Kullback-Leibler function on $\mathcal{S}$, it satisfies

$$s = \underset{t\in\mathcal{S}}{\operatorname{argmin}} \int -\ln\{t(x)\}s(x)\,dx.$$

Consequently, s is also a minimizer over $\mathcal{S}$ of the expectation of the empirical contrast defined by

$$\gamma_n(t) = -\frac{1}{n}\sum_{i=1}^{n} \ln\left\{t(\mathbf{y}_i)\right\}.$$

A minimizer of the empirical contrast γ_n over a model S, a subspace of $\mathcal{S}$, is denoted $\hat{s}$. Substituting the empirical criterion γ_n to its expectation and minimizing γ_n on S, it is expected to obtain a sensible estimator of s, at least if s belongs (or is close enough) to model S.

A countable collection of models $(S_m)_{m\in\mathcal{M}}$ with a corresponding collection $(\hat{s}_m)_{m\in\mathcal{M}}$ of estimators is considered. The best model is the one presenting the smallest risk

$$m(s) = \underset{m\in\mathcal{M}}{\operatorname{argmin}}\ \mathbb{E}[\mathrm{KL}(s,\hat{s}_m)].$$

However, the function $\hat{s}_{m(s)}$, called oracle, is unknown since it depends on the true density s. Nevertheless, this oracle is a benchmark: A data-driven criterion is then found to select an estimator such that its risk is close to the oracle risk. The model selection via penalization procedure consists of considering some proper penalty function pen : $m \in \mathcal{M} \mapsto \mathrm{pen}(m) \in \mathbb{R}^+$ and of selecting $\hat{m}$ minimizing the associated penalized criterion

$$\mathrm{crit}(m) = \gamma_n(\hat{s}_m) + \mathrm{pen}(m).$$

The resulting selected estimator is denoted $\hat{s}_{\hat{m}}$. The final purpose of this non asymptotic approach is to obtain a penalty function and an associated oracle inequality, allowing to compare the risk of the penalized MLE $\hat{s}_{\hat{m}}$ with the benchmark $\inf_{m\in\mathcal{M}} \mathbb{E}[\mathrm{KL}(s,\hat{s}_m)]$.

Commonly, in order to find a suitable penalty function, one begins by writing the following inequality (see Massart, 2007, p.9): For all $m \in \mathcal{M}$ and $s_m \in S_m$,

$$\mathrm{KL}(s,\hat{s}_{\hat{m}}) \le \mathrm{KL}(s,s_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + \bar{\gamma}_n(s_m) - \bar{\gamma}_n(\hat{s}_{\hat{m}})$$

where $\bar{\gamma}_n$ is the centered empirical process defined by $\bar{\gamma}_n(t) = \gamma_n(t) - \mathbb{E}[\gamma_n(t)]$. The penalty function has to be chosen to annihilate the fluctuation of $\bar{\gamma}_n(s_m) - \bar{\gamma}_n(\hat{s}_{\hat{m}})$. The aim is to obtain an uniform control of $\bar{\gamma}_n(s_m) - \bar{\gamma}_n(\hat{s}_{m'})$ with respect to m' in $\mathcal{M}$. This quantity

is controlled by its expectation using a Talagrand's inequality (see Talagrand, 1995, 1996; Massart, 2007, for an overview). Next, two different situations occur. In some situations, the expectation in the Talagrand's inequality can be efficiently connected to the model dimension, and an oracle inequality with explicit constants is deduced. This is the case in the context of histogram density estimation (Castellan, 1999) and of density estimation via exponential model (Castellan, 2003). For situations when these sharp calculations are impossible to obtain, Massart (2007) proposes a general theorem which gives the form of penalties and associated oracle inequalities in terms of the Kullback-Leibler and Hellinger losses. This theorem is based on the centered process control with the bracketing entropy, allowing to evaluate the "size" of models. For Gaussian mixture models, we can only follow the second alternative because of the non linear behavior of the logarithm function on Gaussian mixture densities. Moreover, being impossible to bound uniformly all the ratios of two Gaussian mixtures in our context, a hypothesis of boundness as for all $t \in S_{m'}$ $\|\bar{\gamma}_n(s_m) - \bar{\gamma}_n(t)\|_\infty$ is bounded by a constant, which is required to apply concentration inequalities, cannot be fulfilled.

Before stating the general MLE selection model theorem (Massart, 2007, Theorem 7.11) in a restricted form which is sufficient for our study, the definition of the Hellinger distance and some notation are specified. The norm $\|\sqrt{f} - \sqrt{g}\|_2$ between two nonnegative functions f and g of $\mathbb{L}_1$ is denoted $d_H(f,g)$. We note that if f and g are two densities with respect to the Lebesgue measure on $\mathbb{R}^Q$, $d_H(f,g)$ is the Hellinger distance between f and g. In the following, $d_H(f,g)$ is improperly called Hellinger distance even if f and g are not density functions. An ε-bracketing for a subset S of $\mathcal{S}$ with respect to d_H is a set of integrable function pairs $(l_1,u_1),\ldots,(l_N,u_N)$ such that for each $f \in S$, there exists $j \in \{1,\ldots,N\}$ such that $l_j \leq f \leq u_j$ and $d_H(l_j,u_j) \leq \varepsilon$. The bracketing number $\mathcal{N}_{[.]}(\varepsilon,S,d_H)$ is the smallest number of ε-brackets necessary to cover S and the bracketing entropy is defined by $\mathcal{H}_{[.]}(\varepsilon,S,d_H) = \ln\left\{\mathcal{N}_{[.]}(\varepsilon,S,d_H)\right\}$. Since $\mathcal{S}$ is the density set, the bracket extremities can be chosen as nonnegative functions in $\mathbb{L}_1$. The notation $a \wedge b$ is the infimum between a and b.

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 Lebesgue measure. Firstly, the following separability assumption allows to avoid measurability problems. For each model S_m, assume that there exists some countable subset S'_m of S_m such that for all $t \in S_m$, there exists a sequence $(t_k)_{k\geq 1}$ of elements of S'_m such that for $x \in \mathbb{R}^Q$, $\ln\{t_k(x)\}$ tends to $\ln\{t(x)\}$ when k tends to infinity. Secondly $\sqrt{\mathcal{H}_{[.]}(\varepsilon,S_m,d_H)}$ is assumed to be integrable at 0 for each m and we also assume that there exists a function Ψ_m on $\mathbb{R}_+$ fulfilling the following properties

[I]: Ψ_m is nondecreasing, $x \to \Psi_m(x)/x$ is nonincreasing on $]0,+\infty[$ and for $\xi \in \mathbb{R}_+$ and all $u \in S_m$, denoting $S_m(u,\xi) = \{t \in S_m; d_H(t,u) \leq \xi\}$,

$$\int_0^\xi \sqrt{\mathcal{H}_{[.]}(x,S_m(u,\xi),d_H)}\,dx \leq \Psi_m(\xi).$$

Theorem 7.2.1 (Massart (2007))**.** *Let* $\mathbf{y}_1, \ldots, \mathbf{y}_n$ *be i.i.d. random variables with unknown density* s *with respect to Lebesgue measure on* $\mathbb{R}^Q$*. Let* $(S_m)_{m\in\mathcal{M}}$ *be some at most countable collection of models fulfilling the previous properties and let* $(\hat{s}_m)_{m\in\mathcal{M}}$ *be the corresponding collection of MLEs. Let* $(\rho_m)_{m\in\mathcal{M}}$ *be some family of nonnegative numbers such that*

$$\sum_{m\in\mathcal{M}} e^{-\rho_m} = \Upsilon < \infty.$$

For every $m \in \mathcal{M}$*, considering* Ψ_m *with properties* ***[I]****,* ξ_m *denotes the unique positive solution of the equation*

$$\Psi_m(\xi) = \sqrt{n}\,\xi^2.$$

Let $\text{pen} : \mathcal{M} \to \mathbb{R}_+$ *and consider the penalized loglikelihood criterion*

$$\text{crit}(m) = \gamma_n(\hat{s}_m) + \text{pen}(m).$$

Then, there exists some absolute constants κ *and* C *such that whenever for all* $m \in \mathcal{M}$*,*

$$\text{pen}(m) \geq \kappa \left(\xi_m^2 + \frac{\rho_m}{n}\right)$$

some random variable $\hat{m}$ *minimizing* crit *over* $\mathcal{M}$ *does exist and moreover, whatever the density* s*,*

$$\mathbb{E}\left[d_H^2(s, \hat{s}_{\hat{m}})\right] \leq C \left[\inf_{m\in\mathcal{M}} \{\text{KL}(s, S_m) + \text{pen}(m)\} + \frac{\Upsilon}{n}\right], \tag{7.8}$$

where $\text{KL}(s, S_m) = \inf_{t\in S_m} \text{KL}(s, t)$ *for every* $m \in \mathcal{M}$*.*

Inequality (7.8) is not exactly an oracle inequality since the Hellinger risk is upper bounded by the Kullback bias $\text{KL}(s, S_m)$. Nevertheless, this last term is of the order of $d_H^2(s, S_m)$ if $\ln(\|s/t\|_\infty)$ is uniformly bounded on $\cup_{m\in\mathcal{M}} S_m$ according to Lemma 7.23 in Massart (2007). In our context, this condition can be achieved if all densities are assumed to be bounded and defined on a compact support, the Gaussian mixtures being truncated on this compact support.

7.3 Main results

As announced previously, Theorem 7.2.1 is applied to our specific framework described in Section 7.2.1. The ensuing theoretical results are now addressed, for the ordered and non-ordered variable cases separately: For each one, a non asymptotic penalized criterion is provided to select the number of clusters K and the variable subset $\mathbf{v}$ used for Gaussian mixtures. Moreover, these results give an oracle inequality which is fulfilled by the associated penalized estimator.

7.3.1 Ordered variable case

In this section, variables are assumed to be ordered and the model collection is denoted $(\mathcal{S}_{(K,\alpha)})_{(K,\alpha)\in\mathcal{M}}$. In the four types of Gaussian mixtures, the following theorem gives the form of penalty functions and the associated oracle inequalities.

Theorem 7.3.1. *For the four Gaussian mixture collections, there exists two absolute constants κ and C such that, if*

$$\mathrm{pen}(K,\alpha) \geq \kappa \frac{D(K,\alpha)}{n}\left\{1 + 2\,A + \ln\left(\frac{1}{1 \wedge \frac{D(K,\alpha)}{n}\,A}\right)\right\}$$

where the constant A is a function of Q, λ_m, λ_M, a, and also β_m, β_M for the $[LB_k]$ collection, such that $A = O(\sqrt{\ln Q})$ as Q tends to infinity, then the model $(\hat{K},\hat{\alpha})$ minimizing

$$\mathrm{crit}(K,\alpha) = \gamma_n(\hat{s}_{(K,\alpha)}) + \mathrm{pen}(K,\alpha)$$

over $\mathcal{M}$ exists and

$$\mathbb{E}\left[d_H^2(s,\hat{s}_{(\hat{K},\hat{\alpha})})\right] \leq C\left[\inf_{(K,\alpha)\in\mathcal{M}}\{\mathrm{KL}(s,\mathcal{S}_{(K,\alpha)}) + \mathrm{pen}(K,\alpha)\} + \frac{1}{n}\right].$$

This theorem is proved in Appendix 7.A.1. It requires to control the bracketing entropy of Gaussian mixture families. This problem is recast into the control for Gaussian density families. Appendices 7.B.1, 7.B.2, 7.B.3 and 7.B.4 are then devoted to the bracketing entropy control of Gaussian density families $\mathcal{F}_{(\alpha)}$ for the $[L_kB_k]$ and $[L_kC_k]$ collections and of $\mathcal{F}_{(K,\alpha)}$ for the $[LB_k]$ and $[LC]$ collections respectively. Note that in order to apply Theorem 7.2.1, the local bracketing entropy $\mathcal{H}_{[.]}(x,\mathcal{S}_{(K,\alpha)}(u,\xi),d_H)$ has to be controlled. Nevertheless, it is difficult to characterize the subset $\mathcal{S}_{(K,\alpha)}(u,\xi)$ in function of the parameters of its mixtures. Therefore a global study of the entropy bracketing is proposed in the theorem proof since $\mathcal{H}_{[.]}(x,\mathcal{S}_{(K,\alpha)}(u,\xi),d_H) \leq \mathcal{H}_{[.]}(x,\mathcal{S}_{(K,\alpha)},d_H)$.

Several remarks can be given about this result. First, the deduced penalty function has an expected form since it is proportional to the model dimension $D(K,\alpha)$. This shows the interest of considering separately the four collections since the model dimensions are different. For instance, for the $[L_kB_k]$ mixture family, the risk bound is less accurate when this family is considered as a subset of the $[L_kC_k]$ collection. Second, the constant A is made explicit in the theorem proof (see Appendix 7.A.1) and its expression is different for each mixture collection (see Equations (7.12), (7.13) and (7.14)). It depends on parameters λ_{m}, λ_{M}, a, Q and also β_{m}, β_{M} for the $[LB_k]$ collection and $A = O(\sqrt{\ln Q})$ as Q tends to infinity. This number of variables Q has to have a reasonable order in the constant A so that the upper bound in the oracle inequality remains meaningful. Contrary to classical criteria for which Q is fixed and n tends to infinity, our result allows to study cases for which Q increases with n. For specific clustering problems where the number of variables Q is of the order of n or even larger than n, the oracle inequality is still significant. Third, since the multiplicative constants are not explicit, a practical method is necessary. This is addressed in Chapter 8.

7.3.2 Non-ordered variable case

Theorem 7.3.1 can be generalized to the non-ordered variable case. In this context, a model $\mathcal{S}_{(K,\mathbf{v})}$ is characterized by its number of mixture components K and its subset $\mathbf{v} \in \mathcal{V}$ of relevant variable indexes. This model is related to the model $\mathcal{S}_{(K,\alpha)}$ of the ordered case by

$$\mathcal{S}_{(K,\mathbf{v})} = \{x \in \mathbb{R}^Q \mapsto f \circ \tau(x),\ f \in \mathcal{S}_{(K,\alpha)}\}$$

where τ is a permutation such that $(\tau(x)_1, \ldots, \tau(x)_\alpha)' = x_{[\mathbf{v}]}$ and has the same dimension $D(K,\alpha)$. Consequently, the model $\mathcal{S}_{(K,\mathbf{v})}$ has the same complexity as $\mathcal{S}_{(K,\alpha)}$ and thus has the same bracketing entropy. However, the model set $\{\mathcal{S}_{(K,\mathbf{v})}\}_{(K,\mathbf{v})\in\mathcal{M}}$ contains more models per dimension than in the ordered case. This richness of the model family involves to define following new weights in penalty function:

$$\rho_{(K,\mathbf{v})} = \frac{D(K,\alpha)}{2} \ln\left[\frac{8eQ}{\{D(K,\alpha)-1\} \wedge (2Q-1)}\right].$$

Consequently, in the following theorem which is the analog of Theorem 7.3.1 for the non-ordered case, the associated penalty functions have an additional logarithm term depending on the dimension.

Theorem 7.3.2. *For the four Gaussian mixture collections, there exists two absolute constants κ and C such that, if*

$$\mathrm{pen}(K,\mathbf{v}) \geq \kappa \frac{D(K,\alpha)}{n}\left(2\,A + \ln\left\{\frac{1}{1 \wedge \frac{D(K,\alpha)}{n}\,A}\right\} + \frac{1}{2}\ln\left[\frac{8eQ}{\{D(K,\alpha)-1\} \wedge (2Q-1)}\right]\right)$$

where A is the same constant as the ordered case, then the model $(\hat{K},\hat{\mathbf{v}})$ minimizing $\mathrm{crit}(K,\mathbf{v}) = \gamma_n(\hat{s}_{(K,\mathbf{v})}) + \mathrm{pen}(K,\mathbf{v})$ *on $\mathcal{M}$ exists and*

$$\mathbb{E}\left[d_H^2(s,\hat{s}_{(\hat{K},\hat{\mathbf{v}})})\right] \leq C\left[\inf_{(K,\mathbf{v})\in\mathcal{M}}\{\mathrm{KL}(s,\mathcal{S}_{(K,\mathbf{v})}) + \mathrm{pen}(K,\mathbf{v})\} + \frac{2}{n}\right].$$

The theorem proof given in Appendix 7.A.2 only consists of justifying the form of new weights and finding an upper bound of the weight sum since $\mathcal{S}_{(K,\mathbf{v})}$ has the same bracketing entropy as $\mathcal{S}_{(K,\alpha)}$. This non-ordered case is more attractive for practical use but Theorem 7.3.2 is difficult to apply when the number of variables becomes too large since an exhaustive research of the best model is then untractable.

7.4 Discussion

In this chapter, specific Gaussian mixtures are considered to take the role of variables in the clustering process into account. Main results are stated for four Gaussian mixture forms for ordered and non-ordered variables. A non asymptotic penalized criterion is proposed to

select the number of clusters and the clustering relevant variable subset. Oracle inequalities satisfied by the associated estimator $\hat{s}_{(\hat{K},\hat{\mathbf{v}})}$ are also obtained. The main interest of these results is to give the form of an adequate penalty in this particular framework. Proofs of these results require to control the bracketing entropy of multidimensional Gaussian density families and to determinate weights taking the richness of the model collection into account. Similar results for non-Gaussian mixtures can be obtained as soon as the bracketing entropy of the new component density families can be controlled. To give more interest to our results, we have attempted to establish an adaptive property of our penalized maximum likelihood estimators in a minimax sense. In order to find target classes of densities which are closely approximated by Gaussian mixtures, we have searched a theoretical result about the estimation rate of regular functions by Gaussian mixtures, expressed according to the regularity coefficient and the sample size. Unfortunately, such a result of approximation theory for Gaussian mixtures lacks to make the link between model selection and adaptive estimation.

A complete collection of twenty eight parsimonious models is available, used for instance in MIXMOD software (Biernacki et al., 2006). These models are obtained by imposing conditions on the proportions and the elements of variance matrix eigenvalue decomposition (see Chapter 1). In this chapter, we focus on four mixture forms but similar results can be stated for other mixture forms. Without difficulty, the results can be extended from the present work to all spherical mixtures, diagonal mixtures and mixtures with the form type $[p\_L\_C\_]$ (see Table 1.1). On the contrary, it is presently impossible to extend the results for mixture types $[p\_DA_kD']$ and $[p\_D_kAD'_k]$ because of the difficulty to construct a countable covering over the orthogonal matrices.

Usually, the Gaussian mixture clustering problem involves the selection of the number of mixture components and besides, of the mixture shape among a mixture shape collection. Commonly, an asymptotic criterion as BIC (Schwarz, 1978) or ICL (Biernacki et al., 2000) is used to solve this model selection problem. Our main results allow us to propose a non asymptotic criterion to select the number of clusters, the subset $\mathbf{v}$ being fixed to the complete variable set. It would be possible to extend our penalized criterion to select the mixture form as well. We get back to this topic in Chapter 9.

For practical purposes, theoretical results stated in this chapter cannot be immediately used since they depend on unknown constants and mixture parameters are not bounded. Nevertheless, they are required to justify the shape of penalties and allow that the number of variables Q can be large. Birgé and Massart (2006) propose their so-called "slope heuristics" (see also Massart, 2007, Section 8.5) to calibrate these constants. The topic of the next chapter is to carry out this heuristics in our framework to allow for a practical use of our results.

Appendices

7.A Proofs of the main results

7.A.1 Proof of Theorem 7.3.1

We consider the Gaussian mixture models $\mathcal{S}_{(K,\alpha)}$ where variables are assumed to be ordered. The aim of this section is to apply the general MLE selection model theorem (Theorem 7.2.1) in order to prove Theorem 7.3.1. This requires to determine a suitable function $\Psi_{(K,\alpha)}$ fulfilling properties **[I]**. Thus, the first step consists of controlling the bracketing entropy of the Gaussian mixture families $\mathcal{S}_{(K,\alpha)}$.

Control of the bracketing entropy of mixture model families:

Ghosal and van der Vaart (2001) and Genovese and Wasserman (2000) have proposed an upper bound of the bracketing entropy of unidimensional Gaussian mixtures in order to obtain convergence rates in Hellinger distance for density estimation using the Gaussian mixtures. We first tried to follow the strategy proposed in Ghosal and van der Vaart (2001) to control the bracketing entropy of our multidimensional Gaussian mixture models. But the control obtained this way has a too large dependency in Q: The constant A depends on a power of Q, allowing not that Q is of the order of n or even larger than n in particular. We propose instead a method inspired by the work of Genovese and Wasserman (2000). The key idea is given by their theorem stated hereafter: The control of the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$ can be recast into the control of the bracketing entropies of the associated mixture component density families. For all k in $\{1,\dots,K\}$, let $\mathcal{C}_k = \{f_{\theta_k}, \theta_k \in \Theta_k\}$ be a family of densities with respect to Lebesgue measure on $\mathbb{R}^Q$. The following family of mixture distributions based on $\mathcal{C}_k$ is considered

$$\mathcal{W}_K := \left\{ \sum_{k=1}^{K} p_k f_{\theta_k},\ \theta_k \in \Theta_k\ \forall k = 1,\dots,K,\ \mathbf{p} = (p_1,\dots,p_K) \in \mathcal{P}_{K-1} \right\}$$

where $\mathcal{P}_{K-1}$ is the $K-1$ dimensional simplex defined by

$$\mathcal{P}_{K-1} := \Big\{ \mathbf{p} = (p_1,\dots,p_K),\ \forall k = 1,\dots,K,\ p_k \geq 0,\ \sum_{k=1}^{K} p_k = 1 \Big\}.$$

Theorem 7.A.1. *With the previous notation, for all K and all $\varepsilon \in (0,1]$,*

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{W}_K, d_H) \leq \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{P}_{K-1}, d_H\right) \prod_{k=1}^{K} \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{C}_k, d_H\right)$$

where

$$\mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{P}_{K-1}, d_H\right) \leq K(2\pi\, e)^{\frac{K}{2}} \left(\frac{3}{\varepsilon}\right)^{K-1}.$$

In our context, we want to take the specific form of the studied multidimensional mixtures into account. Recall that two situations occur, depending on whether $\mathcal{F}_{(K,\alpha)}$ can be written as the cartesian product of K times a set $\mathcal{F}_{(\alpha)}$ or not. For these two situations, a new result is deduced from Theorem 7.A.1.

For the $[L_kB_k]$ and $[L_kC_k]$ collections, Equation (7.3) gives that an element $f \in \mathcal{S}_{(K,\mathbf{v})}$ can be written, for all $x \in \mathbb{R}^Q$,

$$f(x) = \Phi(x_{[\mathbf{v}^c]}|0, \omega^2 I_{Q-\alpha}) \sum_{k=1}^{K} p_k \Phi(x_{[\mathbf{v}]}|\mu_k, \Sigma_k)$$

where $\Phi(.|0, \omega^2 I_{Q-\alpha})$ belongs to $\mathcal{G}_{(\alpha)}$ and where Gaussian densities $\Phi(.|\mu_k, \Sigma_k)$ belong to $\mathcal{F}_{(\alpha)}$ (see Section 7.2.1). According to Theorem 7.A.1, the bracketing entropy of the mixture family $\mathcal{L}_{(K,\alpha)}$ is related to the one of $\mathcal{F}_{(\alpha)}$ ($\mathcal{C}_k = \mathcal{F}_{(\alpha)},\ \forall k \in \{1, \ldots, K\}$). The following proposition is deduced from Theorem 7.A.1 and is proved in Appendix 7.C. It allows us to bound the bracketing entropy of the mixture family $\mathcal{S}_{(K,\mathbf{v})}$ by a product of the bracketing entropies of the simplex $\mathcal{P}_{K-1}$, of $\mathcal{G}_{(\alpha)}$ and of $\mathcal{F}_{(\alpha)}$.

Proposition 7.A.1. *For the $[L_kB_k]$ and $[L_kC_k]$ mixture collections, for all $\varepsilon \in (0, 1]$, the bracketing number of the density family $\mathcal{S}_{(K,\mathbf{v})}$ is bounded by*

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq K(2\pi\, e)^{\frac{K}{2}} \left(\frac{9}{\varepsilon}\right)^{K-1} \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) \mathcal{N}_{[.]}\left(\frac{\varepsilon}{9}, \mathcal{F}_{(\alpha)}, d_H\right)^K.$$

It is then deduced that the bracketing entropy of $\mathcal{S}_{(K,\mathbf{v})}$ is bounded by

$$\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq C(K) + (K-1)\ln\left(\frac{1}{\varepsilon}\right) + \mathcal{H}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) + K\,\mathcal{H}_{[.]}\left(\frac{\varepsilon}{9}, \mathcal{F}_{(\alpha)}, d_H\right) \quad (7.9)$$

with $C(K) = \ln(K) + \frac{K}{2}\ln(2\pi e) + (K-1)\ln(9)$.

As explained in Section 7.2.1, the variance matrices have a common element in the $[LB_k]$ and $[LC]$ collections: They have the same volume for the $[LB_k]$ collection and are equal for the $[LC]$ collection. The family $\mathcal{F}_{(K,\alpha)}$ cannot be assimilated to one Gaussian density family and thus Theorem 7.A.1 cannot be applied in this case. Nevertheless the following proposition is a variant, allowing us to take the specific form of studied mixtures into account. Its proof, stated in Appendix 7.C, is established along the line of the proof of Theorem 2 in Genovese and Wasserman (2000). This proposition recasts the problem to upper bound the bracketing entropy of $\mathcal{S}_{(K,\mathbf{v})}$ into the study of the bracketing entropy of the simplex, $\mathcal{G}_{(\alpha)}$ and $\mathcal{F}_{(K,\alpha)}$.

Proposition 7.A.2. *For the $[LB_k]$ and $[LC]$ mixture collections, for all $\varepsilon \in (0,1]$, the bracketing number of $\mathcal{S}_{(K,\mathbf{v})}$ is upper bounded by*

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq K(2\pi\, e)^{\frac{K}{2}} \left(\frac{9}{\varepsilon}\right)^{K-1} \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) \mathcal{N}_{[.]}\left(\frac{\varepsilon}{9}, \mathcal{F}_{(K,\alpha)}, d_H\right).$$

Hence with $C(K) = \ln(K) + \frac{K}{2}\ln(2\pi e) + (K-1)\ln(9)$,

$$\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq C(K) + (K-1)\ln\left(\frac{1}{\varepsilon}\right) + \mathcal{H}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) + \mathcal{H}_{[.]}\left(\frac{\varepsilon}{9}, \mathcal{F}_{(K,\alpha)}, d_H\right). \tag{7.10}$$

Finally, the control of the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$ is recast into the one of $\mathcal{G}_{(\alpha)}$ and $\mathcal{F}_{(\alpha)}$ or $\mathcal{F}_{(K,\alpha)}$ according to the considered mixture collection. This control is stated in Proposition 7.B.1, Proposition 7.B.2, Proposition 7.B.7 and Proposition 7.B.8 for the $[L_kB_k]$, $[L_kC_k]$, $[LB_k]$ and $[LC]$ collections respectively.

Determination of a function $\Psi_{(K,\alpha)}$:

Admitting the bracketing entropy upper bounds given in Appendix 7.B, a convenient function $\Psi_{(K,\alpha)}$, fulfilling properties **[I]**, has to be determined in order to apply Theorem 7.2.1. First, the $[L_kB_k]$ collection is considered. According to Proposition 7.B.1, for all positive real number ξ,

$$\begin{aligned}\int_0^{\xi} \sqrt{\mathcal{H}_{[.]}(x, \mathcal{S}_{(K,\alpha)}, d_H)}\, dx \;\leq\; & \xi \left\{ \sqrt{C(K)} + \sqrt{K\alpha \ln\left(a\sqrt{\frac{8}{c_1 \lambda_{\mathrm{m}}}}\right)} + \sqrt{(K\alpha+1)\ln\left(8\frac{\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)} \right\} \\ & + \;\; \xi\sqrt{(2K\alpha+1)\ln(9\sqrt{2}\,Q)} \;+\; \int_0^{\xi\wedge 1} \sqrt{D(K,\alpha)\ln\left(\frac{1}{x}\right)}\, dx. \end{aligned} \tag{7.11}$$

In order to control the last term of the right-hand side of Inequality (7.11), the following technical result is considered:

Lemma 7.A.3. *For all $\varepsilon \in (0,1]$,*

$$\int_0^{\varepsilon} \sqrt{\ln\left(\frac{1}{x}\right)}\, dx \leq \varepsilon \left\{ \sqrt{\ln\left(\frac{1}{\varepsilon}\right)} + \sqrt{\pi} \right\}.$$

Proof. This inequality is deduced from an integration by part and the following concentration inequality (see Massart, 2007, p.19): If Z is a centered standard Gaussian variable then $P(Z \geq c) \leq e^{-\frac{c^2}{2}}$ for all $c > 0$. □

Thus, using Lemma 7.A.3, we get

$$\begin{aligned}\int_0^\xi \sqrt{\mathcal{H}_{[.]}(x,\mathcal{S}_{(K,\alpha)},d_H)}\,dx &\leq \xi\left\{\sqrt{C(K)}+\sqrt{K\alpha\ln\left(a\sqrt{\frac{8}{c_1\lambda_{\mathrm{m}}}}\right)}+\sqrt{(K\alpha+1)\ln\left(8\frac{\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)}\right\}\\ &\quad+\xi\left\{\sqrt{(2K\alpha+1)\ln(9\sqrt{2}\,Q)}\right\}\\ &\quad+\xi\sqrt{D(K,\alpha)}\left\{\sqrt{\ln\left(\frac{1}{1\wedge\xi}\right)}+\sqrt{\pi}\right\}\\ &\leq \xi\sqrt{D(K,\alpha)}\left\{(\square)+\sqrt{\ln\left(\frac{1}{1\wedge\xi}\right)}\right\}\end{aligned}$$

with

$$\begin{aligned}(\square) &= \sqrt{\frac{C(K)}{D(K,\alpha)}}+\sqrt{\frac{K\alpha}{D(K,\alpha)}\ln\left(a\sqrt{\frac{8}{c_1\lambda_{\mathrm{m}}}}\right)}+\sqrt{\frac{(K\alpha+1)}{D(K,\alpha)}\ln\left(8\frac{\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)}\\ &\quad+\sqrt{\frac{(2K\alpha+1)}{D(K,\alpha)}\ln(9\sqrt{2}\,Q)}+\sqrt{\pi}.\end{aligned}$$

Moreover, since $\frac{C(K)}{D(K,\alpha)} \leq \ln(18\pi e^2)$ and $\frac{K\alpha}{D(K,\alpha)}$, $\frac{K\alpha+1}{D(K,\alpha)}$ and $\frac{2K\alpha+1}{D(K,\alpha)}$ are all smaller than 1, $(\square)$ is bounded by a constant $\mathcal{A}(\lambda_{\mathrm{m}},\lambda_{\mathrm{M}},a,Q)$ denoted only $\mathcal{A}$ hereafter and defined by

$$\mathcal{A}(\lambda_{\mathrm{m}},\lambda_{\mathrm{M}},a,Q) := \sqrt{\pi}+\sqrt{\ln(18\pi e^2)}+\sqrt{\ln\left(a\sqrt{\frac{8}{c_1\lambda_{\mathrm{m}}}}\right)}+\sqrt{\ln\left(8\frac{\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)}+\sqrt{\ln(9\sqrt{2}\,Q)}. \tag{7.12}$$

In the same way, using Proposition 7.B.2, Proposition 7.B.7 and Proposition 7.B.8 for the $[L_kC_k]$, $[LB_k]$ and $[LC]$ mixture collections respectively, we obtain that for all $\xi > 0$

$$\int_0^\xi \sqrt{\mathcal{H}_{[.]}(x,\mathcal{S}_{(K,\alpha)},d_H)}\,dx \leq \xi\sqrt{D(K,\alpha)}\left\{\mathcal{A}+\sqrt{\ln\left(\frac{1}{1\wedge\xi}\right)}\right\}$$

where the constant $\mathcal{A}$ is given by,

- for the $[L_kC_k]$ and $[LC]$ collections:

$$\begin{aligned}\mathcal{A} &= \sqrt{\ln(18\pi e^2)}+\sqrt{\ln(Q^2)}+\sqrt{\pi}\\ &\quad+\sqrt{\ln\left(\frac{24\sqrt{2}\;\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)}+\sqrt{\ln\left(\frac{54\sqrt{3}\,\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)}+\sqrt{\ln\left(\frac{54\,a}{\sqrt{\lambda_{\mathrm{m}}}}\right)}.\end{aligned} \tag{7.13}$$

- for the $[LB_k]$ collection:

$$\begin{aligned}\mathcal{A} &= \sqrt{\ln(18\pi e^2)} + \sqrt{\pi} + \sqrt{\ln(Q)} \\ &\quad + \sqrt{\ln\left(\frac{216\sqrt{2}\,\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)} + \sqrt{\ln\left(\frac{6a}{\sqrt{\beta_{\mathrm{M}}(1-2^{-\frac{1}{4}})}}\right)} + \sqrt{\ln\left(\frac{24\,\beta_{\mathrm{M}}}{\beta_{\mathrm{m}}}\right)}. \quad (7.14)\end{aligned}$$

Consequently, for the four collections, the following function

$$\Psi_{(K,\alpha)} : \xi \in \mathbb{R}_+^\star \mapsto \xi\sqrt{D(K,\alpha)}\left\{\mathcal{A} + \sqrt{\ln\left(\frac{1}{1\wedge\xi}\right)}\right\}$$

which satisfies condition **[I]** of Theorem 7.2.1 can be considered.

End of the proof:

To continue the proof, we need to find $\xi_\star$ such that $\Psi_{(K,\alpha)}(\xi_\star) = \sqrt{n}\,\xi_\star^2$ to deduce the penalty function. This is equivalent to solving

$$\sqrt{\frac{D(K,\alpha)}{n}}\left\{\mathcal{A} + \sqrt{\ln\left(\frac{1}{1\wedge\xi_\star}\right)}\right\} = \xi_\star.$$

Noticing that the quantity $\tilde{\xi} = \sqrt{\frac{D(K,\alpha)}{n}}\,\mathcal{A}$ satisfies $\tilde{\xi} \le \xi_\star$, we get

$$\xi_\star \le \sqrt{\frac{D(K,\alpha)}{n}}\left\{\mathcal{A} + \sqrt{\ln\left(\frac{1}{1\wedge\tilde{\xi}}\right)}\right\}$$

and thus

$$\xi_\star^2 \le \frac{D(K,\alpha)}{n}\left\{2\,\mathcal{A}^2 + \ln\left(\frac{1}{1\wedge\frac{D(K,\alpha)}{n}\,\mathcal{A}^2}\right)\right\}.$$

Finally, according to the lower bound of penalty functions in Theorem 7.2.1, it remains to define the weights $\rho_{(K,\alpha)}$. The considered weights $\rho_{(K,\alpha)} = D(K,\alpha)$ depend on the model dimension and their sum Υ is equal to 1 since

$$\mathrm{card}\,\{(K,\alpha) \in \mathbb{N}^\star \times \{1,\dots,Q\};\ D(K,\alpha) = D\} \le D$$

and $\sum_{(K,\alpha)} e^{-\rho_{(K,\alpha)}} \le \sum_{D\ge 1} D\,e^{-D} \le 1$. Therefore according to Theorem 7.2.1, if the penalty function satisfies the inequality

$$\mathrm{pen}(K,\alpha) \ge \kappa\frac{D(K,\alpha)}{n}\left\{1 + 2\,\mathcal{A}^2 + \ln\left(\frac{1}{1\wedge\frac{D(K,\alpha)}{n}\,\mathcal{A}^2}\right)\right\},$$

a minimizer $(\hat{K},\hat{\alpha})$ of $\mathrm{crit}(K,\alpha) = \gamma_n(\hat{s}_{(K,\alpha)}) + \mathrm{pen}(K,\alpha)$ on $\mathcal{M}$ exists and

$$\mathbb{E}\left[d_H^2(s,\hat{s}_{(\hat{K},\hat{\alpha})})\right] \le C\left[\inf_{(K,\alpha)\in\mathcal{M}}\{\mathrm{KL}(s,\mathcal{S}_{(K,\alpha)}) + \mathrm{pen}(K,\alpha)\} + \frac{1}{n}\right].$$

7.A.2 Proof of Theorem 7.3.2

Apart from the weight definition step, the proof of Theorem 7.3.2 is the same as in the ordered case. The following Lemma is used to define weights for this richer family. Recall that $D(K,\alpha)$ denotes the dimension of $\mathcal{S}_{(K,\mathbf{v})}$ and is given for each collection in Section 7.2.1.

Lemma 7.A.4. *For the four collections, the quantity* $\operatorname{card}\{(K,\mathbf{v}) \in \mathbb{N}^\star \times \mathcal{V};\ D(K,\alpha) = D\}$ *is upper bounded by*

$$\begin{cases} 2^Q & \text{if } Q \leq \frac{D-1}{2} \\ \left(\frac{2eQ}{D-1}\right)^{\frac{D-1}{2}} & \text{otherwise} \end{cases}.$$

Proof. For the $[L_k B_k]$ collection,

$$\begin{aligned} \operatorname{card}\{(K,\mathbf{v}) \in \mathbb{N}^\star \times \mathcal{V};\ D(K,\alpha) = D\} &= \operatorname{card}[(K,\mathbf{v}) \in \mathbb{N}^\star \times \mathcal{V};\ K\{2\operatorname{card}(\mathbf{v}) + 1\} = D] \\ &= \sum_{K=1}^{\infty} \sum_{\alpha=1}^{Q} \binom{Q}{\alpha} \mathbb{1}_{K(2\alpha+1)=D} \\ &\leq \sum_{\alpha=1}^{\infty} \binom{Q}{\alpha} \mathbb{1}_{\alpha \leq Q \wedge \lfloor \frac{D-1}{2} \rfloor}. \end{aligned}$$

If $Q \leq \lfloor \frac{D-1}{2} \rfloor$, $\sum_{\alpha=1}^{\infty} \binom{Q}{\alpha} \mathbb{1}_{\alpha \leq Q \wedge \lfloor \frac{D-1}{2} \rfloor} = 2^Q$. Otherwise, according to Proposition 2.5 in Massart (2007),

$$\sum_{\alpha=1}^{\infty} \binom{Q}{\alpha} \mathbb{1}_{\alpha \leq Q \wedge \lfloor \frac{D-1}{2} \rfloor} \leq f\left(\left\lfloor \frac{D-1}{2} \right\rfloor\right)$$

where $f(x) = \left(\frac{eQ}{x}\right)^x$ is an increasing function on $[1, Q]$. Noticing that Q is an integer, it leads that

$$\sum_{\alpha=1}^{Q \wedge \lfloor \frac{D-1}{2} \rfloor} \binom{Q}{\alpha} \leq \begin{cases} 2^Q & \text{if } Q \leq \frac{D-1}{2} \\ \left(\frac{2eQ}{D-1}\right)^{\frac{D-1}{2}} & \text{otherwise} \end{cases}.$$

For the $[L_k C_k]$ collection, $\operatorname{card}\left\{(K,\mathbf{v}) \in \mathbb{N}^\star \times \mathcal{V};\ K\left[1 + \operatorname{card}(\mathbf{v}) + \frac{\operatorname{card}(\mathbf{v})\{\operatorname{card}(\mathbf{v})+1\}}{2}\right] = D\right\}$ is upper bounded by

$$\sum_{\alpha=1}^{Q} \binom{Q}{\alpha} \mathbb{1}_{1+\frac{3}{2}\alpha+\frac{\alpha^2}{2} \leq D} \leq \sum_{\alpha=1}^{Q} \binom{Q}{\alpha} \mathbb{1}_{\alpha \leq \frac{D-1}{2}}$$

hence the result is the same as for the $[L_k B_k]$ collection. An analogous proof gives the result for the two other collections $[LB_k]$ and $[LC]$. □

Proposition 7.A.5. *Consider the following weight family* $\left\{\rho_{(K,\mathbf{v})}\right\}_{(K,\mathbf{v}) \in \mathbb{N}^\star \times \mathcal{V}}$ *defined by*

$$\rho_{(K,\mathbf{v})} = \frac{D(K,\alpha)}{2} \ln\left[\frac{8eQ}{\{D(K,\alpha)-1\} \wedge (2Q-1)}\right].$$

Then we have $\sum_{(K,\mathbf{v}) \in \mathbb{N}^\star \times \mathcal{V}} e^{-\rho_{(K,\mathbf{v})}} \leq 2.$

Proof. Let $Y(D) = \exp\left[-\frac{D}{2}\ln\left\{\frac{8eQ}{(D-1)\wedge(2Q-1)}\right\}\right]$. According to Lemma 7.A.4,

$$\begin{aligned}
\sum_{(K,\mathbf{v})\in\mathbb{N}^\star\times\mathcal{V}} e^{-\rho_{(K,\mathbf{v})}} &= \sum_{D=3}^{\infty} Y(D)\,\mathrm{card}\{(K,\mathbf{v});\ D(K,\mathbf{v})=D\} \\
&\leq \sum_{D=3}^{\infty} Y(D)\left\{2^Q \mathbb{1}_{Q\leq\frac{D-1}{2}} + \left(\frac{2eQ}{D-1}\right)^{\frac{D-1}{2}} \mathbb{1}_{\frac{D-1}{2}<Q}\right\} \\
&\leq \sum_{D=3}^{2Q} \exp\left\{-\frac{D}{2}\ln\left(\frac{8eQ}{D-1}\right) + \frac{D-1}{2}\ln\left(\frac{2eQ}{D-1}\right)\right\} \\
&+ \sum_{D=2Q+1}^{\infty} \exp\left\{-\frac{D}{2}\ln\left(\frac{8eQ}{2Q-1}\right) + Q\ln(2)\right\}.
\end{aligned}$$

For the term in the exponential function of the first sum,

$$\begin{aligned}
-\frac{D}{2}\ln\left(\frac{8eQ}{D-1}\right) + \frac{D-1}{2}\ln\left(\frac{2eQ}{D-1}\right) &= -\frac{D}{2}\ln(4) - \frac{1}{2}\ln\left(\frac{2eQ}{D-1}\right) \\
&\leq -(D-1)\ln(2)
\end{aligned}$$

since $D \leq 2Q$. For the term in the exponential function of second sum, since $D \geq 2Q+1$,

$$\begin{aligned}
-\frac{D}{2}\ln\left(\frac{8eQ}{2Q-1}\right) + Q\ln(2) &= -\frac{3D}{2}\ln(2) + Q\ln(2) - \frac{D}{2}\ln\left(\frac{eQ}{2Q-1}\right) \\
&\leq \left(Q - \frac{D-1}{2}\right)\ln(2) - (D-1)\ln(2) \\
&\leq -(D-1)\ln(2).
\end{aligned}$$

Then

$$\begin{aligned}
\sum_{(K,\mathbf{v})\in\mathbb{N}^\star\times\mathcal{V}} e^{-\rho_{(K,\mathbf{v})}} &\leq \sum_{D=3}^{\infty}\left(\frac{1}{2}\right)^{D-1} \\
&\leq 2.
\end{aligned}$$

□

7.B Tools: bound on bracketing entropies of mixture density families

7.B.1 Control of the bracketing entropy for the $[L_k B_k]$ collection

In this section, we consider the case of the $[L_k B_k]$ Gaussian mixture family (see the description in Section 7.2.1). The following proposition gives an upper bound of the bracketing

entropy of the two families $\mathcal{F}_{(\alpha)}$ and $\mathcal{G}_{(\alpha)}$ defined by (7.4) and (7.2) respectively. It allows us to deduce an upper bound of the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$ according to Inequality (7.9).

Proposition 7.B.1. *Set* $c_1 = 5\left(1-2^{-\frac{1}{4}}\right)/8$. *For all* $\varepsilon \in (0,1]$,

$$\mathcal{H}_{[.]}(\varepsilon, \mathcal{F}_{(\alpha)}, d_H) \leq \alpha \ln\left(2\,a\sqrt{\frac{2}{c_1\,\lambda_m}}\right) + \alpha \ln\left(8\frac{\lambda_M}{\lambda_m}\right) + 2\alpha\ln(\sqrt{2}\,Q) + 2\alpha\ln\left(\frac{1}{\varepsilon}\right) \quad (7.15)$$

and

$$\mathcal{H}_{[.]}(\varepsilon, \mathcal{G}_{(\alpha)}, d_H) \leq \ln\left(8\frac{\lambda_M}{\lambda_m}\right) + \ln\left(\sqrt{2}\,Q\right) + \ln\left(\frac{1}{\varepsilon}\right). \quad (7.16)$$

Thus,

$$\begin{aligned}\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\alpha)}, d_H) \;\leq\; & C(K) + \alpha(2K\alpha+1)\ln(9\sqrt{2}\,Q) + (K\alpha+1)\ln\left(8\frac{\lambda_M}{\lambda_m}\right) \\ & + K\alpha\ln\left(a\sqrt{\frac{8}{c_1\lambda_m}}\right) + D(K,\alpha)\ln\left(\frac{1}{\varepsilon}\right) \end{aligned} \quad (7.17)$$

where $C(K) = \ln(K) + \frac{K}{2}\ln(2\pi e) + (K-1)\ln(9)$.

Proof. According to Assertion (7.9), the upper bound of the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$, given by Inequality (7.17), is deduced from upper bounds of the bracketing entropy of $\mathcal{F}_{(\alpha)}$ and $\mathcal{G}_{(\alpha)}$, respectively expressed in Inequalities (7.15) and (7.16). These two inequalities are now proved.

The proof of Inequality (7.15) is adapted from Genovese and Wasserman (2000) who prove similar results for unidimensional Gaussian mixture families. The main idea is to define a lattice over the parameter space $\mathcal{B} = \{(\mu, \sigma_1^2, \ldots, \sigma_\alpha^2) \in [-a,a]^\alpha \times [\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]^\alpha\}$ and next to deduce a bracket covering of $\mathcal{F}_{(\alpha)}$ according to the Hellinger distance.

First, consider $\varepsilon \in (0,1]$ and $\delta = \varepsilon/(\sqrt{2}Q)$. For all $j \in \{2,\ldots,r\}$, set

$$b_j^2 = (1+\delta)^{1-\frac{j}{2}}\,\lambda_{\mathrm{M}}$$

with $r = \left\lceil 2\,\frac{\ln\left\{\frac{\lambda_{\mathrm{M}}(1+\delta)}{\lambda_{\mathrm{m}}}\right\}}{\ln(1+\delta)}\right\rceil$ in order to have $b_r^2 \leq \lambda_m < \lambda_M = b_2^2$ ($\lceil h \rceil$ denotes the smallest integer greater than or equal to h). Then, for all $J = (j(1),\ldots,j(\alpha)) \in \{2,\ldots,r\}^\alpha$, a diagonal matrix B_J is defined by

$$B_J = \mathrm{diag}(b_{j(1)}^2, \ldots, b_{j(\alpha)}^2)\ .$$

We also consider vectors

$$\nu_J = (\nu_1^{(J)}, \ldots, \nu_\alpha^{(J)}) \in [-a,a]^\alpha$$

such that

$$\forall q \in \{1,\ldots,\alpha\},\ \nu_q^{(J)} = \sqrt{c_1\,\lambda_{\mathrm{M}}}\,\delta\,(1+\delta)^{\frac{1-j(q)}{4}}\,s_q,$$

where $s_q \in \mathbb{Z} \cap [-A, A]$ with $A = \left\lfloor \frac{a\,\delta^{-1}\,(1+\delta)^{-\frac{1-j(q)}{4}}}{\sqrt{c_1\,\lambda_{\mathrm{M}}}} \right\rfloor$. Thus, the set $\mathcal{R}(\varepsilon, \alpha)$ of all such couples (ν_J, B_J) forms a lattice on $\mathcal{B}$.

This set $\mathcal{R}(\varepsilon, \alpha)$ allows to construct brackets that cover $\mathcal{F}_{(\alpha)}$. For a function $f(.) = \Phi(.|\mu, \Sigma)$ of $\mathcal{F}_{(\alpha)}$, the two following functions are considered:

$$\left\{ \begin{array}{l} l(x) = (1+\delta)^{-\alpha}\Phi(x|\nu_J, (1+\delta)^{-\frac{1}{4}}\,B_{J+1}) \\ u(x) = (1+\delta)^{\alpha}\Phi(x|\nu_J, (1+\delta)\,B_J). \end{array} \right.$$

The index set $J = (j(1), \ldots, j(\alpha))$ is taken to satisfy $b^2_{j(q)+1} \leq \sigma^2_q \leq b^2_{j(q)}$ for all q in $\{1, \ldots, \alpha\}$ and ν_J can be chosen such that

$$(\mu - \nu_J)' B^{-1}_{J+1}(\mu - \nu_J) \leq c_1\,\alpha\,\delta^2 \tag{7.18}$$

where $J + 1 := (j(1) + 1, \ldots, j(\alpha) + 1)$. Then we check that the bracket $[l, u]$ contains f. Inequality (7.18) implies that

$$(\mu - \nu_J)' B^{-1}_{J}(\mu - \nu_J) \leq \frac{\alpha}{4}\delta^2. \tag{7.19}$$

The use of Corollary 7.D.2, which allows to bound the ratio of two Gaussian densities with diagonal variance matrices, together with (7.19) leads to

$$\begin{array}{rcl} \dfrac{f(x)}{u(x)} & = & \dfrac{\Phi(x|\mu, B)}{(1+\delta)^{\alpha}\,\Phi(x|\nu_J, (1+\delta)\,B_J)} \\ & \leq & (1+\delta)^{-\frac{\alpha}{4}} \exp\left[\dfrac{1}{2\,\delta}\,(\mu - \nu_J)'\,B_J^{-1}\,(\mu - \nu_J)\right] \\ & \leq & 1. \end{array}$$

The function $h : \delta \mapsto 1 - (1+\delta)^{-\frac{1}{4}}$ being concave, it yields $1 - (1+\delta)^{-\frac{1}{4}} \geq \delta(1 - 2^{-\frac{1}{4}})$. With Corollary 7.D.2 and (7.18), this shows that $l \leq f$ since

$$\begin{array}{rcl} \dfrac{l(x)}{f(x)} & = & \dfrac{(1+\delta)^{-\alpha}\Phi(x|\nu_J, (1+\delta)^{-\frac{1}{4}}\,B_{J+1})}{\Phi(x|\mu, B)} \\ & \leq & (1+\delta)^{-\frac{5\alpha}{8}} \exp\left[\dfrac{(\mu - \nu_J)'\,B^{-1}_{J+1}(\mu - \nu_J)}{2[1 - (1+\delta)^{-\frac{1}{4}}]}\right] \\ & \leq & 1. \end{array}$$

Therefore, $[l, u]$ contains the function f. To prove that $[l, u]$ is an ε-bracket, it remains to check that $d_H(l, u) \leq \varepsilon$. According to Corollary 7.D.4,

$$\begin{array}{rcl} d^2_H(l, u) & = & d^2_H\left((1+\delta)^{-\alpha}\Phi(.|\nu_J, (1+\delta)^{-\frac{1}{4}}\,B_{J+1}), (1+\delta)^{\alpha}\Phi(.|\nu_J, (1+\delta)\,B_J)\right) \\ & = & (1+\delta)^{-\alpha} + (1+\delta)^{\alpha} - 2\left\{\dfrac{2}{(1+\delta)^{-\frac{7}{8}} + (1+\delta)^{\frac{7}{8}}}\right\}^{\frac{\alpha}{2}} \\ & = & \underbrace{2\cosh(\alpha \ln[1+\delta]) - 2}_{(i)} + \underbrace{2 - 2\left[\cosh\left\{\dfrac{7}{8}\ln(1+\delta)\right\}\right]^{-\frac{\alpha}{2}}}_{(ii)}. \end{array}$$

The upper bounds of terms (i) and (ii) separately lead to

$$\begin{aligned} d_H^2(l,u) &\leq \left\{\sinh(1)+\frac{49}{128}\right\}\alpha^2\delta^2 \\ &\leq 2\,\alpha^2\delta^2 \\ &\leq \varepsilon^2. \end{aligned}$$

Consequently, the parameter family $\mathcal{R}(\varepsilon,\alpha)$ induces an ε-bracketing family over $\mathcal{F}_{(\alpha)}$.

An upper bound of the bracketing number of $\mathcal{F}_{(\alpha)}$ is then deduced from an upper bound of the cardinal of $\mathcal{R}(\varepsilon,\alpha)$

$$\begin{aligned} \mathcal{N}_{[.]}(\varepsilon,\mathcal{F}_{(\alpha)},d_H) &\leq \mathrm{Card}(\mathcal{R}(\varepsilon,\alpha)) \\ &\leq \sum_{J\in\{2,\ldots,r\}^\alpha}\prod_{q=1}^{\alpha}\left\{\frac{2\,a}{\sqrt{c_1\,\lambda_{\mathrm{M}}}\,\delta\,(1+\delta)^{\frac{1-j(q)}{4}}}\right\} \\ &\leq \left\{\frac{2\,a(1+\delta)^{\frac{r-1}{4}}}{\sqrt{c_1\,\lambda_{\mathrm{M}}}\delta}\right\}^{\alpha}(r-1)^{\alpha}. \end{aligned}$$

According to the definition of r, $(1+\delta)^{\frac{r-1}{4}} \leq \sqrt{\lambda_{\mathrm{M}}(1+\delta)/\lambda_{\mathrm{m}}}$. Hence,

$$\begin{aligned} \mathcal{N}_{[.]}(\varepsilon,\mathcal{F}_{(\alpha)},d_H) &\leq \left(\frac{2\,a}{\delta}\sqrt{\frac{1+\delta}{c_1\,\lambda_{\mathrm{m}}}}\right)^{\alpha}\left[2\,\frac{\ln\left\{\frac{\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}(1+\delta)\right\}}{\ln(1+\delta)}\right]^{\alpha} \\ &\leq \left(\frac{2\sqrt{2}\,a}{\sqrt{c_1\lambda_{\mathrm{m}}}}\right)^{\alpha}\left(\frac{8\,\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)^{\alpha}\delta^{-(2\alpha)} \\ &\leq \left(\frac{2\sqrt{2}\,a}{\sqrt{c_1\lambda_{\mathrm{m}}}}\right)^{\alpha}\left(\frac{8\,\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)^{\alpha}\left(\frac{\sqrt{2}Q}{\varepsilon}\right)^{2\alpha} \end{aligned}$$

that implies Inequality (7.15).

Using a similar proof, the upper bound of the bracketing entropy of $\mathcal{G}_{(\alpha)}$ given by Inequality (7.16) is obtained. To check this result, the variance family

$$\{b_j^2 = (1+\delta)^{1-\frac{j}{2}}\,\lambda_{\mathrm{M}},\ \forall\, 2\leq j\leq r\}$$

and brackets $[\tilde{l},\tilde{u}]$ defined on $\mathbb{R}^{Q-\alpha}$ by

$$\begin{cases} \tilde{l}(x) = (1+\delta)^{-(Q-\alpha)}\,\Phi(x|0,(1+\delta)^{-\frac{1}{4}}\,b_{j+1}^2 I_{Q-\alpha}) \\ \tilde{u}(x) = (1+\delta)^{Q-\alpha}\,\Phi(x|0,(1+\delta)\,b_j^2 I_{Q-\alpha}) \end{cases}$$

are considered. □

7.B.2 Control of the bracketing entropy for the $[L_kC_k]$ collection

We now consider the case of the $[L_kC_k]$ Gaussian mixture collection. The following proposition gives an upper bound of the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$, based on the Gaussian density family $\mathcal{F}_{(\alpha)}$ defined by (7.5).

Proposition 7.B.2. *For all $\varepsilon \in (0,1]$,*

$$\begin{aligned}\mathcal{H}_{[.]}\left(\varepsilon, \mathcal{F}_{(\alpha)}, d_H\right) &\leq \frac{\alpha(\alpha+1)}{2}\ln\left(\frac{6\sqrt{3}\,\lambda_M}{\lambda_m}\right) + \alpha\ln\left(\frac{6\,a}{\sqrt{\lambda_m}}\right) \\ &+ \left\{\frac{\alpha(\alpha+1)}{2}+\alpha\right\}\ln(Q^2) + \left\{\frac{\alpha(\alpha+1)}{2}+\alpha\right\}\ln\left(\frac{1}{\varepsilon}\right). \quad (7.20)\end{aligned}$$

Thus,

$$\begin{aligned}\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\alpha)}, d_H) &\leq C(K) + \ln\left(\frac{24\sqrt{2}\;\lambda_M}{\lambda_m}\right) + K\frac{\alpha(\alpha+1)}{2}\ln\left(\frac{54\sqrt{3}\,\lambda_M}{\lambda_m}\right) + K\alpha\ln\left(\frac{54\,a}{\sqrt{\lambda_m}}\right) \\ &+ V(K,\alpha)\ln(Q^2) + D(K,\alpha)\ln\left(\frac{1}{\varepsilon}\right) \quad (7.21)\end{aligned}$$

where $C(K) = \ln(K) + \frac{K}{2}\ln(2\pi e) + (K-1)\ln(9)$ and $V(K,\alpha) = K\,\frac{\alpha(\alpha+1)}{2} + K\alpha + 1$.

The result (7.20) together with Inequality (7.9) and the upper bound of $\mathcal{G}_{(\alpha)}$ (see Inequality (7.16)) gives the upper bound (7.21). To prove Inequality (7.20), the method used in the diagonal case cannot be extended to this general situation. Considering the eigenvalue decomposition of the variance matrices, a countable covering on the spectrum could be build as in the diagonal case. An explicit countable covering over the orthogonal matrix set is also necessary to obtain an upper bound of the bracketing entropy of $\mathcal{F}_{(\alpha)}$. Nevertheless, this last point is tricky thus an alternative method is proposed. It consists of defining an adequate covering over the space $\mathcal{D}^+_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}})$ with respect to the uniform norm, and then using it to construct a bracket covering of $\mathcal{F}_{(\alpha)}$. The following notation is used for matrix norms: $\|B\|_\infty = \max_{1\leq i,j\leq\alpha}|B_{ij}|$ and $|||B||| = \sup_{\|x\|_2=1}|x'Bx| = \sup_{\lambda\in \mathrm{vp}(B)}|\lambda|$ where $\mathrm{vp}(B)$ denotes the spectrum of B.

The variance matrix lattice

Let $\beta > 0$ and let $\mathcal{R}(\beta)$ be a β-covering on $\mathcal{D}^+_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}})$ for the uniform norm $\|.\|_\infty$, composed of symmetric matrices and defined by

$$\mathcal{R}(\beta) = \left\{A = (A_{ij})_{1\leq i,j\leq\alpha};\; A_{ij} = a_{ij}\beta;\; a_{ij} = a_{ji} \in \mathbb{Z} \cap \left[-\left\lfloor\frac{\lambda_{\mathrm{M}}}{\beta}\right\rfloor, \left\lfloor\frac{\lambda_{\mathrm{M}}}{\beta}\right\rfloor\right]\right\}.$$

Thus, for all Σ in $\mathcal{D}^+_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}})$, there exists A in $\mathcal{R}(\beta)$ such that

$$\|A - \Sigma\|_\infty \leq \beta. \quad (7.22)$$

The following lemma allows to compare the eigenvalues of Σ with respect to those of its associated matrix A.

Lemma 7.B.3. *Let $\Sigma \in \mathcal{D}^+_{(\alpha)}(\lambda_m, \lambda_M)$ and $A \in \mathcal{R}(\beta)$ such that $\|\Sigma - A\|_\infty \leq \beta$. Let $\lambda_1, \ldots, \lambda_\alpha$ and $\tau_1, \ldots, \tau_\alpha$ be respectively the eigenvalues of Σ and A, ranked in increasing order and counted with their multiplicity.Then, for all $q \in \{1, \ldots, \alpha\}$,*

$$\tau_q - \beta\alpha \leq \lambda_q \leq \tau_q + \beta\alpha.$$

Proof. Since $\|\Sigma - A\|_\infty \leq \beta$, we have $|||\Sigma - A||| \leq \beta\alpha$. Moreover, according to Theorem of Rayleigh, given for instance in Serre (2002, Theorem 3.3.2 p49),

$$\lambda_q = \min_{\dim(F)=q} \max_{x \in F\setminus\{0\}} \frac{x'\Sigma x}{\|x\|_2^2} \text{ and } \tau_q = \min_{\dim(F)=q} \max_{x \in F\setminus\{0\}} \frac{x'Ax}{\|x\|_2^2}$$

where F is a linear subspace of $\mathbb{R}^\alpha$. Then, for all $q \in \{1, \ldots, \alpha\}$, $\tau_q - \beta\alpha \leq \lambda_q \leq \tau_q + \beta\alpha$. □

Covering $\mathcal{F}_{(\alpha)}$ with a family of ε-brackets

Based on the set $\mathcal{R}(\beta)$, ε-brackets for the Gaussian density family $\mathcal{F}_{(\alpha)}$ are now constructed. Consider $f = \Phi(.|\,\mu, \Sigma)$ be a function of $\mathcal{F}_{(\alpha)}$ with $\mu \in [-a, a]^\alpha$ and $\Sigma \in \mathcal{D}^+_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}})$. For $\beta > 0$, there exists a matrix $A \in \mathcal{R}(\beta)$ such that $\|A - \Sigma\|_\infty \leq \beta$ according to (7.22). Then the two following functions are considered

$$u(x) = (1 + 2\delta)^\alpha \, \Phi\left(x|\,\nu, (1 + \delta)A\right) \tag{7.23}$$

and

$$l(x) = (1 + 2\delta)^{-\alpha} \, \Phi\left(x|\,\nu, (1 + \delta)^{-1}A\right) \tag{7.24}$$

where the vector ν and the positive number δ are adjusted later in order that $[l, u]$ is an ε-bracket of $\mathcal{F}_{(\alpha)}$ containing the function f.

Next lemma allows to fulfill hypothesis necessary to use Proposition 7.D.1. The resulting bounds on Gaussian density ratios are given in Lemma 7.B.5.

Lemma 7.B.4. *Assume that $0 < \beta < \lambda_m/(3\alpha)$ and set $\delta = 3\beta\alpha/\lambda_m$. Then, $(1 + \delta)A - \Sigma$ and $\Sigma - (1 + \delta)^{-1}A$ are both positive definite matrices. Moreover, for all x in $\mathbb{R}^\alpha$,*

$$x'\{(1 + \delta)A - \Sigma\}x \geq \beta\alpha\|x\|_2^2 \tag{7.25}$$

and

$$x'\{\Sigma - (1 + \delta)^{-1}A\}x \geq \beta\alpha\|x\|_2^2. \tag{7.26}$$

Proof. For all $x \neq 0$, since $|||A - \Sigma||| \leq \alpha\beta$,

$$\begin{aligned} x'\{(1 + \delta)A - \Sigma\}x &= (1 + \delta)x'(A - \Sigma)x + \delta x'\Sigma x \\ &\geq -(1 + \delta)\,|||A - \Sigma|||\,\|x\|_2^2 + \delta\,\lambda_{\mathrm{m}}\,\|x\|_2^2 \\ &\geq \{\delta\,\lambda_{\mathrm{m}} - (1 + \delta)\alpha\beta\}\|x\|_2^2 \\ &\geq \left(\frac{2}{3}\delta\,\lambda_{\mathrm{m}} - \alpha\beta\right)\|x\|_2^2 \end{aligned}$$

because $\alpha\beta \leq \lambda_{\mathrm{m}}/3$. Then $x'\{(1+\delta)A - \Sigma\}x \geq \alpha\beta\|x\|_2^2 > 0$ according to the definition of δ. Similarly,

$$\begin{aligned} x'\{\Sigma - (1+\delta)^{-1}A\}x &= (1+\delta)^{-1}x'(\Sigma - A)x + \{1-(1+\delta)^{-1}\}x'\Sigma x \\ &\geq \left(\frac{\delta\,\lambda_{\mathrm{m}} - \alpha\beta}{1+\delta}\right)\|x\|_2^2 \\ &\geq \frac{2\alpha\beta}{1+\delta}\|x\|_2^2 \\ &\geq \alpha\beta\|x\|_2^2 > 0. \end{aligned}$$

□

Lemma 7.B.5. *Assume that $\beta < \lambda_m/(3\alpha)$ and set $\delta = 3\beta\alpha/\lambda_m$. Then,*

$$\frac{f(x)}{u(x)} \leq (1+2\delta)^{-\frac{\alpha}{2}} \exp\left(\frac{\|\mu-\nu\|_2^2}{2\beta\alpha}\right)$$

and

$$\frac{l(x)}{f(x)} \leq (1+2\delta)^{-\frac{\alpha}{2}} \exp\left(\frac{\|\mu-\nu\|_2^2}{2\beta\alpha}\right).$$

Proof. According to Proposition 7.D.1, since $(1+\delta)A - \Sigma$ is a positive definite matrix from Lemma 7.B.4,

$$\frac{f(x)}{u(x)} \leq (1+2\delta)^{-1}\sqrt{\frac{|(1+\delta)A|}{|\Sigma|}} \exp\left[\frac{1}{2}(\mu-\nu)'\{(1+\delta)A-\Sigma\}^{-1}(\mu-\nu)\right].$$

Inequality (7.25) implies that $|||\{(1+\delta)A-\Sigma\}^{-1}||| = \{\inf\lambda\}^{-1} \leq (\beta\alpha)^{-1}$ where the infimum is taken over all eigenvalues of $(1+\delta)A - \Sigma$. Then, since

$$(\mu-\nu)'\{(1+\delta)A-\Sigma\}^{-1}(\mu-\nu) \leq |||\{(1+\delta)A-\Sigma\}^{-1}|||\,\|\mu-\nu\|_2^2,$$

this leads to

$$(\mu-\nu)'\{(1+\delta)A-\Sigma\}^{-1}(\mu-\nu) \leq \frac{\|\mu-\nu\|_2^2}{\alpha\beta}.$$

Moreover, according to Lemma 7.B.3,

$$\begin{aligned} \frac{|(1+\delta)A|}{|\Sigma|} &= (1+\delta)^{\alpha}\prod_{q=1}^{\alpha}\frac{\tau_q}{\lambda_q} \\ &\leq (1+\delta)^{\alpha}\prod_{q=1}^{\alpha}\left(1+\frac{\beta\alpha}{\lambda_q}\right) \\ &\leq (1+\delta)^{\alpha}\left(1+\frac{\beta\alpha}{\lambda_{\mathrm{m}}}\right)^{\alpha} \\ &\leq (1+2\delta)^{\alpha}. \end{aligned}$$

Then

$$\frac{f(x)}{u(x)} \leq (1+2\delta)^{-\frac{\alpha}{2}} \exp\left(\frac{\|\mu-\nu\|_2^2}{2\beta\alpha}\right).$$

Similarly, using Proposition 7.D.1, (7.26) and Lemma 7.B.3, we obtain

$$\frac{l(x)}{f(x)} \leq (1+2\delta)^{-\frac{\alpha}{2}} \exp\left(\frac{\|\mu-\nu\|_2^2}{2\beta\alpha}\right).$$

□

Next proposition terminates the construction of an ε-bracket covering of $\mathcal{F}_{(\alpha)}$.

Proposition 7.B.6. *For all $\varepsilon \in (0,1]$, we define $\delta = \varepsilon/(\sqrt{3}\,\alpha)$ and $\beta = \lambda_m\,\varepsilon/(3\sqrt{3}\alpha^2)$. The following set*

$$\left\{[l,u];\ \begin{array}{l} u(x) = (1+2\delta)^{\alpha}\,\Phi\left(x|\,\nu,(1+\delta)A\right) \\ l(x) = (1+2\delta)^{-\alpha}\,\Phi\left(x|\,\nu,(1+\delta)^{-1}A\right) \end{array}\ ;\ A \in \mathcal{R}(\beta),\ \nu \in \mathcal{X}(\varepsilon,a,\lambda_m,\alpha)\right\}$$

where

$$\mathcal{X}(\varepsilon,a,\lambda_m,\alpha) = \left\{\nu=(\nu_1,\ldots,\nu_\alpha);\ \nu_q = \frac{\sqrt{\lambda_m}\,\varepsilon}{3\alpha}\,s_q;\ s_q \in \mathbb{Z} \cap \left[-\left\lfloor\frac{3\,a\,\alpha}{\sqrt{\lambda_m}\,\varepsilon}\right\rfloor, \left\lfloor\frac{3\,a\,\alpha}{\sqrt{\lambda_m}\,\varepsilon}\right\rfloor\right]\right\},$$

is an ε-bracket set over $\mathcal{F}_{(\alpha)}$.

Proof. Let $f(x) = \Phi(x|\mu,\Sigma)$ be a function of $\mathcal{F}_{(\alpha)}$ where $\mu \in [-a,a]^\alpha$ and $\Sigma \in \mathcal{D}^+_{(\alpha)}(\lambda_{\rm m},\lambda_{\rm M})$. There exists A in $\mathcal{R}(\beta)$ such that $\|\Sigma - A\|_\infty \leq \beta$ and ν in $\mathcal{X}(\varepsilon,a,\lambda_{\rm m},\alpha)$ satisfying, for all q in $\{1,\ldots,\alpha\}$, $|\mu_q-\nu_q| \leq \sqrt{\lambda_{\rm m}}\varepsilon/(3\alpha)$. Consider the two associated functions l and u defined in (7.23) and (7.24) respectively. Since $\|\mu-\nu\|_2^2 \leq \lambda_{\rm m}\,\varepsilon^2/(9\alpha)$, using Lemma 7.B.5,

$$\frac{f(x)}{u(x)} \leq (1+2\delta)^{-\frac{\alpha}{2}} \exp\left(\frac{\sqrt{3}\,\varepsilon}{6}\right).$$

Thus, noting that for all x in $[0,2]$, $\ln(1+x) \geq x/2$, it leads to

$$\begin{aligned} \ln\left\{\frac{f(x)}{u(x)}\right\} &\leq -\frac{\alpha}{2}\ln\left(1+\frac{2\,\varepsilon}{\sqrt{3}\alpha}\right) + \frac{\sqrt{3}\,\varepsilon}{6} \\ &\leq -\frac{\alpha}{2}\frac{\varepsilon}{\sqrt{3}\alpha} + \frac{\varepsilon}{2\sqrt{3}} \leq 0. \end{aligned}$$

Similarly, $\ln\{l(x)/f(x)\} \leq 0$ and thus for all $x \in \mathbb{R}^\alpha$, $l(x) \leq f(x) \leq u(x)$. It remains to bound the size of bracket $[l,u]$ with respect to Hellinger distance. According to Proposition.7.D.3,

$$\begin{aligned} d_H^2(l,u) &= (1+2\delta)^{\alpha} + (1+2\delta)^{-\alpha} - \left\{2 - d_H^2(\Phi(.|\nu,(1+\delta)A),\Phi(.|\nu,(1+\delta)^{-1}A))\right\} \\ &= 2\left(\cosh\{\alpha\ln(1+2\delta)\} - 1 + 1 - [\cosh\{\ln(1+\delta)\}]^{-\frac{\alpha}{2}}\right) \\ &\leq 2\left(\sinh(1)\alpha^2\delta^2 + \frac{1}{4}\alpha^2\delta^2\right) \\ &\leq 3\alpha^2\delta^2 = \varepsilon^2. \end{aligned}$$

□

Proof of Proposition 7.B.2

Proof. Since the set of ε-brackets over $\mathcal{F}_{(\alpha)}$, described in the Proposition 7.B.6 is totally defined by the parameter spaces $\mathcal{R}(\beta)$ and $\mathcal{X}(\varepsilon, a, \lambda_{\mathrm{m}}, \alpha)$, an upper bound of the bracketing entropy of $\mathcal{F}_{(\alpha)}$ is deduced from an upper bound of the two set cardinals.

$$\begin{aligned}
\mathcal{N}_{[.]}\left(\varepsilon, \mathcal{F}_{(\alpha)}, d_H\right) &\leq \operatorname{card}\{\mathcal{R}(\beta)\} \times \operatorname{card}\{\mathcal{X}(\varepsilon, a, \lambda_{\mathrm{m}}, \alpha)\} \\
&\leq \left(\frac{2\,\lambda_{\mathrm{M}}}{\beta}\right)^{\frac{\alpha(\alpha+1)}{2}} \left(\frac{6\,a\alpha}{\sqrt{\lambda_{\mathrm{m}}}\,\varepsilon}\right)^{\alpha} \\
&\leq \left(\frac{6\sqrt{3}\,\lambda_{\mathrm{M}}\,\alpha^2}{\varepsilon\,\lambda_{\mathrm{m}}}\right)^{\frac{\alpha(\alpha+1)}{2}} \left(\frac{6\,a\alpha}{\sqrt{\lambda_{\mathrm{m}}}\,\varepsilon}\right)^{\alpha}.
\end{aligned}$$

Thus, since $\ln(\alpha)$ and $\ln(\alpha^2)$ are bounded by $\ln(Q^2)$,

$$\begin{aligned}
\mathcal{H}_{[.]}\left(\varepsilon, \mathcal{F}_{(\alpha)}, d_H\right) &\leq \frac{\alpha(\alpha+1)}{2} \ln\left(\frac{6\sqrt{3}\,\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right) + \alpha \ln\left(\frac{6\,a}{\sqrt{\lambda_{\mathrm{m}}}}\right) \\
&+ \left\{\frac{\alpha(\alpha+1)}{2} + \alpha\right\} \ln(Q^2) + \left\{\frac{\alpha(\alpha+1)}{2} + \alpha\right\} \ln\left(\frac{1}{\varepsilon}\right).
\end{aligned}$$

□

7.B.3 Control of the bracketing entropy for the $[LB_k]$ collection

In this section, we want to upper bound the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$ according to Inequality (7.10). Recall that it is sufficient to control the bracketing entropy of the family $\mathcal{F}_{(K,\alpha)}$, defined by (7.6), and used this of the family $\mathcal{G}_{(\alpha)}$ given by (7.16).

Proposition 7.B.7. *For all $\varepsilon \in (0, 1]$, the bracketing number of the set $\mathcal{F}_{(K,\alpha)}$ is upper bounded by*

$$\mathcal{N}_{[.]}\left(\mathcal{F}_{(K,\alpha)}\right) \leq A_2(a, \lambda_m, \lambda_M, \beta_m, \beta_M, K, \alpha) \left(\frac{\alpha}{\varepsilon}\right)^{K(2\alpha-1)+1}$$

where $c_1 = 1 - 2^{-\frac{1}{4}}$ and

$$A_2(a, \lambda_m, \lambda_M, \beta_m, \beta_M, K, \alpha) = \left(\frac{24\,\lambda_M}{\lambda_m}\right)^{K(\alpha-1)} \left(\frac{6a}{\sqrt{\beta_M\,c_1}}\right)^{K\alpha} \left(\frac{24\,\beta_M}{\beta_m}\right)^{\frac{K\alpha}{2}+1}.$$

Hence

$$\begin{aligned}
\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\alpha)}, d_H) &\leq C(K) + \{K(\alpha-1)+1\} \ln\left(\frac{216\sqrt{2}\,\lambda_M}{\lambda_m}\right) + K\alpha \ln\left(\frac{54a}{\sqrt{\beta_M\,c_1}}\right) \\
&+ \left(\frac{K\alpha}{2} + 1\right) \ln\left(\frac{24\,\beta_M}{\beta_m}\right) + \{K(2\alpha-1)+2\} \ln(Q) + D(K,\alpha) \ln\left(\frac{1}{\varepsilon}\right)
\end{aligned}$$

where $D(K, \alpha) = 2K\alpha + 1$.

Proof. The proof for the unidimensional case ($\alpha = 1$) is already available in Genovese and Wasserman (2000). Let $\varepsilon \in (0, 1]$ and assume $K \geq 2$ and $\alpha \geq 2$ fixed. Let $\delta = \varepsilon/(3\alpha)$. For $j \in \{2, \ldots, r\}$, we define

$$b_j^2 = (1+\delta)^{1-\frac{j}{2}}\ \lambda_{\mathrm{M}}$$

where $r = \left\lceil 2\,\frac{\ln\left\{\frac{\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}(1+\delta)\right\}}{\ln(1+\delta)} \right\rceil$ in order to have $b_r^2 \leq \lambda_{\mathrm{m}} \leq b_2^2 = \lambda_{\mathrm{M}}$. $\lceil h \rceil$ denotes the smallest integer greater than or equal to h. For $z \in \{0, \ldots, r'\}$ we consider

$$\beta_z = (1+\delta)^{-z}\beta_{\mathrm{M}}$$

where $r' = \left\lceil \frac{\ln\left\{\frac{\beta_{\mathrm{M}}}{\beta_{\mathrm{m}}}\right\}}{\ln(1+\delta)} \right\rceil$ in order to have $\beta_{r'} \leq \beta_{\mathrm{m}} \leq \beta_0 = \beta_{\mathrm{M}}$.
For a vector $J = (j(1), \ldots, j(\alpha-1)) \in \{2, \ldots, r\}^{\alpha-1}$, the diagonal matrices B_J^l and B_J^u are defined by

$$B_J^l = \mathrm{diag}\left(b_{j(1)+1}^2, \ldots, b_{j(\alpha-1)+1}^2, \lambda_{\mathrm{M}}^{1-\alpha}(1+\delta)^{\frac{S_J}{2}-(\alpha-1)}\right)$$

and

$$B_J^u = \mathrm{diag}\left(b_{j(1)}^2, \ldots, b_{j(\alpha-1)}^2, \lambda_{\mathrm{M}}^{1-\alpha}(1+\delta)^{\frac{S_J}{2}-\frac{\alpha-1}{2}}\right)$$

with $S_J = \sum_{q=1}^{\alpha-1} j(q)$. The q^{th} diagonal coefficients of these matrices are denoted $B_{J,q}^l$ and $B_{J,q}^u$ respectively.

First, a function $\Phi(.|\mu, \beta B)$ such that $\beta \in [\beta_m, \beta_{\mathrm{M}}]$, $\mu \in [-a, a]^\alpha$ and $B \in \Delta_{(\alpha)}^1(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}})$ is considered. Let $z \in \{0, \ldots, r'\}$ be the unique integer of $\{0, \ldots, r'\}$ such that $\beta_{z+1} < \beta \leq \beta_z$ and let J be the unique vector of $\{2, \ldots, r\}^{\alpha-1}$ such that $\forall q \in \{1, \ldots, \alpha-1\}$, $B_{J,q}^l \leq B_{qq} \leq B_{J,q}^u$. Hence for all $q \in \{1, \ldots, \alpha\}$,

$$\beta_{z+1}\, B_{J,q}^l \leq \beta\, \Sigma_{qq} \leq \beta_z\, B_{J,q}^u.$$

For a couple (z, J), we also consider a regular lattice of mean vector $\nu^{(z,J)} = \left(\nu_1^{(z,J)}, \ldots, \nu_\alpha^{(z,J)}\right) \in [-a, a]^\alpha$ such that for all $q \in \{1, \ldots, \alpha-1\}$,

$$\nu_q^{(z,J)} = (1+\delta)^{-\frac{j(q)+1}{4}-\frac{z}{2}}\sqrt{\lambda_{\mathrm{M}}\ \beta_{\mathrm{M}}\ c_1}\,\delta s_q,$$

with $s_q \in \{-N_q, \ldots, N_q\}$ where $N_q = \left\lfloor \frac{a(1+\delta)^{\frac{j(q)+1}{4}+\frac{z}{2}}}{\sqrt{\lambda_{\mathrm{M}}\ \beta_{\mathrm{M}}\ c_1}\,\delta} \right\rfloor$,

$$\nu_\alpha^{(z,J)} = (1+\delta)^{\frac{S_J}{4}-\frac{\alpha+z}{2}}\sqrt{\lambda_{\mathrm{M}}^{1-\alpha}\,\beta_{\mathrm{M}}\ c_1}\,\delta s_\alpha,$$

with $s_\alpha \in \{-N_\alpha, \ldots, N_\alpha\}$ where $N_\alpha = \left\lfloor \frac{a(1+\delta)^{\frac{\alpha+z}{2}-\frac{S_J}{4}}}{\sqrt{\lambda_{\mathrm{M}}^{1-\alpha}\,\beta_{\mathrm{M}}\ c_1}\delta} \right\rfloor$ and $c_1 := 1 - 2^{-\frac{1}{4}}$. For a given couple (z, J), this insures that for all vectors $\mu \in [-a, a]^\alpha$, there exists a vector $\nu^{(z,J)}$ of this lattice such that

$$\frac{(1+\delta)^z}{\beta_{\mathrm{M}}\,\lambda_{\mathrm{M}}}\left\{\sum_{q=1}^{\alpha-1}\left(\nu^{(z,J)}-\mu\right)_q^2(1+\delta)^{\frac{j(q)+1}{2}} + \left(\nu^{(z,J)}-\mu\right)_\alpha^2(1+\delta)^{-\frac{S_J}{2}+\alpha}\lambda_{\mathrm{M}}^\alpha\right\} \leq c_1\,\alpha\,\delta^2. \quad (7.27)$$

For a couple (z, J) and a vector $\nu^{(z,J)}$ defined as before, the two following associated functions are considered

$$\left\{ \begin{array}{l} l(x) = (1+\delta)^{-2\alpha}\, \Phi\left(x|\nu^{(z,J)}, (1+\delta)^{-\frac{1}{4}}\, \beta_{z+1}\, B_J^l\right) \\ u(x) = (1+\delta)^{2\alpha}\, \Phi\left(x|\nu^{(z,J)}, (1+\delta)\, \beta_z\, B_J^u\right). \end{array} \right. \tag{7.28}$$

Second, we check that for all $x \in \mathbb{R}^Q$, $l(x) \leq \Phi(x|\mu, \beta B) \leq u(x)$. According to Proposition 7.D.1 which allows to upper bound the ratio of two Gaussian densities, we get

$$\begin{aligned} \frac{\Phi(x)}{u(x)} &\leq (1+\delta)^{-2\alpha} \sqrt{\frac{|(1+\delta)\beta_z B_J^u|}{|\beta B|}}\, \exp\left[\frac{1}{2}(\nu^{(z,J)} - \mu)' \left\{(1+\delta)\beta_z B_J^u - \beta B\right\}^{-1} (\nu^{(z,J)} - \mu)\right] \\ &\leq (1+\delta)^{-\frac{3\alpha+1}{4}}\, \exp\left\{\frac{1}{2\delta}(\nu^{(z,J)} - \mu)'(\beta_z B_J^u)^{-1}(\nu^{(z,J)} - \mu)\right\} \end{aligned}$$

and

$$\begin{aligned} \frac{l(x)}{\Phi(x)} &\leq (1+\delta)^{-2\alpha} \sqrt{\frac{|\beta B|}{|(1+\delta)^{-\frac{1}{4}}\beta_{z+1} B_J^l|}} \\ &\qquad \times \exp\left[\frac{1}{2}(\nu^{(z,J)} - \mu)' \left\{\beta B - (1+\delta)^{-\frac{1}{4}}\beta_{z+1} B_J^l\right\}^{-1} (\nu^{(z,J)} - \mu)\right] \\ &\leq (1+\delta)^{-(\frac{9\alpha}{8}+\frac{1}{4})}\, \exp\left\{\frac{1}{2\delta(1-2^{-\frac{1}{4}})}(\nu^{(z,J)} - \mu)'(\beta_{z+1} B_J^l)^{-1}(\nu^{(z,J)} - \mu)\right\} \end{aligned}$$

using the concavity of the function $\delta \mapsto 1 - (1+\delta)^{-\frac{1}{4}}$. The following inequalities

$$(\nu^{(z,J)} - \mu)' \left(\beta_z B_J^u\right)^{-1} (\nu^{(z,J)} - \mu) \leq \frac{\delta^2}{4}(3\alpha + 1) \tag{7.29}$$

and

$$(\nu^{(z,J)} - \mu)'(\beta_{z+1} B_J^l)^{-1}(\nu^{(z,J)} - \mu) \leq \delta^2(1 - 2^{-\frac{1}{4}})\frac{9\alpha + 2}{8} \tag{7.30}$$

are then sufficient to have $l \leq \Phi \leq u$. We can check that condition (7.27) implies the two inequalities (7.29) and (7.30).

Third, we show that $d_H(u, l) \leq \varepsilon$. According to Proposition 7.D.3, we have that the $d_H^2(l, u) = (1+\delta)^{-2\alpha} + (1+\delta)^{2\alpha} - (\blacksquare)$ where

$$\begin{aligned} (\blacksquare) &= 2^{\frac{\alpha}{2}+1} \left|(1+\delta)\beta_z B_J^u (1+\delta)^{-\frac{1}{4}}\beta_{z+1} B_J^l\right|^{-\frac{1}{4}} \left|\frac{(B_J^u)^{-1}}{(1+\delta)\beta_z} + \frac{(1+\delta)^{\frac{1}{4}}(B_J^l)^{-1}}{\beta_{z+1}}\right|^{-\frac{1}{2}} \\ &= 2\left\{\frac{(1+\delta)^{\frac{11}{8}} + (1+\delta)^{-\frac{11}{8}}}{2}\right\}^{-\frac{\alpha-1}{2}} \left\{\frac{(1+\delta)^{\frac{2\alpha+7}{8}} + (1+\delta)^{-\frac{2\alpha+7}{8}}}{2}\right\}^{-\frac{1}{2}}. \end{aligned}$$

Hence

$$\begin{aligned} d_H^2(l,u) &= [2\cosh\{2\alpha\ln(1+\delta)\}-2] + \left[2-2\left\{\cosh\left(\frac{11}{8}\ln(1+\delta)\right)\right\}^{-\frac{\alpha-1}{2}}\right] \\ &\quad + 2\left\{\cosh\left(\frac{11}{8}\ln(1+\delta)\right)\right\}^{-\frac{\alpha-1}{2}}\left[1-\left\{\cosh\left(\frac{7+2\alpha}{8}\ln(1+\delta)\right)\right\}^{-\frac{1}{2}}\right] \\ &\leq 4\sinh(1)\alpha^2\delta^2 + 2\frac{\alpha-1}{2}\frac{11}{8}\delta^2 + 2\frac{7+2\alpha}{8}\frac{1}{2}\delta^2 \leq 9\alpha^2\delta^2 = \varepsilon^2. \end{aligned}$$

Finally, we can construct an ε-bracket family (with respect to d_H) to cover $\mathcal{F}_{(K,\alpha)}$. Let $(\Phi(.|\mu_1,\beta B_1),\ldots,\Phi(.|\mu_K,\beta B_K))$ be an element of $\mathcal{F}_{(K,\alpha)}$. Let $z\in\{0,\ldots,r'\}$ and $J_1,\ldots,J_K$ in $\{2,\ldots,r\}^{\alpha-1}$ such that for all $k\in\{1,\ldots,K\}$ and all $q\in\{1,\ldots,\alpha\}$,

$$\beta_{z+1}\,B^l_{J_k,q} \leq \beta\,B_{k,qq} \leq \beta_z\,B^u_{J_k,q}.$$

For all k, there exists a vector $\nu^{(z,J_k)}$ such that condition (7.27) is satisfied for the mean vector μ_k. For z, J_k and $\nu^{(zJ_k)}$, the two associated functions defined by (7.28) are denoted u_k and l_k. Then we define $L:=(l_1,\ldots,l_K)$ and $U:=(u_1,\ldots,u_K)$. The set of all such brackets $[L,U]$ covers the family $\mathcal{F}_{(K,\alpha)}$ and is denoted $\mathcal{R}(\varepsilon,K,\alpha)$. An upper bound of the bracketing number of $\mathcal{F}_{(K,\alpha)}$ is thus determined by the computation of the cardinal of $\mathcal{R}(\varepsilon,K,\alpha)$. Then

$$\begin{aligned} \mathcal{N}_{[.]}(\varepsilon,\mathcal{F}_{(K,\alpha)},d_H) &\leq \text{card }\mathcal{R}(K,\varepsilon,\alpha) \\ &\leq \sum_{z=0}^{r'}\left[\sum_J\left\{\prod_{q=1}^{\alpha-1}\frac{2\,a\,(1+\delta)^{\frac{j(q)+1}{4}+\frac{z}{2}}}{\sqrt{\beta_M\,\lambda_{\rm M}\,c_1}\,\delta}\right\}\left\{\frac{2\,a\,(1+\delta)^{\frac{\alpha+z}{2}-\frac{S_J}{4}}}{\sqrt{\beta_{\rm M}\,c_1\,{\lambda_{\rm M}}^{1-\alpha}}\,\delta}\right\}\right]^K \\ &\leq \sum_{z=0}^{r'}\left[(r-1)^{\alpha-1}\left(\frac{2a}{\sqrt{\beta_{\rm M}\,c_1}\delta}\right)^{\alpha}(1+\delta)^{\frac{3\alpha-1}{4}+\frac{z}{2}}\right]^K \\ &\leq (r-1)^{K(\alpha-1)}\left(\frac{2a}{\sqrt{\beta_{\rm M}\,c_1}\delta}\right)^{K\alpha}(1+\delta)^{K(\frac{3\alpha-1}{4})}(1+\delta)^{\frac{K\alpha r'}{2}}(r'+1) \\ &\leq A_2(a,\lambda_{\rm m},\lambda_{\rm M},\beta_{\rm m},\beta_{\rm M},K,\alpha)\left(\frac{\alpha}{\varepsilon}\right)^{K(2\alpha-1)+1} \end{aligned}$$

where

$$A_2(a,\lambda_{\rm m},\lambda_{\rm M},\beta_{\rm m},\beta_{\rm M},K,\alpha) = \left(\frac{24\,\lambda_{\rm M}}{\lambda_{\rm m}}\right)^{K(\alpha-1)}\left(\frac{6a}{\sqrt{\beta_{\rm M}\,c_1}}\right)^{K\alpha}\left(\frac{24\,\beta_{\rm M}}{\beta_{\rm m}}\right)^{\frac{K\alpha}{2}+1},$$

using $1+\delta\leq 2$ and the definitions of r, r' and δ. □

7.B.4 Control of the bracketing entropy for the $[LC]$ collection

This section is devoted to upper bound the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$ for the $[LC]$ Gaussian mixture collection. According to Inequality (7.10), it remains to control the bracketing entropy of the family $\mathcal{F}_{(K,\alpha)}$ defined by (7.7).

Proposition 7.B.8. *For all $\varepsilon \in (0,1]$, the bracketing number of $\mathcal{F}_{(K,\alpha)}$ is controlled by*

$$\mathcal{N}_{[.]}\left(\varepsilon, \mathcal{F}_{(K,\alpha)}, d_H\right) \leq \left(\frac{6\sqrt{3}\,\lambda_M\,\alpha^2}{\lambda_m}\right)^{\frac{\alpha(\alpha+1)}{2}} \left(\frac{6a\alpha}{\sqrt{\lambda_m}}\right)^{K\alpha} \left(\frac{1}{\varepsilon}\right)^{K\alpha+\frac{\alpha(\alpha+1)}{2}}$$

where $K\alpha + \frac{\alpha(\alpha+1)}{2}$ is the dimension of $\mathcal{F}_{(K,\alpha)}$. Hence the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$ is upper bounded by

$$\begin{aligned}\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\alpha)}, d_H) &\leq C(K) + \ln\left(\frac{24\sqrt{2}\;\lambda_M}{\lambda_m}\right) + K\frac{\alpha(\alpha+1)}{2}\ln\left(\frac{54\sqrt{3}\,\lambda_M}{\lambda_m}\right) + K\alpha\ln\left(\frac{54\,a}{\sqrt{\lambda_m}}\right) \\ &+ V(K,\alpha)\ln(Q^2) + D(K,\alpha)\ln\left(\frac{1}{\varepsilon}\right)\end{aligned}$$

where $C(K) = \ln(K) + \frac{K}{2}\ln(2\pi e) + (K-1)\ln(9)$ and $V(K,\alpha) = \frac{\alpha(\alpha+1)}{2} + K\alpha + 1$.

Proof. This result is obtained by considering the following bracket family. Its construction is inspired by the bracket family used in the study of the bracketing entropy of $\mathcal{F}_{(\alpha)}$ for the $[L_k C_k]$ collection (see Appendix 7.B.2).

For all $\varepsilon \in (0,1]$, let $\delta = \frac{\varepsilon}{\sqrt{3}\,\alpha}$ and $\beta = \frac{\lambda_{\mathrm{m}}\,\varepsilon}{3\sqrt{3}\alpha^2}$. The following set

$$\left\{([l_1,u_1],\ldots,[l_K,u_K]);\ \begin{array}{l} u_k(x) = (1+2\delta)^{\alpha}\Phi\left(x|\nu_k,(1+\delta)A\right) \\ l_k(x) = (1+2\delta)^{-\alpha}\Phi\left(x|\nu_k,(1+\delta)^{-1}A\right)\end{array};\ A \in \mathcal{R}(\beta), \nu_k \in \mathcal{X}(\varepsilon,a,\lambda_{\mathrm{m}},\alpha)\right\}$$

where

$$\mathcal{R}(\beta) = \left\{A = (A_{ij})_{1\leq i,j\leq\alpha};\ A_{ij} = a_{ij}\beta;\ a_{ij} = a_{ji} \in \mathbb{Z} \cap \left[-\left\lfloor\frac{\lambda_{\mathrm{M}}}{\beta}\right\rfloor, \left\lfloor\frac{\lambda_{\mathrm{M}}}{\beta}\right\rfloor\right]\right\}$$

and

$$\mathcal{X}(\varepsilon,a,\lambda_{\mathrm{m}},\alpha) = \left\{\nu = (\nu_1,\ldots,\nu_\alpha);\ \nu_q = \frac{\sqrt{\lambda_{\mathrm{m}}}\,\varepsilon}{3\alpha}\,s_q;\ s_q \in \mathbb{Z} \cap \left[-\left\lfloor\frac{3\,a\,\alpha}{\sqrt{\lambda_{\mathrm{m}}}\,\varepsilon}\right\rfloor, \left\lfloor\frac{3\,a\,\alpha}{\sqrt{\lambda_{\mathrm{m}}}\,\varepsilon}\right\rfloor\right]\right\},$$

is an ε-bracket set over $\mathcal{F}_{(K,\alpha)}$. Finally the bracketing number of $\mathcal{F}_{(K,\alpha)}$ is upper bounded by

$$\begin{aligned}\mathcal{N}_{[.]}(\varepsilon, \mathcal{F}_{(K,\alpha)}, d_H) &\leq \mathrm{card}(\mathcal{R}(\beta)) \times \mathrm{card}(\mathcal{X}(\varepsilon,a,\lambda_{\mathrm{m}},\alpha))^K \\ &\leq \left(\frac{2\,\lambda_{\mathrm{M}}}{\beta}\right)^{\frac{\alpha(\alpha+1)}{2}} \times \left(\frac{6a\alpha}{\sqrt{\lambda_{\mathrm{m}}}\varepsilon}\right)^{K\alpha} \\ &\leq \left(\frac{6\sqrt{3}\,\lambda_{\mathrm{M}}\,\alpha^2}{\varepsilon\,\lambda_{\mathrm{m}}}\right)^{\frac{\alpha(\alpha+1)}{2}} \times \left(\frac{6a\alpha}{\sqrt{\lambda_{\mathrm{m}}}\varepsilon}\right)^{K\alpha} \\ &\leq \left(\frac{6\sqrt{3}\,\lambda_{\mathrm{M}}\,\alpha^2}{\lambda_{\mathrm{m}}}\right)^{\frac{\alpha(\alpha+1)}{2}} \times \left(\frac{6a\alpha}{\sqrt{\lambda_{\mathrm{m}}}}\right)^{K\alpha} \times \left(\frac{1}{\varepsilon}\right)^{K\alpha+\frac{\alpha(\alpha+1)}{2}}.\end{aligned}$$

□

7.C Proof of Propositions 7.A.1 and 7.A.2

Proof of Proposition 7.A.1:

According to Theorem 7.A.1, for all $\delta \leq 1$,

$$\mathcal{N}_{[.]}(\delta, \mathcal{L}_{(K,\alpha)}, d_H) \leq K(2\pi e)^{\frac{K}{2}} \left(\frac{3}{\delta}\right)^{K-1} \prod_{k=1}^{K} \mathcal{N}_{[.]}\left(\frac{\delta}{3}, \mathcal{F}_{(\alpha)}, d_H\right) .$$

If we prove that for all $\varepsilon \leq 1$,

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{L}_{(K,\alpha)}, d_H\right) \tag{7.31}$$

then we obtain the result

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq K(2\pi\, e)^{\frac{K}{2}} \left(\frac{9}{\varepsilon}\right)^{K-1} \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) \mathcal{N}_{[.]}\left(\frac{\varepsilon}{9}, \mathcal{F}_{(\alpha)}, d_H\right)^K .$$

Thus, it remains to check Inequality (7.31). It is done by the following adaptation of a result proof given by Genovese and Wasserman (2000).

Let $\delta \in [0,1]$ and $h \in \mathcal{S}_{(K,\mathbf{v})}$, decomposed into $h(x) = f(x_{[\mathbf{v}]})g(x_{[\mathbf{v}^c]})$ where $f \in \mathcal{L}_{(K,\alpha)}$ and $g \in \mathcal{G}_{(\alpha)}$. Let $[l,u]$ and $[\tilde{l},\tilde{u}]$ be two δ-brackets of $\mathcal{L}_{(K,\alpha)}$ and $\mathcal{G}_{(\alpha)}$ containing f and g respectively. Then, the two functions defined by

$$L(x) = l(x_{[\mathbf{v}]})\tilde{l}(x_{[\mathbf{v}^c]}) \text{ and } U(x) = u(x_{[\mathbf{v}]})\tilde{u}(x_{[\mathbf{v}^c]}) \tag{7.32}$$

constitute a bracket of $\mathcal{S}_{(K,\mathbf{v})}$ containing h. The size of this bracket is now calculated. First of all, Lemma 3 from Genovese and Wasserman (2000) gives that

$$\begin{cases} \int u(x_{[\mathbf{v}]})\mathrm{d}x_{[\mathbf{v}]} \leq 1 + 3\delta \\ \int \tilde{u}(x_{[\mathbf{v}^c]})\mathrm{d}x_{[\mathbf{v}^c]} \leq 1 + 3\delta. \end{cases} \tag{7.33}$$

Then the squared Hellinger distance between L and U is equal to

$$\begin{aligned}
d_H^2(L,U) &= \int \left\{\sqrt{u(x_{[\mathbf{v}]})\tilde{u}(x_{[\mathbf{v}^c]})} - \sqrt{l(x_{[\mathbf{v}]})\tilde{l}(x_{[\mathbf{v}^c]})}\right\}^2 \mathrm{d}x \\
&= \int \left[\sqrt{\tilde{u}(x_{[\mathbf{v}^c]})}\left\{\sqrt{u(x_{[\mathbf{v}]})} - \sqrt{l(x_{[\mathbf{v}]})}\right\}\right. \\
&\qquad \left. + \left\{\sqrt{\tilde{u}(x_{[\mathbf{v}^c]})} - \sqrt{\tilde{l}(x_{[\mathbf{v}^c]})}\right\}\sqrt{l(x_{[\mathbf{v}]})}\right]^2 \mathrm{d}x \\
&= \int \tilde{u}(x_{[\mathbf{v}^c]})\,\mathrm{d}x_{[\mathbf{v}^c]} \int \left\{\sqrt{u(x_{[\mathbf{v}]})} - \sqrt{l(x_{[\mathbf{v}]})}\right\}^2 \mathrm{d}x_{[\mathbf{v}]} \\
&\quad + \int \left\{\sqrt{\tilde{u}(x_{[\mathbf{v}^c]})} - \sqrt{\tilde{l}(x_{[\mathbf{v}^c]})}\right\}^2 \mathrm{d}x_{[\mathbf{v}^c]} \int l(x_{[\mathbf{v}]})\,\mathrm{d}x_{[\mathbf{v}]} \\
&\quad + 2\int \sqrt{\tilde{u}(x_{[\mathbf{v}^c]})}\left\{\sqrt{\tilde{u}(x_{[\mathbf{v}^c]})} - \sqrt{\tilde{l}(x_{[\mathbf{v}^c]})}\right\}\mathrm{d}x_{[\mathbf{v}^c]} \\
&\qquad \times \int \left\{\sqrt{u(x_{[\mathbf{v}]})} - \sqrt{l(x_{[\mathbf{v}]})}\right\}\sqrt{l(x_{[\mathbf{v}]})}\,\mathrm{d}x_{[\mathbf{v}]}.
\end{aligned}$$

According Cauchy-Schwarz inequality and (7.33),

$$\begin{aligned}\int \left\{\sqrt{u(x_{[\mathbf{v}]})} - \sqrt{l(x_{[\mathbf{v}]})}\right\} \sqrt{l(x_{[\mathbf{v}]})}\, \mathrm{d}x_{[\mathbf{v}]} &\leq 1 \times d_H(l,u) \\ &\leq \delta\end{aligned}$$

and

$$\begin{aligned}\int \sqrt{\tilde{u}(x_{[\mathbf{v}^c]})} \left\{\sqrt{\tilde{u}(x_{[\mathbf{v}^c]})} - \sqrt{\tilde{l}(x_{[\mathbf{v}^c]})}\right\} \mathrm{d}x_{[\mathbf{v}^c]} &\leq \sqrt{1+3\delta} \times d_H(\tilde{l},\tilde{u}) \\ &\leq 2\,\delta\ .\end{aligned}$$

Thus,

$$\begin{aligned}d_H^2(L,U) &\leq d_H^2(l,u) \int \tilde{u}(x_{[\mathbf{v}^c]})\, \mathrm{d}x_{[\mathbf{v}^c]} + d_H^2(\tilde{l},\tilde{u}) + 4\delta^2 \\ &\leq (1+3\delta)\,\delta^2 + \delta^2 + 4\delta^2 \\ &\leq 9\delta^2\ .\end{aligned}$$

Finally, with $\delta = \varepsilon/3$ and according to the bracket definition (7.32), the number of brackets for $\mathcal{S}_{(K,\mathbf{v})}$ is upper bounded by $\mathcal{N}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) \times \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{L}_{(K,\alpha)}, d_H\right)$.

Proof of Proposition 7.A.2:

According to (7.31) in the proof of Proposition 7.A.1,

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{L}_{(K,\alpha)}, d_H\right).$$

Then we can prove along the line of the proof of Theorem 2 in Genovese and Wasserman (2000) that

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{L}_{(K,\alpha)}, d_H) \leq \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{P}_{K-1}, d_H\right) \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{F}_{(K,\alpha)}, d_H\right)$$

where $\mathcal{P}_{K-1}$ is the $K-1$ dimensional simplex. Sketch of the proof: Consider an $\varepsilon/3$-bracketing $\{[a_1 b_1], \ldots, [a_N, b_N]\}$ with $a_j, b_j \in [0,1]^K$, for the simplex $\mathcal{P}_{K-1}$ and an $\varepsilon/3$-bracketing for $\mathcal{F}_{(K,\alpha)}$. This last family is a set of K-tuples $([l_1,u_1], \ldots, [l_K,u_K])$ such that $d_H(l_k,u_k) \leq \varepsilon/3$ for all k. Then the family of brackets $[L,U]$ defined by $L(x) = \sum_{k=1}^K a_{jk} l_k(x)$ and $U(x) = \sum_{k=1}^K b_{jk} u_k(x)$ is an ε-bracketing of $\mathcal{L}_{(K,\alpha)}$.

7.D Results for multivariate Gaussian densities

7.D.1 Ratio of two Gaussian densities

Proposition 7.D.1. *Let $\Phi(.|\mu_1,\Sigma_1)$ and $\Phi(.|\mu_2,\Sigma_2)$ be two Gaussian densities . If $\Sigma_2 - \Sigma_1$ is a positive definite matrix then for all $x \in \mathbb{R}^Q$,*

$$\frac{\Phi(x|\mu_1,\Sigma_1)}{\Phi(x|\mu_2,\Sigma_2)} \leq \sqrt{\frac{|\Sigma_2|}{|\Sigma_1|}}\, \exp\left\{\frac{1}{2}(\mu_1-\mu_2)'(\Sigma_2-\Sigma_1)^{-1}(\mu_1-\mu_2)\right\}.$$

Proof. The ratio between the two Gaussian densities is equal to

$$\frac{\Phi(x|\mu_1,\Sigma_1)}{\Phi(x|\mu_2,\Sigma_2)} = \sqrt{\frac{|\Sigma_2|}{|\Sigma_1|}} \exp\left[-\frac{1}{2}\left\{(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) - (x-\mu_2)'\Sigma_2^{-1}(x-\mu_2)\right\}\right].$$

Proposition 7.D.1 is proved if

$$(x-\mu_2)'\Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) \leq (\mu_1-\mu_2)'(\Sigma_2-\Sigma_1)^{-1}(\mu_1-\mu_2). \quad (7.34)$$

The matrix $\Sigma_1^{-1} - \Sigma_2^{-1} = \Sigma_1^{-1}(\Sigma_2-\Sigma_1)\Sigma_2^{-1}$ is a positive definite matrix as the product of three positive definite matrices. Defining $\mu^\star = (\Sigma_1^{-1}-\Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2)$,

$$\begin{aligned}(x-\mu_2)'\Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) &= (\mu_1-\mu_2)'(\Sigma_2-\Sigma_1)^{-1}(\mu_1-\mu_2)\\ &- (x-\mu^\star)'(\Sigma_1^{-1}-\Sigma_2^{-1})(x-\mu^\star).\end{aligned}$$

Since the matrix $(\Sigma_1^{-1}-\Sigma_2^{-1})$ is a positive definite matrix, $(x-\mu^\star)'(\Sigma_1^{-1}-\Sigma_2^{-1})(x-\mu^\star) \geq 0$ and it leads to Inequality (7.34). □

Corollary 7.D.2. *Let $\Phi(.|\mu_1,B_1)$ and $\Phi(.|\mu_2,B_2)$ be two Gaussian densities. Their variance matrice are assumed to have the following diagonal form $B_i = \operatorname{diag}(b_{i1}^2,\ldots,b_{iQ}^2)$ for all $i=1,2$ such that $b_{2q}^2 > b_{1q}^2 > 0$ for all $q \in \{1\ldots,Q\}$. Then, for all $x \in \mathbb{R}^Q$, the ratio of the two densities is bounded by*

$$\left(\prod_{q=1}^{Q}\frac{b_{2q}}{b_{1q}}\right)\exp\left\{\frac{1}{2}(\mu_1-\mu_2)'\operatorname{diag}\left(\frac{1}{b_{21}^2-b_{11}^2},\ldots,\frac{1}{b_{2Q}^2-b_{1Q}^2}\right)(\mu_1-\mu_2)\right\}.$$

7.D.2 Hellinger distance between two Gaussian densities

The following proposition gives the expression of the Hellinger distance between two Gaussian densities.

Proposition 7.D.3. *Let $\Phi(.|\mu_1,\Sigma_1)$ and $\Phi(.|\mu_2,\Sigma_2)$ be two Gaussian densities. The squared Hellinger distance between these two densities has the following expression:*

$$2\left[1 - 2^{\frac{Q}{2}}|\Sigma_1\Sigma_2|^{-\frac{1}{4}}|\Sigma_1^{-1}+\Sigma_2^{-1}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{4}(\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)\right\}\right].$$

Proof. According to the definition of the Hellinger distance,

$$d_H^2(\Phi(.|\mu_1,\Sigma_1),\Phi(.|\mu_2,\Sigma_2)) = 2 - 2\int\sqrt{\Phi(x|\mu_1,\Sigma_1)\,\Phi(x|\mu_2,\Sigma_2)}\mathrm{d}x\,.$$

Furthermore,

$$\Phi(x|\mu_1,\Sigma_1)\,\Phi(x|\mu_2,\Sigma_2) = (2\pi)^{-Q}\,|\Sigma_1\Sigma_2|^{-\frac{1}{2}}\,\exp\left[-\frac{1}{2}(\diamondsuit)\right]$$

where the quantity $(\diamondsuit) := \big\{(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) + (x-\mu_2)'\Sigma_2^{-1}(x-\mu_2)\big\}$. Defining $\mu^\star = (\Sigma_1^{-1}+\Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2)$, we deduce that

$$(\diamondsuit) = (x-\mu^\star)'(\Sigma_1^{-1}+\Sigma_2^{-1})(x-\mu^\star) + (\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2).$$

Finally, the squared distance $d_H^2(\Phi(.|\mu_1,\Sigma_1),\Phi(.|\mu_2,\Sigma_2))$ is equal to

$$\begin{aligned}
&2 - 2(2\pi)^{-\frac{Q}{2}}|\Sigma_1\Sigma_2|^{-\frac{1}{4}}\,\exp\left\{-\frac{1}{4}(\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)\right\}\\
&\times\int\exp\left\{-\frac{1}{4}(x-\mu^\star)'(\Sigma_1^{-1}+\Sigma_2^{-1})(x-\mu^\star)\right\}\,\mathrm{d}x\\
&= 2 - 2(2\pi)^{-\frac{Q}{2}}|\Sigma_1\Sigma_2|^{-\frac{1}{4}}\,\exp\left\{-\frac{1}{4}(\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)\right\}\\
&\times(4\pi)^{\frac{Q}{2}}|\Sigma_1^{-1}+\Sigma_2^{-1}|^{-\frac{1}{2}}
\end{aligned}$$

that entails the concluding result. □

Corollary 7.D.4. *Using the notation of the Corollary 7.D.2, the Hellinger distance of two Gaussian densities with diagonal variance matrices is given by the following expression*

$$2 - 2\left(\prod_{q=1}^{Q}\frac{2\,b_{1q}\,b_{2q}}{b_{1q}^2+b_{2q}^2}\right)^{\frac{1}{2}}\exp\left[-\frac{1}{4}\,(\mu_1-\mu_2)'\,\mathrm{diag}\left\{\left(\frac{1}{b_{1q}^2+b_{2q}^2}\right)_{1\le q\le Q}\right\}\,(\mu_1-\mu_2)\right].$$

Chapter 8

Slope heuristics for a practical use of our penalized criterion

Résumé: Dans le chapitre précédent, nous avons considéré des mélanges gaussiens de formes spécifiques pour résoudre un problème de sélection de variables en classification non supervisée. Un critère de vraisemblance pénalisée a alors été proposé pour sélectionner le nombre de composantes du mélange et le sous-ensemble des variables significatives pour la classification. Mais ce critère dépend de constantes multiplicatives inconnues qui doivent être évaluées en pratique. Nous proposons dans ce chapitre d'utiliser une méthode heuristique dite "de la pente" pour résoudre ce problème. Cette procédure est appliquée sur des données simulées, sur un ensemble de courbes et sur des données transcriptomes pour mettre en évidence son intérêt.

8.1 Introduction

Variable selection for clustering, recast into a model selection problem, is tackled with a non asymptotic point of view in this part II. A penalized likelihood criterion is theoretically constructed in Chapter 7 to select the "best" model among a Gaussian mixture model collection. It allows us to specify the general shapes of the penalty functions but is not usable in practice since it depends on unknown constants. The aim of this chapter is to calibrate the penalty functions for a practical use of our non asymptotic penalized criterion. In such situations, Birgé and Massart (2006) propose their so-called "slope heuristics" (see also Massart, 2007, Section 8.5). The rule of thumb of this heuristics consists of assuming that twice the minimal penalty is almost the optimal penalty. It is theoretically proved in few frameworks but is the subject of several practical studies. This heuristics, described in detail in Section 8.3, is here considered in our framework.

First, our resulting methodology is applied to an oil production curve clustering problem. Curve clustering deals with the problem of identifying homogeneous groups in a set of functional data. This situation occurs in many areas of sciences, for instance in genetics,

neuroscience, economics and engineering. Many methods of curve clustering are based on different versions of the k-means algorithm. A widely used technique consists of finding a convenient projection of the functional data into a finite dimensional subspace, and next of applying a k-means procedure on the finite dimensional data obtained. In this context, B-spline bases are currently used, see for instance Abraham et al. (2003) and García-Escudero and Gordaliza (2005). An other approach proposed by Tarpey and Kinateder (2003) is to adapt the k-means algorithm for functional spaces. With a different point of view, Ma et al. (2006) use mixture models on B-splines coefficients, as in the work of James and Sugar (2003) for sparsely sampled functional data. In most of cited work, each curve is described with a coefficient vector and in practice, the number of these coefficients can be of the same order as the curve number. This high dimensional context makes our method desirable to solve curve clustering problems.

Next, our method is applied to the transcriptome data clustering in a particular context. During these last years, biologists are interested in determining biological functions of genes. In this aim, clustering methods such as hierarchical clustering or k-means algorithm are commonly applied to find clusters of coexpressed genes (see for instance Sharan et al., 2002; Jiang et al., 2004, and references therein). Since the experiment number increases in available transcriptome datasets, variable selection is more and more considered in order to lead to reliable and interpretable clustering for biologists. In the model-based clustering context, Maugis et al. (2008) apply their variable selection method for a transcriptome dataset analysis for instance.

The chapter is organized as follows: In Section 8.2, the theoretical results of Chapter 7, given the shape of the penalized criterion to select a model into the considered collections of Gaussian mixture models are recalled. Section 8.3 is devoted to the description of the slope heuristics and its practical use in our specific context. Simulations and applications for a curve clustering problem and for a genomics study are presented in Section 8.4 to highlight the interest of this method.

8.2 Recall of the theoretical results

8.2.1 Framework

Suppose that we observe a centered sample $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, with $\mathbf{y}_i \in \mathbb{R}^Q$, from an unknown probability distribution with density s. This target density is estimated using a specific collection of Gaussian mixture models. Each model $\mathcal{S}_{(K,\mathbf{v})}$ corresponds to a particular clustering situation where the cluster number is $K \in \mathbb{N}^\star$ and $\mathbf{v}$ is the relevant clustering variable subset, included into $\{1, \ldots, Q\}$. Recall that the cardinal of $\mathbf{v}$ is denoted α, its complement is denoted $\mathbf{v}^c$ and a vector $x \in \mathbb{R}^Q$ is decomposed into $x_{[\mathbf{v}]}$ and $x_{[\mathbf{v}^c]}$. The dimension of the model $\mathcal{S}_{(K,\mathbf{v})}$, denoted $D(K, \alpha)$, corresponds to the free parameter number of Gaussian mixture densities in this model. A density t belonging to $\mathcal{S}_{(K,\mathbf{v})}$ is decomposed into a Gaussian mixture density with K components on the relevant clustering variable subset $\mathbf{v}$ and a multidimensional spherical Gaussian density on the other variables: For all

$x \in \mathbb{R}^Q$,

$$t(x) = \sum_{k=1}^{K} p_k \Phi(x_{[\mathbf{v}]}|\mu_k, \Sigma_k) \times \ \Phi(x_{[\mathbf{v}^c]}|0, \omega^2 I_{Q-\alpha}) \tag{8.1}$$

where $\forall k \in \{1, \ldots, K\}$, $\mu_k \in [-a, a]^\alpha (a \in \mathbb{R}_+^\star), p_k \in]0, 1[$ such that $\sum_{k=1}^K p_k = 1$ and $\omega^2 \in [\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$ $(0 < \lambda_{\mathrm{m}} < \lambda_{\mathrm{M}})$. The variance matrices $(\Sigma_1, \ldots, \Sigma_K)$ belong to a family of K-uples of $\alpha \times \alpha$ symmetric positive definite matrices, which is related to the Gaussian mixture shape. Recall that the following four collections of Gaussian mixtures are considered:

- For the $[L_k B_k]$ collection, the variance matrices of mixtures are assumed to be diagonal and free. Their eigenvalues are assumed to be in the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$ and the associated dimension of model $\mathcal{S}_{(K,\mathbf{v})}$ is equal to $D(K, \alpha) = K(2\alpha + 1)$.

- For the $[L_k C_k]$ collection, the variance matrices are assumed to be totally free. Thus, the variance matrices are $\alpha \times \alpha$ positive definite matrices whose eigenvalues are assumed to belong to the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. The associated model dimension is $D(K, \alpha) = K[1 + \alpha + \frac{\alpha(\alpha+1)}{2}]$.

- For the $[L B_k]$ collection, the variance matrices are assumed to be diagonal and to have the same volume. The variance matrices are decomposed into $\Sigma_k = \beta B_k$ where the common volume β belongs to $[\beta_{\mathrm{m}}, \beta_{\mathrm{M}}]$ and B_k is a diagonal matrix with a determinant 1 and with diagonal coefficients in the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. Here, the model dimension is equal to $D(K, \alpha) = 2K\alpha + 1$.

- For the $[LC]$ collection, the variance matrices are all equal to a free positive definite matrix Σ whose eigenvalues are assumed to be in the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. The model dimension is $D(K, \alpha) = K(1 + \alpha) + \frac{\alpha(\alpha+1)}{2}$.

Moreover, for each of the four possible model collections, the variables can be assumed to be ordered or not ordered. If variables are ordered, the relevant variable subset is $\mathbf{v} = \{1, \ldots, \alpha\}$ and can be assimilated to its cardinal α.

These Gaussian mixture families recast the clustering and variable selection problems into a global model selection problem since this selection automatically leads to a data clustering and a variable selection.

8.2.2 The theoretical penalized likelihood criterion

In this section, we go back on the theoretical results obtained in Chapter 7 to make the link with their practical use. Recall that the considered contrast is $\gamma(t, \cdot) = -\ln\{t(\cdot)\}$, leading to the maximum likelihood criterion and the corresponding loss function is the Kullback-Leibler information. For our countable collection of models $\{\mathcal{S}_{(K,\mathbf{v})}\}_{(K,\mathbf{v})\in\mathcal{M}}$, $\hat{s}_{(K,\mathbf{v})}$ is a minimizer of the empirical contrast

$$\gamma_n(t) = -\frac{1}{n} \sum_{i=1}^{n} \ln\{t(\mathbf{y}_i)\}$$

over the model $\mathcal{S}_{(K,\mathbf{v})}$. The model we want to select is the one presenting the smallest risk

$$(K^\star, \mathbf{v}^\star) = \underset{(K,\mathbf{v})\in\mathcal{M}}{\operatorname{argmin}}\ \mathbb{E}[\mathrm{KL}(s, \hat{s}_{(K,\mathbf{v})})].$$

However, the function $\hat{s}_{(K^\star,\mathbf{v}^\star)}$, called oracle, is unknown since it depends on the true density s. Thus a penalized criterion is proposed to select a model $(\hat{K}, \hat{\mathbf{v}})$, for which the associated selected estimator $\hat{s}_{(\hat{K},\hat{\mathbf{v}})}$ has a risk as close as possible as the benchmark $\inf_{(K,\mathbf{v})\in\mathcal{M}} \mathbb{E}[\mathrm{KL}(s, \hat{s}_{(K,\mathbf{v})})]$, its behaviour being justified by the oracle inequality. Recall that the norm $\|\sqrt{f} - \sqrt{g}\|_2$ between two nonnegative functions f and g of $\mathbb{L}_1$ is denoted $d_H(f, g)$ and is improperly called Hellinger distance. The theoretical results, established in Chapter 7 and valid for the four collections of Gaussian mixture models, are summarized in the following theorem.

Theorem. *For the four Gaussian mixture collections,*

1. *If the variables are ordered, there exists two absolute constants κ and C such that, if*

$$\mathrm{pen}(K, \mathbf{v}) \geq \kappa \frac{D(K,\alpha)}{n} \left[2\,A + \ln\left(\frac{1}{1 \wedge \frac{D(K,\alpha)}{n} A} \right) + 1 \right] \tag{8.2}$$

   then the model $(\hat{K}, \hat{\mathbf{v}})$ minimizing $\mathrm{crit}(K, \mathbf{v}) = \gamma_n(\hat{s}_{(K,\mathbf{v})}) + \mathrm{pen}(K, \mathbf{v})$ on $\mathcal{M}$ exists and

$$\mathbb{E}\left[d_H^2(s, \hat{s}_{(\hat{K},\hat{\mathbf{v}})} \right] \leq C \left\{ \inf_{(K,\mathbf{v})\in\mathcal{M}} [\mathrm{KL}(s, \mathcal{S}_{(K,\mathbf{v})}) + \mathrm{pen}(K, \mathbf{v})] + \frac{1}{n} \right\}.$$

2. *If the variables are not ordered, there exists two absolute constants κ and C such that, if*

$$\mathrm{pen}(K, \mathbf{v}) \geq \kappa \frac{D(K,\alpha)}{n} \left\{ 2\,A + \ln\left[\frac{1}{1 \wedge \frac{D(K,\alpha)}{n} A} \right] + \frac{1}{2} \ln\left[\frac{8 \exp(1) Q}{(D(K,\alpha) - 1) \wedge (2Q - 1)} \right] \right\}, \tag{8.3}$$

   then the model $(\hat{K}, \hat{\mathbf{v}})$ minimizing $\mathrm{crit}(K, \mathbf{v}) = \gamma_n(\hat{s}_{(K,\mathbf{v})}) + \mathrm{pen}(K, \mathbf{v})$ on $\mathcal{M}$ exists and

$$\mathbb{E}\left[d_H^2(s, \hat{s}_{(\hat{K},\hat{\mathbf{v}})} \right] \leq C \left\{ \inf_{(K,\mathbf{v})\in\mathcal{M}} [\mathrm{KL}(s, \mathcal{S}_{(K,\mathbf{v})}) + \mathrm{pen}(K, \mathbf{v})] + \frac{2}{n} \right\}.$$

In the four cases, A is a function of parameters λ_m, λ_M, a, Q and also β_m and β_M for the $[LB_k]$ collection such that $A = O(\sqrt{\ln Q})$ as Q tends to infinity.

The penalty functions take into account the model complexity through $D(K, \alpha)$ and the richness of model family. Since the number of models with the same dimension is larger in the non-ordered variable case, the associated penalty functions have an additional logarithm term, depending on the dimension.

The other logarithm term, common to both cases, is probably not necessary to define efficient penalties. As explained in the previous chapter, the reason for that is certainly that the general model selection theorem for MLE is stated in a local version whereas we only manage to apply the global version in our framework. Logarithm terms are not detected in practice as shown in Section 8.4 and thus only the preponderant term in $D(K,\alpha)/n$ is retained in the penalty shape.

Contrary to classical criteria for which Q is fixed and n tends to infinity, our result allows to study cases for which Q increases with n. For specific clustering problems where the number of variables Q is of the order of n or even larger than n, the oracle inequality is still relevant.

The point we want to stress here is that the theoretical results (8.2) and (8.3) give the general shape of penalty functions but is not totally explicit since the results depend on absolute unknown constants. Consequently a method is to be applied to calibrate the penalty function for a practical use of these results. Moreover, we note that mixture parameters are not bounded in practice.

8.3 Slope heuristics

Since the lower bounds on penalty functions in (8.2) and (8.3) are defined up to an unknown multiplicative constant, this theorem does not provide directly an usable model selection criterion. Last years, some efforts have been paid to overcome such a difficulty. Birgé and Massart (2006) propose a practical method based on a mixture of theoretical and heuristic ideas for defining efficient penalty functions from the data. This heuristics is only proved in Birgé and Massart (2006) in the framework of Gaussian regression with a homoscedastic fixed design and more recently generalized by Arlot and Massart (2008) in the heteroscedastic random-design case. Nevertheless applications of this method are developed in many other frameworks: For instance, in multiple change points detection by Lebarbier (2005), in genomics applications by Villers (2007) and in Gaussian Markov random fields by Verzelen (2008). This section first describes the main ideas of this heuristics, called the "*slope heuristics*", and next details its practical use in our framework.

8.3.1 Rationale for the slope heuristics

In many situations, the considered model collection contains several models with the same dimension. In order to penalize each model of dimension D in the same way, a new collection $(\mathcal{S}_D)_{D\in\mathcal{D}}$ is considered such that $\mathcal{S}_D$ is the union of all the models $\mathcal{S}_{(K,\mathbf{v})}$ having the same dimension D. A minimizer of $\mathrm{KL}(s,\cdot)$ on $\mathcal{S}_D$ is denoted

$$s_D = \underset{t\in\mathcal{S}_D}{\operatorname{argmin}}\ \mathrm{KL}(s,t)$$

and a minimizer of $\gamma_n(\cdot)$ on $\mathcal{S}_D$ is denoted

$$\hat{s}_D = \underset{t\in\mathcal{S}_D}{\operatorname{argmin}}\ \gamma_n(t).$$

As for criteria due to Mallows (1973) and Akaike (1973, 1974), Birgé and Massart's criterion is based on an unbiased risk estimation. The ideal model to estimate s is the one minimizing the risk $\mathbb{E}\left[\mathrm{KL}(s, \hat{s}_D)\right]$. Nevertheless, it is impossible to choose this optimal model since s is unknown. A solution is to find a penalty function, called *optimal penalty* such that the empirical risk is as close as possible to the benchmark $\inf_{D\in\mathcal{D}} \mathbb{E}\left[\mathrm{KL}(s, \hat{s}_D)\right]$. To express the risk of each estimator $\hat{s}_D$, the following decomposition is considered

$$\begin{aligned} \mathrm{KL}(s, \hat{s}_D) &= \int \ln\left[\frac{s(x)}{s_D(x)}\right] s(x)dx + \int \ln\left[\frac{s_D(x)}{\hat{s}_D(x)}\right] s(x)dx \\ &= b_D + V_D \end{aligned} \tag{8.4}$$

where $b_D := \mathrm{KL}(s, s_D)$ is the bias term and $V_D := \int \ln(s_D/\hat{s}_D)s$ is the variance term. Note that the bias b_D decreases whereas the variance term V_D tends to increase when the dimension D increases. Then, taking the expectation of (8.4), it leads to

$$\mathbb{E}\left[\mathrm{KL}(s, \hat{s}_D)\right] = b_D + \mathbb{E}[V_D].$$

Among the model collection $\mathcal{D}$, the selected model $\hat{D}$ is the one minimizing the criterion

$$D \mapsto \gamma_n(\hat{s}_D) + \mathrm{pen}(D). \tag{8.5}$$

Defining $\hat{b}_D := \gamma_n(s_D) - \gamma_n(s)$ and $\widehat{V}_D := \gamma_n(s_D) - \gamma_n(\hat{s}_D)$, the selected model is also a minimizer of

$$\begin{aligned} \gamma_n(\hat{s}_D) - \gamma_n(s) + \mathrm{pen}(D) &= \gamma_n(\hat{s}_D) - \gamma_n(s_D) + \gamma_n(s_D) - \gamma_n(s) + \mathrm{pen}(D) \\ &= \hat{b}_D - \widehat{V}_D + \mathrm{pen}(D). \end{aligned} \tag{8.6}$$

Then, introducing the term of interest (8.4) into (8.6), it leads to

$$\begin{aligned} \gamma_n(\hat{s}_D) - \gamma_n(s) + \mathrm{pen}(D) &= b_D + V_D + (\hat{b}_D - b_D) - (V_D + \widehat{V}_D) + \mathrm{pen}(D) \\ &= \mathrm{KL}(s, \hat{s}_D) + (\hat{b}_D - b_D) - (V_D + \widehat{V}_D) + \mathrm{pen}(D). \end{aligned}$$

Because of the law of large numbers, it is reasonable to assume that $\hat{b}_D - b_D \approx 0$. Furthermore, concentration arguments allow us to suppose that $\mathrm{KL}(s, \hat{s}_D)$ is close to its expectation. Thus

$$\gamma_n(\hat{s}_D) - \gamma_n(s) + \mathrm{pen}(D) \approx \mathbb{E}[\mathrm{KL}(s, \hat{s}_D)] - (V_D + \widehat{V}_D) + \mathrm{pen}(D). \tag{8.7}$$

In order to make (8.7) close to the risk $\mathbb{E}\left[\mathrm{KL}(s, \hat{s}_D)\right]$, the *optimal penalty* is defined by

$$\mathrm{pen}_{\mathrm{opt}}(D) = V_D + \widehat{V}_D.$$

Next, the main point of this heuristics is to assume that $\widehat{V}_D \approx V_D$. An argument to justify this assumption is that the probability measure and the corresponding empirical measure play a similar role, in the expressions of V_D and $\widehat{V}_D$. If one permutes these measures inside

the definitions of V_D and $\widehat{V}_D$, and also in the definitions of s_D and $\hat{s}_D$, then V_D is changed in $\widehat{V}_D$ and reciprocally. Finally, this assumption leads to $\text{pen}_{\text{opt}}(D) = 2\widehat{V}_D$. Turning back on the expression of $\widehat{V}_D$, it can be written

$$\begin{aligned}\widehat{V}_D &= \gamma_n(s_D) - \gamma_n(s) + \gamma_n(s) - \gamma_n(\hat{s}_D) \\ &= \hat{b}_D + \gamma_n(s) - \gamma_n(\hat{s}_D).\end{aligned}$$

For large dimensions, the bias term stabilizes itself since the approximation of the model cannot be appreciably improved. Thus, the behavior of $\widehat{V}_D$ according to the model dimension is known for large dimensions via $-\gamma_n(\hat{s}_D)$. In our framework, penalty functions could be regarded as proportional to the dimension (see remarks after the recall of the theoretical results in Section 8.2) hence

$$\text{pen}_{\text{opt}}(D) = 2\widehat{V}_D = 2C_{\text{opt}}D$$

where C_{opt} is a constant. In order to use the slope heuristics to calibrate the penalty, a required condition is to observe a linear behaviour of $D \mapsto -\gamma_n(\hat{s}_D)$ for large dimension. If this condition is fulfilled, C_{opt} can be estimated by the slope $\hat{C}$ of the linear part of $D \mapsto -\gamma_n(\hat{s}_D)$ and the final penalty is

$$\text{pen}(D) = 2\hat{C}D.$$

8.3.2 Using the slope heuristics

This section details how the slope heuristics is applied to select a Gaussian mixture model among a family $\left(\mathcal{S}_{(K,\mathbf{v})}\right)_{(K,\mathbf{v})\in\mathcal{M}}$ with $\mathcal{M} := \{(K,\mathbf{v});\ 2 \leq K \leq K_{\max}, \mathbf{v} \in \mathcal{V}\}$ where the maximum number of mixture components $K_{\max}$ and the mixture shape are fixed by the user. Our model selection procedure, based on the slope heuristics, is decomposed into the three following steps:

1. *Estimation step*: The maximum likelihood estimator $\hat{s}_{(K,\mathbf{v})}$ is computed for each model $\mathcal{S}_{(K,\mathbf{v})}$. According to (8.1), the mixture parameters and ω^2 can be independently estimated. Thus

$$\hat{s}_{(K,\mathbf{v})}(x) = \sum_{k=1}^{K} \hat{p}_k \Phi\left(x_{[\mathbf{v}]}|\hat{\mu}_k, \widehat{\Sigma}_k\right) \times \Phi(x_{[\mathbf{v}^c]}|0, \hat{\omega}^2 I_{Q-\alpha})$$

   where the estimated mixture parameters $\left(\hat{p}_1, \ldots, \hat{p}_K, \hat{\mu}_1, \ldots, \hat{\mu}_K, \widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_K\right)$ are computed with the Expectation Maximization (EM) algorithm (Dempster et al., 1977) using MIXMOD software (Biernacki et al., 2006) and $\hat{\omega}^2 = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_{i[\mathbf{v}^c]}\|^2$.

2. *Penalty determination step*: First, models are grouped according to their dimension in order to obtain the model collection $(\mathcal{S}_D)_{D\in\mathcal{D}}$. For each dimension $D \in \mathcal{D}$, $\hat{s}_D$

is the estimator providing the greatest loglikelihood value among the estimators associated to a model of dimension D. The associated model is denoted $(K_D, \mathbf{v}_D)$ ($\hat{s}_D = \hat{s}_{(K_D,\mathbf{v}_D)}$) and the estimated parameter vector, obtained in the first step, is preserved for the third step.

The function $D \mapsto -\gamma_n(\hat{s}_D)$ is plotted. The user has to check that this function has a linear behaviour for large dimensions. If this condition is not fulfilled, the slope heuristics cannot be applied. This situation may occur when the estimation step is performed not for large enough dimensions or when the choice of the used Gaussian mixture collection among the four possibilities (see Section 8.2) is not adapted to the studied dataset.

Assume that the linear behaviour of $D \mapsto -\gamma_n(\hat{s}_D)$ is observed for large dimensions. From the graphical representation, the user has to choose a large enough threshold D_0 such that the restriction of $-\gamma_n(\hat{s}_D)$ on the set of dimensions greater than D_0 has a linear behaviour. Then the slope $\hat{C}$ of this linear part is evaluated. Since possible estimation errors in the first step can damage the slope estimation, a robust regression procedure (Huber, 1981) is used to attenuate this influence. This robust regression method using an iteratively weighted least squares algorithm, is expected to give lower weight to suboptimal and spurious parameter estimates. Finally, the calibrated penalty function is $\text{pen}(D) = 2\hat{C}D$.

3. *Model selection step*: The minimizer $\hat{D}$ of the criterion $D \mapsto \gamma_n(\hat{s}_D) + 2\hat{C}D$ is determined and the model $(\hat{K}, \hat{\mathbf{v}}) = (K_{\hat{D}}, \mathbf{v}_{\hat{D}})$ is selected. Finally, the estimated parameter vector associated to $(\hat{K}, \hat{\mathbf{v}})$ provide a data clustering using the MAP rule (see Chapter 1).

Remark about estimated oracle model: When simulated datasets (density s is known) are studied, the model $(\hat{K}, \hat{\mathbf{v}})$ selected with our penalized criterion can be compared to the oracle model $(K^\star, \mathbf{v}^\star)$ fulfilling

$$(K^\star, \mathbf{v}^\star) = \underset{(K,\mathbf{v})\in\mathcal{M}}{\text{argmin}}\ \mathbb{E}\left[-\int \ln\{\hat{s}_{(K,\mathbf{v})}(x)\}s(x)dx\right] \tag{8.8}$$

with a Monte Carlo procedure. A first sample $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, with a large size n, is simulated from the density s in order to approximate the integral of the right-hand side of (8.8) by

$$f_{\mathbf{x}}(\hat{s}_{(K,\mathbf{v})}) = -\frac{1}{n}\sum_{i=1}^{n} \ln\{\hat{s}_{(K,\mathbf{v})}(\mathbf{x}_i)\}.$$

A collection of M samples from s is considered. For the j^{th} sample, the maximum likelihood estimators $(\hat{s}^{(j)}_{(K,\mathbf{v})})_{(K,\mathbf{v})\in\mathcal{M}}$ are evaluated. Thus the estimated oracle model is defined by

$$(\hat{K}_{\text{oracle}}, \hat{\mathbf{v}}_{\text{oracle}}) = \underset{(K,\mathbf{v})\in\mathcal{M}}{\text{argmin}}\ \frac{1}{M}\sum_{j=1}^{M} f_{\mathbf{x}}(\hat{s}^{(j)}_{(K,\mathbf{v})}).$$

8.4 Applications

This section is devoted to the application of our method on simulated and real datasets. We show that our method allows us to choose the variable role to improve the clustering. Moreover, we check that the penalized estimator mimics the oracle. The slope heuristics is compared with the classical criteria used for Gaussian mixture model selection: AIC, BIC and ICL. They are respectively defined by

$$\text{crit}_{\text{AIC}}(D) = \gamma_n(\hat{s}_D) + \frac{D}{n},$$

$$\text{crit}_{\text{BIC}}(D) = \gamma_n(\hat{s}_D) + \frac{D\ln(n)}{2n}$$

and

$$\text{crit}_{\text{ICL}}(D) = \text{crit}_{\text{BIC}} + \frac{\text{ENT}}{n}$$

with the entropy term $\text{ENT} = -\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\ln(t_{ik})$ where z is given by the MAP rule and

$$t_{ik} = \frac{\hat{p}_k\Phi(\mathbf{y}_i|\hat{\mu}_k,\hat{\Sigma}_k)}{\sum_{l=1}^{K}\hat{p}_l\Phi(\mathbf{y}_i|\hat{\mu}_l,\hat{\Sigma}_l)}.$$

The reader is respectively referred to Akaike (1973, 1974), Schwarz (1978) and Biernacki et al. (2000) for more details on these criteria.

The method is applied on simulated datasets in Sections 8.4.1 and 8.4.2 and then on real datasets. In Section 8.4.3, our procedure is carried out on a curve clustering example for oil production profiles. In Section 8.4.4, a transcriptome dataset is studied with our method to obtain coexpressed gene clusters.

8.4.1 Assessment of the slope heuristics

The aim of this first example is to check the validity of the linear penalty shape assumption, to compare the slope estimator with other penalized estimators and to study its behavior with respect to the oracle. The dataset consists of $n = 2000$ points described by $Q = 32$ variables. The data are simulated according to a mixture of four equiprobable Gaussian distributions $\mathcal{N}(\mu_k, \Sigma_k)$ where

$$\mu_1 = (3, 2, 1, 0.7, 0.3, 0.2, 0.1, 0.07, 0.05, 0.025),\ \mu_2 = 0_{10},\ \mu_3 = -\mu_1,$$
$$\mu_4 = (3, -2, 1, -0.7, 0.3, -0.2, 0.1, -0.07, -0.05, -0.025),$$

and

$$\Sigma_1 = \Sigma_3 = \Sigma_4 = I_{10} \text{ and } \Sigma_2 = \text{diag}(2, 1.9, 1.8, \ldots, 1.1).$$

The vector 0_p denotes the null vector of length p. Twenty two independent variables sampled from a $\mathcal{N}(0, 1)$ are appended. Consequently, the true density belongs to the model $\mathcal{S}_{(K_0,\mathbf{v}_0)}$ where $K_0 = 4$ and $\mathbf{v}_0 = \{1, \ldots, 10\}$ ($\alpha_0 = 10$) and the variables are ordered.

Note that the discriminant power of the relevant variables decreases with respect to the variable index. In other words the four subpopulations of the mixtures are progressively gathered together into a unique Gaussian distribution, as shown in Figure 8.1.

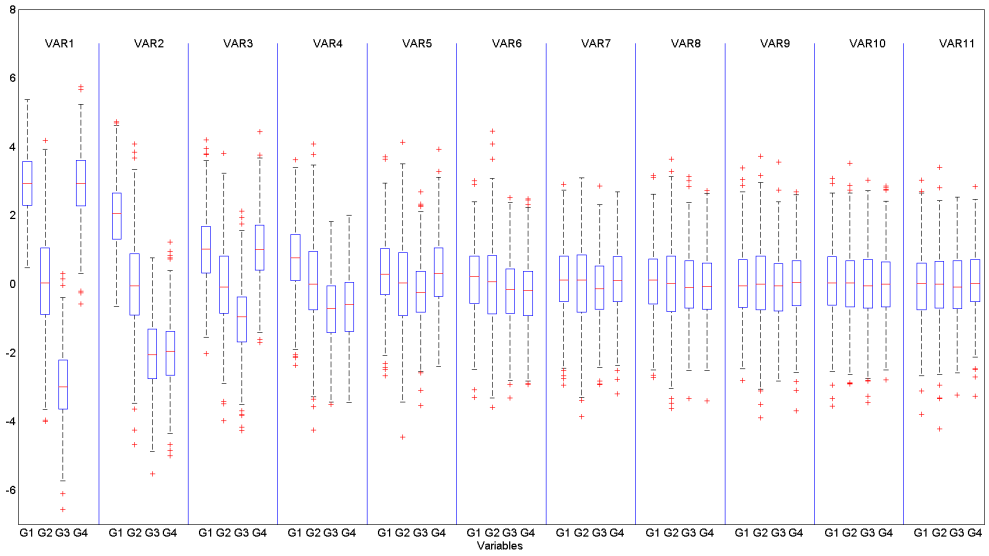


Figure 8.1: Boxplots of the first eleven variables (VAR1,...,VAR11) on the four mixture components (G1,G2,G3,G4).

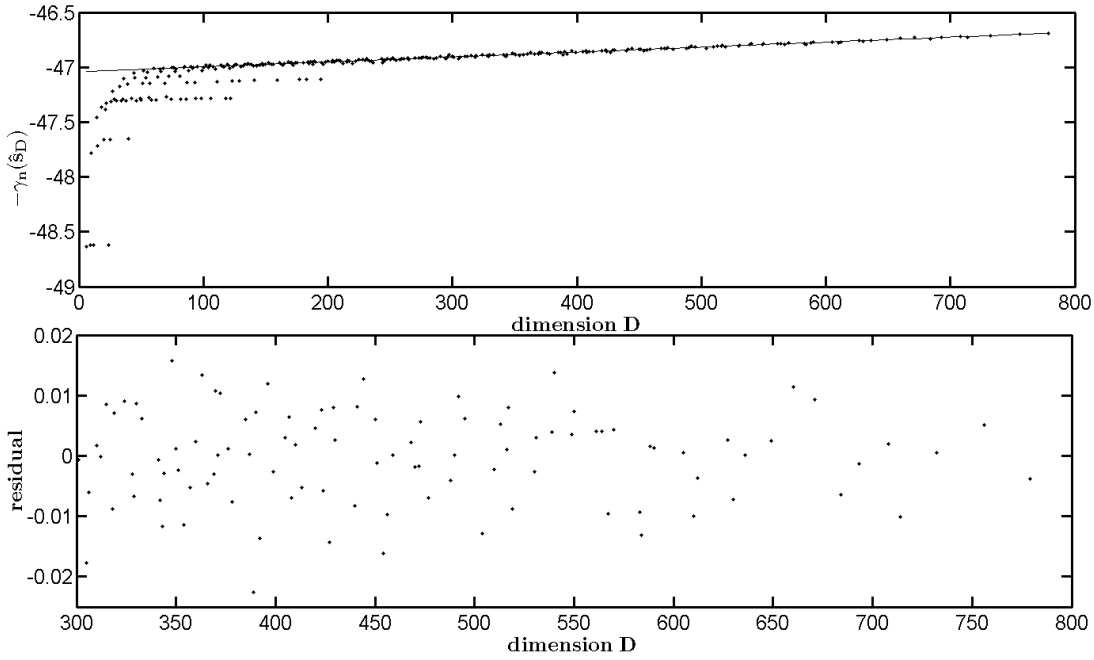


Figure 8.2: On the top graph, the function $D \mapsto -\gamma_n(\hat{s}_D)$ is plotted. The linear regression is performed for $D \geq D_0 = 300$. The associated residuals are drawn on the bottom graph.

The model collection associated to the $[L_k B_k]$ Gaussian mixture shape is considered and variables are assumed to be ordered. After the estimation step, the function $D \mapsto -\gamma_n(\hat{s}_D)$ is plotted on the top of Figure 8.2. For $D \geq D_0 = 300$, we observe that the function $D \mapsto -\gamma_n(\hat{s}_D)$ has a linear behavior as expected (see Section 8.3.1). The residuals of the linear regression are plotted on the bottom of Figure 8.2. This defends the use of penalties proportional to the dimension since no trend can be observed in the residuals. The estimation of $\hat{C}$ leads to the penalty choice in criterion (8.5) and the selected model according to this penalized criterion is $\hat{K}_{\text{slope}} = 4$ and $\hat{\alpha}_{\text{slope}} = 7$.

In order to compare the behavior of the slope estimator with the oracle and with the behaviors of three other estimators given by AIC, BIC and ICL criteria, 1000 simulated datasets are considered. The distribution used for simulations is the same as before except that the last ten variables have been removed in order to reduce the computation times. Consequently, the "true model" still corresponds to $(K_0, \alpha_0) = (4, 10)$ but the total number of variables is now $Q = 22$. Since the true density is known, a Monte Carlo procedure (see Section 8.3.2) gives the following oracle model estimation $\hat{K}_{\text{oracle}} = 4$ and $\hat{\alpha}_{\text{oracle}} = 9$. Note that even if the true density belongs to the density collection, the oracle model is not equal to the corresponding true model. The results are summarized in Table 8.1. The AIC criterion selects too many components and too many relevant variables since it underpenalizes models in the mixture context. The two criteria BIC and ICL select a model with four components and most of the times with six relevant variables. It is shown in Keribin (2000) that a model selection procedure using BIC is consistent to find the number of components of a Gaussian mixture when the component densities are bounded. But as far as we know, there is no consistency result for such a variable selection and clustering problem. The results of Table 8.1 show that the model selected by BIC is not the true model. In this context, even if BIC tries to find the true model, the asymptotics could be not achieved.The behavior of the ICL method is not surprising since the aim of this criterion is to provide a mixture model leading to a sensible partitioning of the data. From a clustering point of view, BIC, ICL and the slope method have similar performances. The interest of this first simulated example is to illustrate the behavior of different criteria. As expected, the slope method selects a model close to the oracle model.

criterion	$\hat{K}$	$\hat{\alpha}$								
		5	6	7	8	9	10	11	12	≥ 13
ICL	4	22	**792**	184	2					
BIC	4	29	**859**	111	1					
AIC	≤ 5			2	21	26	11	3	2	
	6			7	42	62	26	9	2	3
	≥ 7			57	155	**237**	170	93	41	41
Slope Heuristics	4		43	417	**456**	58	1			
	5		8	13	13	4				

Table 8.1: For each criterion, number of times that a model (K, α) is selected among the 1000 simulations.

8.4.2 Waveform dataset

The waveform dataset, available at the UCI repository Blake et al. (1999), is composed of three groups based on a random convex combination of two of three wave functions sampled at the integers from 1 to 21, with added noise. A detailed description is available in Breiman et al. (1984). The dataset consists of 5000 observations described by 40 variables. The last nineteen are noisy variables, sampled from a $\mathcal{N}(0,1)$ density. By construction, Variables 1 and 21 have the same distribution $\mathcal{N}(0,1)$. Consequently they are both irrelevant for the clustering and thus there are 19 variables which are potentially relevant for clustering.

First, the data have been centered. Contrary to the previous example, the variables are not ordered. Ideally, the model collection should be based on all the possible relevant variable subsets $\mathbf{v}$. Nevertheless, the selection among this model family is impossible because of the large cardinal of this collection and the resulting computation times. To get round this problem, the variables are ordered by decreasing order of their variances. With this ordering, the last twenty-one variables are the variables sampled from the $\mathcal{N}(0,1)$ density.

The model selection is performed with the two mixture shapes $[L_k B_k]$ and $[L_k C_k]$. The plots of $D \mapsto -\gamma_n(\hat{s}_D)$ for the estimation of $\hat{C}$ are given on the top of Figure 8.3 for these two collections of models. To compare the two model collections, both corresponding fittings of the dimension model surfaces on maximum loglikelihood surfaces are presented on the bottom of Figure 8.3. As expected, we observe that the fitting is dramatically better for the model collection associated to the $[L_k C_k]$ collection. Indeed the relevant variables are dependent by construction. The $[L_k B_k]$ model collection is too simple for this problem. This $[L_k C_k]$ collection leads to select a model with $\hat{K} = 3$ clusters and $\hat{\alpha} = 19$ relevant variables. Despite the simulated data do not follow a Gaussian mixture, the procedure provides a stable and sensible solution. As to other criteria, they all select $\hat{\alpha} = 19$ with respectively $\hat{K} = 2$, 3 and 10 for ICL, BIC and AIC.

Table 8.2 gives the contingency table for the waveform data clustering obtained with our procedure. The three clusters are well related with the three true groups, with an error rate of 14.3%. Figure 8.4 plots the error rate in function of the number α of relevant variables, each curve corresponding to a fixed number of components in the mixture. Choosing a model with an ill-chosen subset of relevant variables deteriorates the clustering performance.

	cluster 1	cluster 2	cluster 3	total
group 1	**1331**	185	176	1692
group 2	95	99	**1459**	1653
group 3	65	**1494**	96	1655
total	1491	1778	1731	5000

Table 8.2: Contingency table for the clustering obtained with the slope heuristics.

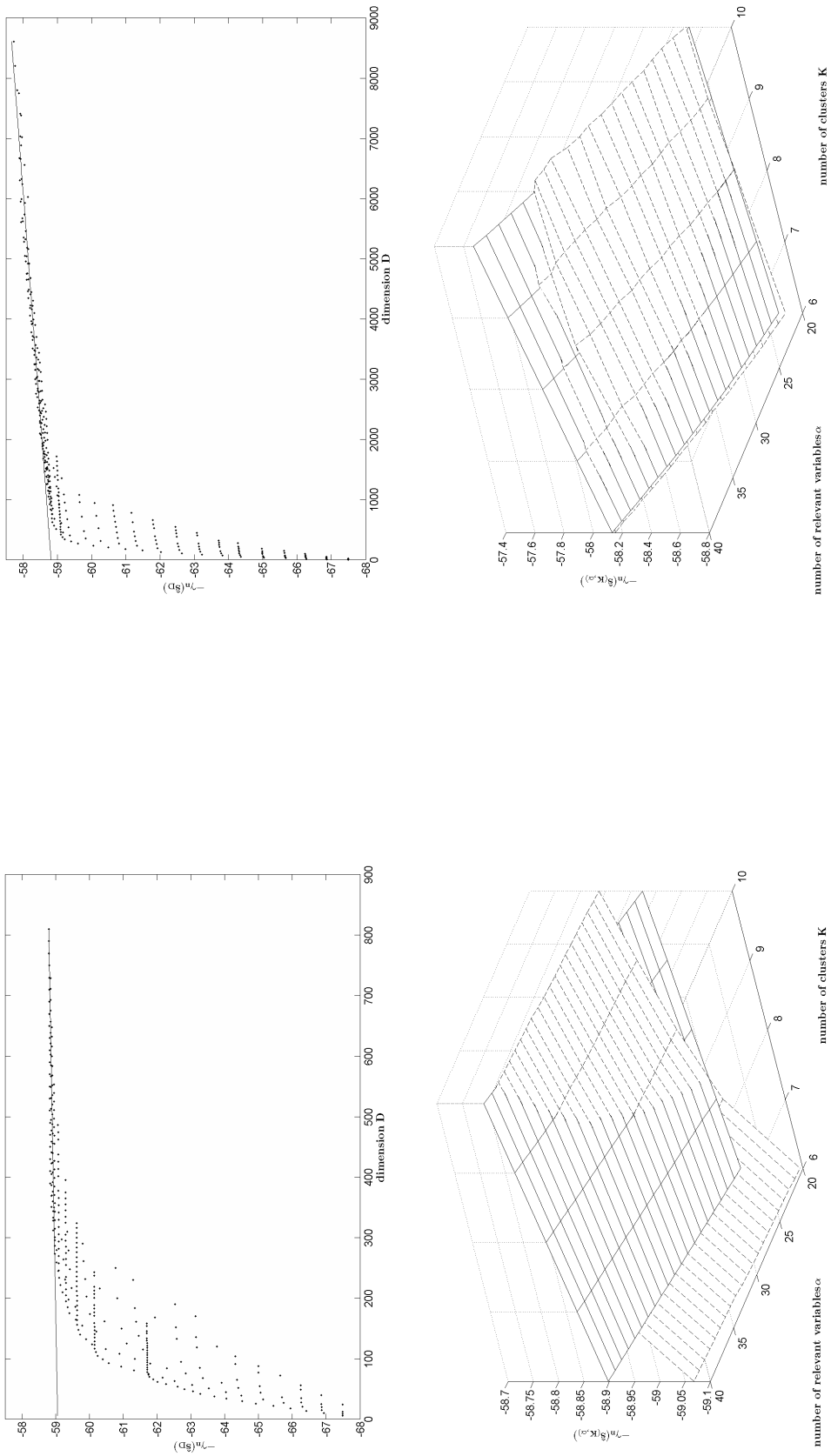


Figure 8.3: On the top, graphical representation of $D \mapsto -\gamma_n(\hat{s}_D)$ leading the estimation of $\hat{C}$ for the $[L_k B_k]$ collection (on the left) and the $[L_k C_k]$ collection (on the right). On the bottom, fitting of the dimension surface on $(K, \alpha) \in [|6, 10|] \times [|20, 40|] \mapsto -\gamma_n(\hat{s}_{(K,\alpha)})$ for the $[L_k B_k]$ collection (on the left) and the $[L_k C_k]$ collection (on the right).

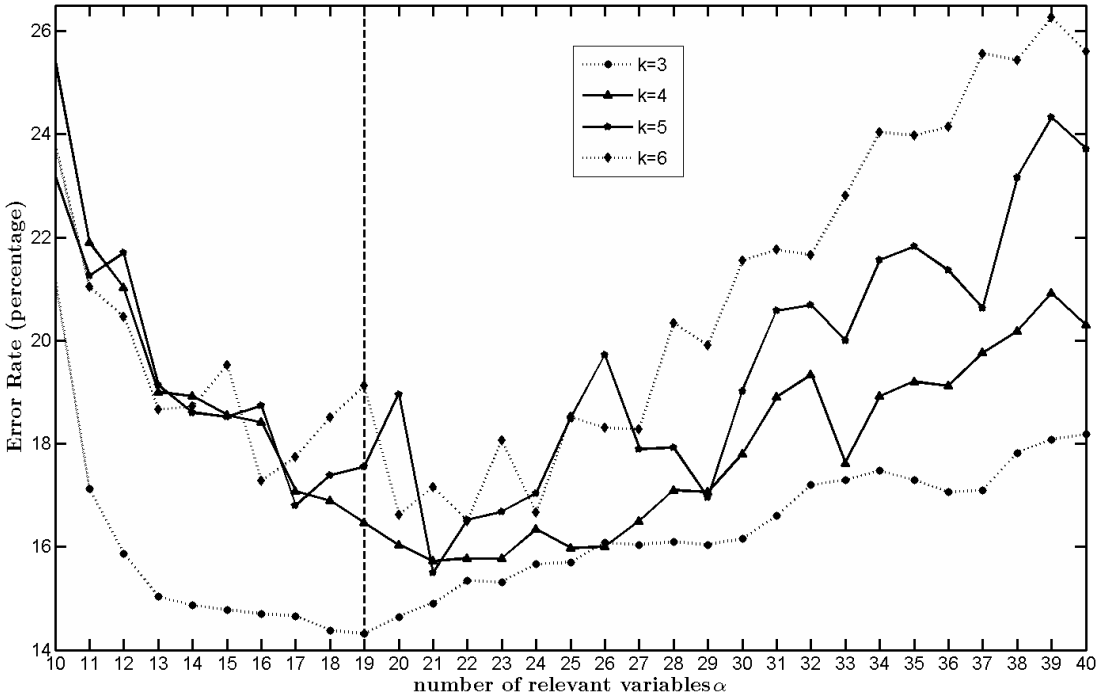


Figure 8.4: Evolution of the clustering error rate in function of the chosen number α of relevant variables. Each curve corresponds to a fixed number K of mixture components.

8.4.3 Curve clustering

An oil field production profile is the curve of oil production versus time. In the following, the term reserves (or ultimate reserves) denotes the amount of oil that is produced during the exploitation of an oil field. The reader is referred to Babusiaux et al. (2007) for more details about the exploration and the production of oil. It is well known by the oil industry that production profiles of large fields have a different shape than production profiles of little fields. Indeed, little fields tend to produce their reserves in a short time and early pass the production peak. On the contrary, large fields slowly produce their reserves and their production presents a plateau during several years at the top level. Figure 8.5 illustrates this behaviour with productions of three fields of different size. In order to compare the production profile shapes, we consider production profiles normalized by the amount of reserves contained in each field. The study's aim is to validate that a clustering of normalized production profiles is consistent with the values of the reserves variable.

The database is composed of several hundred of oil production profiles corresponding to hydrocarbon layers in the North Sea[1]. The data used in the procedure are obtained from

[1]Data is available on the website of the Norwegian Petroleum Directorate : `www.npd.no/engelsk/cwi/pbl/en/index.htm`, and the website of the English Department of Trade and Industry (DTI): `www.og.dti.gov.uk/fields/fields_index.htm`.

the original curves as follows. First, each production profile is normalized by the reserves of the corresponding field[2]. Ideally, it is desirable to proceed to the clustering on complete production profile. This is impossible since most of the fields are still in production. Figure 8.5 suggests that the beginning of the production curve is sufficient to distinguish different shapes in the curve family. Thus, we only consider the subsample composed of 180 fields which have started their production more than 64 months ago. Next, a discrete wavelet transform (DWT) is proceeded on each of these normalized curves. This decomposition has the advantage of giving information on each curve at different resolution levels. This transformation has already been used in curve classification (see for instance Berlinet et al., 2008). The reader is referred to Percival and Walden (2000) for details on the DWT. Let $\mathbf{W}_i$ be the wavelet coefficient vector of the i^{th} curve. Since the length of each curve is 64, the dimension of $\mathbf{W}_i$ is also 64. The vector $\mathbf{W}_i$ is defined by

$$\mathbf{W}_i = (\mathbf{V}'_{i6}, \mathbf{W}'_{i6}, \ldots, \mathbf{W}'_{i1})'$$

where $\mathbf{W}_{ij}$ is a vector of length $64/2^j$ which is composed of all the wavelet coefficients corresponding to the scale j. The coefficient $\mathbf{V}_{i6}$ is equal to the mean of the curve i divided by $\sqrt{64}$. The hierarchical structure of the DWT suggests a natural order of the wavelet coefficient variables according to their resolution. Indeed, $\mathbf{V}_{i6}$ and $\mathbf{W}_{i6}$ give informations about the general shape of the curve i whereas $\mathbf{W}_{i1}$ and $\mathbf{W}_{i2}$ give informations about details on it. We do not use the coefficients in $\mathbf{W}_{i1}$ and $\mathbf{W}_{i2}$ since they correspond to the finer resolution. We will see that the remaining coefficients are sufficient to propose a sensible clustering. Moreover, all the wavelet coefficient variables are centered and scaled to unit variance to make easier the fitting of the multidimensional Gaussian distribution $\mathcal{N}(0, \omega^2 I_{Q-\alpha})$ on the coefficient vectors which are not used for the clustering. These new coefficients are denoted $\widetilde{\mathbf{V}}_{i6}$ and $\widetilde{\mathbf{W}}_{ij}$ where $j \in \{3, \ldots, 6\}$. The procedure is performed on the sample $\mathbf{y}$ where $\mathbf{y}_i = \left(\widetilde{\mathbf{V}}'_{i6}, \widetilde{\mathbf{W}}'_{i6}, \ldots, \widetilde{\mathbf{W}}'_{i3}\right)'$ for an ordered model collection $[LB_k]$. This mixture collection avoids estimation problems when the variances are too small.

Figure 8.6 clearly shows the expected linear behavior of $D \mapsto -\gamma_n(\hat{s}_D)$ in large dimensions. The selected model minimizing the penalized criterion deduced from the slope heuristics has $\hat{K} = 3$ components and $\hat{\alpha} = 20$ clustering variables. Finally, the MAP rule gives a clustering of the curves. The clusters contain 31, 140 and 9 curves respectively. Figure 8.7 displays the mean cluster of the normalized production profiles in each cluster. Boxplot of the logarithm of the reserves variable for each cluster are displayed in Figure 8.8. The second cluster mainly corresponds to the large fields whereas the first and the third clusters contain fields of medium size and small size respectively. As expected, the shape of normalized production profiles can be explained by the reserves variable. The reader is referred to Michel (2008) for more details.

[2]The DTI does not provide estimations of reserves of their fields, consequently we use the IHS database `http://energy.ihs.com` for the english fields of the sample

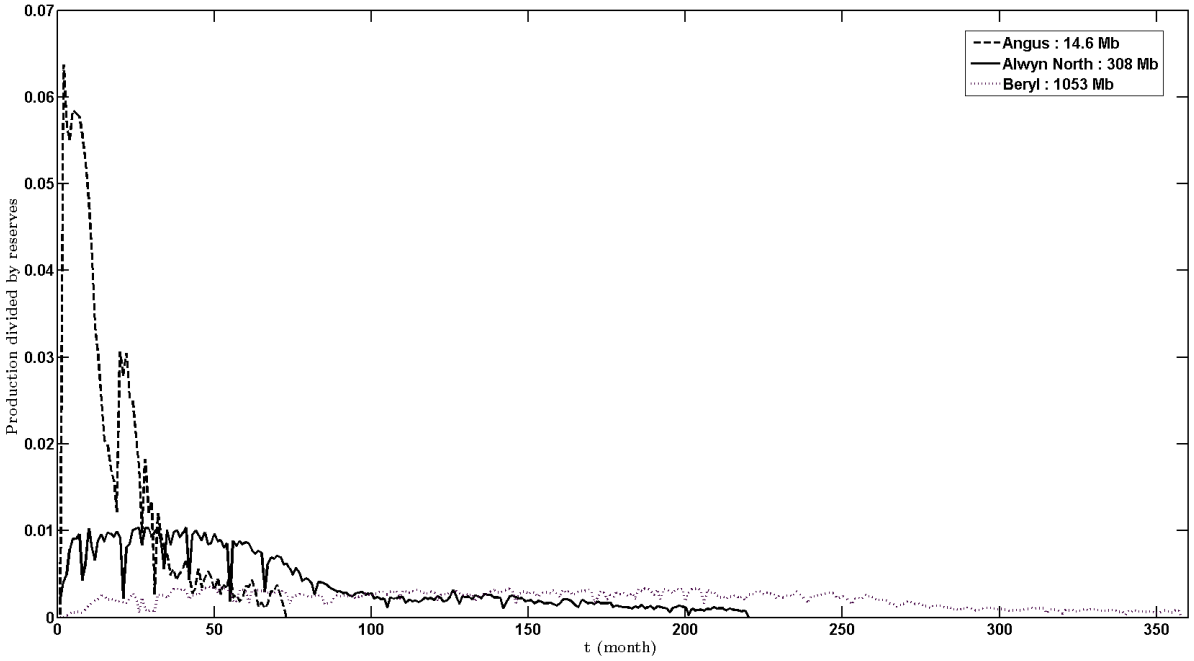


Figure 8.5: Oil production profiles normalized by reserves of three fields located in the North Sea.

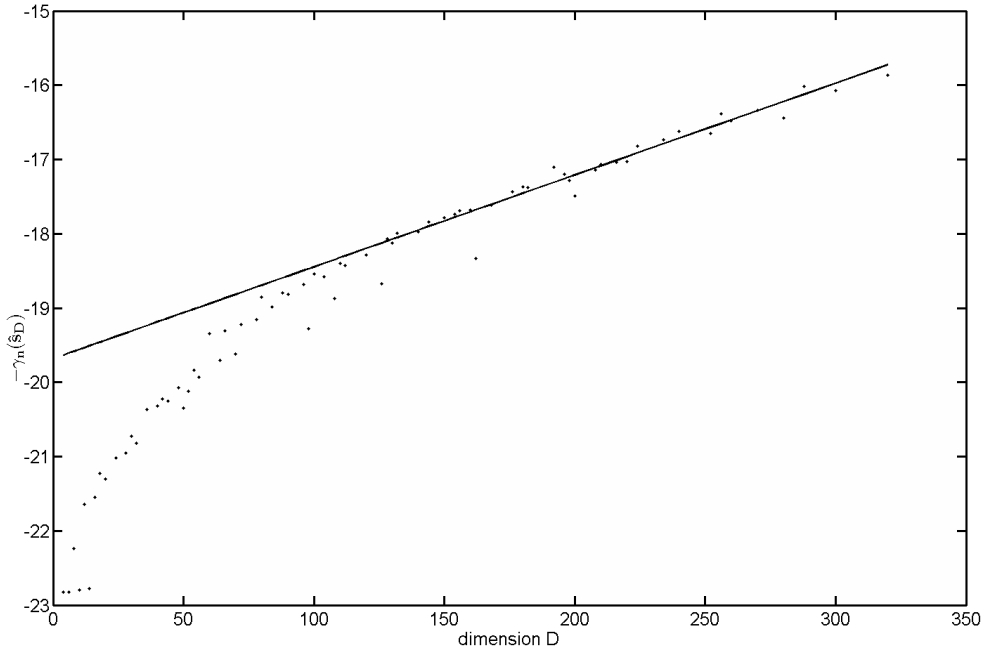


Figure 8.6: Slope method applied to the $[LB_k]$ collection for the wavelet coefficient curve data.

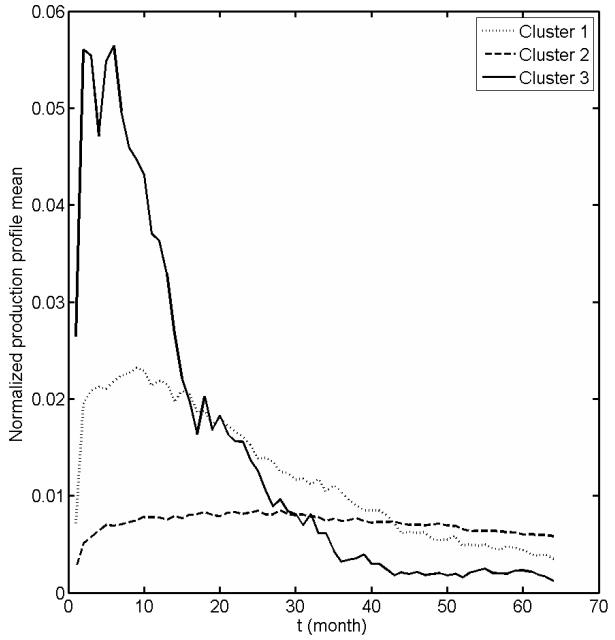


Figure 8.7: Normalized production profile means for each cluster.

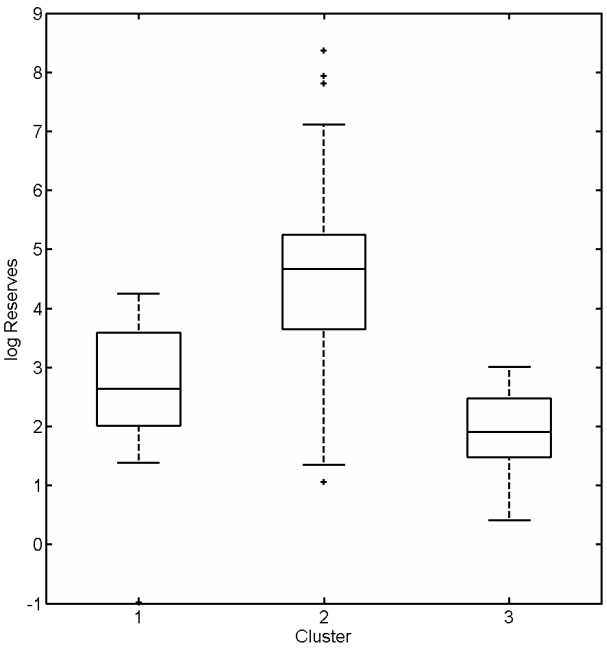


Figure 8.8: Boxplots of the logarithm of the reserves variable for each cluster.

8.4.4 Analysis of a transcriptome dataset

Currently, biologists are interested in gene functional analysis. It is usually considered that coexpressed genes are often implicated in the same biological function and consequently are potential candidate to be co-regulated genes. Thus biologists try to extract groups of coexpressed genes according to transcriptome datasets in order to characterize more precisely their biological functions. Moreover an experiment selection for the clustering is desirable to improve the clustering and its interpretation with a biological point of view.

Here we study a transcriptome dataset of *Arabidopsis thaliana* extracted from the database CATdb (Gagnot et al., 2008). To build this database, an identical statistical analysis for all transcriptome experiments has been performed to remove the technical biases (normalization) and to determine the gene significantly differentially expressed (differential analysis) between two conditions. In this differential analysis, we test if a gene is non-differentially expressed or not in the experiment j. For this test, a test statistic corresponding to the normalized differential expression and a p-value adjusted by the Bonferroni method are determined. Then a gene is declared to be differentially expressed when its Bonferroni p-value is lower than 0.05. The reader is referred to Gagnot et al. (2008) for a description of such an analysis and Lurin et al. (2004) for an application.

We focus on 305 genes of *Arabidopsis thaliana* studied on ten experiments which correspond to mutant conditions or different stress situations. These genes are declared differentially expressed in the two last experiments and non-differentially expressed in five

experiments. Table 8.3 gives the number of differentially expressed genes per experiment. Each gene is described with a vector $\mathbf{y}_i \in \mathbb{R}^{10}$, the component y_{ij} corresponding to the test statistic calculated in the experiment j for the differential analysis.

experiment number	1	2	3	4	5	6	7	8	9	10
number of differentially expressed genes	0	0	207	0	219	118	0	0	305	305

Table 8.3: Number of differentially expressed genes per experiment.

Since there is not a natural way to order the variables, our procedure for non-ordered variables is performed with the $[LC]$ mixture collection. The maximum number of components is fixed to $K_{\max} = 40$. After the estimation step, we notice that the function $D \mapsto -\gamma_n(\hat{s}_D)$ has a linear behavior for $D \geq 220$ (see Figure 8.9), thus the slope heuristics can be applied. The procedure selects a clustering with $\hat{K} = 8$ clusters based on the seven variables $\hat{\mathbf{v}} = \{3, 5, 6, 7, 8, 9, 10\}$. The eight clusters have different size (see Table 8.4) and the clustering shows some interesting similar behaviors of expression profiles (see Figure 8.10). A similar clustering can be found if all the variables are considered ($\alpha = 10$ fixed) but with the variable selection, the interpretation of the clustering is made clearer.

cluster	1	2	3	4	5	6	7	8
number of genes	60	39	47	12	82	51	9	5

Table 8.4: Number of genes per cluster.

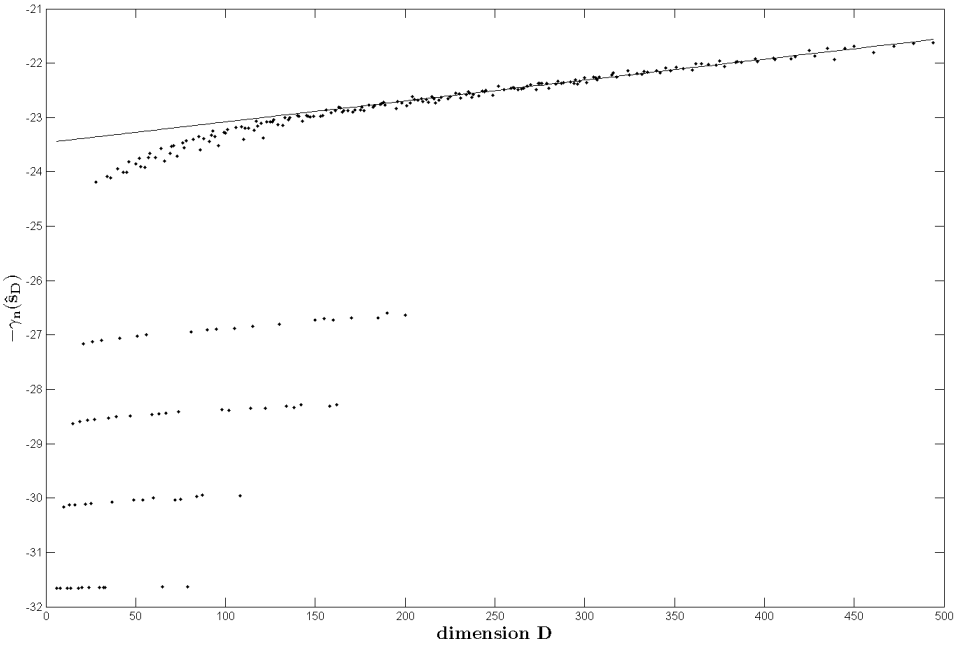


Figure 8.9: Penalty determination on the linear behavior of the function $D \mapsto -\gamma_n(\hat{s}_D)$.

First, we note that the two benchmark experiments (9 and 10) where all genes are differentially expressed are selected. Moreover, the three variables which are not selected for the clustering are three variables where all genes are non-differentially expressed. The average behavior of genes per cluster is the same in the irrelevant experiments 1, 2 and 4 since it is concentrated around zero. On the contrary, genes of Cluster 2 have a particular behavior in Experiments 7 and 8. Their expression difference decreases between the two experiments (7 and 8) whereas the genes of the other clusters have the same expression in these two experiments (see Figure 8.10). This remark may explain why the two Experiments 7 and 8 where all genes are non-differentially expressed are selected for the clustering while Experiments 1, 2 and 4 are not. To validate this explanation, t-test between Experiments 7 and 8 for the eight clusters have been performed at level 0.05. Only the test for Cluster 2 is significant (p-value$< 5.10^{-4}$). This clustering can help biologists to find gene biological functions. For instance, 12 genes for which biologists know any biological function are clustered with 27 other genes in Cluster 2. According to biological knowledge, the 27 genes have the same subcellular localisation (in plastid) and are involved in the photosynthesis of *Arabidopsis thaliana*. Thus the function of the 12 unknown genes is certainly related to the photosynthesis. This hypothesis remains to be confirmed by biologists experimentally.

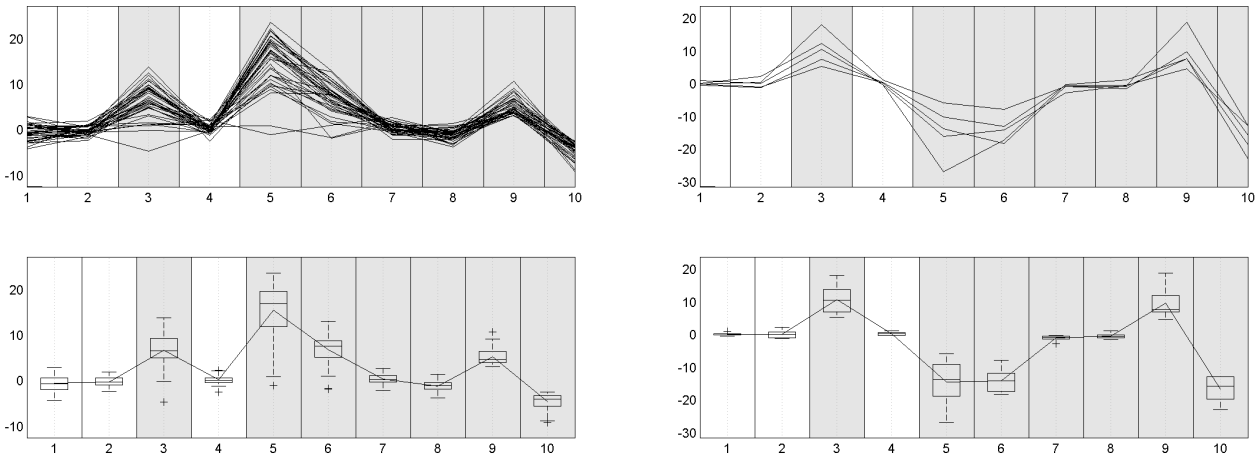


Figure 8.10: Graphical representation of gene profiles in Clusters 2 (on the left) and 8 (on the right). Relevant experiments are colored in grey.

8.5 Discussion

In this chapter, a methodology has been proposed to take into account the variable role in a clustering process in a model-based cluster analysis setting. The interest of our approach is to recast these two problems into a model selection problem where model collections are indexed by two quantities (K and $\mathbf{v}$). The practical use of the penalized likelihood criterion, proposed in Chapter 7, is based on a slope heuristics method allowing to calibrate

the multiplicative constant in the penalty term. The behaviour of this slope heuristics method has been studied on simulated and real datasets. It has been compared with standard asymptotic criteria, BIC, ICL, AIC, in a Gaussian mixture clustering context.

To calibrate the penalty function, the slope $\hat{C}$ is estimated on the restriction of the function $D \mapsto -\gamma_n(\hat{s}_D)$ to $D \geq D_0$, namely where this function has a linear behaviour. A robust regression is used, allowing to attenuate the influence of possible estimation errors. The threshold D_0 affecting the estimated slope $\hat{C}$, an unsuitable user choice of this threshold can imply an overpenalization or underpenalization. The slope $\hat{C}$ can be controlled by plotting the estimated slope in function of different values of the threshold. In a neighbourhood of $\hat{C}$, the slope is expected to be stable.

Commonly, the penalty functions are calibrated differently. The rule of thumb of the slope heuristics consists of assuming that the optimal penalty is twice the minimal penalty. This minimal penalty pen_{min} is such that the selected model has a too large dimension if the penalty $\text{pen} < \text{pen}_{\text{min}}$ and has a reasonable dimension if $\text{pen} > \text{pen}_{\text{min}}$. This *dimension jump* is a key phenomenon to determine the constant $\hat{C}$ corresponding to the minimal penalty. Thus it could be used to calibrate the penalty by $2\hat{C}$. In practice, the selected model dimension is plotted according to constant values $C \mapsto D(\hat{K}_{(C)}, \hat{\mathbf{v}}_{(C)})$ (see Figure 8.11) and $\hat{C}$ is determined by observing the dimension jump. This calibration method was applied for instance by Lebarbier (2005), Arlot (2007) and Villers (2007). In this chapter, we do not use this method to determine $\hat{C}$ since the dimension jump is often not quite clear (see Figure 8.11). Nevertheless, the estimated constant obtained with our procedure can be confirmed with the dimension plot.

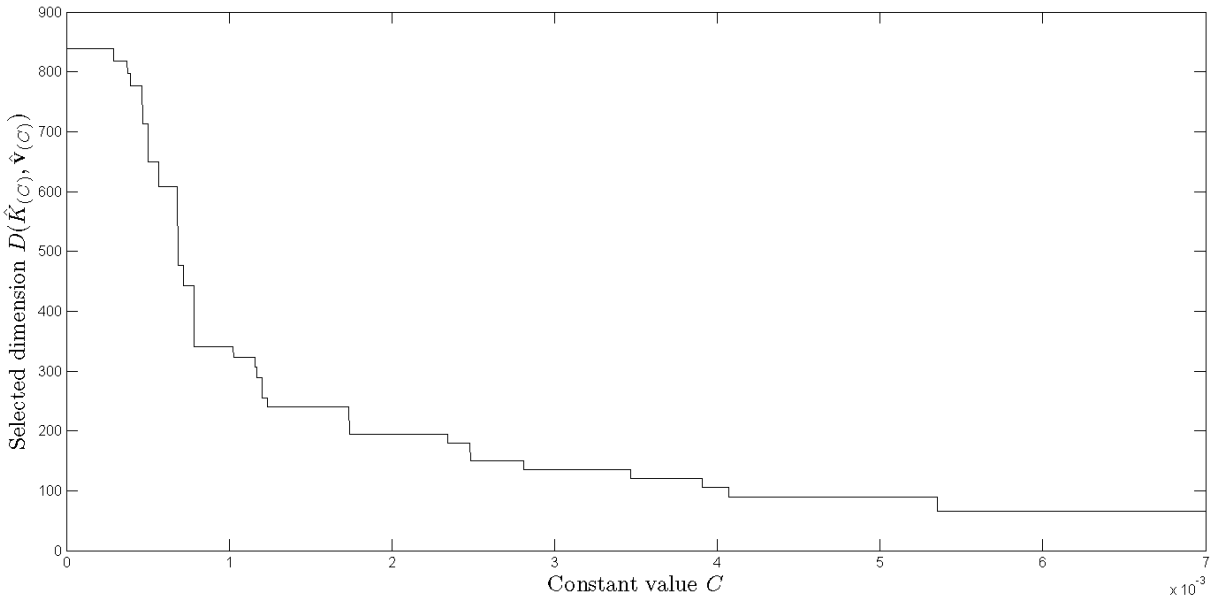


Figure 8.11: Graphical representation of $C \mapsto D(\hat{K}_{(C)}, \hat{\mathbf{v}}_{(C)})$ where $(\hat{K}_{(C)}, \hat{\mathbf{v}}_{(C)})$ is the selected model associated to the slope value C.

Our variable selection method can be efficiently applied when the variables could be ordered as illustrated in the curve clustering study in Section 8.4.3. When the variables cannot be ordered in a natural way, our method is difficult to use when the number of variables is too large. An exhaustive research of the best model becomes untractable. A possible way to circumvent this problem is to find a convenient strategy allowing to run the estimation step only for a model subset with a reasonable size.

In all cases, the user has to check that the linear behaviour of $D \mapsto -\gamma_n(\hat{s}_D)$ is observed and also that the dimension model surface fits the maximum likelihood surface on high dimensions (see Figure 8.3). If it is not the case, several explanations can be given. First, the model dimension can be too low, namely the maximum number of components $K_{\max}$ has to be increased. The problem can be also related to the model family choice. Roughly speaking, the family model leads to a stabilization of the bias in large dimension only if the family model efficiently approaches the true density. Otherwise, the collection of models has to be changed in order to obtain a better fitting between the two surfaces.

Our theoretical and practical results could be extended for most of the twenty-eight Gaussian mixture shapes proposed by Celeux and Govaert (1995). In particular, without variable selection ($\alpha = Q$), our method allows us to select the number of clusters which is the fundamental problem in model-based clustering. It could be envisaged to adapt our works to select also the Gaussian mixture form. Some preliminary results on simulated examples are presented in Chapter 9.

Chapter 9

Conclusion and perspectives

The construction of a non asymptotic penalized criterion is proposed in Chapter 7 in the framework of variable selection for clustering with Gaussian mixture models. This approach to study a Gaussian mixture model selection problem with a non asymptotic point of view is new. This theoretical construction requires to control the bracketing entropies of Gaussian density families. The theoretical results proved for four Gaussian mixture forms can be extended for spherical mixtures, diagonal mixtures and mixtures with the form $[p\_L\_C\_]$ (see Table 1.1). The extension to the mixture forms $[p\_L\_D\_A\_D'\_]$ needs to construct a countable covering on the orthogonal matrices with a convenient cardinal in order to obtain the model dimension term in the penalty function.

This penalized criterion depending on an unknown multiplicative constant, the slope heuristics proposed by Birgé and Massart (2006) is performed to calibrate the penalty function. For the variable selection, our procedure which is competitive for ordered variables becomes useless if the variables are non-ordered and their number is large. A strategy allowing to visit cleverly only a subset of models remains to be constructed in order to study more complex datasets with our method.

Difficulties of the slope heuristics carrying out:

The slope heuristics proposed by Birgé and Massart (2006) is performed to calibrate the unknown multiplicative constant in order to use our model selection criterion. This heuristics is applied in a framework where the model families are indexed by two entities. In Chapter 8, we gave some practical rules for an efficient use of this heuristics. But some encountered numerical difficulties need additional attention. First, this method is sensitive to the optimality of parameter estimates. And, the robust regression, used to reduce the influence of suboptimal estimates on the slope calibration, is not always sufficient. Second, parameter estimation has to be carried out for models with large dimensions in order to observe the linear part, requiring samples with large sizes. Third, the slope calibration depends on the threshold D_0 chosen by the user. Tools have to be made available in order to control such a choice. The improvement of the slope heuristics use taking all those points into account in the mixture context is the subject of a work in progress, in collaboration

with J.-P. Baudry and B. Michel.

Now, we focus on the potential use of our method for the selection of the number of mixture component and the mixture form in a Gaussian mixture context. Some exploratory examples are presented.

Selection of the number of mixture components

In the context of clustering with Gaussian mixture models without variable selection, the theoretical results in Chapter 7 allow us to specify the penalized criterion form for the selection of the number of components, conditionally to a fixed mixture form. In practice, the slope heuristics can still be used to calibrate the unknown constant. This determination of the component number for the mixture is a fundamental problem in clustering. We propose a non asymptotic penalized criterion to solve this problem, where asymptotic criteria as BIC and ICL are usually used. The following example illustrates the behaviour differences between BIC, ICL and our criterion.

The dataset consists of n data points from a mixture of seven equiprobable Gaussian distributions $\mathcal{N}(\mu_k, \Sigma_k)$ with $\mu_1 = (2,1)$, $\mu_2 = (2,8)$, $\mu_3 = (2,8)$, $\mu_4 = (8,4)$, $\mu_5 = (9,4)$, $\mu_6 = (10,4)$, $\mu_7 = (9,8)$ and $\Sigma_1 = \mathrm{diag}(1,1)$, $\Sigma_2 = \mathrm{diag}(0.5,3)$, $\Sigma_3 = \mathrm{diag}(3,0.5)$, $\Sigma_4 = \mathrm{diag}(0.25,4) = \Sigma_5 = \Sigma_6$, $\Sigma_7 = \mathrm{diag}(2,0.5)$. The true number of components is $K_0 = 7$ and the true mixture form is $m_0 = [pL_kB_k]$. This mixture form m_0 is fixed and the sample size n takes its value among $\{200, 500, 1000, 2000\}$. A graphical representation of the dataset with $n = 2000$ is given in Figure 9.1. For each value of the sample size n, 1000 samples are studied and the oracle estimator is evaluated using a Monte Carlo procedure. In the four scenarii, this oracle estimator is equal to $\hat{K}_{\text{oracle}} = 7$.

The selected numbers of components $\hat{K}$ obtained with the three criteria in the four scenarii are given in Table 9.1. Our non asymptotic penalized criterion achieves its aim (the oracle) easier than BIC for samples with small sizes (recall that the selected number of components converges to the true model $K_0 = 7$). The selection with ICL leads to a clustering with four clusters, its objective being to choose a mixture given a good data clustering and not to select the true model. Groups 2 and 3 constitute a cluster and two clusters are constructed with Groups 4, 5, 6 and 7. Note that the choices of BIC and ICL are more and more stable with the increase of the sample size whereas the slope heuristics tends to select sometimes a greater number of components. It results from an underpenalization due to the estimation problems as previously mentioned.

Selection of the number of mixture components and the mixture form

Theoretically, the penalty functions constructed in Chapter 7 have the same form for each studied mixture forms, apart an unknown constant which has to be calibrated in practice. Can we select the number of components K and the form m of a mixture with a non asymptotic penalized criterion of the form

$$(K, m) \mapsto \gamma_n(\hat{s}_{(K,m)}) + 2\hat{C}D(K, m)$$

where the constant $\hat{C}$ is evaluated using the slope heuristics? By way of exploratory analysis, we study the previous example varying the mixture form as well. The selected

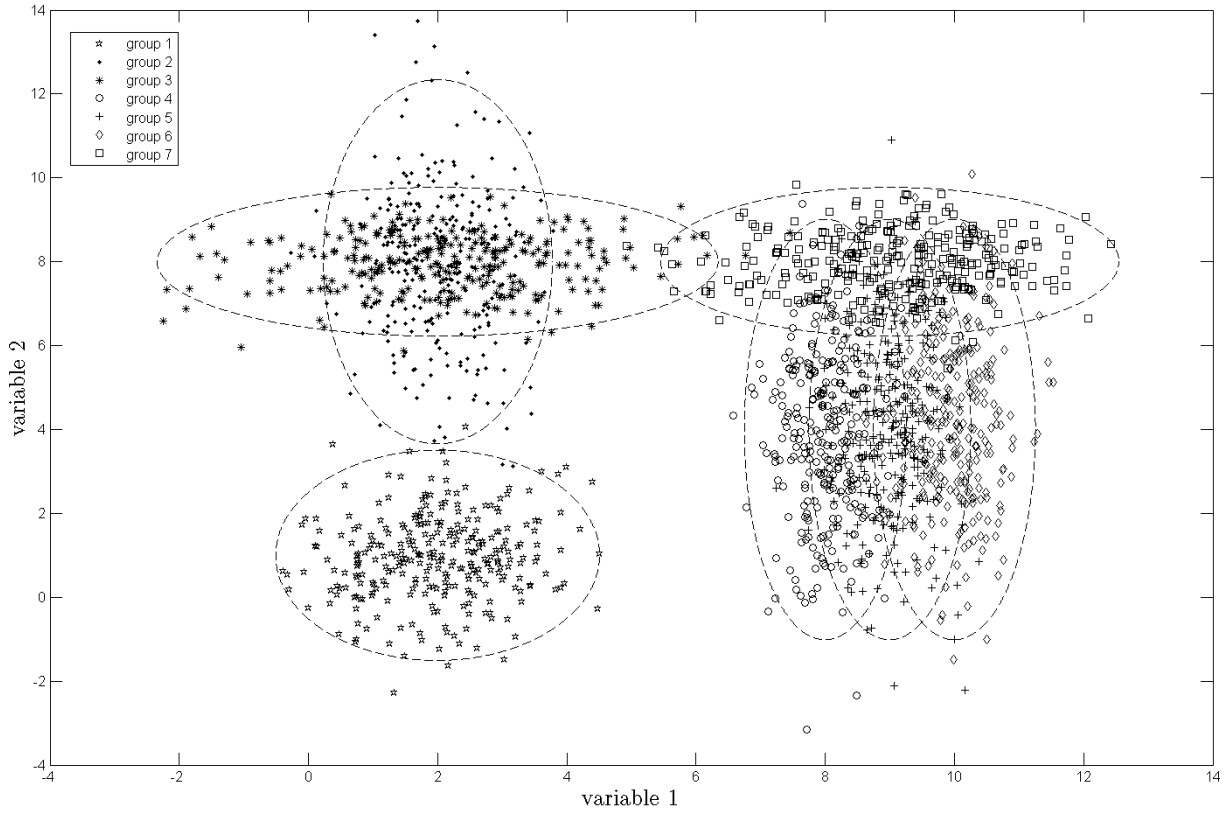


Figure 9.1: Representation of the dataset according to the two first variables with $n = 2000$ points.

n	$\hat{K}$	2	3	4	5	6	7	8	9	10	11	12	13	14
200	ICL	367	169	**436**	27	1								
	BIC		4	402	**433**	149	11	1						
	our criterion			74	268	**369**	233	32	12	9	3			
500	ICL	193	119	**663**	25									
	BIC			8	133	**520**	334	5						
	our criterion				15	216	**639**	92	20	8	7	2	1	
1000	ICL	77	74	**843**	6									
	BIC				3	238	**731**	28						
	our criterion					55	**806**	93	25	3	8	9	1	
2000	ICL	25	49	**923**	3									
	BIC					39	**902**	55	4					
	our criterion					11	**818**	97	32	5	14	8	9	6

Table 9.1: Number of times among 1000 tests the number of mixture components $\hat{K}$ is selected by BIC, ICL and our penalized criterion.

criterion	$\hat{K}$	$\hat{m}$	n=500	n=1000	n=2000
ICL	3	$[p_k L_k B_k]$	**95**	**100**	**100**
		$[pL_k C_k]$	1		
	4	$[pLB_k]$	1		
		$[p_k LB_k]$	1		
		$[p_k L_k B_k]$	2		
BIC	4	$[p_k LB_k]$	2		
		$[p_k L_k B_k]$	5		
	5	$[pL_k I]$	5		
		$[p_k L_k I]$	3		
		$[pLB_k]$	1		
	6	$[pL_k I]$	3		
		$[pLB_k]$	36	13	2
		$[p_k LB_k]$	1	3	6
	7	$[pL_k I]$	2		
		$[pLB_k]$	**39**	**84**	**77**
		$[pL_k B_k]$			12
		$[pLDA_k D']$			1
	8	$[pL_k I]$	2		
		$[pLB_k]$	1		2
Our criterion	4	$[p_k L_k B_k]$	3		
	5	$[p_k L_k I]$	1		
		$[p_k L_k B_k]$	1		
	6	$[pLB_k]$	19	4	
		$[pL_k B_k]$	1		
		$[p_k LB_k]$	4	4	5
		$[p_k L_k B_k]$	2		2
		$[pL_k C_k]$			1
	7	$[pLB_k]$	**59**	**81**	**48**
		$[pL_k B_k]$		5	40
		$[p_k LB_k]$		3	
		$[p_k L_k B_k]$			1
		$[pLDA_k D']$		2	2
	8	$[pL_k I]$	1		
		$[p_k L_k I]$	2		
		$[pLB_k]$	3	1	
		$[p_k L_k B_k]$			1
	12	$[p_k L_k I]$	1		
	13	$[pL_k I]$	1		
	14	$[p_k L_k I]$	1		

Table 9.2: Number of times among 100 tests the mixture model $(\hat{K}, \hat{m})$ is selected by BIC, ICL and our penalized criterion.

model $(\hat{K}, \hat{m})$ obtained with the three criteria for 100 samples of size n belonging to $\{500, 1000, 2000\}$ are given in Table 9.2. ICL selects a mixture with three clusters and the form $[p_k L_k B_k]$, which is not the true mixture form. The more the sample size increases, the more BIC selects a mixture with seven clusters and a form $[pLB_k]$. It does not select the true mixture form but we can note that the volumes of the true components are very close. Beyond its numerical problems explained previously, our penalized criterion has a similar behaviour than BIC for samples with size $n = 500$ and $n = 1000$. On the contrary, for $n = 2000$, our penalized criterion hesitates over the mixture form, and seems to be more sensitive to the studied dataset.

We want to continue this work for the selection of the number of components but also for the choice of the mixture form, with both a theoretical and a practical point of views.

Bibliographie

Abraham, C., Cornillon, P. A., Matzner-Løber, E., and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics. Theory and Applications*, 30(3) :581–595.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19 :716–723. System identification and time-series analysis.

Alizadeh, A. A., Eiden, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J. J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403 :503–511.

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *PNAS*, 97(18) :10101–10106.

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley, 3rd edition.

Arlot, S. (2007). *Réechantillonnage et sélection de modèles*. PhD thesis, Université Paris-Sud 11.

Arlot, S. and Massart, P. (2008). Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*. To appear.

Babusiaux, D., Barreau, S., and Bauquis, P.-R. (2007). *Oil and gas exploration and production, reserves, costs, contracts*. Technip, Paris.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3) :803–821.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113 :301–413.

Ben-Dor, A., Shamir, R., and Yakhini, Z. (1999). Clustering gene expression patterns. *Journal of Computational Biology*, 6(3-4) :281–297.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B. Methodological*, 57(1) :289–300.

Berlinet, A., Biau, G., and Rouvière, L. (2008). Functional classification with wavelets. *Annales de l'ISUP*. To appear.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7) :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 and Data Analysis*, 41(3-4) :561–575.

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.

Birgé, L. and Massart, P. (2001a). Gaussian model selection. *Journal of the European Mathematical Society*, 3(3) :203–268.

Birgé, L. and Massart, P. (2001b). A generalized C_p criterion for Gaussian model selection. Prépublication n°647, Universités de Paris 6 et Paris 7.

Birgé, L. and Massart, P. (2006). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1-2) :33–73.

Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York.

Blake, C., Keogh, E., and Merz, C. (1999). *UCI Repository of Machine Learning Algorithms Databases*.

Bø T. H., Dysvik, B., and Jonassen, I. (2004). LSimpute : accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Research*, 32(3) :e34.

Bouveyron, C., Girard, S., and Schmid, C. (2007). High-Dimensional Data Clustering. *Computational Statistics and Data Analysis*, 52(1) :502–519.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International, Belmont, California.

Brock, G. N., Shaffer, J. R., Blakesley, R. E., Lotz, M. J., and Tseng, G. C. (2008). Which missing value imputation method to use in expression profiles : a comparative study and two selection schemes. *BMC Bioinformatics*, 9(12).

Brusco, M. J. and Cradit, J. D. (2001). A variable selection heuristic for k-means clustering. *Psychometrika*, 66(2) :249–270.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference : A Practical Information-Theoretic Approach*. Springer-Verlag, New York, 2nd edition.

Castellan, G. (1999). Modified Akaike's criterion for histogram density estimation. Technical report, Université Paris-Sud 11.

Castellan, G. (2003). Density estimation via exponential model selection. *IEEE Transactions on Information Theory*, 49(8) :2052–2060.

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 :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.

Crowe, M. L., Serizet, C., Thareau, V., Aubourg, S., Rouze, P., Hilson, P., Beynon, J., Weisbeek, P., van Hummelen, P., Reymond, P., Paz-Ares, J., Nietfeld, W., and Trick, M. (2003). CATMA : a Complete Arabidopsis GST database. *Nucleic Acids Research*, 31(1) :156–158.

Dash, M., Choi, K., Scheuermann, P., and Liu, H. (2002). Feature Selection for Clustering - A Filter Solution. *Proceedings of the Second IEEE International Conference on Data Mining*, pages 115–122.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B.*, 39(1) :1–38.

Devaney, M. and Ram, A. (1997). Efficient feature selection in conceptual clustering. *Machine Learning : Proceedings of the Fourteenth International Conference*, pages 92–97.

Dopazo, J. and Carozo, J. M. (1997). Phylogenetic reconstruction using a growing neural network that adopts the topology of a phylogenetic tree. *Journal of molecular evolution*, 44(2) :226–233.

Dudoit, S., Shaffer, J.-P., and Boldrick, J.-C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1) :71–103.

Dudoit, S. and van der Laan, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer, New York.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25) :14863–14868.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture distributions*. Chapman & Hall, London.

Foissac, S., Bardou, P., Moisan, A., Cros, M. J., and Schiex, T. (2003). EUGENE'HOM : A generic similarity-based gene finder using multiple homologous sequences. *Nucleic Acids Research*, 31(13) :3742–3745.

Fowlkes, E. B., Gnanadesikan, R., and Kettenring, J. R. (1988). Variable selection in clustering. *Journal of Classification*, 5(2) :205–228.

Fraley, C. and Raftery, A. E. (2003). Enhanced software for model-based clustering, density estimation, and discriminant analysis : MCLUST. *Journal of Classification*, 20(2) :263–286.

Friedland, S., Niknejad, A., and Chihara, L. (2006). A simultaneous reconstruction of missing data in DNA microarrays. *Linear Algebra and its applications*, 416 :8–28.

Friedman, J. H. and Meulman, J. J. (2004). Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society, Series B*, 66(4) :815–849.

Gagnot, S., Tamby, J.-P., Martin-Magniette, M.-L., Bitton, F., Taconnat, L., Balzergue, S., Aubourg, S., Renou, J.-P., Lecharny, A., and Brunaud, V. (2008). CATdb : a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic Acids Research*, 36(Database Issues) :986–990.

Gan, X., Liew, A. W.-C., and Yan, H. (2006). Microarray missing data imputation based on a set theoretic framework and biological knowledge. *Nucleic Acids Research*, 34(5) :1608–1619.

García-Escudero, L. A. and Gordaliza, A. (2005). A proposal for robust curve clustering. *Journal of Classification*, 22(2) :185–201.

Gasch, A. and Eisen, M. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3(59) :1–22.

Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *The Annals of Statistics*, 28(4) :1105–1127.

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5) :1233–1263.

Ghosh, D. and Chinnaiyan, A. (2002). Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, 18(2) :275–286.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3 :1157–1182.

Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., and Brown, P. (2000). Gene Shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1(2) :1–21.

Herrero, J., Valencia, A., and Dopazo, J. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17(2) :126–136.

Hilson, P., Allemeersch, J., Altmann, T., Aubourg, S., Avon, A., Beynon, J., Bhalerao, R. P., Bitton, F., Caboche, M., Cannoot, B., Chardakov, V., Cognet-Holliger, C., Colot, V., Crowe, M., Darimont, C., Durinck, S., Eickhoff, H., de Longevialle, A. F., Farmer, E. E., Grant, M., Kuiper, M. T., Lehrach, H., Léon, C., Leyva, A., Lundeberg, J., Lurin, C., Moreau, Y., Nietfeld, W., Paz-Ares, J., Reymond, P., Rouzé, P., Sandberg, G., Segura, M., Serizet, C., Tabrett, A., Taconnat, L., Thareau, V., Van Hummelen, P., Vercruysse, S., Vuylsteke, M., Weingartner, M., Weisbeek, P., Wirta, V., Wittink, F., Zabeau, M., and Small, I. (2004). Versatile gene-specific sequence tags for Arabidopsis functional genomics : transcript profiling and reverse genetics applications. *Genome Research*, 14(10B) :2176–2189.

Hu, J., Li, H., Waterman, M., and Zhou, X. (2006). Integrative missing value estimation for microarray data. *BMC Bioinformatics*, 7 :449.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons Inc., New York.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data Clustering : A Review. *ACM Computing Surveys*, 31(3) :254–323.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462) :397–408.

Jammes, F., Lecomte, P., Almeida-Engler, J., Bitton, F., Martin-Magniette, M.-L., Renou, J.-P., Abad, P., and Favery, B. (2005). Genome-wide expression profiling of the host response to root-knot nematode infection in Arabidopsis. *The Plant Journal*, 44(3) :447–458.

Jiang, D., Pei, J., and Zhang, A. (2003). DHC : A Density-based Hierarchical Clustering Method for Time Series Gene Expression Data. *Proceeding of the 3rd IEEE Symposium on Bioinformatics and Bioengineering*.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data : A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11) :1370–1386.

Jornsten, R., Wang, H.-Y., Welsh, W. J., and Ouyang, M. (2005). DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics*, 21(22) :4155–4161.

Jouve, P.-E. and Nicoloyannis, N. (2005). A filter feature selection method for clustering. *Proceedings of International Symposium on Methodologies for Intelligent Systems*, pages 583–593.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430) :773–795.

Kaufman, L. and Rousseeuw, P. J. (1987). Clustering by means of Medoids. *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, pages 405–416.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā. The Indian Journal of Statistics. Series A*, 62(1) :49–66.

Kerr, M. K., Afshari, A., Bennett, L., Bushel, P., Martinez, J., Walker, N. J., and Churchill, A. (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica*, 12(1) :203–217.

Kim, D.-W., Lee, K.-Y., Lee, K. H., and Lee, D. (2007). Towards clustering of incomplete microarray data without the use of imputation. *Bioinformatics*, 23(1) :107–113.

Kim, H., Golub, G. H., and Park, H. (2005). Missing value estimation for DNA microarray gene expression data : local least squares imputation. *Bioinformatics*, 21(2) :187–198.

Kim, S., Tadesse, M. G., and Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, 93(4) :877–893.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2) :273–324.

Kohonen, T. (1997). *Self-Organizing Maps*. Springer-Verlag, Berlin, 2nd edition.

Law, M. H., Figueiredo, M. A. T., and Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9) :1154–1166.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85(4) :717–736.

Lebarbier, E. and Mary-Huard, T. (2006). Une introduction au critère BIC : fondements théoriques et interprétation. *Journal de la SFdS*, 147(1) :39–57.

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics*, 21 :20–24.

Little, R. J. A. and Rubin, D. B. (1986). *Statistical analysis with missing data*. John Wiley & Sons, Inc., New York, USA.

Lurin, C., Andréas, C., Aubourg, S., Bellaoui, M., Bitton, F., Bruyère, C., Caboche, M., Debast, C. Gualberto, J., Hoffmann, B., Lecharny, A. Le Ret, M., Martin-Magniette, M.-L., Mireau, H., Peeters, N., Renou, J.-P., Szurek, B., Taconnat, L., and Small, I. (2004). Genome-wide analysis of arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell*, 16(8) :2089–2103.

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research*, 34(4) :1261–1269.

Mallows, C. (1973). Some comments on C_p. *Technometrics*, 37(4) :362–372.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. Academic Press, London.

Martin-Magniette, M.-L. and Robin, S. (2004). *Informatique pour l'analyse du transcriptome*. Hermès Science. Chapter 3 : Techniques statistiques pour l'analyse du transcriptome.

Massart, P. (2007). *Concentration inequalities and model selection*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2008). Variable Selection for Clustering with Gaussian Mixture Models. *Biometrics*. To appear.

McLachlan, G., Bean, R., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3) :413–422.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley-Interscience, New York.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models : Inference and Applications to clustering*. Marcel Dekker Inc., New York.

McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and extensions*. John Wiley & Sons Inc., New York.

McLachlan, G. J., Peel, D., Basford, K. E., and Adams, P. (1999). The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software*, 4(2) :1–14.

McQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1 :281–297.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9) :1194–1206.

Michel, B. (2008). *Modélisation de la production d'hydrocarbures dans un bassin pétrolier*. PhD thesis, Université Paris-Sud 11.

Miller, A. J. (1990). *Subset Selection in Regression*. Chapman and Hall, London.

Newcomb, S. (1886). A Generalized Theory of the Combination of Observations so as to Obtain the Best Result. *American Journal of Mathematics*, 8(4) :343–366.

Oba, S., Sato, M. A., Takemasa, I., Monden, M., Matsubara, K., and Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16) :2088–2096.

Ouyang, M., Welsh, W. J., and Georgopoulos, P. (2004). Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, 20(6) :917–923.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., and Brazma, A. (2007). ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Research*, 35(Database issue) :747–750.

Parmigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger, S. L. (2003). *The analysis of gene expression data*. Springer, New York.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, A 185 :71–110.

Percival, D. B. and Walden, A. T. (2000). *Wavelet methods for time series analysis*. Cambridge University Press, Cambridge.

Raftery, A. E. and Dean, N. (2006a). CLUSTVARSEL : Variable Selection for Model-Based Clustering. `http://cran.r-project.org/web/packages/clustvarsel/index.html`.

Raftery, A. E. and Dean, N. (2006b). Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, 101(473) :168–178.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3) :581–592.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall, London.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2) :461–464.

Sehgal, M. S. B., Gondal, I., and Dooley, L. S. (2005). Collateral missing value imputation : a new robust missing value estimation algorithm for microarray data. *Bioinformatics*, 21(10) :2417–23.

Serre, D. (2002). *Matrices*. Springer-Verlag, New York.

Sharan, R., Elkon, R., and Shamir, R. (2002). Cluster analysis and its applications to gene expression data. In *Ernst Schering Workshop on Bioinformatics and Genome Analysis*. Springer Verlag.

Sharan, R. and Shamir, R. (2000). CLICK : A clustering algorithm with applications to gene expression analysis. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 307–316.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470) :602–617.

Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Institut des Hautes Études Scientifiques. Publications Mathématiques*, 81 :73–205.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126(3) :505–563.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps : Methods and application to hematopoietic differentiation. *PNAS*, 96(6) :2907–2912.

Tarpey, T. and Kinateder, K. K. J. (2003). Clustering functional data. *Journal of Classification*, 20(1) :93–114.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics*, 22 :281–285.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, Chichester.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6) :520–525.

Tuikkala, J., Elo, L., Nevalainen, O., and Aittokallio, T. (2006). Improving missing value estimation in microarray data with gene ontology. *Bioinformatics*, 22(5) :566–572.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Verzelen, N. (2008). Data-driven neghborhood selection of a Gaussian field. In preparation.

Villers, F. (2007). *Tests et sélection de modèles pour l'analyse de données protéomiques et transcriptomiques*. PhD thesis, Université Paris-Sud 11.

Wallace, C. S. and Dowe, D. L. (1994). Intrinsic classification by MML- the Snob program. *In the 7th Australian Joint Conference on Artificial Intelligence*, pages 37–44.

Wang, X., Li, A., Jiang, Z., and Feng, H. (2006). Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme. *BMC Bioinformatics*, 7(32).

Xu, Y., Olman, V., and Xu, D. (2001). Minimum Spanning Trees for Gene Expression Data Clustering. *Genome Informatics*, 12 :24–33.

Yang, Y., Dudoit, S., Luu, P., Lin, D., Peng, V., Ngai, J., and Speed, T. (2002). Normalization for cDNA microarray data : a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(15) :1–10.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzz, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10) :977–987.

Zhou, X., Wang, X., and Dougherty, E. R. (2003). Missing-value estimation using linear and non-linear regression with Bayesian gene selection. *Bioinformatics*, 19(17) :2302–2307.

TITRE : Sélection de variables pour la classification non supervisée par mélanges gaussiens. Application à l'étude de données transcriptomes.

Résumé : Nous nous intéressons à la sélection de variables en classification non supervisée par mélanges gaussiens. Ces travaux sont en particulier motivés par la classification de gènes à partir de données transcriptomes. Dans les deux parties de cette thèse, le problème est ramené à celui de la sélection de modèles.
Dans la première partie, le modèle proposé, généralisant celui de Raftery et Dean (2006b) permet de spécifier le rôle des variables vis-à-vis du processus de classification. Ainsi les variables non significatives peuvent être dépendantes d'une partie des variables retenues pour la classification. Ces modèles sont comparés grâce à un critère de type BIC. Leur identifiabilité est établie et la consistance du critère est démontrée sous des conditions de régularité. En pratique, le statut des variables est obtenu grâce à un algorithme imbriquant deux algorithmes descendants de sélection de variables pour la classification et pour la régression linéaire. L'intérêt de cette procédure est en particulier illustré sur des données transcriptomes. Une amélioration de la modélisation du rôle des variables, consistant à répartir les variables déclarées non significatives entre celles dépendantes et celles indépendantes des variables significatives pour la classification, est ensuite proposée pour pallier une surpénalisation de certains modèles. Enfin, la technologie des puces à ADN engendrant de nombreuses données manquantes, une extension de notre procédure tenant compte de l'existence de ces valeurs manquantes est suggérée, évitant leur estimation préalable.
Dans la seconde partie, des mélanges gaussiens de formes spécifiques sont considérés et un critère pénalisé non asymptotique est proposé pour sélectionner simultanément le nombre de composantes du mélange et l'ensemble des variables pertinentes pour la classification. Un théorème général de sélection de modèles pour l'estimation de densités par maximum de vraisemblance, proposé par Massart (2007), est utilisé pour déterminer la forme de la pénalité. Ce théorème nécessite le contrôle de l'entropie à crochets des familles de mélanges gaussiens multidimensionnels étudiées. Ce critère dépendant de constantes multiplicatives inconnues, l'heuristique dite « de la pente » est mise en œuvre pour permettre une utilisation effective de ce critère.
Mots clés : Sélection de variables, classification non supervisée, mélanges gaussiens, données transcriptomes.

TITLE : Variable selection for model-based clustering. Application for transcriptome data analysis.

Abstract : We are interested in variable selection for clustering with Gaussian mixture models. This research is motivated by the clustering of genes described by transcriptome datasets in particular. In the two parts, this problem is regarded as a model selection problem in a model-based cluster analysis framework.
In the first part, the proposed model, generalizing the one of Raftery and Dean (2006b), specifies the variable role for the clustering process. The irrelevant clustering variables can be dependent to a relevant variable subset. Models are compared with a BIC-like criterion. The model identifiability is established and the consistency of the criterion is proved under regularity conditions. In practice, the variable role is obtained through an algorithm embedding two backward stepwise algorithms for variable selection for the clustering and the linear regression. The interest of this procedure is highlighted by a transcriptome dataset application especially. An improvement of the variable role modelling, consisting of partitioning the irrelevant variables according to their dependence or independence with some relevant clustering variables, is suggested to avoid an overpenalization of some models. Finally, the DNA microarray technology generating many missing values, an extension of our variable selection procedure taken into account the existence of missing entries is proposed. It avoids the missing entry imputation usually used in preprocessing.
In the second part, specific Gaussian mixtures are considered and a non asymptotic penalized criterion is proposed to select the number of mixture components and the relevant clustering variable subset. A general model selection theorem for maximum likelihood estimation, proposed by Massart (2007), is used to obtain the penalty function form. This theorem requires to control the bracketing entropy of studied Gaussian mixture families. This criterion depending on unknown constants, the "slope heuristics" method is carried out to allow the practical use of this criterion.
Keywords : Variable Selection, Model-based Clustering, Gaussian Mixtures, Transcriptome Data.