Modèles linéaires généralisés à effets aléatoires : contributions au choix de modèle et au modèle de mélange

Marie-José Martinez

UNIVERSITÉ MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

THÈSE

pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ MONTPELLIER II

Discipline	**:**	Mathématiques appliquées
Formation doctorale	**:**	Biostatistique
École doctorale	**:**	Information, Structures et Systèmes

présentée et soutenue publiquement
par

Marie-José MARTINEZ

le 29 septembre 2006

MODÈLES LINÉAIRES GÉNÉRALISÉS À EFFETS ALÉATOIRES : CONTRIBUTIONS AU CHOIX DE MODÈLE ET AU MODÈLE DE MÉLANGE

JURY

Gilles DUCHARME	Président
Avner BAR-HEN	Rapporteur
Denys POMMERET	Rapporteur
Christèle ROBERT-GRANIÉ	Examinateur
Christian LAVERGNE	Directeur de thèse
Catherine TROTTIER	Co-Directeur de thèse

Remerciements

Cette thèse résulte de plusieurs années de travail effectuées au sein de l'équipe de Probabilités et Statistique de l'Institut de Mathématiques et de Modélisation de Montpellier. Nombreuses sont les personnes qui ont contribué, à divers titres, à l'élaboration de ce travail. Je voudrais profiter de l'occasion qui m'est donnée ici pour les remercier.

En premier lieu, je tiens à exprimer toute ma reconnaissance à Christian et Catherine pour m'avoir encadrée durant cette thèse. Christian a suggéré puis dirigé ce travail tout en me laissant une certaine liberté d'action. Toujours disponible, je le remercie pour les discussions que nous avons eues : elles m'ont beaucoup appris et m'ont aidé à avancer dans ce milieu nouveau pour moi. Son dynamisme et son enthousiasme ont également été une source de motivation et de persévérance dans les moments de doute. Quant à Catherine, elle a constamment été présente et je lui suis extrêmement reconnaissante de ses conseils, son soutien et ses encouragements tout au long de ces années. Son sens de la rigueur et son souci de perfection ont contribué à l'accomplissement de ce travail. Je leur adresse, à tous les deux, un grand merci et j'espère que notre collaboration ne s'arrêtera pas avec cette thèse.

J'adresse mes sincères remerciements à Avner Bar-Hen et Denys Pommeret pour l'intérêt qu'ils ont manifesté pour mon travail en acceptant d'en être rapporteur malgré leurs nombreuses charges respectives. Je les remercie pour leurs remarques constructives et les échanges que nous avons eus.

Je tiens également à remercier Gilles Ducharme qui m'a fait l'honneur de présider ce jury mais aussi pour m'avoir transmis son goût pour les statistiques et la recherche durant ma formation.

Je tiens aussi à remercier Christèle Robert-Granié pour avoir examiné ce travail.

Pendant ces années de thèse, j'ai aussi eu le plaisir d'enseigner dans le cadre de mon monitorat. Je remercie donc en particulier Marc Joannides, Benoît Cadre et Ali Gannoun pour m'avoir fait confiance et m'avoir accompagnée dans mes premiers pas d'enseignante ainsi que Mohammed Mellouk pour avoir été mon tuteur.

Nombreux sont encore ceux qui m'ont aidé ou encouragé, chacun à leur manière, tout au long de ces années. Aussi, je souhaite encore adresser un merci :

aux membres de l'équipe de Probabilités et Statistique et plus particulièrement à Ludovic Menneteau pour sa gentillesse et son soutien lors de mes premières communications orales, à Jean-Noël Bacro pour son dynamisme et sa bonne humeur, à Gérard Biau et André Mas pour leurs encouragements et leur sympathie.

à Baptiste Chapuisat pour son aide en informatique qui m'a permis de venir à bout de nombreuses difficultés

au personnel administratif et plus particulièrement à Pierrette Arnaud pour son aide administrative et sa gentillesse. Elle est partie à la retraite le jour de ma soutenance et je lui souhaite donc une bonne retraite bien méritée ! Je n'oublie pas de remercier Sarah pour sa gentillesse et j'en profite pour lui souhaiter bonne chance dans sa nouvelle vie...

à toute l'équipe pédagogique de l'IUT STID de Carcassonne qui m'a chaleureusement accueillie en tant qu'ATER durant ma dernière année de thèse. Je remercie en particulier Sylvie Viguier-Pla pour son aide précieuse dans la préparation des enseignements. J'ai beaucoup apprécié ses conseils et nos discussions. Merci également à Denys Chaffardon, Thierry Spinosa, Mouna Kamel, Evelyne Amstrong et Gabriel Fraisse.

aux anciens et nouveaux doctorants du bureau 14 qui ont rendu le quotidien de cette thèse plus agréable. Je pense en particulier à Sandie pour son soutien dans les moments difficiles ainsi que pour la complicité et les bons moments que nous avons partagés. Merci également à Kevin pour avoir accompagné mes derniers pas ainsi que pour ses relectures attentives en anglais.

aux anciens du DEA de Biostatistique qui se sont lancés, eux aussi, dans cette aventure bien particulière qu'est une thèse. Merci en particulier à Laurent, Thomas, Edith

et Céline pour nos nombreuses conversations. Merci également à Olivier pour ses conseils, ses encouragements et pour m'avoir accueilli chez lui lors de mon séjour en Ecosse.

aux thésards des équipes ACSIOM et GTA qui ont égaillé les journées de travail.

Je tiens à remercier les membres de ma famille et mes amis qui sont venus me soutenir le jour J et qui ont su m'encourager tout au long de ces années. Une dédicace particulière à mon amie de toujours Audrey. Qu'il est long le chemin parcouru depuis la passerelle de la Digne...

Je remercie du fond du coeur Christophe pour sa patience et son soutien constant. Il a vécu cette thèse à mes côtés jour après jour et a su, tour à tour, être à l'écoute, me soutenir et même me secouer parfois quand il le fallait. Je tiens à remercier également ses parents, Bernadette et Daniel, son frère Anthony et Mamie de leur soutien et je profite de ces quelques lignes pour leur exprimer mon attachement.

Pour finir, c'est à mes parents que j'adresse un immense merci pour m'avoir toujours soutenue et encouragée. Je ne les remercierai jamais assez pour leur soutien si précieux durant toutes ces années d'étude.

Table des matières

Liste des tableaux

Introduction

Lorsqu'il s'agit d'analyser des données groupées, comme par exemple des données longitudinales ou des données de mesures répétées, émanant de divers domaines tels que l'agriculture, la biologie, l'économie ou encore la géophysique, les modèles à effets aléatoires constituent un outil puissant et flexible. Ce type de modélisation permet en effet d'introduire deux niveaux de lecture du comportement des individus : un niveau global traduit par les effets fixes et un niveau individuel traduit par les effets aléatoires. Les paramètres d'effets fixes sont communs à l'ensemble des individus alors que les effets aléatoires varient avec les individus.
D'autre part, en modélisation statistique, la loi normale s'impose dans de nombreuses situations. Malgré cela, un certain nombre de phénomènes observés sont difficilement modélisables par cette loi. Citons, par exemple, les cas de relevés de durées de vie de matériels, de l'observation du nombre d'individus dans une population ayant une certaine caractéristique, ou encore du décompte d'événements rares. Afin de permettre l'analyse de telles données, non gaussiennes, une extension en termes de loi des modèles linéaires classiques a conduit au développement de la classe plus large que sont les modèles linéaires généralisés.
La combinaison de l'introduction des effets aléatoires et de la généralisation de la loi dans les modèles linéaires a donné naissance aux modèles linéaires généralisés à effets aléatoires. Ces modèles sont l'objet central d'étude de notre travail.

Dans un premier chapitre, nous revenons sur la question de l'estimation des paramètres de tels modèles. Pour cela, la théorie statistique nous conduit généralement vers la maximisation de la fonction de vraisemblance. Cependant, les modèles linéaires généralisés à effets aléatoires contiennent deux sources de variation : celle due aux effets aléatoires et celle due aux erreurs. Les hypothèses sur la distribution de la variable à expliquer

ne peuvent être formulées correctement que conditionnellement aux effets aléatoires. Par conséquent, la distribution marginale de la variable aléatoire modélisant le phénomène observé est difficilement descriptible excepté pour certains choix de lois au sein de la famille exponentielle ou pour une modélisation particulière des effets aléatoires. Cela constitue l'obstacle principal au développement de procédures d'estimation dans la mesure où les effets aléatoires se réalisent au cours de l'expérience et ne sont pas directement observés. Dans ce chapitre, nous avons voulu aborder le problème d'un point de vue global en considérant, de plus, l'hypothèse classique de distribution normale des effets aléatoires.
Pour faire face à ce problème, différentes approches ont été développées menant, en tout état de cause, à des méthodes non exactes par le biais d'approximations réalisées à différents niveaux selon les raisonnements. La première partie de notre travail s'inscrit donc dans un objectif d'étude de méthodes d'estimation. Après une description précise des modèles linéaires généralisés à effets aléatoires, nous nous intéressons tout particulièrement à une méthode proposée par Schall [51]. Cette méthode est basée sur une linéarisation du modèle, conditionnellement aux effets aléatoires. Cette linéarisation se réalise par l'introduction d'une variable dépendante, technique propre à l'estimation des paramètres par maximum de vraisemblance dans les modèles linéaires généralisés classiques. Le modèle linéarisé ainsi obtenu est alors traité comme un modèle linéaire à effets aléatoires avec la particularité qu'une partie de la structure de variance est connue. Nous revenons également sur d'autres démarches dites "de linéarisation" proposées par divers auteurs qui vont s'avérer être différentes façons de justifier la méthode de Schall.

À la question de l'estimation des paramètres d'un modèle succède celle sur la possibilité d'établir des méthodes de choix de modèles. La deuxième partie de notre travail s'inscrit ainsi dans un objectif de mise en place de critères de sélection de modèles au sein des modèles linéaires généralisés à effets aléatoires. Ce sujet est d'autant plus délicat que les méthodes d'estimation sont, elles mêmes, approchées.
De nombreux critères permettant de choisir parmi plusieurs modèles statistiques en compétition ont été établis. Dans ce travail, nous nous intéressons aux critères de sélection de modèles usuels basés sur la fonction de vraisemblance. Comme pour l'estimation des paramètres, l'utilisation de ces critères est confrontée, dans le cadre des modèles linéaires généralisés à effets aléatoires, au problème du calcul de la fonction de vraisemblance.
Dans le chapitre 2, nous proposons de traiter la question du choix de modèles conjointe-

ment à la procédure d'estimation des paramètres. Nous revenons sur la méthode de Schall qui mène à la construction d'un modèle linéarisé et nous proposons un critère basé sur la vraisemblance marginale calculée dans le modèle linéarisé obtenu à la convergence de la procédure d'estimation. Cette approche est envisageable quelque soit le modèle linéaire généralisé à effets aléatoires considéré.
Dans ce même chapitre, guidé par le problème de la modélisation de données de défaillance en fiabilité, nous nous intéressons plus particulièrement au cas du modèle exponentiel mixte lien logarithme. Nous proposons un autre critère de choix de modèles construit de la même façon conjointement à une méthode d'estimation spécifique à la loi exponentielle. Pour finir, nous comparons sur simulations ces deux critères basés sur la linéarisation du modèle à un troisième critère construit sur une approximation directe de la vraisemblance.

Le troisième et dernier chapitre s'inscrit dans un cadre légèrement différent. En effet, nous enrichissons les hypothèses de modélisation pour introduire la notion d'hétérogénéité dans les modèles linéaires généralisés à effets aléatoires.
Lorsqu'une population se divise en sous-groupes, il est important de prendre en compte dans la modélisation les différences de comportement d'un sous-groupe à l'autre. Dans certaines situations, le découpage en sous-groupes n'est pas connu précisément, autrement dit on ne sait pas si telle donnée appartient ou non à tel sous-groupe. Les modèles de mélange sont alors un outil naturel permettant de modéliser cette hétérogénéité dans les données. Des lois sont supposées pour chaque classe, et en affectant à chaque donnée une certaine probabilité d'appartenir aux différentes classes, ces modèles permettent de prendre en compte la non connaissance exacte du découpage.
Dans le chapitre 3, nous nous plaçons dans ce contexte et nous nous intéressons à une nouvelle classe de modèles que sont les mélanges finis de modèles linéaires généralisés à effets aléatoires. Dans ces modèles, chaque classe est défini par un modèle linéaire généralisé à effets aléatoires. Nous sommes donc ici en présence de deux sources d'information cachées : les appartenances des données à l'un des composants du mélange et les effets aléatoires.
Pour aborder l'estimation des paramètres au sein de ces modèles, nous commençons par nous intéresser au cas des mélanges finis de modèles gaussiens à effets aléatoires. L'introduction du modèle de mélange pour modéliser l'hétérogénéité implique que certaines démarches usuelles ne sont plus directement envisageable. L'algorithme EM constitue alors un outil essentiel. Nous étudions sa mise en place dans le cas gaussien. Puis nous revenons à

l'objet principal de ce chapitre en nous plaçant dans le cadre des mélanges finis de modèles linéaires généralisés à effets aléatoires. Plus particulièrement, nous nous intéressons au cas d'un mélange de modèles exponentiels à effets aléatoires et nous voyons pourquoi l'application directe de l'algorithme EM n'est pas envisageable dans ce cas. Nous proposons alors une méthode d'estimation, alliant la technique de linéarisation spécifique à la loi exponentielle du chapitre 2 et l'utilisation de l'algorithme EM dans le cas d'un mélange de modèles linéaires à effets aléatoires. Nous proposons également une seconde méthode plus générale puisque s'appliquant à un mélange de modèles à effets aléatoires quelconques. Cette méthode s'appuie sur une étape de Metropolis-Hastings pour construire un algorithme de type MCEM.

Chapitre 1

Méthodes d'estimation dans les GL2M

Les modèles linéaires généralisés à effets aléatoires (ou modèles linéaires généralisés mixtes) notés GL2M (Generalized Linear Mixed Models) constituent un ensemble assez vaste de modèles. Ces modèles ont vu le jour dans les années 80. Ils sont à la croisée de deux types d'extension des modèles linéaires classiques notés LM (Linear Models). La première est une extension en termes de loi et donne naissance à la classe des modèles linéaires généralisés, désignés par GLM (Generalized Linear Models). La deuxième est une extension en termes d'introduction d'effets aléatoires qui aboutit, quant à elle, à la classe des modèles linéaires mixtes, notés L2M (Linear Mixed Models).

Ce chapitre est donc consacré, dans un premier temps, à une description succincte de ces deux classes de modèles : GLM et L2M. Une fois l'aspect modélisation ayant été décrit, nous revenons sur l'estimation des paramètres de ces modèles. Tout cela nous conduira naturellement aux GL2M, objet central de cette thèse. Après avoir défini la classe des GL2M, la question naturelle qui en découle est celle de l'estimation des paramètres. De nombreux travaux ont été réalisés concernant l'estimation dans les GL2M. En abordant le problème sous des angles différents, des auteurs ont proposé diverses méthodes d'estimation. Nous reviendrons sur certaines d'entre elles et présenterons tout particulièrement différentes méthodes de linéarisation qui, en définitive, aboutissent aux mêmes équations.

1.1 Retour sur l'estimation dans les GLM

Cette section est consacrée, dans un premier temps, à une description des modèles linéaires généralisés. Nous abordons ensuite de façon succincte la question de l'estimation des paramètres de ces modèles par maximum de vraisemblance.

1.1.1 Définition des GLM

La classe des GLM est une extension des modèles linéaires classiques en termes de loi. Elle permet l'analyse de données discrètes mais aussi de données continues pour lesquelles la loi normale n'est pas des plus adaptées. Elle a une place importante dans la modélisation statistique, trouvant son intérêt dans différents domaines d'application. Dans leur ouvrage, McCullagh et Nelder [38] en font une présentation complète.

Notons y le vecteur des observations de taille n, réalisation du vecteur aléatoire Y, variable à expliquer. Un GLM se caractérise par les trois hypothèses suivantes :

• On suppose que les composantes Y_i $(i = 1, \ldots, n)$ de Y sont indépendantes et distribuées selon une loi appartenant à la famille exponentielle au sens de Nelder et Wedderburn [42], c'est-à-dire que la fonction de densité de la variable aléatoire Y_i s'écrit :

$$f_{Y_i}(y_i, \theta_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}$$

où θ_i est un paramètre canonique et ϕ un paramètre de dispersion. Les fonctions b et c sont spécifiques à chaque distribution et la fonction a_i s'écrit : $a_i(\phi) = \dfrac{\phi}{w_i}$ où w_i est un poids connu associé à la réalisation y_i.

L'espérance et la variance de la variable associée s'exprime à l'aide des fonctions a_i et b et leurs dérivées :

$$\begin{aligned} E(Y_i) &= b'(\theta_i) \\ \mathrm{Var}(Y_i) &= a_i(\phi)b''(\theta_i) \end{aligned}$$

Il existe donc une relation directe entre l'espérance (notée μ_i) et la variance de Y_i :

$$\begin{aligned} \mathrm{Var}(Y_i) &= a_i(\phi)b''(b'^{-1}(\mu_i)) \\ &= a_i(\phi)v(\mu_i) \end{aligned}$$

Cette fonction $v = b'' \circ b'^{-1}$ est appelée *fonction de variance*.

• On définit le prédicteur linéaire :

$$\eta = X\beta$$

où β est un vecteur de paramètres inconnus de taille p et X la matrice des variables explicatives fixée par l'expérience de dimension $n \times p$.

• Le lien entre l'espérance de Y_i et la $i^{ème}$ composante du prédicteur linéaire est réalisé par la fonction g (monotone et différentiable) appelée *fonction de lien* :

$$\eta_i = g(\mu_i)$$

Une fonction de lien pour laquelle $\eta_i = \theta_i$ est appelée *fonction de lien canonique*.

Pour résumer, les modèles linéaires généralisés sont caractérisés par deux fonctions :
- la fonction de lien, spécifiant l'introduction de la linéarité,
- la fonction de variance, spécifiant la relation entre l'espérance et la variance.

1.1.2 Estimation dans les GLM

Nous nous intéressons ici à l'estimation du vecteur des paramètres β, de dimension p, coefficients de la combinaison linéaire des covariables permettant d'expliquer le vecteur Y. Nous décrivons, pour cela, brièvement la procédure classique ML (Maximum Likelihood) permettant d'atteindre l'estimateur du maximum de vraisemblance.

Avec l'hypothèse d'indépendance des composantes de Y, la log-vraisemblance du vecteur des paramètres canoniques θ s'écrit :

$$L(\theta; y) = \sum_{i=1}^{n} \Big[\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i, \phi) \Big] = \sum_{i=1}^{n} L_i(\theta_i; y_i).$$

Pour obtenir les équations du maximum de vraisemblance pour l'estimation de β, nous dérivons la log-vraisemblance L du vecteur de paramètres β par rapport à ses différentes composantes. On obtient alors : $\forall i \in \{1, \ldots, n\}, \forall j \in \{1, \ldots, p\}$,

$$\begin{aligned} \frac{\partial L_i}{\partial \beta_j} &= \frac{\partial \eta_i}{\partial \beta_j} \frac{d\mu_i}{d\eta_i} \frac{d\theta_i}{d\mu_i} \frac{\partial L_i}{\partial \theta_i} \\ &= X_{ij} \frac{1}{g'(\mu_i)} \frac{1}{b''(\theta_i)} \frac{y_i - \mu_i}{\phi/w_i}, \end{aligned}$$

D'où

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} X_{ij} \frac{1}{g'(\mu_i)^2 \mathrm{var}(Y_i)} g'(\mu_i)(y_i - \mu_i).$$

Ainsi, en considérant les matrices diagonales définies par

$$W_\beta = \mathrm{diag}\{\mathrm{var}(Y_i) g'(\mu_i)^2\}_{i=1,\ldots,n} = \mathrm{diag}\{\frac{\phi}{w_i} v(\mu_i) g'(\mu_i)^2\}_{i=1,\ldots,n}$$

et

$$\frac{d\eta}{d\mu} = \mathrm{diag}\{\frac{d\eta_i}{d\mu_i}\}_{i=1,\ldots,n} = \mathrm{diag}\{g'(\mu_i)\}_{i=1,\ldots,n},$$

les équations du maximum de vraisemblance pour β s'écrivent :

$$X' W_\beta^{-1} \frac{d\eta}{d\mu}(y - \mu) = 0 \tag{1.1}$$

Ce système d'équations n'étant pas linéaire en β, une résolution itérative est mise en place. L'algorithme itératif usuel est l'algorithme des scores de Fisher dont les équations itérées sont données par :

$$\begin{aligned} \beta^{[t+1]} &= \beta^{[t]} - \left(E\Big[\frac{\partial^2 L}{\partial \beta \partial \beta'}\Big]^{[t]} \right)^{-1} \frac{\partial L^{[t]}}{\partial \beta} \\ &= \beta^{[t]} + (X' W_{\beta^{[t]}}^{-1} X)^{-1} X' W_{\beta^{[t]}}^{-1} \frac{d\eta^{[t]}}{d\mu}(y - \mu^{[t]}) \\ &= (X' W_{\beta^{[t]}}^{-1} X)^{-1} X' W_{\beta^{[t]}}^{-1} z^{[t]} \end{aligned}$$

où $z^{[t]} = X\beta^{[t]} + \dfrac{d\eta^{[t]}}{d\mu}(y - \mu^{[t]})$.

En introduisant le vecteur dépendant défini par :

$$z = \eta + \frac{d\eta}{d\mu}(y - \mu) = X\beta + \frac{d\eta}{d\mu}(y - \mu),$$

les équations (1.1) s'écrivent :

$$X' W_\beta^{-1}(z - X\beta) = 0. \tag{1.2}$$

Ainsi, le même algorithme est décrit en résolvant itérativement les équations (1.2) comme des équations normales. À chaque itération, la valeur courante de β est utilisée pour le calcul de la matrice des poids W_β et du vecteur dépendant z. Cela permet ensuite par résolution de ce système linéarisé d'obtenir une nouvelle valeur de β.

Cette réécriture (1.2) permet une interprétation de type linéaire qui sera exploitée plus particulièrement dans le cadre des modèles linéaires généralisés à effets aléatoires. À β fixé, en considérant z comme un nouveau vecteur de données et W_β comme une matrice de poids fixés, on reconnait dans le système (1.2) les équations classiques des moindres carrés généralisés associées au modèle :

$$Z = X\beta + e$$

où $E(e) = 0$ et $\mathrm{Var}(e) = W_\beta$. La $i^{ème}$ composante du vecteur aléatoire $Z = X\beta + g'(\mu)(Y - \mu)$ a pour variance : $\mathrm{var}(Z_i) = g'(\mu_i)^2 \mathrm{var}(Y_i)$. Ainsi W_β correspond bien à la matrice de variance de Z.

Pour conclure, l'estimation du maximum de vraisemblance de β dans le GLM écrit sous la forme :

$$Y = g^{-1}(\eta) + \varepsilon \quad \text{où} \quad \begin{array}{l} E(\varepsilon) = 0 \\ \mathrm{Var}(\varepsilon) = \mathrm{diag}\{a(\phi)v(\mu_i)\}_{i=1,\dots,n} \end{array}$$

est équivalent à l'estimation successive du maximum de vraisemblance dans le LM $\mathcal{M}^{[t]}$ défini à l'étape $[t]$ par :

$$Z = \eta + e^{[t]} \qquad E(e^{[t]}) = 0 \quad \text{et} \quad \mathrm{Var}(e^{[t]}) = W_{\beta^{[t]}}$$

pour le vecteur de données $z^{[t]}$.

1.1.3 Simplification dans le cas d'un lien canonique

Le lien canonique se définit par :

$$\eta_i = \theta_i = g(\mu_i) = X_i\beta.$$

Dans le cas d'un lien canonique, nous avons :

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i) = \frac{\mathrm{Var}(Y_i)}{a_i(\phi)}.$$

Nous obtenons ainsi :

$$\frac{\partial L_i}{\partial \beta_j} = X_{ij} \frac{1}{a_i(\phi)} (y_i - \mu_i).$$

Par suite, les termes de la matrice hessienne s'écrivent :

$$\begin{aligned} -\frac{\partial^2 L_i}{\partial\beta_j\partial\beta_k} &= -\frac{\partial}{\partial\beta_k}\big[X_{ij}\frac{1}{a_i(\phi)}(y_i-\mu_i)\big] \\ &= X_{ij}X_{ik}\frac{\mathrm{Var}(Y_i)}{a_i(\phi)^2} \end{aligned}$$

Et les termes de la matrice d'information de Fisher s'écrivent :

$$\begin{aligned} E\Big[-\frac{\partial^2 L_i}{\partial\beta_j\partial\beta_k}\Big] &= E\Big[\big(\frac{\partial L_i}{\partial\beta_j}\big)\big(\frac{\partial L_i}{\partial\beta_k}\big)\Big] \\ &= E\Big[X_{ij}X_{ik}\frac{1}{a_i(\phi)^2}(y_i-\mu_i)^2\Big] \\ &= X_{ij}X_{ik}\frac{\mathrm{Var}(Y_i)}{a_i(\phi)^2} \end{aligned}$$

Ainsi, dans le cas d'un lien canonique, l'algorithme des scores de Fisher est identique à l'algorithme de Newton-Raphson.

1.1.4 Propriétés asymptotiques

Concernant les propriétés asymptotiques de l'estimateur du maximum de vraisemblance, dans le cadre général des GLM, Fahrmeir et Kaufmann [19] démontrent différents résultats dont, en particulier, un résultat sur la normalité asymptotique de $\hat{\beta}_n$, solution des équations du maximum de vraisemblance pour un jeu de données de taille n. Nous ne reprenons pas ici l'énoncé de ce théorème. Notons uniquement qu'il repose sur des hypothèses concernant les matrices hessienne et d'information de Fisher.

1.2 Retour sur l'estimation dans les L2M

Dans cette section, nous donnons une description des modèles linéaires mixtes. Nous abordons ensuite la question de l'estimation des paramètres au sein de ces modèles et présentons deux méthodes ou algorithmes d'estimation : l'algorithme EM (Expectation Maximisation) et la méthode dite "de Henderson".

1.2.1 Définition des L2M

Dans toute expérience statistique, les données présentent une certaine variabilité. L'intérêt d'une étude statistique consiste notamment à en déterminer la nature, l'importance,

les facteurs... L'histoire du modèle mixte remonte aux travaux de Fisher sur l'analyse de la variance qui tentent de cloisonner les différentes sources de variation et de répondre notamment à des questions sur la significativité de différences observées entre moyennes de sous-groupes de données. Mais l'introduction d'effets aléatoires dans la modélisation constitue un moyen plus élaboré d'étudier cette variabilité en précisant les diverses sources de variation. Cela permet de séparer la variance totale en deux parties : la variation due aux effets aléatoires et celle que l'on affecte à l'erreur. On est donc plus précis quant à son origine.
Ainsi, les modèles linéaires mixtes sont en fait une extension des modèles linéaires classiques : aux effets fixes de ces derniers, viennent s'ajouter des effets aléatoires. Les L2M se présentent alors de la manière suivante :

$$Y = \underbrace{X\beta}_{\text{partie effets fixes}} + \underbrace{U\xi}_{\text{partie effets aléatoires}} + \varepsilon \tag{1.3}$$

où

- Y : vecteur aléatoire à expliquer de taille n,

- β : vecteur de paramètres inconnus des effets fixes, de taille p, et X, de dimension $n \times p$, sa matrice d'incidence supposée fixe et connue,

- ξ : vecteur d'effets aléatoires de taille q. En toute généralité, ce vecteur se décompose en K parties $\xi = (\xi_1', \ldots, \xi_K')'$ où K est le nombre d'effets aléatoires considérés dans le modèle. Chaque composante ξ_j est un vecteur aléatoire modélisant un effet aléatoire de dimension q_j ($\sum_{j=1}^{K} q_j = q$).
  On suppose en général une distribution normale centrée réduite des effets aléatoires, c'est-à-dire : $\forall j \in \{1, \ldots, K\}$, $\xi_j \sim \mathcal{N}_{q_j}(0, \sigma_j^2 A_j)$ avec A_j matrice de dimension $q_j \times q_j$ supposée connue. D'autre part, $\forall i, j \in \{1, \ldots, K\}^2$ $i \neq j$, ξ_i et ξ_j sont indépendants. Donc $\xi \sim \mathcal{N}_q(0, D)$ où D est une matrice diagonale par blocs : $D = \text{diag}\{\, \sigma_j^2 A_j \,\}_{j=1,\ldots,K}$.
  La matrice d'incidence U de dimension $n \times q$ est connue et formée des matrices d'incidence U_j de dimension $n \times q_j$ de chaque effet aléatoire : $U = [U_1 \vdots \ldots \vdots U_K]$.

- ε : vecteur aléatoire d'erreurs de taille n. On suppose : $\varepsilon \sim \mathcal{N}_n(0, \sigma_0^2 V_0)$. On notera aussi $R = \sigma_0^2 V_0$. On suppose que $\forall j \in \{1, \ldots, K\}$, ε et ξ_j sont indépendants.

Sous ces différentes hypothèses, on a :

- $\begin{aligned} E(Y|\xi) &= X\beta + U\xi \\ \mathrm{Var}(Y|\xi) &= R = \sigma_0^2 V_0 \end{aligned}$

- $\begin{aligned} E(Y) &= X\beta \\ \mathrm{Var}(Y) &= R + UDU' \\ &= \sigma_0^2 V_0 + \sum_{j=1}^{K} \sigma_j^2 U_j A_j U_j' \\ &= \Gamma \end{aligned}$

ce qui, avec $V_j = U_j A_j U_j'$ pour tout $j \in \{1, \ldots, K\}$, peut se mettre sous la forme :

$$\Gamma = \sum_{j=0}^{K} \sigma_j^2 V_j.$$

La variance totale se retrouve ainsi scindée en plusieurs composantes σ_j^2 que l'on appelle *composantes de la variance*. Le vecteur des effets fixes β ainsi que le vecteur des paramètres de variance $\sigma^2 = (\sigma_0^2, \ldots, \sigma_K^2)'$ sont inconnus et il s'agit de les estimer.
Il est important de noter ici, qu'en pratique, nous n'observons pas directement les effets aléatoires ξ. Ils sont indirectement observés dans les données.

Dans ces modèles, nous nous intéressons à la fois à l'estimation de l'effet fixe ainsi qu'à celle des composantes de la variance. De nombreux travaux sur les L2M ont été réalisés sous des formes et selon des approches différentes. Citons par exemple Pinheiro et Bates [45] ou encore Searle, Casella et McCulloch [53] qui y consacrent leur ouvrage. Mais ce sont Hartley et Rao [25] en 1967 qui, les premiers, ont donné un formalisme à l'estimation des composantes de la variance. Par la suite, différentes méthodes d'estimation ont été proposées.
L'approche ML (Maximum Likelihood) utilise le concept classique de la fonction de vraisemblance. Une autre approche est celle du maximum de vraisemblance restreint qui, comme son nom l'indique, reste apparentée à celle du maximum de vraisemblance. On se focalise davantage, dans ce cas, sur l'estimation des composantes de la variance : on fait disparaître momentanément les effets fixes pour ne maximiser que la partie de la vraisemblance concernant les composantes.
Cette méthode REML (Restricted Maximum Likelihood) a été proposée par Anderson et

Bancroft [4] et Thompson [60] pour l'analyse de dispositifs équilibrés, puis généralisée à un modèle mixte gaussien quelconque par Patterson et Thompson [43] dans les années 70. Foulley, Delmas et Robert-Granié [20] présentent une synthèse sur l'utilisation de ces méthodes du maximum de vraisemblance.

L'estimation des composantes de la variance par ces approches ML et REML conduit à des systèmes non linéaires avec contraintes. Outre le fait que rien ne nous assure la positivité des estimations pas à pas, il n'est pas certain non plus que ces systèmes mènent à un maximum global de la fonction de vraisemblance. D'autres alternatives à la résolution itérative de ces systèmes ont été proposées : l'algorithme EM mis en place par Dempster, Laird et Rubin [17] et l'algorithme de Henderson [29]. Ce sont ces deux méthodes d'estimation des composantes de la variance que nous présentons maintenant.

1.2.2 L'algorithme EM

L'algorithme EM constitue un outil permettant d'atteindre les estimations ML ou REML. Il permet de contourner la difficulté de l'obtention de la vraisemblance des observations lorsque la distribution marginale de ces observations est difficile à spécifier. Il réalise cela par l'introduction de données manquantes qui ne sont pas directement observées au cours de l'expérience, mais dont on connait la vraisemblance jointe aux données observées.

Dans le cadre des L2M, les données observées (ou incomplètes suivant la terminologie EM) sont constituées du vecteur des observations y. Le vecteur des effets aléatoires $\xi = (\xi_1', \ldots, \xi_K')'$ n'étant pas directement observé, il constitue les données manquantes. Le vecteur x des données complètes est alors défini par $x = (y', \xi')'$. Ainsi, l'algorithme EM va tirer partie de l'information manquante sur les effets aléatoires ξ non directement observés en considérant l'espérance conditionnelle de la vraisemblance de l'échantillon complet x sachant les données observées y.

D'après les propriétés usuelles de conditionnement de la loi normale, la distribution du vecteur des données complètes $x = [y', \xi_1', \ldots, \xi_K']'$ est une loi normale d'espérance

$$\mu = \begin{bmatrix} X\beta \\ 0 \\ \vdots \\ 0 \end{bmatrix} \text{ et de matrice de variance-covariance}$$

$$\Sigma = \begin{bmatrix} \Gamma & \{\sigma_j^2 U_j A_j\}_{j=1,\dots,K} \\ \{\sigma_j^2 A_j U_j'\}'_{j=1,\dots,K} & \text{diag}\{\ \sigma_j^2 A_j\ \}_{j=1,\dots,K} \end{bmatrix}.$$

La fonction de vraisemblance associée aux données complètes s'écrit alors :

$$l(\beta, \sigma^2; x) = -\frac{1}{2}\,(\sum_{j=0}^{K} q_j)\ln(2\pi) - \frac{1}{2}\sum_{j=0}^{K}(q_j \ln(\sigma_j^2) + \ln(|A_j|)) - \frac{1}{2}\sum_{j=0}^{K}\frac{\xi_j' A_j^{-1}\xi_j}{\sigma_j^2}$$

avec $\xi_0 = \varepsilon = y - X\beta - \sum_{j=1}^{K} U_j\xi_j = y - X\beta - U\xi$, $q_0 = n$ et $A_0 = V_0$.

Si l'on observait une réalisation du vecteur aléatoire ξ, les estimateurs du maximum de vraisemblance associée aux données complètes s'obtiendraient naturellement par :

$$\begin{cases} \hat{\sigma}_j^2 &= \dfrac{\xi_j' A_j^{-1}\xi_j}{q_j} \qquad \forall j \in \{0,\dots,K\} \\ (X'V_0^{-1}X)\hat{\beta} &= X'V_0^{-1}(y - U\xi) \end{cases}$$

Or, comme on ne dispose que des observations y, l'algorithme EM consiste à remplacer les statistiques exhaustives $\xi_j' A_j^{-1}\xi_j$ et $(y - U\xi)$ par leurs espérances conditionnelles sachant y et les valeurs courantes des paramètres. On utilise pour cela les résultats sur le conditionnement des variables aléatoires normales. Pour $j \in \{0,\dots,K\}$, on obtient :

$$\begin{aligned} E(\xi_j|y) &= \sigma_j^2 A_j U_j' \Gamma^{-1}(y - X\beta) \\ E(\xi_j A_j^{-1}\xi_j|y) &= \sigma_j^4 (y - X\beta)'\Gamma^{-1}V_j\Gamma^{-1}(y - X\beta) + \text{tr}(\sigma_j^2 Id_{q_j} - \sigma_j^4 U_j'\Gamma^{-1}U_j A_j) \\ &= \sigma_j^4 (y - X\beta)'\Gamma^{-1}V_j\Gamma^{-1}(y - X\beta) + q_j\sigma_j^2 - \sigma_j^4 \text{tr}(\Gamma^{-1}V_j) \end{aligned}$$

$$\begin{aligned} E(Y - U\xi|y) &= y - \textstyle\sum_{j=1}^{K} \sigma_j^2 V_j \Gamma^{-1}(y - X\beta) \\ &= y - (\Gamma - \sigma_0^2 V_0)\Gamma^{-1}(y - X\beta) \\ &= X\beta + \sigma_0^2 V_0 \Gamma^{-1}(y - X\beta). \end{aligned}$$

Ainsi, l'algorithme EM pour l'estimation ML mène au schéma itératif suivant. À partir des estimations obtenues à l'itération $[t]$, les nouvelles valeurs des paramètres à l'étape

$[t+1]$ sont données par :

$$\left\{ \begin{array}{rcl} q_j\sigma_j^{2[t+1]} & = & \sigma_j^{4[t]}(y - X\beta^{[t]})'\Gamma^{-1[t]}V_j\Gamma^{-1[t]}(y - X\beta^{[t]}) + q_j\sigma_j^{2[t]} - \sigma_j^{4[t]}\mathrm{tr}(\Gamma^{-1[t]}V_j) \\ X\beta^{[t+1]} & = & X(X'V_0^{-1}X)^{-1}X'V_0^{-1}(X\beta^{[t]} + \sigma_0^{2[t]}V_0\Gamma^{-1[t]}(y - X\beta^{[t]})). \end{array} \right.$$

Nous avons présenté ici l'algorithme EM dans le cadre particulier des L2M en suivant une démarche basée sur la notion de statistique exhaustive. La méthodologie EM peut également être adapté au maximum de vraisemblance restreint. Cette approche est notamment présentée en détails dans l'ouvrage de Searle et al [53].

1.2.3 La méthode de Henderson

Cette méthode d'estimation des composantes de la variance se présente comme un sous-produit de la résolution des équations de Henderson. Au cours de ses travaux, Henderson a été amené à prédire des réalisations non observées d'un effet aléatoire à l'intérieur d'un modèle mixte. Ainsi la prédiction de ξ devient un élément important et indispensable. Cette prédiction de ξ est ensuite utilisée pour l'estimation des composantes de la variance.

1.2.3.1 Les équations de Henderson

Il existe plusieurs manières de prédire ξ . Celle considérée ici est nommée BLUP (Best Linear Unbiaised Predictor). Cette prédiction $\tilde{\xi}$ est une fonction linéaire des données, non biaisée ($E(\tilde{\xi}) = E(\xi)$) et la meilleure au sens des carrés moyens (pour toute matrice symétrique A définie positive, $E((\tilde{\xi} - \xi)'A(\tilde{\xi} - \xi))$ est minimum).

La méthode de Henderson propose des équations permettant d'obtenir simultanément l'estimation BLUE (Best Linear Unbiaised Estimator) de β (notée $\hat{\beta}$, équivalente au maximum de vraisemblance sous des hypothèses de normalité adéquates) et la prédiction BLUP de ξ. Pour former ce système d'équations, la distribution jointe de Y et ξ est maximisée en β et ξ. Ainsi, après avoir utilisé sa distribution pour construire la fonction de vraisemblance, ξ joue alors le rôle de paramètre.

Compte tenu des résultats suivants :

- $Y \sim \mathcal{N}_n(X\beta, \Gamma)$,
- $\xi \sim \mathcal{N}_q(0, D)$,
- $Y|\xi \sim \mathcal{N}_n(X\beta + U\xi, R)$,

la distribution jointe s'écrit :

$$f(y,\xi) = \frac{1}{(2\pi)^{\frac{n+q}{2}}|R|^{\frac{1}{2}}|D|^{\frac{1}{2}}} \exp\{-\frac{1}{2}[(y - X\beta - U\xi)^{'}R^{-1}(y - X\beta - U\xi) + \xi^{'}D^{-1}\xi]\},$$

on en déduit alors le système d'équations :

$$\begin{pmatrix} X^{'}R^{-1}X & X^{'}R^{-1}U \\ U^{'}R^{-1}X & U^{'}R^{-1}U + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \xi \end{pmatrix} = \begin{pmatrix} X^{'}R^{-1}y \\ U^{'}R^{-1}y \end{pmatrix}. \tag{1.4}$$

Ces équations sont souvent appelées *équations du modèle mixte* ou MME (*Mixed Model Equations*) ou encore *équations de Henderson*. Remarquons que sans la présence de D^{-1} dans la partie inférieure droite de ce système, il correspondrait aux équations du maximum de vraisemblance lorsqu'on traite ξ comme un effet fixe. Donc par l'introduction de D^{-1}, on prend en compte en partie la nature aléatoire de ξ.

Ce système est équivalent à :

$$\begin{cases} X^{'}\Gamma^{-1}X\beta &= X^{'}\Gamma^{-1}y \\ \xi &= DU^{'}\Gamma^{-1}(y - X\beta) = E(\xi|y) \end{cases} \tag{1.5}$$

Le système (1.5) permet d'obtenir l'estimation BLUE de β et la prédiction BLUP de ξ. Cependant, il nécessite l'inversion de Γ non diagonale et d'ordre n. Ainsi, les équations de Henderson qui ne nécessitent que l'inversion des matrices R et D (souvent diagonales) et celle du système (d'ordre $p+q$ souvent plus petit que n) représentent une alternative intéressante à la résolution directe de ce système.

Ayant obtenu $\hat{\beta}$ et $\tilde{\xi}$, il reste à estimer les composantes de la variance.

1.2.3.2 Estimation ML et REML par Henderson

Dans le système (1.4), les matrices R et D dépendent respectivement des valeurs σ_0^2 et σ_1^2,..., σ_K^2 toutes inconnues. L'estimation de ces composantes est donc nécessaire. Les valeurs de ξ et β, obtenues par résolution du système de Henderson, vont alors permettre de calculer les estimations ML et REML dans un schéma itératif (Harville [26]). À partir des équations ML et REML (Searle, Casella et McCulloch [53]), on construit les procédures itératives suivantes :

ML

$$\left[\begin{array}{lcl} \sigma_j^{2[t+1]} & = & \dfrac{\xi_j^{'[t]} A_j^{-1} \xi_j^{[t]}}{q_j - \mathrm{tr}(P_{jj}^{[t]})} \\ \\ \sigma_0^{2[t+1]} & = & \dfrac{y' V_0^{-1}(y - X\beta^{[t]} - U\xi^{[t]})}{n} \end{array}\right. \tag{1.6}$$

où $P = (I_q + U'R^{-1}UD)^{-1}$ avec I_q la matrice identité d'ordre q.

P_{jj} est la $j^{ème}$ sous-matrice de P.

REML

$$\left[\begin{array}{lcl} \sigma_j^{2[t+1]} & = & \dfrac{\xi_j^{'[t]} A_j^{-1} \xi_j^{[t]}}{q_j - \mathrm{tr}(Q_{jj}^{[t]})} \\ \sigma_0^{2[t+1} & = & \dfrac{y' V_0^{-1}(y - X\beta^{[t]} - U\xi^{[t]})}{n - \mathrm{rg}(X)} \end{array}\right. \tag{1.7}$$

où $Q = (I_q + U'SUD)^{-1}$

Q_{jj} est la $j^{ème}$ sous-matrice de Q.

$S = R^{-1}(I_n - X(X'R^{-1}X)^{-1}X'R^{-1})$

De façon équivalente, on peut aussi utiliser les schémas itératifs suivants qui s'avèrent être plus utiles d'un point de vue pratique :

ML

$$\left[\begin{array}{lcl} \sigma_j^{2[t+1]} & = & \dfrac{\xi_j^{'[t]} A_j^{-1} \xi_j^{[t]}}{q_j - \dfrac{\mathrm{tr}(A_j^{-1} C_{jj}^{*[t]})}{\sigma_j^{2[t]}}} \\ \\ \sigma_0^{2[t+1]} & = & \dfrac{(y - X\beta^{[t]} - U\xi^{[t]})' V_0^{-1}(y - X\beta^{[t]} - U\xi^{[t]})}{n - \sum_{j=1}^{K}(q_j - \dfrac{\mathrm{tr}(A_j^{-1} C_{jj}^{*[t]})}{\sigma_j^{2[t]}})} \end{array}\right. \tag{1.8}$$

où C^* : est l'inverse de la matrice formée par les q dernières lignes et colonnes de la matrice des coefficients du système de Henderson (1.4) : $C^* = (U' R^{-1} U + D^{-1})^{-1}$

C_{jj}^* : est la $j^{ème}$ sous-matrice de C^*, correspondant au $j^{ème}$ effet aléatoire.

REML

$$\left[\begin{array}{rcl} \sigma_j^{2[t+1]} & = & \dfrac{\xi_j^{'[t]} A_j^{-1} \xi_j^{[t]}}{q_j - \dfrac{\mathrm{tr}(A_j^{-1} C_{jj}^{[t]})}{\sigma_j^{2[t]}}} \\ \\ \sigma_0^{2[t+1]} & = & \dfrac{(y - X\beta^{[t]} - U\xi^{[t]})' V_0^{-1} (y - X\beta^{[t]} - U\xi^{[t]})}{n - \mathrm{rg}(X) - \sum_{j=1}^{K} (q_j - \dfrac{\mathrm{tr}(A_j^{-1} C_{jj}^{[t]})}{\sigma_j^{2[t]}})} \end{array}\right. \tag{1.9}$$

où C : est la matrice formée des q dernières lignes et colonnes de l'inverse de la matrice des coefficients du système de Henderson (1.4),

C_{jj} : est la $j^{ème}$ sous matrice de C, correspondant au $j^{ème}$ effet aléatoire.

La procédure d'estimation alterne alors entre :

1. pour des valeurs de σ_j^2 connues (fixant les valeurs de R et D), la résolution de (1.4),
2. pour des valeurs de β et ξ, la résolution de (1.6) ou (1.7) (resp. (1.8) ou (1.9)).

1.2.4 Propriétés asymptotiques

Sous des conditions de régularité concernant la fonction de vraisemblance, les estimateurs du maximum de vraisemblance et du maximum de vraisemblance restreint possèdent des propriétés de convergence presque sûre et de normalité asymptotique. L'énoncé de ces résultats avec les différentes conditions nécessaires peuvent être trouvés dans l'ouvrage de Rao et Kleffe [47].

1.3 Définition des GL2M

De même que les effets aléatoires ont été introduits dans les modèles linéaires définissant ainsi les L2M, ils peuvent l'être au sein des modèles linéaires généralisés pour donner naissance aux modèles linéaires généralisés mixtes notés GL2M. C'est alors dans l'expression du prédicteur linéaire qu'une partie aléatoire vient s'ajouter à la partie fixe. En gardant les notations de la section précédente concernant le vecteur des paramètres d'effets fixes β et sa matrice associée X ainsi que le vecteur des effets aléatoires ξ et sa matrice associée U, le prédicteur linéaire s'exprime de la façon suivante :

$$\eta_\xi = X\beta + U\xi.$$

Pour bien insister sur l'introduction des effets aléatoires dans ce prédicteur, nous l'avons indicé par ξ. Nous maintenons, de plus, l'hypothèse de normalité sur ξ : $\xi \sim \mathcal{N}(0, D)$ où $D = \text{diag}\{\, \sigma_j^2 A_j \,\}_{j=1,\dots,K}$.

Comme dans les GLM, les fonctions de lien et de variance sont importantes dans les GL2M mais dans un raisonnement conditionnel à ξ. En effet, la fonction de lien g relie le prédicteur linéaire à l'espérance conditionnelle que nous indiçons également par ξ :

$$\eta_\xi = g(\mu_\xi) \qquad \text{où} \qquad \mu_\xi = E(Y|\xi).$$

La fonction de variance v intervient, quant à elle, dans l'expression de la variance conditionnelle :

$$\forall i \in \{1, \dots, n\}, \qquad \text{var}(Y_i|\xi) = a_i(\phi) v(\mu_{\xi,i}).$$

De plus, l'hypothèse de distribution est ici formulée sur la loi de Y conditionnelle à ξ. On suppose,

- d'une part que, conditionnellement à ξ, les composantes de Y sont indépendantes ; la matrice de variance conditionnelle est donc diagonale :

$$\text{var}(Y|\xi) = \text{diag}\{\, a_i(\phi) v(\mu_{\xi,i}) \,\}_{i=1,\dots,n}.$$

- d'autre part, $\forall i \in \{1, \dots, n\}$, $Y_i|\xi$ est supposée distribuée selon une loi issue de la famille exponentielle.

Ainsi, conditionnellement à ξ, le GL2M conserve toutes les propriétés du GLM. Par conséquent, le GL2M se trouve principalement défini dans un raisonnement conditionnel à ξ. Il peut se résumer de la façon suivante :

- les composantes de Y sont, conditionnellement à ξ, indépendantes et de loi appartenant à la famille exponentielle,

- le prédicteur linéaire s'écrit : $\eta_\xi = X\beta + U\xi$,

- l'espérance conditionnelle de Y est reliée au prédicteur linéaire par la fonction de lien : $\eta_\xi = g(\mu_\xi)$.

Rappelons que nous disposons, en pratique, d'observations y_i des Y_i mais nous n'observons pas directement les effets aléatoires ξ réalisés au cours de l'expérience.

Dans le cas de la loi normale, nous retrouvons bien la définition précédente du L2M avec un lien identité. Ainsi, comme le LM est un cas particulier des GLM, le L2M en est un des GL2M. Cependant, il est important de noter que pour les L2M, il y a conservation de loi lors du passage des lois de $Y|\xi$ et ξ à la loi marginale de Y. Cette propriété est spécifique à la loi normale. Elle ne se retrouve pas pour d'autres lois de façon générale avec une hypothèse gaussienne sur les effets aléatoires. Le L2M est donc un cas tout particulier de GL2M.

Les applications où les GL2M ont trouvé leur utilité sont multiples. Un exemple d'application est issu de la génétique quantitative. Des données, présentées par Shaeffer et Wilton [54] à l'origine, puis réutilisées par Gianola et Foulley [23], illustrent l'étude de l'évaluation génétique pour la facilité de vêlage. Dans 2 troupeaux, des génisses et des vaches sont croisées à 4 pères donnant naissance à 20 veaux mâles et femelles. Pour chacun, on enregistre la difficulté de vêlage selon 2 catégories : facile/difficile. La variable d'intérêt est donc l'information sur la difficulté de vêlage qui peut prendre 2 valeurs possibles. Il s'agit donc d'une variable discrète. On enregistre également l'âge de la mère, le numéro du troupeau et le sexe du veau. On considère alors une modélisation avec 3 effets fixes :

- l'âge de la mère à 2 niveaux : génisse ou vache
- le numéro du troupeau à 2 niveaux : troupeau 1 ou troupeau 2
- le sexe du veau à 2 niveaux : mâle ou femelle

Comme effet aléatoire, on considère l'effet père dont on observe 4 réalisations. On suppose qu'il est distribué selon une loi $\mathcal{N}(0, \sigma^2)$ et on ne cherche pas à connaître l'effet de chacun des 4 niveaux observés mais plutôt à estimer σ^2, la variabilité due à cet effet. Le modèle considéré est donc un modèle linéaire généralisé à un effet aléatoire.

1.4 Estimation dans les GL2M

1.4.1 Introduction

N'existant pas de méthode standard contrairement aux GLM ou aux L2M, la question de l'estimation des effets fixes et des composantes de la variance dans les GL2M a été

considérée dans de nombreux travaux par divers auteurs. Parmi ces travaux, un certain nombre ne concernent qu'un cas particulier de loi au sein de la famille exponentielle ou une modélisation particulière des effets aléatoires (surdispersion ou effets aléatoires emboîtés par exemple). Moins nombreux sont ceux qui s'intéressent à ces modèles de façon générale. Nous donnerons ici un aperçu de ces travaux dont un certain nombre sont présentés par Trottier [63].

Un GL2M est correctement défini conditionnellement aux effets aléatoires ξ. Ceci constitue l'obstacle principal à la mise en place de procédures d'estimation dans la mesure où ces effets aléatoires qui se réalisent au cours de l'expérience ne sont pas observés directement. Cet obstacle est d'autant plus important que l'on cherche à estimer les paramètres de leur distribution. Comme nous ne connaissons que la loi des observations conditionnellement aux effets aléatoires, la fonction de log-vraisemblance des paramètres β et σ^2 s'obtient par intégration :

$$L(\beta, \sigma^2; y_1, \ldots, y_n) = \int_{\mathbb{R}^q} \prod_{i=1}^{n} f(y_i|\xi)\ \varphi(\xi)\ \mathrm{d}\xi.$$

Mais ce calcul d'intégrale multiple n'est pas envisageable de façon explicite. Différentes approches ont été proposées menant, en tout état de cause, à des méthodes non exactes par le biais d'approximations réalisées à différents niveaux selon les raisonnements.

Une démarche classique consiste en l'obtention de la fonction de vraisemblance marginale et en sa maximisation moyennant des techniques d'intégration numériques. Les différentes intégrales sont ainsi approchées numériquement par quadrature gaussienne par exemple. Cette démarche a été notamment adoptée par Anderson et Aitkin [3] pour des données binaires. Mais ces méthodes d'intégration multiple sont numériquement exigeantes et sont difficilement pratiquables en toute généralité malgré le développement des capacités informatiques. En effet, elles donnent des résultats plutôt satisfaisants dans certains cas (dimension q faible, effets emboîtés) mais se heurtent à des problèmes de calcul dès que la dimension des effets aléatoires devient grande.
Les méthodes de Monte Carlo par chaînes de Markov sont également utilisées comme alternative. Zeger et Karim [69] ont, ainsi, proposé un algorithme de Gibbs sampling pour l'estimation des paramètres. McCulloch [39] propose également une méthode s'appuyant sur une étape de Metropolis-Hastings conduisant à la construction d'un algorithme de type Monte Carlo EM. En effet, du fait de la non accessibilité de la distribution conditionnelle

des effets aléatoires sachant les données observées, l'utilisation directe de l'algorithme EM se trouve confrontée au problème du calcul de l'espérance conditionnelle de la vraisemblance des données complètes sachant les données observées. Pour contourner cette difficulté, McCulloch propose alors une variante de l'algorithme EM qui introduit un algorithme de Metropolis-Hastings dans le but d'approcher par Monte Carlo l'espérance de l'étape E.
Puisque la distribution marginale des observations est très difficile à atteindre, une autre démarche est de s'inscrire dans un raisonnement conditionnel. C'est ce que fait Schall [51], par exemple, en effectuant une linéarisation du modèle. Ainsi replongé dans le cadre linéaire, le problème du calcul intégral est alors contourné. D'autres démarches, s'inscrivant également dans un raisonnement conditionnel puisqu'elles n'abordent pas directement le déconditionnement, ont été développées.

Dans la section suivante, nous revenons en détails sur l'algorithme Monte Carlo EM proposé par McCulloch [39]. Nous avons choisi de décrire cette méthode car nous serons amenés à l'adapter dans le cadre de l'estimation des paramètres d'un mélange quelconque de modèles linéaires généralisés mixtes. Nous nous focalisons ensuite sur les approches dites "de linéarisation". Après avoir décrit précisément, dans un premier temps, la méthode proposée par Schall, nous revenons sur quelques-unes de ces méthodes qui évitent le calcul intégral et qui vont s'avérer être différentes façons de justifier la méthode de Schall.

Dans toute la suite, on considère le vecteur d'observations $y = (y_1, \ldots, y_n)'$, réalisation du vecteur aléatoire Y modélisé par un GL2M de fonction de lien g et de fonction de variance v. On notera $X = (x_1, \ldots, x_n)'$ et $U = (u_1, \ldots, u_n)'$ les matrices d'incidence supposées connues de dimensions respectives $n \times p$ et $n \times q$ associées respectivement aux effets fixes β et aux effets aléatoires ξ. Le vecteur ξ regroupera K effets aléatoires : $\xi = (\xi_1', \ldots, \xi_K')'$ et on supposera : $\forall j \in \{1, \ldots, K\}$, $\xi_j \sim \mathcal{N}_{q_j}(0, \sigma_j^2 A_j)$ avec les matrices A_j, de dimension $q_j \times q_j$, $(q = \sum_{j=1}^{K} q_j)$ connues. Ainsi $\xi \sim \mathcal{N}_q(0, D)$ avec $D = \text{diag}\{\, \sigma_j^2 A_j \,\}_{j=1,\ldots,K}$. Les composantes de la variance, notées $\sigma^2 = (\sigma_1^2, \ldots, \sigma_K^2)$, sont des paramètres inconnus. Remarquons que dans les GL2M, une information supplémentaire est apportée par la fonction de variance. Elle implique, en effet, qu'il n'y a pas de paramètre σ_0^2 associé à la variance des résidus contrairement aux L2M. Cependant, notons tout de même, qu'il nous est aussi laissé la possibilité d'estimer un paramètre de surdis-

persion au même titre que les composantes de la variance si cela est nécessaire.

1.4.2 L'algorithme Monte Carlo EM

1.4.2.1 Limites de l'algorithme EM dans les GL2M

Depuis sa mise en place par Dempster, Laird et Rubin [17] en 1977, l'algorithme EM a permis, dans des contextes variés, de résoudre de nombreux problèmes liés à l'estimation des paramètres. En effet, cet algorithme constitue un outil conceptuellement simple pour obtenir des estimations du maximum de vraisemblance. Il permet, dans diverses situations, de contourner la difficulté d'obtention de la vraisemblance des observations lorsque la distribution marginale de ces observations est difficile à spécifier. Il réalise cela par l'introduction de données manquantes, que l'on n'observe pas directement au cours de l'expérience, mais dont on connait la vraisemblance jointe aux données observées.

Dans les modèles à effets aléatoires, la distribution conditionnelle du vecteur réponse conditionnellement aux effets aléatoires et la distribution marginale des effets aléatoires sont connues. La distribution jointe du vecteur réponse Y et des effets aléatoires ξ s'obtient donc facilement, ce qui est loin d'être le cas de la distribution marginale de Y.
Les effets aléatoires n'étant pas observées, ils joueront logiquement le rôle des données manquantes. La distribution jointe précédente constitue alors la distribution des données complètes. En notant θ le vecteur des paramètres à estimer, cela s'écrit :

$$f(Y, \xi|\theta) = f(Y|\xi, \theta).f(\xi|\theta)$$

où on adopte la notation générique f comme fonction de densité des lois des variables indiquées.
L'algorithme est itératif et se décompose, à chaque itération, en deux étapes. Soit $\theta^{[t]}$ la valeur des paramètres à l'itération $[t]$. Les deux étapes de l'algorithme EM peuvent se résumer de la façon suivante :

- **Étape E :**
  Les effets aléatoires n'étant pas observés, on remplace la log-vraisemblance des données complètes par son espérance selon la distribution conditionnelle des effets aléatoires sachant les données observées :

$$Q(\theta|\theta^{[t]}) = E\big[\ln f(Y, \xi|\theta)|y, \theta^{[t]}\big]$$

– **Étape M :**
On maximise $Q(\theta|\theta^{[t]})$ pour obtenir $\theta^{[t+1]}$:

$$\theta^{[t+1]} = \text{argmax}\, Q(\theta|\theta^{[t]})$$

On itère ces deux étapes jusqu'à convergence. De nombreux travaux ont été réalisés pour étudier les conditions de convergence de cet algorithme (cf Wu [68]). Nous n'insistons pas ici sur ce point.

Comme nous l'avons déjà vu, dans le cadre des L2M, cet algorithme permet d'obtenir des estimations du maximum de vraisemblance ou du maximum de vraisemblance restreint. Par contre, dans le cadre des GL2M, le raisonnement EM n'est pas directement applicable : on butte sur l'obstacle du calcul de l'espérance à l'étape E, réalisable avec la loi normale grâce aux propriétés de déconditionnement mais plus difficile de façon générale avec d'autres lois. Devant la difficulté de ce calcul intégral, McCulloch [39] propose un algorithme où l'étape E est réalisée par une méthode de Monte Carlo via l'algorithme de Metropolis-Hastings. Après une brève description de l'algorithme de Metropolis-Hastings, nous décrivons plus précisément cette démarche dans la sous-section 1.4.2.3.

1.4.2.2 L'algorithme de Metropolis-Hastings

L'algorithme de Metropolis-Hastings, développé en premier lieu par des physiciens en 1940, a été adapté à la simulation statistique en 1970 par Hastings [28]. Cet algorithme est, sans aucun doute, l'une des méthodes MCMC (Markov Chain Monte Carlo) les plus connues et les plus utilisées dans la littérature. L'objectif des méthodes MCMC est de générer des échantillons selon une densité de probabilité "cible" π non calculable de façon explicite. Plus précisément, on appelle algorithme MCMC toute méthode produisant une chaîne de Markov ergodique de loi stationnaire la distribution d'intérêt π. Ainsi, à partir d'un nombre important d'itérations (on parle d'étape de burn-in), les échantillons simulés sont supposés distribués suivant la distribution d'intérêt.

Supposons que l'on sache générer aisément des échantillons selon une densité h que l'on appelle densité "instrumentale" ou "candidate". L'algorithme de Metropolis-Hastings va alors permettre de sélectionner, parmi les échantillons fournis par h, des échantillons représentatifs de π par une méthode d'acceptation/rejet. Nous en décrivons ci-dessous le principe algorithmique (Robert et Casella [50]) :

Algorithme de Metropolis-Hastings

Étant donné $x^{[t]}$,

1. Générer $Y_t \sim h(y|x^{[t]})$
2. Prendre

$$X^{[t+1]} = \begin{cases} Y_t & \text{avec probabilité} \quad \rho(x^{[t]}, Y_t) \\ x^{[t]} & \text{avec probabilité} \quad 1 - \rho(x^{[t]}, Y_t) \end{cases}$$

où

$$\rho(x, y) = \min\left\{1, \frac{\pi(y)h(x|y)}{\pi(x)h(y|x)}\right\}$$

Une question naturelle se pose : quels choix possibles pour h. Plusieurs possibilités sont envisageables. Nous retiendrons ici le cas où h est indépendante de l'événement $x^{[t]}$: on parle alors d'algorithme de Metropolis-Hastings indépendant.

1.4.2.3 Description de l'algorithme MCEM

Pour contourner le problème du calcul de l'espérance de l'étape E de l'algorithme EM, McCulloch [39] introduit une étape de Metropolis-Hastings afin de simuler des effets aléatoires à partir de la distribution conditionnelle des effets aléatoires ξ sachant le vecteur réponse Y. Les effets aléatoires ainsi simulés permettront ensuite d'approcher l'espérance par une méthode de Monte Carlo.

L'algorithme de Metropolis-Hastings nécessite de choisir une distribution instrumentale h à partir de laquelle seront générées des valeurs "potentielles" des effets aléatoires. McCulloch propose de prendre pour h la distribution marginale des effets aléatoires. Notons ici le vecteur de taille q des effets aléatoires $\xi = (\xi_1, \ldots, \xi_q)$ correspondant à la dernière valeur générée selon la distribution conditionnelle de ξ sachant Y et générons une nouvelle valeur, ξ_k^*, pour la $k^{ème}$ composante de ξ, à partir de la distribution instrumentale. La probabilité d'accepter la nouvelle valeur $\xi^* = (\xi_1, \ldots, \xi_{k-1}, \xi_k^*, \xi_{k+1}, \ldots, \xi_q)$ s'écrit :

$$\rho(\xi, \xi^*) = \min\left\{1, \frac{f(\xi^*|y, \theta)h(\xi)}{f(\xi|y, \theta)h(\xi^*)}\right\}$$

où le second terme se simplifie par :

$$\begin{aligned}\frac{f(\xi^*|y,\theta)h(\xi)}{f(\xi|y,\theta)h(\xi^*)} &= \frac{f(\xi^*|y,\theta)f(\xi|\theta)}{f(\xi|y,\theta)f(\xi^*|\theta)} \\ &= \frac{f(y|\xi^*,\theta)f(\xi^*|\theta)f(\xi|\theta)}{f(y|\xi,\theta)f(\xi|\theta)f(\xi^*|\theta)} \\ &= \frac{f(y|\xi^*,\theta)}{f(y|\xi,\theta)} \\ &= \frac{\prod_{i=1}^n f(y_i|\xi^*,\theta)}{\prod_{i=1}^n f(y_i|\xi,\theta)}.\end{aligned}$$

Dans cette approche, cette simplification est essentielle. En effet, l'expression ci-dessus ne dépend plus de la distribution conditionnelle des effets aléatoires ξ sachant le vecteur réponse Y qui est inconnue. En fait, elle nécessite uniquement de connaître la distribution conditionnelle du vecteur réponse Y sachant les effets aléatoires ξ.

L'algorithme MCEM proposé par McCulloch peut se résumer, à l'itération $[t]$, de la façon suivante :

1. Générer M valeurs $\xi^{[1]}, \xi^{[2]}, \ldots, \xi^{[M]}$ à partir de la distribution conditionnelle de ξ sachant les données observées et la valeur courante $\theta^{[t]}$ des paramètres par l'algorithme de Metropolis-Hastings précédent,
2. Calculer $\theta^{[t+1]}$ qui maximise l'approximation de Monte Carlo de $E\big[\ln f(y,\xi|\theta)|y,\theta^{[t]}\big]$ définie par :
$$\frac{1}{M}\sum_{m=1}^{M} \ln f(y,\xi^{[m]}|\theta).$$

On itère ces deux étapes jusqu'à convergence. Dans l'article de McCulloch [39], l'algorithme MCEM est testé et illustré sur simulations mais aucun résultat théorique de convergence ne vient appuyer ces simulations. Les conditions de convergence de cet algorithme ont été étudiées par Sherman et al. [55]. Il est important de souligner que cette méthode est numériquement exigeante du fait qu'elle nécessite de réaliser un nombre important de simulations à chaque étape de l'algorithme. Afin d'éviter ce type d'algorithmes très lourds d'un point de vue numérique, d'autres méthodes d'estimation ont été développées qui effectuent une linéarisation du modèle. Ainsi replongé dans un cadre linéaire, le problème du calcul intégral est alors contourné. Nous revenons ici sur un certain nombre de ces méthodes de linéarisation.

1.4.3 Méthodes de linéarisation

1.4.3.1 La méthode de Schall

En 1991, Schall [51] propose une méthode d'estimation des paramètres dans un GL2M général, c'est-à-dire sans spécification particulière de la loi ou de la modélisation des effets aléatoires. Sa démarche est itérative et consiste, à chaque itération, en une linéarisation du modèle conditionnellement aux effets aléatoires, puis en l'estimation des paramètres par utilisation des équations de Henderson pour les modèles linéaires mixtes.

Étape de linéarisation

Dans un premier temps, Schall se place conditionnellement à ξ. Si ξ était un paramètre fixé, le modèle considéré serait alors un GLM qui pourrait s'écrire :

$$Y = g^{-1}(\eta_\xi) + \varepsilon,$$

avec ε centrée et en utilisant les hypothèses de lois conditionnelles adéquates. Sachant ξ, en reprenant la démarche classique d'estimation des paramètres d'un GLM, on introduit la variable dépendante $z = X\beta + U\xi + (y - \mu_\xi)g^{'}(\mu_\xi)$. Notons que cette variable peut être vue comme un développement de Taylor d'ordre 1 de la fonction de lien g comme fonction des observations y en μ_ξ. Cela conduit Schall à considérer le *modèle linéarisé* que l'on note $\mathcal{M}_\xi$:

$$Z = X\beta + U\xi + e, \tag{1.10}$$

$$\begin{aligned}
\text{où} \quad E(Z|\xi) &= X\beta + U\xi \\
\\
\text{et} \quad \text{var}(Z|\xi) &= \text{var}(e|\xi) \\
&= \text{var}\big((y - \mu_\xi)g'(\mu_\xi)|\xi\big) \\
&= \text{var}(\varepsilon\, g^{'}(\mu_\xi)|\xi) \\
&= \text{diag}\{\, g^{'}(\mu_{\xi,i})^2 \,\text{var}(\varepsilon_i|\xi)\,\}_{i=1,\ldots,n} \\
&= W_\xi.
\end{aligned}$$

Dans le cas d'un lien canonique, W_ξ n'est autre que $\text{diag}\{\, a_i(\phi)\, g^{'}(\mu_{\xi,i})\,\}_{i=1,\ldots,n}$.

En se plaçant conditionnellement à ξ, Schall fait perdre momentanément à l'effet aléatoire sa nature aléatoire.

Étape d'estimation

C'est maintenant l'aspect extension d'un L2M qui est utilisée. Dans le modèle linéarisé $\mathcal{M}_\xi : Z = X\beta + U\xi + e$, on adopte alors la structure d'un L2M où :

$$\begin{aligned} E(Z) &= X\beta \\ \text{var}(Z) &= UDU' + W_\xi \;=\; \Gamma_\xi. \end{aligned}$$

La matrice de variance des erreurs de ce modèle linéaire mixte est donc W_ξ. L'analyse de ce modèle comme un L2M implique alors que ξ retrouve sa nature aléatoire mais seulement partiellement puisque, à l'intérieur de la structure de variance, la matrice de variance conditionnelle des résidus est maintenue. L'utilisation de W_ξ au lieu de $E(W_\xi)$ se justifie dans l'utilisation des équations de Henderson pour obtenir les estimations dans le L2M associé. En effet, la construction de ces équations se base sur la loi du couple (Z, ξ) comme produit de la loi conditionnelle de Z sachant ξ et la loi de ξ. Ainsi dans l'approximation normale de cette loi conditionnelle, c'est bien la matrice W_ξ qui intervient et c'est donc bien celle que l'on retrouve lors de la dérivation des équations.

Le système de Henderson pour le modèle linéarisé de Schall est alors :

$$\begin{pmatrix} X'W_\xi^{-1}X & X'W_\xi^{-1}U \\ U'W_\xi^{-1}X & U'W_\xi^{-1}U + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \xi \end{pmatrix} = \begin{pmatrix} X'W_\xi^{-1}z \\ U'W_\xi^{-1}z \end{pmatrix}. \tag{1.11}$$

Finalement, Schall obtient l'estimation des composantes de la variance correspondant à l'estimation ML et/ou REML par Henderson dans le modèle gaussien défini sur z.

L'algorithme

L'algorithme va suivre les deux étapes de linéarisation puis d'estimation. La dépendance de W_ξ, Γ_ξ et z aux valeurs courantes des paramètres implique un algorithme itératif.

Soient $\beta^{[t]}$, $\xi^{[t]}$ et $\sigma^{2[t]} = (\sigma_1^{2[t]}, \ldots, \sigma_K^{2[t]})'$ les valeurs courantes à l'étape t, l'algorithme effectue les pas suivants :

1. réactualisation des données : on calcule

$$z^{[t]} = X\beta^{[t]} + U\xi^{[t]} + (y - \mu_\xi^{[t]})\, g'(\mu_\xi^{[t]}),$$

où $\mu_\xi^{[t]} = g^{-1}(X\beta^{[t]} + U\xi^{[t]})$

2. calcul de $W_\xi^{[t]}$ et $\Gamma_\xi^{[t]}$,
le modèle $\mathcal{M}^{[t]} : Z^{[t]} = X\beta + U\xi + e^{[t]}$ est alors défini,

3. résolution du système de Henderson :

$$\begin{pmatrix} X'W_\xi^{[t]-1}X & X'W_\xi^{[t]-1}U \\ U'W_\xi^{[t]-1}X & U'W_\xi^{[t]-1}U + D^{[t]-1} \end{pmatrix} \begin{pmatrix} \beta \\ \xi \end{pmatrix} = \begin{pmatrix} X'W_\xi^{[t]-1}z^{[t]} \\ U'W_\xi^{[t]-1}z^{[t]} \end{pmatrix} \tag{1.12}$$

$\beta^{[t+1]}$ et $\xi^{[t+1]}$ sont solutions du système.

4. calcul de $\sigma^{2[t+1]}$
- *ML*

$$\sigma_j^{2[t+1]} = \frac{\xi_j^{'[t+1]} A_j^{-1} \xi_j^{[t+1]}}{q_j - \dfrac{\mathrm{tr}(A_j^{-1} C_{jj}^{*[t]})}{\sigma_j^{2[t]}}}, \qquad j = 1, \ldots, K$$

où C^* : est l'inverse de la matrice formée par les q dernières lignes et colonnes de la matrice des coefficients du système de Henderson (1.12) :
$C^* = (U'W_\xi^{-1}U + D^{-1})^{-1}$,
C_{jj}^* : est la $j^{ème}$ sous-matrice de C^*, correspondant au $j^{ème}$ effet aléatoire.

- *REML*

$$\sigma_j^{2[t+1]} = \frac{\xi_j^{'[t+1]} A_j^{-1} \xi_j^{[t+1]}}{q_j - \dfrac{\mathrm{tr}(A_j^{-1} C_{jj}^{[t]})}{\sigma_j^{2[t]}}}, \qquad j = 1, \ldots, K$$

où C : est la matrice formée des q dernières lignes et colonnes de l'inverse de la matrice des coefficients du système de Henderson (1.12),
C_{jj} : est la $j^{ème}$ sous-matrice de C, correspondant au $j^{ème}$ effet aléatoire.

On itère ce processus jusqu'à convergence de β et σ^2.

Dans son article, Schall justifie cette heuristique, en annexe, en se plongeant dans le cadre bayésien et en maximisant la vraisemblance a posteriori sous un a priori gaussien pour les effets aléatoires. Nous en discuterons plus en détails en section 1.4.4.

1.4.3.2 La méthode de Engel et Keen

La méthode proposée par Engel et Keen [18] consiste, comme celle de Schall, en deux étapes de linéarisation et d'estimation et la mise en place d'une procédure itérative. Elle ne se distingue de la précédente qu'au niveau de la procédure d'estimation.

Lors de la linéarisation, la variable dépendante est introduite de façon identique,

$$z = X\beta + U\xi + (y - \mu_\xi)g^{'}(\mu_\xi).$$

Cependant, le modèle linéaire adopté par la suite diffère. Ici, contrairement à Schall, l'effet aléatoire retrouve entièrement sa nature aléatoire avec le calcul de la variance marginale des erreurs et l'on plonge alors le modèle dans la structure d'un L2M où

$$\begin{aligned} E(Z) &= X\beta \\ \text{var}(Z) &= \text{var}(E(Z|\xi)) + E(\text{var}(Z|\xi)) \\ &= UDU^{'} + E(W_\xi). \end{aligned}$$

L'estimation des paramètres des effets fixes et des composantes de la variance se fait par le même système d'équations (1.11) utilisé par Schall en remplaçant W_ξ par $E(W_\xi)$.

Si le calcul de $E(W_\xi)$ ne pose pas de problème pour les fonctions de liens canoniques, il n'est pas toujours réalisable analytiquement dans les cas de liens non canoniques. Les résultats de cette variance marginale sont présentés pour les lois classiques de la famille exponentielle par Trottier [63].

1.4.3.3 La méthode de Breslow et Clayton

La méthode d'estimation proposée par Breslow et Clayton [7], dite méthode Penalized Quasi-Likelihood (ou méthode de la quasi-vraisemblance pénalisée, notée PQL), est basée sur une approximation de la vraisemblance marginale par approximation de Laplace d'ordre 1. Pour autant, cette démarche s'inscrit davantage dans un raisonnement conditionnel car, plutôt que de faire disparaître les effets aléatoires, la quasi-vraisemblance pénalisée en définitive rajoute un terme les concernant.

Dans un premier temps, Breslow et Clayton définissent une fonction de quasi-vraisemblance marginale en intégrant la quasi-vraisemblance conditionnelle de Y sachant ξ par

rapport à la loi de ξ :

$$Q(\beta,\sigma^2;y) \;\propto\; |D|^{-\frac{1}{2}} \int_{\mathbb{R}^q} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n} d_i(y_i,\mu_{\xi,i}) - \frac{1}{2}\,\xi^{'} D^{-1}\xi \right\} \mathrm{d}\xi, \tag{1.13}$$

où

$$d_i(y_i,\mu_{\xi,i}) \;=\; -2\int_{y_i}^{\mu_{\xi,i}} \frac{y_i - t}{a_i(\phi)v(t)}\,\mathrm{d}t$$

est le logarithme de la fonction de quasi-vraisemblance pour la loi conditionnelle de y_i sachant l'effet aléatoire ξ à un facteur -2 près.

Nous parlerons par la suite de quasi-vraisemblance pénalisée puisqu'un terme de pénalité $-\frac{1}{2}\,\xi^{'} D^{-1}\xi$ a été introduit.

L'équation (1.13) étant de la forme $c|D|^{-\frac{1}{2}} \int e^{-k(\xi)}\,\mathrm{d}\xi$, une log-quasi-vraisemblance pénalisée approchée est obtenue par approximation de Laplace d'ordre 1 en $\overline{\xi}$ solution de $k^{'}(\xi) = 0$:

$$q(\beta,\,\sigma^2) \;\simeq\; -\frac{1}{2}\,\log|D| \;-\; \frac{1}{2}\log|k^{''}(\overline{\xi})| \;-\, k\,(\overline{\xi}), \tag{1.14}$$

avec $\quad k(\xi) = \dfrac{1}{2}\displaystyle\sum_{i=1}^{n} d_i(y_i,\mu_{\xi,i}) + \frac{1}{2}\,\xi^{'} D^{-1}\xi,$

$$k^{'}(\xi) = -\sum_{i=1}^{n} \frac{(y_i-\mu_{\xi,i})u_i}{a_i(\phi)v(\mu_{\xi,i})g^{'}(\mu_{\xi,i})} + D^{-1}\xi \text{ où } u_i' \text{ est la } i^{ème} \text{ ligne de } U,$$

$$k^{''}(\xi) = \sum_{i=1}^{n} \frac{u_i u_i^{'}}{a_i(\phi)v(\mu_{\xi,i})g^{'}(\mu_{\xi,i})^2} + D^{-1} - \sum_{i=1}^{n}(y_i-\mu_{\xi,i})\,u_i\,\frac{\partial}{\partial\xi}\left[\frac{1}{a_i(\phi)v(\mu_{\xi,i})g^{'}(\mu_{\xi,i})}\right]^{'}.$$

Le dernier terme de $k^{''}(\xi)$ est d'espérance nulle et est égal à zéro pour la fonction de lien canonique. En négligeant ce terme, on obtient l'approximation de $k^{''}(\xi)$ suivante :

$$\begin{aligned} k^{''}(\xi) &= U' W_\xi^{-1} U + D^{-1} \\ &= \left(U' W_\xi^{-1} U D + Id_q\right) D^{-1} \end{aligned} \tag{1.15}$$

avec

$$W_\xi = \mathrm{diag}\big\{\, a_i(\phi)v(\mu_{\xi,i})g^{'}(\mu_{\xi,i})^2 \,\big\}_{i=1,\ldots,n}.$$

Finalement, en combinant (1.14) - (1.15), on obtient l'approximation de la log-quasi-vraisemblance pénalisée suivante :

$$q(\beta,\,\sigma^2) \;\simeq\; -\frac{1}{2}\,\log|Id_q + U^{'} W_{\overline{\xi}}^{-1} U D| \;-\; \frac{1}{2}\sum_{i=1}^{n} d_i(y_i,\mu_{\overline{\xi},i}) - \frac{1}{2}\,\overline{\xi}^{'} D^{-1}\overline{\xi}. \tag{1.16}$$

En supposant que $W_{\bar{\xi}}^{-1}$ varie de façon négligeable en fonction des paramètres, on néglige le premier terme et les paramètres qui maximisent l'expression (1.16) sont alors ceux qui maximisent la log-quasi-vraisemblance pénalisée de Green [24] :

$$-\frac{1}{2}\sum_{i=1}^{n} d_i(y_i, \mu_{\xi,i}) - \frac{1}{2}\,\xi' D^{-1}\xi.$$

Or, l'écriture des équations de maximisation en β et ξ de cette log-quasi-vraisemblance pénalisée conduit (cf Breslow et Clayton [7] p.11) au système résolu itérativement de Henderson dans le modèle linéarisé (1.10) de Schall :

$$\begin{pmatrix} X'W_{\xi}^{-1}X & X'W_{\xi}^{-1}U \\ U'W_{\xi}^{-1}X & U'W_{\xi}^{-1}U + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \xi \end{pmatrix} = \begin{pmatrix} X'W_{\xi}^{-1}z \\ U'W_{\xi}^{-1}z \end{pmatrix}.$$

où z est la variable dépendante définie par : $z = X\beta + U\xi + (y - \mu_\xi)g'(\mu_\xi)$.

De la même façon, après plusieurs autres approximations, Breslow et Clayton proposent une estimation des composantes de la variance identique à l'estimation REML par Henderson dans un modèle gaussien défini sur la variable dépendante z.

1.4.3.4 La méthode de Wolfinger

Wolfinger [67] propose une méthode d'estimation des paramètres dans un GL2M. Sa démarche consiste principalement en une linéarisation du modèle par utilisation d'une pseudo-vraisemblance définie après diverses approximations. Dans son article, il considère le GL2M défini par $y = g^{-1}(\eta_\xi) + \varepsilon$ avec $E(\xi) = 0$ et $\mathrm{cov}(\xi) = D$ (D est supposée inconnue). Il suppose, de plus, que $E(\varepsilon|\xi) = 0$ et $\mathrm{var}(\varepsilon|\xi) = R_{\mu_\xi}^{1/2} R R_{\mu_\xi}^{1/2}$, R_{μ_ξ} étant une matrice diagonale associée aux valeurs $v(\mu_{\xi,i})$ de la fonction de variance et R une matrice inconnue.

Dans un premier temps, en considérant $\hat{\beta}$ et $\hat{\xi}$ des estimations connues de β et ξ et en définissant $\hat{\mu}_\xi$ par $\hat{\mu}_\xi = g^{-1}(X\hat{\beta} + U\hat{\xi})$, Wolfinger définit :

$$\tilde{\varepsilon} = y - \hat{\mu}_\xi - (g^{-1})'(X\hat{\beta} + U\hat{\xi})(X\beta - X\hat{\beta} + U\xi - U\hat{\xi}),$$

où $\tilde{\varepsilon}$ peut être vu comme un développement de Taylor d'ordre 1 de $\varepsilon = y - \mu_\xi$ en $X\hat{\beta} + U\hat{\xi}$.

Ensuite la distribution de $\tilde{\varepsilon}$ conditionnellement à β et ξ est approximée par une distribution gaussienne de moyenne nulle et de variance $R_{\mu_\xi}^{1/2} R R_{\mu_\xi}^{1/2}$:

$$\tilde{\varepsilon}|\beta, \xi \sim \mathcal{N}(0\,,\, R_{\mu_\xi}^{1/2} R R_{\mu_\xi}^{1/2}).$$

Une dernière approximation analytique consiste à substituer $\hat{\mu}_\xi$ à μ_ξ dans l'expression de la variance. Pour chaque composante i, nous avons :

$$(g^{-1})^{'}(x_i'\hat{\beta} + u_i'\hat{\xi}) = \frac{1}{g^{'}(\hat{\mu}_{\xi,i})},$$

où x_i' et u_i' sont les $i^{èmes}$ lignes de X et U respectivement.
On peut donc écrire :

$$g^{'}(\hat{\mu}_\xi)(y - \hat{\mu}_\xi)|\beta, \xi \ \sim \ \mathcal{N}(X\beta - X\hat{\beta} + U\xi - U\hat{\xi} \ , \ g^{'}(\hat{\mu}_\xi)R^{1/2}_{\hat{\mu}_\xi}RR^{1/2}_{\hat{\mu}_\xi}g^{'}(\hat{\mu}_\xi)).$$

En définissant la variable z par $z = g(\hat{\mu}_\xi) + g^{'}(\hat{\mu}_\xi)(y - \hat{\mu}_\xi)$, on obtient de manière équivalente :

$$z|\beta, \xi \ \sim \ \mathcal{N}(X\beta + U\xi \ , \ g^{'}(\hat{\mu}_\xi)R^{1/2}_{\hat{\mu}_\xi}RR^{1/2}_{\hat{\mu}_\xi}g^{'}(\hat{\mu}_\xi)).$$

Finalement, en supposant β inconnu et en faisant l'hypothèse de normalité des effets aléatoires $\xi \sim \mathcal{N}(0\,,\, D)$, Wolfinger se ramène au modèle linéarisé :

$$z = X\beta + U\xi + e$$

où

$$\begin{aligned} \text{var}(z|\xi) &= g^{'}(\hat{\mu}_\xi)R^{1/2}_{\hat{\mu}_\xi}RR^{1/2}_{\hat{\mu}_\xi}g^{'}(\hat{\mu}_\xi) \\ &= W_\xi^{1/2}RW_\xi^{1/2}, \end{aligned}$$

et

$$W_\xi = g^{'}(\hat{\mu}_\xi)^2\, R_{\hat{\mu}_\xi}.$$

La variable dépendante z peut être, ici aussi, vue comme un développement de Taylor d'ordre 1 de $g(y)$.

Les paramètres β et ξ sont estimés par utilisation des équations de Henderson dans le modèle linéarisé $z = X\beta + U\xi + e$ où $\text{var}(z) = UDU' + W_\xi^{1/2}RW_\xi^{1/2}$:

$$H\begin{pmatrix} \beta \\ \xi \end{pmatrix} = \begin{pmatrix} X'W_\xi^{-1/2}R^{-1}W_\xi^{-1/2}z \\ U'W_\xi^{-1/2}R^{-1}W_\xi^{-1/2}z \end{pmatrix}.$$

où

$$H = \begin{pmatrix} X'W_\xi^{-1/2}R^{-1}W_\xi^{-1/2}X & X'W_\xi^{-1/2}R^{-1}W_\xi^{-1/2}U \\ U'W_\xi^{-1/2}R^{-1}W_\xi^{-1/2}X & U'W_\xi^{-1/2}R^{-1}W_\xi^{-1/2}U + D^{-1} \end{pmatrix}$$

Notons qu'en prenant R égal à la matrice identité, le système d'équations ci-dessus est équivalent au système d'équations 1.12 de Schall.

Finalement, Wolfinger obtient une estimation des composantes de la variance (et donc de R et D) correspondant à l'estimation ML et/ou REML dans le modèle gaussien défini sur z. Ces deux procédures sont ainsi les procédures d'estimation appelées PL et REPL respectivement. Notons que la démarche de Wolfinger donne la possibilité d'estimer un paramètre additionnel de dispersion ϕ.

1.4.4 Discussion

Toutes ces démarches (l'algorithme MCEM mis à part) s'inscrivent dans un raisonnement conditionnel puisqu'elles mènent à la maximisation de la vraisemblance jointe de (Y, ξ). Et même si la méthode de Breslow et Clayton se justifie par une approximation de la vraisemblance marginale, cela concerne surtout le paramètre β et moins les composantes de la variance qui sont estimées dans une étape ultérieure. En tout cas, toutes ces méthodes évitent le calcul intégral.

La démarche de Breslow et Clayton et celle de Wolfinger se sont avérées être deux façons de justifier la méthode de Schall mise en place en 1991 puisqu'elles aboutissent aux mêmes équations. La méthode de Engel et Keen, quant à elle, diffère au niveau de la procédure d'estimation avec le calcul de la matrice de variance marginale des erreurs. Notons, cependant, que dans toutes ces méthodes, l'estimation des paramètres des effets fixes et des composantes de la variance se fait moyennant une prédiction de l'effet aléatoire ξ.

Comme nous l'avons déjà évoqué, la méthode de quasi-vraisemblance pénalisée de Breslow et Clayton [7] s'appuie sur le fait que les paramètres maximisant la vraisemblance marginale approchée par approximation de Laplace sont ceux qui maximisent la quasi-vraisemblance pénalisée de Green [24]. Notons que cette approximation de Laplace a été reprise dans le cadre général des HGLM (Hierarchical Generalized Linear Models) qui englobent les GL2M par Lee et Nelder [35]. Dans ces modèles, ils définissent la h-vraisemblance comme la vraisemblance jointe définie à partir de la loi jointe comme le produit de la loi de $Y|\xi$ et de ξ. Dans le cas des GL2M, la maximisation de cette vraisemblance jointe est identique à la maximisation de la quasi-vraisemblance pénalisée. Or, Lee

et Nelder montrent que la solution en β de la maximisation de la h-vraisemblance maximise la vraisemblance marginale approchée par approximation de Laplace. La solution en ξ est ensuite utilisée pour l'estimation des composantes de la variance par une procédure de nouveau équivalente dans le cas des GL2M à celle de Schall [51].

Il est également intéressant de souligner ici que la méthode de Schall trouve une justification du point de vue bayésien considéré par Stiratelli, Laird et Ware [57] qui s'intéressent au mode a posteriori lorsque l'on suppose un a priori non informatif uniforme sur β. En effet, comme l'a suggéré Schall [51], en notant $g(\beta|G)$ la distribution a priori normale centrée, de variance G du vecteur de paramètre β, et indépendante de $g(\xi|D)$, la distribution normale des effets aléatoires considérée comme distribution a priori, nous avons :

$$\begin{aligned} f(\beta, \xi|Y, G, D) &\propto f(Y|\beta, \xi)g(\beta|G)g(\xi|D) \\ &\propto f(Y, \xi|\beta, D)g(\beta|G). \end{aligned}$$

Ainsi, en prenant un a priori non informatif (avec une variance infinie ou plus exactement en terme matriciel tel que $G^{-1} \to 0$), on a :

$$f(\beta, \xi|Y, G, D) \propto f(Y, \xi|\beta, D).$$

La maximisation de $f(\beta, \xi|Y, G, D)$ (la densité a posteriori de β et ξ) correspond alors à la maximisation de la vraisemblance jointe de (Y, ξ). Ce raisonnement bayésien conduit donc aux mêmes équations (1.11).

Par la suite, nous reviendrons sur la méthode de Schall présentée précédemment et basée sur un principe de linéarisation du modèle. On cherchera précisément à définir des critères de choix de modèle à partir des modèles linéarisés obtenus à convergence de la procédure d'estimation.

Chapitre 2

Sélection de modèle dans le modèle exponentiel mixte lien log

2.1 Introduction

Dans ce chapitre, nous nous intéressons au problème du choix de modèle dans les GL2M. Plus précisément, nous nous plaçons ici dans le cadre particulier du modèle exponentiel lien logarithme à effets aléatoires gaussiens. Ce problème a été guidé, au départ, par la modélisation de données de défaillance en fiabilité sur lesquelles nous revenons en section 2.2.

Pour traiter de la question de la sélection de modèle en statistique, il existe de nombreux critères permettant de choisir le "meilleur" modèle parmi plusieurs modèles considérés. Parmi eux, un certain nombre sont basés sur la vraisemblance dont les plus connus sont les critères AIC et BIC. Nous présentons ainsi en section 2.3 les principaux fondements de ces deux critères classiques de sélection de modèle.
Dans ce travail, nous cherchons à utiliser ces critères d'information usuels basés sur la vraisemblance. Par commodité, nous noterons, par la suite, IC le critère général d'information défini par :

$$IC = -2 \log f(y|\hat{\theta}) + pen\, K$$

où pen désigne la pénalité imposée pour chaque paramètre introduit dans le modèle et K le nombre de paramètres du modèle. Notons que la démarche que nous proposons ici et que nous présentons dans le cadre du critère général IC s'adaptera tout naturellement à

n'importe quel critère classique basé sur la vraisemblance.

L'utilisation de ces critères est confrontée ici et de façon générale dans les GL2M au problème de la non accessibilité de la vraisemblance marginale. Nous ne pouvons donc pas appliquer directement le critère IC dans le cadre des GL2M et devons chercher à approximer la vraisemblance. Pour résoudre ce problème, nous proposons de traiter la question du choix de modèle conjointement à la procédure d'estimation des paramètres. En effet, différentes démarches d'estimation sont envisagées menant toutes à la construction de modèles linéarisés. Pour chacune de ces démarches, nous définissons alors des critères de choix de modèle basés sur la vraisemblance marginale calculée dans le modèle linéarisé obtenu à la convergence de la procédure d'estimation utilisée.

Dans le cadre des modèles exponentiels mixtes lien log, nous envisageons plusieurs méthodes d'estimation et proposons, pour chacune d'entre elles, un critère associé. Tout d'abord, nous nous intéressons, dans la section 2.4, à une méthode présentée au chapitre 1 et utilisable quelque soit le GL2M : la méthode de Schall. Nous présentons ensuite en section 2.5 une seconde méthode spécifique à la loi exponentielle et développons un critère associé à cette méthode. Dans ces deux approches, les approximations ne sont pas faites aux mêmes endroits, ce qui induit un calcul différent des critères de choix de modèle. Des résultats de simulations sont présentés pour observer le comportement des deux critères développés. Pour finir, en section 2.7, nous définissons un autre critère basé, cette fois-ci, sur une approximation directe de la vraisemblance. Nous le comparons sur simulations aux critères précédents afin de mesurer l'impact des linéarisations sur la qualité de la sélection.

2.2 Modélisation de données de défaillance

L'objet de cette section est de revenir sur le contexte des données qui nous ont conduits à nous placer dans le cadre particulier du modèle exponentiel mixte lien logarithme. En effet, nous avons été amenés à nous intéresser à la modélisation de données de défaillance de matériels supposés sans vieillissement et réparables selon l'hypothèse "as good as new". Plus précisément, nous nous sommes placés dans le cadre particulier où l'on observe plusieurs défaillances par matériel. L'enjeu est alors de modéliser ces répétitions de défaillance sur des matériels que chaque réparation est supposée remettre à neuf.

Lorsqu'on s'intéresse à la modélisation de données de défaillance de matériels c'est-à-dire des durées de fonctionnement avant une panne, la modélisation par la loi exponentielle traduit une hypothèse de non vieillissement : la probabilité de défaillance à un instant donné est indépendante du temps passée en état de fonctionnement (loi sans mémoire). La loi exponentielle va donc jouer ici un rôle central et la paramétrisation considérée est :

$$Y \sim \mathcal{E}xp(\lambda) \quad \begin{cases} \text{Fdr} & F(y) = P(Y < y) = 1 - \exp(-\frac{y}{\lambda}) \\ \text{Fiabilité} & R(y) = 1 - F(y) = \exp(-\frac{y}{\lambda}) \\ \text{Densité} & f(y) = \frac{1}{\lambda}\exp(-\frac{y}{\lambda}), \end{cases}$$

d'où $E(Y) = \lambda$ et $\text{var}(Y) = \lambda^2$.

Si l'on observait de façon indépendante une seule défaillance pour n matériels de même type, l'échantillon $(y_1, y_2, ..., y_n)$ serait constitué de n réalisations indépendantes de la loi $\mathcal{E}(\lambda)$. La log-vraisemblance serait alors :

$$L(\lambda, y) = -n\ln(\lambda) - \sum_{i=1}^{n} \frac{y_i}{\lambda},$$

et l'estimation du paramètre λ par maximum de vraisemblance :

$$\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Dans une situation où on dispose de I matériels réparables et où on observe plusieurs défaillances pour chacun d'entre eux (n_i pour le matériel i), les observations ne peuvent plus être considérées a priori indépendantes même si on suppose que chaque réparation remet le matériel dans son état de fonctionnement d'origine ("as good as new"). En effet, il intervient un effet "matériel". Cet effet matériel est inconnu et nous ne nous intéressons pas à l'estimation de ses niveaux. Au contraire, nous les considérons comme échantillonnés d'une famille infinie de niveaux possibles. L'effet matériel est alors modélisé comme effet aléatoire par une variable aléatoire ξ_i que nous supposerons suivre une loi normale. Ainsi l'effet aléatoire introduit une dépendance marginale entre les données de défaillance pour un même matériel. Par contre, pour un matériel donné, les défaillances successives seront supposées indépendantes entre elles et de loi exponentielle de même paramètre, ce qui traduit l'hypothèse "as good as new".

On aboutit à la modélisation des n_i défaillances y_{ij} du matériel i ($i = 1, \ldots, I$; $j = 1, \ldots, n_i$) :

$$Y_{ij}|\xi_i \sim \mathcal{E}xp(\mu_{\xi,ij})$$

en considérant $\mu_{\xi,ij}$ comme une variable aléatoire qui est ici modélisée par :

$$\mu_{\xi,ij} = \exp(\Theta_{ij}) \text{ avec } \Theta_{ij} \sim \mathcal{N}(x'_{ij}\beta, \sigma^2),$$

ou encore :

$$\mu_{\xi,ij} = \exp(x'_{ij}\beta + \xi_i) \text{ avec } \xi_i \sim \mathcal{N}(0, \sigma^2),$$

les ξ_i étant des variables aléatoires indépendantes et β un paramètre inconnu de dimension $p \times 1$ associé au vecteur de covariables x_{ij} de dimension $p \times 1$.

Notons que le fait de relier $\mu_{\xi,ij}$ aux effets par la fonction exponentielle assure la positivité du paramètre de la loi exponentielle. Rappelons également que les coefficients β (effets fixes), identiques quels que soient les matériels, mesurent l'effet moyen des covariables. L'effet aléatoire, quant à lui, permet de mesurer la variabilité inter-matériel. Par la suite, dans les simulations, nous considèrerons la modélisation la plus simple des effets fixes par une constante commune à toutes les observations. Evidemment, ces effets fixes peuvent être modélisés de façon plus approfondie si on dispose d'information supplémentaire au travers de covariables.

De façon équivalente, en considérant pour le matériel i les observations $y_i = (y_{i1}, \ldots, y_{in_i})'$ réalisation du vecteur aléatoire Y_i, on peut résumer le modèle de la façon suivante pour tout $i = 1 \ldots, I$:

$$Y_i|\xi_i \sim \mathcal{E}xp(\mu_{\xi,i}) \text{ avec } \begin{cases} \mu_{\xi,i} = \exp(X_i\beta + U_i\xi_i) \\ \xi_i \sim \mathcal{N}_(0, \sigma^2) \end{cases}$$

où X_i est la matrice d'incidence de dimension $n_i \times p$ associée aux effets fixes, U_i le vecteur de 1 de dimension $n_i \times 1$ associé au matériel i et β le vecteur des paramètres fixes inconnus de taille p.
Notons que toutes les opérations définies sur le vecteur sont ici des opérations effectuées terme à terme.

Finalement, en utilisant une notation matricielle, nous avons Y de taille $n = \sum_{i=1}^{I} n_i$, ξ

de taille I et μ_ξ de taille n. Notons $X = \begin{bmatrix} X_1 \\ - \\ \vdots \\ - \\ X_I \end{bmatrix}$ la matrice d'incidence associée aux effets fixes de dimension $n \times p$ et $U = \text{diag}\{U_i\}_{i=1,\ldots,I}$ la matrice de dimension $n \times I$. On écrit alors :

$$Y|\xi \sim \mathcal{E}xp(\mu_\xi) \text{ avec } \begin{cases} \mu_\xi = \exp(X\beta + U\xi) \\ \xi \sim \mathcal{N}_{\mathbb{R}^I}(0, D) \text{ où } D = \sigma^2 I_I \end{cases}$$

où I_I est la matrice identité d'ordre I.
Le vecteur aléatoire ξ, à I composantes, est le vecteur de l'effet aléatoire. Il se réalise en $\xi = (\xi_1, \ldots, \xi_I)'$, réalisations dont on rappelle qu'elles ne sont pas directement observées. L'effet "matériel" ξ est ainsi considéré dans ce modèle comme un effet aléatoire qui associe une réalisation pour chaque matériel, autrement dit, il est le même pour toutes les données d'un même matériel.

On définit ainsi un modèle linéaire généralisé mixte caractérisé par :
• conditionnellement à l'effet aléatoire ξ, les défaillances sont indépendantes entre elles et de loi exponentielle,
• les composantes de l'effet aléatoire sont indépendantes et de loi normale centrée de variance σ^2, paramètre à estimer,
• la fonction de lien g est la fonction logarithme.

Avant de nous intéresser à la question du choix de modèle dans le modèle exponentiel mixte lien logarithme, nous revenons dans la section suivante sur les principaux fondements des critères AIC et BIC.

2.3 Principaux fondements des critères classiques AIC et BIC

Dès qu'un phénomène, qu'il soit physique, biologique ou autre, est trop complexe pour accéder à une description analytique débouchant sur une modélisation déterministe, on tente d'en modéliser au mieux le comportement à partir d'observations. À cette fin, un ensemble de méthodes statistiques ont été développées. Tous les auteurs s'accordent

néanmoins pour souligner l'importance qu'il y a à construire des modèles parcimonieux. Plus un modèle est complexe, plus il intègre de paramètres et plus il est capable de s'ajuster correctement aux données. Par contre, un tel modèle pourra s'avérer être défaillant lorsqu'il s'appliquera à des données qui n'ont pas servi à son estimation, comme par exemple pour faire de la prédiction. Ce problème de parcimonie peut s'illustrer en régression classique. Ajouter des variables explicatives dans un modèle réduit l'erreur d'ajustement et réduit le biais. Mais ajouter des variables fait aussi augmenter la variance des estimateurs et dégrade donc les prédictions. L'erreur quadratique de prédiction s'exprime comme le carré du biais plus la variance. Il est donc primordial de trouver un équilibre entre le biais et la variance en contrôlant le nombre de variables. Ces remarques conduisent à la définition de critères de choix de modèle tels que le C_p de Mallows [37], le critère d'information d'Akaike (AIC) [1, 2] ou encore le critère de Schwarz (BIC) [52] pour ne citer que les plus connus, tous basés sur ce principe de parcimonie.

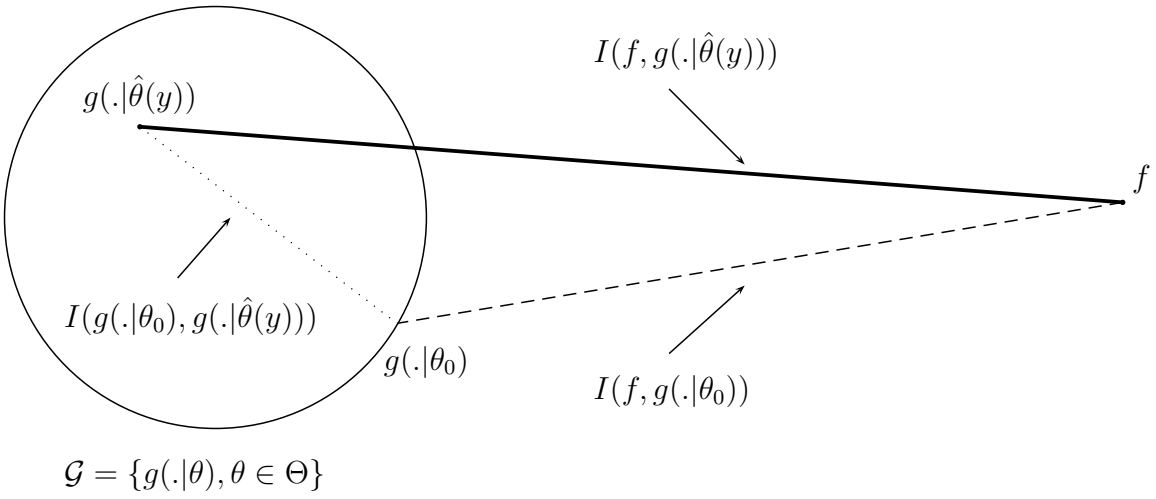
Dans la littérature, de nombreux ouvrages, par exemple [36, 41], sont consacrés à la sélection de modèle. Dans un premier temps, nous revenons sur le principe général qui sous-tend cette sélection et plus précisément sur cet équilibre biais-variance. Puis, après avoir défini brièvement l'information de Kullback-Leibler, nous présentons deux critères classiques de choix de modèle respectant ce principe de parcimonie et basés sur le calcul de la vraisemblance : le critère AIC et le critère BIC.

2.3.1 Le principe de sélection de modèle

En général en statistique, nous disposons d'observations y. Ces observations sont des réalisations d'une variable aléatoire de densité f inconnue. L'objectif de la sélection de modèle en statistique paramétrique est d'approcher f à partir d'un modèle défini par une famille paramétrée de densités $\mathcal{G} = \{g(.|\theta), \theta \in \Theta\}$, Θ étant de dimension finie.
La densité f étant inconnue, il est difficile de savoir si le modèle considéré est approprié. Le problème de la sélection de modèle est alors d'évaluer l'éloignement de l'estimateur $g(.|\hat{\theta}(y))$ à la réalité f au moyen d'une mesure d'information $I(f, g(.|\hat{\theta}(y)))$ interprétée comme une "distance" entre deux distributions.
Dans le cas où le modèle considéré n'est pas approprié, l'inférence basée sur le modèle sera mauvaise. Ainsi, plus le modèle est pauvre, plus le risque de considérer une loi non appropriée est grand. Par contre, plus le modèle est riche et plus on a de chance de disposer,

au sein de la famille paramétrée définissant le modèle considéré, d'une loi $g(.|\theta_0)$ proche de la réalité. Cette qualité d'approximation du modèle se mesure par la quantité $I(f, g(.|\theta_0))$, souvent appelée "terme de biais". Ce terme est nul lorsque la vraie loi appartient au modèle considéré. Cependant, plus le modèle considéré est riche et plus $g(.|\hat{\theta}(y))$ risque d'être loin de $g(.|\theta_0)$ et donc de f en terme d'estimation. Cette erreur d'estimation est d'autant plus grande que le modèle considéré est complexe. Elle se traduit par la quantité $I(g(.|\theta_0), g(.|\hat{\theta}(y)))$, souvent appelée "terme de variance". L'objectif de la sélection de modèle est donc de définir le modèle qui réalise le meilleur compromis entre le terme de biais et le terme de variance. La figure 2.1 illustre ce problème de parcimonie.

FIG. 2.1 – Principe de parcimonie



En définitive, choisir un modèle parmi K modèles en compétition définis par des familles paramétrées de densités $\mathcal{G}_k = \{g_k(.|\theta), \theta \in \Theta\}$, $k \in \{1, \ldots, K\}$, revient à déterminer, pour chaque modèle, la quantité $I(f, g_k(.|\hat{\theta}(y)))$. Le modèle sélectionné sera alors celui dont l'estimateur $g_k(.|\hat{\theta}(y)))$ minimise la "distance" $I(f, g_k(.|\hat{\theta}(y)))$ considérée. On parle de "meilleur modèle" au sens de cette "distance".

2.3.2 L'information de Kullback-Leibler

La théorie de l'information et plus particulièrement l'information de Kullback-Leibler [33] est le principal fondement théorique de la sélection de modèle. L'information de Kullback-Leibler se présente comme l'opposée de l'entropie de Boltzman développée en physique et thermodynamique. L'information de Kullback-Leibler entre deux modèles

représentés par les densités f et g supposées continues se définit par la relation :

$$I(f,g) = \int f(x)\, \log\left\{\frac{f(x)}{g(x)}\right\} \mathrm{d}x.$$

Cette quantité mesure l'information perdue quand g est utilisée pour approcher f. Elle est toujours positive et ne s'annule que lorsque $f \equiv g$. Cette information entre deux modèles f et g n'est pas, au sens exact, une distance puisque la mesure entre f et g est différente de la mesure entre g et f.

Nombreux sont les critères qui se basent sur cette information. On peut citer notamment le critère d'Akaike présenté dans la section suivante qui définit le "meilleur modèle" au sens de l'information de Kullback-Leibler.

2.3.3 Le critère AIC

Pour définir le critère AIC (Akaike's Information Criterion) comme critère de sélection de modèle, Akaike [1, 2] s'est basé sur l'information de Kullback-Leibler. Nous reprenons ici les principales étapes de construction du critère AIC en nous appuyant sur les travaux de Burnham et Anderson [8]. D'autres auteurs se sont également intéressés à ce critère. Citons notamment Chow [15], Stone [58] ou encore Bozdogan [6].

2.3.3.1 Principe d'Akaike

Nous disposons d'un échantillon $y = (y_1, \ldots, y_n)'$, réalisations i.i.d. d'une variable aléatoire de densité f inconnue. Ces observations doivent être modélisées en choisissant entre différents modèles en compétition. Il s'agit alors de mettre en place un critère de sélection parmi ces modèles.

Pour le modèle $\mathcal{M}$ défini par une famille paramétrée de densités $\{g(.|\theta), \theta \in \Theta\}$, Θ étant de dimension finie, la meilleure densité qui approxime f au sens de l'information de Kullback-Leibler est définie par $g(.|\theta_0)$ où :

$$\begin{aligned} \theta_0 &= \arg\min_{\theta \in \Theta} I(f,g) \\ &= \arg\min_{\theta \in \Theta} \int f(x) \log\left\{\frac{f(x)}{g(x|\theta)}\right\} \mathrm{d}x. \end{aligned}$$

Notons que le paramètre θ_0 est solution de l'équation en θ :

$$E_{f(x)}\left[\frac{\partial}{\partial\theta}\log g(x|\theta)\right] = 0. \tag{2.1}$$

Par la suite, pour simplifier les notations, on écrira de façon générale $E_{f(x)}\left[\frac{\partial}{\partial\theta}\log g(x|\theta_0)\right]$ au lieu de $E_{f(x)}\left[\frac{\partial}{\partial\theta}\log g(x|\theta)\right]_{|\theta=\theta_0}$.

Puisque f et θ_0 sont inconnus, les données sont utilisées pour estimer θ_0.
Notons que $\hat{\theta} = \hat{\theta}(y)$, l'estimateur du maximum de vraisemblance pour y sous le modèle $g(y|\theta) = g(y_1, \ldots, y_n|\theta) = \prod_{i=1}^n g(y_i|\theta)$, vérifie l'équation :

$$\frac{\partial}{\partial\theta}\log g(y|\hat{\theta}(y)) = 0.$$

Par la loi forte des grands nombres, nous pouvons écrire :

$$\forall\theta \qquad \frac{1}{n}\sum_{i=1}^{n}\log g(y_i|\theta) \longrightarrow \int f(x)\log g(x|\theta)\,\mathrm{d}x \qquad \text{p.s.} \tag{2.2}$$

Par conséquent, en regroupant (2.1)-(2.2) et sous des conditions de régularité faibles, $\hat{\theta}$ converge presque sûrement vers θ_0. De plus, par le théorème central limite, on a (Ripley [49]) :

$$\hat{\theta} \xrightarrow{\mathcal{L}} \mathcal{N}(\theta_0, I(\theta_0)^{-1}J(\theta_0)I(\theta_0)^{-1}) \tag{2.3}$$

$$\text{avec} \begin{cases} I(\theta_0) &= E_{f(y)}\left[-\dfrac{\partial^2\log g(y|\theta_0)}{\partial\theta^2}\right], \\ J(\theta_0) &= \mathrm{Var}_{f(y)}\left[\dfrac{\partial\log g(y|\theta_0)}{\partial\theta}\right] = E_{f(y)}\left[\left[\dfrac{\partial\log g(y|\theta_0)}{\partial\theta}\right]\left[\dfrac{\partial\log g(y|\theta_0)}{\partial\theta}\right]'\right]. \end{cases}$$

Remarquons que, les espérances étant prises sur f, nous n'avons pas l'égalité $I(\theta_0) = J(\theta_0)$ dans le cas général où f n'est pas un sous-modèle de g (c'est-à-dire $f \subset g$ ou $f = g$).

La quantité $I(f, g(.|\theta_0))$ est ainsi approchée par :

$$I(f, g(.|\hat{\theta}(y))) = \int f(x)\log\left\{\frac{f(x)}{g(x|\hat{\theta}(y))}\right\}\mathrm{d}x.$$

Remarquons que toute valeur $\hat{\theta}$ autre que θ_0 résulte dans $I(f, g(.|\hat{\theta})) > I(f, g(.|\theta_0))$. D'autre part, notons que si nous avions la vraie structure du modèle et si on pouvait trouver θ_0 qui minimise l'information de Kullback-Leibler, le modèle parfait serait tel que $I(f, g) = 0$. Mais comme il est nécessaire d'estimer θ_0, il faut raisonner, dans tous les cas, en terme d'information de Kullback-Leibler prenant, en moyenne, une valeur strictement positive. Et, plutôt que de raisonner sur $I(f, g(.|\hat{\theta}(y)))$ qui est fonction des observations y, Akaike propose de prendre l'espérance sur tous les échantillons possibles. Ainsi, le problème n'est plus la minimisation de $I(f, g(.|\theta_0))$, mais celle de la valeur légèrement plus grande, donnée par

$$E_{f(y)}\big[I(f, g(.|\hat{\theta}(y)))\big] > I(f, g(.|\theta_0)).$$

Le critère devient alors "sélectionner g qui minimise $E_{f(y)}\big[I(f, g(.|\hat{\theta}(y)))\big]$". En écrivant

$$\begin{aligned} E_{f(y)}\big[I(f, g(.|\hat{\theta}(y)))\big] &= \int f(x)\log f(x)\,\mathrm{d}x - E_{f(y)}\big[\int f(x)\log g(x|\hat{\theta}(y))\,\mathrm{d}x\big] \\ &= \text{cste} - E_{f(y)}\big[\int f(x)\log g(x|\hat{\theta}(y))\,\mathrm{d}x\big] \\ &= \text{cste} - E_{f(y)}E_{f(x)}\big[\log g(x|\hat{\theta}(y))\big], \end{aligned}$$

le problème se ramène à la maximisation de $T = E_{f(y)}E_{f(x)}\big[\log g(x|\hat{\theta}(y))\big]$. Cette quantité n'est pas directement calculable et nécessite d'être estimée.

Définissons alors la quantité $T_n = E_{f(y)}E_{f(x)}\big[\log g(x|\hat{\theta}(y))\big]$ où x désigne de la même façon un n-échantillon i.i.d. et indépendant de y. Nous établissons la relation entre la quantité T_n et la quantité T que nous cherchons à estimer :

$$\begin{aligned} T_n &= E_{f(y)}E_{f(x)}\left[\log g(x|\hat{\theta}(y))\right] \\ &= E_{f(y)}\left\{\int_{\mathbb{R}^n}\log g(x|\hat{\theta}(y))f(x)\,\mathrm{d}x\right\} \\ &= E_{f(y)}\left\{\int_{\mathbb{R}^n}\log\left(\prod_{i=1}^n g(x_i|\hat{\theta}(y))\right)\left(\prod_{i=1}^n f(x_i)\right)\mathrm{d}x_1\ldots\mathrm{d}x_n\right\} \\ &= E_{f(y)}\left\{\sum_{i=1}^n\int_{\mathbb{R}^n}\log g(x_i|\hat{\theta}(y))\prod_{i=1}^n f(x_i)\,\mathrm{d}x_1\ldots\mathrm{d}x_n\right\} \\ &= E_{f(y)}\left\{\sum_{i=1}^n\int_{\mathbb{R}}\log g(x_i|\hat{\theta}(y))\,f(x_i)\,\mathrm{d}x_i\right\} \\ &= n\,E_{f(y)}E_{f(x^{(1)})}\left[\log g(x^{(1)}|\hat{\theta}(y))\right] \\ &= n\,T \end{aligned}$$

où $x^{(1)}$ désigne de façon générique un élément quelconque de l'échantillon x.

Nous nous intéressons finalement, de façon équivalente, à cette quantité T_n définie pour un échantillon et cherchons donc à l'estimer.

2.3.3.2 Estimation de $T_n = E_{f(y)}E_{f(x)}\big[\log g(x|\hat{\theta}(y))\big]$

Rappelons ici que x et y sont deux échantillons i.i.d.. Dans un premier temps, il s'agit d'approximer $E_{f(x)}\big[\log g(x|\hat{\theta}(y))\big]$ en utilisant un développement de Taylor de $\log g(x|\hat{\theta}(y))$ d'ordre 2 au voisinage de θ_0 et en calculant son espérance. Nous serons amenés à utiliser la distribution asymptotique de $\hat{\theta}(y)$.

Ainsi, le développement de Taylor de $E_{f(x)}\big[\log g(x|\hat{\theta}(y))\big]$ autour de θ_0 conduit à :

$$\begin{aligned} E_{f(x)}\big[\log g(x|\hat{\theta}(y))\big] &\approx E_{f(x)}\big[\log g(x|\theta_0)\big] + E_{f(x)}\left[\frac{\partial \log g(x|\theta_0)}{\partial \theta}\right]'(\hat{\theta}(y)-\theta_0) \\ &\quad + \frac{1}{2}(\hat{\theta}(y)-\theta_0)' E_{f(x)}\left[\frac{\partial^2 \log g(x|\theta_0)}{\partial \theta^2}\right](\hat{\theta}(y)-\theta_0) \\ &\approx E_{f(x)}\big[\log g(x|\theta_0)\big] - \frac{1}{2}(\hat{\theta}(y)-\theta_0)' I(\theta_0)(\hat{\theta}(y)-\theta_0), \end{aligned}$$

on en déduit :

$$T_n \approx E_{f(x)}\big[\log g(x|\theta_0)\big] - \frac{1}{2}\operatorname{tr}\Big\{ I(\theta_0)\mathrm{Var}_{f(y)}\big[(\hat{\theta}(y)-\theta_0)\big]\Big\}.$$

En remplaçant alors la variance de $\hat{\theta}(y)$ par la variance asymptotique $I(\theta_0)^{-1}J(\theta_0)I(\theta_0)^{-1}$ d'après (2.3), on obtient :

$$T_n \approx E_{f(x)}\big[\log g(x|\theta_0)\big] - \frac{1}{2}\operatorname{tr}\big[I(\theta_0)^{-1}J(\theta_0)\big]. \tag{2.4}$$

Dans un second temps, on effectue un développement de Taylor d'ordre 2 de $\log g(x|\theta_0)$ au voisinage, cette fois-ci, de $\hat{\theta} = \hat{\theta}(x)$, estimateur du maximum de vraisemblance de θ pour un échantillon x quelconque. Nous obtenons ainsi :

$$\begin{aligned} \log g(x|\theta_0) &\approx \log g(x|\hat{\theta}(x)) + \left[\frac{\partial \log g(x|\hat{\theta}(x))}{\partial \theta}\right]'(\theta_0 - \hat{\theta}(x)) \\ &\quad + \frac{1}{2}(\theta_0-\hat{\theta}(x))'\left[\frac{\partial^2 \log g(x|\hat{\theta}(x))}{\partial \theta^2}\right](\theta_0-\hat{\theta}(x)). \end{aligned}$$

Nous prenons ensuite l'espérance sur tous les échantillons possibles :

$$E_{f(x)}\big[\log g(x|\theta_0)\big] \approx E_{f(x)}\big[\log g(x|\hat{\theta}(x))\big] - \frac{1}{2}\,\mathrm{tr}\Big\{E_{f(x)}\Big[-\frac{\partial^2 \log g(x|\hat{\theta}(x))}{\partial\theta^2}\,(\theta_0-\hat{\theta}(x))(\theta_0-\hat{\theta}(x))'\Big]\Big\}.$$

En approximant $\hat{I}(\hat{\theta}) = -\dfrac{\partial^2 \log g(x|\hat{\theta}(x))}{\partial\theta^2}$ par $I(\theta_0)$, nous aboutissons à l'équation suivante :

$$E_{f(x)}\big[\log g(x|\theta_0)\big] \approx E_{f(x)}\big[\log g(x|\hat{\theta}(x))\big] - \frac{1}{2}\,\mathrm{tr}\big[I(\theta_0)^{-1}J(\theta_0)\big]. \tag{2.5}$$

Finalement, en substituant (2.5) dans (2.4), nous obtenons le résultat final :

$$T_n \approx E_{f(x)}\big[\log g(x|\hat{\theta}(x))\big] - \mathrm{tr}\big[I(\theta_0)^{-1}J(\theta_0)\big],$$

avec

$$I(\theta_0) = E_{f(x)}\Big[-\frac{\partial^2 \log g(x|\theta_0)}{\partial\theta^2}\Big] \qquad \text{et} \qquad J(\theta_0) = E_{f(x)}\Big[\Big[\frac{\partial \log g(x|\theta_0)}{\partial\theta}\Big]\Big[\frac{\partial \log g(x|\theta_0)}{\partial\theta}\Big]'\Big].$$

Un critère de sélection de modèle cherchera alors à maximiser une approximation de T_n de la forme :

$$\hat{T}_n \approx \log g(x|\hat{\theta}(x)) - \hat{\mathrm{tr}}\big[I(\theta_0)^{-1}J(\theta_0)\big]$$

où $\hat{\mathrm{tr}}$ dénote un estimateur de la trace.

Notons que $\log g(x|\hat{\theta}(x)) - \hat{\mathrm{tr}}\big[I(\theta_0)^{-1}J(\theta_0)\big]$ est un estimateur de T_n mais c'est $\dfrac{1}{n} \times \big\{\log g(x|\hat{\theta}(x)) - \hat{\mathrm{tr}}\big[I(\theta_0)^{-1}J(\theta_0)\big]\big\}$ qui est un estimateur de T.

En conclusion, le meilleur modèle sera celui qui a la plus grande valeur de $\hat{T}_n$. Le critère est souvent établi comme minimisant

$$-2\log g(x|\hat{\theta}(x)) + 2\,\hat{\mathrm{tr}}\big[I(\theta_0)^{-1}J(\theta_0)\big]. \tag{2.6}$$

2.3.3.3 Résultat

Dans le cas où f est un sous-modèle de g (c'est-à-dire $f \subset g$ ou $f = g$), nous avons l'égalité $I(\theta_0) = J(\theta_0)$ et par conséquent, en notant K la dimension de θ, on a :

$$\mathrm{tr}\big[I(\theta_0)^{-1}J(\theta_0)\big] = K.$$

Le critère AIC est donc un cas particulier de (2.6) et est alors établi comme minimisant :

$$AIC = -2 \log g(x|\hat{\theta}(x)) + 2K,$$

où K correspond au nombre de paramètres à estimer. Il est intéressant de souligner que même si g est juste un bon modèle (c'est-à-dire une bonne approximation) pour f, la littérature soutient l'idée que le meilleur estimateur est probablement $\hat{\text{tr}}\big[I(\theta_0)^{-1}J(\theta_0)\big] = K$ (cf Shibata [56]).

Pour finir, il est important de rappeler que le critère AIC est un résultat asymptotique au sens où toutes les approximations faites ne sont justifiées que dans le cas où n, la taille de l'échantillon considéré, est grand. Plusieurs auteurs [30] ont montré qu'il avait une tendance au sur-ajustement dans le cas de petits échantillons. En réponse à ce problème, Sugiura [59] et Hurvich et Tsai [30] ont proposé des versions corrigées du critère AIC adaptées à des petites tailles d'échantillons par rapport au nombre de paramètres K à estimer. Hurvich et Tsai [30] ont ainsi développé, dans un contexte gaussien, le critère AICc défini par :

$$\text{AICc} = \text{AIC} + \frac{2K(K+1)}{n-K-1}. \tag{2.7}$$

Ce critère conduit donc à utiliser le terme $\dfrac{2K(K+1)}{n-K-1}$ comme un terme de correction de biais dans le cas de petits échantillons. Notons cependant que le résultat (2.7) n'est pas un résultat général. En effet, il suppose un modèle linéaire à effets fixes avec des erreurs normales. Sous l'hypothèse d'un autre modèle, le terme de correction de biais sera différent.

2.3.4 Le critère BIC

Nous nous intéressons maintenant au critère BIC (Bayesian Information Criterion) qui se place dans un contexte bayésien de sélection de modèle. Cette approche repose en particulier sur la notion de facteur de Bayes. Après avoir défini précisément cette notion, nous présentons les principales étapes de construction de ce critère en nous appuyant sur les travaux de Raftery [46] et Kass et Raftery [31].

2.3.4.1 Facteur de Bayes

Supposons que nous cherchons à utiliser les données y pour comparer deux modèles M_1 et M_2 de paramètres respectifs θ_1 et θ_2 appartenant respectivement à des espaces Θ_1

et Θ_2.
Soient $p(M_1)$ et $p(M_2)$ les probabilités a priori des deux modèles.
D'après le théorème de Bayes, la probabilité a posteriori du modèle M_i $(i = 1, 2)$ est :

$$p(M_i|y) = \frac{p(y|M_i)p(M_i)}{p(y)}. \tag{2.8}$$

Dans l'équation (2.8), la probabilité $p(y|M_i)$ s'obtient par intégration sur les valeurs de θ_i :

$$p(y|M_i) = \int_{\Theta_i} p(y|\theta_i, M_i)p(\theta_i|M_i)\,\mathrm{d}\theta_i \tag{2.9}$$

où $p(y|\theta_i, M_i)$ est la vraisemblance de θ_i correspondant au modèle M_i et $p(\theta_i|M_i)$ la distribution a priori des paramètres du modèle M_i. Cette probabilité $p(y|M_i)$ est appelée la vraisemblance intégrée pour le modèle M_i. Elle est aussi appelée vraisemblance marginale ou probabilité marginale des données sous le modèle M_i car on intègre la densité jointe $p(y, \theta_i|M_i)$ par rapport à θ_i.

Á partir de l'équation (2.8), on peut déterminer dans quelle mesure les données y favorisent le modèle M_2 par rapport au modèle M_1 en calculant le rapport de leur probabilité a posteriori :

$$\frac{p(M_2|y)}{p(M_1|y)} = \frac{p(y|M_2)}{p(y|M_1)} \times \frac{p(M_2)}{p(M_1)}.$$

La quantité $\dfrac{p(y|M_2)}{p(y|M_1)}$ est appelée le facteur de Bayes. La quantité $\dfrac{p(M_2)}{p(M_1)}$ est l'odd-ratio a priori et est souvent pris égal à 1 quand il n'y a pas de préférence a priori pour l'un des deux modèles c'est-à-dire quand $p(M_1) = p(M_2)$. Ainsi, le facteur de Bayes est égal au rapport des probabilités a posteriori quand l'odd-ratio a priori est égal à 1. Son évaluation nécessite de calculer la vraisemblance intégrée (2.9). Or, le calcul exact de cette vraisemblance n'est possible que dans quelques cas particuliers comme, par exemple, le cas de la famille exponentielle avec des lois a priori conjuguées. Différentes approximations analytiques et numériques ont été proposées. Nous nous intéressons ici à l'approximation particulière conduisant au critère BIC.

Notons que d'autres aspects du facteur de Bayes sont présentés en détails par Kass et Raftery [31]. Ils considèrent notamment que la log-vraisemblance intégrée peut être aussi vue comme un score prédictif pour le modèle.

2.3.4.2 Approximation BIC

Dans cette section, nous approchons la vraisemblance intégrée (2.9) pour un modèle M et nous montrons ensuite comment cette approximation conduit au critère BIC.

Nous considérons un développement de Taylor de la fonction $g(\theta) = \log\{p(y|\theta, M)p(\theta|M)\}$ d'ordre 2 au voisinage de $\tilde{\theta}$, valeur de θ qui maximise $g(\theta)$ et donc annule $g'(\theta)$:

$$g(\theta) = g(\tilde{\theta}) + (\theta - \tilde{\theta})' g'(\tilde{\theta}) + \frac{1}{2}(\theta - \tilde{\theta})' g''(\tilde{\theta}) (\theta - \tilde{\theta}) + o(\|\theta - \tilde{\theta}\|^2),$$

où $g'(\theta)$ est le vecteur des dérivées premières et $g''(\theta)$ la matrice hessienne des dérivées secondes de $g(\theta)$. Puisque $g'(\tilde{\theta}) = 0$, on obtient l'approximation de $g(\theta)$ suivante :

$$g(\theta) \approx g(\tilde{\theta}) + \frac{1}{2}(\theta - \tilde{\theta})' g''(\tilde{\theta}) (\theta - \tilde{\theta}). \tag{2.10}$$

À partir de (2.10), on peut écrire :

$$\begin{aligned} p(y|M) &= \int \exp[g(\theta)] \, \mathrm{d}\theta \\ &\approx \exp[g(\tilde{\theta})] \int \exp\left[\frac{1}{2}(\theta - \tilde{\theta})' g''(\tilde{\theta}) (\theta - \tilde{\theta})\right] \mathrm{d}\theta. \end{aligned} \tag{2.11}$$

Une approximation de Laplace donne alors :

$$p(y|M) \approx \exp[g(\tilde{\theta})] \, (2\pi)^{K/2} \, |A_{\tilde{\theta}}|^{-1/2}, \tag{2.12}$$

en faisant apparaître dans (2.11) une densité de loi normale multivariée de moyenne $\tilde{\theta}$ et de matrice de variance-covariance inverse $A_{\tilde{\theta}} = -g''(\tilde{\theta})$ et en notant K le nombre de paramètres du modèle. Notons que l'erreur d'approximation dans (2.12) est de l'ordre de $O(\frac{1}{n})$ (Tierney et Kadane [61]). On obtient finalement :

$$\log p(y|M) = \log p(y|\tilde{\theta}, M) + \log p(\tilde{\theta}|M) + \frac{K}{2} \log(2\pi) - \frac{1}{2} \log |A_{\tilde{\theta}}| + O(\frac{1}{n}) \tag{2.13}$$

La difficulté maintenant est l'évaluation de $\tilde{\theta}$ et $A_{\tilde{\theta}}$.
Lorsque n est grand, $\log\{p(y|\theta, M)p(\theta|M)\}$ se comporte comme $\log p(y|\theta, M)$ qui croît avec n alors que $\log p(\theta|M)$ reste constant. Asymptotiquement, $\tilde{\theta}$ peut donc être remplacé par l'estimateur du maximum de vraisemblance $\hat{\theta}$:

$$\hat{\theta} = \mathrm{argmax}_\theta \log p(y|\theta, M),$$

et $A_{\tilde{\theta}}$ par la matrice d'information de Fisher que l'on définit, en supposant que les observations $y = \{y_1, \ldots, y_n\}$ sont indépendantes et identiquement distribuées, par $I = nI_0$ avec $I_0 = E\Big[- \dfrac{\partial^2 \log p(y^{(1)}|\theta, M)}{\partial \theta^2}_{|\theta=\hat{\theta}} \Big]$, la matrice d'information de Fisher de dimension $K \times K$ pour une observation ($y^{(1)}$ désigne de façon générique un élément quelconque de l'échantillon y).

Ces deux approximations introduisent un terme d'erreur en $O(\frac{1}{\sqrt{n}})$ (cf par exemple Lebarbier et Mary-Huard [34]) dans l'équation (2.13). Nous obtenons l'approximation suivante :

$$\begin{aligned} \log p(y|M) &= \overbrace{\log p(y|\hat{\theta}, M) - \frac{K}{2}\log(n)}^{O(n)} \\ &\quad \underbrace{+ \log p(\hat{\theta}|M) + \frac{K}{2}\log 2\pi - \frac{1}{2}\log|I_0|}_{\text{reste borné : } O(1)} + O(\frac{1}{\sqrt{n}}) \end{aligned} \tag{2.14}$$

En négligeant les termes d'erreur $O(1)$ et $O(\frac{1}{\sqrt{n}})$, nous obtenons :

$$\log p(y|M) \approx \log p(y|\hat{\theta}, M) - \frac{K}{2}\log(n) \tag{2.15}$$

Ainsi la log-vraisemblance intégrée $\log p(y|M)$ est égale à la log-vraisemblance maximisée $\log p(y|\hat{\theta}, M)$ moins un terme correcteur.

L'équation (2.15) est l'approximation sur laquelle le critère BIC est basé. L'erreur en $O(\frac{1}{\sqrt{n}})$ dans l'égalité (2.14) est négligeable lorsque n tend vers l'infini. Par contre, l'erreur d'approximation en $O(1)$ est une approximation assez grossière qui signifie, en général, que le terme d'erreur ne diminue pas avec n. Néanmoins, notons que pour certaines distributions a priori sur le paramètre θ, le terme d'erreur peut être plus petit que $O(1)$. Dans le cas d'une distribution multinormale par exemple, le terme d'erreur est de l'ordre de $O(\frac{1}{\sqrt{n}})$ (Raftery [46]).

2.3.4.3 Résultat

Comme nous l'avons souligné dans la section précédente, l'équation (2.15) mise en place par Schwarz [52] est l'approximation qui permet de définir le critère BIC. Plus précisément, pour le modèle M, il correspond à l'approximation de $-2\log p(y|M)$ et est donc défini par :

$$BIC = -2\log p(y|\hat{\theta}, M) + \log(n)\, K.$$

Lors de la comparaison entre différents modèles, le modèle sélectionné sera celui pour lequel la valeur du BIC est minimale.

2.3.5 Comparaison des critères AIC et BIC

Les critères AIC et BIC ont souvent fait l'objet de comparaisons empiriques (Bozdogan [6], Burnham et Anderson [8]). Il a été observé que le critère BIC sélectionne des modèles de dimension plus petite que le critère AIC, ce qui n'est pas surprenant puisque BIC tend à pénaliser plus lourdement les modèles complexes que ne le fait le critère AIC. D'un point de vue théorique, il a été montré que la probabilité pour le critère BIC de choisir le "bon" modèle tend vers 1 lorsque n tend vers l'infini (cf par exemple Lebarbier et Mary-Huard [34]). Ce n'est pas le cas du critère AIC qui tend alors à choisir des modèles plus complexes. Néanmoins, à taille fini, le critère BIC risque de se limiter à des modèles trop simples.
Pour finir, il est important de remarquer qu'en pratique les résultats sur données simulées montrent à quel point les performances de ces deux critères sont fonction de la complexité du vrai modèle et des modèles candidats, et de la taille de l'échantillon. Ces considérations montrent finalement qu'il n'existe pas de critère universellement meilleur et que seuls l'objectif de l'expérimentateur et la connaissance des données peuvent conditionner le choix d'un critère de sélection de modèles.

2.4 Sélection de modèle dans un GL2M quelconque

Dans cette section, nous nous intéressons à la question du choix de modèle dans les GL2M. Nous développons un critère simple de sélection de modèle associé à la méthode d'estimation de Schall utilisable pour un GL2M quelconque c'est-à-dire sans spécification particulière de loi. Pour des raisons de clarté, nous considérons, dans tout ce chapitre, des modèles avec un seul effet aléatoire (à I réalisations). Notons néanmoins que ceci peut se réécrire dans le cas de plusieurs effets aléatoires.

2.4.1 Modèle et notations

Nous considérons le GL2M dont nous rappelons brièvement les hypothèses :

- soit Y le vecteur à expliquer et y son observation dont la composante y_i contient les observations liées au même individu,
- on suppose que, conditionnellement à l'effet aléatoire, les composantes Y_i de Y sont indépendantes et distribuées selon une loi de la famille exponentielle pour laquelle on a :

$$\begin{aligned} E(Y_i|\xi_i) &= \mu_{\xi,i} \\ \text{et} \quad \text{Var}(Y_i|\xi_i) &= a_i(\phi)v(\mu_{\xi,i}) \end{aligned}$$

  où v est la fonction de variance, a_i une fonction connue et ϕ un paramètre de dispersion.
- on considère le prédicteur linéaire :

$$\eta_{\xi,i} = X_i\beta + U_i\xi_i,$$

- on relie ce prédicteur linéaire à l'espérance conditionnelle $\mu_{\xi,i}$ par la fonction de lien g ($h = g^{-1}$) :

$$\eta_{\xi,i} = g(\mu_{\xi,i})$$

2.4.2 Estimation

Nous considérons ici la méthode d'estimation proposée par Schall [51] et présentée au chapitre 1. Rappelons que cette démarche consiste en une alternance entre la linéarisation du modèle conditionnellement à l'effet aléatoire pour une valeur courante des paramètres et l'estimation des paramètres par utilisation des équations de Henderson dans le L2M ainsi obtenu. Cela conduit à l'algorithme itératif suivant pour l'estimation des effets fixes et des composantes de la variance à l'itération $[t]$.

2.4.2.1 L'algorithme

- *Pas 1 :* À partir du prédicteur linéaire : $\eta_\xi^{[t]} = X\beta^{[t]} + U\xi^{[t]}$, définir le vecteur de travail $Z^{[t]}$ selon la technique des GLM :

$$Z^{[t]} = \eta_\xi^{[t]} + (Y - \mu_\xi^{[t]})g'(\mu_\xi^{[t]}),$$

  où $\mu_\xi^{[t]} = g^{-1}(X\beta^{[t]} + U\xi^{[t]})$ et calculer sa réalisation $z^{[t]}$.

- *Pas 2 :* Considérer le modèle linéarisé $\mathcal{M}^{[t]}$ pour les données $z^{[t]}$:

$$\mathcal{M}^{[t]} \; : \quad Z^{[t]} = X\beta + U\xi + \varepsilon^{[t]}$$

avec $W^{[t]}$ la matrice de variance-covariance des erreurs de dimension $n \times n$ définie par :

$$\begin{aligned} W^{[t]} &= \mathrm{Var}(\varepsilon^{[t]}|\xi) \\ &= \mathrm{Var}((Y - \mu_{\xi}^{[t]})g'(\mu_{\xi}^{[t]})|\xi) \\ &= \mathrm{diag}\{\mathrm{var}(Y_i|\xi_i)\; g'(\mu_{\xi,i}^{[t]})^2\}_{i=1,\ldots,I} \\ &= \mathrm{diag}\{\mathrm{var}(Y_{ij}|\xi_i)\; g'(\mu_{\xi,ij}^{[t]})^2\}_{i=1,\ldots,I;j=1,\ldots,n_i} \end{aligned}$$

ce qui dans le cas d'un lien canonique pour lequel $g'(\mu_{\xi,i}) = v(\mu_{\xi,i})^{-1}$ (McCullagh et Nelder [38]) n'est autre que :

$$\begin{aligned} W^{[t]} &= \mathrm{diag}\{a_i(\phi)g'(\mu_{\xi,i}^{[t]})\}_{i=1,\ldots,I} \\ &= \mathrm{diag}\{a_i(\phi)g'(\mu_{\xi,ij}^{[t]})\}_{i=1,\ldots,I;j=1,\ldots,n_i} \end{aligned}$$

- *Pas 3 :* Résoudre les équations de Henderson associées au modèle $\mathcal{M}^{[t]}$ considéré maintenant comme un L2M :

$$\begin{pmatrix} X'W^{[t]-1}X & X'W^{[t]-1}U \\ U'W^{[t]-1}X & U'W^{[t]-1}U + D^{[t]-1} \end{pmatrix} \begin{pmatrix} \beta \\ \xi \end{pmatrix} = \begin{pmatrix} X'W^{[t]-1}z^{[t]} \\ U'W^{[t]-1}z^{[t]} \end{pmatrix} \tag{2.16}$$

pour obtenir $\beta^{[t+1]}$ and $\xi^{[t+1]}$.

- *Pas 4 :* À l'aide de $\xi^{[t+1]}$, obtenir une nouvelle estimation $\sigma^{2[t+1]}$ suivant le schéma ML (Harville [26], Searle et al. [53]) :

$$\sigma^{2[t+1]} = \frac{\xi'^{[t+1]}\xi^{[t+1]}}{\{I - \mathrm{tr}(C^{[t]})/\sigma^{2[t]}\}},$$

C étant l'inverse de la matrice formée par les I dernières lignes et colonnes de la matrice des coefficients du système de Henderson (2.16). Cette valeur $\sigma^{2[t+1]}$ est ensuite utilisée pour obtenir les quantités $z^{[t+1]}$, $W^{[t+1]}$ et $D^{[t+1]}$.

Les pas 1, 2, 3 et 4 sont itérés jusqu'à la convergence de β et σ^2.

2.4.2.2 Remarques

L'algorithme présenté ci-dessus est un algorithme général développé par Schall pour l'estimation des paramètres d'un GL2M quelconque. Son adaptation au modèle exponentiel mixte lien logarithme (lien non canonique) défini en section 2.2 conduit à l'algorithme suivant à l'itération $[t]$:

- *Pas 1 :* À partir du prédicteur linéaire : $\eta_\xi^{[t]} = X\beta^{[t]} + U\xi^{[t]}$, définir le vecteur de travail $Z^{[t]}$:

$$\begin{aligned} Z^{[t]} &= \eta_\xi^{[t]} + (Y - \mu_\xi^{[t]})g'(\mu_\xi^{[t]}) \\ &= \eta_\xi^{[t]} + \frac{Y - \exp(\eta_\xi^{[t]})}{\exp(\eta_\xi^{[t]})} \end{aligned}$$

  et calculer sa réalisation $z^{[t]}$.

- *Pas 2 :* Considérer le modèle linéarisé $\mathcal{M}^{[t]}$ pour les données $z^{[t]}$:

$$\mathcal{M}^{[t]} \ : \quad Z^{[t]} = X\beta + U\xi + \varepsilon^{[t]}$$

  avec $W^{[t]}$ la matrice de variance-covariance des erreurs définie par :

$$\begin{aligned} W^{[t]} &= \text{diag}\{\text{var}(Y_i|\xi_i)\ g'(\mu_{\xi,i}^{[t]})^2\}_{i=1,\ldots,I} \\ &= \text{diag}\{\text{var}(Y_{ij}|\xi_i)\ g'(\mu_{\xi,ij}^{[t]})^2\}_{i=1,\ldots,I;j=1,\ldots,n_i} \\ &= \text{diag}\Big\{\mu_{\xi,ij}^{[t]\,2}\frac{1}{\mu_{\xi,ij}^{[t]\,2}}\Big\}_{i=1,\ldots,I;j=1,\ldots,n_i} \\ &= I_n \end{aligned}$$

- *Pas 3 :* Résoudre les équations de Henderson associées au modèle $\mathcal{M}^{[t]}$ considéré comme un L2M :

$$\begin{pmatrix} X'X & X'U \\ U'X & U'U + D^{[t]-1} \end{pmatrix} \begin{pmatrix} \beta \\ \xi \end{pmatrix} = \begin{pmatrix} X'z^{[t]} \\ U'z^{[t]} \end{pmatrix} \tag{2.17}$$

  avec

$$U'U + D^{[t]-1} = \begin{pmatrix} n_1 + 1/\sigma^{2[t]} & 0 & \ldots & 0 \\ 0 & n_2 + 1/\sigma^{2[t]} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & n_I + 1/\sigma^{2[t]} \end{pmatrix}$$

et

$$U'z^{[t]} = \begin{pmatrix} \sum_{j=1}^{n_1} z_{1j}^{[t]} \\ \sum_{j=1}^{n_2} z_{2j}^{[t]} \\ \vdots \\ \sum_{j=1}^{n_I} z_{Ij}^{[t]} \end{pmatrix}$$

- *Pas 4 :* À l'aide de $\xi^{[t+1]}$, obtenir une nouvelle estimation $\sigma^{2[t+1]}$:

$$\sigma^{2[t+1]} = \frac{\xi'^{[t+1]}\xi^{[t+1]}}{I - \mathrm{tr}(C^{[t]})/\sigma^{2[t]}},$$

avec $C^{[t]} = (U'U + D^{[t]-1})^{-1} = \mathrm{diag}\Big\{\frac{1}{n_i + 1/\sigma^{2[t]}}\Big\}_{i=1\ldots,I}$

2.4.3 Critère de sélection de modèle

La question qui nous préoccupe maintenant est de pouvoir définir un critère de choix de modèle dans les GL2M associé à la méthode d'estimation présentée précédemment. Le problème qui se pose ici est le même que pour l'estimation : nous ne disposons pas d'expression de la log-vraisemblance marginale. Nous proposons alors un critère de sélection de modèle conceptuellement simple mais néanmoins général qui tire partie de la linéarisation introduite dans l'étape d'estimation. Ce critère, que nous appellerons IC^S, est défini pour le modèle linéarisé final $\mathcal{M}^{[f]}$ c'est-à-dire le modèle linéarisé obtenu à la convergence de l'algorithme d'estimation de Schall. Il s'obtient simplement en calculant la vraisemblance définie avec une hypothèse gaussienne pour les données de travail finales $z^{[f]}$ c'est-à-dire les données de travail obtenues à la convergence de l'algorithme d'estimation. L'expression de la log-vraisemblance est alors donnée par :

$$L^S_{GL2M}(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\hat{\Gamma}|) - \frac{1}{2}(z^{[f]} - X\hat{\beta})'\hat{\Gamma}^{-1}(z^{[f]} - X\hat{\beta})$$

où $\hat{\beta}$ et $\hat{\sigma}^2$ sont les estimations de β et σ^2, et $\hat{\Gamma} = \hat{W} + U\hat{D}U'$ est obtenue à partir de ces estimations.

Ainsi, IC^S est défini par :

$$IC^S_{GL2M} = n\log(2\pi) + \log(|\hat{\Gamma}|) + (z^{[f]} - X\hat{\beta})'\hat{\Gamma}^{-1}(z^{[f]} - X\hat{\beta}) + pen\ K$$

où *pen* est le terme de pénalité et $K = p + 1$ est le nombre de paramètres du modèle (β vecteur des paramètres fixes inconnus de taille p et σ^2 le paramètre de variance). Ainsi,

parmi tous les GL2M considérés, le modèle sélectionné sera celui pour lequel la valeur de IC^S est minimale.

Dans nos simulations (voir section 2.6), nous chercherons à comparer des modèles avec et sans effet aléatoire. Par conséquent, nous serons amenés à comparer des valeurs du critère IC entre des GLM et des GL2M.
Dans le cas d'un GLM, nous pouvons directement obtenir l'expression de la log-vraisemblance. En effet, considérons $y = (y_1, \ldots, y_I)$ le vecteur des observations, réalisation du vecteur aléatoire $Y = (Y_1, \ldots, Y_I)$. Les composantes Y_{ij}, $j = 1, \ldots, n_i$, de Y_i sont supposées indépendantes et distribuées selon une loi appartenant à la famille exponentielle, c'est-à-dire que, pour tout $i = 1, \ldots, I$ et $j = 1, \ldots, n_i$, la fonction de densité de la variable aléatoire Y_{ij} s'écrit :

$$f_{Y_{ij}}(y_{ij}, \theta_{ij}) = \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a_{ij}(\phi)} + c(y_{ij}, \phi)\right\}$$

où θ_{ij} est un paramètre canonique et ϕ un paramètre de dispersion. Les fonctions b, c et a_{ij} sont spécifiques à chaque distribution. Notons, de plus, μ_{ij} l'espérance de Y_{ij} liée au prédicteur linéaire $\eta_{ij} = x'_{ij}\beta$ par la fonction de lien g : $\eta_{ij} = g(\mu_{ij})$.

Le critère IC pour un GLM peut alors s'écrire :

$$IC_{GLM} = -2\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left[\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a_{ij}(\phi)} + c(y_{ij}, \phi)\right] + pen\ K$$

où K est ici égal à p, la longueur du vecteur de paramètres β.
Cependant, la valeur de ce critère ne peut pas directement être comparée aux valeurs du critère IC^S car IC et IC^S ne sont pas établis sur la même échelle. Le critère IC est calculé au niveau des données observées y alors que le critère IC^S est calculé au niveau des données de travail z. Il est donc nécessaire d'adapter le critère IC^S au cas du GLM en se plaçant sur la même échelle au niveau du modèle linéarisé. Pour cela, nous devons calculer la log-vraisemblance définie avec une hypothèse gaussienne pour les données de travail $z^{[f]} = X\hat{\beta} + g'(\hat{\mu})(y - \hat{\mu})$ avec $\hat{\mu} = g^{-1}(X\hat{\beta})$ obtenues à la convergence de l'algorithme d'estimation itératif du GLM. En prenant en compte le fait que la matrice de variance-covariance de Z est définie par $W = \text{diag}\{a_{ij}(\phi)v(\mu_{ij})g'(\mu_{ij})^2\}_{i=1,\ldots,I;j=1,\ldots,n_i}$, l'expression de la log-vraisemblance peut alors s'écrire :

$$L^S_{GLM}(\hat{\beta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\hat{W}|) - \frac{1}{2}(z^{[f]} - X\hat{\beta})'\hat{W}^{-1}(z^{[f]} - X\hat{\beta}).$$

Le critère IC^S pour un GLM s'écrit alors :

$$IC^S_{GLM} = n\log(2\pi) + \log(|\hat{W}|) + (z^{[f]} - X\hat{\beta})'\hat{W}^{-1}(z^{[f]} - X\hat{\beta}) + pen\ K$$

avec $K = p$.
Notons que ce $\hat{\beta}$ obtenu à partir de la procédure GLM coïncide avec le $\hat{\beta}$ estimé à partir de la procédure GL2M lorsque $\sigma^2 = 0$. Par conséquent, nous avons $L^S_{GLM}(\hat{\beta}) = L^S_{GL2M}(\hat{\beta}, 0)$ et les valeurs du critère IC^S pour le GLM et pour le GL2M coincident lorsque $\sigma^2 = 0$.

Le critère IC^S est, rappelons-le, un critère général développé à partir du critère IC pour la sélection de modèle dans les GL2M sans aucune spécification de loi ou d'effets particuliers. Nous utiliserons ce critère pour comparer des modèles avec des structures différentes d'effets fixes ou aléatoires mais aussi pour comparer des modèles avec et sans effet aléatoire. Concernant la modélisation des données de défaillance de la section 2.2 pour lesquelles nous avons proposé un modèle à effet aléatoire adapté aux données répétées, comparer les valeurs du critère IC^S pour le modèle avec effet aléatoire et le modèle sans effet aléatoire nous permettra de savoir si oui ou non les données indiquent que le modèle doit prendre en compte les répétitions. Dans le cas où le modèle retenu serait celui sans effet aléatoire, on pourra considérer que les défaillances sont indépendantes entre elles.

2.5 Cas particulier du modèle exponentiel mixte

Dans cette section, nous considérons le modèle exponentiel mixte lien logarithme défini en section 2.2 pour lequel une méthode d'estimation basée sur une propriété spécifique à la loi exponentielle peut être développée (Gaudoin, Lavergne et Soler [21]). Nous l'appellerons la méthode "Gumbel". Comme dans la méthode de Schall, cette procédure d'estimation conduit à un modèle linéarisé permettant de développer un critère de sélection de modèle basé sur le critère IC pour le modèle linéarisé obtenu à la convergence de l'algorithme d'estimation.

2.5.1 Modèle et notations

Nous considérons le modèle exponentiel mixte lien logarithme dont nous rappelons brièvement les hypothèses :

- les composantes de Y de taille n_i sont, conditionnellement à l'effet aléatoire, indépendantes et de loi :

$$\forall i \in \{1, \ldots, I\}, \qquad Y_i|\xi_i \sim \mathcal{E}xp(\mu_{\xi,i}).$$

  Ce qui implique donc notamment $E(Y_i|\xi_i) = \mu_{\xi,i}$ et $\mathrm{Var}(Y_i|\xi_i) = \mu_{\xi,i}^2$
- Chacun des $\mu_{\xi,i}$ est relié à la $i^{\text{ème}}$ composante du prédicteur linéaire $\eta_{\xi,i} = X_i\beta + U_i\xi_i$ par la fonction de lien logarithme :

$$\eta_{\xi,i} = \log(\mu_{\xi,i}) \Leftrightarrow \mu_{\xi,i} = \exp(X_i\beta + U_i\xi_i)$$

- On garde la même distribution normale pour l'effet aléatoire.

2.5.2 Estimation

Dans le cadre du modèle mixte exponentiel lien log, nous avons donc :

$$\forall i \in \{1, \ldots, I\} \qquad Y_i|\xi_i \sim \mathcal{E}xp(\mu_{\xi,i}).$$

Nous pouvons écrire de façon équivalente :

$$\forall i \in \{1, \ldots, I\} \qquad \frac{Y_i}{\mu_{\xi,i}} \sim \mathcal{E}xp(1),$$

ou encore

$$\forall i \in \{1, \ldots, I\} \qquad \log(Y_i) - \log(\mu_{\xi,i}) \sim \mathcal{G}umbel,$$

la densité d'une loi de Gumbel étant définie par : $\forall t \in \mathbb{R}\ f(t) = \exp(t - \exp(t))$. Rappelons également qu'une variable aléatoire de loi de Gumbel a pour espérance la constante d'Euler $\gamma = -0.57722$ et pour variance $\frac{\pi^2}{6}$, ce qui nous permet d'écrire :

$$\log(Y_i) - \log(\mu_{\xi,i}) = \gamma + \varepsilon_i \quad \text{avec} \quad E(\varepsilon_i) = 0_{n_i} \quad \text{et} \quad \mathrm{Var}(\varepsilon_i) = \frac{\pi^2}{6} I_{n_i},$$

ou encore, en posant $Z_i = \log(Y_i) - \gamma$:

$$Z_i = X_i\beta + U_i\xi_i + \varepsilon_i.$$

En utilisant une notation matricielle, nous obtenons :

$$Z = X\beta + U\xi + \varepsilon$$

Finalement, en approchant la loi des erreurs par une loi gaussienne centrée et de matrice de variance-covariance $\frac{\pi^2}{6} I_n$, on définit un modèle linéaire mixte $\mathcal{M}$ "non standard"[1] pour les données $z = \log(y) - \gamma$

Nous procédons alors à l'estimation selon l'algorithme suivant :

- *Pas 1 :* Définir la variable :

$$Z = \log(Y) - \gamma$$

- *Pas 2 :* Résoudre le système de Henderson décrit à l'instant $[t]$ en multipliant toutes les lignes du système par $\frac{\pi^2}{6}$:

$$\begin{pmatrix} X'X & X'U \\ U'X & U'U + \frac{\pi^2}{6} D^{[t]-1} \end{pmatrix} \begin{pmatrix} \beta \\ \xi \end{pmatrix} = \begin{pmatrix} X'z \\ U'z \end{pmatrix}$$

avec

$$U'U + \frac{\pi^2}{6} D^{[t]-1} = \begin{pmatrix} n_1 + \frac{\pi^2}{6\sigma^{2[t]}} & 0 & \dots & 0 \\ 0 & n_2 + \frac{\pi^2}{6\sigma^{2[t]}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & n_I + \frac{\pi^2}{6\sigma^{2[t]}} \end{pmatrix}$$

et

$$U'z = \begin{pmatrix} \sum_{j=1}^{n_1} z_{1j} \\ \sum_{j=1}^{n_2} z_{2j} \\ \vdots \\ \sum_{j=1}^{n_I} z_{Ij} \end{pmatrix}$$

pour obtenir $\beta^{[t+1]}$ et $\xi^{[t+1]}$.

- *Pas 3 :* À l'aide de $\xi^{[t+1]}$, obtenir une nouvelle estimation $\sigma^{2[t+1]}$ suivant le même schéma ML que dans l'algorithme de Schall :

$$\sigma^{2[t+1]} = \frac{\xi'^{[t+1]} \xi^{[t+1]}}{I - \mathrm{tr}(C^{[t]})/\sigma^{2[t]}},$$

avec $C^{[t]} = (U'U + \frac{\pi^2}{6} D^{[t]-1})^{-1} = \mathrm{diag}\Big\{ \frac{1}{n_i + \pi^2/(6\sigma^{2[t]})} \Big\}_{i=1\dots,I}$

[1] modèle linéaire mixte à variance résiduelle connue

À partir de la valeur de $\sigma^{2[t+1]}$ ainsi obtenue, on itère en reprenant la boucle à l'étape 2 avec une nouvelle matrice $D^{[t+1]}$. Les pas 2 et 3 sont itérés jusqu'à convergence mais notons que les données de travail z restent les mêmes au cours des différentes itérations.

2.5.3 Critère de sélection de modèle

Nous cherchons à définir un critère de choix de modèle associé à la méthode d'estimation "Gumbel" présentée ci-dessus. Nous nous appuyons, cette fois-ci, sur le modèle linéaire mixte $\mathcal{M}$ défini précédemment. Nous procédons ainsi au calcul de la log-vraisemblance définie avec une hypothèse gaussienne pour les données $z = \ln(y) - \gamma$ en les paramètres $\hat{\beta}$ et $\hat{\sigma}^2$ obtenus à la convergence de l'algorithme d'estimation précédent. Nous obtenons alors :

$$L^G_{GL2M}(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\hat{\Gamma}|) - \frac{1}{2}(z - X\hat{\beta})'\hat{\Gamma}^{-1}(z - X\hat{\beta})$$

avec $\hat{\Gamma} = \text{diag}\{\frac{\pi^2}{6}\} + U\hat{D}U'$ et $\hat{D}$ définie à partir de $\hat{\sigma}^2$.

Ainsi, cette log-vraisemblance nous conduit à définir le critère suivant :

$$IC^G_{GL2M} = n\log(2\pi) + \log(|\hat{\Gamma}|) + (z - X\hat{\beta})'\hat{\Gamma}^{-1}(z - X\hat{\beta}) + pen\ K$$

avec $K = p + 1$.

Comme dans la section précédente, considérons maintenant le cas du GLM. Ici, la procédure classique d'estimation dans le GLM ne coïncide plus avec la méthode d'estimation "Gumbel" dans le cas où $\sigma^2 = 0$. Il ne s'agit donc plus, comme précédemment, d'adapter uniquement le critère dans le cas GLM en se plaçant sur la même échelle au niveau du modèle linéarisé. En effet, pour rester cohérent, il est nécessaire de faire aussi coïncider les deux procédures d'estimation afin que les deux valeurs du critère pour le GLM et pour le GL2M coïncident bien lorsque $\sigma^2 = 0$. Ainsi, pour le GLM, nous écartons la procédure d'estimation classique et nous utilisons la linéarisation précédente pour estimer le paramètre β. Nous calculons alors la log-vraisemblance définie avec une hypothèse gaussienne pour les données $z = \log(y) - \gamma$ à partir de l'expression suivante :

$$L^G_{GLM}(\hat{\beta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\frac{\pi^2}{6}) - \frac{3(z - X\hat{\beta})'(z - X\hat{\beta})}{\pi^2},$$

et le critère associé IC^G est donné par :

$$IC^G_{GLM} = n\log(2\pi) + n\log(\frac{\pi^2}{6}) + \frac{6(z - X\hat{\beta})'(z - X\hat{\beta})}{\pi^2} + pen\ K$$

avec $K = p$.

2.6 Résultats de simulations

Nous venons de présenter deux critères de sélection de modèle IC^S et IC^G associés à deux méthodes d'estimation des paramètres. Nous disposons donc des quatre critères de sélection AIC^S, AIC^G, BIC^S et BIC^G. Nous les mettons maintenant à l'épreuve sur des données simulées de façon à étudier leur comportement. Nous considérons plus particulièrement le modèle exponentiel mixte lien logarithme pour lequel nous comparons les critères de sélection associés aux deux procédures d'estimation "Schall" et "Gumbel". Dans un premier temps, l'étude numérique du comportement de ces critères est développée pour la sélection d'une structure d'effets fixes. Dans un second temps, on s'intéressera plus particulièrement à mesurer la capacité qu'ont ces critères à détecter la présence d'effets aléatoires. Cette question est déjà une question délicate à traiter dans le cadre des L2M pour lesquels nous n'utilisons aucune approximation. En effet, rappelons que, dans ce cadre, la vraisemblance marginale est calculable directement, ce qui nous permet de calculer la valeur exacte du critère IC. C'est la raison pour laquelle, dans toutes les simulations effectuées, nous considèrerons les résultats obtenus pour le L2M comme des résultats de référence et nous comparerons les performances des critères développés avec celles du cas gaussien. Soulignons également que, pour faire ces comparaisons, nous avons pris soin, dans le cas gaussien, de simuler des données avec une variance de l'erreur du même ordre de grandeur que les variances de l'erreur des modèles linéarisés construits dans le cas exponentiel.

2.6.1 Comparaison de modèles pour les effets fixes

Dans cette section, nous nous intéressons à la sélection de la structure d'effets fixes. Nous considérons un plan d'expérience avec 12 individus et 8 observations par individu, ce qui revient à : $I = 12$ et pour tout $i = 1, \ldots, 12$, $n_i = J = 8$. Il s'agit d'un plan d'expérience équilibré avec un seul effet aléatoire à 12 réalisations. Nous simulons 200 jeux de données à partir du modèle exponentiel mixte lien logarithme défini par :

$$Y_{ij}|\xi_i \sim \mathcal{E}xp(\mu_{\xi,ij}) \qquad i = 1, \ldots, 12;\ j = 1, \ldots, 8$$

avec

$$\log(\mu_{\xi,ij}) = \beta_0 + \beta_1 x_{ij}^{[1]} + \xi_i \quad \text{et} \quad x_{ij}^{[1]} = \begin{cases} 0 \text{ pour } i = 1, \ldots, 6 \\ 1 \text{ pour } i = 7, \ldots, 12 \end{cases}$$

L'effet aléatoire ξ_i est généré selon une distribution normale centrée de variance $\sigma^2 = 0.5$. Le paramètre β_0 est pris égal à 0.5 et les jeux de données sont simulés pour différentes valeurs de β_1.

Nous étudions les performances relatives des critères AIC^S et AIC^G (respectivement BIC^S et BIC^G) définis dans les sections précédentes en comparant trois modèles $\mathcal{M}_0$, $\mathcal{M}_1$ et $\mathcal{M}_2$. Le modèle $\mathcal{M}_1$ est le modèle qui génère les données. Les modèles $\mathcal{M}_0$ et $\mathcal{M}_2$ diffèrent de $\mathcal{M}_1$ par leur prédicteur linéaire défini respectivement par $\beta_0 + \xi_i$ et $\beta_0 + \beta_1 x_{ij}^{[1]} + \beta_2 x_{ij}^{[2]} + \xi_i$ pour lequel

$$x_{ij}^{[1]} = \begin{cases} 0 \text{ pour } i = 1, \ldots, 6 \\ 1 \text{ pour } i = 7, \ldots, 9 \\ 0 \text{ pour } i = 10, \ldots, 12 \end{cases} \quad \text{et } x_{ij}^{[2]} = \begin{cases} 0 \text{ pour } i = 1, \ldots, 6 \\ 0 \text{ pour } i = 7, \ldots, 9 \\ 1 \text{ pour } i = 10, \ldots, 12. \end{cases}$$

Les résultats sont présentés dans le tableau (2.1). Ce tableau présente également les résultats obtenus dans le cas gaussien.

Nous observons une bonne performance globale des critères puisque le modèle $\mathcal{M}_1$ est clairement préféré la plupart du temps et ce d'autant plus que β_1 est grand mais avec une tendance naturelle à sélectionner le modèle $\mathcal{M}_0$ lorsque β_1 devient faible. De plus, les résultats obtenus pour les GL2M sont très similaires à ceux obtenus pour les L2M. Cela doit être souligné puisque, dans le cas particulier des L2M, aucune approximation n'a été nécessaire.

Nous notons également que, de façon globale, les trois critères AIC, AIC^S et AIC^G ont une tendance à sur-ajuster la structure correcte des effets fixes en choisissant le modèle $\mathcal{M}_2$ plus fréquemment que ne le font les critères BIC, BIC^S et BIC^G. En effet, pour $\beta_1 = 4$ par exemple, les proportions de sur-ajustement des critères AIC, AIC^S et AIC^G sont respectivement 16%, 20% et 21.5% alors que celles des critères BIC, BIC^S et BIC^G sont respectivement 5.5%, 6.5% et 5.5%.

Pour finir, notons que nous obtenons le même type de conclusions en faisant varier la taille de l'échantillon n ou la variance de l'effet aléatoire σ^2.

TAB. 2.1 – Sélection de la structure d'effets fixes sur 200 jeux de données

	Estim.	Cas gaussien		Cas exponentiel			
	Modèle	AIC	BIC	AIC^S	AIC^G	BIC^S	BIC^G
$\beta_1 = 0.1$	$\mathcal{M}_0$	145	186	147	142	189	190
	$\mathcal{M}_1^*$	29	9	28	33	8	8
	$\mathcal{M}_2$	26	5	25	25	3	2
$\beta_1 = 1$	$\mathcal{M}_0$	28	80	44	33	106	81
	$\mathcal{M}_1^*$	122	111	114	123	82	106
	$\mathcal{M}_2$	50	9	42	44	12	13
$\beta_1 = 2$	$\mathcal{M}_0$	0	3	5	1	16	9
	$\mathcal{M}_1^*$	153	187	153	150	168	177
	$\mathcal{M}_2$	47	10	42	49	16	14
$\beta_1 = 4$	$\mathcal{M}_0$	0	0	0	0	0	0
	$\mathcal{M}_1^*$	168	189	160	157	187	189
	$\mathcal{M}_2$	32	11	40	43	13	11

* Modèle simulé

2.6.2 Comparaison de modèles pour les effets aléatoires

Nous nous intéressons maintenant à évaluer la capacité qu'ont les critères développés à détecter la présence ou l'absence d'effets aléatoires dans les modèles. Pour cela, nous réalisons, dans un premier temps, des simulations à partir du modèle exponentiel mixte lien logarithme défini par :

$$\mathcal{M}_1 \ : \ \begin{cases} Y_{ij}|\xi_i \sim \mathcal{E}xp(\mu_{\xi,ij}) \\ \xi_i \sim \mathcal{N}(0,\sigma^2) \end{cases} \qquad i = 1,\ldots,12;\ j = 1,\ldots,8$$

avec $\log(\mu_{\xi,ij}) = \beta_0 + \xi_i$. Après avoir fixé $\beta_0 = 1$, nous avons réalisé, pour différentes valeurs de σ^2, 200 simulations. Nous ajustons alors aux données simulées le modèle $\mathcal{M}_1$ et le modèle à effet fixe $\mathcal{M}_0$ pour lequel $\sigma^2 = 0$. Dans un second temps, nous inversons les rôles de $\mathcal{M}_0$ et $\mathcal{M}_1$, c'est-à-dire que nous générons les données à partir du modèle $\mathcal{M}_0$ et nous ajustons de la même façon aux données les deux modèles $\mathcal{M}_0$ et $\mathcal{M}_1$.

Le tableau (2.2) donne le nombre de fois où chaque modèle en compétition a été sélectionné par les critères AIC^S et AIC^G (respectivement BIC^S et BIC^G). Ce tableau présente également les résultats obtenus par les critères AIC et BIC dans le cas gaussien.

A partir du tableau (2.2), nous remarquons que le nombre de fois où le modèle simulé est sélectionné augmente de façon naturelle avec la variance de l'effet aléatoire σ^2 quel que soit le critère considéré. Pour $\sigma^2 = 2$ par exemple, les critères sélectionnent presque toujours le modèle à effet aléatoire $\mathcal{M}_1$. Cependant, comme on pouvait s'y attendre, les critères de type BIC tendent à être plus conservateurs que les critères de type AIC. En effet, ils tendent à choisir plus souvent le modèle plus simple $\mathcal{M}_0$.

Si on s'intéresse aux résultats obtenus dans le cas où les données sont générées à partir du modèle à effet fixe $\mathcal{M}_0$, nous notons que les critères développés à partir de la linéarisation de Schall (AIC^S ou BIC^S) ont tendance à sélectionner un modèle plus complexe. Dans ce cas précis par exemple, AIC^S détecte à tort un effet aléatoire dans 21.5% des cas. Au contraire, les critères développés à partir de la linéarisation "Gumbel" conduisent à des résultats qui s'apparentent fortement aux résultats obtenus dans le cas gaussien.

En ce qui concerne les simulations faites en générant les données à partir du modèle à effet aléatoire $\mathcal{M}_1$, quelle que soit la valeur de σ^2, le modèle simulé $\mathcal{M}_1$ est toujours plus souvent sélectionné par les critères développés à partir de la linéarisation de Schall. Cela ne veut pas

TAB. 2.2 – Sélection de l'effet aléatoire sur 200 jeux de données

	Estim.	Cas gaussien		Cas exponentiel			
	Modèle	AIC	BIC	AIC^S	AIC^G	BIC^S	BIC^G
$\sigma^2 = 0$	$\mathcal{M}_0^*$	191	198	157	191	174	199
	$\mathcal{M}_1$	9	2	43	9	26	1
$\sigma^2 = 0.01$	$\mathcal{M}_0$	190	198	137	182	154	195
	$\mathcal{M}_1^*$	10	2	63	18	46	5
$\sigma^2 = 0.05$	$\mathcal{M}_0$	165	189	111	173	123	195
	$\mathcal{M}_1^*$	35	11	89	27	77	5
$\sigma^2 = 0.1$	$\mathcal{M}_0$	133	172	82	162	98	180
	$\mathcal{M}_1^*$	67	28	118	38	102	20
$\sigma^2 = 0.5$	$\mathcal{M}_0$	24	43	9	38	18	75
	$\mathcal{M}_1^*$	176	157	191	162	182	125
$\sigma^2 = 1$	$\mathcal{M}_0$	0	1	3	4	3	14
	$\mathcal{M}_1^*$	200	199	197	196	197	186
$\sigma^2 = 2$	$\mathcal{M}_0$	0	1	0	0	0	5
	$\mathcal{M}_1^*$	200	199	200	200	200	195

* Modèle simulé

dire pour autant que ces critères ont un comportement plus satisfaisant. Premièrement, le modèle simulé n'est pas nécessairement le "meilleur" modèle (particulièrement pour $\sigma^2 = 0.01$). Deuxièmement, en comparant les résultats de ces critères avec ceux obtenus dans le cas gaussien, on remarque que ces critères développés à partir de la linéarisation de Schall (et plus particulièrement AIC^S) tendent à sur-sélectionner le modèle à effet aléatoire, en particulier lorsque σ^2 est petit. A l'inverse, les critères développés à partir de la linéarisation "Gumbel" ont tendance à sous-sélectionner le modèle à effet aléatoire lorsque σ^2 croît. Cette tendance devient plus prononcée pour le critère BIC^G.

2.6.3 Discussion

Dans ce chapitre, nous portons un regard différent sur deux procédures d'estimation des paramètres dans les modèles linéaires généralisés mixtes : la méthode proposée par Schall [51] qui peut s'appliquer à n'importe quel modèle, et la procédure "Gumbel" développée dans le cadre bien précis du modèle exponentiel mixte lien logarithme. Pour ces deux différentes méthodes, nous insistons sur leur approche commune de linéarisation qui nous permet de proposer des critères simples de sélection de modèle. Ces critères se basent sur les critères généraux d'information adaptés au modèle linéarisé. Ils peuvent ainsi être facilement utilisés dans le cadre des modèles linéaires généralisés mixtes.
Les simulations effectuées dans le cadre du modèle exponentiel mixte lien logarithme montrent que ces critères ont un comportement similaire aux critères d'information utilisés dans le cas gaussien aussi bien en ce qui concerne la sélection de la structure d'effet fixe qu'aléatoire. Ces simulations indiquent également que la procédure de Schall associée à ce type de critère a tendance à sur-détecter la présence d'un effet aléatoire tandis que le critère basé sur l'approche "Gumbel" semble conduire à des résultats plus proches de ceux du cas gaussien. Cependant, quand la variance de l'effet aléatoire augmente, l'approche "Gumbel" peut s'avérer trop conservatrice envers le modèle à effet fixe.
En pratique, les différents critères peuvent être utilisés. AIC^S sera alors le critère le moins conservateur envers le modèle à effet non aléatoire et BIC^G le critère le moins conservateur envers le modèle à effet aléatoire. En d'autres termes, nous aurons confiance en une conclusion du critère AIC^S en faveur du modèle à effet fixe ou en une conclusion du critère BIC^G en faveur du modèle à effet aléatoire.

2.7 Comparaison de ces critères avec un troisième critère

La comparaison des critères développés en section 2.4 et 2.5 avec des critères construits sur d'autres approximations de la vraisemblance est maintenant une perspective intéressante à ce premier travail. Dans cette section, nous proposons ainsi de comparer les résultats obtenus avec les critères précédents aux résultats obtenus à partir d'un autre critère construit sur une approximation directe de la vraisemblance. Dans un premier temps, nous décrivons l'algorithme MCEM dans le cadre particulier du modèle exponentiel mixte lien logarithme. Nous définissons ensuite un critère de sélection de modèle construit sur une approximation directe de la vraisemblance à partir des paramètres estimés via l'algorithme MCEM. Pour finir, des résultats de simulations seront présentés pour vérifier, dans un premier temps, le bon comportement de l'algorithme MCEM en ce qui concerne l'estimation des paramètres d'effets fixes et des paramètres de variance, puis pour comparer le comportement des différents critères développés.

2.7.1 Algorithme MCEM pour le modèle exponentiel mixte

Si l'on applique directement le raisonnement EM dans le cadre du modèle exponentiel mixte lien logarithme, on butte, comme nous l'avons déjà souligné, sur le calcul de l'espérance de l'étape E. Pour contourner ce problème, nous avons vu que l'algorithme MCEM consistait à introduire une étape de Metropolis-Hastings dans le but de simuler des effets aléatoires à partir de la distribution conditionnelle des effets aléatoires ξ sachant le vecteur réponse Y. Les effets aléatoires ainsi générés permettent alors d'approcher cette espérance par Monte Carlo. Nous décrivons ici plus précisément cette démarche dans le cas du modèle exponentiel mixte lien logarithme.

Considérons le modèle exponentiel mixte lien log défini par :

$\forall i \in \{1, \ldots, I\},\ \forall j \in \{1, \ldots, n_i\}$,

$$Y_{ij}|\xi_i \sim \mathcal{E}xp(\mu_{\xi,ij}) \text{ avec } \begin{cases} \mu_{\xi,ij} = \exp(x'_{ij}\beta + \xi_i) \\ \xi_i \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

Rappelons, qu'en pratique, nous disposons d'observations y_{ij} des Y_{ij} mais nous n'observons pas les effets aléatoires ξ_i.

Remarque : Par la suite, on note $\theta = (\beta', \sigma^2)'$ le vecteur des paramètres à estimer et on adopte la notation générique de f pour désigner les différentes fonctions de densité des lois mises en jeu.

Le vecteur des données complètes x est défini par $x = (y', \xi')'$ avec $\xi = (\xi_1, \ldots, \xi_I)'$. La log-vraisemblance associée aux données complètes s'écrit alors :

$$\begin{aligned} L(\theta|x) &= \sum_{i=1}^{I} \ln f(y_i, \xi_i|\theta) \\ &= \sum_{i=1}^{I} \Big[\ln f(y_i|\xi_i, \theta) + \ln f(\xi_i|\theta) \Big] \end{aligned}$$

avec

- $$\begin{aligned} \ln f(y_i|\xi_i, \theta) &= \sum_{j=1}^{n_i} \ln f(y_{ij}|\xi_i, \theta) \\ &= -\sum_{j=1}^{n_i} \Big[x'_{ij}\beta + \xi_i + \frac{y_{ij}}{\exp(x'_{ij}\beta + \xi_i)} \Big] \end{aligned}$$

puisque les y_{ij} sont indépendants conditionnellement à ξ_i,

- $$\ln f(\xi_i|\theta) = -\frac{1}{2}\Big[\ln 2\pi + \ln \sigma^2 + \frac{\xi_i^2}{\sigma^2} \Big]$$

La première étape de l'algorithme EM nous conduit à considérer l'espérance de $L(\theta|x)$ par rapport à la distribution des données manquantes ξ sachant les observations y et la valeur courante des paramètres $\theta^{[t]}$:

$$\begin{aligned} Q(\theta|\theta^{[t]}) &= E\Big[L(\theta|x)|y, \theta^{[t]} \Big] \\ &= -\sum_{i=1}^{I} \sum_{j=1}^{n_i} \Big[x'_{ij}\beta + E_C^{[t]}(\xi_i) + \exp(-x'_{ij}\beta) E_C^{[t]}\big[\exp(-\xi_i) \big]\, y_{ij} \Big] \\ &\quad -\frac{1}{2}\sum_{i=1}^{I} \Big[\ln 2\pi + \ln \sigma^2 + \frac{E_C^{[t]}(\xi_i^2)}{\sigma^2} \Big] \end{aligned}$$

où $E_C^{[t]}(.)$ désigne l'espérance prise par rapport à la densité des données manquantes sachant les données observées et la valeur courante des paramètres.

Lors de la seconde étape, c'est la maximisation en θ de cette fonction $Q(\theta|\theta^{[t]})$ qui nous intéresse. On dérive donc $Q(\theta|\theta^{[t]})$ par rapport à β et σ^2. On obtient :

$$\begin{aligned}
\frac{\partial Q}{\partial \sigma^2} &= -\frac{1}{2}\sum_{i=1}^{I}\Big[\frac{1}{\sigma^2} - \frac{E_C^{[t]}(\xi_i^2)}{\sigma^4}\Big] \\
\frac{\partial Q}{\partial \beta} &= \sum_{i=1}^{I}\sum_{j=1}^{n_i} x_{ij}\Big[\exp(-x'_{ij}\beta)E_C^{[t]}\big[\exp(-\xi_i)\big]\, y_{ij} - 1\Big]
\end{aligned}$$

Par annulation des dérivées, il vient :

$$\sigma^{2[t+1]} = \frac{1}{I}\sum_{i=1}^{I} E_C^{[t]}(\xi_i^2)$$

Pour β, nous pouvons écrire :

$$\begin{aligned}
\frac{\partial Q}{\partial \beta} &= \sum_{i=1}^{I}\sum_{j=1}^{n_i} x_{ij}\Big[\exp(-x'_{ij}\beta)E_C^{[t]}\big[\exp(-\xi_i)\big]\, y_{ij} - 1\Big] \\
&= \sum_{i=1}^{I} X_i'\Big[\exp(-X_i\beta) * E_C^{[t]}\big[\exp(-U_i\xi_i)\big] * y_i - \mathbb{1}_{n_i}\Big]
\end{aligned}$$

où $*$ représente une multiplication terme à terme (on ne fera plus la distinction par la suite) et $\mathbb{1}_{n_i}$ un vecteur de 1 de longueur n_i. En notant $\tilde{y}_i = E_C^{[t]}\big[\exp(-U_i\xi_i)\big]y_i$, l'équation précédente devient :

$$\frac{\partial Q}{\partial \beta} = \sum_{i=1}^{I} X_i'\Big[\exp(-X_i\beta)\tilde{y}_i - \mathbb{1}_{n_i}\Big]$$

Par annulation des dérivées, on obtient :

$$\begin{aligned}
\sum_{i=1}^{I} X_i'\Big[\exp(-X_i\beta)\tilde{y}_i - \mathbb{1}_{n_i}\Big] &= 0 \\
\iff \sum_{i=1}^{I} X_i'\Big[X_i\beta + \exp(-X_i\beta)\tilde{y}_i - \mathbb{1}_{n_i} - X_i\beta\Big] &= 0
\end{aligned}$$

En introduisant le vecteur dépendant de taille n_i défini par

$$z_{\beta,i}^{[t]} = X_i\beta^{[t]} + \exp(-X_i\beta^{[t]})\tilde{y}_i - \mathbb{1}_{n_i},$$

on peut résoudre en β :

$$\Big(\sum_{i=1}^{I} X_i'X_i\Big)\beta^{[t+1]} = \sum_{i=1}^{I} X_i' z_{\beta,i}^{[t]}$$

Finalement, en considérant les matrices $X = \begin{bmatrix} X_1 \\ \hline \vdots \\ \hline X_I \end{bmatrix}$, $U = \operatorname{diag}\{U_i\}_{i=1,\ldots,I}$ et le vecteur de l'effet aléatoire $\xi = (\xi_1, \ldots, \xi_I)'$ de taille I, on aboutit à l'écriture matricielle suivante :

$$\beta^{[t+1]} = \big(X'X\big)^{-1} X' z_\beta^{[t]}$$

où $z_\beta^{[t]}$ est le vecteur dépendant de taille n défini par $z_\beta^{[t]} = X\beta^{[t]} + \exp(-X\beta^{[t]})\tilde{y} - \mathbb{1}_n$ et $\tilde{y} = E_C^{[t]}\big[\exp(-U\xi)\big]y$. Il est intéressant de remarquer qu'en considérant les données transformées $\tilde{y} = E_C^{[t]}\big[\exp(-U\xi)\big]y$, nous retrouvons ici le même type de données de travail que dans le GLM défini par la loi exponentielle et la fonction de lien log.

Cette maximisation conduit ainsi à des expressions qui dépendent des espérances conditionnelles $E_C^{[t]}(\xi_i^2)$, et $E_C^{[t]}\big[\exp(-\xi_i)\big]$. À ce stade, l'algorithme EM se trouve confronté au problème du calcul de ces espérances pour lesquelles il n'est pas envisageable d'obtenir des expressions explicites. L'algorithme MCEM permet alors de générer des effets aléatoires à partir de la distribution conditionnelle de ξ_i sachant les données y_i et la valeur courante des paramètres $\theta^{[t]}$ par l'intermédiaire d'un algorithme de Metropolis-Hastings introduit à chaque itération de l'algorithme. Les effets aléatoires simulés permettent alors d'approcher par Monte Carlo ces espérances.

Pour mettre en place l'algorithme de Metropolis-Hastings, il est nécessaire de définir une distribution instrumentale h à partir de laquelle sont générées les valeurs "potentielles" des effets aléatoires ξ_i. Nous prenons ici pour h la distribution marginale des effets aléatoires :

$$\begin{aligned} h(\xi_i) &= f(\xi_i|\theta) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{\xi_i^2}{2\sigma^2} \right\} \end{aligned}$$

Dans le cadre du modèle exponentiel mixte, l'algorithme de Metropolis-Hastings se résume de la façon suivante, pour $i = 1, \ldots, I$:

Étant donné $\xi_i^{[m]}$ la dernière valeur générée,

1. Générer $\xi_i^* \sim f(\xi_i|\theta)$

2. Prendre

$$\xi_i^{[m+1]} = \begin{cases} \xi_i^* & \text{avec probabilité} \quad \rho(\xi_i^{[m]}, \xi_i^*) \\ \xi_i^{[m]} & \text{avec probabilité} \quad 1 - \rho(\xi_i^{[m]}, \xi_i^*) \end{cases}$$

où

$$\begin{aligned} \rho(\xi_i^{[m]}, \xi_i^*) &= \min\left\{1, \frac{f(\xi_i^*|y_i,\theta) f(\xi_i^{[m]}|\theta)}{f(\xi_i^{[m]}|y_i,\theta) f(\xi_i^*|\theta)}\right\} \\ &= \min\left\{1, \frac{f(y_i|\xi_i^*,\theta)}{f(y_i|\xi_i^{[m]},\theta)}\right\} \end{aligned}$$

Pour finir, l'algorithme MCEM se résume, à l'itération $[t+1]$, de la façon suivante :

1. Pour $i = 1, \ldots, I$, on génère M valeurs $\xi_i^{[1]}, \ldots, \xi_i^{[M]}$ à partir de la distribution de ξ_i sachant Y_i en la valeur courante $\theta^{[t]}$ des paramètres par l'algorithme de Metropolis-Hastings précédent et on approche les espérances conditionnelles de la fonction $Q(\theta|\theta^{[t]})$ par :

$$\begin{aligned} E_C^{[t]}(\xi_i^2) &\simeq \frac{1}{M}\sum_{m=1}^{M} \xi_i^{[m]2} \\ E_C^{[t]}\big[\exp(-\xi_i)\big] &\simeq \frac{1}{M}\sum_{m=1}^{M} \exp(-\xi_i^{[m]}) \end{aligned}$$

2. On maximise ensuite la fonction $Q(\theta|\theta^{[t]})$ pour obtenir la nouvelle valeur des paramètres $\theta^{[t+1]}$ à l'itération $[t+1]$.

2.7.2 Critère de sélection de modèle

Dans l'esprit de cette procédure d'estimation MCEM, nous cherchons à définir un critère de choix de modèle basé sur une approximation directe de la vraisemblance. Nous construisons un critère de sélection de modèle basé sur une approximation de Monte Carlo de la vraisemblance marginale à partir des estimations $\hat{\theta} = (\hat{\beta}, \hat{\sigma}^2)$ obtenues via l'algorithme MCEM.

L'approximation de Monte Carlo de la vraisemblance est donnée par :

$$\begin{aligned} l_{GL2M}(\hat{\beta},\hat{\sigma}^2) &= \prod_{i=1}^{I} f(y_i|\hat{\beta},\hat{\sigma}^2) \\ &= \prod_{i=1}^{I} \int f(y_i|\xi_i,\hat{\beta},\hat{\sigma}^2) f(\xi_i|\hat{\sigma}^2)\, d\xi_i \\ &= \prod_{i=1}^{I} \int \prod_{j=1}^{n_i} f(y_{ij}|\xi_i,\hat{\beta},\hat{\sigma}^2) f(\xi_i|\hat{\sigma}^2)\, d\xi_i \\ &\approx \prod_{i=1}^{I} \left\{ \frac{1}{N} \sum_{n=1}^{N} \Big[\prod_{j=1}^{n_i} f(y_{ij}|\xi_i^{[n]},\hat{\beta},\hat{\sigma}^2) \Big] \right\} \end{aligned} \tag{2.18}$$

où les $\xi_i^{[n]}$ sont générés à partir de leur distribution connue en la valeur estimée $\hat{\sigma}^2$ et N est le nombre de valeurs générées.

À partir de l'expression (2.18), on obtient l'approximation de la log-vraisemblance suivante :

$$\begin{aligned} \hat{L}_{GL2M}(\hat{\beta},\hat{\sigma}^2) &= \sum_{i=1}^{I} \log \left\{ \frac{1}{N} \sum_{n=1}^{N} \Big[\prod_{j=1}^{n_i} f(y_{ij}|\xi_i^{[n]},\hat{\beta},\hat{\sigma}^2) \Big] \right\} \\ &= \sum_{i=1}^{I} \log \left\{ \frac{1}{N} \sum_{n=1}^{N} \left[\prod_{j=1}^{n_i} \frac{1}{\exp(x'_{ij}\hat{\beta}+\xi_i^{[n]})} \exp\Big\{ -\frac{y_{ij}}{\exp(x'_{ij}\hat{\beta}+\xi_i^{[n]})} \Big\} \right] \right\} \end{aligned}$$

et le critère associé est alors défini par :

$$\widehat{IC}_{GL2M} = -2\, \hat{L}_{GL2M}(\hat{\beta},\hat{\sigma}^2) + pen\ K$$

où *pen* est le terme de pénalité et K le nombre de paramètres du modèle.

Dans les simulations, nous cherchons à comparer des modèles avec et sans effet aléatoire. Contrairement aux deux critères développés précédemment, ce critère est basé sur une approximation directe de la vraisemblance. Il est ainsi calculé sur l'échelle des données observées y. La valeur de ce critère peut donc être directement comparée à la valeur classique du critère dans le cas du GLM afin de choisir entre les modèles avec et sans effet aléatoire. Dans le cas d'un modèle exponentiel, rappelons que la log-vraisemblance

est donnée par :

$$\begin{aligned} L_{GLM}(\hat{\beta}) &= \sum_{i=1}^{I} \log \left\{ \prod_{j=1}^{n_i} \frac{1}{\exp(x'_{ij}\hat{\beta})} \exp \left\{ - \frac{y_{ij}}{\exp(x'_{ij}\hat{\beta})} \right\} \right\} \\ &= - \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left[x'_{ij}\hat{\beta} + \frac{y_{ij}}{\exp(x'_{ij}\hat{\beta})} \right], \end{aligned}$$

où $\hat{\beta}$ est obtenu par la procédure classique d'estimation des paramètres d'un GLM.
Le critère associé est alors défini par :

$$IC_{GLM} = -2\, L_{GLM}(\hat{\beta}) + pen\ K.$$

2.7.3 Simulations

2.7.3.1 Préliminaires

Afin de vérifier le bon comportement de l'algorithme MCEM, nous avons fait tourner en parallèle les méthodes d'estimation de Schall, "Gumbel" et l'algorithme MCEM. De nombreuses simulations numériques ont ainsi été effectuées. Notons que l'algorithme MCEM étant numériquement très exigeant, les résultats relatifs à cette approche ont été obtenus par l'utilisation d'algorithmes implémentés en C.

Dans le tableau (2.3), nous présentons les résultats de simulation obtenus dans le cas d'un modèle exponentiel mixte avec lien logarithme. Nous considérons un plan d'expérience avec un seul effet aléatoire ayant 4 réalisations selon un plan équilibré. Pour chaque valeur de la variance σ^2 de cet effet (allant de 0.01 à 2), nous simulons 200 vecteurs de données de longueur 40. La valeur de β est prise égale à 1. Le tableau contient alors le résumé (moyenne, écart-type) des 200 simulations pour chacune des trois méthodes.

À l'aide de ce tableau et d'autres simulations réalisées, nous pouvons faire les remarques suivantes.

- On constate un bon comportement général des trois méthodes, que ce soit pour l'estimation des paramètres d'effets fixes ou pour celle des paramètres de variance. Notons néanmoins que les trois méthodes ont tendance à sous-estimer σ^2 quand il est grand et à le sur-estimer quand il est petit. On observe également une tendance générale à sous-estimer β.

TAB. 2.3 – Résultats d'estimation des paramètres obtenus par les trois méthodes dans le cas d'un modèle exponentiel - lien log

	Valeurs estimées					
	Schall		Gumbel		MCEM	
Valeurs simulées	$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{\beta}$	$\hat{\sigma}^2$
$\beta = 1$ $\sigma^2 = 0.01$	0.9621	0.0361	0.9899	0.0730	0.9703	0.0544
	(0.1639)	(0.0710)	(0.1980)	(0.1442)	(0.1577)	(0.0038)
$\beta = 1$ $\sigma^2 = 0.05$	1.0075	0.0545	1.0326	0.0655	1.0184	0.0769
	(0.1920)	(0.0908)	(0.2064)	(0.1162)	(0.1857)	(0.0067)
$\beta = 1$ $\sigma^2 = 0.5$	0.9495	0.3619	1.0109	0.3938	1.0104	0.3686
	(0.2580)	(0.2938)	(0.2885)	(0.3763)	(0.2564)	(0.0796)
$\beta = 1$ $\sigma^2 = 1$	0.8954	0.9089	0.9822	0.8912	0.9851	0.9062
	(0.3758)	(0.6221)	(0.4151)	(0.6653)	(0.3732)	(0.3720)
$\beta = 1$ $\sigma^2 = 2$	0.8694	1.7708	0.9740	1.7769	0.9798	1.7558
	(0.4941)	(0.9480)	(0.4958)	(0.9970)	(0.4911)	(0.8831)

- Bien entendu, plus la valeur simulée de la composante de la variance est importante, plus la précision de ces estimations diminue au sens où les écarts-types augmentent.

- On note une meilleure précision des estimations obtenues par l'algorithme MCEM en comparaison avec la précision des estimations obtenues par la méthode de Schall ou "Gumbel".

2.7.3.2 Comparaison de modèles pour les effets fixes

Comme dans la section 2.6.1, nous nous intéressons ici à la sélection de la structure d'effets fixes. Nous considérons un plan d'expérience avec $I = 12$ individus et $J = 6$ observations par individu. Nous générons 200 jeux de données à partir du modèle exponentiel mixte lien log défini par :

$$Y_{ij}|\xi_i \sim \mathcal{E}xp(\mu_{\xi,ij}) \qquad i = 1, \ldots, 12;\ j = 1, \ldots, 6$$

avec

$$\log(\mu_{\xi,ij}) = \beta_0 + \beta_1 x_{ij}^{[1]} + \xi_i \quad \text{et} \quad x_{ij}^{[1]} = \begin{cases} 0 \text{ pour } i = 1, \ldots, 6 \\ 1 \text{ pour } i = 7, \ldots, 12 \end{cases}$$

L'effet aléatoire ξ_i est généré selon une distribution normale centrée de variance $\sigma^2 = 0.5$ et β_0 est fixé à 0.5.

Nous comparons le comportement du critère basé sur une approximation directe de la vraisemblance $\widehat{AIC}$ (respectivement $\widehat{BIC}$) aux critères AIC^S et AIC^G (respectivement BIC^S et BIC^G). Nous comparons les valeurs des critères pour les trois modèles $\mathcal{M}_0$, $\mathcal{M}_1$ et $\mathcal{M}_2$ définis dans la section 2.6.1. Le tableau (2.4) présente les résultats obtenus pour différentes valeurs de β_1.

2.7.3.3 Comparaison de modèles pour les effets aléatoires

Comme dans la section 2.6.2, nous nous intéressons ici à évaluer la capacité qu'ont les trois critères développés à détecter la présence ou l'absence d'un effet aléatoire. Nous considérons un plan d'expérience avec $I = 12$ individus et $J = 6$ observations par individu. Nous générons 200 jeux de données à partir du modèle exponentiel mixte lien log défini

TAB. 2.4 – Sélection de la structure d'effets fixes sur 200 jeux de données avec $I = 12$, $J = 6$, $\beta_0 = 0.5$ et $\sigma^2 = 0.5$

	Modèle	AIC^S	AIC^G	$\widehat{AIC}$	BIC^S	BIC^G	$\widehat{BIC}$
$\beta_1 = 0.1$	$\mathcal{M}_0$	151	149	149	181	181	178
	$\mathcal{M}_1^*$	31	34	34	15	18	19
	$\mathcal{M}_2$	18	17	17	4	1	3
$\beta_1 = 1$	$\mathcal{M}_0$	79	59	52	119	117	104
	$\mathcal{M}_1^*$	88	102	108	70	73	86
	$\mathcal{M}_2$	33	39	40	11	10	10
$\beta_1 = 2$	$\mathcal{M}_0$	5	1	0	17	7	2
	$\mathcal{M}_1^*$	138	138	141	157	170	175
	$\mathcal{M}_2$	57	61	59	26	23	23
$\beta_1 = 4$	$\mathcal{M}_0$	0	0	0	0	0	0
	$\mathcal{M}_1^*$	150	159	157	186	182	184
	$\mathcal{M}_2$	50	41	43	14	18	16

* Modèle simulé

TAB. 2.5 – Sélection de l'effet aléatoire sur 200 jeux de données avec $I = 12$, $J = 6$ et $\beta_0 = 1$

	Modèle	AIC^S	AIC^G	$\widehat{AIC}$	BIC^S	BIC^G	$\widehat{BIC}$
$\sigma^2 = 0$	$\mathcal{M}_0^*$	163	194	194	173	198	199
	$\mathcal{M}_1$	37	6	6	27	2	1
$\sigma^2 = 0.01$	$\mathcal{M}_0$	144	184	181	159	191	196
	$\mathcal{M}_1^*$	56	16	19	41	9	4
$\sigma^2 = 0.05$	$\mathcal{M}_0$	127	184	180	149	193	193
	$\mathcal{M}_1^*$	73	16	20	51	7	7
$\sigma^2 = 0.1$	$\mathcal{M}_0$	80	156	138	92	181	166
	$\mathcal{M}_1^*$	120	44	62	108	19	34
$\sigma^2 = 0.5$	$\mathcal{M}_0$	8	27	19	9	71	28
	$\mathcal{M}_1^*$	192	163	181	191	129	172
$\sigma^2 = 1$	$\mathcal{M}_0$	2	4	1	2	13	4
	$\mathcal{M}_1^*$	198	196	199	198	187	196
$\sigma^2 = 2$	$\mathcal{M}_0$	0	1	0	0	1	1
	$\mathcal{M}_1^*$	200	199	200	200	199	199

* Modèle simulé

par :

$$\mathcal{M}_1 \ : \ \begin{cases} Y_{ij}|\xi_i \sim \mathcal{E}xp(\mu_{\xi,ij}) \\ \xi_i \sim \mathcal{N}(0, \sigma^2) \end{cases} \qquad i = 1, \ldots, 12; \ j = 1, \ldots, 6$$

avec $\log(\mu_{\xi,ij}) = \beta_0 + \xi_i$ et $\beta_0 = 1$. Nous ajustons aux données simulées le modèle $\mathcal{M}_1$ et le modèle à effet fixe $\mathcal{M}_0$ pour lequel $\sigma^2 = 0$. Comme dans la section 2.6.2, nous inversons ensuite les rôles de $\mathcal{M}_0$ et $\mathcal{M}_1$. Le tableau (2.5) présente les résultats obtenus pour différentes valeurs de la variance σ^2.

À partir des tableaux (2.4) et (2.5), nous remarquons que les trois critères se comportent globalement de la même façon aussi bien en ce qui concerne la sélection de la structure d'effet fixe que celle de l'effet aléatoire. Notons néanmoins que le critère $\widehat{AIC}$ (resp. $\widehat{BIC}$) a un comportement qui s'apparente davantage au comportement du critère AIC^G (resp. BIC^G), et ceci est d'autant plus flagrant dans le tableau (2.5). De plus, on retrouve ici la

tendance au sur-ajustement et surtout à la sur-sélection du modèle à effet aléatoire des critères développés à partir de la linéarisation de Schall.
Dans le cadre de la sélection des effets fixes, ces résultats de simulation renforcent l'idée que les deux critères développés à partir des linéarisations de Schall et "Gumbel" sont satisfaisants. La comparaison avec le troisième critère construit sur une approximation directe de la vraisemblance nous permet d'affirmer que les linéarisations utilisées n'ont pas d'impact réellement négatif sur la qualité de la sélection. Ces critères présentent en plus l'avantage d'une mise en place pratique nettement plus simple et plus rapide. Ceci sera d'autant plus appréciable que la structure des effets aléatoires sera complexe.
Pour la sélection de l'effet aléatoire, dans le cas exponentiel lien log, les critères AIC^G et BIC^G semblent nettement à favoriser. Dans un cas plus général de GL2M, on peut imaginer une première sélection rapide à partir de AIC^S ou BIC^S, puis une étude plus raffinée basée sur $\widehat{AIC}$ ou $\widehat{BIC}$.

Chapitre 3

Modèles de mélange fini pour des données répétées de loi exponentielle

3.1 Introduction

Dans ce chapitre, nous nous intéressons à une nouvelle catégorie de modèles que sont les mélanges finis de modèles linéaires généralisés à effets aléatoires. Comme leur nom l'indique, nous restons dans le cadre d'une modélisation liant effets aléatoires et modèles linéaires généralisés. Cependant, nous introduisons ici de nouveaux paramètres afin de modéliser une hétérogénéité dans une population provenant de l'existence de différents sous-groupes non contrôlés. Notre attention se porte sur l'estimation des paramètres d'effets fixes et de variance des effets aléatoires ainsi que sur les proportions de mélange, le but de ce chapitre étant d'en proposer des méthodes d'estimation.

Comme nous venons de l'évoquer, les modèles de mélange permettent de modéliser une hétérogénéité dans une population provenant de l'existence de différents sous-groupes d'individus à comportement distinct. Les composants du mélange traduisent différents statuts possibles des individus en supposant que ce statut reste inchangé au cours de la répétition d'observations pour un même individu. Nous nous intéressons plus particulièrement dans ce chapitre à des données issues d'une loi exponentielle. La dépendance et l'extra-variabilité des données sont modélisées par l'introduction d'effets aléatoires. Nous présentons ainsi un modèle de mélange dans lequel chaque composant est défini par un modèle exponentiel mixte. Un premier exemple d'application est la modélisation

de durées de séjours hospitaliers répétés pour lesquels on cherche à mettre en évidence des classes de patients non contrôlées a priori. Un second exemple concerne l'analyse des temps d'élimination après plusieurs absorptions répétées d'un produit chez des patients. En fiabilité, une application est la modélisation de données de défaillance de matériels supposés sans vieillissement et réparables selon l'hypothèse "as good as new" (cf chapitre 2 section 2.2). On pourra alors supposer l'existence de différentes conditions d'exploitation.

Dans ce chapitre et pour atteindre notre objectif, l'algorithme EM va s'avérer être un outil essentiel. Ainsi, après des rappels généraux sur les modèles de mélange en section 3.2, puis un retour sur les mélanges de GLM en section 3.3, l'objet de la section 3.4 est de revenir sur l'utilisation de l'algorithme EM dans le cadre des mélanges de L2M. Après avoir décrit cet algorithme dans ce cas précis, nous considérons dans la section 3.5 la classe des mélanges de GL2M, et plus particulièrement le cas exponentiel. Nous verrons qu'il semble peu envisageable de prolonger, sans autre détour, la démarche pour un mélange de modèles exponentiels à effets aléatoires. Nous proposons alors, dans un premier temps, une méthode d'estimation des paramètres de ce modèle qui combine à la fois la linéarisation "Gumbel" du modèle exponentiel à effets aléatoires et l'utilisation de l'algorithme EM au sein des mélanges de L2M. Nous proposons ensuite une seconde méthode d'estimation des paramètres basée sur une étape de Metropolis-Hastings pour construire un algorithme de type MCEM. Ceci permet de contourner le problème rencontré par l'algorithme EM lié à la non accessibilité de la distribution marginale de chaque composant du mélange. Cette méthode, contrairement à la précédente, est applicable dans le cadre d'un mélange quelconque de modèles à effets aléatoires c'est-à-dire pour toute spécification de loi dans la famille exponentielle. Pour finir, des résultats de simulations seront présentés dans le but de comparer les différentes méthodes proposées. Des simulations intermédiaires effectuées dans le cas de mélanges de GLM et de L2M permettront également d'évaluer l'impact de l'introduction des effets aléatoires et de la généralisation de la loi sur la qualité des estimations dans les modèles de mélange.

3.2 Rappels généraux sur les modèles de mélange

3.2.1 Définition d'un mélange de lois

Une des tâches courantes de la statistique consiste à déduire d'un échantillon observé une distribution de probabilité sur la population-mère dont l'échantillon est tiré. Selon la nature physique du phénomène observé ou selon l'échantillon lui-même, on peut être amené à estimer une distribution de type connu. Mais bien souvent, l'existence de plusieurs sources de variation à l'origine du phénomène ou bien l'observation d'un histogramme multimodal indiquent que la distribution de la population-mère peut être un mélange de lois de probabilité simples et régulières.

Une des premières analyses majeures impliquant l'utilisation des modèles de mélange est due au biométricien Karl Pearson. Dans son article [44], Pearson crée un mélange de deux densités de probabilité gaussiennes d'espérances μ_1 et μ_2 et de variances σ_1^2 et σ_2^2 dans les proportions p_1 et p_2 pour un jeu de données concernant des mesures de proportions de la taille du crâne par rapport à la longueur du corps de 1000 crabes. L'asymétrie dans l'histogramme des données laissait supposer l'existence de deux sous-espèces. L'approche de Pearson pour résoudre ce problème fut le modèle de mélange. Il utilisa alors la méthode des moments pour estimer les paramètres de ce mélange de gaussiennes en résolvant un polynôme de degré neuf. Depuis cette première tentative d'analyse d'un modèle de mélange, l'étude des mélanges de lois est devenue un domaine à part entière de la statistique moderne. De nombreux ouvrages de référence existent sur le sujet. Le plus récent est celui de McLachlan et Peel [40] qui fait une présentation complète des différentes approches développées à ce jour.

La formalisation d'un mélange de densités est la suivante. Considérons $Y_1, \ldots, Y_n$ un échantillon aléatoire de taille n où Y_i est une variable aléatoire de densité $f(y_i)$. Notons $Y = (Y_1, \ldots, Y_n)'$ le n-uplet représentant l'échantillon complet. Le modèle de mélange suppose que la densité $f(y_i)$ de Y_i peut s'écrire sous la forme :

$$f(y_i) = \sum_{k=1}^{K} p_k f_k(y_i|\theta_k), \tag{3.1}$$

avec

$$\forall k = 1, \ldots, K,\ 0 < p_k < 1 \text{ et } \sum_{k=1}^{K} p_k = 1,$$

où $f_k(.|\theta_k)$ est une densité de probabilité de paramètre θ_k de $\mathbb{R}^s$ (s étant la dimension du paramètre). Le nombre de composants K est fixé. Les quantités $p_1, \ldots, p_K$ sont les proportions du mélange avec p_k la probabilité qu'une variable quelconque de l'échantillon suive la loi de densité $f_k(.|\theta_k)$.

Certains auteurs ont tenté de définir des critères pour obtenir le nombre optimal de composants. Citons par exemple Celeux et Soromenho [14] qui proposent un critère d'entropie ou encore Biernacki, Celeux et Govaert [5] qui tentent d'améliorer ce même critère.

Les modèles de mélange ont ainsi fourni une approche mathématique à la modélisation statistique d'une large variété de phénomènes aléatoires. Grâce à leur flexibilité, les applications des modèles de mélange se sont considérablement développées dans des domaines très variés tels que la biologie, la génétique, l'ingénierie ou encore l'économie. Nous précisons maintenant quelques aspects théoriques des modèles de mélanges de lois et envisageons la question de l'estimation des paramètres.

3.2.2 Inférence dans les mélanges de lois

La formulation paramétrique des mélanges de lois a permis l'émergence de nombreuses techniques d'inférence ayant pour but essentiel l'estimation des paramètres. La première méthode utilisée pour ce faire, basée sur l'étude des moments, connaît un regain d'intérêt récent avec des auteurs comme Craigmile et Titterington [16] par exemple. Il existe aussi des méthodes graphiques, mais les plus utilisées sont les méthodes bayésiennes ainsi que celles utilisant le maximum de vraisemblance. C'est cette dernière approche par maximum de vraisemblance que nous présentons plus particulièrement ici.

Les approches relevant des techniques d'estimation par maximum de vraisemblance visent à estimer les paramètres $(\theta_k, p_k)_{k=1,\ldots,K}$ en résolvant itérativement les équations de vraisemblance pour un échantillon $(y_1, \ldots, y_n)$ donné. Le logarithme de la vraisemblance est donné par :

$$L(\theta_1, \ldots, \theta_K, p_1, \ldots, p_K | y_1, \ldots, y_n) = \sum_{i=1}^{n} \ln \big(\sum_{k=1}^{K} p_k f_k(y_i|\theta_k) \big) \tag{3.2}$$

Pour plus de clarté, nous noterons $L(\phi|y)$ la log-vraisemblance précédente avec $\phi = (\theta_1, \ldots, \theta_K, p_1, \ldots, p_K)$. Lorsque nous n'avons pas l'expression analytique de la log-vraisem-

blance, les algorithmes les plus efficaces pour obtenir malgré tout les estimations du maximum de vraisemblance sont les algorithmes de type EM (cf Dempster, Laird et Rubin [17], Redner et Walker [48]). Nous présentons ici l'algorithme EM et un certain nombre de ses variantes. Nous ne reviendrons pas sur les propriétés des estimateurs du maximum de vraisemblance qui sont maintenant bien connues et qui sont détaillées par exemple dans McLachlan et Peel [40].

3.2.2.1 Une structure de données incomplètes

Dans le cadre du modèle de mélange, les données sont considérées comme incomplètes car nous ne connaissons pas les composants auxquels appartiennent les observations. Chaque variable aléatoire Y_i qui compose l'échantillon Y provient d'un des composants du mélange. On associe à chaque Y_i une variable aléatoire discrète Z_i prenant des valeurs dans $\{1, \ldots, K\}$ avec les probabilités $p_1, \ldots, p_K$ respectivement :

$$\forall k = 1, \ldots, K \quad P(Z_i = k) = p_k.$$

On suppose que la densité conditionnelle de Y_i sachant $Z_i = k$ est $f_k(y_i|\theta_k)$. La densité de Y_i est donc $f(y_i)$. Dans ce contexte, la variable aléatoire Z_i peut être vu comme l'étiquette du composant (on parle de *label component*) de Y_i car elle définit le composant auquel il appartient. Par la suite, on définit un vecteur d'étiquettes $Z = (Z_1, \ldots, Z_n)$ de composantes K-dimensionnelles où, pour $k = 1, \ldots, K$, le $k^{ème}$ élément de Z_i est défini par :

$$Z_{ik} = (Z_i)_k = \begin{cases} 1 & \text{si l'individu } i \text{ est issu du } k^{ème} \text{ composant} \\ 0 & \text{sinon} \end{cases}$$

Ainsi, Z_i est distribuée selon une loi multinomiale consistant en un tirage parmi K catégories de probabilités respectives $p_1, \ldots, p_K$:

$$P(Z_i = z_i) = p_1^{z_{i1}} p_2^{z_{i2}} \ldots p_K^{z_{iK}}$$

On écrit classiquement : $Z_i \sim \mathcal{M}ult_K(1, p)$ avec $p = (p_1, \ldots, p_K)$.

Le vecteur des données complètes est donc défini par $(y, z) = \big((y_i, z_i)\big)_{1 \leq i \leq n}$. Les vecteurs $z_1, \ldots, z_n$ sont des réalisations des variables aléatoires $Z_1, \ldots, Z_n$ supposées indépendantes et identiquement distribuées. Notons que z définit une partition $P = (P_1, \ldots, P_K)$ des données observées y avec $P_k = \{y_i / z_{ik} = 1\}$.

La vraisemblance associée aux données complètes s'écrit alors :

$$l(\phi|y,z) = \prod_{i=1}^{n}\prod_{k=1}^{K} p_k^{z_{ik}}\, f_k(y_i|\theta_k)^{z_{ik}}$$

où, pour $i = 1,\ldots,n$, et $k = 1,\ldots,K$, z_{ik} est une réalisation non observée de la variable indicatrice Z_{ik}. Cela conduit finalement à l'écriture de la log-vraisemblance des données complètes :

$$L(\phi|y,z) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\ln p_k + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\ln f_k(y_i|\theta_k). \tag{3.3}$$

3.2.2.2 L'algorithme EM

a) Présentation

L'algorithme EM trouve ici son intérêt puisqu'il tire partie de l'information manquante en considérant l'espérance conditionnelle de la log-vraisemblance de l'échantillon complet $L(\phi|y,z)$ sachant les données observées. À l'itération $[t]$, l'étape E consiste à calculer l'espérance de la log-vraisemblance associée aux données complètes conditionnellement aux données observées y et aux valeurs courantes des paramètres $\phi^{[t]} = (\theta^{[t]}, p^{[t]})$:

$$Q(\phi|\phi^{[t]}) = \sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i)\ln p_k + \sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i)\ln f_k(y_i|\theta_k) \tag{3.4}$$

avec

$$\begin{aligned}
t_k^{[t]}(y_i) &= E(Z_{ik}|y,\phi^{[t]}) \\
&= P(Z_{ik}=1|y,\phi^{[t]}) \\
&= P(Z_{ik}=1|y_i,\phi^{[t]}) \text{ car } \forall i \neq i' Z_i \text{ est indépendant de } Y_{i'} \\
&= \frac{p_k^{[t]}\, f(y_i|z_{ik}=1,\phi^{[t]})}{f(y_i|\phi^{[t]})} \\
&= \frac{p_k^{[t]}\, f_k(y_i|\theta_k^{[t]})}{\sum_{l=1}^{K} p_l^{[t]}\, f_l(y_i|\theta_l^{[t]})}
\end{aligned}$$

Notons que la probabilité $t_k^{[t]}(y_i)$ représente la probabilité a posteriori que l'individu i appartienne au $k^{ème}$ composant du mélange à l'itération $[t]$. L'étape M consiste ensuite à maximiser $Q(\phi|\phi^{[t]})$ par rapport à ϕ.

Ainsi, à partir d'une solution initiale $\phi^{[0]} = (p_k^{[0]}, \theta_k^{[0]})_{k=1,\ldots,K}$, l'algorithme EM peut se résumer de la façon suivante, à l'itération $[t]$:

Étape E :

Pour $i = 1, \ldots, n$, et $k = 1, \ldots, K$, calcul des probabilités a posteriori $t_k^{[t]}(y_i)$.

Étape M :

Pour $k = 1, \ldots, K$, calcul des valeurs maximisant $Q(\phi|\phi^{[t]})$ pour les proportions du mélange

$$p_k^{[t+1]} = \frac{1}{n} \sum_{i=1}^{n} t_k^{[t]}(y_i),$$

et résolution des équations pour les paramètres θ_k :

$$\forall k = 1, \ldots, K;\ j = 1, \ldots, s \quad \sum_{i=1}^{n} t_k^{[t]}(y_i) \frac{\partial \ln f_k(y_i|\theta_k^{[t+1]})}{\partial \theta_{kj}} = 0.$$

où $\theta_k = (\theta_{kj}; j = 1, \ldots, s)$.

Par exemple, dans le cas d'un mélange gaussien, nous avons $\theta_k = (\mu_k, \sigma_k^2)$ où μ_k et σ_k^2 sont respectivement l'espérance et la variance associées au $k^{ème}$ composant du mélange. Les équations précédentes donnent alors :

$$\mu_k^{[t+1]} = \frac{1}{\sum_{i=1}^{n} t_k^{[t]}(y_i)} \sum_{i=1}^{n} t_k^{[t]}(y_i)\ y_i,$$

$$\sigma_k^{2[t+1]} = \frac{1}{\sum_{i=1}^{n} t_k^{[t]}(y_i)} \sum_{i=1}^{n} t_k^{[t]}(y_i)(y_i - \mu_k^{[t+1]})^2.$$

Ces équations montrent en particulier que les probabilités a posteriori peuvent être interprétées comme des poids.

b) Comportement théorique

La principale caractéristique de l'algorithme EM est de faire croître la vraisemblance $l(\phi|y)$ à chaque itération. Plus précisément, on a le théorème suivant (Dempster, Laird et Rubin [17]) :

Théorème

Toute suite $(\phi^{[t]})$ engendrée par EM vérifie $l(\phi^{[t+1]}|y) \geq l(\phi^{[t]}|y)$ avec l'égalité si et seulement si $Q(\phi^{[t+1]}|\phi^{[t]}) = Q(\phi^{[t]}|\phi^{[t]})$.

Des résultats sur la convergence de l'algorithme EM peuvent être trouvés dans Wu [68]. Dans le cadre de mélanges de distributions appartenant à une famille exponentielle, nous voulons indiquer un résultat asymptotique et local obtenu par Redner et Walker [48] :

Théorème

Si les proportions du mélange sont strictement positives et la matrice d'information de Fisher associée au vrai ϕ est définie positive, alors l'unique solution consistante des équations de vraisemblance existe presque sûrement et, pour n assez grand, $\phi^{[t]}$ converge linéairement vers cette solution pour peu que la valeur initiale en soit assez proche.

c) Comportement pratique

Le nombre très important de publications, dans des domaines très variés, sur l'algorithme EM témoigne de sa souplesse et de son efficacité. Néanmoins, bon nombre d'articles analysent des défauts bien réels de l'algorithme EM que l'on peut résumer en trois points :

- forte dépendance parfois par rapport à sa position initiale,
- convergence vers des solutions hautement sous-optimales,
- situations de convergence très lentes.

Dans le cas des mélanges qui, incontestablement, constitue un terrain d'applications privilégié pour EM, les auteurs s'accordent pour caractériser les performances de l'algorithme EM de la façon suivante (cf. par exemple Titterington, Smith et Makov [62]) : l'algorithme EM donne des résultats parfaitement satisfaisants et dans des temps compétitifs

- pour des mélanges dont le nombre de composants est connu,
- si les composants sont bien séparés,
- si les proportions ne sont pas trop déséquilibrées,
- et si la position initiale des paramètres n'est pas trop loin des vraies valeurs.

S'il est initialisé avec un nombre erroné de composants, l'algorithme EM ne permet pas de déceler cette erreur de diagnostic. Si l'une des trois autres conditions n'est pas vérifiée, la solution de l'algorithme EM peut dépendre fortement de sa position initiale et peut même converger vers un col de la fonction de vraisemblance ou passer un temps très long près d'un tel col. Pratiquement, ce sont surtout les deux conditions, connaissance du nombre de composants et bonnes valeurs initiales des paramètres, qui s'avèrent cruciales pour assurer de bonnes performances de l'algorithme EM. On trouvera dans Celeux et Diebolt [11] des simulations qui illustrent ces dires.

Pour s'affranchir du caractère local du maximum atteint, une façon de procéder en pratique consiste à faire tourner l'algorithme EM un certain nombre de fois à partir de valeurs initiales différentes de manière à avoir de plus grandes chances d'atteindre

le maximum global de vraisemblance. Plusieurs versions stochastiques de l'algorithme EM visant à répondre à ces limitations ont également été développées. Nous présentons maintenant un certain nombre d'entre elles.

3.2.2.3 L'algorithme SEM

L'algorithme SEM développé par Celeux et Diebolt [10] utilise de manière complémentaire la construction de partitions et les étapes de l'algorithme EM. Il s'agit en fait d'un algorithme EM auquel est ajouté une étape d'apprentissage probabiliste. D'où son nom, algorithme SEM : Stochastique, Estimation, Maximisation.

Initialisation

Au départ, on fixe le paramètre K majorant supposé du nombre de composants du mélange et un seuil $c(n)$ compris entre 0 et 1. En chaque point y_i, $i = 1, \ldots, n$, on choisit (en général au hasard) les probabilités initiales d'appartenance à l'un des composants :

$$t_k^{[0]}(y_i), k = 1, \ldots, K \text{ avec } 0 < t_k^{[0]}(y_i) < 1 \text{ et } \sum_{k=1}^{K} t_k^{[0]}(y_i) = 1.$$

Itération [t] (t $\geq$ 1)

Étape S (stochastique)

On tire en chaque point y_i la variable aléatoire multinomiale $e^{[t]}(y_i) = (e_k^{[t]}(y_i); k = 1, \ldots, K)$ d'ordre un et de paramètres $(t_k^{[t]}(y_i); k = 1, \ldots, K)$. Les réalisations $e^{[t]}(y_i)$ définissent une partition $P^{[t]} = (P_1^{[t]}, \ldots, P_K^{[t]})$ de l'échantillon avec :

$$P_k^{[t]} = \{y_i / e_k^{[t]}(y_i) = 1\}.$$

Si pour un certain k, card$(P_k^{[t]})$ est plus petit que $nc(n)$, l'algorithme est ré-initialisé.

Étape M (maximisation)

On calcule les estimations du maximum de vraisemblance $\phi_k^{[t+1]} = (p_k^{[t+1]}, \theta_k^{[t+1]})$ des paramètres du mélange sur la base des sous-échantillons $(P_k^{[t]}; k = 1, \ldots, K)$.

On a :

$$p_k^{[t+1]} = \frac{1}{n} \sum_{i=1}^{n} e_k^{[t]}(y_i).$$

L'estimation des $\theta_k^{[t+1]}$ dépend de la famille paramétrée, posée a priori, de distributions des composants du mélange.

Étape E (estimation)
À partir des $\phi_k^{[t+1]} = (p_k^{[t+1]}, \theta_k^{[t+1]})$, on calcule, pour $i = 1, \ldots, n$ et $k = 1, \ldots, K$:

$$t_k^{[t+1]}(y_i) = \frac{p_k^{[t+1]} f_k(y_i|\theta_k^{[t+1]})}{\sum_{l=1}^{K} p_l^{[t+1]} f_l(y_i|\theta_l^{[t+1]})}.$$

À la stabilité de l'algorithme, on obtient une classe de partitions statistiquement admissibles pour les estimations des paramètres du mélange. Ces estimations sont précises et asymptotiquement sans biais (cf. Celeux et Diebolt [10]). Le type de convergence obtenue est une convergence en loi correspondant à la stationnarité de la suite des estimés $\phi^{[t]} = (p^{[t]}, \theta^{[t]})$. Par ailleurs, les perturbations introduites à chaque itération par les tirages aléatoires empêchent la convergence vers un maximum local instable de la vraisemblance comme cela peut être le cas pour l'algorithme EM. Cet algorithme fournit le nombre exact de composants pourvu qu'il soit initialisé avec un majorant de ce nombre. Il converge notablement plus rapidement que l'algorithme EM quelle que soit la configuration initiale, les tirages aléatoires l'empêchant de stationner trop longtemps loin de la solution limite. Cependant, notons que pour de petits échantillons, l'algorithme SEM risque de sous-estimer le nombre de composants, les aléas introduits prenant trop d'importance.

3.2.2.4 L'algorithme SAEM

L'algorithme SAEM (Stochastic Approximation EM) (Celeux et Diebolt [12]) est une modification de l'algorithme SEM telle que la convergence en loi peut être remplacée par la convergence presque sûre. Le comportement erratique possible de SEM pour de petits échantillons peut être atténué sans pour autant sacrifier la nature stochastique de l'algorithme. Cela est réalisé en utilisant une suite de réels positifs (γ_t) décroissante vers 0 (avec $\gamma_0 = 1$). Plus précisément, si $\phi^{[t]}$ est le paramètre courant estimé par SAEM, l'approximation $\phi^{[t+1]}$ de ϕ est

$$\phi^{[t+1]} = (1 - \gamma_{t+1})\phi_{EM}^{[t+1]} + \gamma_{t+1}\phi_{SEM}^{[t+1]},$$

avec $\phi_{EM}^{[t+1]}$ l'approximation de ϕ par EM et $\phi_{SEM}^{[t+1]}$ l'approximation de ϕ par SEM.

L'algorithme SAEM progresse ainsi d'un pur SEM au début vers un pur EM à la fin. Notons que le choix du taux de convergence vers 0 de γ_t est important. Un faible taux

de convergence est nécessaire pour de bons résultats. D'un point de vue pratique, il est important que γ_t reste près de $\gamma_0 = 1$ durant les premières itérations pour éviter les valeurs stationnaires de $L(\phi)$.

3.2.2.5 L'algorithme MCEM

L'algorithme MCEM est un algorithme développé par Wei et Tanner [66] introduisant une étape de Monte Carlo à l'étape E. Il s'agit de remplacer le calcul de $Q(\phi|\phi^{[t]})$ par celui d'une version empirique $Q_{t+1}(\phi|\phi^{[t]})$ basée sur M ($M \gg 1$) réalisations des données manquantes z à partir de $f(z|y, \phi^{[t]})$ la densité conditionnelle de z sachant y et la valeur courante des paramètres $\phi^{[t]}$. Formellement, $f(z|y, \phi) = \dfrac{f(y, z|\phi)}{f(y|\phi)}$.

Plus précisément, à l'itération $[t]$, l'algorithme se présente de la façon suivante :

1. Générer un échantillon indépendant et identiquement distribué $z^{[t]}(1), \ldots, z^{[t]}(M)$ à partir de $f(z|y, \phi^{[t]})$
2. Calculer l'approximation courante de $Q(\phi|\phi^{[t]})$ par
$$Q_{t+1}(\phi|\phi^{[t]}) = \frac{1}{M} \sum_{m=1}^{M} \ln f(y, z^{[t]}(m)|\phi).$$
3. L'étape M est donnée par : $\phi^{[t+1]} = \text{argmax}_{\phi} Q_{t+1}(\phi|\phi^{[t]})$

Si $M = 1$, MCEM est réduit à SEM. Si M est très grand, MCEM marche approximativement comme EM. Wei et Tanner [66] ont motivé l'introduction de l'algorithme MCEM comme une alternative qui remplace le calcul analytique de l'intégrale de l'étape E par le calcul numérique d'une approximation de Monte Carlo de cette intégrale. En pratique, Wei et Tanner recommandent de démarrer avec une petite valeur de M et de l'augmenter au fur et à mesure que $\phi^{[t]}$ se rapproche du vrai maximiseur de $L(\phi)$. Plus précisément, si nous sélectionnons une suite (M_t) d'entiers tels que $M_0 = 1$ et M_t croit vers l'infini quand t tend vers l'infini, nous allons d'un pur SEM ($M_0 = 1$) vers un pur EM ($M = \infty$) quand t tend vers l'infini.

3.3 Mélange de GLM

Nous nous intéressons maintenant plus particulièrement aux mélanges de modèles linéaires généralisés tels que nous les avons décrits au chapitre 1 c'est-à-dire aux mélanges

TAB. 3.1 – Nombre d'avis de décès dans The Times entre 1910 et 1912

nombre d'avis de décès observés	0	1	2	3	4	5	6	7	8	9
fréquence	162	267	271	185	111	61	27	8	3	1

de densités issues de la même famille exponentielle. La table (3.1) présente un exemple souvent utilisé dans la littérature (Hasselblad [27], Titterington, Smith et Makov [62]). Il s'agit du nombre d'avis de décès des femmes âgées de 80 ans et plus, apparaissant dans le journal londonien The Times, chaque jour, pendant trois années consécutives (1910, 1911 et 1912). Pour modéliser ces données, Hasselblad [27] suggère d'utiliser un mélange à deux composants de distributions de Poisson, les deux composants correspondant alors à deux niveaux différents de mortalité en été et en hiver.

3.3.1 Le modèle

Considérons $y = (y_1, \ldots, y_n)$ le vecteur de taille n des observations, réalisation du vecteur aléatoire $Y = (Y_1, \ldots, Y_n)$ et supposons l'existence d'une structure de classes cachées $\mathcal{C}_k, k = 1, \ldots, K$.

Connaissant le composant $\mathcal{C}_k$ dont est issue l'unité statistique i, l'observation y_i est supposée modélisée par le modèle linéaire généralisé suivant :

$$f_k(y_i|\theta_{ik}, \lambda_k) = \exp\Big\{\frac{y_i\theta_{ik} - b(\theta_{ik})}{a_i(\lambda_k)} + c(y_i, \lambda_k)\Big\}$$

où θ_{ik} est un paramètre canonique et λ_k est un paramètre de dispersion supposé être constant pour toutes les observations issus du composant $\mathcal{C}_k$.

Connaissant le composant $\mathcal{C}_k$ dont est issue y_i, l'espérance et la variance de la variable associée s'écrit :

$$\begin{aligned} E(Y_i) &= \mu_{ik} = b'(\theta_{ik}) \\ \mathrm{Var}(Y_i) &= a_i(\lambda_k)b''(\theta_{ik}) \\ &= a_i(\lambda_k)v(\mu_{ik}). \end{aligned}$$

On définit également le prédicteur linéaire η_{ik} et une fonction de lien g tel que :

$$\begin{aligned} \eta_{ik} &= g(\mu_{ik}) \\ &= X_i'\beta_k \end{aligned}$$

où $X_i = (X_{i1}, \ldots, X_{ip})'$ est le vecteur $p \times 1$ de covariables associées à l'unité i et $\beta_k = (\beta_{k1}, \ldots, \beta_{kp})'$ le vecteur $p \times 1$ des paramètres associés au composant $\mathcal{C}_k$.

Notons que le lien entre β_k et θ_{ik} est décrit par la relation :

$$X_i'\beta_k = g(b'(\theta_{ik})).$$

La densité d'une observation y_i est donc un mélange défini par :

$$f(y_i|\phi) = \sum_{k=1}^{K} p_k f_k(y_i|\beta_k, \lambda_k)$$

où $\forall k = 1, \ldots, K\ 0 < p_k < 1$ et $\sum_{k=1}^{K} p_k = 1$,
et $\phi = (p', \beta', \lambda')'$ avec $p = (p_1, \ldots, p_K)'$, $\beta = (\beta_1', \ldots, \beta_K')'$ et $\lambda = (\lambda_1, \ldots, \lambda_K)'$.

3.3.2 Estimation des paramètres par l'algorithme EM

Comme précédemment, on introduit, pour $i = 1, \ldots, n$, le vecteur de données manquantes $z_i = (z_{i1}, \ldots, z_{iK})$ caractérisant l'appartenance de l'unité statistique i à l'un des composants du mélange : $z_{ik} = 1$ si l'unité i appartient au composant $\mathcal{C}_k$ et 0 sinon. Dans ce contexte, la log-vraisemblance associée aux données complètes (y, z) s'écrit de la manière suivante :

$$L(\phi|y, z) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \Big[\ln p_k + \ln f_k(y_i|\beta_k, \lambda_k) \Big]$$

3.3.2.1 Étape E

On s'intéresse à l'espérance de la log-vraisemblance des données complètes $L(\phi|y, z)$ par rapport à la distribution des données manquantes sachant les données observées y et la valeur courante des paramètres $\phi^{[t]}$:

$$\begin{aligned} Q(\phi|\phi^{[t]}) &= E\Big[L(\phi|y, z)|y, \phi^{[t]} \Big] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} t_k^{[t]}(y_i) \Big[\ln p_k + \ln f_k(y_i|\beta_k, \lambda_k) \Big] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} t_k^{[t]}(y_i)\ \ln p_k + \sum_{i=1}^{n} \sum_{k=1}^{K} t_k^{[t]}(y_i)\ \ln f_k(y_i|\beta_k, \lambda_k) \end{aligned}$$

où, comme précédemment,

$$t_k^{[t]}(y_i) = E(Z_{ik}|y, \phi^{[t]}) = \frac{p_k^{[t]}\ f_k(y_i|\beta_k^{[t]}, \lambda_k^{[t]})}{\sum_{l=1}^{K} p_l^{[t]}\ f_l(y_i|\beta_l^{[t]}, \lambda_l^{[t]})}$$

3.3.2.2 Étape M

On s'intéresse maintenant à la maximisation de $Q(\phi|\phi^{[t]})$ par rapport à ϕ. Le maximum de $Q(\phi|\phi^{[t]})$ par rapport à p doit respecter la contrainte $\sum_{k=1}^{K} p_k = 1$ que l'on introduit par un multiplicateur de Lagrange ψ. Il s'obtient donc par maximisation de la fonction

$$\sum_{i=1}^{n}\sum_{k=1}^{K} t_k^{[t]}(y_i)\ \ln p_k - \psi\big(\sum_{k=1}^{K} p_k - 1\big) \tag{3.5}$$

Par dérivation de (3.5) puis annulation de la dérivée, on obtient :

$$p_k^{[t+1]} = \frac{1}{n}\sum_{i=1}^{n} t_k^{[t]}(y_i).$$

En supposant que deux composants n'ont pas d'éléments en commun, maximiser $Q(\phi|\phi^{[t]})$ par rapport à β et λ est équivalent à maximiser indépendamment chacune des K expressions :

$$L_k^* = \sum_{i=1}^{n} t_k^{[t]}(y_i)\ \ln f_k(y_i|\beta_k, \lambda_k)$$

par rapport à β_k et λ_k.
La maximisation de L_k^* est équivalente à un problème de maximisation dans un GLM, excepté que chaque observation y_i contribue ici à la log-vraisemblance du composant $\mathcal{C}_k$ avec un poids connu $t_k^{[t]}(y_i)$ obtenu à l'étape E.

Pour $k = 1, \ldots, K$, pour obtenir les équations du maximum de vraisemblance pour l'estimation de β_k, nous dérivons la log-vraisemblance L_k^* par rapport à ses différentes composantes :
$\forall k \in \{1, \ldots, K\}, \forall j \in \{1, \ldots, p\}$,

$$\frac{\partial L_k^*}{\partial \beta_{kj}} \quad = \quad \sum_{i=1}^{n} t_k^{[t]}(y_i)\ \frac{\partial \ln f_k(y_i|\beta_k, \lambda_k)}{\partial \beta_{kj}}.$$

On a :
$\forall i \in \{1,\dots,n\}, \forall k \in \{1,\dots,K\}, \forall j \in \{1,\dots,p\}$,

$$\begin{aligned}\frac{\partial \ln f_k(y_i|\beta_k,\lambda_k)}{\partial \beta_{kj}} &= \frac{\partial \eta_{ik}}{\partial \beta_{kj}}\frac{d\mu_{ik}}{d\eta_{ik}}\frac{d\theta_{ik}}{d\mu_{ik}}\frac{\partial \ln f_k(y_i|\beta_k,\lambda_k)}{\partial \theta_{ik}} \\ &= X_{ij}\frac{1}{g'(\mu_{ik})}\frac{1}{b''(\theta_{ik})}\frac{(y_i - b'(\theta_{ik}))}{a_i(\lambda_k)} \\ &= X_{ij}\frac{1}{g'(\mu_{ik})v(\mu_{ik})}\frac{(y_i-\mu_{ik})}{a_i(\lambda_k)}\end{aligned}$$

D'où

$$\begin{aligned}\frac{\partial L_k^*}{\partial \beta_{kj}} &= \sum_{i=1}^{n} t_k^{[t]}(y_i)\, X_{ij}\frac{1}{g'(\mu_{ik})v(\mu_{ik})}\frac{(y_i-\mu_{ik})}{a_i(\lambda_k)} \\ &= \sum_{i=1}^{n} t_k^{[t]}(y_i)\, X_{ij}\frac{1}{a_i(\lambda_k)g'(\mu_{ik})^2 v(\mu_{ik})}g'(\mu_{ik})(y_i-\mu_{ik}) \\ &= \sum_{i=1}^{n} X_{ij}\frac{1}{\frac{a_i(\lambda_k)}{t_k^{[t]}(y_i)}g'(\mu_{ik})^2 v(\mu_{ik})}g'(\mu_{ik})(y_i-\mu_{ik})\end{aligned}$$

Ainsi, en considérant les matrices diagonales définies par :

$$W_{\beta_k} = \mathrm{diag}\Big\{\frac{a_i(\lambda_k)}{t_k^{[t]}(y_i)}v(\mu_{ik})g'(\mu_{ik})^2\Big\}_{i=1,\dots,n}$$

et

$$\frac{d\eta_k}{d\mu_k} = \mathrm{diag}\Big\{\frac{d\eta_{ik}}{d\mu_{ik}}\Big\}_{i=1,\dots,n} = \mathrm{diag}\{g'(\mu_{ik})\}_{i=1,\dots,n},$$

les équations du maximum de vraisemblance pour β_k s'écrivent :

$$X' W_{\beta_k}^{-1}\frac{d\eta_k}{d\mu_k}(y-\mu_k) = 0 \tag{3.6}$$

En définissant la fonction a_i par $a_i(\lambda_k) = \dfrac{\lambda_k}{w_i}$ où w_i est un poids connu associé à y_i, la matrice W_{β_k} s'écrit :

$$W_{\beta_k} = \mathrm{diag}\Big\{\frac{\lambda_k}{w_i\, t_k^{[t]}(y_i)}v(\mu_{ik})g'(\mu_{ik})^2\Big\}_{i=1,\dots,n}.$$

Pour chaque composant $\mathcal{C}_k$, L_k^* peut être maximisée par une procédure itérative classique pour l'estimation du maximun de vraisemblance dans un GLM, en associant à chaque observation y_i le poids additionnel $t_k(y_i)$.

3.3.3 Réécriture en terme d'équations normales

Pour $k = 1, \ldots, K$, considérons la dérivée de L_k^* par rapport à β_k :

$$\frac{\partial L_k^*}{\partial \beta_k} = X' W_{\beta_k}^{-1} \frac{d\eta_k}{d\mu_k}(y - \mu_k)$$

où W_{β_k}, η_k, μ_k ont été définis précédemment.
Rappelons que les équations (3.6) du maximum de vraisemblance de l'étape M pour l'estimation de β_k s'écrivent :

$$\frac{\partial L_k^*}{\partial \beta_k} = 0 \quad \Leftrightarrow \quad X' W_{\beta_k}^{-1} \frac{d\eta_k}{d\mu_k}(y - \mu_k) = 0$$

Ce système d'équations n'étant pas linéaire en β_k, une résolution itérative est mise en place. L'algorithme usuel est l'algorithme des scores de Fisher dont les équations sont ici données par :

$$\begin{aligned}
\beta_k^{[t+1]} &= \beta_k^{[t]} - \left(E\Big[\frac{\partial^2 L_k^*}{\partial \beta_k \partial \beta_k'} \Big]^{[t]} \right)^{-1} \frac{\partial L_k^{*[t]}}{\partial \beta_k} \\
&= \beta_k^{[t]} + (X' W_{\beta_k^{[t]}}^{-1} X)^{-1} X' W_{\beta_k^{[t]}}^{-1} \frac{d\eta_k^{[t]}}{d\mu_k}(y - \mu_k^{[t]}) \\
&= (X' W_{\beta_k^{[t]}}^{-1} X)^{-1} X' W_{\beta_k^{[t]}}^{-1} z_k^{[t]}
\end{aligned}$$

où $z_k^{[t]} = X\beta_k^{[t]} + \dfrac{d\eta_k^{[t]}}{d\mu_k}(y - \mu_k^{[t]})$.

En introduisant le vecteur dépendant défini par :

$$z_k = \eta_k + \frac{d\eta_k}{d\mu_k}(y - \mu_k) = X\beta_k + \frac{d\eta_k}{d\mu_k}(y - \mu_k),$$

les équations (3.6) s'écrivent :

$$X' W_{\beta_k}^{-1}(z_k - X\beta_k) = 0. \tag{3.7}$$

Ainsi, le même algorithme est décrit en résolvant itérativement les équations (3.7) comme des équations normales. À chaque itération, la valeur courante de β_k permet le calcul de la matrice des poids W_{β_k} et du vecteur dépendant z_k. Cela permet ensuite par résolution de ce système linéarisé d'obtenir une nouvelle valeur de β_k.
Cette réécriture (3.7) permet une interprétation de type linéaire. À β_k fixé, en considérant z_k comme un nouveau vecteur de données et W_{β_k} comme une matrice de poids fixés, on

TAB. 3.2 – Résultats d'estimation des paramètres d'un mélange à 2 composants de GLM par l'algorithme EM sur 100 simulations : cas Poisson - lien log

		Valeurs estimées	
Valeurs simulées		moyenne	écart-type
$\mathcal{C}_1$	$p_1 = 0.7$	0.7132	0.0459
	$\beta_{11} = 0$	0.1256	0.2404
	$\beta_{12} = -1$	-1.2673	0.5496
$\mathcal{C}_2$	$p_2 = 0.3$	0.2868	0.0459
	$\beta_{21} = 1$	1.0344	0.2160
	$\beta_{22} = 0.5$	0.4239	0.3344

reconnait dans le système (3.7) les équations classiques des moindres carrés généralisés associées au modèle :

$$Z_k = X\beta_k + e_k$$

où $E(e_k) = 0$
et $\text{Var}(e_k) = W_{\beta_k} = \text{diag}\{\frac{\lambda_k}{w_i\, t_k(y_i)} v(\mu_{ik}) g'(\mu_{ik})^2\}_{i=1,\ldots,n}$. Les poids initiaux w_i associés aux observations y_i, $i = 1, \ldots, n$, sont ici pondérés par les probabilités a posteriori $t_k(y_i)$.

3.3.4 Des résultats de simulation

Dans le but d'étudier le comportement de l'algorithme EM dans le cadre d'un mélange de GLM, nous procédons à quelques simulations dans le cas des lois Poisson et exponentielle. La taille de l'échantillon est fixé à 200. Nous considérons un mélange à 2 classes avec les proportions du mélange fixées à $p_1 = 0.7$ et $p_2 = 0.3$. Les paramètres d'effets fixes sont $\beta_1 = (0, -1)$ et $\beta_2 = (1, 0.5)$. La matrice d'incidence X associée à ces effets fixes est définie par une première colonne de 1 et une seconde colonne générée à partir d'une loi uniforme sur $[0, 1]$. Nous présentons les résultats obtenus sous forme de tableaux donnant les moyennes et les écart-types des estimations calculés sur une suite de 100 réalisations. Les tableaux (3.2) et (3.3) présentent respectivement les résultats obtenus dans les cas Poisson - lien log et exponentiel - lien log.

Dans les deux situations, on constate un bon comportement général de l'algorithme EM,

TAB. 3.3 – Résultats d'estimation des paramètres d'un mélange à 2 composants de GLM par l'algorithme EM sur 100 simulations : cas exponentiel - lien log

		Valeurs estimées	
Valeurs simulées		moyenne	écart-type
	$p_1 = 0.7$	0.6836	0.0481
$\mathcal{C}_1$	$\beta_{11} = 0$	-0.0670	0.2659
	$\beta_{12} = -1$	-0.9405	0.4980
	$p_2 = 0.3$	0.3164	0.0481
$\mathcal{C}_2$	$\beta_{21} = 1$	0.9327	0.1969
	$\beta_{22} = 0.5$	0.5602	0.3308

que ce soit pour l'estimation des proportions du mélange ou pour celle des paramètres $\beta_k = (\beta_{k1}, \beta_{k2})$, $k = 1, 2$. Notons qu'ici les proportions sont très différentes, ce qui augmente la difficulté pour une bonne estimation. D'autres simulations ont été réalisées pour étudier l'impact des divers paramètres sur la qualité des estimations. On constate notamment une diminution des écart-types lorsque l'écart entre les composants augmente. De nombreux auteurs ont étudié sur la base de simulations de Monte Carlo intensives le comportement de cet algorithme pour différentes distributions. Citons, par exemple, Wedel et DeSarbo [65] ou encore Celeux, Chauveau et Diebolt [9].

3.4 Mélange de L2M

Dans cette section, nous nous intéressons aux mélanges de modèles linéaires mixtes introduit par Celeux, Martin et Lavergne [13] pour l'analyse de profils d'expression de gènes. Ces modèles ont pour objectif de bien cerner les sources de variabilité inter-individuelle et de dépendance des observations pour un même individu par des modèles linéaires mixtes construits conditionnellement aux composants du mélange. Nous présentons dans un premier temps le modèle et nous détaillons ensuite comment les paramètres de ce modèle peuvent être estimés par l'algorithme EM. Nous présentons enfin des résultats de simulation.

3.4.1 Le modèle

Supposons que l'on dispose d'observations $y = (y_1', \ldots, y_I')'$ avec y_i correspondant à l'individu i. Chaque y_i est un vecteur de n_i éléments y_{ij} correspondant à des répétitions effectuées sur l'unité i. On suppose, de plus, l'existence d'une structure de classes cachées $\mathcal{C}_k$, $k = 1, \ldots, K$ et l'appartenance de chaque observation à la classe $\mathcal{C}_k$ est caractérisée par le vecteur $z_i = (z_{i1}, \ldots, z_{iK})$ avec $z_{ik} = 1$ si l'unité i appartient à la classe $\mathcal{C}_k$ et 0 sinon. On admet que toutes les observations d'une même unité statistique appartiennent à la même classe. Conditionnellement aux classes, nous introduisons un modèle linéaire mixte en considérant un effet aléatoire "individu" sur les mesures y_{ij}.

Conditionnellement au fait que les observations de l'individu i appartiennent à la classe $\mathcal{C}_k$, on suppose qu'elles sont issues du modèle

$$(Y_i | Z_{ik} = 1) \quad = \quad X_i\beta_k \; + \; U_i\xi_i \; + \; \varepsilon_i$$

où

- $\xi_i | Z_{ik} = 1 \sim \mathcal{N}(0, \sigma_k^2)$,
- $\varepsilon_i \sim \mathcal{N}_{n_i}(0, \tau^2 \, I_{n_i})$,
- On suppose que $\forall i \in \{1, \ldots, I\}$, ε_i et ξ_i sont indépendants et $\forall i, j \in \{1, \ldots, I\}^2 \; i \neq j$, ξ_i et ξ_j, respectivement ε_i et ε_j, sont indépendants,
- β_k est le vecteur des paramètres fixes inconnus de taille q spécifique à la classe $\mathcal{C}_k$,
- σ_k^2 est la variance de l'effet aléatoire spécifique à la classe $\mathcal{C}_k$,
- $X_i = (x_{i1} \vdots \ldots \vdots x_{in_i})'$ est la matrice $n_i \times q$ associée aux effets fixes et $U_i = (u_{i1} \vdots \ldots \vdots u_{in_i})'$ la matrice $n_i \times 1$ associée à l'effet aléatoire ξ_i (en toute généralité, nous considérons ici les éléments u_{ij} quelconques).

La distribution des y_i est donc un mélange défini par :

$$f(y_i | \theta, p) = \sum_{k=1}^{K} p_k \, f_k(y_i | \theta_k)$$

où $f_k(y_i|\theta_k) = f(y_i | z_{ik} = 1, \theta_k)$, $p = (p_1, \ldots, p_K)$, $p_k \; \geq 0$, $k = 1, \ldots, K$ sont les proportions du mélange vérifiant $\sum_{k=1}^{K} p_k = 1$ et $\theta = (\theta_1, \ldots, \theta_K)$ avec $\theta_j = (\beta_j, \sigma_j^2, \tau^2)$.

Nous sommes donc dans le cadre d'un problème de classification de données gaussiennes répétées que nous abordons dans un contexte de modélisation par mélange.

Il s'agit maintenant d'estimer le paramètre θ et le vecteur p des proportions du mélange. Pour cela, nous considérons une approche du maximum de vraisemblance par utilisation de l'algorithme EM qui tire partie de la structure incomplète des données (cf. Dempster, Laird et Rubin [17]). Les données manquantes sont ici de deux types :

- les variables indicatrices z_i d'appartenance des individus à l'un des composants du mélange,
- les effets aléatoires ξ_i.

Notons que le modèle considéré suppose que les paramètres β_k et σ_k^2 dépendent du composant $\mathcal{C}_k$ alors que la variance résiduelle τ^2 est identique pour tous les composants. Dans certaines situations, il peut être utile d'envisager d'autres modèles en contraignant par exemple la variance résiduelle à dépendre de k ou, au contraire, en supposant que certains paramètres du mélange sont indépendants des composants.

3.4.2 Estimation des paramètres par l'algorithme EM

Nous exposons ici la méthodologie EM pour le modèle de mélange présenté en section 3.4.1. Les données complètes x sont formées par les triplets d'objets (y_i, ξ_i, z_i) et leur densité s'écrit :

$$\begin{aligned} f(x;\theta,p) &= \prod_{i=1}^{I} f(y_i,\xi_i,z_i) \\ &= \prod_{i=1}^{I}\prod_{k=1}^{K} \left[P(z_{ik}=1) f(y_i,\xi_i|z_{ik}=1)\right]^{z_{ik}} \end{aligned}$$

On en déduit la log-vraisemblance des données complètes :

$$\begin{aligned} L(\theta,p|x) &= \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\left[\ln P(z_{ik}=1) + \ln f(y_i,\xi_i|z_{ik}=1)\right] \\ &= \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\ln p_k + \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\ln f(y_i,\xi_i|z_{ik}=1) \end{aligned}$$

où $\ln f(y_i,\xi_i|z_{ik}=1)$ se décompose de la façon suivante :

$$\ln f(y_i,\xi_i|z_{ik}=1) = \ln f(y_i|\xi_i,z_{ik}=1) + \ln f(\xi_i|z_{ik}=1)$$

avec

- $\ln f(y_i|\xi_i, z_{ik}=1) = -\frac{1}{2}\Big[n_i \ln 2\pi + n_i \ln \tau^2 + \frac{\varepsilon_i'\varepsilon_i}{\tau^2}\Big]$
  $= -\frac{1}{2}\Big[n_i \ln 2\pi + n_i \ln \tau^2 + \frac{(y_i - X_i\beta_k - U_i\xi_i)'(y_i - X_i\beta_k - U_i\xi_i)}{\tau^2}\Big]$

- $\ln f(\xi_i|z_{ik}=1) = -\frac{1}{2}\Big[\ln 2\pi + \ln \sigma_k^2 + \frac{\xi_i^2}{\sigma_k^2}\Big]$

En prenant l'espérance de $L(\theta, p|x)$ par rapport à la distribution conditionnelle des données manquantes (ξ_i, z_i) sachant les observations y et la valeur courante des paramètres $(\theta^{[t]}, p^{[t]})$, on obtient :

$$Q(\theta, p|\theta^{[t]}, p^{[t]}) = \sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i)\ln p_k - \frac{1}{2}\sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i)\Big[(n_i+1)\ln 2\pi + n_i \ln \tau^2 + \ln \sigma_k^2 + \frac{E_{Ck}^{[t]}(\varepsilon_i'\varepsilon_i)}{\tau^2} + \frac{E_{Ck}^{[t]}(\xi_i^2)}{\sigma_k^2}\Big]$$

où

• $E_{Ck}^{[t]}(.)$ désigne l'espérance conditionnelle prise par rapport à la densité des effets aléatoires sachant les données observées, les appartenances aux classes et la valeur courante des paramètres

- $t_k^{[t]}(y_i) = \dfrac{p_k^{[t]}\, f_k(y_i|\theta_k^{[t]})}{\sum_{l=1}^{K} p_l^{[t]}\, f_l(y_i|\theta_l^{[t]})}$

- $E_{Ck}^{[t]}(\xi_i) = E(\xi_i|y_i, z_{ik}=1, \theta^{[t]})$
  $= \sigma_k^{2[t]}\, U_i'\, \Gamma_{k,i}^{[t]-1}\, (y_i - X_i\beta_k^{[t]})$
  où $\Gamma_{k,i}^{[t]} = \tau^{2[t]} Id_{n_i} + \sigma_k^{2[t]} U_i U_i'$

- $\mathrm{Var}_{Ck}^{[t]}(\xi_i) = \mathrm{Var}(\xi_i|y_i, z_{ik}=1, \theta^{[t]})$
  $= \sigma_k^{2[t]} - \sigma_k^{4[t]}\, U_i'\, \Gamma_{k,i}^{[t]-1}\, U_i$

- $E_{Ck}^{[t]}(\xi_i^2) = E(\xi_i|y_i, z_{ik}=1, \theta^{[t]})'\, E(\xi_i|y_i, z_{ik}=1, \theta^{[t]}) + \mathrm{tr}\big[\mathrm{Var}_{Ck}^{[t]}(\xi_i)\big]$
  $= \sigma_k^{4[t]}\, (y_i - X_i\beta_k^{[t]})'\, \Gamma_{k,i}^{[t]-1} U_i\, U_i'\, \Gamma_{k,i}^{[t]-1}\, (y_i - X_i\beta_k^{[t]})$
  $+\sigma_k^{2[t]} - \sigma_k^{4[t]} \mathrm{tr}\big[\Gamma_{k,i}^{[t]-1} U_i U_i'\big]$

• $E_{Ck}^{[t]}(\varepsilon_i'\varepsilon_i) = \big[y_i - X_i\beta_k - U_i E_{Ck}^{[t]}(\xi_i)\big]'\, \big[y_i - X_i\beta_k - U_i E_{Ck}^{[t]}(\xi_i)\big] + \mathrm{tr}\big[U_i\, \mathrm{Var}_{Ck}^{[t]}(\xi_i)\, U_i'\big]$

Il faut maintenant maximiser $Q(\theta, p|\theta^{[t]}, p^{[t]})$ par rapport à $(\theta,\ p)$ ou plus exactement $Q^*(\theta, p|\theta^{[t]}, p^{[t]}) = Q(\theta, p|\theta^{[t]}, p^{[t]}) - \psi(\sum_{k=1}^{K} p_k - 1)$ où ψ est un multiplicateur de Lagrange.

On obtient par dérivation :

$$\frac{\partial Q^*}{\partial p_k} = \frac{\sum_{i=1}^{I} t_k^{[t]}(y_i)}{p_k} - \psi \tag{3.8}$$

$$\frac{\partial Q^*}{\partial \sigma_k^2} = -\frac{1}{2}\sum_{i=1}^{I} t_k^{[t]}(y_i)\Big[\frac{1}{\sigma_k^2} - \frac{E_{Ck}^{[t]}(\xi_i^2)}{\sigma_k^4}\Big] \tag{3.9}$$

$$\frac{\partial Q^*}{\partial \tau^2} = -\frac{1}{2}\sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i)\Big[\frac{n_i}{\tau^2} - \frac{E_{Ck}^{[t]}(\varepsilon_i'\varepsilon_i)}{\tau^4}\Big] \tag{3.10}$$

La dérivation par rapport à β_k nous conduit à :

$$\begin{aligned}\frac{\partial Q^*}{\partial \beta_k} &= -\frac{1}{2}\sum_{i=1}^{I} t_k^{[t]}(y_i)\,\frac{1}{\tau^2}\,\frac{\partial E_{Ck}^{[t]}(\varepsilon_i'\varepsilon_i)}{\partial \beta_k} \\ &= \sum_{i=1}^{I} t_k^{[t]}(y_i)\,\frac{1}{\tau^2}\,X_i'\Big[y_i - X_i\beta_k - U_i E_{Ck}^{[t]}(\xi_i)\Big].\end{aligned} \tag{3.11}$$

Par annulation des dérivées (3.8) et(3.9), il vient immédiatement :

$$p_k^{[t+1]} = \frac{\sum_{i=1}^{I} t_k^{[t]}(y_i)}{I}$$

$$\begin{aligned}\sigma_k^{2[t+1]} &= \frac{\sum_{i=1}^{I} t_k^{[t]}(y_i) E_{Ck}^{[t]}(\xi_i^2)}{\sum_{i=1}^{I} t_k^{[t]}(y_i)} \\ &= \frac{1}{\sum_{i=1}^{I} t_k^{[t]}(y_i)}\sum_{i=1}^{I} t_k^{[t]}(y_i)\Big[\sigma_k^{4[t]}\,(y_i - X_i\beta_k^{[t]})'\,\Gamma_{k,i}^{[t]-1}U_i\,U_i'\,\Gamma_{k,i}^{[t]-1}\,(y_i - X_i\beta_k^{[t]}) \\ &\qquad + \sigma_k^{2[t]} - \sigma_k^{4[t]}\mathrm{tr}(\Gamma_{k,i}^{[t]-1}U_iU_i')\Big]\end{aligned}$$

Par annulation, l'équation obtenue à partir de (3.11) ne dépend pas de τ^2 et on peut résoudre en β_k :

$$\begin{aligned}\Big(\sum_{i=1}^{I} t_k^{[t]}(y_i)X_i'\,X_i\Big)\,\beta_k^{[t+1]} &= \sum_{i=1}^{I} t_k^{[t]}(y_i)\,X_i'\Big[y_i - U_iE_{Ck}^{[t]}(\xi_i)\Big] \\ &= \sum_{i=1}^{I} t_k^{[t]}(y_i)\Big[\tau^{2[t]}\,X_i'\,\Gamma_{k,i}^{[t]-1}\,(y_i - X_i\beta_k^{[t]}) + X_i'\,X_i\beta_k^{[t]}\Big]\end{aligned} \tag{3.12}$$

On aboutit à l'expression de β_k suivante :

$$\beta_k^{[t+1]} = \big(\sum_{i=1}^{I} t_k^{[t]}(y_i)X_i'\,X_i\big)^{-1}\sum_{i=1}^{I} t_k^{[t]}(y_i)\Big[\tau^{2[t]}\,X_i'\,\Gamma_{k,i}^{[t]-1}\,(y_i - X_i\beta_k^{[t]}) + X_i'\,X_i\beta_k^{[t]}\Big]$$

Par annulation de l'équation (3.10), en notant $n = \sum_{i=1}^{I} n_i$, on obtient :

$$\tau^{2[t+1]} = \frac{1}{n} \sum_{i=1}^{I} \sum_{k=1}^{K} t_k^{[t]}(y_i) E_{Ck}^{[t]}(\varepsilon_i' \varepsilon_i)$$

où, à chaque itération, $E_{Ck}^{[t]}(\varepsilon_i' \varepsilon_i)$ sera calculée en la valeur courante des paramètres par :

$$\begin{aligned} E_{Ck}^{[t]}(\varepsilon_i' \varepsilon_i) &= \left[y_i - X_i \beta_k^{[t]} - U_i E_{Ck}^{[t]}(\xi_i)\right]' \left[y_i - X_i \beta_k^{[t]} - U_i E_{Ck}^{[t]}(\xi_i)\right] + \mathrm{tr}\left[U_i \, \mathrm{Var}_{Ck}^{[t]}(\xi_i) \, U_i'\right] \\ &= \tau^{4[t]} (y_i - X_i \beta_k^{[t]})' \, \Gamma_{k,i}^{[t]-1} \, \Gamma_{k,i}^{[t]-1} \, (y_i - X_i \beta_k^{[t]}) + n_i \tau^{2[t]} - \tau^{4[t]} \mathrm{tr}(\Gamma_{k,i}^{[t]-1}) \end{aligned}$$

Finalement, cela conduit à l'expression de τ^2 suivante :

$$\begin{aligned} \tau^{2[t+1]} = \frac{1}{n} \sum_{i=1}^{I} \sum_{k=1}^{K} t_k^{[t]}(y_i) \Big[& \tau^{4[t]} (y_i - X_i \beta_k^{[t]})' \, \Gamma_{k,i}^{[t]-1} \, \Gamma_{k,i}^{[t]-1} \, (y_i - X_i \beta_k^{[t]}) \\ & + n_i \tau^{2[t]} - \tau^{4[t]} \mathrm{tr}(\Gamma_{k,i}^{[t]-1}) \Big] \end{aligned}$$

3.4.3 Notations matricielles

En définissant la matrice $W_{k,i}^{[t]} = \dfrac{1}{t_k^{[t]}(y_i)} Id_{n_i}$, l'équation (3.12) peut s'écrire :

$$\Big(\sum_{i=1}^{I} X_i' W_{k,i}^{[t]\,-1} X_i \Big) \beta_k^{[t+1]} = \sum_{i=1}^{I} X_i' W_{k,i}^{[t]\,-1} \Big(y_i - U_i E_{Ck}^{[t]}(\xi_i) \Big)$$

En considérant les matrices $X = \begin{bmatrix} X_1 \\ - \\ \vdots \\ - \\ X_I \end{bmatrix}$, $U = \mathrm{diag}\{ U_i \}_{i=1,\ldots,I}$, $W_k = \mathrm{diag}\{ W_{k,i}^{[t]} \}_{i=1,\ldots,I}$ et le vecteur de l'effet aléatoire $\xi = (\xi_1, \ldots, \xi_I)'$ de taille I, on aboutit à l'écriture matricielle suivante :

$$(X' W_k^{[t]\,-1} X) \beta_k^{[t+1]} = X' W_k^{[t]\,-1} [y - U E_{Ck}^{[t]}(\xi)].$$

Si l'on introduit le vecteur défini par $\tilde{y} = y - U E_{Ck}^{[t]}(\xi)$, les équations précédentes deviennent alors :

$$(X' W_k^{[t]\,-1} X) \beta_k^{[t+1]} = X' W_k^{[t]\,-1} \tilde{y}. \tag{3.13}$$

En considérant les données transformées $\tilde{y}$, on reconnait alors dans le système (3.13) les équations classiques des moindres carrés généralisés obtenus en pondérant chaque individu par sa probabilité a posteriori d'appartenir à la classe $\mathcal{C}_k$.

3.4.4 Remarques

Concernant l'estimation des paramètres des mélanges de L2M, nous avons présenté l'algorithme EM afin de maximiser la vraisemblance. Une alternative au maximum de vraisemblance pour l'estimation des paramètres d'un L2M est la méthode du maximum de vraisemblance restreint. Cette méthode permet de se focaliser davantage sur les composantes de la variance en ne maximisant que la partie de la vraisemblance concernant les composantes de la variance et indépendante des effets fixes. Dans le contexte de mélange, l'estimation des paramètres par l'approche REML n'a pas été ici considérée. Cependant, soulignons que l'approche REML pourrait s'avérer intéressante pour un mélange comprenant certaines proportions faibles. En effet, dans ce contexte précis, l'approche REML risque de fournir des estimations plus fiables que l'approche ML.

3.4.5 Des résultats de simulation

Nous présentons des résultats de simulations afin d'évaluer la capacité de l'algorithme EM à estimer correctement les paramètres d'un mélange de L2M. Nous fixons le nombre d'individus statistiques $I = 200$ et nous prenons le même nombre de répétitions par individu, c'est-à-dire, pour tout $i = 1, \ldots, I,\ n_i = J = 4$. Nous considérons un mélange à 3 classes avec les proportions du mélange $p_1 = 0.3$, $p_2 = 0.5$ et $p_3 = 0.2$. Les variances des effets aléatoires sont $\sigma_1^2 = 0.2$, $\sigma_2^2 = 0.5$ et $\sigma_3^2 = 0.8$. Nous considérons ici un seul paramètre d'effet fixe par composant défini par : $\beta_1 = -2$, $\beta_2 = 2$ et $\beta_2 = 6$. Nous générons ainsi des données selon deux modèles (A) et (A') définis ci-dessus pour lesquels seule la variance résiduelle τ^2 diffère. Pour le modèle (A), nous prenons $\tau^2 = 2$ et pour le modèle (A'), $\tau^2 = 3$.

Le tableau (3.4) donne les moyennes et les écart-types des estimations obtenus sur 100 jeux de données générés à partir des modèles (A) et (A'). Dans toutes nos simulations, l'algorithme EM a été initialisé à partir de la méthode des K-means. Le tableau (3.6) donne la moyenne des taux de bon classement calculés en utilisant la méthode du maximum a posteriori (MAP) à partir des estimations $\hat{p}$, $\hat{\theta}$ obtenues par l'algorithme EM. Cette méthode consiste à affecter toutes les observations de l'individu i au composant $\mathcal{C}_k$ le plus probable a posteriori, c'est-à-dire tel que :

$$k = \text{argmax}_{1 \leq l \leq K} \widehat{t_l(y_i)}$$

TAB. 3.4 – Résultats d'estimation des paramètres obtenus par l'algorithme EM sur 100 simulations des modèles de mélange (A) et (A')

		Valeurs estimées			
		Modèle (A) : $\tau^2 = 2$		Modèle (A') : $\tau^2 = 3$	
Valeurs simulées		moyenne	écart-type	moyenne	écart-type
	$p_1 = 0.3$	0.3009	0.0114	0.3020	0.0207
$\mathcal{C}_1$	$\beta_1 = -2$	-1.9778	0.1134	-1.9633	0.1781
	$\sigma_1^2 = 0.2$	0.2051	0.1395	0.2273	0.2182
	$p_2 = 0.5$	0.4995	0.0247	0.4947	0.0495
$\mathcal{C}_2$	$\beta_2 = 2$	2.0071	0.1324	2.0224	0.1494
	$\sigma_2^2 = 0.5$	0.5396	0.2856	0.4632	0.3837
	$p_3 = 0.2$	0.1996	0.0195	0.2023	0.0380
$\mathcal{C}_3$	$\beta_3 = 6$	6.0109	0.2313	6.0128	0.4295
	$\sigma_3^2 = 0.8$	0.7957	0.4457	0.7633	0.6318
	τ^2	1.9755	0.1175	3.0195	0.1768

avec $\widehat{t_l(y_i)} = P(Z_{il} = 1|y_i, \hat{p}, \hat{\theta})$.

Le tableau (3.4) montre que, globalement, l'algorithme EM estime en moyenne raisonnablement bien les valeurs des paramètres des modèles simulés. Globalement, la précision des estimations dépend de la variance des effets aléatoires : plus cette variance est grande, plus l'écart-type des estimations est important. De la même façon, le tableau (3.6) montre des taux moyens de bon classement globalement satisfaisants mais qui se détériorent quand la variance de l'effet aléatoire augmente. Notons également que l'écart-type des estimations augmente avec la variance résiduelle τ^2 (modèle (A')) même si globalement les estimations restent très correctes en moyenne. En comparant les résultats du tableau (3.6) pour les modèles (A) et (A'), on constate également que le taux de bon classement diminue lorsque la variance résiduelle τ^2 augmente. Ces différentes remarques traduisent un bon comportement de la procédure d'estimation des paramètres du modèle suivant la nature des variations prises en considération.

Afin d'évaluer la qualité des estimations obtenues selon que les composants sont plus

TAB. 3.5 – Résultats d'estimation des paramètres obtenus par l'algorithme EM sur 100 simulations des modèles de mélange (B) et (B')

		Valeurs estimées			
		Modèle (B) : $\tau^2 = 2$		Modèle (B') : $\tau^2 = 3$	
Valeurs simulées		moyenne	écart-type	moyenne	écart-type
	$p_1 = 0.3$	0.3640	0.1449	0.3420	0.1391
$\mathcal{C}_1$	$\beta_1 = -1$	-0.8840	0.4414	-0.9998	0.4640
	$\sigma_1^2 = 0.2$	0.2366	0.3369	0.1134	0.3321
	$p_2 = 0.5$	0.3370	0.1229	0.3796	0.1207
$\mathcal{C}_2$	$\beta_2 = 1$	1.0436	0.5998	1.0484	0.6215
	$\sigma_2^2 = 0.5$	0.0914	0.4216	0.1254	0.6298
	$p_3 = 0.2$	0.2990	0.1787	0.2784	0.1721
$\mathcal{C}_3$	$\beta_3 = 3$	2.7618	0.8356	2.8579	0.7935
	$\sigma_3^2 = 0.8$	0.7698	0.6948	0.6380	0.8414
	τ^2	2.0264	0.1297	3.0045	0.1726

TAB. 3.6 – Moyennes des taux de bon classement (en %) obtenues par l'algorithme EM sur 100 simulations des modèles (A), (A'), (B) et (B')

	Modèle (A) $\beta = (-2, 2, 6)$ $\tau^2 = 2$	Modèle (A') $\beta = (-2, 2, 6)$ $\tau^2 = 3$	Modèle (B) $\beta = (-1, 1, 3)$ $\tau^2 = 2$	Modèle (B') $\beta = (-1, 1, 3)$ $\tau^2 = 3$
$\mathcal{C}_1$	97.98	96.13	84.57	79.67
$\mathcal{C}_2$	96.70	94.00	51.82	55.50
$\mathcal{C}_3$	93.67	90.60	73.85	68.42

ou moins séparés, nous avons effectué d'autres simulations en considérant les modèles (B) et (B') définis par : $\beta_1 = -1$, $\beta_2 = 1$ et $\beta_2 = 3$. Comme précédemment, les modèles (B) et (B') diffèrent uniquement par la valeur de la variance résiduelle τ^2. Pour le modèle (B), nous prenons $\tau^2 = 2$ et pour le modèle (B'), $\tau^2 = 3$. Le tableau (3.5) donne les moyennes et les écart-types des estimations obtenus sur 100 jeux de données générés à partir des modèles (B) et (B'). Les moyennes des taux de bon classement associés à ces modèles sont également présentés dans le tableau (3.6).

Le tableau (3.5) montre clairement que les estimations obtenues sont moins bonnes quand les composants sont moins séparés. En particulier, on note une difficulté pour l'algorithme à estimer correctement les paramètres du composant intermédiaire. Les taux de bon classement présentés dans le tableau (3.6) témoignent également de cette difficulté avec un taux de bon classement moyen pour la composante $\mathcal{C}_2$ de $51,82\%$ (respectivement 55.50%) pour le modèle (B) (respectivement (B')).

Pour finir, afin d'évaluer le rôle des répétitions dans l'estimation des paramètres d'un mélange de L2M, nous réalisons quelques simulations complémentaires. Nous générons ainsi 100 jeux de données à partir du modèle (B) avec un nombre de répétitions $J = 4$, $J = 8$ et $J = 12$. Nous notons (C) et (D) les deux derniers modèles qui diffèrent du modèle (B) par le nombre de répétitions. Les différents résultats sont présentés dans les tableaux (3.7) et (3.8).

Le tableau (3.7) montre clairement que les estimations obtenues sont meilleures quand le nombre de répétitions augmente. Pour $J = 8$ et $J = 12$, les moyennes des estimations sont nettement plus proches des valeurs simulées avec une meilleure précision pour $J = 12$. Ces résultats sont cohérents : il semble en effet normal qu'une augmentation des répétitions entraîne une meilleure précision des estimations. De la même façon, les résultats du tableau (3.8) indiquent que le taux de bon classement augmente avec le nombre de répétitions. Bien qu'il s'agisse de simulations, la prise en compte des répétitions semble ainsi être une source d'information importante pour la validation des classes.

TAB. 3.7 – Résultats d'estimation des paramètres obtenus par l'algorithme EM sur 100 simulations des modèles de mélange (B), (C) et (D)

		Valeurs estimées					
		Modèle (B) : $J = 4$		Modèle (C) : $J = 8$		Modèle (D) : $J = 12$	
Valeurs simulées		moy.	e.t.	moy.	e.t.	moy.	e.t.
	$p_1 = 0.3$	0.3632	0.1290	0.3238	0.1036	0.3253	0.0695
$\mathcal{C}_1$	$\beta_1 = -1$	-0.8842	0.3812	-0.9406	0.2793	-0.9320	0.1953
	$\sigma_1^2 = 0.2$	0.2442	0.3331	0.2421	0.2065	0.2291	0.1461
	$p_2 = 0.5$	0.3159	0.1211	0.4306	0.1726	0.4589	0.1382
$\mathcal{C}_2$	$\beta_2 = 1$	1.0441	0.5489	1.0270	0.3231	1.0681	0.2246
	$\sigma_2^2 = 0.5$	0.0542	0.3191	0.4113	0.4222	0.4198	0.3264
	$p_3 = 0.2$	0.3209	0.1768	0.2456	0.1504	0.2158	0.1098
$\mathcal{C}_3$	$\beta_3 = 3$	2.6623	0.7732	2.9923	0.7802	3.0593	0.5755
	$\sigma_3^2 = 0.8$	0.9895	0.7662	0.7869	0.6685	0.6634	0.4114
	$\tau^2 = 2$	2.0138	0.1163	2.0121	0.0777	1.9987	0.0578

TAB. 3.8 – Moyennes des taux de bon classement (en %) obtenues par l'algorithme EM sur 100 simulations des modèles (B), (C) et (D)

	Modèle (B) $\beta = (-1, 1, 3)$, $\tau^2 = 2$ $J = 4$	Modèle (C) $\beta = (-1, 1, 3)$, $\tau^2 = 2$ $J = 8$	Modèle (D) $\beta = (-1, 1, 3)$, $\tau^2 = 2$ $J = 12$
$\mathcal{C}_1$	85.72	87.60	91.30
$\mathcal{C}_2$	50.99	70.31	78.91
$\mathcal{C}_3$	76.87	71.97	72.77

3.5 Mélange de GL2M : cas exponentiel - lien log

3.5.1 Introduction

Après avoir traité le cas gaussien, nous en arrivons tout naturellement au mélange de GL2M. Dans cette section, nous nous intéressons plus particulièrement aux modèles de mélange pour des données répétées issues d'une loi exponentielle. Nous allons constater que l'estimation des paramètres d'un tel modèle par l'utilisation directe de l'algorithme EM n'est pas envisageable. Pourtant, selon les résultats de la section 3.2, la démarche EM semble être un outil intéressant pour faire face au problème des mélanges. Après avoir défini le modèle en section 3.5.2 et avoir souligné les limites de l'algorithme EM en section 3.5.3, nous proposons en section 3.5.4 une première méthode d'estimation combinant la linéarisation "Gumbel" spécifique au modèle exponentiel mixte associé à chaque composant du mélange et l'utilisation de l'algorithme EM pour le mélange obtenu de L2M approchés. Une seconde méthode est proposée en section 3.5.5 introduisant à l'étape E de l'algorithme EM une étape de Metropolis-Hastings pour construire un algorithme de type MCEM.

3.5.2 Le modèle

Conditionnellement au fait que les observations de l'individu i appartiennent à la classe $\mathcal{C}_k$, on suppose maintenant qu'elles sont issues d'un modèle exponentiel mixte. La distribution des y_i est un mélange

$$f(y_i|\theta, p) = \sum_{k=1}^{K} p_k f_k(y_i|\theta_k)$$

où $f_k(.|\theta_k)$ est ici la densité marginale d'un modèle exponentiel mixte de paramètres $\theta_k = (\beta_k, \sigma_k^2)$. Mais, contrairement au cas gaussien, cette densité marginale n'est pas directement accessible. En effet, sachant que les observations de l'individu i appartiennent à la classe $\mathcal{C}_k$, c'est conditionnellement à l'effet aléatoire ξ_i que l'hypothèse de distribution est formulée sur la loi des y_i. Elle peut être résumée de la façon suivante :

$$Y_i|\xi_i, Z_{ik} = 1 \sim \mathcal{E}xp(\mu_{\xi,i}^k) \text{ avec } \begin{cases} \mu_{\xi,i}^k = \exp(X_i\beta_k + U_i\xi_i) \\ \xi_i|Z_{ik} = 1 \sim \mathcal{N}(0, \sigma_k^2) \end{cases}$$

où, comme précédemment,

- on suppose que $\forall i, j \in \{1, \ldots, I\}^2$ $i \neq j$, ξ_i et ξ_j sont indépendants,
- β_k est le vecteur des paramètres d'effets fixes inconnus de taille q spécifique à la classe $\mathcal{C}_k$,
- σ_k^2 est la variance de l'effet aléatoire spécifique à la classe $\mathcal{C}_k$,
- $X_i = (x_{i1} \vdots \ldots \vdots x_{in_i})'$ est la matrice $n_i \times q$ associée aux effets fixes et $U_i = (u_{i1} \vdots \ldots \vdots u_{in_i})'$ la matrice $n_i \times 1$ associée à l'effet aléatoire ξ_i.

Notons que nous considérons ici le cas du lien logarithme, et non pas le lien canonique. Une raison qui justifie le choix de ce lien est notamment le fait que ce lien assure la positivité du paramètre de la loi exponentielle.

3.5.3 Limites de l'algorithme EM dans le cas exponentiel

L'utilisation de l'algorithme EM nous conduit à nous intéresser à la log-vraisemblance des données complètes :

$$L(\theta, p|x) = \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik} \ln p_k + \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik} \ln f(y_i, \xi_i | z_{ik} = 1)$$

où $\ln f(y_i, \xi_i | z_{ik} = 1)$ se décompose de la façon suivante :

$$\ln f(y_i, \xi_i | z_{ik} = 1) = \ln f(y_i | \xi_i, z_{ik} = 1) + \ln f(\xi_i | z_{ik} = 1)$$

avec

- $$\begin{aligned} \ln f(y_i|\xi_i, z_{ik} = 1) &= \sum_{j=1}^{n_i} \ln f(y_{ij}|\xi_i, z_{ik} = 1), \\ &= -\sum_{j=1}^{n_i} \Big[x'_{ij}\beta_k + u_{ij}\xi_i + \frac{y_{ij}}{\exp(x'_{ij}\beta_k + u_{ij}\xi_i)} \Big] \end{aligned}$$

puisque les y_{ij} sont indépendants conditionnellement à ξ_i.

- $$\ln f(\xi_i|z_{ik} = 1) = -\frac{1}{2}\Big[\ln 2\pi + \ln \sigma_k^2 + \frac{\xi_i^2}{\sigma_k^2} \Big]$$

En prenant l'espérance de $L(\theta, p|x)$ par rapport à la distribution conditionnelle des données manquantes (ξ_i, z_i) sachant les observations y et la valeur courante des paramètres

$(\theta^{[t]}, p^{[t]})$, on obtient :

$$\begin{aligned} Q(\theta, p|\theta^{[t]}, p^{[t]}) &= \sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i) \ln p_k \\ &\quad - \sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i) \sum_{j=1}^{n_i} \Big[x'_{ij}\beta_k + u_{ij} E_{Ck}^{[t]}(\xi_i) + \exp(-x'_{ij}\beta_k) E_{Ck}^{[t]}\big[\exp(-u_{ij}\xi_i)\big]\, y_{ij} \Big] \\ &\quad - \frac{1}{2}\sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i) \Big[\ln 2\pi + \ln \sigma_k^2 + \frac{E_{Ck}^{[t]}(\xi_i^2)}{\sigma_k^2} \Big] \end{aligned}$$

où $t_k^{[t]}(y_i) = \dfrac{p_k^{[t]}\, f_k(y_i|\theta_k^{[t]})}{\sum_{l=1}^{K} p_l^{[t]}\, f_l(y_i|\theta_l^{[t]})}$.

La maximisation de $Q(\theta, p|\theta^{[t]}, p^{[t]})$ conduit alors à des expressions qui dépendent des espérances conditionnelles $E_{Ck}^{[t]}(\xi_i^2)$, $E_{Ck}^{[t]}\big[\exp(-u_{ij}\xi_i)\big]$ et des probabilités a posteriori $t_k^{[t]}(y_i)$ pour $k = 1, \ldots, K$. À ce stade, les limites de l'algorithme EM sont de deux sortes.

- Premièrement, on ne sait pas atteindre la distribution marginale $f_k(.|\theta_k^{[t]})$ pour chaque composant du mélange nécessaire aux calculs des $t_k^{[t]}(y_i)$.
- Deuxièmement, on est confronté au problème du calcul des différentes espérances mises en jeu qui sont réalisables avec la loi normale grâce aux règles de conditionnement mais qui pose problème pour d'autres lois.

Dans les deux sections suivantes, nous proposons alors deux méthodes d'estimation des paramètres permettant de contourner ces problèmes liés à l'algorithme EM.

3.5.4 Méthode d'estimation basée sur une linéarisation

La méthode que nous proposons combine la linéarisation "Gumbel" spécifique au modèle exponentiel mixte (Gaudoin, Lavergne et Soler [21]) associé à chaque composant du mélange et l'utilisation de l'algorithme EM pour un mélange de L2M (Celeux, Martin et Lavergne [13]).

Conditionnellement au composant $\mathcal{C}_k$, l'hypothèse de distribution pour un individu i est définie par :

$$\begin{aligned} Y_i|\xi_i, Z_{ik} = 1 \sim \mathcal{E}xp(\mu_{\xi,i}^k) \quad \text{d'où} \qquad & \frac{Y_i}{\mu_{\xi,i}^k} \sim \mathcal{E}xp(1) \\ \text{ou encore} \quad & \ln(Y_i) - \ln(\mu_{\xi,i}^k) \sim \mathcal{G}umbel. \end{aligned}$$

En notant 0_{n_i} le vecteur $(0,\ldots,0)'$ de taille n_i, on peut écrire dans le composant $\mathcal{C}_k$:

$$\ln(Y_i) - \ln(\mu^k_{\xi,i}) = \gamma + \varepsilon_i \quad \text{avec} \quad E(\varepsilon_i) = 0_{n_i} \quad \text{et} \quad \text{var}(\varepsilon_i) = \frac{\pi^2}{6} Id_{n_i}$$

ou encore, en posant $D_i = \ln(Y_i) - \gamma$:

$$D_i = X_i\beta_k + U_i\xi_i + \varepsilon_i.$$

Ainsi, en approchant la loi des erreurs par une loi gaussienne centrée et de matrice de variance $\frac{\pi^2}{6} Id_{n_i}$, on définit un modèle linéaire mixte $\mathcal{M}_k$ pour les données $d_i = \ln(y_i) - \gamma$ connaissant le composant $\mathcal{C}_k$ dont est issu l'individu i. Notons que le modèle $\mathcal{M}_k$ est un modèle linéaire mixte à variance résiduelle connue.

Nous procédons maintenant à l'estimation des paramètres par l'algorithme EM dans le mélange de modèles linéaires mixtes défini pour l'individu i par :

$$h(d_i|\theta, p) = \sum_{k=1}^{K} p_k h_k(d_i|\theta_k)$$

où $d_i = \ln(y_i) - \gamma$ et $h_k(d_i|\theta_k)$ est la densité gaussienne multivariée de paramètre $\theta_k = (\beta_k, \sigma^2_k)$ d'espérance $X_i\beta_k$ et de matrice de variance $\Gamma_{k,i} = \sigma^2_k U_i U_i' + \frac{\pi^2}{6} Id_{n_i}$. Dans cette approche, le vecteur d_i est défini à partir des données y_i indépendamment du composant $\mathcal{C}_k$ dont est issu l'individu et des valeurs courantes des paramètres.

L'estimation par l'algorithme EM des paramètres de ce mélange de modèles linéaires mixtes conduit aux expressions explicites suivantes pour $k = 1,\ldots,K$:

$$p_k^{[t+1]} = \frac{\sum_{i=1}^{I} t_k^{[t]}(d_i)}{I}$$

$$\sigma_k^{2[t+1]} = \frac{1}{\sum_{i=1}^{I} t_k^{[t]}(d_i)} \sum_{i=1}^{I} t_k^{[t]}(d_i)\Big[\sigma_k^{4[t]} (d_i - X_i\beta_k^{[t]})' \Gamma_{k,i}^{[t]-1} U_i U_i' \Gamma_{k,i}^{[t]-1} (d_i - X_i\beta_k^{[t]}) + \sigma_k^{2[t]} - \sigma_k^{4[t]} \text{tr}(\Gamma_{k,i}^{[t]-1} U_i U_i')\Big]$$

$$\beta_k^{[t+1]} = \Big(\sum_{i=1}^{I} t_k^{[t]}(d_i) X_i' X_i\Big)^{-1} \sum_{i=1}^{I} t_k^{[t]}(d_i)\Big[\frac{\pi^2}{6} X_i' \Gamma_{k,i}^{[t]-1}(d_i - X_i\beta_k^{[t]}) + X_i' X_i \beta_k^{[t]}\Big]$$

Notons que les expressions ci-dessus correspondent aux expressions obtenues à la section 3.4 dans le cas d'un mélange de L2M pour les nouvelles données d_i. La seule différence

réside dans le fait qu'ici il n'y a pas de variance résiduelle à estimer puisque celle-ci est connue et égale à $\frac{\pi^2}{6}$. Dans l'expression de $\beta_k^{[t+1]}$, nous retrouvons ainsi la valeur $\frac{\pi^2}{6}$ et la matrice de variance est égale à $\Gamma_{k,i}^{[t]} = \sigma_k^{2[t]} U_i U_i' + \frac{\pi^2}{6} Id_{n_i}$.

Pour résumer, la première méthode d'estimation proposée se décompose en deux étapes :

- la linéarisation utilisant une propriété spécifique au modèle exponentiel conduisant à la définition des données de travail $d_i = \ln(y_i) - \gamma$
- l'estimation par l'algorithme EM des paramètres du mélange de modèles linéaires mixtes à variance résiduelle connue obtenu à l'étape précédente pour les nouvelles données d_i.

3.5.5 Algorithme de type MCEM

Nous avons vu en section 3.5.3 que l'utilisation de l'algorithme EM conduit à des expressions qui dépendent des espérances conditionnelles $E_{Ck}^{[t]}(\xi_i^2)$, $E_{Ck}^{[t]}\big[\exp(-u_{ij}\xi_i)\big]$ et des probabilités a posteriori $t_k^{[t]}(y_i)$ pour $i = 1, \ldots, I$ et $k = 1, \ldots, K$. Du fait de la non-accessibilité de la distribution marginale de Y_i pour chaque composant du mélange d'une part, et de la distribution de ξ_i sachant Y_i pour chaque composant du mélange d'autre part, ces quantités ne sont pas calculables de façon formelle. L'utilisation directe de l'algorithme EM n'est donc pas directement envisageable. Nous proposons ainsi une seconde méthode contournant les problèmes liés à l'algorithme EM. Cette méthode est une adaptation de l'algorithme MCEM mis en place par McCulloch [39] pour l'estimation des paramètres d'un GL2M et décrit au chapitre 1.

3.5.5.1 Le principe

Puisque les espérances conditionnelles $E_{Ck}^{[t]}(\xi_i^2)$, $E_{Ck}^{[t]}\big[\exp(-u_{ij}\xi_i)\big]$ et les probabilités a posteriori $t_k^{[t]}(y_i)$ pour chaque individu i et chaque composant du mélange $\mathcal{C}_k$ ne sont pas calculables de façon formelle, nous allons chercher à les approcher.

Pour approcher les espérances conditionnelles, nous proposons d'introduire, à l'itération $[t]$ de l'étape E de l'algorithme EM, une étape de Metropolis-Hastings (ne nécessitant pas de connaître les densités marginales) afin de simuler, pour chaque individu i et chaque composant $\mathcal{C}_k$, des effets aléatoires à partir de la distribution de ξ_i sachant Y_i et $Z_{ik} = 1$ en la valeur courante du paramètre $\theta_k^{[t]}$. Les effets aléatoires ainsi simulés permettent d'approcher les espérances conditionnelles par des méthodes de Monte Carlo.

De la même façon, pour approcher les probabilités a posteriori, nous proposons de simuler, à l'itération $[t]$ de l'étape E, pour chaque individu i et chaque composant $\mathcal{C}_k$, des effets aléatoires à partir, cette fois-ci, de la distribution (connue) de ξ_i sachant $Z_{ik} = 1$ en la valeur courante du paramètre $\theta_k^{[t]}$ afin d'approcher par Monte Carlo les densités marginales $f_k(y_i|\theta_k^{[t]})$ nécessaires au calcul des différentes probabilités a posteriori $t_k^{[t]}(y_i)$.

3.5.5.2 L'étape de Metropolis-Hastings

Avant de présenter l'algorithme de type MCEM proposé, nous revenons sur l'étape de Metropolis-Hastings introduit à l'itération $[t]$ pour $i = 1, \ldots, I$ et $k = 1, \ldots, K$. Nous proposons d'utiliser la distribution marginale de $\xi_i|Z_{ik} = 1$ en la valeur courante $\theta_k^{[t]}$ comme distribution instrumentale h à partir de laquelle seront générées les valeurs potentielles des effets aléatoires. Rappelons qu'il s'agit ici de la loi normale centrée de variance $\sigma_k^{2[t]}$.

Notons $\xi_i^{[m]}$ la dernière valeur générée à partir de la distribution de $\xi_i|Y_i, Z_{ik} = 1$ en la valeur courante $\theta_k^{[t]}$. La probabilité d'accepter la nouvelle valeur ξ_i^* générée à partir de la distribution h s'écrit :

$$\rho(\xi_i^{[m]}, \xi_i^*) = \min\left\{1, \frac{f(\xi_i^*|y_i, z_{ik} = 1, \theta_k^{[t]})h(\xi_i^{[m]})}{f(\xi_i^{[m]}|y_i, z_{ik} = 1, \theta_k^{[t]})h(\xi_i^*)}\right\}$$

où le second terme se simplifie par :

$$\begin{aligned}
\frac{f(\xi_i^*|y_i, z_{ik} = 1, \theta_k^{[t]})h(\xi_i^{[m]})}{f(\xi_i^{[m]}|y_i, z_{ik} = 1, \theta_k^{[t]})h(\xi_i^*)} &= \frac{f(\xi_i^*|y_i, z_{ik} = 1, \theta_k^{[t]})f(\xi_i^{[m]}|z_{ik} = 1, \theta_k^{[t]})}{f(\xi_i^{[m]}|y_i, z_{ik} = 1, \theta_k^{[t]})f(\xi_i^*|z_{ik} = 1, \theta_k^{[t]})} \\
&= \frac{f(y_i|\xi_i^*, z_{ik} = 1, \theta_k^{[t]})f(\xi_i^*|z_{ik} = 1, \theta_k^{[t]})f(\xi_i^{[m]}|z_{ik} = 1, \theta_k^{[t]})}{f(y_i|\xi_i^{[m]}, z_k = 1, \theta_k^{[t]})f(\xi_i^{[m]}|z_{ik} = 1, \theta_k^{[t]})f(\xi_i^*|z_{ik} = 1, \theta_k^{[t]})} \\
&= \frac{f(y_i|\xi_i^*, z_{ik} = 1, \theta_k^{[t]})}{f(y_i|\xi_i^{[m]}, z_{ik} = 1, \theta_k^{[t]})}.
\end{aligned}$$

Finalement, l'étape de Metropolis-Hastings nécessite uniquement de connaître la distribution de Y_i conditionnellement à ξ_i sachant que l'individu i appartient à la classe $\mathcal{C}_k$. Or, sachant que l'individu i appartient à la classe $\mathcal{C}_k$, c'est justement sur la loi de Y_i conditionnellement à ξ_i qu'est formulée l'hypothèse de distribution (cf section 3.5.2).

3.5.5.3 L'algorithme proposé

L'algorithme proposé peut maintenant se résumer de la façon suivante à l'itération $[t]$:

1. On génère pour $i = 1, \ldots, I$ et $k = 1, \ldots, K$:

   - M valeurs $\xi_i^{[1]}, \ldots, \xi_i^{[M]}$ à partir de la distribution de $\xi_i | Y_i, Z_{ik} = 1$ en la valeur courante $\theta_k^{[t]}$ des paramètres par l'algorithme de Metropolis-Hastings précédent pour approcher les espérances conditionnelles de la fonction $Q^*(\theta, p | \theta^{[t]}, p^{[t]})$ par :

$$\begin{aligned} E_{Ck}^{[t]}(\xi_i^2) &\simeq \frac{1}{M} \sum_{m=1}^{M} \xi_i^{[m]2} \\ E_{Ck}^{[t]}\big[\exp(-u_{ij}\xi_i)\big] &\simeq \frac{1}{M} \sum_{m=1}^{M} \exp(-u_{ij}\xi_i^{[m]}) \end{aligned}$$

   - N valeurs $\xi_i^{[1]}, \ldots, \xi_i^{[N]}$ à partir de la distribution connue de $\xi_i | Z_{ik} = 1$ en la valeur courante $\theta_k^{[t]}$ des paramètres pour approcher les densités marginales par :

$$\begin{aligned} f_k(y_i | \theta_k^{[t]}) &= f(y_i | z_{ik} = 1, \theta_k^{[t]}) \\ &= \int \prod_{j=1}^{n_i} f(y_{ij} | \xi_i, z_{ik} = 1, \theta_k^{[t]})\, f(\xi_i | z_{ik} = 1, \theta_k^{[t]})\, d\xi_i \\ &\approx \frac{1}{N} \sum_{n=1}^{N} \Big[\prod_{j=1}^{n_i} f(y_{ij} | \xi_i^{[n]}, z_{ik} = 1, \theta_k^{[t]}) \Big] \end{aligned}$$

   et obtenir ensuite une approximation des probabilités a posteriori $t_k^{[t]}(y_i)$.

2. On maximise ensuite la fonction $Q^*(\theta, p | \theta^{[t]}, p^{[t]})$ pour obtenir de nouvelles valeurs des paramètres $\theta^{[t+1]}$ et $p^{[t+1]}$.

3.6 Des résultats de simulation

3.6.1 Préliminaires

Afin d'évaluer le comportement de l'algorithme de type MCEM développé en section 3.5.5, nous nous intéressons dans un premier temps à son utilisation dans le cadre d'un mélange de L2M. En effet, dans ce cas, les performances de l'algorithme de type MCEM peuvent facilement être comparées à celles de l'algorithme EM.

Par simplicité, nous considérons ici un mélange à 2 classes. Nous fixons le nombre d'individus statistiques $I = 100$ et le nombre de répétitions par individu $J = 6$. Les proportions du mélange sont fixées à $p_1 = 0.6$ et $p_2 = 0.4$. Les variances des effets

TAB. 3.9 – Résultats d'estimation des paramètres obtenus par les algorithmes EM et MCEM dans le cas gaussien sur 100 simulations

		EM		MCEM	
Valeurs simulées		moyenne	écart-type	moyenne	écart-type
	$p_1 = 0.6$	0.6006	0.0171	0.6006	0.0170
$\mathcal{C}_1$	$\beta_1 = -2$	-1.9872	0.1141	-1.9873	0.1140
	$\sigma_1^2 = 0.2$	0.1994	0.1201	0.1991	0.1166
	$p_2 = 0.4$	0.3994	0.0171	0.3994	0.0170
$\mathcal{C}_2$	$\beta_2 = 2$	2.0289	0.1737	2.0292	0.1733
	$\sigma_2^2 = 0.8$	0.7657	0.2897	0.7605	0.2855
	$\tau^2 = 2$	2.0210	0.1340	2.0216	0.1339

aléatoires sont $\sigma_1^2 = 0.2$ et $\sigma_2^2 = 0.8$ et la variance résiduelle est $\tau^2 = 2$. Nous considérons un seul paramètre d'effet fixe par composant : $\beta_1 = -2$ et $\beta_2 = 2$. Le tableau (3.9) donne les moyennes et les écart-types des estimations obtenus par les algorithmes EM et MCEM sur 100 jeux de données.

Le tableau (3.9) montre clairement que les résultats obtenus par l'algorithme de type MCEM sont semblables à ceux obtenus par l'algorithme EM. Cependant, nous mettons ici en évidence l'inconvénient notable de l'algorithme de type MCEM qui est l'allongement du temps de calcul qu'il engendre, vu le nombre important de simulations qu'il faut réaliser à chaque itération de l'algorithme. Pour donner un ordre d'idée, le temps de calcul de l'algorithme EM implémenté sous Splus est ici de quelques minutes contre plusieurs heures pour l'algorithme de type MCEM implémenté en C.

3.6.2 Comparaison des deux méthodes proposées

Dans un second temps, nous présentons des résultats de simulations afin d'évaluer la capacité des deux méthodes proposées à estimer correctement les paramètres d'un mélange de modèles exponentiels mixtes lien logarithme. Comme précédemment, nous considérons un mélange à 2 classes. Nous fixons le nombre d'individus statistiques $I = 100$. Les proportions du mélange sont fixées à $p_1 = 0.6$ et $p_2 = 0.4$ et les variances des effets aléatoires sont prises égales à $\sigma_1^2 = 0.2$ et $\sigma_2^2 = 0.8$. Nous considérons ici aussi un seul

paramètre d'effet fixe par composant.

Afin d'étudier le comportement des méthodes d'estimation selon le niveau de séparation des composants, nous générons des données selon :

- le modèle (A) défini par : $\beta_1 = -3$ et $\beta_2 = 3$,
- le modèle (B) défini par : $\beta_1 = -1$ et $\beta_2 = 1$.

Enfin, dans le but d'étudier l'impact du nombre de répétitions sur la qualité des estimations, pour chaque modèle (A) et (B), nous faisons également varier le nombre J de répétitons par individu : nous prenons $J = 4$ et $J = 8$.
Les tableaux (3.10) et (3.11) donnent les moyennes et les écart-types des estimations obtenus par les deux méthodes sur 100 simulations générées à partir des différents modèles. Dans les différents tableaux, nous désignons par "Linear." la méthode basée sur la linéarisation du modèle et par "MCEM" l'algorithme de type MCEM. Le tableau (3.12) donne la moyenne des taux de bon classement calculés en utilisant la méthode du maximum a posteriori définie en section 3.4.5 à partir des estimations $\hat{\beta}$ et $\hat{\sigma}^2$ obtenues.

À partir du tableau (3.10), on constate que les estimations obtenues sont en moyenne très proches des valeurs simulées quelle que soit la méthode utilisée avec, néanmoins, des résultats sensiblement meilleurs pour l'algorithme de type MCEM (estimations légèrement plus proches des valeurs simulées et meilleure précision). Notons que, de façon générale, la précision des estimations diminue quand la variance de l'effet aléatoire augmente. En effet, les valeurs des écart-types augmentent avec la variance de l'effet aléatoire. Bien entendu, on note également que les résultats s'améliorent quand le nombre de répétitions augmente.

Le tableau (3.11) montre que les résultats obtenus lorsque les composants du mélange sont moins séparés sont globalement moins bons pour les deux méthodes. Cependant, ils restent très raisonnables. Les remarques faites à partir du tableau (3.10) restent encore valables. Notons néanmoins, dans ce cas, un impact plus marqué du nombre de répétitions sur la qualité des estimations.

Pour finir, à l'aide du tableau (3.12), nous pouvons faire les remarques suivantes.

- On constate un pourcentage moyen de bon classement satisfaisant pour les deux méthodes.

TAB. 3.10 – Résultats d'estimation des paramètres obtenus par les deux méthodes proposées sur 100 simulations des modèles définis par $\beta_1 = -3$ et $\beta_2 = 3$

		Modèle (A) $J = 4$		Modèle (A') $J = 8$	
Valeurs simulées		Linear.	MCEM	Linear.	MCEM
	$p_1 = 0.6$	0.6017	0.5999	0.6002	0.6001
		(0.0039)	(0.0023)	(0.0012)	(0.0010)
$\mathcal{C}_1$	$\beta_1 = -3$	-2.9990	-2.9945	-3.0154	-3.0057
		(0.0878)	(0.0815)	(0.0921)	(0.0817)
	$\sigma_1^2 = 0.2$	0.2166	0.1960	0.1986	0.1977
		(0.1227)	(0.0834)	(0.0749)	(0.0645)
	$p_2 = 0.4$	0.3983	0.4001	0.3998	0.3999
		(0.0039)	(0.0023)	(0.0012)	(0.0010)
$\mathcal{C}_2$	$\beta_2 = 3$	3.0181	3.0042	3.0105	3.0121
		(0.1765)	(0.1641)	(0.1555)	(0.1588)
	$\sigma_2^2 = 0.8$	0.7419	0.7953	0.7798	0.7893
		(0.2490)	(0.2588)	(0.2455)	(0.2280)

TAB. 3.11 – Résultats d'estimation des paramètres obtenus par les deux méthodes proposées sur 100 simulations des modèles définis par $\beta_1 = -1$ et $\beta_2 = 1$

		Modèle (B) $J = 4$		Modèle (B') $J = 8$	
Valeurs simulées		Linear.	MCEM	Linear.	MCEM
	$p_1 = 0.6$	0.6634	0.6536	0.6481	0.6330
		(0.1340)	(0.0851)	(0.0856)	(0.0797)
$\mathcal{C}_1$	$\beta_1 = -1$	-0.8908	-0.9334	-0.9485	-0.9601
		(0.1845)	(0.1494)	(0.1599)	(0.1371)
	$\sigma_1^2 = 0.2$	0.2760	0.2376	0.2523	0.2330
		(0.2251)	(0.1466)	(0.1324)	(0.1119)
	$p_2 = 0.4$	0.3366	0.3464	0.3519	0.3670
		(0.1340)	(0.0851)	(0.0856)	(0.0797)
$\mathcal{C}_2$	$\beta_2 = 1$	1.2909	1.2202	1.2036	1.1504
		(0.5317)	(0.3047)	(0.3426)	(0.2895)
	$\sigma_2^2 = 0.8$	0.5437	0.6173	0.6195	0.6795
		(0.4568)	(0.3591)	(0.3257)	(0.3274)

TAB. 3.12 – Moyennes des taux de bon classement (en %) obtenues par les deux méthodes proposées sur 100 simulations

		Modèle (A) $\beta = (-3, 3)$ $J = 4$	Modèle (A') $\beta = (-3, 3)$ $J = 8$	Modèle (B) $\beta = (-1, 1)$ $J = 4$	Modèle (B') $\beta = (-1, 1)$ $J = 8$
Linear.	$\mathcal{C}_1$	99.98	100.00	93.37	96.38
	$\mathcal{C}_2$	99.55	99.95	67.87	76.95
MCEM	$\mathcal{C}_1$	99.96	100.00	96.15	96.52
	$\mathcal{C}_2$	99.92	99.97	73.17	79.95

- Bien entendu, plus la variance de l'effet aléatoire augmente, plus le taux de bon classement se détériore.
- De façon générale, le taux de bon classement augmente avec le nombre de répétitions.
- On note que les taux de bon classement obtenus par l'algorithme de type MCEM sont sensiblement meilleurs avec une différence plus marquée dans le cas où les composants sont moins séparés.

3.6.3 Remarques

La mise en place de l'algorithme de type MCEM en pratique soulève plusieurs interrogations et nécessite, à ce titre, de faire quelques remarques. Notons, tout d'abord, que cet algorithme est présenté ici pour un mélange de modèles exponentiels mixtes mais il est facilement adaptable quelque soit le mélange de GL2M considéré.
Dans cet algorithme, une étape de Metropolis-Hastings est introduite à chaque étape de l'algorithme EM afin de générer, pour chaque unité statistique i, des effets aléatoires selon la distribution conditionnelle de ξ_i sachant Y_i dans chaque composant $\mathcal{C}_k$. Dans cette étape de Metropolis-Hastings, le nombre d'itérations à partir duquel nous pouvons supposer que les échantillons simulés sont distribués suivant la distribution d'intérêt fait débat. On parle d'étape de burn-in. Divers auteurs se sont intéressés à la question, mais selon Robert et Casella [50], il n'existe pas de méthode optimale pour choisir la longueur du burn-in. Dans nos simulations, nous avons fait ce choix de façon empirique. Après plusieurs essais, nous avons finalement fixer systématiquement l'étape de burn-in à 500 itérations.
En ce qui concerne le nombre d'itérations effectuées à chaque étape de Metropolis-Hastings, nous avons effectué différentes simulations en faisant varier ce nombre. Finalement, les résultats présentés ici ont été obtenus en effectuant 4000 itérations à chaque étape de Metropolis-Hastings. Une seconde façon de procéder peut être envisagée. Elle consiste à augmenter le nombre d'itérations de l'algorithme de Metropolis-Hastings au fur et à mesure des itérations de l'algorithme EM. Cette approche est notamment utilisée par McCulloch [39].
Généralement dans les méthodes MCMC, il est également nécessaire de tenir compte des impératifs d'indépendances entre les valeurs simulées. On parle d'étape de thinning. Gelman et al. [22] proposent de ne garder les estimations que toutes les j itérations. Après plusieurs essais qui n'ont pas révélé de différences significatives entre les différents résultats obtenus, nous avons choisi par simplicité de conserver la totalité des valeurs générées par

l'algorithme de Metropolis-Hastings pour approximer les espérances mises en jeu à l'étape E de l'algorithme.

3.7 Discussion

Dans ce chapitre, nous considérons une nouvelle classe de modèles que sont les mélanges de GL2M. Nous proposons deux méthodes d'estimation des paramètres de ces modèles : l'algorithme de type MCEM qui peut s'appliquer à n'importe quel mélange de GL2M, et la méthode basée sur une linéarisation du modèle développée dans le cadre bien précis d'un mélange de modèles exponentiels mixtes.

Les simulations effectuées dans le cadre du mélange de modèles exponentiels mixtes montrent que ces deux méthodes ont un comportement globalement satisfaisant. Elles indiquent également que les estimations obtenues par l'algorithme de type MCEM sont sensiblement meilleures que celles obtenues par la méthode basée sur la linéarisation du modèle. Cette différence de comportement s'observe plus particulièrement dans les situations délicates. Il s'agit par exemple du cas où les composants du mélange sont moins séparés ou encore le cas où le nombre de répétitions par unité statistique est faible. Néanmoins, il est important de mettre en évidence l'inconvénient notable de l'algorithme de type MCEM développé qui est l'allongement du temps de calcul qu'il engendre vu le nombre important de simulations qui nécessitent d'être réalisées à chaque itération. En pratique, cet algorithme a ainsi nécessité d'être implémenté en langage C afin de réduire les temps de calculs et les rendre acceptables. Par exemple, pour les simulations présentées dans ce chapitre, les temps de calculs sont compris entre 7h et 9h. Concernant l'algorithme de type MCEM, nous souhaitons également souligner que, même si cet algorithme semble bien se comporter en pratique, nous n'avons pas établi de résultat théorique de convergence à ce jour.
Notons que contrairement à l'algorithme de type MCEM, la méthode basée sur la linéarisation du modèle est très simple à implémenter et très rapide. Pour les mêmes simulations, les temps de calculs sont ici de quelques minutes. Néanmoins, cette méthode est basée sur une propriété spécifique à la loi exponentielle. Elle n'est donc applicable que dans le cadre particulier des mélanges de modèles exponentiels mixtes.

Pour finir, notons que la comparaison des résultats obtenus dans le cas des mélanges de GLM, L2M et GL2M n'est pas chose aisée. Nous n'avons présenté dans le cas GL2M

que des résultats concernant des mélanges à deux composants. Cependant, même en considérant le même nombre de composants dans les différentes situations, il est difficile d'évaluer correctement l'impact de l'introduction des effets aléatoires et de la généralisation de la loi sur la qualité des estimations. En effet, d'une part, il n'est pas facile de comparer les différents niveaux de séparation des composants des cas gaussien et exponentiel, les échelles ne sont pas les mêmes. D'autre part, la taille des échantillons considérés est difficilement comparable entre les situations à effet fixe et à effet aléatoire. Dans le cadre des GLM, on dispose d'une seule observation par individu conduisant, pour n individus, à un échantillon de n observations indépendantes. Par contre, dans le cadre des L2M et GL2M, pour n individus, on dispose de n p-uplets indépendants constitués des observations répétées. Il n'est donc pas évident de mettre en correspondance le nombre d'observations dans ces deux situations. Néanmoins, il semble, au vu des résultats, que la prise en compte des répétitions soit une source d'information importante pour la validation des classes. Quelques simulations à ce sujet nous laissent penser que la qualité des estimations dans les mélanges de L2M est de même ordre que celle dans les mélanges de GL2M.

Conclusion

De nouvelles façons d'envisager l'explication d'une variable observée en modélisation statistique ont vu le jour grâce au développement des modèles à effets aléatoires. Les domaines d'application de ces modèles sont nombreux et variés, ce qui explique sans doute la multitude des travaux effectués sur le sujet. Nous nous sommes intéressés dans ce travail à la classe des modèles linéaires généralisés à effets aléatoires.

Dans le chapitre 1, nous avons considéré la question de l'estimation des paramètres d'effets fixes et des composantes de la variance de tels modèles. Dans le cas d'un modèle linéaire, l'estimation des paramètres par maximum de vraisemblance peut se faire par des méthodes classiques telles que l'algorithme EM. Dès lors que le modèle n'est plus linéaire, i.e. que l'hypothèse gaussienne ne s'avère plus adéquate, ces méthodes ne peuvent plus être appliquées. Nous avons vu, par exemple, que l'espérance conditionnelle intervenant dans l'étape E de l'algorithme EM n'admet plus d'expression analytique simple. Des solutions très variées ont émergé pour permettre une estimation dans les modèles linéaires généralisés à effets aléatoires.
Dans ce chapitre, nous sommes revenus sur certaines de ces méthodes. Nous nous sommes plus particulièrement intéressés aux approches basées sur une linéarisation du modèle. Nous avons présenté la méthode développée par Schall [51] ainsi que plusieurs autres approches. Ces dernières se sont avérées être différentes démarches permettant d'aboutir aux mêmes équations. Nous avons également présenté une méthode s'appuyant sur une démarche différente basée sur la simulation et développée par McCulloch [39] : la méthode MCEM. Elle permet d'approcher, par des méthodes de Monte Carlo, l'espérance conditionnelle intervenant dans l'étape E de l'algorithme EM via un algorithme de Metropolis-Hastings. L'explicitation de ces différentes méthodes a été nécessaire pour les chapitres suivants.

Dans le chapitre 2, nous avons cherché à développer des critères de sélection de modèles pour les modèles linéaires généralisés à effets aléatoires en nous appuyant sur les critères d'information usuels basés sur la vraisemblance. Notre attention s'est portée plus précisément sur l'adaptation des critères AIC et BIC bien que les approches que nous avons développées soient applicables à n'importe quel critère basé sur la vraisemblance.
À partir de deux méthodes d'estimation basées sur une linéarisation du modèle : la méthode proposée par Schall [51] utilisable pour un GL2M quelconque et la méthode "Gumbel" développée dans le cas particulier d'un modèle exponentiel mixte lien logarithme, nous avons proposé un critère construit sur la vraisemblance marginale du modèle linéarisé obtenu à la convergence de la procédure d'estimation.
Les simulations effectuées dans le cadre du modèle exponentiel mixte montrent que les critères développés ont globalement un comportement semblable aux critères d'information utilisés dans le cas gaussien en ce qui concerne la sélection aussi bien de la structure d'effet fixe que de l'effet aléatoire. Les simulations nous ont également permis de mettre en lumière des différences de comportement selon la linéarisation utilisée, la linéarisation de Schall ayant tendance à sur-sélectionner le modèle à effets aléatoires.
Nous avons aussi cherché à comparer ces approches basées sur la linéarisation à une troisième approche basée sur l'approximation par Monte Carlo de la vraisemblance marginale. Nous avons pu constater que la linéarisation n'avait pas d'effet négatif sur la sélection et qu'elle était plus facile à mettre en oeuvre quelque soit la complexité de la structure des effets aléatoires. Il serait maintenant intéressant de travailler davantage sur le terme de pénalité et de chercher à définir un critère dont la pénalité serait peut-être davantage appropriée aux modèles à effets aléatoires. Certains auteurs ont commencé récemment à s'y intéresser. Citons en particulier Vaida et Blanchard [64].

Dans le chapitre 3, nous avons défini une nouvelle classe de modèles que sont les mélanges finis de modèles linéaires généralisés à effets aléatoires. Cette classe de modèles permet d'introduire la notion d'hétérogénéité au sein des modèles linéaires généralisés à effets aléatoires. L'outil principal de ce chapitre est l'algorithme EM.
Nous avons développé deux méthodes d'estimation des paramètres : un algorithme de type MCEM utilisable pour un mélange de GL2M quelconque et une méthode basée sur une linéarisation du modèle développée dans le cas particulier d'un mélange de modèles exponentiels mixtes. Ces deux méthodes sont des adaptations de l'algorithme EM per-

mettant de contourner les problèmes liés à l'utilisation directe de ce dernier.
Les simulations effectuées dans le cadre du mélange de modèles exponentiels mixtes montrent que les méthodes développées ont un comportement globalement satisfaisant. Elles montrent également que l'algorithme de type MCEM fournit des estimations sensiblement meilleures que la méthode basée sur la linéarisation du modèle. Cependant, nous n'avons pas fait d'étude théorique de cet algorithme. Une poursuite intéressante pourrait donc être d'étudier de façon plus approfondie les propriétés d'un tel algorithme.
De plus, les simulations nous ont permis de mettre en évidence un autre inconvénient de l'algorithme de type MCEM développé qui est son temps de calcul extrêmement long, vu le nombre important de simulations réalisées à chaque étape de l'algorithme. Pour éviter des simulations trop nombreuses, il serait intéressant de proposer une version intermédiaire utilisant la simulation via une approximation stochastique. Dans ce sens, une poursuite intéressante serait de chercher à adapter la méthode développée par Kuhn et Lavielle [32]. Dans leur article, Kuhn et Lavielle proposent une variante de l'algorithme SAEM en le combinant à une méthode de Monte Carlo par chaînes de Markov.
Pour finir, dans tout le chapitre 3, nous avons supposé que le nombre de composants du mélange était connu, ce qui n'est pas toujours le cas en pratique. Une autre perspective intéressante à ce travail serait d'établir des méthodes de choix de modèles dans le but de déterminer le nombre approprié de composants du mélange. Ces méthodes de sélection semblent pouvoir être développées rapidement en s'appuyant sur les travaux réalisés au chapitre 2.

Bibliographie

[1] H. Akaike. Information theory as an extension of the maximum likelihood principle. In B. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Akademiai Kiado, Budapest, 1973.

[2] H. Akaike. A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, 19 :716–723, 1974.

[3] D. A. Anderson and M. Aitkin. Variance components models with binary response : interviewer variability. *Journal of the Royal Statistical Society B*, 47(2) :203–210, 1985.

[4] R. L. Anderson and T. A. Bancroft. *Statistical Theory in Research*. McGraw-Hill Book Company, New-York - Toronto - London, 1952.

[5] C. Biernacki, G. Celeux, and G. Govaert. An improvement of the NEC criterion for assessing the number of clusters arising from a mixture. *Pattern Recognition Letters*, 20 :267–272, 1999.

[6] H. Bozdogan. Model selection and Akaike's information criterion (AIC) : the general theory and its analytical extensions. *Psychometrika*, 52 :345–370, 1987.

[7] N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88 :9–25, 1993.

[8] K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference : a practical information-theoretic approach*. Springer-Verlag, New York, seconde edition, 2002.

[9] G. Celeux, D. Chauveau, and J. Diebolt. On stochastic versions of the EM algorithm. Technical Report RR-2514, INRIA, 1995.

[10] G. Celeux and J. Diebolt. L'algorithme SEM : un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densités. *Revue de Statistique Appliquée*, 34(2) :35–52, 1986.

[11] G. Celeux and J. Diebolt. The EM and the SEM algorithms for mixtures : statistical and numerical aspects. *Cahiers du CERO*, 32 :135–151, 1990.

[12] G. Celeux and J. Diebolt. A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and stochastics reports*, 41 :119–134, 1992.

[13] G. Celeux, O. Martin, and C. Lavergne. Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling*, 5 :243–267, 2005.

[14] G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13 :195–212, 1996.

[15] G. C. Chow. A comparaison of the information and posterior probability criteria for model selection. *Journal of Economics*, 16 :21–33, 1981.

[16] P. F. Craigmile and D. M. Titterington. Parameter estimation for finite mixtures of uniform distributions. *Communications in Statistics - Theory and Methods*, 26(8) :1981–1995, 1997.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39 :1–38, 1977.

[18] B. Engel and A. Keen. A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, 48(1) :1–22, 1994.

[19] L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1) :342–368, 1985.

[20] J. L. Foulley, C. Delmas, and C. Robert-Granié. Méthodes du maximum de vraisemblance en modèle linéaire mixte. *Journal de la Société Française de Statistique*, 143(1-2) :5–52, 2002.

[21] O. C. Gaudoin, C. Lavergne, and J. L. Soler. A generalized geometric deeutrophication software reliability model. *IEEE Transactions on Reliability*, 43 :536–541, 1994.

[22] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Texts in Statistical Science Series. Chapman & Hall, London, seconde edition, 1995.

[23] D. Gianola and J. L. Foulley. Sire evaluation for ordered categorical data with a threshold model. *Genetic Selection Evolution*, 15(2) :201–224, 1983.

[24] P. J. Green. Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, 55(3) :245–259, 1987.

[25] H. O. Hartley and J. N. K. Rao. Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54 :93–108, 1967.

[26] D. A. Harville. Maximum-likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72 :320–340, 1977.

[27] V. Hasselblad. Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association*, 64 :1459–1471, 1969.

[28] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57 :97–109, 1970.

[29] C. R. Henderson. Best linear unbiaised estimators and prediction under a selection model. *Biometrics*, 31 :423–447, 1975.

[30] C. M. Hurvich and C. L. Tsai. Regression and times series model selection in small samples. *Biometrika*, 76 :297–307, 1989.

[31] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430) :773–795, 1995.

[32] E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM : Probability and Statistics*, 8 :115–131, 2004.

[33] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22 :79–86, 1951.

[34] E. Lebarbier and T. Mary-Huard. Le critère BIC : fondements théoriques et interprétation. Technical Report RR-5315, INRIA, 2004.

[35] Y. Lee and J. A. Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society B*, 58 :619–678, 1996.

[36] H. Linhart and W. Zucchini. *Model selection*. John Wiley & Sons, New York, 1986.

[37] C. L. Mallows. Some comments on C_p. *Technometrics*, 15 :661–675, 1973.

[38] P. McCullagh and J. A. Nelder. *Generalized linear models*. Chapman and Hall, London, 1989.

[39] C. E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92 :162–170, 1997.

[40] G. McLachlan and D. Peel. *Finite mixture models*. Wiley-Interscience, New York, 2000.

[41] A. D. R. McQuarrie and C-L. Tsai. *Regression and time series model selection*. World Scientific Publishing Company, Singapore, 1998.

[42] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society A*, 135 :370–384, 1972.

[43] H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58 :545–554, 1971.

[44] K. Pearson. Contributions to the theory of mathematical evolution. *Philosophical transactions of the royal society of London*, A 185 :71–110, 1894.

[45] J. C. Pinheiro and D. M. Bates. *Mixed-effects models in S and S-plus*. Statistics and Computing Series. Springer-Verlag, New York, 2000.

[46] A. E. Raftery. Bayesian model selection in social research (with discussion). *Sociological Methodology*, 25 :111–195, 1995.

[47] C. R. Rao and J. Kleffe. *Estimation of variance components and Applications*. North Holland series in statistics and probability. Elsevier, Amsterdam, 1988.

[48] R. A. Redner and H. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26 :195–239, 1984.

[49] B. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.

[50] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New-York, seconde edition, 2004.

[51] R. Schall. Estimation in generalized linear models with random effects. *Biometrika*, 78 :719–727, 1991.

[52] G. Schwarz. Estimationg the dimension of a model. *Annals of Statistics*, 6 :461–464, 1978.

[53] S. R. Searle, G. Casella, and C. E. McCulloch. *Variance components*. John Wiley & Sons, 1992.

[54] L. R. Shaeffer and J. W. Wilton. Methods of sire evaluation for calving ease. *Journal of Dairy Science*, 59 :544–551, 1976.

[55] R. P. Sherman, Y. K. Ho, and S. R. Dalal. Conditions for convergence of Monte Carlo EM sequences with an application to product diffusion modeling. *Econometrics Journal*, 2(2) :248–267, 1999.

[56] R. Shibata. Statistical aspects of model selection. In J. C. Willems, editor, *From data to model*, pages 215–240, London, 1989. Springer-Verlag.

[57] R. Stiratelli, N. Laird, and J. H. Ware. Random-effects models for serial observations with binary response. *Biometrics*, 40 :961–971, 1984.

[58] C. J. Stone. Local asymptotic admissibility of a generalization of Akaike's model selection rule. *Annals of the Institute of Statistical Mathematics*, 34 :123–133, 1982.

[59] N. Sugiura. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics. Theory and Methods A7*, pages 13–26, 1978.

[60] W. A. Thompson. The problem of negative estimates of variance components. *Annals of Mathematical Statistics*, 33 :273–289, 1962.

[61] L. Tierney and J. B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81 :82–86, 1986.

[62] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions*. John Wiley & Sons Ltd, Chichester, 1985.

[63] C. Trottier. *Estimation dans les modèles linéaires généralisés à effets aléatoires*. PhD thesis, Institut National Polytechnique de Grenoble, 1998.

[64] F. Vaida and S. Blanchard. Conditional Akaike information for mixed-effects models. *Biometrika*, 92(2) :351–370, 2005.

[65] M. Wedel and W. S. DeSarbo. A mixture likelihood approach for generalized linear models. *Journal of Classification*, 12 :21–55, 1995.

[66] G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the Poor Man's data augmentation algorithms. *Journal of the American Statistical Association*, 85 :699–704, 1990.

[67] R. Wolfinger. Generalized linear mixed models : a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48 :233–243, 1993.

[68] C. J. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11 :95–103, 1983.

[69] S. L. Zeger and M. R. Karim. Generalized linear models with random effects : a Gibbs sampling approach. *Journal of the American Statistical Association*, 86 :79–86, 1991.

Résumé : Ce travail est consacré à l'étude des modèles linéaires généralisés à effets aléatoires (GL2M). Dans ces modèles, sous une hypothèse de distribution normale des effets aléatoires, la vraisemblance basée sur la distribution marginale du vecteur à expliquer n'est pas, en général, calculable de façon formelle. Dans la première partie de notre travail, nous revisitons différentes méthodes d'estimation non exactes par le biais d'approximations réalisées à différents niveaux selon les raisonnements. La deuxième partie est consacrée à la mise en place de critères de sélection de modèles au sein des GL2M. Nous revenons sur deux méthodes d'estimation nécessitant la construction de modèles linéarisés et nous proposons des critères basés sur la vraisemblance marginale calculée dans le modèle linéarisé obtenu à la convergence de la procédure d'estimation. La troisième et dernière partie s'inscrit dans le cadre des modèles de mélanges de GL2M. Les composants du mélange sont définis par des GL2M et traduisent différents états possibles des individus. Dans le cadre de la loi exponentielle, nous proposons une méthode d'estimation des paramètres du mélange basée sur une linéarisation spécifique à cette loi. Nous proposons ensuite une méthode plus générale puisque s'appliquant à un mélange de GL2M quelconques. Cette méthode s'appuie sur une étape de Metropolis-Hastings pour construire un algorithme de type MCEM. Les différentes méthodes développées sont testées par simulations.

Mots-clés : Modèles linéaires généralisés, Effets aléatoires, Estimation, Sélection de modèle, Modèle de mélange, Algorithme EM, Algorithme de Metropolis-Hastings.

Abstract : This work focuses on generalized linear mixed models (GL2M). In these models, considering a gaussian hypothesis for the random effects distribution, the likelihood based on the marginal distribution of the response cannot be derived in closed form. In the first part of this work, we critically review parameter estimation methods using different kinds of approximations. The second part focuses on model selection for GL2Ms. Two parameter estimation methods are revisited, both leading to iterative model linearisations. We propose simple model selection criteria adapted from classical information criteria and based on the linearised model obtained once the algorithm has converged. In the third and last part, the analysis of mixture models of GL2Ms is considered. The mixture components are defined by GL2Ms and correspond to different possible states of the statistical units. For a mixture of exponential mixed models, we propose a method using a linearisation specific to this distribution. We also propose a second and more general approach which uses a Metropolis-Hastings step to allow construction of an MCEM algorithm. This method can be used for mixtures of any GL2Ms. The different developed methods are tested by simulations.

Discipline : Mathématiques appliquées et applications des mathématiques.

Unité d'accueil : Equipe Probabilités et Statistique, Institut de Mathématiques et de Modélisation de Montpellier, UMR CNRS 5149, CC 051, Université Montpellier II, 34095 Montpellier Cedex 05.