Analyse de données structurées au travers de modèles linéaires mixtes

Benjamin Heuclin, Statisticien, UR AIDA, CIRAD

23/12/2024

Contents

Ce document a été écrit en collaboration avec Sandrine Le Squin (PalmElit), Marie Denis (Cirad) et Albert Flori (Cirad).

Prérequis : Nous supposons que le lecteur maitrise la théorie sur les modèles linéaires. Si besoin, nous invitons le lecteur à se référer aux livres suivants :

- Generalized, Linear, and Mixed Models, Charles E. McCulloch, Shayle R. Searle
- LINEAR MODELS IN STATISTICS, Alvin C. Rencher and G. Bruce Schaalje (https://www.utstat.toronto.edu/~brunner/books/LinearModelsInStatistics.pdf)

Les modèles linéaires mixtes permettent d'analyser des données structurées qui impliquent une corrélation entre les observations ou plusieurs sources d'aléa. Parmi toutes les structures possibles, on retrouve, entre autres, les données groupées, répétées, longitudinales ou encore spatiales :

- **Des données groupées :** Ce sont des données regroupées par une caractéristique commune.
- **Exemple n°1 « Essai agronomique en blocs complets randomisés »:** Dans un essai agronomique où l'on souhaite tester l'effet de 4 traitements sur un caractère d'intérêt (le rendement par exemple), plusieurs répétitions de chaque traitement sont réalisées. Cela peut nécessiter alors une taille de parcelle assez conséquente et l'hypothèse d'homogénéité du sol n'est alors plus assurée. Il est alors possible de diviser la parcelle en blocs de placettes dans lesquelles on retrouvera une fois chaque traitement. Chaque bloc fait alors office de répétition. L'idée est de constituer des blocs de telle sorte que les placettes intra-bloc soient les plus homogènes et les blocs soient les plus hétérogènes entre eux. Ainsi on retrouve dans les données une structuration par bloc avec potentiellement une corrélation intra-bloc et une variabilité inter-bloc.

Bloc 1	Bloc 2	Bloc 3	Bloc 4
T3	T3	T2	T1
T4	T1	T4	T3
T1	T2	T3	T4
T2	T4	T1	T2

Figure 1: Plan du design en blocs complets de l'exemple 1

- **Exemple n°2 « Essai agronomique génétique » :** Dans un essai agronomique sur les plantes, des géniteurs sont alors croisés entre eux pour engendrer une nouvelle génération d'individus sur lesquels on va mesurer une caractéristique d'intérêt (rendement). Les individus peuvent alors être regroupés en fonction de leurs pedigrees. Ainsi les observations faites sont alors plus ou moins corrélées en fonction du degré d'apparentement entre les individus.
- **Exemple n°3 « Essai clinique multicentrique » :** Dans le cadre d'un essai clinique multicentrique visant à évaluer l'efficacité d'une nouvelle procédure chirurgicale, une sélection aléatoire est effectuée parmi un ensemble de cliniques afin de garantir une représentativité de l'ensemble des établissements existants. Cette sélection vise à tester la procédure dans un échantillon diversifié de cliniques. Ainsi, la procédure est mise à l'épreuve dans plusieurs établissements, chacun appliquant les mêmes techniques, mais pouvant apporter de légères modifications au protocole en fonction de leurs préférences spécifiques.
- **Des données répétées :** Ce sont des données mesurées où pour chaque individu, plusieurs mesures sont réalisées dans différentes conditions.

- **Exemple n°4 "Données cliniques répétées" :** Dans un essai clinique sur la santé cardiaque, des médecins ont décidé de tester des patients dans différentes conditions. Tout d'abord, il a été testé la fréquence cardiaque au repos, après la montée d'un étage par les escaliers et après une course de 10 min sur un tapis roulant.
- **Des données longitudinales :** Ce sont des données répétées au cours du temps pour chaque individu.
- **Exemple n°5 "Données cliniques longitudinales" :** L'étude EVA (Epidémiologie du Vieillissement Artériel) avait pour objectif principal de mieux comprendre les processus du vieillissement normal et les conséquences du processus de vieillissement sur les fonctions cognitives et les pathologies cardiovasculaires. Entre juin 1991 et juin 1993, 1389 sujets volontaires (574 hommes et 815 femmes), âgés de 59 à 71 ans, ont été recrutés pour participer à cette étude. Le but de l'analyse était d'étudier l'évolution du sélénium dans le temps et de rechercher les facteurs associés à cette évolution. Dans EVA, le sélénium plasmatique a été mesuré à l'inclusion, 2 ans, 6 ans et 9 ans.
- **Exemple n°6 "Séries temporelles des oiseaux d'eau d'Hawaii" :** Données publiées dans *Time series analysis of Hawaiian waterbirds. Analysing ecological data, Reed, Elphick, Zuur, Ieno, Smith (2007), Springer.* Elles consistent en des données d'abondance de trois espèces d'oiseaux, mesurées sur trois îles d'Hawaii de 1956 à 2003. On a donc plusieurs séries temporelles.
- **Des données spatiales :** Ce sont des données mesurées dans l'espace et qui peuvent donc présenter une corrélation spatiale
- **Exemple n°7 « Essai agronomique spatial » :** Dans un essai sur le palmier à huile plantée au Brésil en 2010, on souhaite étudier l'effet de 8 croisements d'hybrides plantés selon des densités variant de 103 arbres/ha à 143 arbres/ha sur le diamètre des troncs à 12 ans. Il y a 2 parcelles élémentaires de 64 arbres par croisement, dont 49 arbres utiles (on ne tient pas compte des palmiers en bordure de parcelle). Les parcelles sont regroupées en blocs incomplets qui sont eus même regroupés en répétitions complètes (plan Alpha-Lattice). Dans chaque parcelle élémentaire, les densités varient continûment selon un dispositif inventé par Nelder : les lignes s'écartent de plus en plus en allant d'un côté à l'autre de la parcelle et sur chaque ligne, les arbres s'écartent de plus en plus en parcourant la ligne. Seuls 5 à 10 palmiers ont été mesurés par parcelle élémentaire.

Dans chacune de ces situations, les données sont dites structurées et les observations présentes des corrélations entre elles. L'hypothèse d'indépendance des résidus du modèle linéaire ne tient donc plus. Les modèles linéaires mixtes permettent de prendre en compte de telles structures au travers d'effets aléatoires et la considération de matrice de covariance complexe.

1 Définition d'un modèle linéaire mixte

Les modèles linéaires mixtes, également connus sous le nom de modèles à effets mixtes, sont une classe de modèle qui englobe les modèles linéaires et les modèles à effets aléatoires (sans effet fixe). Le principe même des modèles linéaires mixtes se trouve dans leur capacité à intégrer à la fois des composantes fixes représentant les effets moyens des variables explicatives et des composantes aléatoires permettant de prendre en compte la structuration des données.

1.1 Définition d'un effet fixe

Un effet fixe peut-être soit une variable quantitative (ou numérique), soit une variable qualitative nominale ou ordinale (un facteur) avec un nombre de niveaux fini. La prise en compte d'une variable en effet fixe permet d'en estimer son effet sur la variable réponse. Ce sont généralement des variables qui sont contrôlées par l'expérimentateur.

1.2 Définition d'un effet aléatoire

Un effet aléatoire est toujours une variable qualitative/facteur. C'est donc une variable groupante, les niveaux du facteur définisse des groupes d'observations. On suppose qu'il a un nombre infini de niveaux dans la population tout entière et que nous n'en observons qu'un sous-échantillon aléatoire (nous observons un nombre fini de niveau parmi un ensemble infini). Généralement, ce n'est pas un facteur que l'on contrôle et dont on cherche à estimer l'effet de chacun de ses niveaux sur la variable réponse, excepté en génétique où l'on s'intéresse souvent à l'effet spécifique de chaque individu, famille ou lignée sur le caractère étudié. Toutefois on suppose qu'il joue un rôle sur la variabilité de la variable réponse (variabilité inter-groupe non-nulle du facteur aléatoire). L'estimation de cette part de variabilité plutôt que de l'effet de chacun des niveaux présente l'avantage de pouvoir généraliser les résultats quelle que soit la valeur prise par ce facteur aléatoire, même pour des valeurs non observées. Cet objectif d'inférence à des cas non observées nécessite de fixer des conditions sur la distribution des effets aléatoires. En pratique, on supposera qu'un effet aléatoire suivra une loi normale centrée avec une certaine structure de covariance. Cette dernière permet ainsi d'introduire de la corrélation entre les observations. Nous reviendrons sur ce point dans la suite.

Effets aléatoires emboîtés ou croisés

Lorsque l'on souhaite introduire plusieurs effets aléatoires dans le modèle, ils peuvent soit être emboîtés, soit être croisés (indépendants).

Les effets aléatoires emboîtés caractérisent une structure de hiérarchie. Par exemple, sur un essai agronomique en milieu paysan, les fermes sont emboîtées dans des villages qui sont emboîtés dans des régions. Dans un tel dispositif, on pourra prendre en effet aléatoire la ferme, le village et la région.

A l'inverse, les effets aléatoires croisés font référence à des facteurs qui sont associés à plusieurs niveaux de regroupement ou de traitement simultanément. Si l'on reprend l'Exemple n°1 « essai agronomique en blocs complets randomisés » en supposant qu'il est mené sur 10 ans. On peut alors prendre l'année en effet aléatoire pour capturer les variabilités de rendement d'une année sur l'autre dues au conditions climatiques différentes. Il y a aussi les blocs en effets aléatoires. Ces 2 effets sont croisés, car pour chaque année, on retrouve tous les blocs et chaque bloc est présent sur les 10 années. Ils sont indépendants.

1.3 Comment choisir entre effet fixe et aléatoire ?

Différents éléments peuvent orienter la décision de mettre un facteur en effet fixe ou aléatoire. Nous avons déjà énoncé quelques éléments plus haut que nous allons reprendre.

Les règles suivantes sont en faveur d'une prise en compte en effet aléatoire :

- Si le facteur retranscrit une structure de corrélation (corrélation existante entre les observations issues d'un même niveau du facteur)
- Reprenons l'Exemple n°1, où une parcelle est divisée en plusieurs blocs qui sont eux-mêmes divisés en placettes. Nous avons alors une corrélation entre les différentes placettes d'un même bloc, et cela, pour chacun des blocs.
- Si on peut supposer qu'il y a une infinité de niveaux dans la population, mais que l'on n'en observe qu'un échantillon aléatoire.
- Prenons l'exemple n°1, où l'étude se déroule dans quelques cliniques du pays. Ces différentes cliniques ont été choisies au hasard pour tester la nouvelle technique chirurgicale et vont représenter un échantillon. En effet, les cliniques choisies vont servir à tester la technique afin que celle-ci soit généralisable à l'ensemble des cliniques.
- Si la part de variabilité capturé par le facteur importe plus que l'effet de ses niveaux sur la variable et que l'on souhaite généraliser les résultats quelle que soit la valeur prise par ce facteur.

- Prenons l'exemple n°1, où l'étude se déroule dans quelques cliniques du pays. Ces différentes cliniques ont été choisies au hasard pour tester la nouvelle technique chirurgicale et vont représenter un échantillon. En effet, les cliniques choisies vont servir à tester la technique afin que celle-ci soit généralisable à l'ensemble des cliniques.
- Si l'effet de chacun des niveaux nous importe, mais que le nombre de niveaux est conséquent alors il peut être nécessaire de prendre en compte le facteur en effet aléatoire. Effectivement, dans le cas où le facteur serait considéré comme fixe, il y a $p-1$ effet à estimer (p étant le nombre de niveaux) tandis que dans le cas d'un effet aléatoire il y a généralement qu'un seul paramètre de variance à estimer. Cela est plus économique en degré de liberté. En génétique par exemple, on cherche à estimer l'effet des géniteurs sur le caractère d'intérêt pour pouvoir ensuite sélectionner les plus performants. Les géniteurs sont généralement traités comme aléatoires parce qu'ils sont nombreux. De plus, considérer les géniteurs comme effet aléatoires permet de prendre en compte l'apparentement entre eux (via une structure de variance-covariance particulière) grâce à l'utilisation d'un pedigree. Bien que les observations soient souvent réalisées sur les descendants issus de croisements, elles peuvent aussi concerner directement les géniteurs eux-mêmes dans certains cas.

Cependant, il faut bien préciser que les différentes règles présentées ci-dessus sont des règles de décisions. Notons toutefois que lorsque qu'une variable qualitative (facteur) présente peut de niveau (<=3), alors il est déconseillé de la prendre en effet aléatoire car la variance associée à cet effet aléatoire ne pourra pas être corectement estimée, est ce même si le nombre de répétitions par niveau est élevé. La Figure 2 représente l'erreur d'estimation de la variance d'un effet aléatoire (données simulées) en fonction du nombre de niveaux du facteur (abscisse) et du nombre de répétitions par niveau (ordonné). On peut observer que plus le nombre de niveaux va être important, alors meilleure est l'estimation. Le nombre de répétitions quant à lui n'a que peut d'influence sur la qualité d'estimation.

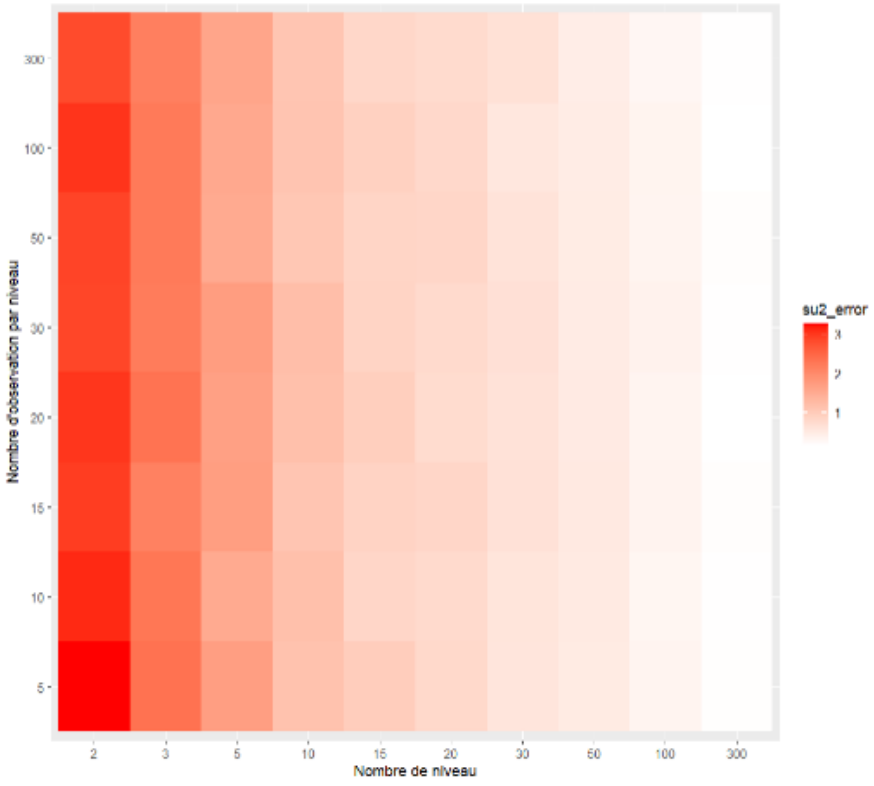


Figure 2: Figure 2 : Graphique de l'erreur (norme L1) d'estimation en fonction du nombre de niveaux du facteur

2 Formulation et Hypothèses

Un modèle linéaire mixte peut être écrit soit sous forme individuelle soit sous forme matricielle. Avant de détailler les formulations, commençons par introduire quelques notations :

- n le nombre total d'observations
- p le nombre de variables à effet fixe
- D le nombre d'effets aléatoires
- Q_d le nombre de niveaux observés du $d^{\text{ème}}$ facteur aléatoire, $d = 1, \ldots, D$
- Q le nombre total de niveaux observés pour l'ensemble des facteurs aléatoires, $Q = \sum_{d=1}^{D} Q_d$

2.1 Forme individuelle

Un modèle linaire mixte peut s'écrire sous forme individuelle, c'est-à-dire à l'échelle d'une observation. Cependant, cette écriture dépend de l'expérience/données et n'est donc pas généralisable.

Prenons l'Exemple n°1 "Essai agronomique en blocs complets randomisés". Ici, une observation correspond au rendement d'une placette qui a subi le traitement i et cette placette se trouve dans le bloc j. Ainsi, le modèle linéaire mixte peut s'écrire au niveau individuel de la façon suivante :

$$Y_{ij} = T_i + u_j + \varepsilon_{ij}$$

Où

- i est l'indice des traitements
- j est l'indice des blocs
- Y_{ij} est l'observation du rendement de la placette du bloc j ayant subi le traitement i
- T_i est l'effet fixe du $i^{\text{ème}}$ traitement
- u_j est l'effet aléatoire du $j^{\text{ème}}$ bloc, tel que le vecteur de l'ensemble des effets des blocs $u = (u_1, u_2, u_3, u_4)^T \sim N_4(0, \sigma_u^2 \mathbf{I}_4)$
- ε_{ij} le résidu associé à l'observation du traitement i du bloc j, tel que le vecteur de l'ensemble des résidus $\varepsilon = (\varepsilon_{11}, \varepsilon_{21}, \ldots, \varepsilon_{44})^T \sim N_{16}(0, \sigma_E^2 \mathbf{I}_{16})$

2.2 Forme matricielle

N'importe quel modèle mixte peut s'écrire sous la forme matricielle suivante :

$$Y = X\beta + ZU + \varepsilon \quad (1)$$

Où :

- Y est la variable réponse de longueur n où n est le nombre total d'observation,
- X est la matrice des effets fixes, pouvant contenir une première colonne de 1 pour modéliser un intercept,

- β est le vecteur des effets fixes associées à X. Il représente les effets moyens des variables explicatives sur la variable réponse,
- Z est la matrice de design permettant de relier chacune des observations à un ou plusieurs niveaux de l'effet aléatoire U (Nous revenons sur la construction de ces matrices dans la suite),
- U est le vecteur d'effet aléatoire,
- ε est le vecteur des résidus.

Cette formulation est plus concise et fait intervenir des matrices de design pour décrire les relations entre les variables explicatives et la variable réponse.

Les hypothèses du modèle linéaire mixte :

- Le vecteur d'effet aléatoire U est supposé suivre une loi normale centrée en zéro : $U \sim N(0, G)$.
- Le vecteur de résidus est supposé suivre une loi normale centrée en zéro : $\varepsilon \sim N(0, R)$.
- L'indépendance entre le vecteur d'effet aléatoire U et les résidus ε

On reviendra plus tard sur les différentes formes des matrices G et R que l'on abordera (voir section 3.1.1 et 3.2).

Les composantes de la variance :

On appellera composantes de la variance l'ensemble des paramètres inconnus dans la partie aléatoire (effet aléatoire (G) et résidus (R)) qui sont à estimer. Ce sont en l'occurrence des paramètres de variance ou de covariance.

2.2.1 Construction de Z et de U

Le vecteur d'effet aléatoire U peut contenir les effets des niveaux d'un seul facteur aléatoire ou être la concaténation des effets des niveaux de plusieurs facteurs aléatoires (ex : modèle génétique père/mère). Il est de longueur D ou D est le nombre total de niveaux observés de l'ensemble des facteurs aléatoires. La matrice Z est une matrice de design ou d'incidence constituée de 0 et de 1. Elle a pour objectif de relier chacune des observations aux niveaux des facteurs aléatoires qui lui corresponde. Ainsi, elle est de taille $n \times D$. Ainsi, la $i^{\text{ème}}$ ligne de cette matrice permet de relier l'observation numéro i aux niveaux du/des facteur.s aléatoire.s. La $j^{\text{ème}}$ colonne permet, quant à elle, de relier l'ensemble des observations au $j^{\text{ème}}$ niveau de l'effet aléatoire U.

Exemple dans le cas où l'on a un seul facteur aléatoire :

Reprenons l'Exemple n°1 « Essai agronomique en blocs complets randomisés ».

Ainsi notre effet aléatoire bloc U serait donné par le vecteur : $U = (u_1, u_2, u_3, u_4)^T$, où u_q est l'effet du bloc q. Pour construire la matrice Z, il nous faut avoir le tableau des données pour voir comment sont organisées les données et ainsi pouvoir relier chacune des observations au bloc dans laquelle elle a été recueillie (voir Tableau 1). Elle doit être de taille 16 x 4. Sur la première colonne, on retrouvera des 1 si les observations proviennent du premier bloc et 0 sinon. Il en est de même pour les colonnes 2, 3 et 4 avec les blocs 2, 3 et 4 respectivement. Cela nous donne la matrice suivante :

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Ainsi le produit ZU nous redonne la colonne « n° de bloc » dans le tableau des données de cet exemple.

2.2.2 Ecriture conditionnelle (*within group*) et marginale (*population*)

Ecriture conditionnelle (*within group*) :

Lorsque l'on travaille en conditionnelle, nous sommes alors conditionnés à l'effet aléatoire U, cela implique que nos résultats vont dépendre des effets des niveaux de U qui ont été estimés.

$$Y|U \sim \mathcal{N}(X\beta + ZU, R),$$
$$U \sim \mathcal{N}(0, G).$$

Ecriture marginale (*population*) :

L'écriture marginale consiste à intégrer l'effet aléatoire U. On a alors accès à la distribution de Y sans avoir à connaître l'effet des niveaux de U :

$$Y \sim \mathcal{N}(X\beta, ZGZ' + R).$$

Pour redémontrer ce résultat, on peut repartir du modèle sous sa forme matricielle (eq. 1) :

$$Y = X\beta + ZU + \varepsilon$$

avec

$U \sim \mathcal{N}(0, G)$ et $\varepsilon \sim \mathcal{N}(0, R)$.

D'après la propriété de transformation affine, on sait que le produit ZU suit également une loi normale de la forme :

$$ZU \sim \mathcal{N}(0, ZGZ')$$

D'après la propriété de somme de deux vecteurs gaussiens indépendants, on sait que $ZU + \varepsilon$ suit également une loi normale de la forme :

$$ZU + \varepsilon \sim \mathcal{N}(0, ZGZ' + R)$$

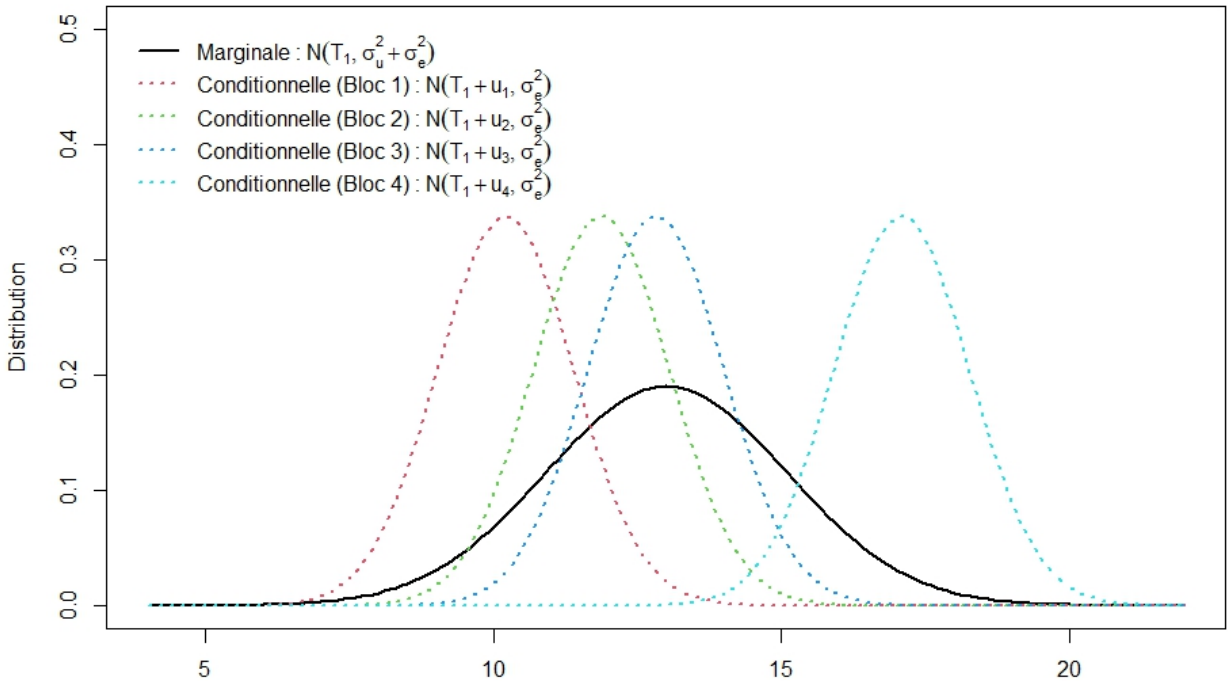
Enfin, d'après la propriété de transformation affine, on peut maintenant en déduire la loi marginale de Y :

$$Y = X\beta + ZU + \varepsilon \sim \mathcal{N}(X\beta, ZGZ' + R)$$

L'écriture marginale permet de généraliser les résultats au niveau de la population. Il est ainsi possible de faire une prédiction pour un nouvel individu dont son niveau d'effet aléatoire serait inconnu ou non-observé sur les données d'ajustement. Cependant, la variance de la distribution marginale est plus élevée que celle de la distribution conditionnelle ce qui implique un intervalle de confiance de la prédiction plus large (on perd en précision).

L'écriture conditionnelle est, quant à elle, plus précise, mais nécessite la prédiction de l'effet des niveaux du facteur aléatoire (BLUP), ce qui peut demander des ressources de calcul plus importantes si D est très grand.

Le graphique suivant illustre cela sur l'Exemple n°1 : "Essai agronomique en blocs complets randomisés".



On peut voir que les distributions conditionnelles sont plus resserrées que la distribution marginale due au fait que les variances conditionnelles sont plus faibles.

Dans le cas où l'on travaillerait avec plusieurs effets aléatoires, on peut en intégrer certains et conditionner le modèle par rapport à d'autres.

3 Prise en compte de la structuration des données au travers d'un modèle mixte

Les effets aléatoires ainsi que les résidus permettent de prendre en compte la structuration et l'hétérogénéité des données au travers des matrices de variance-covariance G et R.

Nous présentons ici les structures de corrélation principales. Cette liste est non exhaustive et nous invitons le lecteur à se référer au livre de Pinheiro & Bates (Mixed-effects models in S and S-PLUS) et à la documentation de la "proc mixed" de SAS.

Commençons par la structuration des observations introduite par un effet aléatoire.

3.1 Structuration introduite par un effet aléatoire

On se place dans le cas où on suppose les résidus *iid* soit $R = \sigma_e^2 I_n$ et où l'on a qu'un seul effet aléatoire.

Dans cette section, nous introduisons les paternes de matrice de variance-covariance pour un effet aléatoire de type Identité, Connue à une constante près, Diagonale, Générale (*Unstructured*), et par Bloc.

3.1.1 Paterne Identité

C'est le cas le plus simple et par défaut dans n'importe quels logiciels/packages. La matrice G est alors un multiple d'une matrice identité :

$$U \sim N(0, \sigma_u^2 Id_Q).$$

Ce type d'effet aléatoire permet d'introduire une corrélation constante entre toutes les observations ayant le même niveau aléatoire. Si on reprend l'Exemple n°1 "Essai agronomique en bloc complet" :

$$Y_{ij} = T_i + u_j + \varepsilon_{ij}$$

alors on a une covariance non-nulle entre les observations provenant d'un même bloc j :

$$cov(y_{ij}, y_{\tilde{i}j}) = \sigma_u^2, \forall i \neq \tilde{i}$$

et 0 pour toutes observations ne provenant pas d'un même bloc (niveau aléatoire) :

$$cov(y_{ij}, y_{\tilde{i}\tilde{j}}) = 0, \forall j \neq \tilde{j}.$$

A partir de l'écriture marginale du modèle multivariée, on peut facilement obtenir la matrice de covariance introduite par l'effet aléatoire.

Toujours dans le cas de l'exemple 2, on a :

$$Y \sim \mathcal{N}(X\beta, \sigma_u^2 ZZ' + \sigma_e^2 I),$$

avec X la matrice de design du facteur étudié (le traitement T), Z la matrice de design de l'effet aléatoire Bloc (u). La matrice de covariance introduite par l'effet aléatoire bloc est alors donnée par le produit $\sigma_u^2 ZZ'$:

σ_u^2	σ_u^2	σ_u^2	σ_u^2	0	0	0	0	0	0	0	0	0	0	0	0
σ_u^2	σ_u^2	σ_u^2	σ_u^2	0	0	0	0	0	0	0	0	0	0	0	0
σ_u^2	σ_u^2	σ_u^2	σ_u^2	0	0	0	0	0	0	0	0	0	0	0	0
σ_u^2	σ_u^2	σ_u^2	σ_u^2	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	σ_u^2	σ_u^2	σ_u^2	σ_u^2	0	0	0	0	0	0	0	0
0	0	0	0	σ_u^2	σ_u^2	σ_u^2	σ_u^2	0	0	0	0	0	0	0	0
0	0	0	0	σ_u^2	σ_u^2	σ_u^2	σ_u^2	0	0	0	0	0	0	0	0
0	0	0	0	σ_u^2	σ_u^2	σ_u^2	σ_u^2	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	σ_u^2	σ_u^2	σ_u^2	σ_u^2	0	0	0	0
0	0	0	0	0	0	0	0	σ_u^2	σ_u^2	σ_u^2	σ_u^2	0	0	0	0
0	0	0	0	0	0	0	0	σ_u^2	σ_u^2	σ_u^2	σ_u^2	0	0	0	0
0	0	0	0	0	0	0	0	σ_u^2	σ_u^2	σ_u^2	σ_u^2	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	σ_u^2	σ_u^2	σ_u^2	σ_u^2
0	0	0	0	0	0	0	0	0	0	0	0	σ_u^2	σ_u^2	σ_u^2	σ_u^2
0	0	0	0	0	0	0	0	0	0	0	0	σ_u^2	σ_u^2	σ_u^2	σ_u^2
0	0	0	0	0	0	0	0	0	0	0	0	σ_u^2	σ_u^2	σ_u^2	σ_u^2

On retrouve une structure de 4 blocs de 4x4 représentant la covariance entre les parcelles issues d'un même bloc.

3.1.2 Paterne G connue à une constante près

La matrice G s'écrit alors $\sigma_u^2 A$, avec A une matrice de variance covariance connue (matrice de parenté par exemple en génétique) :

$$U \sim N(0, \sigma_u^2 A)$$

Ici, les niveaux de l'effet aléatoire ne sont pas indépendants. La matrice de covariance introduite par l'effet aléatoire est alors donnée par $\sigma_u^2 ZAZ'$

On retrouve ce type d'effet aléatoire en génétique pour la prise en compte du pedigree. Cela permet d'introduire une corrélation plus ou moins forte entre les individus en fonction de leur proximité dans l'arbre généalogique. Lorsque la matrice Z est égale à l'identité, c'est-à-dire lorsque chaque niveau du facteur aléatoire correspond directement à chaque individu observé et que le pedigree est disponible pour tous les individus, on parle d'effet aléatoire individuel. C'est notamment le cas dans le modèle dit "animal" couramment utilisé en génétique.

3.1.3 Paterne Diagonale

La matrice de covariance des effets aléatoires G est alors une matrice diagonale. Chaque niveau de l'effet aléatoire à sa propre variance, et il n'y a pas de covariance entre les différents niveaux aléatoires.

La matrice de covariance introduite par l'effet aléatoire est alors donnée par ZGZ' :

$$
\begin{bmatrix}
\sigma_1^2 & 0 & 0 & 0 \\
0 & \sigma_2^2 & 0 & 0 \\
0 & 0 & \sigma_3^2 & 0 \\
0 & 0 & 0 & \sigma_4^2
\end{bmatrix}
$$

Figure 3: Matrice de la structure diagonale

$\sigma^2_{u_1}$	$\sigma^2_{u_1}$	$\sigma^2_{u_1}$	$\sigma^2_{u_1}$	0	0	0	0	0	0	0	0	0	0	0	0
$\sigma^2_{u_1}$	$\sigma^2_{u_1}$	$\sigma^2_{u_1}$	$\sigma^2_{u_1}$	0	0	0	0	0	0	0	0	0	0	0	0
$\sigma^2_{u_1}$	$\sigma^2_{u_1}$	$\sigma^2_{u_1}$	$\sigma^2_{u_1}$	0	0	0	0	0	0	0	0	0	0	0	0
$\sigma^2_{u_1}$	$\sigma^2_{u_1}$	$\sigma^2_{u_1}$	$\sigma^2_{u_1}$	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	$\sigma^2_{u_2}$	$\sigma^2_{u_2}$	$\sigma^2_{u_2}$	$\sigma^2_{u_2}$	0	0	0	0	0	0	0	0
0	0	0	0	$\sigma^2_{u_2}$	$\sigma^2_{u_2}$	$\sigma^2_{u_2}$	$\sigma^2_{u_2}$	0	0	0	0	0	0	0	0
0	0	0	0	$\sigma^2_{u_2}$	$\sigma^2_{u_2}$	$\sigma^2_{u_2}$	$\sigma^2_{u_2}$	0	0	0	0	0	0	0	0
0	0	0	0	$\sigma^2_{u_2}$	$\sigma^2_{u_2}$	$\sigma^2_{u_2}$	$\sigma^2_{u_2}$	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	$\sigma^2_{u_3}$	$\sigma^2_{u_3}$	$\sigma^2_{u_3}$	$\sigma^2_{u_3}$	0	0	0	0
0	0	0	0	0	0	0	0	$\sigma^2_{u_3}$	$\sigma^2_{u_3}$	$\sigma^2_{u_3}$	$\sigma^2_{u_3}$	0	0	0	0
0	0	0	0	0	0	0	0	$\sigma^2_{u_3}$	$\sigma^2_{u_3}$	$\sigma^2_{u_3}$	$\sigma^2_{u_3}$	0	0	0	0
0	0	0	0	0	0	0	0	$\sigma^2_{u_3}$	$\sigma^2_{u_3}$	$\sigma^2_{u_3}$	$\sigma^2_{u_3}$	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	$\sigma^2_{u_4}$	$\sigma^2_{u_4}$	$\sigma^2_{u_4}$	$\sigma^2_{u_4}$
0	0	0	0	0	0	0	0	0	0	0	0	$\sigma^2_{u_4}$	$\sigma^2_{u_4}$	$\sigma^2_{u_4}$	$\sigma^2_{u_4}$
0	0	0	0	0	0	0	0	0	0	0	0	$\sigma^2_{u_4}$	$\sigma^2_{u_4}$	$\sigma^2_{u_4}$	$\sigma^2_{u_4}$
0	0	0	0	0	0	0	0	0	0	0	0	$\sigma^2_{u_4}$	$\sigma^2_{u_4}$	$\sigma^2_{u_4}$	$\sigma^2_{u_4}$

3.1.4 Paterne Générale

La matrice de covariance des effets aléatoires G est une matrice pleine inconnue, c'est-à-dire que tous les éléments de la matrice peuvent être non nuls. Cela signifie que chaque niveau aléatoire peut avoir une variance distincte, et toutes les covariances possibles entre les différents niveaux aléatoires sont également prises en compte. C'est la structure la plus flexible et la plus générale, permettant de modéliser des relations complexes entre les niveaux d'un effet aléatoire, mais elle nécessite l'estimation d'un plus grand nombre de paramètres.

$$
\begin{bmatrix}
\sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\
\sigma_{21} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\
\sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{43} \\
\sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2
\end{bmatrix}
$$

Figure 4: Matrice de covariance générale

3.1.5 Paterne Bloc.

La matrice de covariance des effets aléatoires G est structurée par blocs sur la diagonale. Chaque bloc représente une sous-matrice pleine, indiquant qu'il y a des covariances entre certains niveaux d'effet aléatoire. Cette structuration est utile lorsque l'on souhaite prendre en compte une interaction entre plusieurs effets aléatoires. En effet, cela se traduit par un nouvel effet aléatoire dont la matrice de covariance est le produit de Kronecker entre les matrices de covariances des effets aléatoires en interactions. En génétique par exemple, lorsqu'on souhaite étudier une interaction génotype x environnement dans un essai multisites, on peut faire l'hypothèse que l'effet des génotypes peut ne pas être identique d'un environnement à un autre. Ainsi, on peut introduire un effet aléatoire Site (Environnement) ayant une matrice diagonale (une variance par site) en interaction avec un effet aléatoire génotype ayant une matrice de covariance connue à une constante près. Cette interaction donne un nouvel effet aléatoire ayant pour matrice de covariance :

$$\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix} \otimes A = \begin{bmatrix} \sigma_1^2 A & 0 & 0 \\ 0 & \sigma_2^2 A & 0 \\ 0 & 0 & \sigma_3^2 A \end{bmatrix}$$

.

Les librairies `sommer`, `BGLR`, `BGGE` permettent de facilement mettre en oeuvre ce type de modèle.

Voir section 4.2.2 sur livre de Pinheiro & Bates pour plus de détails.

3.2 Structuration au travers des résidus

3.2.1 Structure diagonale

Cette structure modélise des données où les résidus sont supposés indépendants, mais avec une variance hétérogène en fonction d'une variable groupante.

$$\begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix}$$

Figure 5: Matrice de la structure diagonale

Notons que d'autres formes d'hétérogénéité de la variance résiduelle sont possibles, mais nous ne rentrerons pas plus en détail sur ce point. Nous renvoyons le lecteur vers la section 5.2 du livre "*Mixed-Effects Models in S and S-PLUS*" de Pinheiro et Bates.

3.2.2 Structure Autorégressive d'ordre 1

Dans le cas d'une structuration temporelle des observations, il est possible de prendre en compte cette structuration au travers des résidus à l'aide d'une structuration de type autorégressive d'ordre 1 (AR(1)). L'idée est de relier les observations successives au cours du temps faites sur un même sujet. Ce modèle est défini par un paramètre corrélation, noté ρ, qui représente la force de la corrélation entre les observations successives.

On peut observer que la corrélation entre 2 observations diminue exponentiellement à mesure qu'elles s'éloignent dans le temps.

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Figure 6: Structure AR1

3.2.3 Structure spatiale

De nombreuses structures spatiales ont été proposées dans la littérature. Nous présentons ici les plus couramment proposées dans les différentes librairies abordées.

Deux classes de structures existent, les structures isotropiques pour lesquelles la dépendance spatiale est identique dans toutes les directions et les structures anisotropiques pour lesquelles la dépendance spatiale est plus forte dans une direction plutôt qu'une autre.

3.2.3.1 Structures spatiales isotropiques Pour nous aider à choisir entre les différentes structures possibles, il est conseillé de calculer le semi-variogramme empirique basé sur les distances entre les observations :

$$\gamma(d) = 0.5 * \sqrt{(\varepsilon_x - \varepsilon_y)^2}, \text{ pour tout } x, y \text{ tel que } d < distance(x, y) \leq d + \delta d$$

avec δd étant égal à environ $\frac{1}{20^{\text{ème}}}$ de la distance maximales entre observations et les ε étant les résidus obtenus par l'ajustement d'un modèle sans structure spatiale.

Les distances les plus utilisées sont la distance Euclidienne (norme L2), ou encore la distance de Manhattan (norme L1).

La distance Euclidienne, ou norme L2, est calculée comme la racine carrée de la somme des carrés des différences entre les coordonnées des points. Cette mesure reflète la distance "à vol d'oiseau" dans l'espace, fournissant une mesure directe de la séparation géométrique entre les points. En revanche, la distance de Manhattan, ou norme L1, est obtenue en additionnant les valeurs absolues des différences entre les coordonnées des points. Cette distance est souvent décrite comme la distance parcourue en suivant un chemin en angle droit, comme les rues d'une grille urbaine. Chacune de ces distances présente des avantages selon le contexte de l'analyse : la distance Euclidienne est utile pour capturer des différences globales dans un espace continu, tandis que la distance de Manhattan est souvent préférée dans des contextes où les déplacements se font le long d'axes prédéfinis ou lorsque les données sont naturellement alignées sur des grilles.

Modélisation par un modèle de variogram isotropic :

$$\gamma(d, a) = \sigma_0^2 + \sigma^2(1 - h(d, a))$$

avec h étant la fonction de corrélation, et σ_0^2, σ^2 et a (respectivement : effet de pépite, plateau et portée) étant des paramètres à estimer.

Les principales fonction de corrélation sont : - Exponentielle :

$$h(d, a) = exp(-d/a)$$

- Gaussien :

$$h(d, a) = exp[-(d/a)^2]$$

- Linéaire :

$$h(d, a) = (1 - d/a)I(d < a)$$

- Quadratique rationnelle :

$$h(d,a) = (d/a)^2/[1+(d/a)^2]$$

- Sphérique :

$$h(d,a) = 1 - [1 - 1.5(d/a) + 0.5(d/a)^3]I(d < a)$$

La matrice de covariance des résidus est alors construite de la façon suivante :

$$R = \sigma_e^2 * h(D,a)$$

avec D la matrice de distance entre les observations (norme L1 (Manhatan) ou L2 (Euclidienne) ou autre fonction de distance).

On peut noter que la structure basée sur la fonction Exponentielle est également connue sous le nom de *CAR* pour *Conditionnal AutoRegressiv*.

3.2.3.2 Structures spatiales anisotropiques Les structures anisotropiques supposent que la dépendance spatiale est plus forte dans une direction plutôt qu'une autre. Ainsi, on va supposer une structure sur les lignes, une sur les colonnes et possiblement une pour l'interaction ligne/colonne.

Là encore le calcul du semi-variogramme sur chacune des dimensions peut nous aider à choisir le bon type de structuration.

3.2.3.2.1 Double structuration autorégressive ligne/colonne Dans le cas d'observations espacés régulièrement sur une grille, on peut appliquer une structuration autorégressive sur les lignes et sur les colonnes (AR(1) $\otimes$ AR(1))

3.2.3.2.2 Structuration par P-spline à 2 dimensions ligne/colonne Toujours dans le cas d'observations espacées régulièrement sur une grille, on peut interpoler des P-spline à 2 dimensions sur les lignes et colonnes. Cela ne se fait pas directement sur les résidus, mais au travers d'un effet aléatoire additionnel. Cette approche est présentée par Lee et al (2013) et est proposer par la librairie sommer.

Notons que certaines librairies modélisent la corrélation spatiale et temporelle au travers d'un effet aléatoire indépendant des résidus (supposés iid). C'est notamment le cas d'INLA.

Notons également que la librairie INLA a été conçue pour l'analyse de données spatiales et propose d'autres méthodes non détaillées ici. Nous invitons le lecteur à se rapprocher de la documentation d'INLA.

–>

4 Packages R

Il existe plusieurs librairies sur R pour l'inférence de modèles mixtes. Ici, nous nous intéressons à 9 librairies. Chaque package sera brièvement présenté. Nous renvoyons le lecteur vers les notes pour chacun des packages en annexe B.

4.1 Package nlme

La librairie nlme (*Linear and Nonlinear Mixed Effects Models*) est une des librairies phares de R pour l'ajustement de modèle linéaire mixte. Elle se distingue des autres librairies par un large panel de structure de covariance possible tant sur les effets aléatoires que sur les résidus. Toutefois, il n'est pas possible d'inférer un effet aléatoire dont la matrice de covariance serait connue à une constante près, ne permettant pas de mettre en œuvre un modèle animal par exemple. De plus, nlme est limité aux variables réponses suivant une distribution normale et ne peut pas traiter directement des variables binomiales ou de comptage.

4.2 Package lme4

La librairie lme4 (*Linear Mixed-Effects Models using Eigen and S4*) l'autre librairie phare de R pour l'ajustement de modèle mixte. Contrairement à nlme elle ne permet pas de modéliser des structures de covariance complexes, mais permet de mettre en œuvre des modèles linéaires mixtes généralisés. La librairie lme4 s'accompagne de la libraire lmerTest pour, entre autres, pouvoir réaliser des ANOVA ou de la sélection de variables par approche de type stepwise.

4.3 Package lme4GS

La librairie lme4GS est une extension de lme4 pour sélection génomique. Ainsi, on peut y inférer un effet aléatoire dont la matrice de covariance serait connue à une constante près, tel que rencontré dans le modèle animal.

4.4 Package sommer

Le package sommer (*Solver for Mixed Model Equations in R*) est un outil principalement utilisé pour l'analyse de données génétiques et phénotypiques. Ce package se distingue par sa capacité à gérer des matrices d'apparentement, facilitant ainsi l'estimation des paramètres génétiques tels que l'héritabilité et les valeurs génétiques. Sommer offre une grande flexibilité avec des fonctions comme mmer (*multi-trait mixed model equation*) pour ajuster des modèles multi-traits, et intègre des méthodes avancées pour l'estimation et la prédiction, y compris les modèles de régression linéaire généralisée et les modèles de variance hétérogène.

4.5 Package BGLR et BGGE

BGLR (*Bayesian Generalized Linear Regression*), et BGGE (*Bayesian Genomic Gaussian Processes Regression*) sont des librairies bayésiennes utilisées dans la génomique statistique pour la prédiction des traits quantitatifs. Nottons que BGLR permet de faire de la sélection de variable. BGGE quant à lui est une version optimisée de BGLR pour la prise en compte d'interaction GxE. Il est ainsi plus rapide. Toutefois il ne permet pas de faire de la sélection de variables.

4.6 Package BRMS

BRMS (*Bayesian Regression Models using Stan*) est un package R pour la modélisation bayésienne des données, particulièrement utile pour les modèles mixtes. Les modèles mixtes, qui incluent à la fois des effets fixes et aléatoires, bénéficient grandement de l'approche bayésienne implémentée dans BRMS. Cela permet d'intégrer des incertitudes et des informations a priori dans les estimations des paramètres. BRMS facilite la mise en œuvre de modèles hiérarchiques et de structures de corrélation complexes, rendant l'analyse des modèles mixtes plus flexible et puissante.

BRMS (*Bayesian Regression Models using Stan*) est un package R pour la modélisation bayésienne des données, particulièrement utile pour les modèles mixtes. Ce package est basé sur le logiciel Stan et en offre une utilisation simplifiée. Il offre de larges possibilités de modélisation. Toutefois il peut être lent sur des modèles complexes et/ou des données volumineuses.

4.7 Package INLA

INLA (*Integrated Nested Laplace Approximations*) est une librairie bayesienne permettant de mettre en oeuvre des modèles complexes (génétiques, spatiaux, temporels, ...). Contraîrement aux autres libraries bayésiennes telles ques BGLR, BGGE et brms qui proposent une inférence par méthode MCMC, INLA

propose une inférence par approximation de Laplace. Cela permet une estimation rapide et précise des paramètres et des distributions *a posteriori*. Cela en fait une librairie appropriée pour les grands jeux de données.

4.8 Package AsReml

Le package AsReml, basé sur le logiciel ASReml, est spécialisé dans l'estimation des modèles mixtes linéaires, particulièrement prisé en génétique animale et végétale. Il est reconnu pour ses capacités à traiter de grands ensembles de données et des structures de covariance complexes, fournissant des estimations précises des composantes de variance et des effets aléatoires. AsReml est souvent choisi pour les analyses de performance animale et les essais variétaux en sélection génétique, en raison de son efficacité et de sa précision dans ces contextes.

4.9 En résumé

Le tableau suivant résume les avantages de chaque librairie :

	nlme	lme4	lme4GS	sommer	BGLR	BGGE	brms	INLA	AsReml
Effets aléatoires croisés		x	x	x	x	x	x	x	x
Matrice d'apparentement			x	x	x	x	x	x	x
Variance résiduelle hétérogène	x			x	x		x	x	
Structuration spatiale	x			x	x	x	x	x	x
Structuration temporelle					x	x	x	x	x
Variable réponse non gausienne		x			x		x	x	x
Sélection de variables	x	x			x		x		
Imputation de NA dans var. réponse					x		x	x	x
Imputation de NA dans var. explicatives							x	x	
Grande dimension								x	x

5 Bibliographies

Lee, D.-J., Durban, M., and Eilers, P.H.C. (2013). Efficient two-dimensional smoothing with P-spline ANOVA mixed models and nested bases

McCulloch, C. E., & Searle, S. R. (2004). Generalized, linear, and mixed models. John Wiley & Sons.

Pinheiro, J., & Bates, D. (2000). Mixed-effects models in S and S-PLUS. Springer science & business media.

Rencher, A. C., & Schaalje, G. B. (2008). Linear models in statistics. John Wiley & Sons.