Université de Montpellier
Faculté des Sciences

Collaboration I3M - CIRAD

Extension de la méthode "Régression Linéaire Généralisée sur Composantes Supervisées" (SCGLR) aux données groupées.

Rapport de stage de Master 2 Biostatistique,
2014/2015.

Jocelyn Chauvet

ENCADRANTS :
Mme. Catherine TROTTIER,
M. Xavier BRY, &
M. Frédéric MORTIER

Table des matières

Introduction Générale.

Un problème général, rencontré dans de très nombreux domaines, est de modéliser de façon robuste un ensemble de réponses aux types de distributions variés à partir d'un grand nombre de variables explicatives redondantes. Cette redondance implique une régularisation des coefficients du modèle. D'autre part, le désir d'obtenir un modèle non seulement prédictif mais véritablement explicatif pousse à rechercher des dimensions prédictives interprétables, s'appuyant typiquement sur un nombre assez grand de variables explicatives.

Ce problème est rencontré en particulier en écologique forestière, qui nécessite une modélisation complexe et robuste des communautés d'espèces végétales [1]. Nous décrivons ci-après l'importance de l'obtention d'une telle modélisation en écologie ainsi que les motivations qui ont été les nôtres dans ce contexte.

Plus des deux tiers des forêts tropicales de la planète se regroupent au sein de seulement trois pays : le Brésil, la République Démocratique du Congo et l'Indonésie. La biodiversité y est exceptionnelle tant sur le plan floristique que faunistique : on estime que l'ensemble de ces forêts recèlent au moins 75 % des espèces animales et végétales. De plus, elles apparaissent comme capitales sur plusieurs niveaux. En effet, dans un premier temps, elles assurent un stockage du carbone au niveau mondial. De plus, elles interviennent dans des phénomènes de régulations climatiques à un niveau plutôt régional : elles contribuent à refroidir la planète par leur capacité à favoriser la formation de nuages. Enfin, elles constituent, pour les communautés locales, des moyens essentiels de subsistance.

La préservation de ces forêts est donc essentielle. Cependant, la pression anthropique y est croissante, venant entre autre de l'industrie minière et du développement de différentes infrastructures. Ces forêts sont donc particulièrement intéressantes à observer et à étudier, afin de mieux percevoir dans quelle mesure les changements climatiques et la pression anthropique influencent la conversion de la surface forestière.

En particulier, une question importante qui se pose en écologie forestière tropicale est la modélisation de la structure des communautés d'espèces d'arbres au sein des plus grandes forêts tropicales du monde, dans le but de prévoir les impacts des changements globaux sur la composition floristique de ces peuplements forestiers. Ce stage, en collaboration avec le CIRAD et plus particulièrement avec les projets "CoForTips" et "CoForChange" s'inscrit pleinement dans cette problématique. Les données à disposition du CIRAD dans le cadre

de ces projets concernent les forêts du bassin du Congo, qui constituent le deuxième massif mondial des forêts denses tropicales humides. De nombreux relevés ont été effectués au sein de plusieurs concessions couvrant plusieurs pays de cette région. Une étape nécessaire est alors la mise au point d'outils de modélisation statistique et d'aide à la décision, qui permettront aux experts de mieux comprendre dans quelle mesure les communautés d'espèces floristiques sont tributaires des changements globaux, notamment sur les plans climatiques et anthropologiques. Dans l'idéal, des stratégies à moyen terme pourront alors être élaborées pour atténuer l'impact de ces changements, dont les conséquences pourraient être planétaires sur beaucoup d'aspects.

Globalement, les projets CoForTips et CoForChange ont pour but d'analyser ce qui est, afin de prédire ce qui pourrait être. Plus précisément, il s'agira pour nous de développer des modèles qui prennent appui sur de multiples covariables de natures très diverses pour modéliser et prédire les structures des communautés d'espèces d'arbres dans les forêts du bassin du Congo. Ces outils permettront ensuite aux spécialistes, à travers les scenarii possibles qu'ils peuvent imaginer sur les différentes covariables [2], de se faire une idée assez précise de leurs répercutions sur les abondances des espèces qui peuplent ces forêts.

Jusqu'à présent, dans cette situation, des modèles simples, tels que les modèles linéaires ou les modèles linéaires dans leur version généralisée étaient utilisés. Cependant, ils se heurtent à plusieurs écueils :

i) Avec ce type de modélisation, la prise en compte simultanée de plusieurs espèces n'est pas immédiate.

ii) De plus, la forte colinéarité ou redondance généralement présente au sein des covariables rend les analyses le plus souvent instables, et les qualités de prédiction se voient donc mécaniquement dégradées.

iii) Enfin, même si la stratégie est envisageable, nous ne souhaitons pas pratiquer une sélection de variables, car cela pourrait conduire à un appauvrissement du modèle.

Pour faire face aux distributions diverses des réponses et au nombre et à la redondance des variables explicatives, une méthode a été proposée par Bry et al. [4] : il s'agit de la "Régression Linéaire Généralisée sur Composantes Supervisées" (SCGLR). Cependant, l'hypothèse d'indépendance des mesures - réalisées sur plusieurs localisations - demeure encore dans ce modèle. Or, elles sont structurées en groupe dans l'espace, ce qui y introduit de la dépendance. Une extension du modèle initial est alors nécessaire, car les spécialistes ont besoin de modèles qui tiennent compte des groupes spatiaux au sein des données. Nous étudions ici la prise en compte de cette dépendance spatiale via l'injection, dans le modèle initial, d'un effet aléatoire.

Ce travail se décompose en plusieurs parties. Dans un premier temps, nous exposerons les éléments théoriques nécessaires à la compréhension de la méthode SCGLR dans le chapitre

(1), avant d'évoquer, dans le chapitre (2), les modèles linéaires mixtes simples et généralisés ainsi que les méthodes usuelles d'estimation associées. Ensuite, nous présenterons dans le chapitre (3) l'extension proposée du modèle initial. Pour cela, nous préciserons les arguments principaux qui ont permis l'adaptation de la méthode SCGLR aux données groupées, en y incluant les pseudo-codes généraux. Enfin, dans les chapitres (4) et (5), nous testerons l'algorithme proposé sur des données simulées, avant de présenter l'amélioration apportée sur données réelles.

Chapitre 1

Présentation de la méthode SCGLR.

Dans cette première partie, nous présentons les principaux aspects de la procédure SCGLR qui avait été élaborée pour répondre au problème posé, mais qui suppose l'indépendance des unités statistiques. Les références bibliographiques relatives à l'écriture de cette partie sont [3] et [4] pour l'exposition de la stratégie de "Régression Linéaire Généralisée sur Composantes Supervisées" et [5], [6], [7] et [8] pour ce qui est des Modèles Linéaires Généralisés et l'estimation des paramètres de ces modèles. L'objectif ici est d'exposer les différents outils précédemment mis au point, dont nous présenterons notre extension dans le chapitre 3.

1.1 Le Modèle.

Notre but est de modéliser et prédire q variables réponses $y^1, y^2, \ldots, y^q$ rassemblées dans un tableau noté Y, à l'aide de variables explicatives, de la manière la plus pertinente et la plus interprétable possible. Les réponses Y étant de nature différentes (continues, discrètes, catégorielles), nous adoptons un Modèle Linéaire Généralisé (GLM). En particulier, n désignant le nombre d'unités statistiques considérées, alors pour tout $k \in \{1, 2, \ldots, q\}$ et pour tout $i \in \{1, 2, \ldots, n\}$, la variable Y_i^k est supposée admettre une distribution issue de la famille exponentielle.

D'une manière générale, nous supposons avoir deux catégories de variables explicatives (et prédictives) bien distinctes, et qui joueront des rôles très différents lors de l'ajustement du modèle.

i) Dans un premier temps, nous supposons être en présence de variables préalablement traitées et/ou sélectionnées. Elles ont donc vocation à être peu ou pas redondantes et en général en petit nombre, afin justement de réduire les risques de redondance. Ces variables, notées $t^1, t^2, \ldots, t^r$ seront rassemblées dans un tableau noté T. Elles sont mises dans un tableau à part car les effets propres de chacune d'entre elles auront besoin d'être quantifiés.

ii) D'un autre côté, nous supposons détenir des variables explicatives possiblement très nombreuses et redondantes, qui sont supposées receler un petit nombre de dimensions explicatives latentes.. Elles seront notées quant à elles $x^1, x^2, \ldots, x^p$ et rassemblées dans un tableau X.

Il est important de noter que comme le tableau X contient de nombreuses variables a priori redondantes, il convient de mettre en œuvre une stratégie de réduction de dimension dans ce tableau. Cette réduction dimensionnelle s'effectuera par la détermination de composantes. A contrario, les variables présentes dans T ayant subi un pré-traitement adéquat, la réduction de la dimension est ici non-nécessaire et même non-souhaitable. A cet effet, SCGLR se conçoit comme une stratégie dont le but est de mêler une réduction de dimension au sein des covariables d'une part, et un ajustement par des GLM avec réponse multivariée d'autre part.

En outre, il est important de souligner que les composantes retenues pour X doivent satisfaire des conditions particulières. D'une part, elles doivent avoir un fort pouvoir prédictif vis-à-vis de Y. En conséquence, elles doivent permettre un bon ajustement du modèle de Y, tout en évitant le phénomène de sur-ajustement. D'autre part, leur interprétation doit être aisée, car toute composante sélectionnée a pour but principal d'être interprétée. Donc autant que possible, les composantes sélectionnées dans X ne doivent pas s'ajuster sur du "bruit", mais au contraire s'appuyer de façon assez nette sur un nombre suffisant de variables.

Nous voyons apparaître en filigrane le dilemme central avec lequel SCGLR tente de composer : la qualité de l'ajustement du modèle et la force structurelle des composantes explicatives dans X. De plus, pour éviter toute redondance au sein des composantes, elles seront contraintes à l'orthogonalité deux à deux, ce qui en facilitera aussi leur interprétation graphique.

1.2 Quelques éléments sur les Modèles Linéaires Généralisés.

Avant d'exposer plus en détails la stratégie décrite par SCGLR, nous allons rappeler les principales notions relatives aux GLM que nous utiliserons par la suite. Nous privilégierons en particulier les notations adoptées et les procédures algorithmiques retenues pour en estimer les paramètres.

1.2.1 Principes généraux et notations usuelles.

Nous allons tout d'abord rappeler les hypothèses relatives à la mise en œuvre d'un ajustement par GLM. En supposant avoir rassemblé les variables explicatives dans une matrice X de taille $n \times p$ et en supposant n'avoir qu'une seule variable à expliquer Y (de taille $n \times 1$), nous avons les éléments suivants :

i) Au niveau de la distribution de la variable à expliquer, elle doit appartenir à la famille exponentielle. Plus formellement, chacune des composantes du vecteur $Y = (Y_1, Y_2, \ldots, Y_n)'$ doit admettre une densité pouvant s'écrire :

$$f_{Y_i}(y_i, \theta_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \quad i = 1, \ldots, n. \tag{1.1}$$

Avec l'écriture précédente, θ_i est dit "paramètre canonique", ϕ un paramètre de dispersion et $a_i(\phi) = \frac{\phi}{w_i}$, où w_i représente le poids associé à l'observation i. De plus, les fonctions b et c, dérivables deux fois, sont communes aux composantes de Y et caractérisent la loi retenue.

ii) Ensuite, tout comme dans les modèles linéaires, les variables explicatives interviennent linéairement dans la modélisation. Le prédicteur linéaire η est alors défini comme suit :

$$\eta = X\beta, \tag{1.2}$$

la matrice X étant fixée par l'expérience et β étant un vecteur de paramètres de taille p qu'il faut estimer.

iii) Contrairement au modèle linéaire classique, le lien entre la $i^{ème}$ composante de η et l'espérance de Y_i n'est plus l'identité, mais s'exprime à l'aide d'une fonction monotone et deux fois dérivable g, appelée fonction de lien, de la manière suivante :

$$\eta_i = g(\mu_i), \quad \text{avec } \mu_i = \mathbb{E}(Y_i) \tag{1.3}$$

Rappelons également que nous pouvons exprimer l'espérance et la variance de la variable à expliquer Y au moyen du paramètre canonique θ. En effet, nous avons les relations classiques suivantes, pour tout $i \in \{1, \dots, n\}$:

$$\begin{cases} \mu_i = \mathbb{E}(Y_i) = b'(\theta_i) \\ \mathbb{V}(Y_i) = a_i(\phi) b''(\theta_i) \end{cases} \tag{1.4}$$

Cela nous permet d'introduire une notation très usitée dans le cadre des GLM. En effet, en utilisant la première relation donnée par (1.4), nous avons immédiatement : $\theta_i = {b'}^{-1}(\mu_i)$. En adoptant la notation $v = b'' \circ {b'}^{-1}$, une autre expression de la variance de Y_i est donc :

$$\mathbb{V}(Y_i) = a_i(\phi) v(\mu_i) \tag{1.5}$$

Enfin, une notion importante qu'il faut évoquer est celle de la canonicité du lien. En effet, tout comme il apparaît dans l'écriture de la densité un paramètre dit "canonique", il est communément admis l'existence d'une fonction de lien qualifiée également de canonique, qui est en général privilégiée dans les modélisations. Elle est définie comme étant la fonction de lien qui assure l'égalité entre le prédicteur linéaire et le paramètre canonique. Grâce aux expressions données par (1.3) et (1.4), il vient : $\eta_i = g(b'(\theta_i))$. Par conséquent, en posant $g^\star = {b'}^{-1}$, on a $\eta_i = g^\star(b'(\theta_i)) = \theta_i$, et c'est précisément en ce sens que ${b'}^{-1}$ est qualifiée de "fonction de lien canonique".

1.2.2 Estimation dans le cadre des Modèles Linéaires Généralisés.

En procédant à une approche classique par maximum de vraisemblance, il apparaît une équation non linéaire en les paramètres $\beta_j, j = 1, \dots, p$. En effet, en notant l la log-vraisemblance

du modèle, alors nous avons l'équivalence suivante [9] :

$$\begin{aligned}
\frac{\partial l}{\partial \beta_j} = 0 \quad &\Leftrightarrow \quad \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = 0 \\
&\Leftrightarrow \quad \sum_{i=1}^{n} \left\{ \frac{\partial \eta_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial l_i}{\partial \theta_i} \right\} = 0 \\
&\Leftrightarrow \quad \sum_{i=1}^{n} \left\{ x_{ij} \frac{1}{g'(\mu_i)} \frac{1}{b''(\theta_i)} \frac{y_i - b'(\theta_i)}{a_i(\phi)} \right\} = 0 \\
&\Leftrightarrow \quad \sum_{i=1}^{n} \left\{ x_{ij} \frac{1}{[g'(\mu_i)]^2 \, \mathbb{V}(Y_i)} g'(\mu_i)(y_i - \mu_i) \right\} = 0
\end{aligned}$$

Si l'on veut formuler les choses d'une manière plus concise en utilisant une écriture matricielle, on peut poser d'une part :

$$W_\beta = Diag \left(\frac{1}{[g'(\mu_i)]^2 \, \mathbb{V}(Y_i)} \right)_{i=1,\dots,n}$$

et d'autre part :

$$\frac{d\eta}{d\mu} = Diag \left(\frac{d\eta_i}{d\mu_i} \right)_{i=1,\dots,n} = Diag \left(g'(\mu_i) \right)_{i=1,\dots,n}$$

et observer alors que les équations du maximum de vraisemblance pour le vecteur de paramètres β s'écrivent :

$$\boxed{X' W_\beta \frac{d\eta}{d\mu} (y - \mu) = 0} \tag{1.6}$$

Notons bien que ces équations ne sont pas linéaires en β car ce paramètre est présent dans les matrices W_β et $\dfrac{d\eta}{d\mu}$ ainsi que dans le vecteur μ. C'est pourquoi une stratégie itérative est en général utilisée pour obtenir une estimation $\hat{\beta}$. L'algorithme mis en place pour répondre à cet objectif est l'algorithme des scores de Fisher. L'une des façons de présenter cette procédure est de partir de l'équation (1.6), et de l'assimiler formellement à l'équation normale d'un modèle linéaire dont le vecteur dépendant z_β serait défini par :

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

Les équations décrites par l'écriture (1.6) deviennent alors :

$$\boxed{X' W_\beta \left(z_\beta - X\beta \right) = 0} \tag{1.7}$$

L'idée est alors d'utiliser la valeur courante de β pour calculer la matrice des poids W_β et le vecteur dépendant z_β. Une nouvelle valeur de β est alors accessible, en résolvant le système comme s'il s'agissait d'équations normales. Plus précisément, à l'étape courante $[t]$, on se place sous le modèle linéarisé

$$\begin{aligned} z^{[t]} &= X\beta + \left(\frac{\partial \eta}{\partial \mu}\right)^{[t]} \left(y - \mu^{[t]}\right) \\ &= X\beta + e^{[t]} \end{aligned} \tag{1.8}$$

en ayant posé $e^{[t]} = \left(\frac{\partial \eta}{\partial \mu}\right)^{[t]} \left(y - \mu^{[t]}\right)$.

Sous le modèle défini par (1.8), nous avons les égalités :

$$\begin{cases} \mathbb{E}\left(e^{[t]}\right) = 0 \\ \mathbb{V}\left(e^{[t]}\right) = Diag\left(\mathbb{V}[g'(\mu_i)Y_i]\right)_{i=1,\dots,n} = Diag\left(\left[g'(\mu_i)\right]^2 \mathbb{V}(Y_i)\right)_{i=1,\dots,n} = W_{\beta^{[t]}}^{-1} \end{cases}$$

Les itérations successives consistent alors à poser :

$$X'W_{\beta^{[t]}}\left(z_{\beta^{[t]}} - X\beta^{[t+1]}\right) = 0 \quad \Leftrightarrow \quad \boxed{\beta^{[t+1]} = \left(X'W_{\beta^{[t]}}X\right)^{-1} X'W_{\beta^{[t]}} z_{\beta^{[t]}}}$$

Notons que l'on peut voir l'algorithme des scores de Fisher comme une procédure itérative de moindres carrés pondérés, car la valeur courante $\beta^{[t+1]}$ obtenue est la même que celle qui aurait consisté à poser :

$$\boxed{\beta^{[t+1]} = \underset{\beta \in \mathbb{R}^p}{argmin} \left\{ \left\| z_{\beta^{[t]}} - X\beta \right\|^2_{W_{\beta^{[t]}}} \right\}}$$

Enfin, pour bien clarifier les choses, nous présentons ici un pseudo-code général récapitulant les différentes étapes à mettre en œuvre au sein de l'algorithme des scores de Fisher pour obtenir une estimation numérique du vecteur de paramètres β.

La pertinence de cet algorithme est de plus appuyée par le fait que dans le cas d'un lien canonique, l'algorithme des scores de Fisher est en tout point identique à l'algorithme d'optimisation numérique de Newton-Raphson. [10]

Algorithm 1 ALGORITHME DES SCORES DE FISCHER

Initialiser le vecteur

$$\mu^{[0]} = \left(\mu_1^{[0]}, \ldots, \mu_n^{[0]}\right)'.$$

repeat

Calculer le vecteur dépendant $z_{\beta^{[t]}}$, en posant pour tout $i \in \{1, \ldots, n\}$:

$$z_{\beta^{[t]},i} = g(\mu_i^{[t]}) + g'(\mu_i^{[t]})(y_i - \mu_i^{[t]})$$

Calculer ensuite la matrice des poids $W_{\beta^{[t]}}$ de la manière suivante :

$$W_{\beta^{[t]}} = Diag\left(\frac{1}{\left[g'(\mu_i^{[t]})\right]^2 a_i(\phi) v(\mu_i^{[t]})}\right)_{i=1,\ldots,n}$$

Obtenir alors le vecteur de paramètres courant $\beta^{[t+1]}$, en posant :

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

Récupérer, pour tout $i \in \{1, \ldots, n\}$:

$$\mu_i^{[t+1]} = g^{-1}\left(\sum_{j=1}^{p} X_{ij} \beta_j^{[t+1]}\right).$$

until Convergence

1.3 Définition du critère à maximiser inhérent à SCGLR.

1.3.1 Part du critère liée à la qualité de l'ajustement.

L'objectif est à présent d'adapter la stratégie décrite par l'algorithme des scores de Fisher à notre cadre, qui consiste en somme à mêler une réduction dimensionnelle au sein des covariables X à une modélisation de type GLM pour chacune des variables $y^1, y^2, \ldots, y^q$. Pour cela, nous utilisons conjointement l'information de X et de T afin de définir q prédicteurs linéaires, c'est à dire un pour chacune des variables réponses y^k. Mais la particularité que nous imposons à ces prédicteurs est que leurs parties en X soient toutes colinéaires à la même composante. Les auteurs de [3] proposent alors de définir les prédicteurs linéaires de cette façon :

$$\forall k \in \{1, \ldots, q\}, \eta^k = Xu\gamma_k + T\delta_k, \text{ avec } \|u\| = 1$$

Nous insistons ici sur le fait que la composante calculée Xu ne dépend pas de k. Elle est alors commune à toutes les variables $y^1, y^2, \ldots, y^q$. Par contre, pour tout $k \in \{1, \ldots, q\}$, γ_k et δ_k seront les paramètres à estimer.

En supposant l'indépendance des variables $y^1, \ldots, y^q$ ainsi que des unités statistiques sur lesquelles elles se fondent, la log-densité du modèle peut s'écrire, en notant l_k la log-densité associée à la $k^{ème}$ variable :

$$\begin{aligned} l(y|\eta) &= \sum_{i=1}^{n}\sum_{k=1}^{q} l_k\left(y_i^k|\eta_i^k\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{q}\left\{\frac{y_i^k\theta_{k,i} - b_k(\theta_{k,i})}{a_{k,i}(\phi_k)} + c_k(y_i^k, \phi_k)\right\} \end{aligned}$$

Nous soulignons ici que chaque y^k peut avoir sa propre loi, la seule contrainte étant qu'elle appartienne à la famille exponentielle. Aussi les fonctions b et c sont-elles indicées par k.

Une stratégie analogue au cas univarié consiste à définir les vecteurs dépendants et le modèle linéarisé associé à l'étape courante $[t]$ ainsi :

$$\begin{aligned} \mathcal{LM}^{[t]} : \forall k \in \{1, \ldots, q\}, z^{k[t]} &= Xu\gamma_k + T\delta_k + e^{k[t]}, \\ \text{avec } e^{k[t]} &= \left(\frac{\partial\eta^k}{\partial\mu_k}\right)^{[t]}(y^k - {\mu_k}^{[t]}) \end{aligned}$$

En posant :

$$W_k^{[t]} = Diag\left(\frac{1}{\left[g'(\mu_{k,i}^{[t]})\right]^2 a_{k,i}(\phi_k)v(\mu_{k,i}^{[t]})}\right)_{i=1,\ldots,n}$$

le modèle défini par $(\mathcal{LM}^{[t]})$ implique les égalités suivantes :

$$\begin{cases} \mathbb{E}\left(e^{k[t]}\right) = 0 \quad \forall k \in \{1, \ldots, q\} \\ \mathbb{V}\left(e^{k[t]}\right) = {W_k^{[t]}}^{-1} \quad \forall k \in \{1, \ldots, q\} \end{cases}$$

Poser ce modèle équivaut à effectuer la minimisation suivante, en ayant rajouté une condition sur la norme du vecteur u pour en permettre l'identifiabilité.

$$\underset{\substack{\gamma,\delta \\ u\,:\,u'u=1}}{min} \left\{ \sum_{k=1}^{q} \left\| z^{k[t]} - Xu\gamma_k - T\delta_k \right\|_{W_k^{[t]}}^2 \right\} \tag{1.9}$$

Notons que la technique des moindres carrés alternés [11] serait envisageable ici pour proposer un calcul de la composante Xu, ainsi qu'une estimation de chacun des paramètres γ_k et δ_k, avec $k \in \{1, \ldots, q\}$. Néanmoins, les auteurs rappellent que le programme de minimisation (1.9) est purement axé sur la qualité de l'ajustement, comme le montre la réécriture équivalente suivante :

$$(1.9) \Leftrightarrow \underset{u\,:\,u'u=1}{max} \left\{ \sum_{k=1}^{k} \left\| z^{k[t]} \right\|_{W_k^{[t]}}^2 \cos^2_{W_k^{[t]}}\left(z^{k[t]}; \langle Xu, T \rangle\right) \right\}$$

Or, l'objectif initial est de tenir compte à la fois de la qualité de l'ajustement et de la force structurelle de la composante $f = Xu$. Il apparaît donc nécessaire à ce stade de modifier le programme initial (1.9) pour y greffer une mesure de ladite force structurelle, dans l'espoir d'obtenir un critère dont la maximisation oriente Xu vers des dimensions structurellement plus fortes de X.

1.3.2 Part du critère liée à la pertinence structurelle.

Dans leur article [3], les auteurs proposent une formulation générale de la pertinence structurelle, dont nous allons expliciter le concept. D'une manière générale, on est amené à considérer les éléments suivants :

i) La matrice X de taille $n \times p$ représente toujours les différentes variables $x^1, \ldots, x^p$ potentiellement redondantes mesurées sur n unités statistiques.

ii) On considère également une matrice W codant le poids a priori des n unités statistiques sur lesquelles les variables sont mesurées. Si l'on suppose que lesdites unités ont la même importance, alors il conviendra de poser $W = n^{-1} Id_n$.

iii) On se donne aussi une matrice M de taille $p \times p$, appelée "métrique", dont l'effet est de pondérer les variables de façon adéquate.

iv) Enfin, la composante f définie par $f = Xu$ est contrainte par la relation : $\|u\|^2_{M^{-1}} = 1$. Et c'est précisément cette composante f, ou tout du moins le vecteur u qui lui est associé, que l'on veut pousser à "épouser" les structures internes de X.

En considérant une famille de J matrices de référence symétriques définies positives $N = \{N_1, \ldots, N_J\}$, un système de poids associé $\Omega = \{\omega_1, \ldots, \omega_J\}$, ainsi qu'un scalaire $l \geqslant 1$, on définit une mesure générale de la pertinence structurelle par la formulation suivante :

$$\phi_{N,\Omega,l}(u) = \left(\sum_{j=1}^{J} \omega_j (u' N_j u)^l \right)^{\frac{1}{l}} \tag{1.10}$$

Afin de préciser et de mieux comprendre le rôle de chaque terme, nous pouvons faire les quelques remarques suivantes :

i) En général, on imposera la contrainte supplémentaire $\sum_{j=1}^{J} \omega_j = 1$, de façon à ce que l'expression donnée par (1.10) puisse s'interpréter comme une moyenne pondérée généralisée des formes quadratiques $u' N_j u$.

ii) Notons bien que les matrices N_j sont définies de telle sorte que la forme quadratique $u' N_j u$ ait vocation à mesurer, en un certain sens, la proximité du vecteur u par rapport à une structure de référence S_j (e.g. variable, sous-espace, etc).

iii) Enfin, notons que le scalaire l a pour but de contrôler d'une certaine façon la "largeur" des faisceaux de directions pour lesquels la proximité avec u est mesurée.

Pour plus de détails et quelques exemples illustrant le potentiel de cette formulation générale, nous renvoyons le lecteur à l'annexe (6.1).

1.3.3 Critère définitif : Combinaison de la qualité de l'ajustement et de la pertinence structurelle.

A chaque étape de l'algorithme des scores de Fisher, en notant

$$\psi_{X,T}(u) = \sum_{k=1}^{q} \|z^k\|_{W_k}^2 \cos_{W_k}^2(z^k; \langle Xu, T \rangle),$$

et en se donnant un paramètre $s \in [0, 1]$ dont le but est d'accorder une importance relative à la pertinence structurelle au regard de la qualité de l'ajustement, le programme à résoudre s'écrit finalement :

$$\begin{cases} \textbf{Maximiser} & S_T(u) = [\psi_{X,T}(u)]^{1-s} \, [\phi_{N,\Omega,l}(u)]^s \\ \textbf{sous la contrainte :} & u' M^{-1} u = 1 \end{cases} \tag{1.11}$$

En réécrivant la quantité $\psi_{X,T}(u)$ sous forme matricielle, les auteurs proposent une procédure itérative pour mener la maximisation décrite par (1.11), appelée "PING", pour "Projected Iterated Normed Gradient". Une description de cet algorithme ainsi que ses bonnes

propriétés sont données dans l'annexe (6.2). Son point fort est qu'il est pensé pour résoudre des programmes généraux de la forme :

$$\begin{cases} \textbf{Maximiser} \quad h(u) \\ \textbf{sous les contraintes :} \quad u'M^{-1}u = 1 \text{ et } D'u = 0 \end{cases}$$

ce qui lui permettra d'intervenir à plusieurs reprises au sein de la procédure SCGLR, tant pour la recherche de la première composante que pour les composantes suivantes sous contrainte d'orthogonalité.

1.3.4 Présentation du pseudo-code relatif à SCGLR.

Pour trouver la première composante, $f^{[1]} = Xu^{[1]}$, il faut résoudre le programme défini par (1.11). Pour cela, nous mettons en œuvre l'algorithme PING avec :

i) $h(u) = S_T(u)$.

ii) $D = 0$, car aucune contrainte n'est à rajouter pour trouver la première composante.

Les modifications vont apparaître avec la recherche des composantes de rang supérieur à 1. Supposons pour cela avoir trouvé les h premières composantes et les avoir rassemblées dans une matrice $F^{[h]}$. Nous pouvons alors noter :

$$F^{[h]} = \left[f^{[1]}|f^{[2]}|\ldots|f^{[h]}\right]$$

La composante de rang suivant, notée donc $f^{[h+1]}$ doit venir compléter d'une part les composantes déjà trouvées, mais aussi les variables additionnelles rassemblées dans T. La stratégie consiste alors à utiliser non plus T comme variables additionnelles, mais à utiliser l'extension de T suivante :

$$T^{[h]} = T \cup F^{[h]}$$

Enfin, pour éviter toute redondance entre les différentes composantes trouvées, il convient d'imposer que la nouvelle composante recherchée $f^{[h+1]}$ soit orthogonale à toutes les composantes précédemment trouvées $\left(f^{[1]}, \ldots, f^{[h]}\right)$. Matriciellement, la contrainte d'orthogonalité à imposer est donc :

$$F^{[h]'}Wf^{[h+1]} = 0$$

En notant $f^{[h+1]} = Xu$, la contrainte d'orthogonalité à ajouter au programme s'écrit :

$$\begin{aligned} F^{[h]'}WXu = 0 &\Leftrightarrow (X'WF^{[h]})'u = 0 \\ &\Leftrightarrow D^{[h]'}u = 0, \quad \text{avec } D^{[h]} = X'WF^{[h]}. \end{aligned} \tag{1.12}$$

et nous pouvons alors considérer que le nouveau programme à résoudre est :

$$\begin{cases} \textbf{Maximiser} \quad S_{T^{[h]}}(u) = \left[\psi_{X,T^{[h]}}(u)\right]^{1-s}\left[\phi_{N,\Omega,l}(u)\right]^s \\ \textbf{sous les contraintes :} \quad u'M^{-1}u = 1 \text{ et } D^{[h]'}u = 0 \end{cases}$$

Là encore, l'algorithme PING peut intervenir pour résoudre numériquement ce programme, en appliquant la maximisation générale après avoir posé :

i) $h(u) = S_{T^{[h]}}(u)$, et

ii) $D = D^{[h]} = X'WF^{[h]}$

L'algorithme dont le pseudo-code est présenté ci-après a pour but de déterminer numériquement les H premières composantes de la procédure SCGLR, ainsi que des estimations pour les paramètres $\gamma_k^{[h]}$ et $\delta_k^{[h]}$, pour tout $h \in \{1, \ldots, H\}$ et tout $k \in \{1, \ldots, q\}$. La valeur de H est un paramètre fixé par l'utilisateur ; cependant, dans le package R relatif à la méthode présentée, nommé également "SCGLR" [12], les auteurs proposent, via la fonction "scglrCrossVal", de déterminer le nombre de composantes qui minimise l'erreur de prédiction.

Pour mettre en œuvre l'algorithme, nous clarifions les notations préliminaires suivantes :

i) On pose préalablement $T^{[0]} = T$ afin que la recherche de la première composante $u^{[1]}$ s'effectue bien au moyen de la matrice des covariables additionnelles initiale T.

ii) Par convention, $F^{[0]}$ est considérée comme une matrice "vide", car, à l'étape [0], aucune composante n'a encore été trouvée.

iii) On souligne enfin, toujours par convention, que la matrice $D^{[0]}$ est la matrice nulle. En effet, aucune contrainte supplémentaire d'orthogonalité n'est à rajouter pour la recherche de la première composante.

iv) Si l'on considère que toutes les unités statistiques ont le même poids (i.e. la même importance a priori), il conviendra de considérer : $W = \frac{1}{n} Id_n$.

Algorithm 2 DESCRIPTION DE LA PROCÉDURE SCGLR

for h = 1 à H **do**

Initialiser les vecteurs $z^{k[h]}$ ainsi que la matrice $W_k^{[h]}$, pour tout $k \in \{1, \ldots, q\}$

repeat

Avec l'algorithme PING, poser

$$u^{[h]} = \begin{cases} \arg\max S_{T^{[h-1]}}(u) \\ \text{sous les contraintes} \quad u'M^{-1}u = 1 \text{ et } D^{[h-1]'}u = 0 \end{cases}$$

for k = 1 à q **do**

Régresser $z^{k[h]}$ sur $[Xu^{[h]}, T]$ avec une procédure WLS au sens de $W_k^{[h]}$.
Récupérer alors les coefficients $\gamma_k^{[h]}$ et $\delta_k^{[h]}$.
Poser $\eta^{k[h]} = Xu^{[h]}\gamma_k^{[h]} + T\delta_k^{[h]}$
Poser pour $i = 1, \ldots, n$: $\mu_{k,i}^{[h]} = g_k^{-1}\left(\eta_i^{k[h]}\right)$
Mise à jour de $z^{k[h]}$: $z^{k[h]} = \eta^{k[h]} + Diag\left(g_k{}'(\mu_{k,i}^{[h]})\right)_{i=1,\ldots,n}(y^k - \mu_k^{[h]})$
Mise à jour de $W_k^{[h]}$: $W_k^{[h]} = Diag\left(\dfrac{1}{\left[g_k'(\mu_{k,i}^{[h]})\right]^2 a_{k,i}(\phi_k)v_k(\mu_{k,i}^{[h]})}\right)_{i=1,\ldots,n}$

end for

until Convergence

Récupération de la $h^{ème}$ composante : $f^{[h]} = Xu^{[h]}$
Mise à jour de F : $F^{[h]} = \left[F^{[h-1]} \mid f^{[h]}\right]$
Mise à jour de T : $T^{[h]} = \left[T \mid F^{[h]}\right]$
Mise à jour de D : Poser $D^{[h]} = X'WF^{[h]}$.

end for

Chapitre 2

Principes de l'introduction d'effets aléatoires dans la modélisation.

2.1 Le cas des Modèles Linéaires Mixtes.

2.1.1 Parabole introductive.

La plupart des études statistiques sont motivées par la détection et la quantification de la variabilité potentielle des données. Avec son modèle d'"Analyse de la variance" (ANOVA), Fisher fut l'un des précurseurs dans ce domaine : sa stratégie consistait à cloisonner les différentes sources de variation à l'aide de facteurs "à effets fixes" (en général prédéfinis) afin de déterminer la significativité des différences observées entre les moyennes des sous-groupes de données induites par ces facteurs. Cependant, cette modélisation comporte quelques écueils, notamment quand un facteur contient un nombre tellement grand de niveaux qu'il nous est impossible de tous les inclure dans l'expérience.

Illustrons cette remarque sur un exemple simple : Imaginons qu'un canton français comporte trois lycées : le lycée A, le lycée B et le lycée C, et qu'une association de parents d'élèves ait commandé une étude afin de déterminer les effets des trois lycées sur l'obtention du baccalauréat, afin de savoir dans quel lycée envoyer leurs enfants. Dans ce cadre, une ANOVA classique peut être réalisée car ce sont bien les effets "fixes" desdits lycées qui préoccupent les parents. Par contre, si l'étude se place à l'échelle nationale afin de préciser la variabilité de l'obtention du baccalauréat liée au lycée fréquenté, il ne sera sans doute pas possible (ou trop coûteux) de prendre en compte tous les lycées mais il faudra se contenter d'en sélectionner un échantillon aléatoire. Les effets spécifiques des lycées sélectionnés ne nous intéressent donc pas, car ils sont vus comme des représentants d'un ensemble beaucoup plus vaste de lycées. En ce sens, les effets observés ne sont plus qualifiés d'effets fixes mais plutôt d'effets "aléatoires".

Les L2M, ou LMM pour "Linear Mixed Models", ont vocation à prendre en compte ces deux types de facteurs, afin de préciser les diverses sources de variabilité des données. Une distinction conceptuelle et en terme de modélisation est alors faite entre ces deux types de facteurs.

i) D'une part les facteurs à effets fixes qui comportent un nombre fini de niveaux, et qui sont tous représentés dans le sens où les données se répartissent sur ces différents niveaux. Notre objectif est alors de préciser l'effet de chaque niveau sur la variable d'intérêt.

ii) D'autre part les facteurs à effets aléatoires qui comportent un nombre infini (ou très grand) de niveaux, dont un échantillon seulement est représenté. Il est à noter que la façon dont chacun des facteurs influence les données ne présente pas d'intérêt en soit. Cependant, on aimerait savoir quelle part de la variabilité totale est produite par cet échantillon de niveaux.

Ainsi, les modèles à effets mixtes constituent un moyen plus élaboré d'étudier la variabilité des données, puisqu'une distinction est faite entre : la variabilité liée aux effets fixes, celle liée aux effets aléatoires, et enfin celle que l'on affecte aux erreurs. Les différentes composantes de la variance ainsi considérées rendent alors plus précise son origine.

2.1.2 Écriture générale du modèle et principales propriétés.

En supposant, pour plus de simplicité, n'avoir qu'une seule variable à modéliser Y, le modèle général peut s'écrire de la façon suivante :

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

dans lequel apparaissent les éléments suivants :

i) X est la matrice fixée par l'expérience, relative aux effets fixes : elle contient p variables explicatives. Le nombre d'unités statistiques étant fixées à n, cette matrice sera alors de taille $(n \times p)$.

ii) En supposant avoir q niveaux aléatoires représentés dans l'expérience, le vecteur de paramètres ξ relatifs aux effets aléatoires sera de taille q. Si nous souhaitons distinguer les niveaux correspondant à chacun des effets aléatoires, nous pouvons décomposer le vecteur ξ en K sous-vecteurs, via la relation $\xi = (\xi_1', \ldots, \xi_K')'$, avec K le nombre d'effets aléatoires. En notant q_j le nombre de réalisations du $j^{ème}$ effet aléatoire, alors chacun des ξ_j, pour j allant de 1 à K, est un sous-vecteur de taille q_j, avec la relation $\sum_{j=1}^{K} q_j = q$.
Dans la plupart des modélisations, un seul effet aléatoire est représenté, ce qui simplifie les notations car $K = 1$ et $\xi = \xi_1$.
Par exemple, sur l'exemple des lycées donné dans la section (2.1.1), les trois lycées peuvent être considérés comme trois réalisations du même facteur à effet aléatoire : le facteur "lycée". Dans ce cas, on aura $K = 1$, $\xi = \xi_1$ et $q = q_1 = 3$.

iii) La matrice de design U associée aux effets aléatoires est par conséquent de taille $(n \times q)$, et la même distinction que précédemment peut nous amener à noter $U = (U_1 \mid \ldots \mid U_K)$, où chacune des sous-matrices U_j est de taille $(n \times q_j)$. Dans les cas les plus simples, c'est-à-dire lorsque la modélisation n'implique qu'un seul effet aléatoire, on peut poser $U = U_1$ et préciser alors que dans la plupart des cas, la matrice U_1 a un codage disjonctif complet. Chaque ligne contient ainsi un seul "1" et plusieurs "0", indiquant que la mesure concernée a été prise pour tel niveau du facteur.

Des structures plus complexe existent pour la matrice U, en particulier pour des modélisations incluant des paramètres aléatoires sur les covariables. Dans ce cas, le modèle n'est pus seulement à "intercept aléatoire", mais également à "pente aléatoire".

iv) Enfin, ε est un vecteur aléatoire d'erreurs de taille n.

Essentiellement quatre hypothèses de distribution sont à considérer pour mettre en œuvre une modélisation par un L2M.

i) Dans un premier temps, chacun des effets aléatoires est supposé être distribué selon une loi normale. Ceci implique :

$$\forall j \in \{1, \ldots, q\}, \quad \xi_j \sim \mathcal{N}_{q_j}(0, \sigma_j^2 A_j)$$

avec A_j une matrice $(q_j \times q_j)$ connue et σ_j^2 le paramètre de variance à estimer.

ii) Toujours concernant les effets aléatoires, il sont supposés être indépendants deux à deux. Par conséquent, nous pouvons noter symboliquement, avec D la matrice diagonale par blocs $\left(\sigma_j^2 A_j\right)_{j=1,\ldots,K}$:

$$\xi \sim \mathcal{N}_q(0, D)$$

iii) Dans un modèle linéaire, la distribution des erreurs est également supposée gaussienne :

$$\varepsilon \sim \mathcal{N}_n(0, R) \text{ avec } R = \sigma_0^2 V_0$$

iv) Enfin, on suppose que quel que soit $j \in \{1, \ldots, q\}$, ε et ξ_j sont indépendants.

Dans la plupart des applications, les matrices A_j et V_0 sont diagonales. Cependant, il est possible d'envisager des situations où une dépendance entre les individus soumis à un même facteur d'un effet aléatoire existe.

À ce stade, il est intéressant d'écrire une expression de la matrice de variance-covariance de Y. Nous montrons alors la décomposition de la variabilité totale en plusieurs parties induite par la modélisation selon un L2M. Nous avons en effet :

$$\begin{aligned}
\Gamma = \mathbb{V}(Y) &= \mathbb{V}(X\beta + U\xi + \varepsilon) \\
&= \mathbb{V}(U\xi + \varepsilon) \\
&= \mathbb{V}(U\xi) + \mathbb{V}(\varepsilon) \quad \text{car } \xi \text{ et } \varepsilon \text{ sont indépendants.} \\
&= UDU' + R \\
&= \sum_{j=1}^{K} \sigma_j^2 U_j A_j U_j' + \sigma_0^2 V_0 \quad \text{car D est la matrice diagonale par blocs } \left(\sigma_j^2 A_j\right)_{j=1,\ldots,K}. \\
&= \sum_{j=0}^{K} \sigma_j^2 V_j \quad \text{en posant } V_j = U_j A_j U_j' \text{ pour } j = 1, \ldots, K.
\end{aligned}$$

Nous rappelons de plus quelques résultats sur les distributions jointes, marginales et conditionnelles des vecteur Y et ξ, car ils interviennent lors de l'estimation des paramètres, notamment dans la méthode de Henderson privilégiée par la suite. Ces résultats, en particulier le quatrième qui nécessite un raisonnement basé sur les compléments de Schur, sont démontrés dans [8]. Avec les notations précédemment définies, nous avons alors :

i) $Y|\xi \sim \mathcal{N}_n(X\beta + U\xi, R)$

ii) $Y \sim \mathcal{N}_n(X\beta, \Gamma)$

iii) $\begin{pmatrix} Y \\ \xi \end{pmatrix} \sim \mathcal{N}_{n+q}\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} R + UDU' & UD \\ DU' & D \end{pmatrix}\right)$

iv) $\xi|Y = y \ \sim \mathcal{N}_q\left(DU'\Gamma^{-1}(y - X\beta), D - DU'\Gamma^{-1}UD\right)$

2.1.3 Estimation dans les Modèles Linéaires Mixtes : la méthode de Henderson.

Surprenante dans un premier temps, la stratégie envisagée par Henderdon [13] consiste à considérer la densité jointe de Y et ξ, et de faire jouer à ξ le rôle de paramètre. Grâce à la loi jointe donnée par (iii), et en dérivant la log-vraisemblance correspondante l par rapport à β et à ξ, le système à résoudre sera alors :

$$\begin{cases} \dfrac{\partial l}{\partial \beta}(\beta, \xi) = 0 \\ \dfrac{\partial l}{\partial \xi}(\beta, \xi) = 0 \end{cases} \tag{2.1}$$

Le point fort de cette procédure, curieuse et discutable dans un premier temps sur le plan conceptuel, réside dans le fait que l'estimateur $\hat{\beta}$ et le prédicteur $\tilde{\xi}$ obtenus possèdent des propriétés particulièrement intéressantes que nous évoquerons ensuite, qui rendent possible la mise en œuvre d'une procédure itérative d'estimation des composantes de la variance.

En effet, une réécriture matricielle du système (2.1) permet d'obtenir :

$$\boxed{\begin{pmatrix} X'R^{-1}X & X'R^{-1}U \\ U'R^{-1}X & U'R^{-1}U + D^{-1} \end{pmatrix}\begin{pmatrix} \beta \\ \xi \end{pmatrix} = \begin{pmatrix} X'R^{-1}y \\ U'R^{-1}y \end{pmatrix}} \tag{2.2}$$

La chose fondamentale à remarquer ici est que la résolution des équations de Henderson (2.2) nous permettent de récupérer le "Best Linear Unbiased Estimator" de β ($BLUE(\beta)$) mais également le "Best Linear Unbiased Predictor" de ξ ($BLUP(\xi)$). En effet, dans [8], les auteurs démontrent, toujours en faisant intervenir les compléments de Schur que les valeurs $\hat{\beta}$ et $\tilde{\xi}$ résultant de la résolution du système peuvent se réécrire :

$$\boxed{\begin{cases} \hat{\beta} = \left(X'\Gamma^{-1}X\right)^{-1} X'\Gamma^{-1}y = BLUE(\beta) \\ \tilde{\xi} = DU'\Gamma^{-1}\left(y - X\hat{\beta}\right) = \mathbb{E}\left(\xi|y\right) = BLUP(\xi) \end{cases}}$$

Le gros avantage des équations de Henderson est la réduction du temps de calcul : Il est en effet beaucoup plus simple d'inverser les matrices R et D, souvent diagonales, et de résoudre le système (2.2) de taille $p+q$, que de résoudre le système équivalent, dans lequel la matrice carrée Γ, de taille n (avec souvent $n \gg p+q$), peut s'avérer difficile à inverser.

De plus, en notant C^* l'inverse de la matrice formée des q dernières lignes et colonnes de la matrice des coefficients du système de Henderson (2.2), à savoir $C^* = (U'R^{-1}U + D^{-1})^{-1}$, les auteurs de [8] prolongent l'estimation en établissant les relations :

$$\begin{cases} \hat{\sigma}_0^2 = n^{-1}\left(y'V_0^{-1}(y - X\hat{\beta} - U\tilde{\xi})\right) \\ \hat{\sigma}_j^2 = \dfrac{\tilde{\xi}_j' A_j^{-1}\tilde{\xi}_j}{q_j - \frac{tr(A_j^{-1}C_{jj}^*)}{\hat{\sigma}_j^2}}, \quad \forall j \in \{1,\dots,K\} \end{cases}$$

avec C_{jj}^* la $j^{ème}$ sous-matrice de C^*, i.e. celle correspondant au $j^{ème}$ effet aléatoire.

Le pseudo-code que l'on peut proposer pour récapituler cette procédure est le suivant :

Algorithm 3 Méthode de Henderson

INITIALISATION des paramètres :

Proposer des valeurs initiales

$$\sigma^{2^{[0]}} = \left(\sigma_0^{2^{[0]}}, \sigma_1^{2^{[0]}}, \ldots, \sigma_K^{2\ [0]}\right)$$

repeat

1. Calculer les matrices :

$$\begin{cases} R^{[t]} = \sigma_0^{2^{[t]}} V_0 \\ D^{[t]} = Diag\left(\sigma_j^{2^{[t]}} A_j\right)_{j=1,\ldots,K} \end{cases}$$

2.Résoudre le système :

$$\begin{pmatrix} X' {R^{[t]}}^{-1} X & X' {R^{[t]}}^{-1} U \\ U' {R^{[t]}}^{-1} X & U' {R^{[t]}}^{-1} U + {D^{[t]}}^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \xi \end{pmatrix} = \begin{pmatrix} X' {R^{[t]}}^{-1} y \\ U' {R^{[t]}}^{-1} y \end{pmatrix} \text{ d'inconnues } \beta \text{ et } \xi.$$

Noter $\beta^{[t+1]}$ et $\xi^{[t+1]}$ les solutions.

3. Poser alors :

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

ii) $\sigma_0^{2^{[t+1]}} = \dfrac{y' V_0^{-1} (y - X\beta^{[t+1]} - U\xi^{[t+1]})}{n}$

until Convergence

2.2 Glissement vers les modèles linéaires généralisés à effets aléatoires.

2.2.1 Construction générale.

Cette brève introduction vise à lister les hypothèses qui doivent être considérées pour la mise en œuvre d'un modèle linéaire généralisé avec introduction d'effets aléatoires (GL2M) et les propriétés qui en découlent. Une grande partie d'entre elles se fonde sur une extension des GLM en considérant des raisonnements conditionnels aux effets aléatoires ξ [14]. Pour les lister, nous considérons avoir uniquement une seule variable à expliquer Y et plusieurs variables explicatives rassemblées dans un seul tableau X, U désignant toujours la matrice de design des effets aléatoires.

i) Conditionnellement à ξ, les composantes de Y sont supposées indépendantes et distribuées selon une loi admettant une structure exponentielle. Autrement dit, pour tout $i \in \{1, 2, \dots, n\}$, la variable aléatoire Y_i est supposée être distribuée conditionnellement à ξ selon la loi de densité :

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

Notons que nous avons pris soin d'indicer par ξ les quantités qui en dépendent, et que la même stratégie de notation sera désormais adoptée par la suite.

ii) Les effets aléatoires intervenant pleinement dans le prédicteur linéaire, nous l'indiçons également par ξ. En conséquence, η_ξ est alors défini comme suit :

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

iii) La fonction de lien g ne s'exprime plus ici en fonction de l'espérance marginale de Y mais en fonction de son espérance conditionnelle à ξ. Nous avons alors les expressions suivantes, pour chaque $i \in \{1, 2, \dots, n\}$:

$$\eta_{\xi,i} = g(\mu_{\xi,i}) \text{ avec } \mu_{\xi,i} = \mathbb{E}(Y_i|\xi)$$

iv) Dans la même veine, les quantités $b'(\theta_{\xi,i})$ et $b''(\theta_{\xi,i})$ ne sont plus reliées à l'espérance et à la variance de Y_i, mais respectivement à l'espérance conditionnelle et à la variance conditionnelle à ξ. Une démonstration analogue aux GLM, toujours basée sur la structure exponentielle de la distribution de la variable d'intérêt, permet alors d'écrire :

$$\begin{cases} \mu_{\xi,i} = \mathbb{E}(Y_i|\xi) = b'(\theta_{\xi,i}) \\ \mathbb{V}(Y_i|\xi) = a_i(\phi)b''(\theta_{\xi,i}) \end{cases} \tag{2.3}$$

De plus, tout comme dans les GLM, il existe ici un lien entre $\mathbb{E}(Y_i|\xi)$ et $\mathbb{V}(Y_i|\xi)$. D'après (2.3), et en posant $v = b'' \circ b'^{-1}$, il vient :

$$\mathbb{V}(Y_i|\xi) = a_i(\phi)v(\mu_{\xi,i}).$$

v) La notion de canonicité du lien s'exporte enfin très facilement au cas des GL2M. Nous avons en effet

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

En prenant $g^* = b'^{-1}$, il vient immédiatement :

$$\eta_{\xi,i} = g^*(\mu_{\xi,i}) = b'^{-1}(b'(\theta_{\xi,i})) = \theta_{\xi,i}.$$

C'est pourquoi la fonction b'^{-1} est également qualifiée de lien canonique dans le cas des GL2M.

vi) Enfin, les principales propriétés relatives aux effets aléatoires sont conservées ici. En particulier, nous supposons toujours leur indépendance deux à deux. On a toujours :

$$\xi \sim \mathcal{N}_q(0, D) \text{ avec } D \text{ la matrice diagonale par blocs } \left(\sigma_j^2 A_j\right)_{j=1,\dots,K}$$

Et par indépendance des éléments du vecteur Y, on peut poser :

$$\mathbb{V}(Y|\xi) = Diag\left(a_i(\phi)v(\mu_{\xi,i})\right)_{i=1,\dots,n}$$

2.2.2 Estimation dans les Modèles Linéaires Mixtes Généralisés : l'algorithme de Schall.

La stratégie mise en œuvre par Schall consiste à mêler les approches d'estimations utilisées pour les GLM d'une part et pour les L2M d'autre part. En effet, dans un premier moment, une linéarisation du modèle conditionnellement à ξ est effectuée. Ensuite, nous pourrons utiliser une adaptation des équations de Henderson afin d'obtenir une estimation des paramètres.

Étape de linéarisation.

La même stratégie de linéarisation que celle mise en œuvre pour les GLM peut être proposée ici. En effet, pour tout $i \in \{1, \dots, n\}$, une approximation à l'ordre 1 de $g(y_i)$ au voisinage de $\mu_{\xi,i}$ est :

$$\begin{aligned} g(y_i) &\approx g(\mu_{\xi,i}) + (y_i - \mu_{\xi,i})g'(\mu_{\xi,i}) \\ &= \eta_{\xi,i} + (y_i - \mu_{\xi,i})g'(\mu_{\xi,i}) \end{aligned}$$

On est alors amené à définir le vecteur dépendant $z_{\beta,\xi}$ de la manière suivante :

$$z_{\beta,\xi,i} = \eta_{\xi,i} + (y_i - \mu_{\xi,i})g'(\mu_{\xi,i})$$

En notant e_i la quantité $(y_i - \mu_{\xi,i})g'(\mu_{\xi,i})$, on aboutit au modèle linéarisé mixte $\mathcal{M}_\xi$ suivant :

$$(\mathcal{M}_\xi) : \quad Z_{\beta,\xi} = X\beta + U\xi + e$$

ce qui donne accès aux quantités $\mathbb{E}(Z_{\beta,\xi}|\xi)$ et $\mathbb{V}(Z_{\beta,\xi}|\xi)$. En effet,

i) $\mathbb{E}(Z_{\beta,\xi}|\xi) = X\beta + U\xi$

ii)

$$\begin{aligned}\mathbb{V}(Z_{\beta,\xi}|\xi) &= \mathbb{V}(e|\xi)\\ &= \mathbb{V}\left([(Y-\mu_\xi)g'(\mu_\xi)]\,|\xi\right)\\ &= Diag\left([g'(\mu_{\xi,i})]^2\,\mathbb{V}(Y_i|\xi)\right)_{i=1,\dots,n} \quad \text{avec } \mathbb{V}(Y_i|\xi) = a_i(\phi)b''(\theta_i) = a_i(\phi)v(\mu_{\xi,i}).\\ &= W_{\beta,\xi}^{-1} \quad \text{en notation.}\end{aligned}$$

Enfin, il est possible de simplifier l'écriture $W_{\beta,\xi}$ si l'on adopte le lien canonique dans la modélisation, ce qui peut être intéressant à considérer dans un but algorithmique. Pour $g = g^*$, on a en effet [15] :

$$W_{\beta,\xi} = Diag\left(\frac{1}{a_i(\phi)g'(\mu_{\xi,i})}\right)_{i=1,\dots,n}$$

Étape d'estimation.

Dans un premier temps, on aurait envie de "plonger" le modèle $\mathcal{M}_\xi$ dans la structure d'un L2M, en calculant l'espérance et la variance marginales de la variable $Z_{\beta,\xi}$. La nature aléatoire des effets aléatoires ξ serait alors préservée. Pour simplifier les notations, nous omettons le double indice β, ξ dans la variable Z.

Dans ce cas, on aurait alors d'une part :

$$\mathbb{E}(Z) = X\beta$$

D'autre part, la variance marginale se calculerait comme suit :

$$\begin{aligned}\mathbb{V}(Z) &= \mathbb{E}\left(\mathbb{V}(Z|\xi)\right) + \mathbb{V}\left(\mathbb{E}(Z|\xi)\right)\\ &= \mathbb{E}\left(W_{\beta,\xi}^{-1}\right) + \mathbb{V}\left(U\xi\right)\\ &= W_\beta^{-1} + UDU'\\ &= \Gamma_\beta,\end{aligned}$$

avec $W_\beta^{-1} = \mathbb{E}\left[Diag\left([g'(\mu_{\xi,i})]^2\,\mathbb{V}(Y_i|\xi)\right)_{i=1,\dots,n}\right]$.

En raisonnant sur le couple de vecteurs aléatoires $\begin{pmatrix} Z \\ \xi \end{pmatrix}$, on pourrait alors écrire, d'une manière analogue au raisonnement sur les L2M :

$$\begin{pmatrix} Z \\ \xi \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbb{E}(W_{\beta,\xi}^{-1}) + UDU' & UD \\ DU' & D \end{pmatrix}\right)$$

afin de récupérer les équations de Henderson pour ce modèle. Il est tout à fait possible de le faire, et cette méthode existe et elle a déjà été proposée par Engel et Keen [16].

Cependant, il est à noter que cette technique présente un inconvénient majeur : le calcul de $\mathbb{E}(W_{\beta,\xi}^{-1})$. En effet, si le calcul ne pose aucun problème particulier dans la mesure où l'on adopte le lien canonique, il se peut qu'il n'existe pas d'expression explicite de cette espérance si l'on choisit une fonction de lien non canonique. Schall propose donc d'adopter un point de vue légèrement différent qui règle le problème.

L'idée de Schall consiste à préserver la nature aléatoire de ξ, mais uniquement partiellement, car il propose de maintenir la matrice $W_{\beta,\xi}^{-1}$ au lieu de son espérance. La variance conditionnelle des résidus fait alors son apparition dans les équations. Par conséquent, dans la suite, tout se passe comme si le couple de vecteurs aléatoires $\begin{pmatrix} Z \\ \xi \end{pmatrix}$ était distribué de la façon suivante :

$$\begin{pmatrix} Z \\ \xi \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} W_{\beta,\xi}^{-1} + UDU' & UD \\ DU' & D \end{pmatrix} \right)$$

Au final, en dérivant la log-vraisemblance par rapport à β et ξ, le système de Henderson auquel on aboutit est [17] :

$$\begin{pmatrix} X'W_{\beta,\xi}X & X'W_{\beta,\xi}U \\ U'W_{\beta,\xi}X & U'W_{\beta,\xi}U + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \xi \end{pmatrix} = \begin{pmatrix} X'W_{\beta,\xi}z \\ U'W_{\beta,\xi}z \end{pmatrix} \tag{2.4}$$

Notons que pour l'obtenir, il suffit de remplacer la matrice R^{-1}, présente dans les équations (2.2), par la matrice $W_{\beta,\xi}$. Pour résumer la stratégie adoptée, nous proposons dans la suite un pseudo-code qui récapitule les différentes étapes précédemment exposées, ainsi que l'estimation des composantes de la variance σ_j^2, pour j allant de 1 à K.

2.2.3 Pseudo-code relatif à l'algorithme de Schall.

Notons que dans notre contexte, $C^{*[t]} = \left(U'W_{\beta,\xi}^{[t]}U + D^{[t]^{-1}}\right)^{-1}$, c'est-à-dire l'inverse de la matrice constituée des q dernières lignes et colonnes des coefficients du système de Henderson à l'étape $[t]$. Précisons également que $C_{jj}^{*\,[t]}$ désigne la $j^{ème}$ sous matrice de $C^{*[t]}$, celle correspondant au $j^{ème}$ effet aléatoire de l'expérience.

Algorithm 4 ALGORITHME DE SCHALL

Initialiser les paramètres

$$\beta^{[0]}, \xi^{[0]} \text{ et } \sigma^{2\,[0]} = \left(\sigma_1^{2\,[0]}, \ldots, \sigma_K^{2\,[0]}\right)$$

repeat

1. Calcul du vecteur courant $z^{[t]}$:
Poser tout d'abord : $\eta^{[t]} = X\beta^{[t]} + U\xi^{[t]}$
Poser ensuite, pour $i = 1, \ldots, n$: $\mu_{\xi,i}^{[t]} = g^{-1}\left(\eta_i^{[t]}\right)$
Poser enfin, pour $i = 1, \ldots, n$: $z_i^{[t]} = \eta_i^{[t]} + \left(y_i - \mu_{\xi,i}^{[t]}\right) g'\left(\mu_{\xi,i}^{[t]}\right)$

2. Calcul des matrices $W_{\beta,\xi}^{[t]}$ et $D^{[t]}$:
Poser tout d'abord : $W_{\beta,\xi}^{[t]} = Diag\left(\dfrac{1}{\left[g'(\mu_{\xi,i}^{[t]})\right]^2 a_i(\phi) v(\mu_{\xi,i}^{[t]})}\right)_{i=1,\ldots,n}$
Poser ensuite : $D^{[t]} = Diag(\sigma_j^{2\,[t]} A_j)_{j=1,\ldots,K}$

3. Résolution du système de Henderson courant :
Résoudre le système, d'inconnues β et ξ :

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

Les solutions de ce système seront notées $\beta^{[t+1]}$ et $\xi^{[t+1]}$.

4. Calcul des composantes de la variance :
Poser alors : $\sigma_j^{2\,[t+1]} = \dfrac{\xi_j'^{\,[t+1]} A_j^{-1} \xi_j^{\,[t+1]}}{q_j - \frac{tr(A_j^{-1} C_{jj}^{*\,[t]})}{\sigma_j^{2\,[t]}}}, \quad \forall j \in \{1, \ldots, K\}$

until Convergence

Chapitre 3

Adaptation de la méthode SCGLR dans le cadre de données groupées.

3.1 Explications générales des stratégies adoptées.

Au préalable, précisons que tout comme dans le cadre de la méthode SCGLR initiale, le but est de modéliser non pas une seule variable d'intérêt, mais un ensemble de q variables $(y^1, y^2, \ldots, y^q)$ rassemblées dans un tableau Y de taille $(n \times q)$. Quant aux variables explicatives, nous les segmentons toujours en deux catégories :

i) Le tableau X de taille $(n \times p)$ regroupe toujours un nombre potentiellement important de variables, où la redondance entre les régresseurs est possible.

ii) Le tableau des variables additionnelles T de taille $(n \times r)$ rassemble des variables présélectionnées, peu redondantes, et présentant un intérêt tout particulier aux yeux de l'expérimentateur.

Cette fois-ci par contre, nous incluons dans l'expression des prédicteurs linéaires un terme faisant intervenir les effets aléatoires ξ. Nous supposons pour cela l'existence d'un effet aléatoire pour chacune des covariables y^k, indépendants les uns des autres. Ainsi, pour tout $k \in \{1, \ldots, q\}$, les prédicteurs linéaires sont alors définis par :

$$\eta_\xi^k = X u \gamma_k + T\delta_k + U\xi_k, \quad \|u\| = 1 \tag{3.1}$$

La procédure se place au sein de l'algorithme de Schall, et s'oriente encore en deux étapes. La première a pour but d'effectuer une réduction de dimension au sein du tableau X tandis que la deuxième vise à obtenir des estimations des paramètres γ_k, δ_k et σ_k^2. Pour la première étape, nous utilisons l'algorithme PING, qui s'avère efficace pour rechercher des composantes orthogonales entre elles, satisfaisant un compromis entre la qualité de l'ajustement et la proximité de la composante courante aux structures internes de X. Pour la deuxième étape, nous proposons de résoudre les équations de Henderson impliquant les variables additionnelles T auxquelles on aura greffé la composante retenue dans X.

3.2 Obtention des équations de Henderson dans notre modèle, à u fixé.

Les équations de Henderson s'obtiennent ici en mettant en oeuvre les deux étapes usuelles de linéarisation et d'estimation telles que présentées dans la section (2.2.2). Mais deux différences notables s'observent par rapport au raisonnement mené dans (2.2.2). D'une part, les étapes de linéarisation et d'estimation doivent s'effectuer pour chacune des variables réponses $y^1, \dots, y^k$ et d'autre part, l'écriture des prédicteurs linéaires n'est pas la même.

Étape de linéarisation.

En posant $\mu^k_{\xi,i} = \mathbb{E}(Y^k_i|\xi_k)$, alors pour tout $i \in \{1, \dots, n\}$ et pour chacun des $k \in \{1, \dots, q\}$, une approximation à l'ordre 1 de $g(y^k_i)$ au voisinage de $\mu^k_{\xi,i}$ est donnée par

$$\begin{aligned} g(y^k_i) &\approx g(\mu^k_{\xi,i}) + (y^k_i - \mu^k_{\xi,i})g'(\mu^k_{\xi,i}) \\ &= \eta^k_{\xi,i} + (y^k_i - \mu^k_{\xi,i})g'(\mu^k_{\xi,i}) \end{aligned}$$

avec η^k_ξ défini par l'égalité (3.1).

Cette fois-ci, en après avoir posé $e^k_i = (y^k_i - \mu^k_{\xi,i})g'(\mu^k_{\xi,i})$, on définit le $k^{\text{ème}}$ vecteur dépendant z^k_ξ comme suit :

$$z^k_{\xi,i} = \eta^k_{\xi,i} + e^k_i, \quad i = 1, \dots, n.$$

On aboutit alors à la définition de q modèles "linéarisés" mixtes $\mathcal{M}^1_\xi, \dots, \mathcal{M}^q_\xi$ dont l'expression générale est donnée par :

$$(\mathcal{M}^k_\xi) : \quad Z^k_\xi = Xu\gamma_k + T\delta_k + U\xi_k + e^k, \quad k = 1, \dots, q$$

aussi peut-on calculer, pour chaque k, l'espérance et la variance de Z^k_ξ conditionnellement à l'effet aléatoire ξ_k. On obtient alors :

$$\begin{aligned} \mathbb{E}(Z^k_\xi|\xi_k) &= Xu\gamma_k + T\delta_k + U\xi_k \\ \mathbb{V}(Z^k_\xi|\xi_k) &= \mathbb{V}(e^k|\xi_k) \\ &= \mathbb{V}\left(\left[(Y^k - \mu^k_\xi)g'(\mu^k_\xi)\right]|\xi_k\right) \\ &= Diag\left(\left[g'_k(\mu^k_{\xi,i})\right]^2 \mathbb{V}(Y^k_i|\xi_k)\right)_{i=1,\dots,n} \quad \text{avec } \mathbb{V}(Y^k_i|\xi_k) = a_{k,i}(\phi_k)b''(\theta_{k,i}) = a_{k,i}(\phi_k)v_k(\mu^k_{\xi,i}). \\ &= W^{-1}_k \quad \text{en notation.} \end{aligned}$$

Étape d'estimation.

L'idée consiste alors à présent à produire q systèmes de Henderson, i.e. un pour chacune des variables réponses $y^1, \dots, y^q$, en faisant explicitement la distinction entre les coefficients γ_k (c'est-à-dire ceux liés à la composante $f = Xu$ retenue dans X) et les δ_k, associés quant à

eux aux variables additionnelles T. Pour chaque $k \in \{1, \ldots, q\}$, on considérera que le couple $\begin{pmatrix} Z_\xi^k \\ \xi_k \end{pmatrix}$ est distribué de la manière suivante :

$$\begin{pmatrix} Z_\xi^k \\ \xi_k \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} Xu\gamma_k + T\delta_k \\ 0 \end{pmatrix}, \begin{pmatrix} W_k^{-1} + UD_kU' & UD_k \\ D_kU' & D_k \end{pmatrix} \right)$$

avec D_k la matrice diagonale par blocs $Diag\left(\sigma_{k,j}^2 A_{k,j}\right)_{j=1,\ldots,K}$ de telle sorte que les effets aléatoires liés à la $k^{\text{ème}}$ variable ξ_k soient distribués selon une loi Normale d'espérance 0 et de matrice de variance-covariance D_k.

Après avoir rassemblé la composante retenue Xu et les variables additionnelles T au sein d'une même matrice $[Xu|T]$, les systèmes de Henderson obtenus sont alors, pour $k = 1, \ldots, q$:

$$\begin{pmatrix} [Xu|T]' W_k [Xu|T] & [Xu|T]' W_k U \\ U' W_k [Xu|T] & U'W_kU + D_k^{-1} \end{pmatrix} \begin{pmatrix} \gamma_k \\ \delta_k \\ \xi_k \end{pmatrix} = \begin{pmatrix} [Xu|T]' W_k z_\xi^k \\ U' W_k z_\xi^k \end{pmatrix} \tag{3.2}$$

3.3 Précisions sur le critère à maximiser par PING dans notre contexte, à γ_k, δ_k, σ_k^2 fixés.

La fonction à maximiser à l'aide de l'algorithme PING s'écrit toujours :

$$S_T(u) = [\psi_{X,T}(u)]^{1-s} [\phi_{N,\Omega,l}(u)]^s$$

Nous apportons ici quelques précisions et justifications sur la formes des expressions ψ et ϕ intervenant dans ce critère. Pour plus de lisibilité, nous omettons les indices ξ relatifs aux vecteurs dépendants z_ξ^k.

i) Conditionnellement à l'effet aléatoire ξ, les modèles $(\mathcal{M}_\xi^k)$ s'interprètent comme des GLM. Par conséquent, la part du critère liée à la qualité de l'ajustement s'écrit toujours :

$$\psi_{X,T}(u) = \sum_{k=1}^{q} \|z^k\|_{W_k}^2 \cos_{W_k}^2(z^k; \langle Xu, T\rangle),$$

mais ici, les matrices W_k sont telles que $W_k^{-1} = \mathbb{V}(Z_\xi^k|\xi_k)$. On peut alors noter :

$$W_k = Diag\left(\frac{1}{\left[g_k'(\mu_{\xi,i}^k)\right]^2 a_{k,i}(\phi_k) v_k(\mu_{\xi,i}^k)} \right)_{i=1,\ldots,n}$$

ii) Quant à sa partie concernant la force structurelle de X, on posera (voir (6.1.3) pour plus de détails) :

$$\phi_{N,\Omega,l}(u) = \left(\sum_{j=1}^{p} \omega_j (u'X'Wx^j x^{j'}WXu)^l \right)^{\frac{1}{l}}$$

iii) Notons que dans les simulations qui viendront ainsi que pour traiter les données réelles, nous prendrons $l = 4$ afin que les faisceaux captés soient suffisamment "locaux" et $s = 0.5$ afin de conserver un équilibre entre la qualité de l'ajustement et la force structurelle.

3.4 Quelques pseudo-codes illustratifs.

3.4.1 Présentation du pseudo-code général pour la recherche de la première composante $f^{[1]}$.

Pour ce premier algorithme, aucune convention particulière n'est à adopter. En effet, comme on ne s'intéresse pour le moment qu'à la première composante, il n'y a pas de contrainte d'orthogonalité à considérer.

PROCÉDURE MIXED-SCGLR, cas d'une composante.

INITIALISATION des paramètres :
Proposer des valeurs pour :
- les vecteurs courants $z^k, \forall k \in \{1, \ldots, q\}$
- les matrices de poids $W_k, \forall k \in \{1, \ldots, q\}$
- les matrices relatives aux composantes de la variance

$$D_k = Diag\left(\sigma^2_{k,j} A_{k,j}\right)_{j=1,\ldots,K} \forall k \in \{1, \ldots, q\}$$

repeat

1. Appel à l'algorithme PING :
Poser, via l'algorithme PING,

$$u^{[1]} \leftarrow \begin{cases} \arg\max S_T(u) \\ \text{sous la contrainte :} \quad u'M^{-1}u = 1 \end{cases}$$

Poser alors $f^{[1]} \leftarrow Xu^{[1]}$,
et $\tilde{T} \leftarrow \left[f^{[1]}, T\right]$

2. Résolution des systèmes de Henderson :
Pour $k = 1, \ldots, q$, résoudre les systèmes :

$$\begin{pmatrix} \tilde{T}'W_k\tilde{T} & \tilde{T}'W_kU \\ U'W_k\tilde{T} & U'W_kU + D_k^{-1} \end{pmatrix} \begin{pmatrix} \gamma_k \\ \delta_k \\ \xi_k \end{pmatrix} = \begin{pmatrix} \tilde{T}'W_kz^k \\ U'W_kz^k \end{pmatrix} \text{ d'inconnues } \gamma_k, \delta_k \text{ et } \xi_k.$$

3. Calcul des composantes de la variance :
Poser alors , pour $k = 1, \ldots, q$:

$$\sigma^2_{k,j} \leftarrow \frac{\xi'_{k,j} A^{-1}_{k,j} \xi_{k,j}}{q_j - \frac{tr\left(A^{-1}_{k,j} C^*_{k,jj}\right)}{\sigma^2_{k,j}}}, \quad \forall j \in \{1, \ldots, K\}$$

4. Mise à jour :
Poser alors, pour tout $k \in \{1, \ldots, q\}$:
$\eta^k \leftarrow X u^{[1]} \gamma_k + T\delta_k + U\xi_k$
$\mu_{k,i} \leftarrow g_k^{-1}(\eta_i^k),\ i = 1, \ldots, n$
$z_i^k \leftarrow \eta_i^k + (y_i^k - \mu_{k,i}) g'_k(\mu_{k,i}),\ i = 1, \ldots, n$
$W_k \leftarrow Diag\left(\frac{1}{[g'_k(\mu_{k,i})]^2 a_{k,i}(\phi_k) v_k(\mu_{k,i})}\right)_{i=1,\ldots,n}$
$D_k \leftarrow Diag\left(\sigma^2_{k,j} A_{k,j}\right)_{j=1,\ldots,K}$

until Convergence

3.4.2 Généralisation pour la recherche des H premières composantes.

Ici par contre, plusieurs conventions doivent être rappelées, qui sont essentiellement les mêmes que pour l'algorithme SCGLR sans effets aléatoires (1.3.4).

i) Tout d'abord, nous posons $T^{[0]} = T$ afin que la recherche de la première composante $u^{[1]}$ s'effectue bien au moyen de la matrice des covariables additionnelles initiale T.

ii) $F^{[0]}$ est considérée comme une matrice "vide".

iii) Dans notre cas, $\forall k \in \{1, \ldots, q\}$, les matrices $D_k^{[0]}$ sont nulles.

iv) Enfin, par défaut, $W = \dfrac{1}{n} Id_n$.

PROCÉDURE MIXED-SCGLR

for h = 1 à H **do**

INITIALISATION des paramètres :
Proposer des valeurs pour :
les vecteurs courants $z^{k[h]}, \forall k \in \{1, \ldots, q\}$
les matrices de poids $W_k^{[h]}, \forall k \in \{1, \ldots, q\}$
les matrices relatives aux composantes de la variance

$$D_k^{[h]} = Diag\left(\sigma_{k,j}^{[h]\,2} A_{k,j}\right)_{j=1,\ldots,K} \quad \forall k \in \{1, \ldots, q\}$$

repeat

1. Appel à l'algorithme PING :
Poser, via l'algorithme PING,

$$u^{[h]} \leftarrow \begin{cases} \arg\max S_{T^{[h-1]}}(u) \\ \text{sous les contraintes :} \quad u'M^{-1}u = 1 \text{ et } \left(X'WF^{[h-1]}\right)' u = 0 \end{cases}$$

Poser alors $f^{[h]} \leftarrow Xu^{[h]}$ et
$\tilde{T}^{[h]} \leftarrow \left[f^{[h]}, T\right]$

2. Résolution des systèmes de Henderson :
Pour $k = 1, \ldots, q$, résoudre les systèmes :

$$\begin{pmatrix} \tilde{T}^{[h]'} W_k^{[h]} \tilde{T}^{[h]} & \tilde{T}^{[h]'} W_k^{[h]} U \\ U' W_k^{[h]} \tilde{T}^{[h]} & U' W_k^{[h]} U + {D_k^{[h]}}^{-1} \end{pmatrix} \begin{pmatrix} \gamma_k^{[h]} \\ \delta_k^{[h]} \\ \xi_k^{[h]} \end{pmatrix} = \begin{pmatrix} \tilde{T}^{[h]'} W_k^{[h]} z^{k[h]} \\ U' W_k^{[h]} z^{k[h]} \end{pmatrix}$$

d'inconnues $\gamma_k^{[h]}$, $\delta_k^{[h]}$ et $\xi_k^{[h]}$.

3. Calcul des composantes de la variance :
Poser alors , pour $k = 1, \ldots, q$:

$$\sigma_{k,j}^{[h]\,2} \leftarrow \frac{\xi_{k,j}^{[h]\,\prime} A_{k,j}^{-1} \xi_{k,j}^{[h]}}{q_j - \frac{tr\left(A_{k,j}^{-1} C_{k,jj}^{[h]*}\right)}{\sigma_{k,j}^{[h]\,2}}}, \quad \forall j \in \{1, \ldots, K\}$$

4. Calcul des nouveaux vecteurs courants :
Poser alors, pour tout $k \in \{1, \ldots, q\}$:
$\eta^{k[h]} \leftarrow X u^{[h]} \gamma_k^{[h]} + T \delta_k^{[h]} + U \xi_k^{[h]}$
$\mu_{k,i}^{[h]} \leftarrow g_k^{-1}(\eta_i^{k[h]}),\ i = 1, \ldots, n$
$z_i^{k[h]} \leftarrow \eta_i^{k[h]} + (y_i^k - \mu_{k,i}^{[h]}) g_k'(\mu_{k,i}^{[h]}),\ i = 1, \ldots, n$
$W_k^{[h]} \leftarrow Diag\left(\frac{1}{\left[g_k'(\mu_{k,i}^{[h]})\right]^2 a_{k,i}(\phi_k) v_k(\mu_{k,i}^{[h]})}\right)_{i=1,\ldots,n}$
$D_k^{[h]} \leftarrow Diag\left(\sigma_{k,j}^{[h]\,2} A_{k,j}\right)_{j=1,\ldots,K}$

until Convergence

Récupération de la $h^{ème}$ composante : Poser $f^{[h]} = X u^{[h]}$.
Mise à jour de F : Poser $F^{[h]} = \left[F^{[h-1]} | f^{[h]}\right]$.
Mise à jour de T : Poser $T^{[h]} = \left[T | F^{[h]}\right]$.

end for

Chapitre 4

Premiers essais numériques sur données simulées.

4.1 Plan de simulation.

Dans le but d'évaluer les performances de l'algorithme proposé, nous décidons de considérer :

i) Une unique variable latente ϕ, qui sera l'unique structure principale des covariables X.

ii) Un bloc X de $p = 50$ régresseurs contenant un faisceau X_1 de 25 variables structuré autour de ϕ d'une part, et 25 variables de bruit pur d'autre part.

iii) Un bloc additionnel T constitué de r variables $t^1, \ldots, t^r$ qui pourront être corrélées entre elles et avec ϕ de façon réglable.

Pour faciliter la compréhension de la procédure, nous adoptons les notations suivantes :

i) Le nombre d'individus considérés sera noté N, et le nombre de répétitions associées à chaque individu sera noté R. Donc le nombre total d'observations, correspondant au nombre de lignes des tableaux créés, vaudra $n = N \times R$.

ii) On contrôlera la largeur du faisceau X_1 via un paramètre τ^2. Dans les différents tests, nous prendrons $\tau^2 \in \{0.5, 1, 1.5\}$.

iii) Nous nous permettons de considérer autant d'effets aléatoires que de variables réponses, qui seront au nombre de q. Par conséquent, en notant σ_k l'écart-type de l'effet aléatoire ξ_k, les valeurs choisies seront :

$$\forall k \in \{1, \ldots, q\}, \sigma_k \in \{0.03, 0.3, 0.5\}.$$

Nous nous limiterons à des valeurs volontairement faibles pour des raisons que nous exposerons ensuite.

Pour plus de clarté, voici la manière exacte qui a été utilisée pour générer les prédicteurs linéaires. Premièrement, la variable latente est choisie normale :

$$\phi \sim \mathcal{N}_n(0, Id)$$

Pour générer les 25 premières variables de X structurées autour de ϕ, nous posons, pour $j \in \{1, \ldots, 25\}$:

$$x^j = \frac{1}{\sqrt{1+\beta_j^2\tau^2}}(\phi + \beta_j \varepsilon^j),$$
$$\text{avec } \beta_j \sim \mathcal{U}_{[0,1]} \text{ et } \varepsilon^j \sim \mathcal{N}_n(0, \tau^2 Id)$$

De cette façon, on aboutit à $\mathbb{V}(x^j) = 1 \quad \forall j \in \{1, \ldots, 25\}$.

Les 25 autres variables présentes dans X, quant à elles, sont prises normales et indépendantes de ϕ. On pose ainsi, pour $j \in \{26, \ldots, 50\}$:

$$x^j \sim \mathcal{N}_n(0, Id)$$

Enfin, pour ce qui est des prédicteurs linéaires, q étant le nombre de variables réponses, on pose pour $k = 1, \ldots, q$: $\alpha_k \sim \mathcal{N}_1(0, 0.3^2)$ et $\delta_k \sim \mathcal{N}_r(0, 0.3^2 Id)$, pour finalement obtenir :

$$\eta^k = \alpha_k \phi + T\delta_k + U\xi_k,$$

avec bien sûr $\xi_k \sim \mathcal{N}_N(0, \sigma_k^2 Id)$.

Notons que la variance des coefficients α_k et δ_k est choisie également volontairement faible afin d'obtenir des données manipulables et réalistes, dans la mesure où l'on souhaite générer des données selon des lois de Poisson - lien log. Si l'on ne prend pas cette précaution, l'amplitude des données obtenue est trop grande : l'algorithme encourrait alors des risques d'"explosion" et de non-convergence. Les effets aléatoires simulés doivent donc respecter la même contrainte : il faut que leurs variances respectives soient suffisamment élevées pour que leur introduction dans le modèle soit significative, mais pas trop néanmoins, afin d'éviter l'"explosion". De plus, il faut contourner le cas extrême où l'effet aléatoire prendrait le dessus sur les autres éléments constituant le prédicteur linéaire. C'est pourquoi nous sommes contraints de choisir des valeurs des paramètres σ_k apparemment faibles.

Finalement, pour générer les variables réponses Y, nous choisissons une modélisation selon une loi de Poisson avec un lien log. On les simulera alors de la manière suivante :

$$\forall k \in \{1, \ldots, q\}, \forall i \in \{1, \ldots, n\}, y_i^k \sim \mathcal{P}(\exp(\eta_i^k)).$$

Notons enfin que dans toute la suite des simulations, nous choisissons $q = 20$. Nous aurons donc 20 variables $y^1, \ldots, y^{20}$ à modéliser.

Pour montrer que les valeurs choisies pour les écarts-type de l'effet aléatoire sont cohérentes, nous présentons le tableau suivant, qui a vocation à représenter l'impact relatif des différentes parties intervenant dans notre prédicteur linéaire. Dans la plupart des simulations qui vont suivre, les intervalles de fluctuation seront les suivants :

Quantités intervenant dans le prédicteur linéaire	Intervalles de fluctuation moyen
$\alpha_k \phi$	$[-2, 2]$
$T\delta_k$	$[-2, 2]$
$U\xi_k$, avec $\sigma_k = 0.03$	$[-0.2, 0.2]$
$U\xi_k$, avec $\sigma_k = 0.3$	$[-1, 1]$
$U\xi_k$, avec $\sigma_k = 0.5$	$[-2, 2]$

FIGURE 4.1: *Intervalles de fluctuation moyen des éléments du prédicteur linéaire.*

Comme nous pouvons le constater, dans le cadre de nos simulations, prendre un effet aléatoire d'écart-type 0.03 a un faible impact relatif sur le prédicteur linéaire. Par contre, avec un écart-type égal à 0.3, l'effet aléatoire introduit un impact non-négligeable mais raisonnable. Un écart-type de 0.5 constitue déjà un cas limite où l'effet aléatoire n'est plus du tout raisonnable : il a le même poids que les effets fixes.

4.2 Nos choix pour les critères d'arrêts de l'algorithme et pour les quantités servant à juger de ces performances.

Notre critère d'arrêt pour la procédure mise en œuvre et dont le pseudo-code a été présenté dans la section (2.1) repose sur la stabilité simultanée de plusieurs quantités, comme nous le détaillons ci-après :

i) Dans un premier temps, nous demandons que le vecteur u obtenu par l'algorithme PING soit suffisamment stable. Cette stabilité sera en réalité garante de la stabilité de la composante $f = Xu$. Cela se manifeste par la première condition :

$$\|u_{new} - u_{old}\|^2 < 10^{-6} \tag{4.1}$$

ii) La deuxième chose que nous demandons est une stabilité sur les estimations des composantes de la variance. En notant $\sigma = (\sigma_1, \ldots, \sigma_q)'$, nous demandons que soit vérifiée la condition suivante :

$$\frac{\|\hat{\sigma}^2_{new} - \hat{\sigma}^2_{old}\|^2}{\|\hat{\sigma}^2_{old}\|^2} < 10^{-6} \tag{4.2}$$

iii) La dernière chose que nous imposons pour l'arrêt de la procédure est la vérification d'une certaine stabilité au niveau du prédicteur linéaire faisant intervenir les variables additionnelles T. Notre choix s'est plus précisément porté sur une stabilité moyenne sur les 20 variables, comme le montre la condition suivante :

$$\frac{1}{q}\sum_{k=1}^{q}\left(\frac{\|T\hat{\delta}_{new} - T\hat{\delta}_{old}\|^2}{\|T\hat{\delta}_{old}\|^2}\right) < 10^{-6} \tag{4.3}$$

Pour chacune des situations sur lesquelles la procédure va être testée, 100 échantillons seront simulés. Nous introduisons dans cette optique les notations suivantes :

i) $u^{[s]}$ désignera notre calcul du vecteur u obtenu sur l'échantillon s.

ii) Pour $k = 1, \ldots, 20$ et pour $s = 1, \ldots, 100$, $\hat{\delta}_k^{[s]}$ et $\hat{\sigma}_k^{[s]}$ désigneront les estimations respectives de δ_k et σ_k (relatives à la $k^{ème}$ variable) obtenues pour le $s^{ème}$ échantillon.

iii) Enfin, $Xu^{[s]}\hat{\gamma}_k^{[s]}$ représentera notre estimation du terme $\alpha_k\phi$ relative à la $k^{ème}$ variable et obtenue avec l'échantillon s.

Les quantités qui nous serviront à évaluer les performances de l'algorithme sont les suivantes :

i) En premier lieu, il faudrait que l'algorithme identifie bien l'existence de l'unique variable latente ϕ. Nous évaluerons cela par la quantité :

$$\bar{\rho^2} = \frac{1}{100}\sum_{s=1}^{100}\rho^2\left(\phi, Xu^{[s]}\right)$$

ii) Ensuite, la proximité entre les quantités simulées et estimées du premier terme du prédicteur linéaire doit être évaluée. Cette évaluation s'effectuera à l'aide de la quantité :

$$d_1 = \frac{1}{q}\sum_{k=1}^{q}\left(\frac{1}{100}\sum_{s=1}^{100}\|\alpha_k\phi - Xu^{[s]}\hat{\gamma}_k^{[s]}\|_W^2\right)$$

iii) La même idée appliquée au deuxième terme du prédicteur linéaire nous pousse à définir d_2 de la façon suivante :

$$d_2 = \frac{1}{q}\sum_{k=1}^{q}\left(\frac{1}{100}\sum_{s=1}^{100}\|T\delta_k - T\hat{\delta}_k^{[s]}\|_W^2\right)$$

iv) Enfin, nous voulons évaluer la qualité de nos estimations des composantes de la variance. Par conséquent, le dernier critère nous permettant d'évaluer les performances de la procédure proposée sera défini ainsi :

$$d_3 = \frac{1}{q}\sum_{k=1}^{q}\left(\frac{1}{100}\sum_{s=1}^{100}\|\sigma_k - \hat{\sigma}_k^{[s]}\|^2\right)$$

4.3 Résultats obtenus.

4.3.1 Cas d'une seule covariable additionnelle indépendante de la variable latente.

Dans ce premier cas, on pose $T = t^1$, avec $t^1 \sim \mathcal{N}_n(0, Id)$. Ainsi, pour $k = 1, \ldots, q, \delta_k \sim \mathcal{N}_1(0, 0.3^2)$. Dans nos simulations, nous considérons $N = 50$ individus et $R = 10$ répétitions,

en faisant varier la largeur du faisceau structurant dans X de telle sorte que τ^2 prenne successivement les valeurs 0.5, 1 et 1.5. Nous nous limiterons ici aux cas où l'effet aléatoire, pour chacune des variables réponses, admet pour écart-type des valeurs raisonnables, à savoir 0.03 et 0.3.

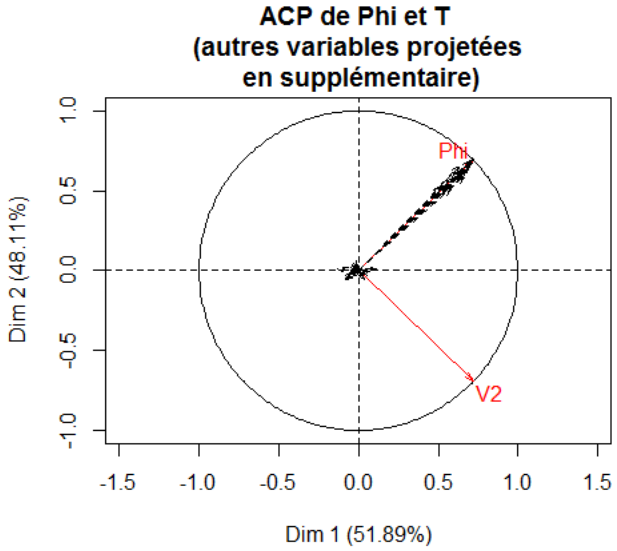


FIGURE 4.2: *Sur le premier plan factoriel présenté ici, l'orthogonalité entre ϕ et T est clairement visible. Nous voyons aussi apparaitre le faisceau des 25 variables de X autour de ϕ et les 25 variables de bruit pur, qui vivent essentiellement dans l'orthogonal de ce premier plan factoriel.*

Les tableaux obtenus sont présentés ci-dessous.

τ^2	0.5	0.5
σ_k	0.03	0.3
$\bar{\rho}^2$	0.994	0.994
$d_1 = \frac{1}{500}\times$	1.08	1.12
$d_2 = \frac{1}{500}\times$	0.88	0.96
$d_3 = 10^{-4}\times$	1.6	14

τ^2	1	1
σ_k	0.03	0.3
$\bar{\rho}^2$	0.988	0.988
$d_1 = \frac{1}{500}\times$	1.27	1.37
$d_2 = \frac{1}{500}\times$	0.85	0.92
$d_3 = 10^{-4}\times$	1.8	14

τ^2	1.5	1.5
σ_k	0.03	0.3
$\bar{\rho^2}$	0.983	0.983
$d_1 = \frac{1}{500}\times$	1.39	1.51
$d_2 = \frac{1}{500}\times$	0.88	0.93
$d_3 = 10^{-4}\times$	1.8	14

- Premièrement, nous observons que lorsque nous augmentons raisonnablement l'écart-type de l'effet aléatoire, l'ensemble des estimations obtenues se dégrade, car les distances d_1, d_2 et d_3 augmentent. Par contre, la corrélation carrée moyenne entre ϕ et Xu n'est pas sensiblement affectée par cette augmentation.
- Ensuite, comme l'on pouvait le prévoir, lorsque la largeur du faisceau structurant de X augmente, la corrélation carrée moyenne entre ϕ et Xu diminue. Cela semble logique : le faisceau étant plus large, la variable latente ϕ qui la structure est moins perceptible. Comme la distance d_1 dépend entre autre de notre estimation de la première composante de X, elle se voit augmentée également. Néanmoins, la dégradation est très faible.
- Au contraire, l'augmentation de la largeur du faisceau ne semble pas affecter notre estimation des $T\delta_k$ et des σ_k, qui sont assez stables.

4.3.2 Cas de 5 covariables additionnelles possiblement corrélées, indépendantes de la variable latente.

Pour modéliser une corrélation entre les différentes variables présentes dans T, nous avons besoin de définir des variables "instrumentales", qui seront également au nombre de 5. On pose alors :

$$\psi^1, \psi^2, \ldots, \psi^5 \overset{ind.}{\sim} \mathcal{N}_n(0, Id)$$

En considérant $x \in [0,1]$ un paramètre ayant pour but de régler le degré de dépendance des variables additionnelles, on pose finalement :

$$\forall l \in \{1,5\}, t^l = \frac{1}{\sqrt{x^2 + (1-x)^2}} \left(x\psi^l + (1-x)\psi^{(l+1)[5]} \right).$$

où [5] désigne le modulo 5.

Nous avons choisi de traiter le cas $x = 0.9$, qui représente une dépendance très modérée entre les variables additionnelles, et le cas $x = 0.5$, qui induit une structure de corrélation assez forte au sein des variables T. Les tableaux récapitulatifs sont alors les suivants, toujours pour $N = 50$ et $R = 10$:

x	0.9	0.9	0.5	0.5
τ^2	0.5	0.5	0.5	0.5
σ_k	0.03	0.3	0.03	0.3
$\bar{\rho^2}$	0.994	0.994	0.994	0.994
$d_1 = \frac{1}{500}\times$	0.90	0.98	0.89	1.00
$d_2 = \frac{1}{500}\times$	3.67	4.26	3.92	4.36
$d_3 = 10^{-4}\times$	0.85	19	1.0	18

x	0.9	0.9	0.5	0.5
τ^2	1	1	1	1
σ_k	0.03	0.3	0.03	0.3
$\bar{\rho^2}$	0.992	0.992	0.992	0.992
$d_1 = \frac{1}{500}\times$	0.92	0.99	1.02	1.12
$d_2 = \frac{1}{500}\times$	3.73	4.27	4.01	4.28
$d_3 = 10^{-4}\times$	0.81	18	1.0	17

x	0.9	0.9	0.5	0.5
τ^2	1.5	1.5	1.5	1.5
σ_k	0.03	0.3	0.03	0.3
$\bar{\rho^2}$	0.982	0.982	0.982	0.982
$d_1 = \frac{1}{500}\times$	1.12	1.22	1.16	1.27
$d_2 = \frac{1}{500}\times$	3.74	4.27	3.89	4.37
$d_3 = 10^{-4}\times$	0.80	17	1.0	16

- Dans un premier temps, il est bon de signaler que "localement", c'est-à-dire pour x fixé, les mêmes phénomènes que ceux décrits dans la section (4.3.1) s'observent : l'augmentation des distances d_1, d_2 et d_3 avec l'augmentation de l'écart-type de l'effet aléatoire ; tout comme l'augmentation de d_1 et la stabilité de d_2 et d_3 lorsque l'on augmente la largeur du faisceau structurant de X.
- Par contre, le fait nouveau que nous apprend ce jeu de simulations est le suivant : lorsqu'on augmente la dépendance entre les variables additionnelles T, la distance d_2 a tendance à augmenter. Ceci semble logique : lorsque la corrélation entre les variables au sein de T devient trop importante, une confusion plus grande entre ces variables apparaît et les estimations associées sont donc plus instables.
- Par contre, l'augmentation du degré de dépendance au sein de T ne semble pas ou peu affecter la distance d_1, qui reste assez stable pour une même valeur de τ^2 et σ_k et lorsque x varie.

4.3.3 Cas d'une seule covariable additionnelle corrélée à la variable latente.

Dans ce cas, on a besoin de simuler uniquement une variable "instrumentale" $t'^1 \sim \mathcal{N}_n(0, Id)$. Le degré de corrélation avec la variable latente ϕ est alors réglé à l'aide du paramètre θ, en définissant T de la façon suivante :

$$T = t^1 = \sin(\theta)\phi + \cos(\theta)t'^1$$

Le cas $\theta = 0$ correspondra à l'absence de corrélation entre ϕ et T ; le choix $\theta = \frac{\pi}{6}$ représentera une corrélation modérée et enfin $\theta = \frac{\pi}{3}$ symbolisera une corrélation forte.

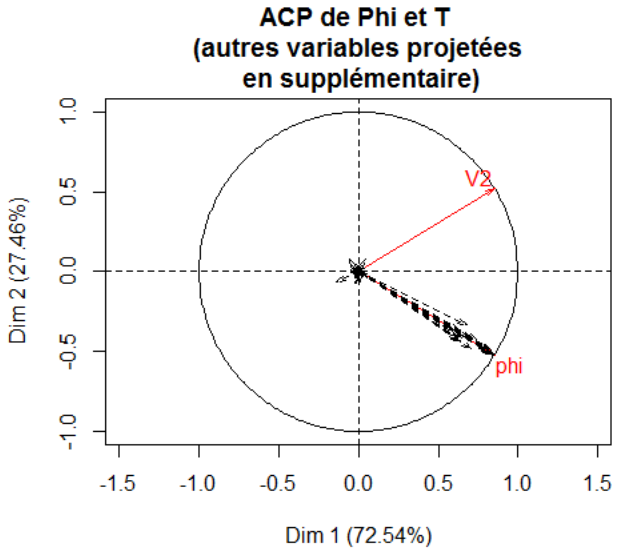


FIGURE 4.3: *Voici le premier plan factoriel avec $\theta = \pi/3$. Contrairement à la figure (4.2), une corrélation visible existe entre la variable latente ϕ et les covariables additionnelles T. On s'attendra alors à observer une dégradation des estimations en particulier pour cette valeur de θ.*

Pour $N = 50$ et $R = 10$, nous obtenons alors les tableaux :

θ	0	0	$\frac{\pi}{6}$	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{3}$
τ^2	0.5	0.5	0.5	0.5	0.5	0.5
σ_k	0.03	0.3	0.03	0.3	0.03	0.3
$\bar{\rho^2}$	0.993	0.993	0.993	0.993	0.993	0.993
$d_1 = \frac{1}{500}\times$	1.12	1.21	1.32	1.38	1.53	1.56
$d_2 = \frac{1}{500}\times$	0.85	0.90	1.04	1.13	3.23	3.27
$d_3 = 10^{-4}\times$	1.7	14	1.6	14	1.8	13

θ	0	0	$\frac{\pi}{6}$	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{3}$
τ^2	1	1	1	1	1	1
σ_k	0.03	0.3	0.03	0.3	0.03	0.3
$\bar{\rho^2}$	0.990	0.990	0.990	0.990	0.990	0.990
$d_1 = \frac{1}{500}\times$	1.23	1.25	1.41	1.46	3.50	3.51
$d_2 = \frac{1}{500}\times$	0.87	0.96	1.07	1.09	3.16	3.20
$d_3 = 10^{-4}\times$	1.7	13	1.7	14	1.8	14

θ	0	0	$\frac{\pi}{6}$	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{3}$
τ^2	1.5	1.5	1.5	1.5	1.5	1.5
σ_k	0.03	0.3	0.03	0.3	0.03	0.3
$\overline{\rho^2}$	0.987	0.987	0.987	0.987	0.987	0.987
$d_1 = \frac{1}{500}\times$	1.28	1.35	1.50	1.60	3.60	3.65
$d_2 = \frac{1}{500}\times$	0.87	0.93	1.03	1.10	3.12	3.22
$d_3 = 10^{-4}\times$	1.7	13	1.7	14	1.7	13

- Pour des valeurs de θ fixées, des phénomènes comparables à ceux observés dans le premier cas s'observent.
- Nous observons ici que lorsque la dépendance entre T et ϕ augmente, les deux distances d_1 et d_2 semblent dégradées. Ceci semble cohérent : dans le cas précédent, la seule confusion qui apparaissait se situait au sein des variables T, ce qui avait pour unique conséquence une augmentation de d_2. Mais là par contre, une confusion entre T et ϕ apparaît, ce qui engendre logiquement une dégradation simultanée de d_1 et d_2.
- Comme précédemment, l'estimation des écart-types de l'effet aléatoire ne semble pas être impactée par la dépendance créée entre T et ϕ.

4.3.4 Cas de 5 covariables additionnelles possiblement corrélées, et corrélées avec la variable latente.

Avec les variables "instrumentales" $\psi^1, \ldots, \psi^5$ définies dans la section (4.3.2), on pose :

$$\forall l \in \{1, 5\}, t^l = \cos(\theta)\psi^l + \sin(\theta)\phi, \quad l = 1, \ldots, 5$$

Il s'agit du pire des cas que l'on puisse imaginer car pour des valeurs de $\theta \neq 0$, chacune des variables dans T est corrélée avec ϕ, et la corrélation avec ϕ induit également une corrélation entre les variables additionnelles. Comme dans le cas d'une seule variable additionnelle potentiellement corrélée avec ϕ, une valeur de θ proche de 0 indique une faible corrélation entre T et ϕ tandis qu'une valeur de θ se rapprochant de $\frac{\pi}{2}$ dénote une très forte corrélation, non seulement au sein des variables additionnelles, mais également entre chacune des variables additionnelles et ϕ. Pour illustrer les résultats obtenus, nous prenons toujours le cas $N = 50$ et $R = 10$.

θ	0	0	$\frac{\pi}{6}$	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{3}$
τ^2	0.5	0.5	0.5	0.5	0.5	0.5
σ_k	0.03	0.3	0.03	0.3	0.03	0.3
$\overline{\rho^2}$	0.993	0.993	0.993	0.993	0.993	0.993
$d_1 = \frac{1}{500}\times$	0.85	1.00	1.28	1.49	9.36	15.09
$d_2 = \frac{1}{500}\times$	2.95	3.95	3.55	4.49	9.77	16.79
$d_3 = 10^{-4}\times$	0.56	26	0.61	25	6.7	140

θ	0	0	$\frac{\pi}{6}$	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{3}$
τ^2	1	1	1	1	1	1
σ_k	0.03	0.3	0.03	0.3	0.03	0.3
$\bar{\rho^2}$	0.985	0.985	0.985	0.985	0.985	0.985
$d_1 = \frac{1}{500}\times$	1.13	1.29	1.55	1.76	9.61	16.26
$d_2 = \frac{1}{500}\times$	3.05	3.80	3.54	4.51	9.43	15.30
$d_3 = 10^{-4}\times$	0.55	23	0.59	24	5.7	120

θ	0	0	$\frac{\pi}{6}$	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{3}$
τ^2	1.5	1.5	1.5	1.5	1.5	1.5
σ_k	0.03	0.3	0.03	0.3	0.03	0.3
$\bar{\rho^2}$	0.982	0.982	0.982	0.982	0.982	0.982
$d_1 = \frac{1}{500}\times$	1.24	1.31	1.60	1.81	11.87	20.69
$d_2 = \frac{1}{500}\times$	2.87	3.74	3.60	4.45	9.52	14.41
$d_3 = 10^{-4}\times$	0.56	22	0.60	20	5.6	180

- La confusion injectée étant plus forte que dans le cas précédent, il est normal d'observer les mêmes augmentations, mais dans des proportions plus importantes.
- Cependant, remarquons tout de même que malgré cela, l'estimation obtenue de la composante semble toujours être d'une très bonne qualité.

En prenant un peu de recul, les résultats semblent cohérents, et aucun contre-sens particulier n'est à signaler... Cependant, notons qu'il est curieux de voir la stabilité de la quantité $\bar{\rho^2}$ lorsqu'on augmente l'écart-type de l'effet aléatoire. En effet, on aurait pu s'attendre à ce que l'estimation de la composante soit dégradée lorsqu'on injecte une part plus conséquente d'aléa. Des tests complémentaires ont permis de voir que cette apparente stabilité est due au nombre important de variables qui structurent le faisceau X_1 autour de ϕ. En effet, le faisceau simulé comportant 25 variables, l'estimation de la variable latente correspondante ne se voit donc que très peu perturbée par l'augmentation de l'écart-type de l'effet aléatoire. En réduisant le faisceau X_1 à 10 variables voire 5, une détérioration significative de $\bar{\rho^2}$ est en effet observée.

4.3.5 Simulations complémentaires.

Nous tenions à revenir sur le choix volontairement faible de l'écart-type de l'effet aléatoire. La double contrainte avec laquelle nous avions à traiter pour la simulation de données Poisson-canonique n'apparaît plus dans le cas gaussien-canonique. En effet, il n'y a plus de risque d'"explosion" car les matrices de poids W_k restent constantes lors des itérations. Nous avons ainsi pu simuler des données gaussiennes avec des valeurs de σ_k dans l'intervalle $[0.5, 1.5]$, avec des résultats comparables à ceux trouvés dans le cadre des données Poisson-canonique.

De plus, d'autres simulations, dont le but était d'analyser l'évolution des quantités $\bar{\rho^2}$, d_1, d_2 et d_3 en faisant varier le nombre d'individus N et le nombre de répétitions R, ont été menées.

i) Lorsque le nombre d'individus N augmente, une légère augmentation de $\bar{\rho}^2$ est observée, tandis qu'une diminution significative de chacune des distances d_1 et d_2 (qui sont des distances au sens de la matrice $W = (NR)^{-1}Id_{NR}$) se produit.

ii) Et comme l'on pouvait s'y attendre, une augmentation de R induit une diminution significative de d_3.

Enfin, il est vrai que toutes les simulations effectuées jusqu'à présent n'impliquaient qu'une seule variante latente ϕ dans X. Seule la largeur du faisceau pouvait varier. Nous avons alors étendu le plan de simulation dans le cas où X était structuré par trois variables latentes orthogonales ϕ_1, ϕ_2, ϕ_3. Des résultats similaires furent observés.

Mais nous voulions surtout vérifier que les composantes calculées $f^1 = Xu^1$, $f^2 = Xu^2$ et $f^3 = Xu^3$ se rapprochaient bien des variables latentes simulées.

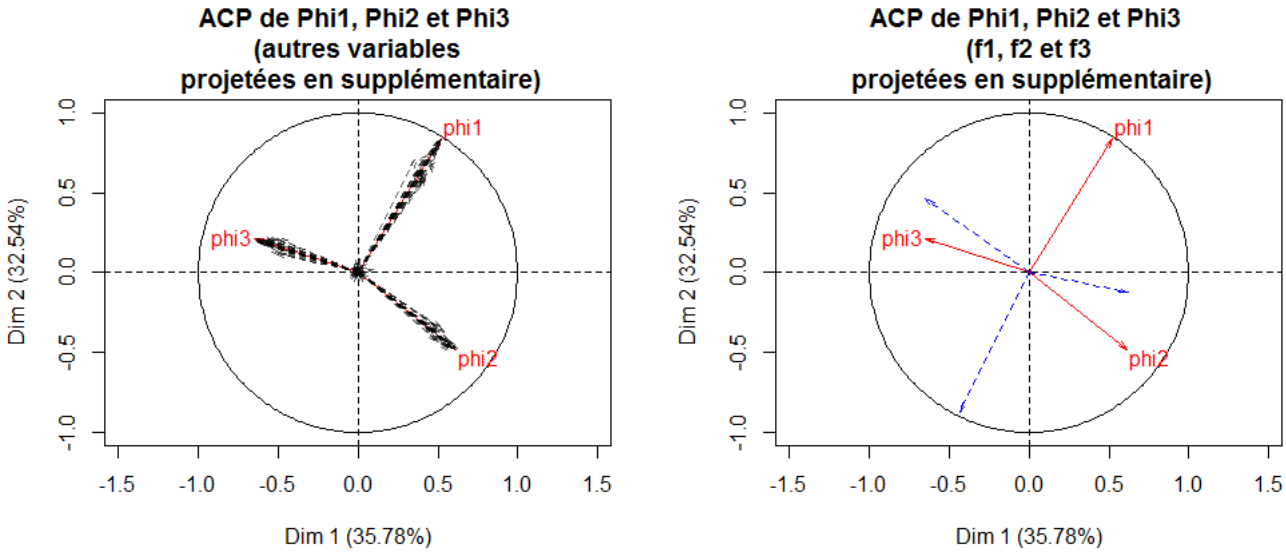


FIGURE 4.4: *Les graphes ci-dessus représentent le premier plan factoriel de l'ACP de* (ϕ_1, ϕ_2, ϕ_3). *Sur le premier d'entre eux, nous avons projeté les variables de* X *en supplémentaires (25 variables autour de chacune des variables latentes avec une largeur de faisceau* $\tau^2 = 1$ *et 25 variables de bruit pur). Sur le deuxième, nous avons projeté en supplémentaire les composantes calculées* f^1, f^2 *et* f^3 *pour évaluer leur proximité avec* ϕ_1, ϕ_2 *et* ϕ_3.

Pour plus de précision, nous exposons ici le tableaux des corrélations carrées entre les différentes variables latentes simulées et les composantes calculées avec notre procédure généralisée à trois composantes.

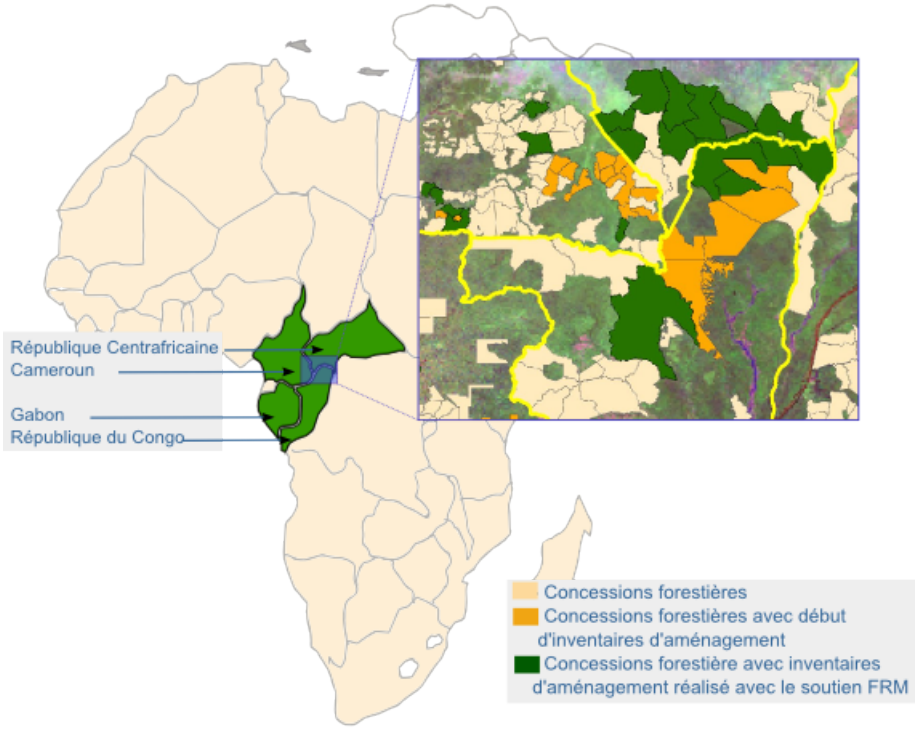
Corrélations carrées	f^1	f^2	f^3
ϕ_1	0.989873917	0.008499248	0.0039704801
ϕ_2	0.002561399	0.977191213	0.0004879874
ϕ_3	0.007026636	0.001481196	0.9783557479

Chapitre 5

Confrontation aux données réelles.

5.1 Description générale des données et justification de la prise en compte d'un effet aléatoire dans la modélisation.

Pour illustrer la méthode mise au point, nous nous sommes basés sur les données fournies par le CIRAD, qui ont entre autre nécessité l'aménagement et l'inventaire de 140.000 parcelles recouvrants quatre pays : la République Centrafricaine, le Congo, le Cameroun et la République Démocratique du Congo. Le but est alors de modéliser et prédire l'abondance des espèces d'arbre au sein de ces forêts tropicales.



Dans ce but, nous avons construit, à partir de la base de données "CoForChange/CoForTips", un sous-jeu de données contenant l'abondance de $q = 94$ genres communs aux forêts du Bassin du Congo et 56 covariables géo-référencées. Au total, $n = 2600$ parcelles de 5km par 5km constituent les observations. Chaque parcelle provient de l'agrégation d'un nombre variable de sous-parcelles de 0.5 ha.

Les variables en question, présentant une certaine redondance, et ayant donc vocation à alimenter notre matrice X, sont de natures très diverses.

i) D'une part, 29 variables décrivent, pour chaque parcelle, les caractéristiques physiques de l'environnement : topographie, pluviométrie et humidité des sols.

ii) D'autre part, la végétation est caractérisée grâce à 25 indices d'activité photosynthétique obtenus par télédétection (EVI (Enhanced Vegetation Index), NIR (Near Infrared channel Index), MIR (Mid-Infrared channel Index))

iii) Enfin, 2 indices décrivent la structure en hauteur des peuplements.

La matrice T, quant à elle, regroupe uniquement $r = 2$ variables faiblement corrélées. La première, sous forme de facteur, a pour but de décrire la géologie des parcelles considérées tandis que l'autre représente un indice de perturbation anthropique calculé comme une distance aux routes ou aux villes les plus proches.

Soulignons enfin que comme la surface échantillonnée n'est pas constante d'une parcelle à l'autre, elle doit être prise en compte dans la modélisation. Nous avons par conséquent affiné le modèle en utilisant la surface comme paramètre d'échelle.

Dans sa version initiale, la procédure SCGLR ne tenait pas compte de la concession dans laquelle les différentes parcelles avaient été référencées : toutes les parcelles en présence étaient considérées indépendantes. La nouvelle procédure a pour but de tenir compte de la dépendance qu'il peut exister au sein des parcelles, en particulier entre celles provenant d'une même concession.

La prise en compte d'un effet aléatoire "concession" présente en effet plusieurs intérêts notables. Le premier d'entre eux est de nature géographique. En effet, toutes les parcelles d'une même concession se situent dans la même région. Elles sont donc "voisines" d'une certaine manière, tout du moins géographiquement proches. Elles auront ainsi tendance à avoir des natures et des caractéristiques similaires. Le deuxième réside dans la façon de répertorier et classer les espèces, qui peut varier d'une concession à l'autre. Ce phénomène peut générer une dépendance encore plus importante entre les parcelles d'une même concession.

L'objectif principal est de montrer l'amélioration des résultats initiaux après la prise en compte de cette dépendance.

5.2 Présentation des résultats obtenus.

5.2.1 Procédure de Validation Croisée.

En premier lieu, avant même la mise en œuvre de notre algorithme sur les données, une question importante restait à résoudre : celle du nombre de composantes à sélectionner dans X pour modéliser et prédire la composition floristique des espèces présentes sur les parcelles. Nous aurions pu nous limiter à deux ou trois composantes, mais il est rare que la complexité des phénomènes naturels que nous étudions se laisse capturer par deux ou trois dimensions. Les résultats ne seraient alors que d'un intérêt limité.

Pour contrer cela, nous avons décidé de mettre en œuvre une stratégie de validation croisée à 10 groupes, en se permettant d'aller jusqu'à 10 composantes. Nous divisons l'ensemble des 2600 parcelles en deux sous-échantillons de manière aléatoire, l'un de taille 2000 et l'autre de taille 600. Sur le premier, nous effectuerons la procédure de validation croisée (échantillon de calibration) tandis que le deuxième sera l'échantillon de test (échantillon de validation). On note Y_{val} les variables réponses de l'échantillon de validation et $\hat{Y}_{val}$ les variables prédites correspondantes. De plus, $y^j_{val,i}$ désignera la $j^{ème}$ variable relative à la $i^{ème}$ parcelle de l'échantillon de validation et $\hat{y}^j_{val,i}$ la variable prédite correspondante. Le critère d'erreur que nous avons choisi pour évaluer les performances prédictives de l'algorithme est appelé "RMSE moyen" (RMSE pour Residual Mean Square Error). Il est en effet préférable de tenir compte de la variabilité des Y prédits pour évaluer l'erreur de prédiction. Dans notre contexte, le RMSE moyen est défini par :

$$\frac{1}{q}\sum_{j=1}^{q}\frac{1}{600}\sum_{i=1}^{600}\frac{\left(y^j_{val,i}-\hat{y}^j_{val,i}\right)^2}{\bar{\hat{y}}^j_{val}}$$

Ceci donne alors accès au graphe suivant :

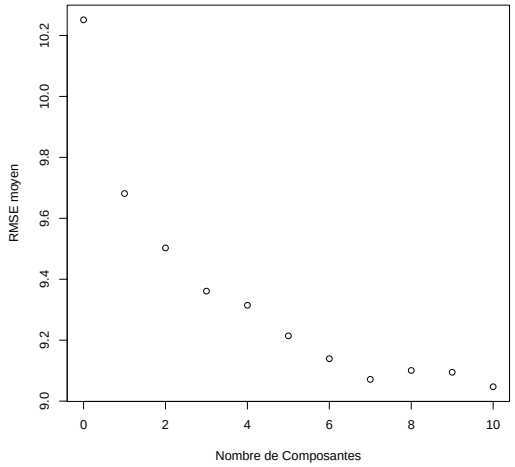


FIGURE 5.1: *Nous précisons que le modèle "nul", c'est-à-dire celui avec un nombre de composantes égal à zéro, correspond en réalité au modèle où seules les variables additionnelles dans T sont prises en compte, en incluant l'effet aléatoire. Au regard du graphe, notre choix s'est porté sur* ***7 composantes****.*

5.2.2 Présentation de l'amélioration apportée.

Notre but est maintenant de comparer les résultats obtenus grâce à notre méthode avec ceux de la version antérieure de SCGLR, et ce en ayant choisi 7 composantes. On calcule alors, pour tout $j \in \{1, \ldots, q\}$:

$$\frac{1}{600}\sum_{i=1}^{600}\rho^2\left(\hat{y}^j_{val,i}\,,\,y^j_{val,i}\right)$$

Nous présentons ci-après les histogrammes qui résument les résultats obtenus :

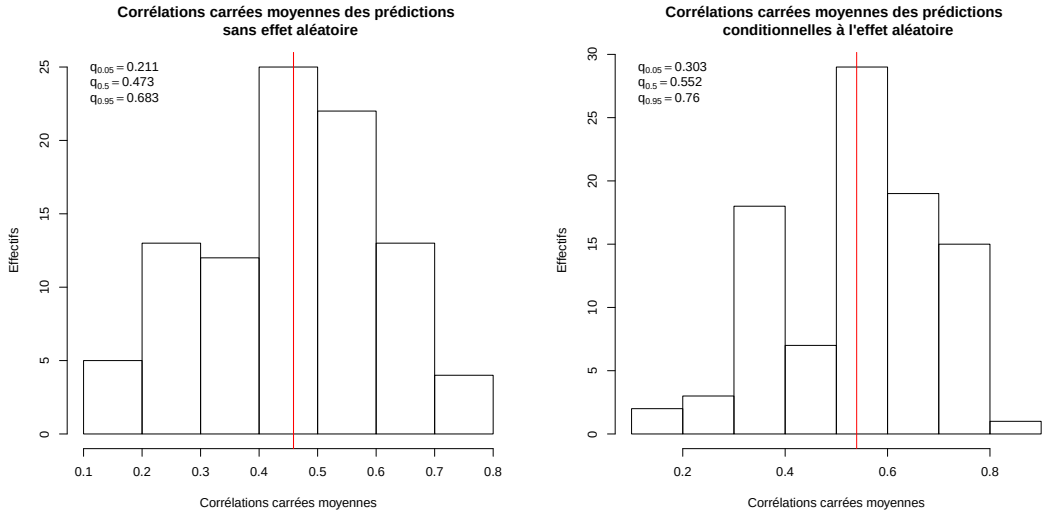


FIGURE 5.2: *Le premier histogramme résume les qualités prédictives de la méthode SCGLR initiale, i.e. avec la modélisation sans effet aléatoire. Le deuxième présente quant à lui les qualités prédictives conditionnelles à l'effet aléatoire obtenues avec l'extension de la méthode proposée. L'amélioration apportée est nette : en particulier, la médiane de la corrélation carrée moyenne est passée de* 0.47 *à* 0.55 *et les valeurs* $q_{0.05}$ *et* $q_{0.95}$ *ont également augmenté.*

Il est apparu également intéressant d'analyser les qualités prédictives marginales obtenues avec la nouvelle version de SCGLR, comparativement à la version initiale.

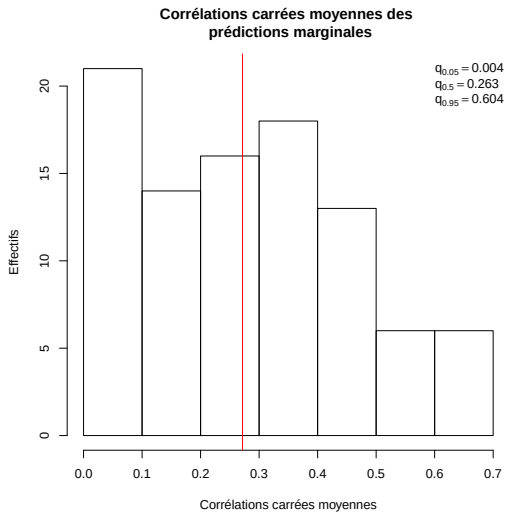


FIGURE 5.3: *Contrairement aux prédictions conditionnelles, les prédictions marginales sont dégradées comparativement à celles obtenues avec la première version de SCGLR.*

La dégradation observée induit deux conséquences.

i) D'un côté, cela traduit qu'il existe bien un effet "concession" non négligeable. En effet, si la concession n'avait aucun impact sur les données, les qualités prédictives des deux modélisations seraient comparables. Plus précisément, ceci montre qu'une partie de la variabilité captée par la version initiale de SCGLR n'était pas due uniquement aux effets fixes.

ii) D'un autre côté, avec la modélisation actuelle, il serait imprudent de prédire les communautés d'espèces sur la base d'observations provenant d'autres concessions... Des structures de dépendance plus fines sont donc à envisager.

5.2.3 Exemples de cartes construites sur la base des prédictions.

Cette section vise à présenter les prédictions spatialement explicites que l'on peut obtenir avec l'extension du modèle proposé. Pour cela, nous nous focaliserons sur deux genres caractéristiques : le Macaranga et le Celtis. Le premier est un genre pionnier et constitue donc un indicateur puissant des perturbations, tandis que le deuxième est un indicateur des forêts semi-décidues matures, c'est-à-dire avec une partie des arbres à feuilles caduques.

Toujours en utilisant 7 composantes dans X, nous présentons les prédictions marginales et les prédictions conditionnelles à la réalisation de l'effet aléatoire "concession". Les cartes obtenues symbolisent l'abondance prédite de l'espèce considérée sur chacune des parcelles des concessions auxquelles nous avons accès.

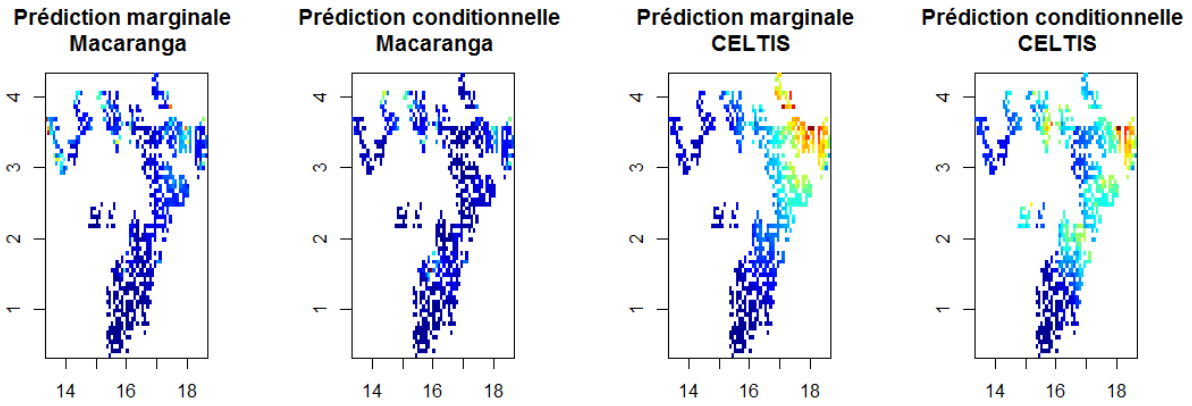


FIGURE 5.4: *Les cartes produites constituent un outil important afin de visualiser les lieux les plus propices à une dégradation de la forêt, qui apparaissent ici en bleu clair ou en rouge. En accord avec l'observation effectuée dans la section (5.2.2), il existe des différences entre les valeurs prédites marginalement et conditionnellement à l'effet aléatoire "concession". Nous privilégions pour l'instant la prédiction conditionnelle.*

5.2.4 Qualités explicatives de la modélisation.

Après avoir montré l'amélioration du modèle initial en terme de prédiction, nous soulignons que ses qualités explicatives sont préservées. En effet, tout comme dans la version initiale de SCGRL, nous pouvons produire les cercles de corrélation ayant vocation d'une part à interpréter les composantes retenues dans X et d'autre part à déterminer les espèces les plus corrélées à ces composantes. Par exemple, le cercle des corrélations sur le plan engendré par les deux premières composantes retenues dans X est le suivant :

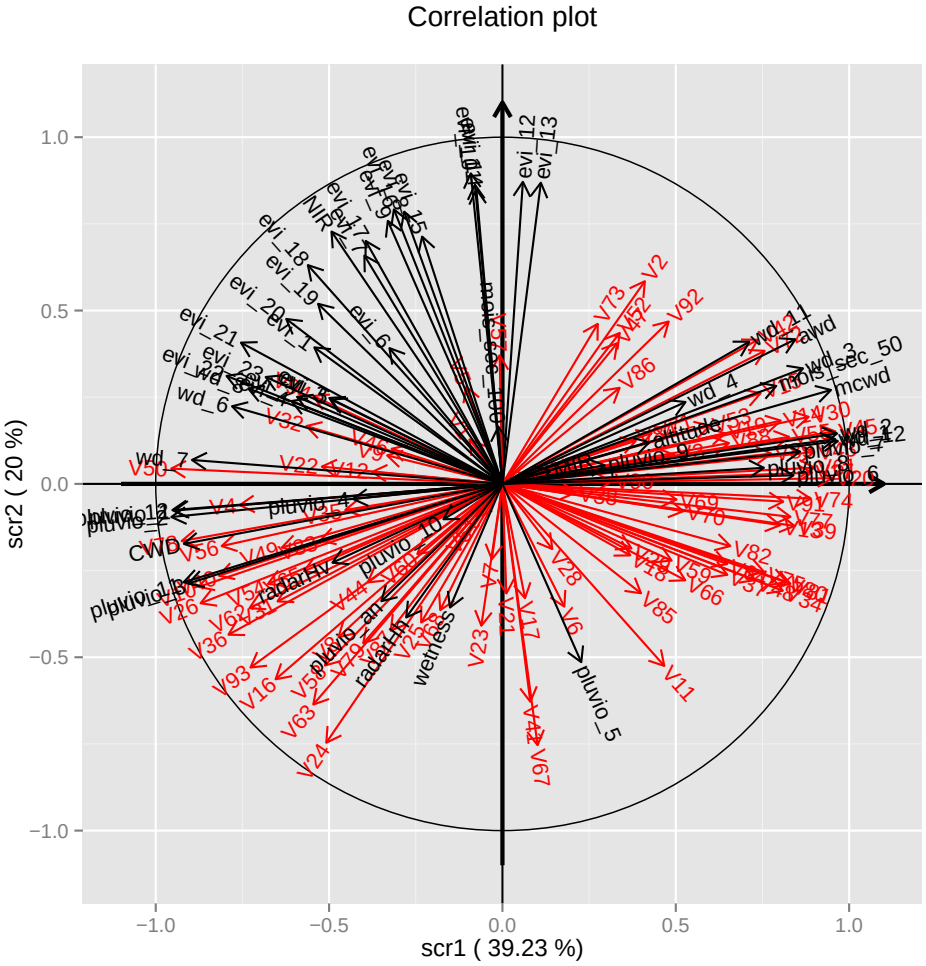


FIGURE 5.5: *Deux faisceaux apparaissent sur le premier plan. Le premier est porté est le faisceau "pluviométrie / intensité du vent" et est porté par les variables "pluvio" et "Wd". Le deuxième, plus large, peut être nommé "axe photosynthétique" et est porté par les variables EVI.*

Conclusion et perspectives.

Dans ce travail, nous avons choisi de combiner la stratégie décrite par SCGLR et une modélisation suivant des GL2M. Les résultats obtenus sur simulations semblent cohérents et une certaine amélioration en terme de prédiction a été observée sur les données réelles. En effet, le choix de modéliser la dépendance spatiale des unités statistiques via l'introduction d'un effet aléatoire "concession" a apporté une amélioration par rapport à la modélisation initiale, démontrée par l'augmentation significative des corrélations carrées des prédictions. Notons que l'injection d'un effet aléatoire n'a pas dégradé les qualités explicatives du modèle, étant donné que les outils interprétatifs que sont les cercles de corrélations restent accessibles.

Cependant, cette modélisation n'était qu'une étape intermédiaire avant d'envisager des modélisations plus complexes. En effet, il y a une chose dont nous n'avons pas tenu compte en particulier dans les données réelles : les phénomènes de compétitions entre espèces. On observe que certaines espèces sont toujours simultanément présentes alors que d'autres, au contraire, s'excluent mutuellement : la présence de l'une induit quasiment mécaniquement l'absence de l'autre. Nous pensons qu'il est possible de tenir compte de ce phénomène en adoptant une modélisation plus fine de l'effet aléatoire, notamment au travers des matrices A_j définies dans la section (2.1.2). La piste consistant à supposer des matrices A_j non nécessairement diagonales sera envisagée incessamment sous peu.

De plus, des modélisations plus fines de la dépendance spatiale pourront être prises en considération. En effet, bien que l'effet "concession" existe, il est sans doute possible d'adopter des modélisations continues de la dépendance spatiale, ce qui faciliterait sans doute la prédiction des communautés d'espèces sur de nouvelles parcelles.

Mais auparavant, avant d'élaboration de nouveaux modèles et de nouvelles extensions, une étape d'ordre technique doit être réalisée : il s'agit de l'intégration de l'extension de SCGLR proposée à la version existante du package. Bien que les codes R développés fonctionnent, un travail d'adaptation doit être effectué pour les généraliser au maximum afin que ces derniers puissent réagir sur toutes les données.

Chapitre 6

Annexes.

6.1 Exemples de mesures de pertinence structurelle.

Le but de cette annexe est de présenter comment adapter la formulation donnée par (1.10) suivant la façon dont on veut contraindre u.

6.1.1 Inertie des unités le long de la direction $\langle u \rangle$.

Par exemple, si l'on veut que la direction donnée par $\langle u \rangle$ soit celle qui capte le maximum d'inertie des unités, alors on peut définir la quantité $\phi(u)$ ainsi :

$$\phi(u) = \|Xu\|_W^2 = u'X'WXu.$$

On voit alors que cette définition de ϕ correspond à $\phi_{N,\Omega,l}$, en ayant choisi la métrique $M = Id$ et en ayant posé :

i) $l = 1$, $J = 1$

ii) $N_1 = X'WX$

iii) $\omega_1 = 1$

Si l'on veut trouver la direction $\langle Xu \rangle$ qui maximise cette quantité, elle est donnée par le premier vecteur propre de l'ACP de (X, Id, W) en se plaçant dans l'espace direct.

6.1.2 Inertie des variables le long de la composante $\langle f \rangle = \langle Xu \rangle$.

Un autre exemple classique consiste à chercher la composante $f = Xu$ qui capte le maximum d'inertie d'un groupe de variables $x^1, x^2, \ldots, x^p$. En imposant, pour identification, une contrainte sur la norme de f comme par exemple $\|f\|_W^2 = 1$, une façon de faire est de définir ϕ de la façon suivante :

$$\begin{aligned}\phi(u) &= \sum_{j=1}^{p}\rho^2(f,x^j) = \sum_{j=1}^{p}\langle f|x^j\rangle_W^2 \\ &= \sum_{j=1}^{p} f'Wx^j x^{j'}Wf \\ &= u'X'W\left(\sum_{j=1}^{p} x^j x^{j'}\right)WXu \\ &= u'\left(X'WXX'WX\right)u = u'(X'WX)^2u\end{aligned}$$

Là par contre, la définition de ϕ adoptée correspond à $\phi_{N,\Omega,l}$ avec la métrique $M = (X'WX)^{-1}$ de telle sorte que :

$$\|f\|_W^2 = 1 = \|u\|_{M^{-1}}^2,$$

et en ayant posé :

i) $l = 1$, $J = 1$

ii) $N_1 = (X'WX)^2$

iii) $\omega_1 = 1$

Parallèlement au cas précédent, la composante $f = Xu$ maximisant $\phi(u)$ est donnée par le premier vecteur propre (dual) de l'ACP du triplet (X, Id, W).

6.1.3 Extension du cas précédent.

On se place dans le même cadre que précédemment, c'est à dire avec des variables numériques $x^1, \ldots, x^p$ rassemblées dans un tableau X. En considérant la composante $f = Xu$, on pose toujours $M = (X'WX)^{-1}$ de telle sorte à respecter l'égalité :

$$\|f\|_W^2 = 1 = \|u\|_{M^{-1}}^2,$$

Mais cette fois-ci, nous définissons $\phi(u)$ de la manière suivante :

$$\begin{aligned}\phi(u) &= \left(\sum_{j=1}^{p}\omega_j\rho^{2l}(f,x^j)\right)^{\frac{1}{l}} = \left(\sum_{j=1}^{p}\omega_j\langle Xu|x^j\rangle_W^{2l}\right)^{\frac{1}{l}} \\ &= \left(\sum_{j=1}^{p}\omega_j(u'X'Wx^j x^{j'}WXu)^l\right)^{\frac{1}{l}}\end{aligned}$$

Notons bien qu'en posant $l = 1$ et $\omega_j = 1$ $\forall j \in \{1, \ldots, j\}$, nous retombons sur le cas précédemment décrit. Cette extension permet en fait de régler le "degré de localité" des faisceaux considérés. En effet, on montre que plus la valeur de l est élevée, plus les faisceaux considérés sont locaux, dans le sens où ils seront plus proches des variables originales.

6.1.4 Mesure de la proximité de $\langle u \rangle$ par rapport à différents sous-espaces.

Enfin, un autre exemple que l'on peut donner est le cas où l'on souhaite déterminer la direction $\langle u \rangle$ qui soit conjointement la plus proche possible d'une famille de sous-espaces $(S_1, \ldots, S_J)$. En prenant $M^{-1} = Id$ et en imposant $\|u\|^2_{M^{-1}} = 1$, une façon de définir cette proximité est de poser :

$$\phi(u) = \sum_{j=1}^{J} \omega_j \cos^2(u, S_j)$$

Or, rappelons que nous pouvons tout à fait transformer l'écriture de $\cos^2(u, S_j)$ de la façon suivante :

$$\begin{aligned}
\cos^2(u, S_j) &= \cos^2\left(u, \prod_{\langle S_j \rangle} u\right) \\
&= \frac{\left\| \prod_{\langle S_j \rangle} u \right\|^2}{\|u\|^2} = \left\| \prod_{\langle S_j \rangle} u \right\|^2 \quad \text{car } \|u\|^2 = 1 \\
&= \langle \prod_{\langle S_j \rangle} u | \prod_{\langle S_j \rangle} u \rangle \\
&= \langle u | \prod_{\langle S_j \rangle} u \rangle \quad \text{car tout projecteur est auto-adjoint et idempotent.} \\
&= u' \prod_{\langle S_j \rangle} u
\end{aligned}$$

Au final, notre définition de ϕ peut se réécrire ainsi :

$$\phi(u) = \sum_{j=1}^{J} \omega_j u' \prod_{\langle S_j \rangle} u$$

ce qui correspond à $\phi_{N,\Omega,l}$, avec la métrique $M = Id$ et en ayant posé :

i) $l = 1$

ii) $N = \left\{ \prod_{\langle S_1 \rangle}, \ldots, \prod_{\langle S_J \rangle} \right\}$

iii) Et avec la condition $\sum_{j=1}^{J} \omega_j = 1$

6.2 Algorithme PING.

6.2.1 Itération générique.

La maximisation sous contrainte que l'algorithme PING a vocation à résoudre s'écrit sous la forme du programme :

$$P : \begin{cases} \textbf{Maximiser} \quad h(u) \\ \textbf{sous les contraintes :} \quad u'M^{-1}u = 1 \text{ et } D'u = 0 \end{cases}$$

On effectue le changement de variable $v = M^{-1/2}u \Leftrightarrow u = M^{1/2}v$. Ce changement de variable implique :

i) $h(u) = h(M^{1/2}v) = g(v)$ en notation.

ii) $u'M^{-1}u = \left(M^{1/2}v\right)' M^{-1}M^{1/2}v = v'M^{1/2}M^{-1}M^{1/2}v = v'v$.
Donc nous avons l'équivalence $u'M^{-1}u = 1 \Leftrightarrow v'v = 1$.

iii) Enfin, $D'u = D'M^{1/2}v = \left(M^{1/2}D\right)' v = 0$
D'où l'équivalence $D'u = 0 \Leftrightarrow C'v = 0$ avec $C = M^{1/2}D$

Le programme défini par (P) est alors strictement équivalent au programme :

$$P' : \begin{cases} \textbf{Maximiser} \quad g(v) \\ \textbf{sous les contraintes :} \quad v'v = 1 \text{ et } C'v = 0 \end{cases}$$

La résolution de (P') nécessite la définition du Lagrangien suivant, et de la recherche de ses points critiques :

$$\mathcal{L}(v, \lambda, \mu) = g(v) - \lambda(v'v - 1) - \mu'C'v$$

La recherche des points critiques nous pousse à définir les équations :

i) $\underset{\lambda,\mu}{\nabla}\mathcal{L}(v, \lambda, \mu) = 0$, ce qui induit immédiatement les égalités suivantes :

$$v'v = 1 \tag{6.1}$$

$$C'v = 0 \tag{6.2}$$

ii) $\underset{v}{\nabla}\mathcal{L}(v, \lambda, \mu) = 0 \Leftrightarrow \underset{v}{\nabla}g(v) - 2\lambda v - C\mu = 0$. Puis en posant $\Gamma(v) = \underset{v}{\nabla}g(v)$, on obtient la relation :

$$\Gamma(v) - 2\lambda v - C\mu = 0 \Leftrightarrow v = \frac{\Gamma(v) - C\mu}{2\lambda} \tag{6.3}$$

En pré-multipliant la relation (6.3) par C', on a les équivalences :

$$\begin{aligned} 2\lambda C'v = C'\Gamma(v) - CC'\mu \Leftrightarrow C'\Gamma(v) - CC'\mu = 0 \quad \text{car } C'v = 0 \\ \Leftrightarrow C'\Gamma(v) = CC'\mu \\ \Leftrightarrow \mu = (C'C)^{-1}C'\Gamma(v) \end{aligned} \tag{6.4}$$

En injectant l'expression (6.4) dans (6.3), on obtient une expression pour v. En effet,

$$\begin{aligned} v &= \frac{1}{2\lambda}\left[\Gamma(v) - C(C'C)^{-1}C'\Gamma(v)\right] \\ &= \frac{1}{2\lambda}\left[Id - C(C'C)^{-1}C'\right]\Gamma(v) \\ &= \frac{1}{2\lambda}\prod_{\langle C\rangle^{\perp}}\Gamma(v) \quad \text{avec } \prod_{\langle C\rangle^{\perp}} = Id - C(C'C)^{-1}C' \end{aligned} \tag{6.5}$$

Enfin, comme on veut $\|v\|^2 = 1$, on obtient finalement :

$$v = \frac{\frac{1}{2\lambda}\prod_{\langle C\rangle^{\perp}}\Gamma(v)}{\left\|\frac{1}{2\lambda}\prod_{\langle C\rangle^{\perp}}\Gamma(v)\right\|} = \frac{\prod_{\langle C\rangle^{\perp}}\Gamma(v)}{\left\|\prod_{\langle C\rangle^{\perp}}\Gamma(v)\right\|}$$

On pourra alors poser, au cours des itérations de l'algorithme PING, :

$$v^{[t+1]} = \frac{\prod_{\langle C\rangle^{\perp}}\Gamma(v^{[t]})}{\left\|\prod_{\langle C\rangle^{\perp}}\Gamma(v^{[t]})\right\|}$$

6.2.2 Propriétés remarquables de l'algorithme PING.

Propriété de "shrinkage".

Rappelons que la détermination de la première composante est basé sur la maximisation suivante :

$$\max_{u'M^{-1}u=1} S_T(u)$$

En posant $u^* = \underset{u'M^{-1}u=1}{argmax}\ S_T(u)$ et $S^* = S(u^*)$, on montre, en utilisant un argument de dualité, que la maximisation précédemment écrite est équivalente à :

$$\min_{S(u)=S^*} u'M^{-1}u$$

Nous pouvons alors voir chaque étape de l'algorithme SCGLR comme la minimisation de $\|u\|_{M^{-1}}$ pour une valeur donnée de $S(u)$. Cette minimisation se rapproche donc des procédures de régression ridge ou lasso qui possède de très bonnes propriétés en termes de variance des estimateurs. Notre procédure tente dons intrinsèquement de minimiser la confusion entre les différentes composantes dans le prédicteur linéaire.

L'algorithme suit une direction de montée.

Nous nous proposons de montrer ici que si $v^{[t+1]}$ est suffisamment proche de $v^{[t]}$, alors on est certain que la fonction g augmente. La stratégie consiste à exhiber une direction w telle que $\langle w|\underset{v}{\nabla} g(v)\rangle$.

Remarquons dans un premier temps que par construction, nous avons l'égalité suivante :

$$v^{[t]} = \prod_{C^\perp} \Gamma(v^{[t]}), \quad \forall t,$$

aussi pouvons nous écrire les égalités suivantes :

$$\begin{aligned}
\langle v^{[t+1]} - v^{[t]} | \Gamma(v^{[t]}) \rangle &= \langle \prod_{C^\perp} \left(v^{[t+1]} - v^{[t]} \right) | \Gamma(v^{[t]}) \rangle \\
&= \langle v^{[t+1]} - v^{[t]} | \prod_{C^\perp}^{*} \Gamma(v^{[t]}) \rangle \\
&= \langle v^{[t+1]} - v^{[t]} | \prod_{C^\perp} \Gamma(v^{[t]}) \rangle \text{ car tout projecteur est auto-adjoint.}
\end{aligned}$$

Or, nous rappelons que par définition de la quantité $v^{[t+1]}$, nous avons :

$$v^{[t+1]} = \frac{\prod_{C^\perp} \Gamma(v^{[t]})}{\left\| \prod_{C^\perp} \Gamma(v^{[t]}) \right\|}$$

Ainsi, nous pouvons remplacer, dans l'expression précédente, la quantité $\prod_{C^\perp} \Gamma(v^{[t]}) \rangle$ par $\left\| \prod_{C^\perp} \Gamma(v^{[t]}) \right\| v^{[t+1]}$, d'où :

$$\begin{aligned}
\langle v^{[t+1]} - v^{[t]} | \Gamma(v^{[t]}) \rangle &= \left\| \prod_{C^\perp} \Gamma(v^{[t]}) \right\| \langle v^{[t+1]} - v^{[t]} | v^{[t+1]} \rangle \\
&= \left\| \prod_{C^\perp} \Gamma(v^{[t]}) \right\| \left(\langle v^{[t+1]} | v^{[t+1]} \rangle - \langle v^{[t]} | v^{[t+1]} \rangle \right) \\
&= \left\| \prod_{C^\perp} \Gamma(v^{[t]}) \right\| \left(1 - \langle v^{[t]} | v^{[t+1]} \rangle \right) \\
&= \left\| \prod_{C^\perp} \Gamma(v^{[t]}) \right\| \left(1 - \cos(v^{[t]}, v^{[t+1]}) \right) \\
&\geqslant 0.
\end{aligned}$$

En posant $\gamma^{[t]} = v^{[t+1]}$, la direction donnée par la corde $\left[v^{[t]}, \gamma^{[t]} \right]$ est bien une direction de montée. Cependant, rien ne nous garantit encore que la fonction g augmente à chaque itération... Notons néanmoins que pour un $\gamma^{[t]}$ suffisamment proche de $v^{[t]}$, la direction donnée par la corde $\left[v^{[t]}, \gamma^{[t]} \right]$ est approximativement la même que celle donnée par le vecteur w, désignant un vecteur tangent à l'arc $\left(v^{[t]}, \gamma^{[t]} \right)$ partant de $v^{[t]}$.

En désignant par ϖ la plan tangent à la sphère en $v^{[t]}$, et pour un $\gamma^{[t]}$ suffisamment proche de $v^{[t]}$, il existe $\tau \in \mathbb{R}_+^*$ tel que : $w = \tau \prod_{\varpi} \gamma^{[t]}$. Ainsi, on observe que :

$$\langle w | \gamma^{[t]} \rangle = \tau \langle \prod_{\varpi} \gamma^{[t]} | \gamma^{[t]} \rangle = \tau \cos(\gamma^{[t]}, \varpi) > 0.$$

Par conséquent, on peut assurer que si l'on reste assez proche du point courant $v^{[t]}$ sur l'arc $\left(v^{[t]}, \gamma^{[t]} \right)$, alors la fonction g va croître. Cependant, en restant trop proche du point courant, l'algorithme peut s'avérer assez lent... Nous proposons alors d'adopter la procédure suivante :

Algorithm 5 Itération générique de l'algorithme PING

Initialiser le vecteur $v^{[0]}$
repeat
  Poser $\gamma^{[t]} = \dfrac{\prod_{C^\perp} \Gamma(v^{[t]})}{\left\| \prod_{C^\perp} \Gamma(v^{[t]}) \right\|}$
  $m \leftarrow \gamma^{[t]}$
  while $g(m) < g(v^{[t]})$ **do**
    $m \leftarrow \dfrac{v^{[t]} + m}{\|v^{[t]} + m\|}$
  end while
  Poser alors $v^{[t+1]} = m$
until Convergence

Bibliographie

[1] Adeline Fayolle, Nicolas Picard, Jean-Louis Doucet, Michael Swaine, Nicolas Bayol, Fabrice Bénédet, and Sylvie Gourlet-Fleury. A new insight in the structure, composition and functioning of central African moist forests. *Forest Ecology and Management*, 329 :195–205, October 2014.

[2] Valéry Gond, Adeline Fayolle, Alexandre Pennec, Guillaume Cornu, Philippe Mayaux, Pierre Camberlin, Charles Doumenge, Nicolas Fauvet, and Sylvie Gourlet-Fleury. Vegetation structure and greenness in Central Africa from Modis multi-temporal data. *Philosophical Transactions of the Royal Society of London B : Biological Sciences*, 368(1625) :20120309, September 2013.

[3] Xavier Bry, Catherine Trottier, and Thomas Verron. Component-based Generalized Linear Regression using a PLS-extended variant of the Fisher scoring algorithm. September 2011.

[4] X. Bry, C. Trottier, T. Verron, and F. Mortier. Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, 119(C) :47–60, 2013.

[5] P. McCullagh and John A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC, Boca Raton, 2 edition edition, August 1989.

[6] James K. Lindsey. *Applying Generalized Linear Models*. Springer Science & Business Media, June 1997.

[7] Julian J. Faraway. *Extending the Linear Model with R : Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC Press, December 2005.

[8] Shayle R. Searle, George Casella, and Charles E. McCulloch. *Variance Components*. John Wiley & Sons, September 2009.

[9] Catherine Trottier. *Estimation dans les modèles linéaires généralisés à effets aléatoires*. phdthesis, Institut National Polytechnique de Grenoble - INPG, July 1998.

[10] R. I. Jennrich and P. F. Sampson. Newton-Raphson and Related Algorithms for Maximum Likelihood Variance Component Estimation. *Technometrics*, 18(1) :11–17, February 1976.

[11] B. D. Marx. Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*, 38(4) :374–381, 1996.

[12] Guillaume Cornu, Frederic Mortier, Catherine Trottier, and Xavier Bry. SCGLR : Supervised Component Generalized Linear Regression, February 2015.

[13] C. R. Henderson, Oscar Kempthorne, S. R. Searle, and C. M. von Krosigk. The Estimation of Environmental and Genetic Trends from Records Subject to Culling. *Biometrics*, 15(2) :192–218, June 1959.

[14] Christian Lavergne and Catherine Trottier. Sur l'estimation dans les modèles linéaires généralisés à effets aléatoires. report, 1999.

[15] Marie-José Martinez. *Modèles linéaires généralisés à effets aléatoires : contributions au choix de modèle et au modèle de mélange*. phdthesis, Université Montpellier II - Sciences et Techniques du Languedoc, September 2006.

[16] B. Engel and A. Keen. A simple approach for the analysis of generalizea linear mixed models. *Statistica Neerlandica*, 48(1) :1–22, March 1994.

[17] Robert Schall. Estimation in generalized linear models with random effects. *Biometrika*, 78(4) :719–727, December 1991.