Introducing complex dependency structures into supervised component-based models

Jocelyn Chauvet

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Biostatistique

École doctorale I2S – Information, Structures, Systèmes

Unité de recherche UMR 5149 – IMAG – Institut Montpelliérain Alexander Grothendieck

Structures de dépendance complexes pour modèles à composantes supervisées

Présentée par Jocelyn CHAUVET

Le 19 avril 2019

Sous la direction de Catherine TROTTIER et Xavier BRY

Devant le jury composé de

Christophe BIERNACKI	Professeur	Université Lille 1	Rapporteur
Xavier BRY	Maître de Conférence	Université de Montpellier	Co–encadrant de thèse
Isabelle CARRIÈRE	Ingénieure de recherche	INSERM	Examinatrice
Brian MARX	Professeur	Louisiana State University	Rapporteur
Jerôme SARACCO	Professeur	Institut Polytechnique de Bordeaux	Président du jury
Catherine TROTTIER	Maîtresse de Conférence	Université Montpellier 3	Directrice de thèse

UNIVERSITÉ DE MONTPELLIER

À la mémoire de mon père.

Remerciements

C'est avec une émotion toute particulière que je m'apprête à écrire les dernières lignes de ce manuscrit. Je voudrais commencer avec un petit proverbe, découvert dans mes années collège — probablement dans les *Sublimes paroles et idioties de Nasr Eddin Hodja* (pardon si je me trompe) —, et que j'ai toujours essayé de garder à l'esprit. Il dit ceci :

Celui qui sait et sait qu'il sait, c'est un savant. Il faut l'écouter.
Celui qui sait et ne sait pas qu'il sait, c'est un sage. Il faut le suivre.
Celui qui ne sait pas et sait qu'il ne sait pas, c'est un chercheur. Il faut l'aider.
Celui qui ne sait pas et ne sait pas qu'il ne sait pas, c'est un idiot. Il faut le fuir.

Quand on est (apprenti) chercheur, on se place naturellement dans la troisième catégorie. Je voudrais donc saisir l'occasion qui m'est donnée pour remercier toutes les personnes qui m'ont aidé d'une façon ou d'une autre, et qui m'ont fait grandir.

Catherine, Xavier, les quelques lignes qui vont suivre seront bien insuffisantes au regard de tout ce que vous m'avez apporté. Il y a quatre ans, vous m'avez accordé votre confiance. Cela signifie déjà beaucoup pour moi et j'espère en avoir été digne. Puis durant toute la durée de la thèse, votre investissement a été sans faille. Entre rigueur et bienveillance, vous avez grandement contribué à mon épanouissement scientifique, et bien plus encore. Vous avez su être à mon écoute, et gérer mes doutes qui ont pu être nombreux, particulièrement dans les périodes charnières de rédaction. Je n'oublie pas que sans vos conseils avisés et relectures attentives, je n'aurais pas pu en arriver là. En bref, et je n'ai pas peur de le dire, votre encadrement a été extraordinaire. Tout au long de cette formidable aventure intellectuelle que vous m'avez donné l'occasion de vivre, vous avez tout simplement été des maîtres et des partenaires exceptionnels. La bonne humeur, les rires et le second degré qui ont ponctué l'intégralité de nos rencontres ont été pour moi des vecteurs essentiels de bien–être, sans lesquels aucune sérénité n'est possible. Pour toutes ces raisons et celles que j'oublie, MERCI !

Je tiens à remercier très chaleureusement Brian Marx [1] et Christophe Biernacki pour avoir accepté de rapporter ce manuscrit, malgré des emplois du temps que j'imagine chargés. Vos rapports m'ont fait très chaud au cœur, et les perspectives de recherche que vous suggérez sont pour moi d'un intérêt primordial. Je remercie également Jérôme Saracco et Isabelle Carrière d'avoir accepté de faire partie du jury de cette thèse. Mais plus généralement, à tous les membres du jury, sachez que votre présence à ma soutenance est déjà un grand honneur pour moi.

Un grand merci à Jean–Noël et Jean–Michel pour avoir participé activement à mes comités de suivi individuels. Vous avez toujours été soucieux du bon déroulement de la thèse, vous m'avez toujours encouragé et fourni de précieux conseils, que ce soit dans le cadre plutôt formel de ces comités, ou lors de rencontres plus informelles. Je n'oublie pas non plus vos grains de folie respectifs, qui m'ont conforté dans l'idée que dans ce monde ahurissant à bien des égards, il est de bon ton de ne pas toujours se prendre au sérieux ! Je voudrais également avoir une pensée particulière pour Bernadette, Éric, Sophie, Carmela, Nathalie et Laurence : je crois bien que c'est en partie grâce à vous que je n'ai pas encore de phobie administrative ! Mais de vous et de Gemma, je retiendrai surtout nos discussions plus ou moins sérieuses et nos délires, qui ont accompagné chacune de mes journées passées au laboratoire, et qui m'ont donné beaucoup de bonheur. Vous êtes inoubliables, et ces quatre années n'auraient pas eu la même saveur sans vous. Vous allez indiscutablement beaucoup me manquer. Je m'en voudrais si j'oubliais de remercier Baptiste et François–David, qui m'ont permis de me familiariser avec les clusters de calcul, et qui de fait ont rendu possible (en un temps de calcul raisonnable) les nombreuses simulations qui émaillent ce manuscrit !

Comme l'enseignement a été pour moi d'une importance capitale pendant ces quatre années, je tenais à remercier tous ceux qui m'ont fait confiance pour assurer cette noble mission. Je pense notamment à Lionel Cucala, Ludovic Menneteau, Élodie Brunel, Mathieu Ribatet et Jorge Ramirez. Vous m'avez permis de prendre beaucoup de plaisir à enseigner à l'université. Un grand merci aussi à Nicolas Saby pour m'avoir proposé la mission diffusion qui m'a ravi pendant deux ans. J'en profite pour te dire que les journées Condorcet que tu as organisées m'ont véritablement conquis. Elles ont définitivement scellé le grand intérêt que j'accorde à la problématique du choix collectif. Enfin, merci à Nicolas Molinari, Véronique Ladret et Kévin Mouzat pour m'avoir fait confiance pour les enseignement relatifs à mon poste d'ATER.

1. I would like to thank Brian Marx very much for his report on my dissertation, which really warmed my heart. The research perspectives you suggest are of primary interest to me. It is an honour to have had my work reviewed by a researcher of your stature.

L'ambiance conviviale et chaleureuse que j'ai ressentie au sein du laboratoire a été essentielle pour moi. Je remercie donc tous les permanents et (post–)doctorants de l'IMAG — qu'ils soient de l'équipe EPS, GTA, ACSIOM ou DEMA — qui ont contribué de près ou de loin à ma gaîté quotidienne ! Vous citer ici de manière exhaustive serait une entreprise périlleuse et je ne m'y risquerai pas. Un grand merci néanmoins pour toutes ces discussions et ces débats, ces pauses repas et pauses café, et tous ces petits moment d'égarement uniques qui m'ont permis une certaine évasion ! Je voudrais quand même avoir une pensée pour mes co–bureaux, incontournables partenaires, qui m'ont accompagné et soutenu de très près durant ces années. Tout d'abord, à Myriam et à Antoine, un grand merci pour avoir facilité mon intégration au sein du laboratoire. Merci à toi Christian, pour nos discussions passionnantes et extraordinairement intelligentes, souvent autour de questions touchant à la démocratie ! Merci à Mario et à Benjamin pour votre bonne humeur de tous les instants. Bien que studieux, vous avez été des fournisseurs incontournables de rires ! Je voudrais enfin avoir une pensée spéciale pour Louis. Même dans les moments où mon humeur était la plus massacrante (vive la rédaction !), tu as toujours été là et tu m'as beaucoup soutenu. Ta présence rassurante et ton calme légendaire ont permis de canaliser mes angoisses, qui ont pu être parfois disproportionnées. Ce manuscrit de thèse ne serait pas ce qu'il est sans toi, tant tu m'as permis de trouver l'abnégation nécessaire à sa rédaction.

Clara, Pauline, Laëti, Morgane, merci pour ce nouvel an mémorable à Barcelone ! Vous avez su me changer les idées au bon moment et je ne vous en serai jamais assez reconnaissant. J'ai été aux abonnés absents ces derniers temps, mais je vais me rattraper, c'est promis !

Enfin, c'est à ma famille que j'adresse ces derniers mots. Maman, Sylvain, vous avez toujours été là pour moi, en toute circonstance et y compris dans mes moments de doute les plus abyssaux. Pour votre amour et votre indéfectible soutien, je tenais à vous dire merci du fond du cœur. J'espère vous apporter autant que ce que vous m'apportez. Gilles, je ne t'oublie pas, cela fait maintenant de nombreuses années que tu es là pour nous, dans les meilleurs moments comme dans les pires. À toi aussi, tout simplement et sincèrement merci !

Papa, il n'y a pas eu un seul jour depuis 18 ans sans que j'aie pensé à toi. Certaines personnes croient en Dieu, d'autres aux forces de l'esprit. Moi, je ne sais pas trop en quoi je crois mais je sais que d'une manière ou d'une autre, tu m'as accompagné toutes ces années. J'espère qu'aujourd'hui tu es fier de moi.

Contents

List of Symbols and Acronyms

Numbers, Matrices and Indexing

a	A scalar (integer or real)
$\boldsymbol{a}$	A vector
a_i	i^{th} element of vector $\boldsymbol{a}$
$\mathbf{1}_k$	All–ones vector of size k
$\mathbf{1}$	All–ones vector, with its size implied by the context
$\boldsymbol{A}$	A matrix
$\boldsymbol{a_j}$	j^{th} column–vector of matrix $\boldsymbol{A}$
$\boldsymbol{a}_{\boldsymbol{i:}}^{\mathrm{T}}$	i^{th} row–vector of matrix $\boldsymbol{A}$
$\boldsymbol{I}_n$	Identity matrix with n rows and n columns
$\boldsymbol{I}$	Identity matrix, with its size implied by the context
$\mathbf{Diag}\left(a_i\right)_{i=1\ldots,n}$	A square, diagonal matrix of size $n \times n$ with i^{th} diagonal entry given by a_i
$\mathbf{bDiag}\left(\boldsymbol{B}_j\right)_{j=1\ldots,p}$	A block diagonal matrix with blocks $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_p$

Linear Algebra

$\boldsymbol{A}^{\mathrm{T}}$	Transpose of matrix $\boldsymbol{A}$
$\boldsymbol{A} \otimes \boldsymbol{B}$	Kronecker product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$
$\lvert\boldsymbol{A}\rvert$	Determinant of matrix $\boldsymbol{A}$
$\mathrm{Trace}\left(\boldsymbol{A}\right)$	Trace of matrix $\boldsymbol{A}$
$\mathrm{span}\left\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_h\right\}$	The space spanned by vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_h$
$\mathrm{span}\left\{\boldsymbol{A}\right\}$	The space spanned by the column–vectors of $\boldsymbol{A}$

Euclidean Geometry

$\langle \boldsymbol{a} \mid \boldsymbol{b} \rangle_{\boldsymbol{M}}$	Scalar product of $\boldsymbol{a}$ and $\boldsymbol{b}$ with respect to $\boldsymbol{M}$, $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{b}$
$\|\boldsymbol{a}\|_{\boldsymbol{M}}$	Euclidean norm of $\boldsymbol{a}$ with respect to metric $\boldsymbol{M}$, $\sqrt{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{a}}$
$\cos_{\boldsymbol{M}}(\boldsymbol{a}, \boldsymbol{b})$	Cosine of the angle between $\boldsymbol{a}$ and $\boldsymbol{b}$ with respect to $\boldsymbol{M}$, $\frac{\langle \boldsymbol{a} \mid \boldsymbol{b} \rangle_{\boldsymbol{M}}}{\|\boldsymbol{a}\|_{\boldsymbol{M}} \|\boldsymbol{b}\|_{\boldsymbol{M}}}$
$\Pi^{\boldsymbol{W}}_{\mathrm{span}\{\boldsymbol{X}\}}$	The $\boldsymbol{W}$–orthogonal projector onto $\mathrm{span}\{\boldsymbol{X}\}$

Calculus

$\frac{\mathrm{d}y}{\mathrm{d}x}$	Derivative of y with respect to x
$\frac{\partial y}{\partial x}$	Partial derivative of y with respect to x
$\int_{\mathbb{S}} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$	Integral with respect to $\boldsymbol{x}$ over the set $\mathbb{S}$
$\boldsymbol{\nabla} f(\boldsymbol{x})$	Gradient vector of f at input point $\boldsymbol{x}$
$\underset{\boldsymbol{x}}{\boldsymbol{\nabla}} f(\boldsymbol{x}, \boldsymbol{y})$	Gradient vector of f with respect to $\boldsymbol{x}$ at input point $(\boldsymbol{x}, \boldsymbol{y})$
$\mathbf{H} f(\boldsymbol{x})$	Hessian matrix of f at input point $\boldsymbol{x}$

Statistical Modelling and Likelihoods

Y	A random vector
$\boldsymbol{y}$	The observed response vector associated with Y
$\boldsymbol{\beta}$	Fixed–effect parameter vector
$\boldsymbol{X}$	Fixed–effect design matrix
$\boldsymbol{\xi}$	Random–effect vector
$\boldsymbol{U}$	Random–effect design matrix
$\boldsymbol{\varepsilon}$	Noise vector
$\mathcal{L}$, $\mathcal{L}_{\mathrm{pen}}$	Likelihood, Penalised likelihood
ℓ, ℓ^{c}, ℓ_{pen}, $\ell^{\mathrm{c}}_{\mathrm{pen}}$	Log–likelihood, Complete log–likelihood, and their penalised versions

Probability and Information Theory

$X \sim F$	Random variable X has distribution F
$\mathbb{E}_f(\phi(\boldsymbol{x}))$ or $\mathbb{E}(\phi(\boldsymbol{x}))$	Expectation of $\phi(\boldsymbol{x})$ under f
$\mathbb{V}_f(\phi(\boldsymbol{x}))$, $\mathbb{V}(\phi(\boldsymbol{x}))$ or $\mathrm{Var}(\phi(\boldsymbol{x}))$	Variance of $\phi(\boldsymbol{x})$ under f
cov, cor	Covariance and correlation operators
$H(q)$	Shannon entropy of distribution q
$D_{KL}(p\|q)$	Kullback–Leibler divergence from p to q

Distributions

$\mathcal{N}_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	N–dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$\mathcal{P}(\boldsymbol{\lambda})$	Multivariate Poisson distribution with parameter $\boldsymbol{\lambda}$
$\mathcal{B}(\boldsymbol{p})$	Multivariate Bernoulli distribution with parameter $\boldsymbol{p}$
$\mathcal{B}\mathrm{in}(\mathbf{trials}, \boldsymbol{p})$	Multivariate binomial distribution with parameters **trials** and $\boldsymbol{p}$

Note that $\boldsymbol{y} \sim \mathcal{P}(\boldsymbol{\lambda})$ means that the y_i's are mutually independent and simulated according to a univariate Poisson distribution with parameter λ_i. The same convention is used for the Gaussian, Bernoulli and Binomial distributions.

Functions and norms

$f(\cdot)$	A function
$f \circ g$	Composition of functions f and g
$f(\boldsymbol{x}; \boldsymbol{\theta})$	A function of $\boldsymbol{x}$ parametrised by $\boldsymbol{\theta}$ ($f(\boldsymbol{x})$ is also used to lighten notation)
$\|\boldsymbol{x}\|_p$	L^p norm of $\boldsymbol{x}$

Sometimes, even if f denotes a function of a scalar argument, we apply it to a vector. $f(\boldsymbol{x})$ then denotes the application of f to the vector element–wise, namely $f(\boldsymbol{x}) = (f(x_1), \ldots, f(x_n))^{\mathrm{T}}$.

Acronyms

AGHQ	Adaptive Gauss–Hermite Quadrature
CV	Cross–Validation
EM	Expectation–Maximisation
FSA	Fisher Scoring Algorithm
GCV	Generalised Cross–Validation
GHQ	Gauss–Hermite Quadrature
GLM	Generalised Linear Model
GLMM	Generalised Linear Mixed Models
iid	independent and identically distributed
LASSO	Least Absolute Shrinkage and Selection Operator
LMM	Linear Mixed Model
MCEM	Monte Carlo Expectation Maximisation
MCLA	Monte Carlo Likelihood Approximation
MCNR	Monte Carlo Newton–Raphson
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
MRSE	Mean Relative Squared Error
MURSE	Mean Upper Relative Squared Error
PCA	Principal Component Analysis
PCR	Principal Component Regression
PING	Projected Iterated Normed Gradient
PLS	Partial Least Squares
PLSR	Partial Least Squares Regression
PQL	Penalised Quasi–Likelihood
REML	REstricted Maximum Likelihood
RMSE	Root Mean Square Error
SC	Supervised Component
SCGLR	Supervised Component–based Generalised Linear Regression
SCEM	Supervised Component EM
SDR	Sufficient Dimension Reduction
SMC	Sequential Monte Carlo
SR	Structural Relevance
StN	Signal to Noise
VPI	Variable–Powered Inertia

List of Figures

List of Tables

List of Algorithms

I

Introduction

Régularisation des Modèles Linéaires Généralisés

Initialement introduits par Nelder and Wedderburn (1972), les modèles linéaires généralisés (GLMs pour *Generalised Linear Models*) sont largement utilisés dans les problèmes de régression, car ils couvrent de nombreuses distributions de réponses appartenant à la famille exponentielle — gaussienne, Bernoulli, Poisson ou multinomiale pour ne citer que quelques exemples. Cette famille de distributions permet de modéliser de nombreux types de sorties (continues, binaires, multi–catégorielles ou de comptage...), et trouve donc des applications dans des domaines variés tels que la biologie, l'épidémiologie, l'écologie, les sciences sociales, l'économie, etc. Par ailleurs, comme la collecte d'une grande quantité de données est désormais la norme, le cadre de modélisation qui nous anime est celui d'un très grand nombre de variables explicatives avec un niveau de redondance possiblement élevé, incluant le cas de la grande dimension (plus de variables explicatives que d'individus). Ces redondances conduisent à des sur–ajustements, et même à des singularités dans le processus d'estimation, ce qui entraîne souvent des prédicteurs linéaires instables voire non identifiés. Pour faire face à ces problèmes, des techniques de régularisation ont été mises au point. Elles consistent à introduire des informations supplémentaires dans le processus d'estimation et permettent ainsi de résoudre un problème mal posé ou d'éviter un sur–ajustement.

Il existe essentiellement deux types de méthodes de régularisation : les méthodes de régularisation par pénalisation et celles fondées sur la construction de composantes.

Méthodes de régularisation par pénalisation. Ces méthodes maximisent la vraisemblance pénalisée par une certaine norme du vecteur des coefficients de régression. Basée sur la norme L_2, la régression ridge introduite par Hoerl and Kennard (1970a,b) est efficace lorsqu'un problème mal conditionné est diagnostiqué, mais que toutes les variables doivent être conservées dans le modèle. D'un autre côté, la régression LASSO (Tibshirani, 1996) introduit la norme L_1, et est donc particulièrement adaptée lorsque le "vrai" vecteur des coefficients de régression est creux (i.e. contenant un grand nombre de zéros). Cette méthode est donc principalement utilisée à des fins de sélection de variables. Malheureusement, le LASSO ne parvient pas à faire de sélections groupées et son efficacité n'est pas assurée dans les cas de grande dimension. Pour surmonter ces limitations, Zou and Hastie (2005) ont proposé l'elastic net qui, en combinant les normes L_2 et L_1, tente de tirer profit des meilleures caractéristiques des méthodes ridge et LASSO. Initialement développées pour le cas gaussien (ou plus largement pour tous les cas où la mise en oeuvre des moindres carrés pénalisés est appropriée), ces méthodes ont ensuite été étendues aux GLMs par Friedman et al. (2010).

Méthodes de régularisation par construction de composantes. Le deuxième type de méthodes consiste à construire des prédicteurs linéaires à partir de quelques composantes explicatives (i.e. des combinaisons linéaires des variables explicatives originales qui synthétisent au mieux la partie utile de l'information qu'elles contiennent). Le principal avantage des méthodes de régularisation par construction de composantes est qu'elles permettent de faciliter l'interprétation du modèle au travers de la décomposition du prédicteur linéaire sur des directions *interprétables*. Introduite par Jolliffe (1982) dans le cadre de la régression ordinaire, la première méthode de régression sur composantes est la régression sur composantes principales (PCR pour *Principal Component Regression*), où la réponse est régressée sur les composantes qui capturent la variabilité maximale dans le sous–espace des variables explicatives. Malheureusement, la PCR ne tient pas compte de la réponse lors de la construction des composantes. Une méthode alternative a ensuite été proposée par Wold (1966) et Wold et al. (1983) — la régression des moindres carrés partiels (PLSR pour *Partial Least Squares Regression*) — dans laquelle les composantes sont optimisées afin de maximiser la covariance empirique avec la réponse. Le premier travail traitant de la régularisation d'un GLM par construction de composantes a été celui de Marx (1996), qui a introduit le mécanisme PLS dans l'algorithme IRLS (pour *Iteratively Reweighted Least Squares*) d'estimation d'un GLM univarié. Dans son sillage et pour des réponses multivariées, Bry et al. (2013) ont développé une méthodologie appelée "régression linéaire généralisée sur composantes supervisées" (SCGLR pour *Supervised Component–based Generalised Linear Regression*), plus tard étendue et affinée par Bry et al. (2014, 2016, 2018). En tant qu'approche de type PLS, les composantes construites dans SCGLR s'appuient sur les structures fortes au sein

des variables explicatives, et doivent également prédire au mieux les réponses. Mais contrairement à PLS, les composantes SCGLR sont construites à l'aide d'un critère beaucoup plus flexible, permettant entre autre de spécifier le type de structures avec lesquelles les composantes doivent s'aligner dans le sous–espace explicatif (des faisceaux particuliers de variables, les composantes principales, etc).

Les méthodes hybrides. Quelques méthodologies combinant les techniques de régularisation par pénalisation et par construction de composantes ont également été développées, afin de tenter de tirer conjointement profit des avantages respectifs des deux types de régularisation. À ce titre, Fort and Lambert-Lacroix (2005) s'intéressent à un problème de classification dans un contexte de grande dimension ($p \gg n$). La méthode qu'ils proposent, combinant séquentiellement régression logistique pénalisée en norme L_2 (Eilers et al., 2001) et construction de composantes PLS , permet de stabiliser le processus d'estimation. Notons cependant que comme la combinaison est séquentielle plutôt qu'itérative, cette méthode ne semble pas mettre correctement à jour les poids d'estimation. Davantage intéressés par des problèmes de sélection de variables, Durif et al. (2017) ont étendu cette approche en construisant des composantes PLS parcimonieuses (Chun and Keleş, 2010) en lieu et place des composantes PLS originales.

Notre travail se concentre sur des situations où les variables explicatives sont nombreuses et considérées comme des proxys de dimensions latentes qui doivent être retrouvées et interprétées. C'est pourquoi nous nous intéressons principalement aux méthodes basées sur la construction de composantes. Cependant, des exemples élémentaires de régression mettent en défaut le pouvoir interprétatif des composantes principales et PLS. Considérons par exemple le modèle linéaire gaussien

$$\boldsymbol{y} \sim \mathcal{N}_n\left(\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta},\, \boldsymbol{\Sigma} = \boldsymbol{I}_n\right),$$

où la matrice de design $\boldsymbol{X}$ contient 20 variables explicatives divisées en deux parties :

- un faisceau de 10 variables corrélées dites "de nuisance", ne jouant donc aucun rôle explicatif

$$\boldsymbol{X_1} = \left[\, \boldsymbol{x_1} \mid \ldots\ldots\ldots\ldots\ldots\ldots \mid \boldsymbol{x_{10}} \,\right],$$

- et deux faisceaux contenant chacun 5 variables corrélées, qui prédisent conjointement la réponse $\boldsymbol{y}$

$$\left[\, \boldsymbol{X_2} \mid \boldsymbol{X_3} \,\right] = \left[\, \boldsymbol{x_{11}} \mid \ldots \mid \boldsymbol{x_{15}} \mid \boldsymbol{x_{16}} \mid \ldots \mid \boldsymbol{x_{20}} \,\right].$$

La Figure 1.1 montre le premier plan factoriel obtenu par la PCR et la PLSR. Pour la PCR, comme le faisceau $\boldsymbol{X_1}$ est celui d'inertie maximale, il apparaît le long de la première composante principale. Les deux faisceaux prédictifs $\boldsymbol{X_2}$ et $\boldsymbol{X_3}$ ne sont que partiellement capturés par la deuxième composante principale. Le premier plan factoriel obtenu par la PCR ne fournit donc que très peu d'informations sur la réponse $\boldsymbol{y}$. En revanche, la première composante PLS combine les deux faisceaux prédictifs en un seul, et la deuxième composante PLS s'aligne avec le faisceau de nuisance. Ainsi, bien que la régression PLS réussisse à détecter les variables pertinentes pour la modélisation de $\boldsymbol{y}$, elle ne détecte pas la présence de deux faisceaux prédictifs distincts, ce qui représente un réel problème pour l'interprétation du modèle.

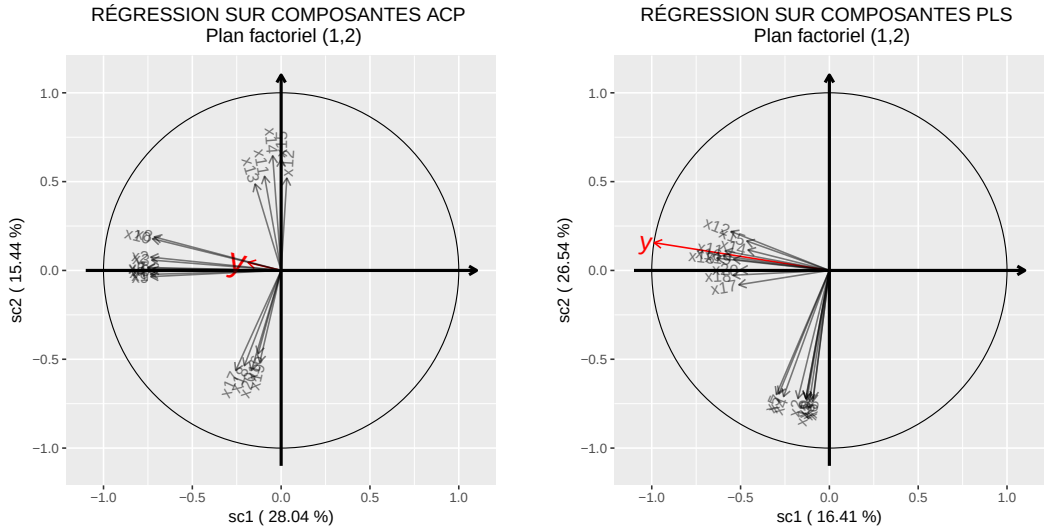


Figure 1.1 – *Échec du pouvoir interprétatif des régressions sur composantes principales et PLS.* *Nous donnons un exemple du premier plan factoriel issu de la régression sur composantes principales (figure de gauche) et de la régression sur composantes PLS (figure de droite). Les flèches noires représentent la projection orthogonale des variables explicatives sur le plan factoriel* $(1, 2)$*, tandis que la flèche rouge représente la projection orthogonale de la réponse* $\boldsymbol{y}$*. Le pourcentage d'inertie capturée par chaque composante est donné entre parenthèses.*

La Figure 1.2, quant à elle, montre le premier plan factoriel issu de la régression sur composantes supervisées, obtenu par la méthode SCGLR dans le cas gaussien. Le package R **SCGLR**, qui construit les composantes supervisées et génère les plans factoriels, est disponible à l'adresse https://scnext.github.io/SCGLR/. Contrairement aux régressions sur composantes principales et PLS, les deux premières composantes supervisées s'alignent sur les deux faisceaux prédictifs $\boldsymbol{X_2}$ et $\boldsymbol{X_3}$ respectivement.

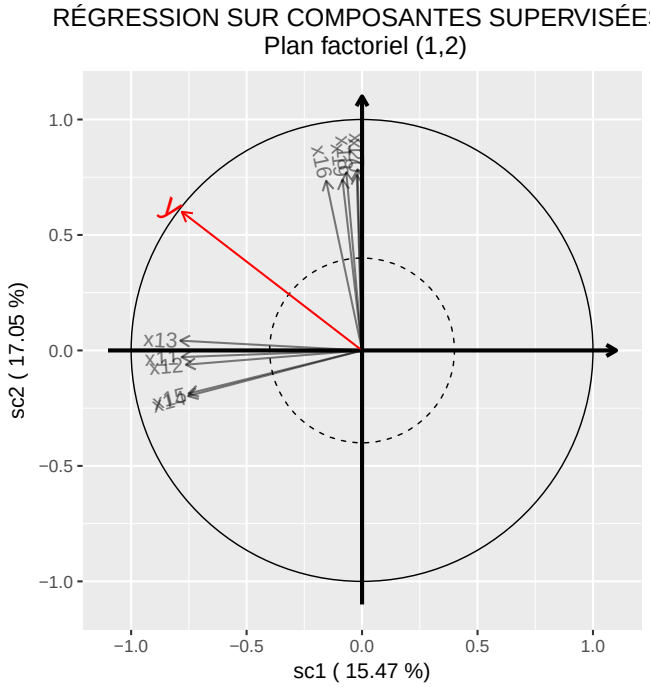


Figure 1.2 – *Pouvoir interprétatif de la régression sur composantes supervisées.* *En utilisant les mêmes données, nous représentons le plan factoriel* $(1, 2)$ *produit par la régression sur composantes supervisées. Pour faciliter l'interprétation du modèle, nous avons caché les variables explicatives ayant un cosinus avec le plan factoriel inférieur à* 0.4.

Tout en préservant les qualités prédictives de la régression PLS, SCGLR améliore considérablement l'interprétation des modèles. C'est la raison pour laquelle notre travail s'appuiera essentiellement sur SCGLR, plus souple et englobant les régressions sur composantes principales et PLS. Mais SCGLR, dans sa version originale, présente quelques limitations. Cette thèse vise donc à surmonter certaines d'entre elles.

(i) Dans la version initiale de SCGLR (Bry et al., 2013), les auteurs ont supposé que les observations étaient indépendantes, ce qui empêche de considérer des structures de dépendance complexes. Dans une perspective de modélisation de données groupées, nous avons choisi d'assouplir cette hypothèse en introduisant dans le modèle un effet aléatoire spécifique au groupe. Cette amélioration étend SCGLR aux modèles linéaires généralisés mixtes (GLMMs pour *Generalised Linear Mixed Models*) multivariés.

(ii) Dans de nombreux domaines (comme l'épidémiologie par exemple), les problèmes rencontrés par les praticiens se situent à l'interface entre régression impliquant des données de panel la modélisation avec un GLMM basé sur un grand nombre de variables explicatives redondantes. Le besoin de régularisation doit alors tenir compte des dépendances in-

duites par des mesures répétées sur chaque individu au cours du temps. Une nouvelle extension de SCGLR a donc été proposée : elle comprend à la fois des effets aléatoires spécifiques aux individus et spécifiques au temps, ces derniers ayant une structure autocorrélée.

(iii) En sciences sociales ou en sciences psychiatriques par exemple, la plupart des études exigent que certaines variables explicatives soient conservées en tant que telles dans le modèle (par exemple le sexe, l'âge, le niveau de scolarité, etc.). En effet, il arrive fréquemment que certaines variables présentent un intérêt particulier pour le praticien — qui souhaite donc estimer leurs effets marginaux avec précision — voire même que certaines variables soient connues pour être des facteurs de confusion. La plupart des problèmes de régression que nous considérons incluent donc deux catégories de variables explicatives. La première catégorie contient des variables abondantes et fortement corrélées nécessitant une réduction dimensionnelle, tandis que la deuxième contient quelques variables faiblement corrélées qui sont d'un intérêt particulier pour le praticien. Ces dernières variables doivent apparaître comme telles dans le modèle, et non par l'intermédiaire des composantes.

Régularisation par construction de composantes des GLMMs multivariés avec effets aléatoires spécifiques à l'individu

Le but de ce travail est de modéliser une réponse — ou un ensemble de réponses avec des distributions variées appartenant à la famille exponentielle — dans un contexte où les observations sont groupées. Comme suggéré plus haut, la modélisation de dépendances entre observations induit souvent l'utilisation d'effets aléatoires, d'où la nécessité de modéliser nos réponses selon des GLMMs. Or, en toute généralité, l'estimation des paramètres d'un GLMM par maximum de vraisemblance est difficile car la fonction de vraisemblance, s'exprimant comme une intégrale par rapport aux effets aléatoires, n'admet pas d'expression analytique. Plusieurs méthodes ont été proposées pour contourner ce problème. Les premières méthodes d'approximation de cette vraisemblance, de nature numériques, ont été mises au point pour la première fois au milieu des années 1980 (voir par exemple Anderson and Aitkin, 1985; Pinheiro and Bates, 1995; Breslow and Clayton, 1993; Shun and McCullagh, 1995). Des méthodes d'approximation stochastiques ont ensuite été développées dès le milieu des années 1990 (par exemple Zeger and Karim, 1991; Clayton, 1996; McCulloch, 1997; Knudson, 2016). Il existe également une

troisième procédure d'estimation, basée sur une approximation linéaire du modèle lui–même. Introduite par Schall (1991), la méthode implique un processus itératif alternant entre la linéarisation du modèle conditionnellement aux effets aléatoires, et l'estimation des paramètres au moyen de méthodes utilisées pour les modèles linéaires mixtes (LMMs pour *Linear Mixed Models*). Outre sa facilité de mise en œuvre et sa rapidité, cette méthode nous fournit un cadre intéressant pour développer notre procédure de régularisation.

Les GLMMs peuvent être vus comme un prolongement important des GLMs pour des observations groupées. Toutefois, leur utilisation est souvent limitée à quelques variables explicatives seulement, principalement parce que la présence de nombreux prédicteurs potentiellement redondants augmente le temps de calcul et produit des estimations instables. Pour surmonter cette limitation, le besoin de réduction dimensionnelle et/ou de régularisation doit tenir compte de la présence d'effets aléatoires dans le modèle. Contrairement aux GLMs, relativement peu d'articles traitent de techniques de régularisation de GLMMs. Pour des réponses gaussiennes, Eliot et al. (2011) ont d'abord proposé d'étendre la régression ridge aux LMMs. Puis plus récemment, dans une optique de sélection de variables, Schelldorfer et al. (2014); Groll and Tutz (2014) ont développé une méthode d'estimation d'un GLMM, qui fait intervenir la norme L_1 des coefficients de régression. Il s'agit donc ici d'une extension du LASSO pour GLMMs. Cependant, à notre connaissance, aucune méthode de régularisation par construction de composantes n'a jusqu'à alors été développée pour les GLMMs. Afin de combler ce vide, nous proposons d'étendre la méthode de Schall en y introduisant la construction de composantes supervisées à chaque itération. Comme dans la Figure 1.2, les composantes supervisées que nous proposons de construire sont conçues pour s'aligner avec les directions les plus prédictives et les plus interprétables dans le sous–espace explicatif.

Le cadre que nous considérons est celui d'un GLMM multivarié avec plusieurs vecteurs réponses $\boldsymbol{Y} = \left[\boldsymbol{y_1} \,|\, \ldots \,|\, \boldsymbol{y_q}\right]$, à expliquer par deux catégories de variables explicatives. La première catégorie est composée de quelques variables faiblement corrélées $\boldsymbol{A} = \left[\boldsymbol{a_1} \,|\, \ldots \,|\, \boldsymbol{a_r}\right]$ dont les effets marginaux doivent être quantifiés avec précision : ces variables doivent donc apparaître dans le modèle sans l'intermédiaire d'une composante. La deuxième catégorie est composée d'un grand nombre de variables corrélées $\boldsymbol{X} = \left[\boldsymbol{x_1} \,|\, \ldots \,|\, \boldsymbol{x_p}\right]$ qui peuvent contenir plusieurs dimensions $K < p$ structurellement pertinentes pour modéliser et prédire $\boldsymbol{Y}$. Les n observations ne sont plus supposées indépendantes mais forment N groupes distincts ($G_1, \ldots, G_N$ dans la Figure 1.3) dans lesquels les observations sont a priori dépendantes. C'est pourquoi, pour chaque réponse $\boldsymbol{y_k}$, un effet aléatoire $\boldsymbol{\xi_k}$ à N niveaux est utilisé pour modéliser la dépendance des observations dans chaque groupe. Le

prédicteur linéaire associé à la réponse $\boldsymbol{y_k}$ s'écrit donc

$$\boldsymbol{\eta_k^\xi} = \sum_{h=1}^{K} (\boldsymbol{X u_h})\, \gamma_{k,h} + \boldsymbol{A \delta_k} + \boldsymbol{U \xi_k},$$

où $\boldsymbol{U}$ est la matrice de design des effets aléatoires. Notez que $\gamma_{k,h}$ est le paramètre de régression associé à la composante $\boldsymbol{f_h} = \boldsymbol{X u_h}$ pour la $k^{\text{ème}}$ réponse, et $\boldsymbol{\delta_k}$ est le paramètre de régression associé aux variables explicatives additionnelles $\boldsymbol{A}$. Nous insistons ici sur le fait que les composantes $\{\boldsymbol{X u_1}, \ldots, \boldsymbol{X u_K}\}$ sont *communes à l'ensemble* des réponses $\boldsymbol{y_k}$, car elles sont précisément conçues pour capturer une dépendance structurelle entre les vecteurs réponses. Les composantes représentent donc les directions les plus pertinentes dans $\boldsymbol{X}$ pour *l'ensemble* des réponses, le paramètre $\gamma_{k,h}$ reflétant la pertinence de la $h^{\text{ème}}$ direction pour la $k^{\text{ème}}$ réponse.

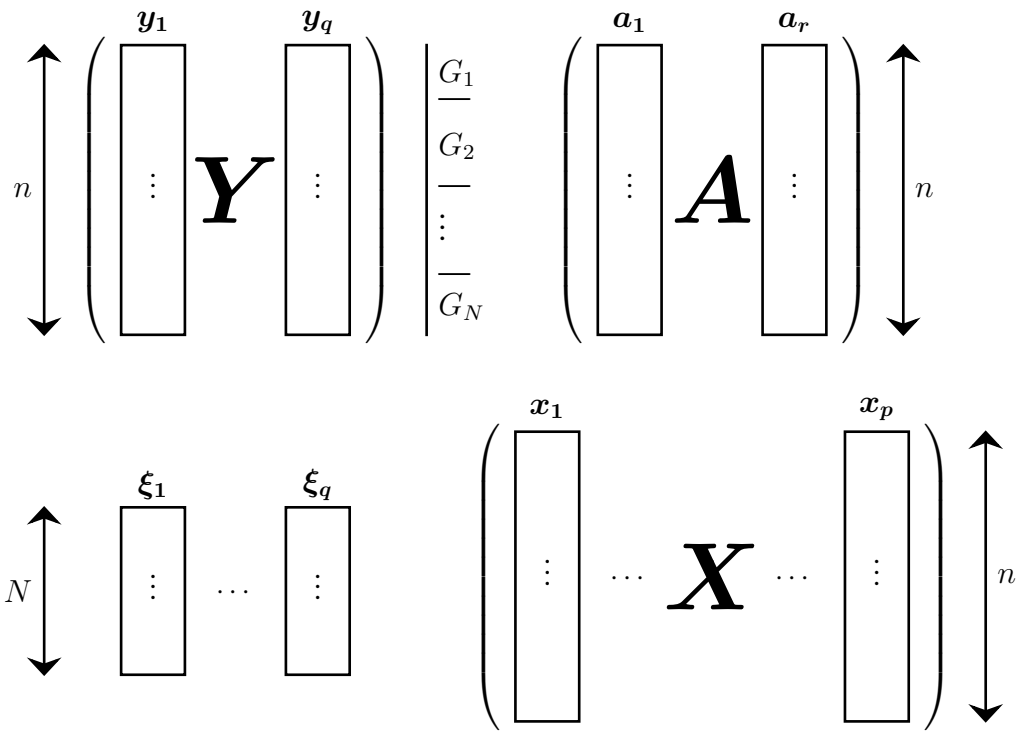


Figure 1.3 – ***Représentation de quelques vecteurs et matrices du modèle avec effets aléatoires spécifiques à l'individu.***

Supposons que les premières $h-1$ composantes sont construites et concaténées dans la matrice $\boldsymbol{F_{h-1}} = \left[\, \boldsymbol{f_1} \,|\, \ldots \,|\, \boldsymbol{f_{h-1}} \,\right]$. Pour calculer la $h^{\text{ème}}$ composante $\boldsymbol{f_h} = \boldsymbol{X u_h}$, notre méthode appelée "mixed–SCGLR" maximise un compromis entre une mesure de pertinence structurelle ϕ et une mesure de qualité d'ajustement ψ. Le programme à résoudre est alors

$$\begin{aligned} \max \quad & [\phi(\boldsymbol{u})]^s \times [\psi(\boldsymbol{u})]^{1-s},\ s \in [0,1], \\ \text{sous les contraintes} \quad & \|\boldsymbol{u}\| = 1 \ \text{ et } \ \boldsymbol{X u} \perp \boldsymbol{F_{h-1}}. \end{aligned}$$

ϕ est un critère de pertinence structurelle qui permet de spécifier le type de structures explicatives avec lesquelles les composantes doivent s'aligner, ainsi que la "résolution" à considérer pour la construction des composantes. Dans notre contexte, ψ est une mesure de qualité d'ajustement du modèle linéarisé qui apparait à chaque itération de l'algorithme de Schall.

Régularisation des GLMMs avec un effet aléatoire autorégressif spécifique au temps

Jusqu'à présent, nous nous sommes concentrés sur les GLMMs multivariées en considérant uniquement des effets aléatoires spécifiques aux individus. Dans cette partie, nous nous intéressons au développement de méthodes de régularisation dans le contexte spécifique des données de panel (impliquant des mesures répétées sur plusieurs individus aux mêmes dates).

Eliot et al. (2011) s'intéressent à la modélisation d'une réponse d'intérêt mesurée de façon répétée sur plusieurs individus au cours du temps, à des intervalles de temps potentiellement inégalement espacés. Afin de faire face aux corrélations potentiellement élevées au sein des variables explicatives, les auteurs proposent d'adapter la régression ridge à ce type de données répétées. Basé sur une log–vraisemblance complétée pénalisée par norme L_2 des coefficients de régression, l'algorithme Espérance–Maximisation (EM) qu'ils proposent permet un réglage adaptatif du paramètre de rétrécissement par validation croisée généralisée à chaque itération. Comme leur papier se focalise sur des données longitudinales, le modèle inclut un effet aléatoire propre à l'individu, mais n'inclut pas un effet aléatoire propre à la date et partagé par tous les individus.

Par ailleurs, on suppose souvent que les effets aléatoires propres à l'individu sont normalement distribués avec des niveaux indépendants. Toutefois, pour les données de panel, l'autocorrélation de l'effet aléatoire spécifique au temps semble naturelle. À ce titre, Karlsson and Skoglund (2004) considèrent à la fois des effets aléatoires propre à l'individu, et des effets aléatoires propre au temps munis d'une structure d'autocorréation. Ces derniers sont considérés comme des phénomènes latents (non pris en compte par les variables explicatives) affectant tous les individus et qui persistent dans le temps. Toutefois, les auteurs ne considèrent aucune situation où il est nécessaire de régulariser le modèle.

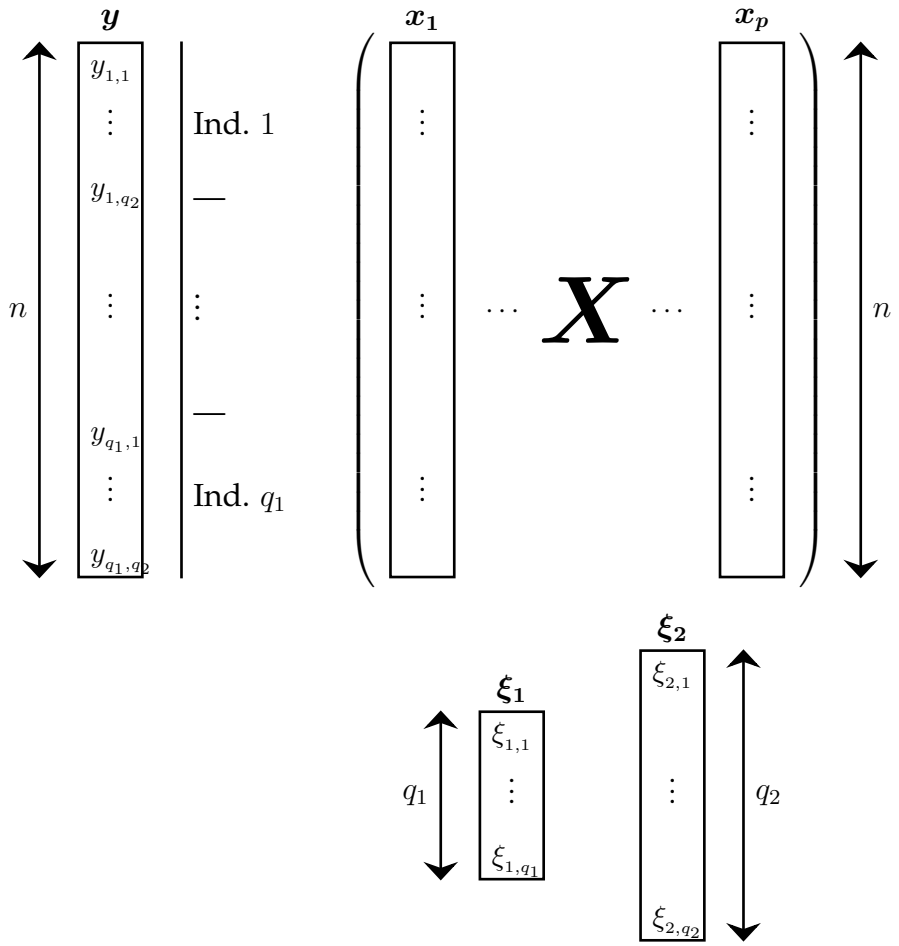


Figure 1.4 – ***Représentation de quelques vecteurs et matrices du modèle avec effets aléatoires spécifiques aux individus et au temps.***

Le défi a donc consisté à développer des méthodes de régularisation dans un contexte où les variables explicatives sont redondantes, pour un modèle avec un effet aléatoire propre à l'individu et un effet aléatoire autocorrélé spécifique au temps. Pour simplifier, nous considérons maintenant un seul vecteur réponse $\boldsymbol{y}$ qui concatène les sorties mesurés sur q_1 individus aux mêmes q_2 dates, et nous omettons les variables explicatives additionnelles. Par contre, deux effets aléatoires sont introduits dans notre modèle (voir la Figure 1.4) :

- $\boldsymbol{\xi_1}$ est l'effet aléatoire (à q_1 niveaux) propre aux individus, qui relie toutes les observations d'un individu à la même réalisation de l'effet aléatoire. Nous supposons $\boldsymbol{\xi_1} \sim \mathcal{N}_{q_1}(\mathbf{0}, \sigma_1^2 \boldsymbol{A_1})$, où σ_1^2 est la composante individuelle de la variance et $\boldsymbol{A_1}$ une matrice connue.
- $\boldsymbol{\xi_2}$ est l'effet aléatoire (à q_2 niveaux) spécifique au temps, qui relie toutes les observations d'une date donnée t à la même réalisation de l'effet aléatoire. Il modélise un phénomène latent possédant une certaine inertie temporelle, et affectant tous les individus. Nous supposons $\boldsymbol{\xi_2} \sim$

$\mathcal{N}_{q_2}\left(\mathbf{0}, \sigma_2^2 \boldsymbol{A_2}(\rho)\right)$, où σ_2^2 est la composante temporelle de la variance et $\boldsymbol{A_2}(\rho) = \left(\frac{\rho^{|i-j|}}{1-\rho^2}\right)_{1 \leqslant i,j \leqslant q_2}$. Le paramètre ρ reflète la structure de corrélation autorégressive d'ordre 1 de $\boldsymbol{\xi_2}$.

Tout d'abord, nous avons jugé nécessaire d'adapter la régression ridge proposée par Eliot et al. (2011) aux modèles comportant les deux types d'effets aléatoires décrits plus haut. En notant $\boldsymbol{\beta}$ le paramètre des effets fixes de dimension p, $\boldsymbol{\xi} := (\boldsymbol{\xi_1}^\mathrm{T}, \boldsymbol{\xi_2}^\mathrm{T})^\mathrm{T}$ et $\boldsymbol{U}$ la matrice de design associée, le modèle que nous considérons s'écrit

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\xi} + \boldsymbol{\varepsilon},$$

où $\boldsymbol{\varepsilon}$ est le vecteur des erreurs gaussiennes. Dans le sillage d'Eliot, nous suggérons d'estimer les paramètres $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_1^2, \sigma_2^2, \sigma_2^2, \rho)$ par un algorithme EM basé sur la log–vraisemblance pénalisée par la norme L_2 des coefficients de régression. La log–vraisemblance complétée pénalisée qui en résulte, $\ell^\mathrm{c}_\mathrm{ridge}$, est donc

$$\ell^\mathrm{c}_\mathrm{ridge}\left(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}\right) = \ell^\mathrm{c}\left(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}\right) - \frac{\lambda}{2}\left\|\boldsymbol{\beta}\right\|_2^2,$$

où ℓ^c désigne la log–vraisemblance complétée et $\lambda \geqslant 0$ le paramètre de rétrécissement usuel.

Le principal inconvénient de cette méthode est qu'elle pénalise les coefficients élevés parce qu'elle considère que les corrélations élevées entre les variables explicatives sont une pure nuisance, en ce sens qu'elles favorisent la confusion des effets. C'est la raison pour laquelle nous proposons ensuite une alternative qui mêle algorithme EM et construction de composantes supervisées (SCEM pour *Supervised Component–based EM*) : au lieu de soustraire un terme de *pénalité* à la log–vraisemblance complétée, nous proposons plutôt d'ajouter un terme *bonus* qui favorise l'alignement des composantes avec les directions les plus interprétables dans le sous–espace explicatif. Lorsqu'une seule composante est construite, le modèle que nous considérons est le suivant :

$$\boldsymbol{y} = (\boldsymbol{X}\boldsymbol{u})\,\gamma + \boldsymbol{U}\boldsymbol{\xi} + \boldsymbol{\varepsilon},$$

où $\boldsymbol{u}$ est un vecteur de pondération des variables explicatives et γ le paramètre de régression associé à la composante $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$. La log–vraisemblance complétée régularisée qui en résulte prend la forme d'un compromis entre ℓ^c et une mesure de pertinence structurelle ϕ. Elle s'écrit plus précisément

$$\ell^\mathrm{c}_\mathrm{SC}\left(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}\right) = (1-s)\,\ell^\mathrm{c}\left(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}\right) + s\,\log\left[\phi\left(\boldsymbol{u}\right)\right],$$

où $\boldsymbol{\theta} = (\boldsymbol{u}, \gamma, \sigma_1^2, \sigma_2^2, \rho)$ et $s \in [0, 1]$. Notez que SCEM est également conçue pour la recherche de plusieurs composantes orthogonales.

Enfin, des extensions des deux méthodes précédentes pour les GLMMs ont également été développées. Comme dans l'algorithme de Schall, nous proposons une méthode qui alterne entre linéarisation du modèle et estimation des paramètres. Mais pour tenir compte à la fois du niveau élevé de redondance dans $\boldsymbol{X}$ et de la structure autocorrélée de l'effet aléatoire spécifique au temps, nous suggérons de remplacer l'étape d'estimation usuelle (impliquant des systèmes de Henderson particuliers) par un algorithme EM incluant une pénalité en norme L_2 ou la construction de composantes supervisées.

Plan de la thèse

En résumé, cette thèse porte sur la régularisation des GLMMs. Les Chapitres 3 et 4 sont de brefs chapitres d'état de l'art : le Chapitre 3 présente les méthodes de régularisation les plus couramment utilisées pour l'estimation d'un GLM, tandis que le Chapitre 4 se concentre sur les méthodes d'estimation usuelles d'un GLMM (sans régularisation). Les chapitres suivants sont consacrés à nos propres contributions aux méthodes de régularisation des GLMMs. Le Chapitre 5 étend la méthode SCGLR aux données groupées dans le contexte des GLMMs multivariés. Nous proposons d'adapter l'algorithme de Schall à des prédicteurs linéaires basés sur des composantes particulières, qui sont construites en tenant compte à la fois de la structure des variables explicatives et de leur capacité à prédire les réponses. Notre méthode, appelée "mixed–SCGLR", est testée sur des données simulées et réelles, et est comparée aux méthodes de régularisation classiques telles que ridge et LASSO. Le Chapitre 6 s'intéresse d'abord aux données de panel gaussiennes, et propose des modèles incluant à la fois des effets aléatoires propres aux individus et des effets aléatoires spécifiques au temps. Dans ce contexte, nous avons d'abord développé un algorithme ridge–EM, puis un algorithme EM basé sur la construction de composantes supervisées qui améliore grandement l'interprétation des modèles. Une extension de ces algorithmes est également proposée dans le cas non gaussien. Enfin, le Chapitre 7 donne un aperçu des travaux en cours et des perspectives.

II

Introduction (for English speakers)

Regularising Generalised Linear Models

Initially introduced by Nelder and Wedderburn (1972), Generalised Linear Models (GLMs) are widely used in regression frameworks, since they cover numerous response distributions belonging to the exponential family — for instance Gaussian, Bernoulli, Poisson or multinomial. This family of distributions allows modelling many types of outcomes (continuous, binary, counts, multi-categorical...), and thus addresses a very large scope of applications in many areas: biology, epidemiology, ecology, social sciences, economy, etc. Besides, since it is nowadays increasingly possible to collect large amounts of data, our modelling framework is that of too many explanatory variables with a possibly high degree of redundancy, up to and including high–dimensional data. These redundancies lead to overfitting or even singularities in the estimation process, coupled with unstable if not unidentified linear predictors. To face this problem, regularisation techniques have been developed, introducing additional information in the estimation process in order to solve an ill–posed problem or to prevent overfitting.

Two types of regularisation methods can be distinguished: penalty–based methods and component–based ones.

Penalty–based methods. These methods maximise the likelihood penalised by some norm of the coefficient vector. Based on the L_2–norm, the ridge regression introduced by Hoerl and Kennard (1970a,b) is useful when an ill–conditioned problem is diagnosed, while all variables must be kept in the

model. On the contrary, the LASSO regression (Tibshirani, 1996) involves the L_1–norm that inherently induces the sparsity of the solution. This method is therefore mainly used for variable selection purposes. Unfortunately, the LASSO fails to do grouped selection and its effective functioning is not ensured in the high–dimensional case. To overcome these limitations, Zou and Hastie (2005) proposed the elastic net which, combining the L_1– and L_2–norms, attempts to take advantage of the best features of both ridge and LASSO methods. Initially developed for the Gaussian case, these methods were generalised to the GLMs by Friedman et al. (2010).

Component–based methods. The second type of methods builds linear predictors from a few explanatory components, i.e. linear combinations of the original explanatory variables which best synthesise the useful part of the information they contain. The main advantage of component–based regularisation methods is that they allow an easy model interpretation through the decomposition of the linear predictor on *interpretable* directions. Introduced by Jolliffe (1982) in the ordinary regression framework, the first component–based regression method is the Principal Component Regression (PCR), where the response is regressed on the components that capture the maximal variability in the explanatory subspace. Unfortunately, the PCR ignores the response while building components. An alternative method was then proposed by Wold (1966); Wold et al. (1983) — the Partial Least Squares (PLS) regression — in which components are optimised so as to maximise the empirical covariance with the response. The first work dealing with component–based regularisation of a GLM was that of Marx (1996), who introduced the PLS mechanism into the Iterative Re–weighted Least Squares algorithm of a univariate GLM. In his wake and for multiple–response settings, Bry et al. (2013) have developed a methodology based on supervised components, named Supervised Component–based Generalised Linear Regression (SCGLR), later extended and refined in Bry et al. (2014, 2016, 2018). As a PLS–type approach, the construction of components in SCGLR is guided both by the correlation structure of the explanatory variables and by the prediction quality of the responses. Nevertheless, unlike PLS, SCGLR involves a general and flexible criterion allowing to specify the type of structure with which components should align in the explanatory subspace (e.g. variable bundles, principal components, etc).

Hybrid methods. Some methodologies combining penalisation techniques and component–based regularisation have also been developed, in order to benefit from the advantages of both frameworks. Fort and Lambert-Lacroix (2005) address for instance the question of classification in a high–dimensional setting ($p \gg n$). The method they proposed, combining PLS and ridge penalised logistic regression sequentially (Eilers et al., 2001), allows to stabilise the estimation process. But because the combination is sequential instead of iterative, this method does not deal with estimation weights properly, never updating them. More focused on variable–selection issues, Durif et al. (2017)

have extended this approach by using a sparse PLS (Chun and Keleş, 2010) instead of PLS.

Our work focuses on situations where the explanatory variables are many and considered as proxies to latent dimensions which must be found and interpreted. We will therefore mainly consider component–based methods. However, there are very simple regression frameworks where the model interpretation provided by classical component–based methods (namely PCR and PLSR) remains difficult. Consider for example the following Gaussian linear model

$$\boldsymbol{y} \sim \mathcal{N}_n\left(\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta},\, \boldsymbol{\Sigma} = \boldsymbol{I}_n\right),$$

where the design matrix $\boldsymbol{X}$ contains 20 explanatory variables divided into two parts:

- a nuisance variable–bundle of 10 correlated variables without explanatory role,
$$\boldsymbol{X_1} = \left[\, \boldsymbol{x_1} \mid \ldots\ldots\ldots\ldots\ldots\ldots \mid \boldsymbol{x_{10}} \,\right],$$
- and two smaller bundles of 5 correlated variables each that together predict $\boldsymbol{y}$,
$$\left[\, \boldsymbol{X_2} \mid \boldsymbol{X_3} \,\right] = \left[\, \boldsymbol{x_{11}} \mid \ldots \mid \boldsymbol{x_{15}} \;\mid\; \boldsymbol{x_{16}} \mid \ldots \mid \boldsymbol{x_{20}} \,\right].$$

Figure 2.1 shows the first component plane obtained by PCA and PLS. For PCA, as bundle $\boldsymbol{X_1}$ is the one with maximum inertia, it appears along the first principal component. The two predictive bundles $\boldsymbol{X_2}$ and $\boldsymbol{X_3}$ are only partially captured by the second principal component. The first component plane obtained by PCA thus provides very little information on response $\boldsymbol{y}$. By contrast, the first PLS component combines the two predictive bundles into a single one, and the second component aligns with the nuisance bundle. Although the PLS regression is successful in detecting variables relevant for modelling $\boldsymbol{y}$, not detecting the presence of two distinct predictive bundles is a real problem for model interpretation.

By contrast, Figure 2.2 shows the first component plane obtained by Supervised Component regression, which is what SCGLR boils down to in the Gaussian case. The R package **SCGLR**, which calculates the components and generates the explanatory component planes, is available on `https://scnext.github.io/SCGLR/`. Unlike PCA and PLS regressions, the first two supervised components align with the predictive variable–bundles $\boldsymbol{X_2}$ and $\boldsymbol{X_3}$ respectively.

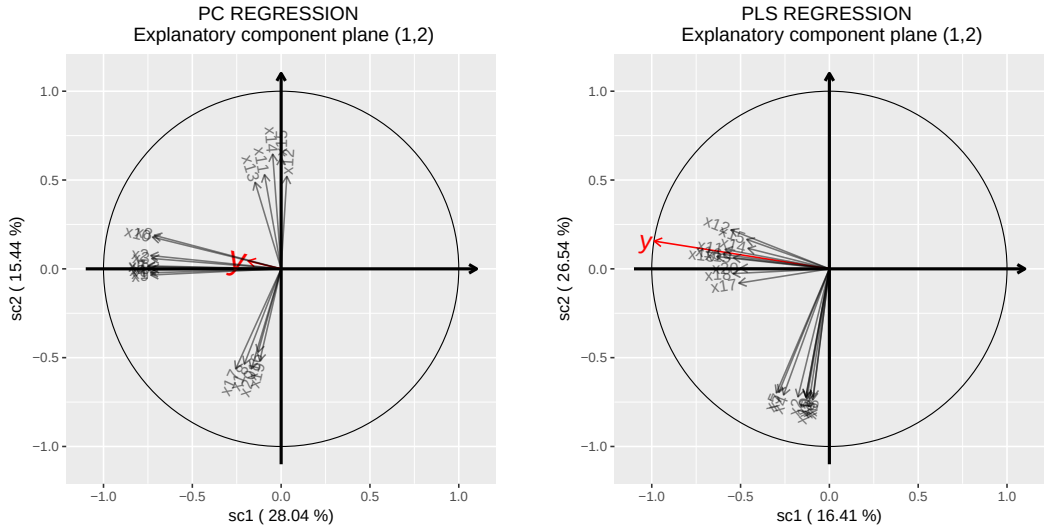


Figure 2.1 – *Failure of the interpretative power of PC and PLS regressions.* *We give an example of the first two–component planes given by the Principal Component Regression (left) and the Partial Least Squares Regression (right). The black arrows represent the orthogonal projection of the explanatory variables on component plane* $(1, 2)$*, and the red one represents the orthogonal projection of* $\boldsymbol{y}$*. The percentage of inertia captured by each component is given in parentheses.*

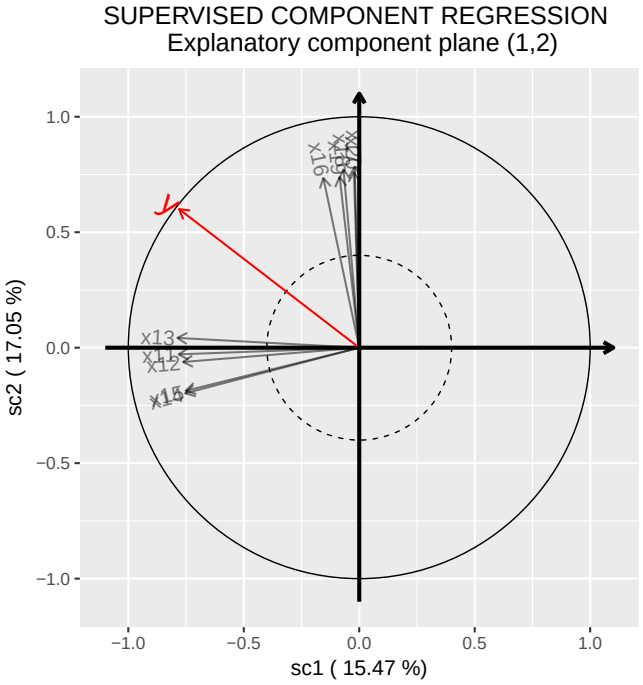


Figure 2.2 – *Interpretative power of the SC regression.* *Using the same data, we present the component plane* $(1, 2)$ *issued from the Supervised Component Regression. For an easier model interpretation, the explanatory variables having a cosine below* 0.4 *with the component plane are hidden.*

While preserving the predictive qualities of PLS regression, SCGLR significantly improves model interpretation. That is why our work stems from SCGLR, which is more flexible and includes the previous two methods. Now, SCGLR, in its original version, does have limitations, and this thesis aims at overcoming some of them.

(i) In the initial version of SCGLR (Bry et al., 2013), the authors assumed that the observations are independent, which prevents considering complex dependence structures. In a perspective of grouped data (or clustered data) modelling, we have chosen to relax this hypothesis by introducing a group–specific random effect into the model. This improvement extends SCGLR to multivariate Generalised Linear Mixed Models (GLMMs).

(ii) In many fields (such as epidemiology for instance), the problems encountered by practitioners are at the interface between regression with panel data and modelling with a GLMM based on a large number of redundant explanatory variables. The need for regularisation has to accommodate the dependencies induced by repeatedly measuring an outcome on each individual over time. A new extension of SCGLR has therefore been proposed: it includes both individual– and time–specific random effects, the latter having an autocorrelated structure.

(iii) In Social Sciences or Psychiatric Sciences for instance, most studies require that some explanatory variables be kept as such in the model (e.g. gender, age, level of education, etc) because they are assumed to be interesting *per se* or because they are known to be confounding factors. Most of the regression frameworks we consider will therefore include two categories of explanatory variables. The first category consists of abundant and highly correlated variables requiring dimension–reduction. The second category consists of few weakly correlated variables selected so as to preclude instability of their estimated coefficients. These variables must appear as such in the model, without being mediated by components.

Component–based regularisation of multivariate GLMMs with individual–specific random effects

In the present work, we aim at modelling a response — or a set of responses with probability distributions in the exponential family — in the framework of grouped (or clustered) data. As suggested above, the use of random effects is widespread in this context, hence the need to consider GLMMs.

Now, for the most general distribution assumptions in such models, parameter estimation faces the intractability of the likelihood expressed as an integral with respect to the random effects. Several methods have been proposed to tackle this issue. Numerical approximation methods of this likelihood were first developed in the mid–1980s (see, for instance Anderson and Aitkin, 1985; Pinheiro and Bates, 1995; Breslow and Clayton, 1993; Shun and McCullagh, 1995), followed by stochastic approximation methods since the mid–1990s (for example Zeger and Karim, 1991; Clayton, 1996; McCulloch, 1997; Knudson, 2016). A third type of estimation procedure is actually available, based on a linear approximation of the model itself. Introduced by Schall (1991), the method involves an iterative process alternating the linearisation of the model conditional on the random effects, and the estimation of the parameters using adapted linear mixed models methods. Besides its ease of implementation and computing speed, this method provides us with an interesting framework to develop our regularisation procedure.

GLMMs are an important extension of GLMs for clustered observations. However, their use is often restricted to few explanatory variables, mainly because the presence of many potentially highly correlated predictors increases the computational costs and yields unstable estimates. To overcome this limitation, the need for dimension–reduction and/or regularisation has to accommodate the presence of random effects in the model. Unlike with GLMs, relatively few articles deal with regularisation techniques for GLMMs. For Gaussian responses, Eliot et al. (2011) first proposed to extend the ridge regression to Linear Mixed Models (LMMs). Then more recently, and with a view towards variable selection, Schelldorfer et al. (2014); Groll and Tutz (2014) proposed an L_1–penalised algorithm for fitting a high–dimensional GLMM, by combining Laplace approximation and cyclic coordinate gradient descent. However, to our knowledge, no component–based regularisation method is available for GLMMs. In order to fill this gap, we propose to combine Schall's iterative model linearisation with a component–based regularisation at each step. As in Figure 2.2, the supervised components we propose to build are intended to align with the most predictive and interpretable directions in the explanatory space.

The framework we consider is that of a multivariate GLMM with multiple response–vectors $\boldsymbol{Y} = \left[\boldsymbol{y_1} \mid \ldots \mid \boldsymbol{y_q}\right]$ to be explained by two categories of explanatory variables. The first category consists of few weakly correlated variables $\boldsymbol{A} = \left[\boldsymbol{a_1} \mid \ldots \mid \boldsymbol{a_r}\right]$ whose marginal effects need to be precisely quantified: they have therefore to appear in the model without the mediation of a component. The second category consists of abundant correlated variables $\boldsymbol{X} = \left[\boldsymbol{x_1} \mid \ldots \mid \boldsymbol{x_p}\right]$ that may contain several unknown structurally relevant dimensions $K < p$ important to model and predict $\boldsymbol{Y}$. The n observations

are no longer assumed independent in that they form N groups ($G_1, \ldots, G_N$ in Figure 2.3) within which the observations are a priori dependent. That is why for each response $\boldsymbol{y_k}$, a N–level random effect $\boldsymbol{\xi_k}$ is used to model the dependence of observations within each group. In this context, the associated component–based linear predictor we consider writes

$$\boldsymbol{\eta_k^{\xi}} = \sum_{h=1}^{K} (\boldsymbol{X u_h})\, \gamma_{k,h} + \boldsymbol{A \delta_k} + \boldsymbol{U \xi_k},$$

where $\boldsymbol{U}$ is the random–effect design matrix. Note that $\gamma_{k,h}$ is the regression parameter associated with component $\boldsymbol{f_h} = \boldsymbol{X u_h}$ for the k–th response, and $\boldsymbol{\delta_k}$ is the regression parameter associated with additional explanatory variables $\boldsymbol{A}$. We would like to emphasise that the components $\{\boldsymbol{X u_1}, \ldots, \boldsymbol{X u_K}\}$ are *common* to *all* the $\boldsymbol{y_k}$'s as they are designed to capture a structural dependence between them. The components thus represent the most relevant directions in $\boldsymbol{X}$ for *all* responses. Parameter $\gamma_{k,h}$ then reflects the relevance of the h–th direction specific to the k–th response.

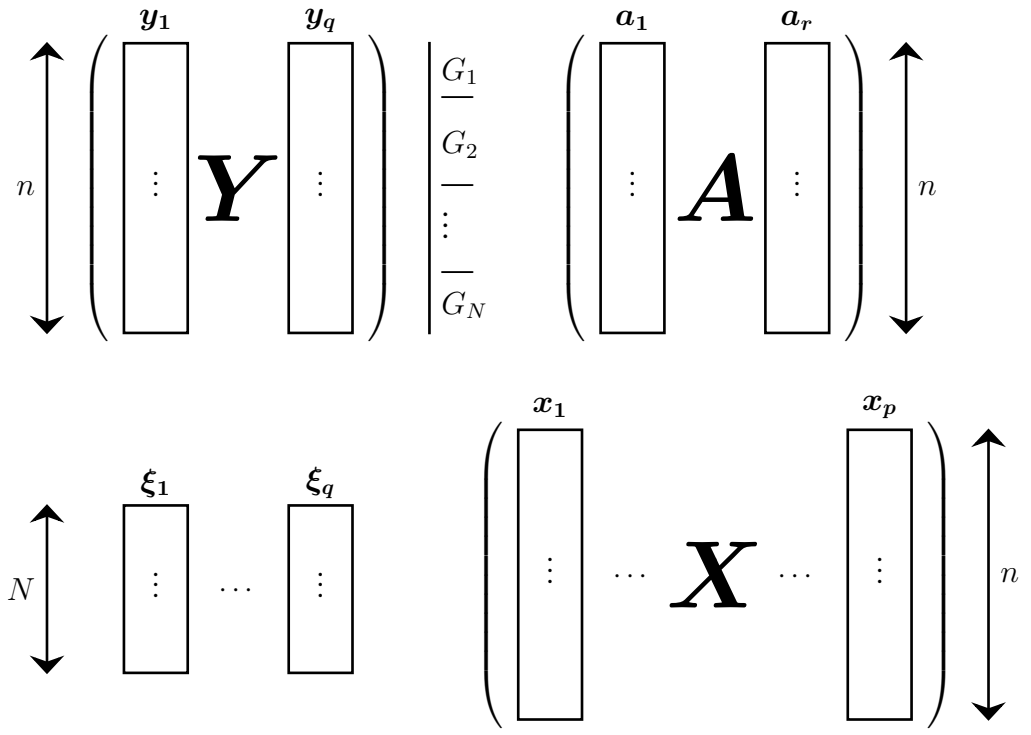


Figure 2.3 – *Some vectors and matrices of the model with individual–specific random effects.*

Suppose the first $h-1$ components are built and concatenated into matrix $\boldsymbol{F_{h-1}} = \left[\boldsymbol{f_1} \mid \ldots \mid \boldsymbol{f_{h-1}}\right]$. In order to compute the h–th component, namely $\boldsymbol{f_h} = \boldsymbol{X u_h}$, our method, named "mixed–SCGLR", attempts a trade–off between a Structural Relevance (SR) measure, ϕ, and a Goodness–of–Fit (GoF)

measure, ψ. The program to be solved is then

$$\begin{aligned} \max \quad & [\phi(\boldsymbol{u})]^s \times [\psi(\boldsymbol{u})]^{1-s}, \ s \in [0,1], \\ \text{subject to} \quad & \|\boldsymbol{u}\| = 1 \ \text{ and } \ \boldsymbol{X}\boldsymbol{u} \perp \boldsymbol{F_{h-1}}. \end{aligned}$$

ϕ is a SR flexible criterion allowing to specify the type of explanatory structures with which components should align, as well as the "resolution" to be considered for the construction of the component. ψ turns out to be a GoF measure of Schall's linearised regression model of the multivariate response on the additional explanatory variables and on the previously calculated components.

Regularisation of GLMMs with an autoregressive time–specific random effect

Up to now, we focused on multivariate GLMMs only considering individual–specific random effects. We are in this part more interested in developing regularisation methods within the specific framework of panel data (involving repeated measures on several individuals at the same time–points).

Eliot et al. (2011) focus on the correlated response setting in which a single outcome of interest is measured repeatedly on several individuals over time, at potentially unevenly spaced time intervals. In order to handle the potentially high correlations between explanatory variables, they propose to extend the ridge regression to this context. Based on a penalised complete log–likelihood, the Expectation–Maximisation (EM) algorithm they suggest includes a new step to find the best shrinkage parameter using a Generalised Cross–Validation (GCV) scheme at each iteration. As their paper considers longitudinal data, the model includes an individual–specific random effect but does not include a time–specific random effect common to all the individuals.

Individual–specific random effects are often assumed normally distributed with independent levels. However, for panel data frameworks, the question of the autocorrelation of the time–specific random effect naturally arises. That is why, for instance, Karlsson and Skoglund (2004) consider both individual– and autocorrelated time–specific random effects. The latter are viewed as latent phenomena (not accounted for by the explanatory variables) affecting all individuals and persistent over time. However, the authors do not consider any situation where it is necessary to regularise the model.

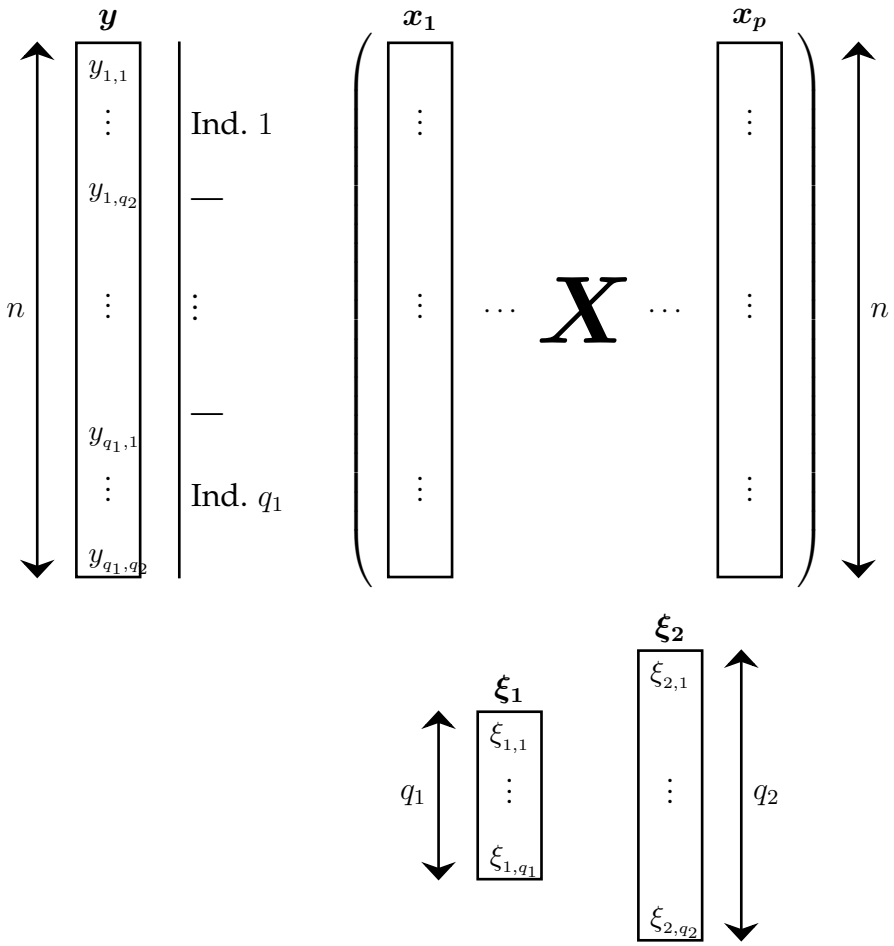


Figure 2.4 – ***Some vectors and matrices of the model with individual– and time–specific random effects.***

The challenge was to combine both issues, by proposing regularisation methods that tackle redundant explanatory variables, in a model with both individual– and autocorrelated time–specific random effects. In order to simplify the framework, let us now consider only one response vector $\boldsymbol{y}$, which concatenates the outputs measured on q_1 individuals at the same q_2 time–points. In addition, to focus on regularisation, let us no longer involve additional explanatory variables. By contrast, two random effects are introduced in our model (see Figure 2.4):

- $\boldsymbol{\xi_1}$ is the q_1–level individual random effect, which links all the observations of an individual to the same realisation of the random effect. We assume $\boldsymbol{\xi_1} \sim \mathcal{N}_{q_1}(\boldsymbol{0}, \sigma_1^2 \boldsymbol{A_1})$, where σ_1^2 is the individual variance component and $\boldsymbol{A_1}$ is a known matrix.
- $\boldsymbol{\xi_2}$ is the q_2–level time–specific random–effect which links all the observations at a time t to the same realisation of the random effect. It models a latent phenomenon with a certain temporal inertia, impacting all the

individuals. We assume $\boldsymbol{\xi_2} \sim \mathcal{N}_{q_2}(\mathbf{0}, \sigma_2^2 \boldsymbol{A_2}(\rho))$, where σ_2^2 is the time–specific variance component and $\boldsymbol{A_2}(\rho) = \left(\frac{\rho^{|i-j|}}{1-\rho^2}\right)_{1 \leqslant i,j \leqslant q_2}$. Parameter ρ reflects the order–1 autoregressive correlation structure of $\boldsymbol{\xi_2}$.

First, we considered necessary to adapt the "mixed ridge regression" proposed by Eliot et al. (2011) to the case of both individual– and autocorrelated time–specific random effects. Let $\boldsymbol{\beta}$ denote the p–dimensional fixed–effect parameter, $\boldsymbol{\xi} = (\boldsymbol{\xi_1}^{\mathrm{T}}, \boldsymbol{\xi_2}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{U}$ the associated design matrix. The model we consider writes

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\xi} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is the vector of Gaussian errors. In Eliot's wake, we suggest to estimate parameters $\boldsymbol{\theta} = (\beta, \sigma_1^2, \sigma_2^2, \rho)$ with an EM algorithm, based on the likelihood penalised by the L_2–norm of the regression coefficients. The resulting penalised complete log–likelihood $\ell^{\mathrm{c}}_{\mathrm{ridge}}$ is therefore

$$\ell^{\mathrm{c}}_{\mathrm{ridge}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) = \ell^{\mathrm{c}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) - \frac{\lambda}{2} \|\boldsymbol{\beta}\|_2^2,$$

where ℓ^{c} refers to the complete log–likelihood and $\lambda \geqslant 0$ is the usual shrinkage parameter.

The main drawback of this method is that it penalises the large coefficients because it considers the high correlations among the explanatory variables to be pure nuisance, since they favour effect–confusion. That is why we then propose a Supervised Component EM (SCEM) as an alternative: instead of subtracting a penalty term to the complete log–likelihood, we rather suggest to add a bonus term favouring the alignment of components with the most interpretable directions in the explanatory subspace. The single–component model we consider is

$$\boldsymbol{y} = (\boldsymbol{X}\boldsymbol{u})\,\gamma + \boldsymbol{U}\boldsymbol{\xi} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{u}$ is a loading–vector and γ is the regression parameter associated with component $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$. The resulting regularised complete log–likelihood, attempting a trade–off between ℓ^{c} and a SR measure ϕ, writes

$$\ell^{\mathrm{c}}_{\mathrm{SC}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) = (1-s)\,\ell^{\mathrm{c}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) + s\,\log\left[\phi(\boldsymbol{u})\right],$$

where $\boldsymbol{\theta} = (\boldsymbol{u}, \gamma, \sigma_1^2, \sigma_2^2, \rho)$ and $s \in [0, 1]$. Note that the SCEM is also designed to search for several orthogonal components.

Finally, extensions of the two previous methods in the GLMM framework are also developed. As in Schall's algorithm, our proposal alternates between linearisation and estimation steps. But in order to take into account both the

high level of redundancy in $\boldsymbol{X}$ and the autocorrelated structure of the time–specific random effect, we suggest to replace the usual estimation step (involving particular Henderson's systems) with a ridge–penalised or SC–regularised EM.

Overview

To sum things up, this thesis deals with the regularisation of Generalised Linear Mixed Models. Chapters 3 and 4 are brief state–of–the–art chapters: Chapter 3 presents some of the regularisation methods commonly used for GLM estimation, and Chapter 4 focuses on the usual estimation methods for GLMMs. Each subsequent chapter is then dedicated to our contributions to GLMM regularisation methods. Chapter 5 extends the Supervised Component–based Generalised Linear Regression (SCGLR) to deal with repeated measures in the context of multivariate GLMMs. Our proposal is to adapt the Schall's algorithm to particular component–based linear predictors, which are constructed taking into account both the explanatory variables structure and the prediction quality of the responses. Our extended method, "mixed–SCGLR", is tested on simulated and real data, and compared to classical regularisation methods such as ridge and LASSO. Chapter 6 focuses on the two–way random effect models (i.e including both individual– and time–specific random effects) for Gaussian panel data. We first developed a ridge–penalised EM for Gaussian panel data, and then a Supervised Component EM as an interesting alternative that greatly improves the model interpretation. An extension of these algorithms is also proposed in the non–Gaussian case. Finally, Chapter 7 outlines some of the ongoing work and perspectives.

III

Regularised GLM estimation

Contents

3.1 Introduction

Ordinary linear regression models express the expected value of a random response variable as a linear combination of observed explanatory variables. Such models inherently assume that changes in an explanatory variable lead to proportional changes in the expected value of the response variable. While this assumption is appropriate when the response variable has a Gaussian distribution, it is much less so for bounded (such as proportions for example), qualitative or discrete responses. Generalised Linear Models (GLM, Nelder and Wedderburn, 1972) cover all these situations by allowing response variables to belong to any distribution from the exponential family. This model class also expresses the expected value of the response variable as a function of a linear combination of the explanatory variables. That is why the GLMs are widely used in many areas. McCullagh and Nelder (1989) provide a complete overview of GLMs, and Fahrmeir and Tutz (1994) extend this overview to multivariate models.

Now, when a phenomenon is richly described through a high number of explanatory variables, the latter tend to have high redundancies. This results in identification troubles and in a severe lack of stability in the estimation of regression models. To make estimation of such models feasible, it is necessary to combine their likelihood with an extra criterion, so that maximising the combination yields regularised estimators. This chapter proposes to explore the main regularisation strategies for GLMs.

The chapter is organised as follows. After providing a description of a GLM in Section 3.2, Section 3.3 recalls some classical estimation methods. Then, Sections 3.4 and 3.5 respectively present a brief state of the art concerning the penalty– and component–based approaches to regularise GLMs. Finally, Section 3.6 discusses hybrid methods which combine both approaches.

3.2 Definition, assumptions and notations

Three elements are necessary to describe a GLM: a probability distribution for the random response variable, a predictor expressed as a linear combination of explanatory variables, and a link function relating the response variable to the explanatory variables.

($\boldsymbol{\mathcal{H}_1}$) Denote $\boldsymbol{y} = (y_1, \ldots, y_n)^\intercal$ the observed response vector, which is a realisation of the random vector $Y = (Y_1, \ldots, Y_n)^\intercal$. The Y_i's are assumed

independent and they have a distribution belonging to the exponential family. The density of Y_i can be expressed in the form

$$p\left(y_i\,;\theta_i\right) = \exp\left\{\frac{y_i\theta_i - b\left(\theta_i\right)}{a_i(\phi)} + c(y_i,\phi)\right\}, \tag{3.1}$$

where θ_i is a canonical parameter and ϕ a dispersion parameter usually known and related to the variance of the distribution. Functions b and c are known and specific to each distribution. The function a_i is of the form $a_i(\phi) = \frac{\phi}{w_i}$, where w_i is a known prior weight associated to the i^{th} observation.

For each distribution described by Equation 3.1, the expectation and variance of Y_i are expressed using functions a_i and b. Indeed, let $\ell\left(\boldsymbol{\theta}\,;\boldsymbol{y}\right) = \log\left[p\left(\boldsymbol{y}\,;\boldsymbol{\theta}\right)\right]$ be the log–likelihood function. The classic results

$$\begin{cases} \mathbf{0} = \mathbb{E}\left[\dfrac{\partial\ell\left(\boldsymbol{\theta}\,;\boldsymbol{y}\right)}{\partial\boldsymbol{\theta}}\right] \\ \mathbf{0} = \mathbb{E}\left[\dfrac{\partial^2\ell\left(\boldsymbol{\theta}\,;\boldsymbol{y}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right] + \mathbb{E}\left[\left(\dfrac{\partial\ell\left(\boldsymbol{\theta}\,;\boldsymbol{y}\right)}{\partial\boldsymbol{\theta}}\right)\left(\dfrac{\partial\ell\left(\boldsymbol{\theta}\,;\boldsymbol{y}\right)}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}}\right] \end{cases}$$

lead to

$$\begin{cases} \mathbb{E}(Y_i) = b'(\theta_i) \\ \mathbb{V}(Y_i) = a_i(\phi)\, b''(\theta_i). \end{cases}$$

There is therefore a direct link between the expectation and the variance of Y_i:

$$\mathbb{V}\left(Y_i\right) = a_i(\phi)\; b'' \circ b'^{-1}\big(\mathbb{E}(Y_i)\big).$$

Denote $\mu_i := \mathbb{E}(Y_i)$ and $v := b'' \circ b'^{-1}$. The independence of the Y_i's finally leads to

$$\mathbb{V}\left(Y\right) = \mathbf{Diag}\Big(a_i(\phi)\; v\left(\mu_i\right)\Big)_{i=1,\ldots,n}.$$

($\boldsymbol{\mathcal{H}_2}$) As in linear models, the explanatory variables are linearly involved in the model. The linear predictor, $\boldsymbol{\eta}$, can be expressed as

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is the fixed–effect parameter vector and $\boldsymbol{X}$ its associated design matrix.

($\boldsymbol{\mathcal{H}_3}$) By contrast, extending linear models, the linear predictor is related to the expected value of the data, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$, through a link function g such that

$$\boldsymbol{\eta} = g\left(\boldsymbol{\mu}\right).$$

Note that this link function g must be strictly monotonic and twice–differentiable.

Section 3.3 discusses the classical estimation methods of a GLM, focusing on maximum (quasi–) likelihood estimation.

3.3 Estimation methods without regularisation

3.3.1 Maximum likelihood estimation

The purpose of this section is to briefly present how to implement the maximum likelihood estimation for a GLM. In view of the independence of the Y_i's and according to (3.1), the log–likelihood function writes

$$\ell\left(\boldsymbol{\theta}\,;\boldsymbol{y}\right)=\sum_{i=1}^{n}\ell_i\left(\theta_i\,;y_i\right),$$

where

$$\ell_i\left(\theta_i\,;y_i\right)=\log\left[p\left(y_i\,;\theta_i\right)\right]=\frac{y_i\theta_i-b(\theta_i)}{a_i(\phi)}+c(y_i,\phi).$$

The maximum likelihood estimation equations for parameter $\boldsymbol{\beta}$ are obtained from the chain derivative rule. For each $i\in\{1,\ldots,n\}$ and for each $j\in\{1,\ldots,p\}$, we have then

$$\begin{aligned}\frac{\partial\ell_i}{\partial\beta_j}&=\frac{\partial\eta_i}{\partial\beta_j}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\ell_i}{\partial\theta_i}\\&=x_{ij}\,\frac{1}{g'(\mu_i)}\,\frac{1}{b''(\theta_i)}\,\frac{y_i-\mu_i}{a_i(\phi)}.\end{aligned}\tag{3.2}$$

As a result,

$$\frac{\partial\ell}{\partial\beta_j}=\sum_{i=1}^{n}x_{ij}\,\frac{1}{g'(\mu_i)}\,\frac{1}{b''(\theta_i)}\,\frac{y_i-\mu_i}{a_i(\phi)}=\sum_{i=1}^{n}x_{ij}\,\frac{1}{\mathbb{V}(Y_i)\left[g'(\mu_i)\right]^2}\,g'(\mu_i)(y_i-\mu_i).$$

Now, let us consider the two matrices

$$\boldsymbol{W}=\mathbf{Diag}\left(\frac{1}{\mathbb{V}(Y_i)\left[g'(\mu_i)\right]^2}\right)_{i=1,\ldots,n}$$

and

$$\frac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}=\mathbf{Diag}\left(\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i}\right)_{i=1,\ldots,n}=\mathbf{Diag}\Big(g'(\mu_i)\Big)_{i=1,\ldots,n}.$$

The maximum likelihood estimate is then the solution of the score equation

$$\boldsymbol{X}^{\mathrm{T}}\,\boldsymbol{W}\,\frac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}\,(\boldsymbol{y}-\boldsymbol{\mu})=\mathbf{0}.\tag{3.3}$$

Now (3.3) is a non–linear equation in $\boldsymbol{\beta}$, since the matrices $\boldsymbol{W}$ and $\dfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}$, as well as the vector $\boldsymbol{\mu}$, depend on $\boldsymbol{\beta}$. Two general methods can then be used: the Newton–Raphson (NR) method and the Fisher Scoring Algorithm (FSA). Depending on the chosen method, at iteration t, the new estimate $\boldsymbol{\beta}^{[t+1]}$ is obtained from the previous estimate $\boldsymbol{\beta}^{[t]}$ by

$$\boldsymbol{\beta}^{[t+1]} = \boldsymbol{\beta}^{[t]} - \left(\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} \right]^{[t]} \right)^{-1} \left(\frac{\partial \ell}{\partial \boldsymbol{\beta}} \right)^{[t]} \qquad \text{for NR,} \tag{3.4}$$

$$\boldsymbol{\beta}^{[t+1]} = \boldsymbol{\beta}^{[t]} - \left(\mathbb{E} \left[\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} \right]^{[t]} \right)^{-1} \left(\frac{\partial \ell}{\partial \boldsymbol{\beta}} \right)^{[t]} \qquad \text{for the FSA.} \tag{3.5}$$

The NR method and the FSA have essentially the same convergence properties, but as the FSA is often easier to compute, it is the most widely used. In addition, for models involving the canonical link, i.e. when $g = b'^{-1}$, they are strictly equivalent.

Equivalence of NR and FSA for a canonical link. As mentioned by Nelder and Wedderburn (1972) and detailed by McCullagh and Nelder (1989), the updates (3.4) and (3.5) are equivalent in the case of a canonical link. Indeed, the canonical link is defined by

$$\forall i \in \{1, \dots, n\}, \quad \eta_i = \theta_i = g(\mu_i) = \boldsymbol{x}_{i:}^{\mathrm{T}} \boldsymbol{\beta}, \text{ i.e. } g = b'^{-1}.$$

In this case, we have

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i)$$

so that the chain derivative rule (3.2) becomes

$$\begin{aligned} \frac{\partial \ell_i}{\partial \beta_j} &= \frac{\partial \eta_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \ell_i}{\partial \theta_i} = \frac{\partial \eta_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \ell_i}{\partial \theta_i} \\ &= x_{ij} \cancel{b''(\theta_i)} \frac{1}{\cancel{b''(\theta_i)}} \frac{y_i - \mu_i}{a_i(\phi)}. \end{aligned} \tag{3.6}$$

The terms involved the Hessian matrix in (3.4) are then expressed as

$$\begin{aligned} -\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k} &= -\frac{\partial}{\partial \beta_k} \left(x_{ij} \frac{y_i - \mu_i}{a_i(\phi)} \right) \\ &= x_{ij} x_{ik} \frac{b''(\theta_i)}{a_i(\phi)} = x_{ij} x_{ik} \frac{\mathbb{V}(Y_i)}{[a_i(\phi)]^2}, \end{aligned}$$

while the terms involved the Fisher information in (3.5) write

$$-\mathbb{E}\left[\frac{\partial^2 \ell_i}{\partial\beta_j \partial\beta_k}\right] = \mathbb{E}\left[\left(\frac{\partial \ell_i}{\partial \beta_j}\right)\left(\frac{\partial \ell_i}{\partial \beta_k}\right)\right]$$
$$= \mathbb{E}\left[x_{ij}x_{ik}\left(\frac{y_i - \mu_i}{a_i(\phi)}\right)^2\right] = x_{ij}x_{ik}\frac{\mathbb{V}(Y_i)}{[a_i(\phi)]^2}.$$

This proves the equivalence of the two methods for the canonical link function.

The FSA as an Iteratively Re–weighted Least Squares algorithm. It turns out that the update given by (3.5) can be written as

$$\boldsymbol{\beta}^{[t+1]} = \left(\boldsymbol{X}^\intercal \boldsymbol{W}^{[t]} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\intercal \boldsymbol{W}^{[t]} \boldsymbol{z}^{[t]}, \tag{3.7}$$

where

$$\boldsymbol{z}^{[t]} = \boldsymbol{X}\boldsymbol{\beta}^{[t]} + \left(\frac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}\right)^{[t]} \left(\boldsymbol{y} - \boldsymbol{\mu}^{[t]}\right)$$
$$\boldsymbol{W}^{[t]} = \mathbf{Diag}\left(\left\{a_i(\phi)\, v\left(\mu_i^{[t]}\right)\left[g'\left(\mu_i^{[t]}\right)\right]^2\right\}^{-1}\right)_{i=1,\ldots,n}.$$

Now, Equation 3.7 is the score equation for a weighted least squares regression of $\boldsymbol{z}^{[t]}$ on $\boldsymbol{X}$ with weights $\boldsymbol{W}^{[t]}$. Hence the estimates can be found using an Iteratively Re–weighted Least Squares (IRLS, Green, 1984) as described in Algorithm 3.1.

Algorithm 3.1: The IRLS algorithm.

Start with an initial guess $\boldsymbol{\mu}^{[0]} = \left(\mu_1^{[0]}, \ldots, \mu_n^{[0]}\right)$ and set $t = 0$

while *some convergence criterion not reached* **do**

Calculate the working response and the weight matrix: Set

$$\boldsymbol{z}^{[t]} = g\left(\boldsymbol{\mu}^{[t]}\right) + \mathbf{Diag}\left(g'\left(\boldsymbol{\mu}^{[t]}\right)\right)\left(\boldsymbol{y} - \boldsymbol{\mu}^{[t]}\right)$$
$$\boldsymbol{W}^{[t]} = \mathbf{Diag}\left(\left\{a_i(\phi)\, v\left(\mu_i^{[t]}\right)\left[g'\left(\mu_i^{[t]}\right)\right]^2\right\}^{-1}\right)_{i=1,\ldots,n}$$

Update the fixed–effect estimate and mean vector: Set

$$\boldsymbol{\beta}^{[t+1]} = \left(\boldsymbol{X}^\intercal \boldsymbol{W}^{[t]} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\intercal \boldsymbol{W}^{[t]} \boldsymbol{z}^{[t]}$$
$$\boldsymbol{\mu}^{[t+1]} = g^{-1}\left(\boldsymbol{X}\boldsymbol{\beta}^{[t+1]}\right)$$

$t \leftarrow t + 1$

end

3.3.2 Maximum quasi–likelihood estimation

In some circumstances (overdispersion for instance), the parametric form of the likelihood is misspecified, making it impossible to implement the maximum likelihood estimation described in Section 3.3.1. In response to this problem, the idea of Wedderburn (1974) was to develop an estimation procedure that only requires specifying the mean function of the response and a relationship between mean and variance functions. Relaxing the assumptions $\boldsymbol{\mathcal{H}_1}$, $\boldsymbol{\mathcal{H}_2}$, $\boldsymbol{\mathcal{H}_3}$, Wedderburn (1974) introduced the "quasi–likelihood" function, which allows the estimation to be performed in a more flexible way. More precisely, he only assumes that

- the observations $y_1, \ldots, y_n$ are independent,
- for each observation i, the mean is some known function of parameter $\boldsymbol{\beta}$, i.e. $\mathbb{E}(Y_i) = \mu_i(\boldsymbol{\beta})$,
- for each observation i, the link between variance and mean writes $\mathbb{V}(Y_i) = a_i(\phi)v(\mu_i)$, where v and a are some known functions.

The construction of the log–quasi–likelihood of observation i, namely $\widetilde{\ell}_i$, is based on its first and second derivatives with respect to μ_i. The author then suggests choosing a function that fulfils

$$\begin{cases} \mathbb{E}\left[\dfrac{\partial \widetilde{\ell}_i\left(\mu_i\,;y_i\right)}{\partial \mu_i}\right] = 0 \\ \mathbb{V}\left[\dfrac{\partial \widetilde{\ell}_i\left(\mu_i\,;y_i\right)}{\partial \mu_i}\right] = -\mathbb{E}\left[\dfrac{\partial^2 \widetilde{\ell}_i\left(\mu_i\,;y_i\right)}{\partial \mu_i^2}\right] = \dfrac{1}{a_i(\phi)v(\mu_i)}, \end{cases}$$

so that $\widetilde{\ell}_i$ has essentially the same properties as a log–likelihood. A function that satisfies these conditions is defined by the relation (Wedderburn, 1974; McCullagh and Nelder, 1989)

$$\frac{\partial \widetilde{\ell}_i\left(\mu_i\,;y_i\right)}{\partial \mu_i} = \frac{y_i - \mu_i}{a(\phi)v(\mu_i)}.$$

In the end, the expression of the log–quasi–likelihood results from the independence of observations:

$$\widetilde{\ell}\left(\boldsymbol{\mu}\,;\boldsymbol{y}\right) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{a_i(\phi)v(t)} dt.$$

It turns out that the same chain derivative rule as described in Section 3.3.1 leads to a "quasi–score" function $\mathcal{U}(\boldsymbol{\beta})$, so that the maximum quasi–likelihood

estimate is the solution of the "quasi–score" equation

$$\mathcal{U}(\beta) = \boldsymbol{G}^{\mathrm{T}}\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}\right) = \mathbf{0},$$

where $\boldsymbol{G} = \dfrac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}$ and $\boldsymbol{V} = \mathbf{Diag}\big(a_i(\phi)v(\mu_i)\big)_{i=1,\ldots,n}$. The sequence of parameter estimates generated by the FSA is then given by

$$\begin{aligned}\boldsymbol{\beta}^{[t+1]} &= \boldsymbol{\beta}^{[t]} - \left(\mathbb{E}\left[\frac{\partial^2 \widetilde{\ell}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\right]^{[t]}\right)^{-1}\left(\frac{\partial \widetilde{\ell}}{\partial \boldsymbol{\beta}}\right)^{[t]} \\ &= \boldsymbol{\beta}^{[t]} + \left(\boldsymbol{G}^{[t]^{\mathrm{T}}}\boldsymbol{V}^{[t]^{-1}}\boldsymbol{G}^{[t]}\right)^{-1}\boldsymbol{G}^{[t]^{\mathrm{T}}}\boldsymbol{V}^{[t]^{-1}}\left(\boldsymbol{y} - \boldsymbol{\mu}^{[t]}\right). \qquad (3.8)\end{aligned}$$

As pointed out by McCullagh and Nelder (1989), if there is a matrix $\boldsymbol{K}^{[t]}$ such as $\boldsymbol{G}^{[t]} = \boldsymbol{K}^{[t]}\boldsymbol{X}$, then the iteration given by (3.8) can be rewritten

$$\boldsymbol{\beta}^{[t+1]} = \left(\boldsymbol{X}^{\mathrm{T}}\widetilde{\boldsymbol{W}}^{[t]}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\widetilde{\boldsymbol{W}}^{[t]}\widetilde{\boldsymbol{z}}^{[t]},$$

where

$$\begin{aligned}\widetilde{\boldsymbol{W}}^{[t]} &= \boldsymbol{K}^{[t]^{\mathrm{T}}}\boldsymbol{V}^{[t]^{-1}}\boldsymbol{K}^{[t]} \\ \widetilde{\boldsymbol{z}}^{[t]} &= \boldsymbol{X}\boldsymbol{\beta}^{[t]} + \boldsymbol{K}^{[t]^{-1}}\left(\boldsymbol{y} - \boldsymbol{\mu}^{[t]}\right).\end{aligned}$$

As a result, the estimation by maximum quasi–likelihood leads to the same IRLS procedure as described in Algorithm 3.1, replacing $\boldsymbol{z}^{[t]}$ by $\widetilde{\boldsymbol{z}}^{[t]}$ and $\boldsymbol{W}^{[t]}$ by $\widetilde{\boldsymbol{W}}^{[t]}$. It is important to note that, if the variance function v used in the model is actually the natural variance function associated with a distribution from the exponential family, and if $\forall i \in \{1,\ldots,n\}, \mu_i = g^{-1}\left(\boldsymbol{x}_{i:}^{\mathrm{T}}\boldsymbol{\beta}\right)$, the maximum likelihood and maximum quasi–likelihood estimations are equivalent. The strength of the maximum quasi–likelihood estimation is that it allows other variance functions to be considered. For instance, it has proven to be extremely useful to estimate the dispersion parameter in quasi–Poisson models (Nelder and Pregibon, 1987; Godambe and Thompson, 1989), which assume $\mathbb{V}(Y_i) > \mathbb{E}(Y_i)$ instead of $\mathbb{V}(Y_i) = \mathbb{E}(Y_i)$.

Alternative estimation methods have also been developed within the Bayesian paradigm. The major difficulty with Bayesian methods in the context of GLMs is that in general, the posterior distribution cannot be found in closed form and so must be approximated, usually using Laplace approximations or some type of MCMC methods such as Gibbs sampling or the Metropolis–Hastings algorithm. The reader can refer to Dey et al. (2000) for detailed descriptions.

The rest of the chapter is devoted to regularisation techniques which are used, among other things, to solve overfitting issues that can occur in GLM.

Section 3.4 then presents the main penalty–based techniques (shrinkage methods) and Section 3.5 focuses on component–based ones.

3.4 Penalty–based regularisation methods

In order to avoid overfitting and reduce the variance of the prediction error, or to handle correlated explanatory variables, the first regularisation strategies available consist in introducing a penalty into the model estimation process. The two most common penalty–based regression techniques are the ridge regression, which involves the L_2–norm of the regression coefficient–vector, and the LASSO regression, which involves its L_1–norm. The elastic net is a trade–off between the two previous regularisation methods that combines L_1 and L_2 penalties. Section 3.4.1 presents these three penalty–based regularisations in the usual linear regression setting, for which parameter estimation can be performed by a penalised least squares approach. Section 3.4.2 then outlines their extensions to the context of GLMs.

3.4.1 Penalised least squares

In this section, we consider the usual setup for linear regression: the purpose is to estimate parameter $\boldsymbol{\beta} \in \mathbb{R}^p$ in the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector, $\boldsymbol{X}_{n\times p} = \left[\boldsymbol{x_1} \mid \ldots \mid \boldsymbol{x_p}\right]$ is the design matrix and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the vector of errors such that $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathbb{V}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$. For simplicity, we also assume that the response vector is centred and that each explanatory variable $\boldsymbol{x_j}$ is normalised, i.e.

$$\sum_{i=1}^{n} y_i = 0,$$
$$\sum_{i=1}^{n} x_{ij} = 0 \text{ and } \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1, \ \forall j \in \{1, \ldots, p\}.$$

3.4.1.1 Ridge regression (Tikhonov regularisation)

In this framework, the ridge regression cost function is a residual sum of squares penalised by the L_2–norm of the regression coefficients. This penalty

shrinks the coefficients towards zero, and also shrinks the coefficients of correlated explanatory variables towards each other. In matrix form, the ridge regression estimate is

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{ridge}} &= \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\} \\ &= \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} \right\},\end{aligned} \tag{3.9}$$

where $\lambda \geqslant 0$ is a tuning parameter that controls the amount of shrinkage. In order to make the size–constraint on the parameter more explicit, the problem (3.9) can be rewritten

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = \begin{cases} \arg\min\limits_{\boldsymbol{\beta}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \\ \text{subject to } \|\boldsymbol{\beta}\|_2^2 \leqslant R_\lambda, \end{cases} \tag{3.10}$$

where the radius R_λ is in bijection with the shrinkage parameter λ in (3.9). When the explanatory variables are highly redundant, their usual coefficients (without regularisation) are often poorly determined since they exhibit high variance. Imposing a size constraint on the coefficients reduces their variance and the effect–confusion.

Since its introduction by Hoerl and Kennard (1970a,b), the ridge regression has been used extensively, probably for three reasons.

(i) Equivalent problems (3.9) and (3.10) admit a closed–form solution, therefore easy to compute:

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}. \tag{3.11}$$

(ii) Like the Ordinary Least Squares (OLS) estimate, the ridge one (3.11) is a linear function of $\boldsymbol{y}$. But unlike the OLS estimate, the solution adds a positive constant to the diagonal of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ before inversion, making the problem non–singular even if $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is not of full rank. The ridge regression is therefore a response to the problem of collinearity of explanatory variables in high–dimensional settings, where the OLS approach fails.

(iii) Finally, Hoerl and Kennard (1970a,b) decompose the Mean Squared Error (MSE) of the ridge estimate into

$$\text{MSE}\left(\widehat{\boldsymbol{\beta}}_{\text{ridge}}\right) = \mathbb{E}\left(\left\|\widehat{\boldsymbol{\beta}}_{\text{ridge}} - \boldsymbol{\beta}\right\|_2^2\right) = \gamma_1(\lambda) + \gamma_2(\lambda),$$

where γ_1 is the total variance of the parameter estimate and γ_2 is a squared bias term, both expressed as a function of shrinkage parameter λ. Underlining that functions γ_1 and γ_2 are respectively monotonically

decreasing and increasing, the authors show that there always exists a $\lambda > 0$ such that

$$\mathrm{MSE}\left(\widehat{\boldsymbol{\beta}}_{\text{ridge}}\right) < \mathrm{MSE}\left(\widehat{\boldsymbol{\beta}}_{\text{OLS}}\right).$$

Interestingly, there is a close relationship between ridge and principal component regressions (introduced in Section 3.5.1.1), which provides a better understanding of how ridge regression works (Friedman et al., 2001). Indeed, let us consider the singular value decomposition of matrix $\boldsymbol{X}$, which has the form

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathsf{T}},$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices such that $\mathrm{span}\{\boldsymbol{U}\} = \mathrm{span}\{\boldsymbol{X}\}$ and $\mathrm{span}\{\boldsymbol{V}\} = \mathrm{span}\{\boldsymbol{X}^{\mathsf{T}}\}$, and where $\boldsymbol{D}$ is the diagonal matrix of the singular values $d_1 \geqslant d_2 \geqslant \ldots \geqslant d_p \geqslant 0$. We can then compare the predictions obtained by the OLS and ridge approaches:

$$\begin{aligned}
\widehat{\boldsymbol{y}}_{\text{OLS}} &= \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{OLS}} \\
&= \boldsymbol{X}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} \\
&= \sum_{j=1}^{p} \boldsymbol{u}_j\,\boldsymbol{u}_j^{\mathsf{T}}\boldsymbol{y} \\
&= \sum_{j=1}^{p} \langle\, \boldsymbol{u}_j \,|\, \boldsymbol{y} \,\rangle\, \boldsymbol{u}_j.
\end{aligned}
\qquad
\begin{aligned}
\widehat{\boldsymbol{y}}_{\text{ridge}} &= \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{ridge}} \\
&= \boldsymbol{X}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda\boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} \\
&= \sum_{j=1}^{p} \boldsymbol{u}_j \frac{d_j^2}{d_j^2+\lambda}\boldsymbol{u}_j^{\mathsf{T}}\boldsymbol{y} \\
&= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda}\,\langle\, \boldsymbol{u}_j \,|\, \boldsymbol{y} \,\rangle\, \boldsymbol{u}_j.
\end{aligned}$$

Now, the fact that $\boldsymbol{u}_j$ is the j^{th} normalised principal component of $\boldsymbol{X}$, associated with eigenvalue d_j^2. As a result, ridge regression slightly shrinks the directions in $\mathrm{span}\{\boldsymbol{X}\}$ having high variance (high d_j^2), but greatly shrinks the directions having small variance (small d_j^2).

In a nutshell, the ridge regression allows to get around the collinearity problems even if the numbers of explanatory variables is large. It is noteworthy that with this method, all the explanatory variables are included in the model. The ridge regression is then inefficient to explain a response when some of the explanatory variables are the "true" ones, surrounded by a high number of irrelevant others. In this case, sparse regression methods are more appropriate because they are able to exactly set the coefficients associated with irrelevant variables to zero. Section 3.4.1.2 briefly presents the most popular of the sparse regression methods: the LASSO.

3.4.1.2 LASSO regression

The LASSO (Tibshirani, 1996) is a shrinkage regression method like ridge, where the L_2–penalty $\|\boldsymbol{\beta}\|_2^2$ is replaced with the L_1–penalty $\|\boldsymbol{\beta}\|_1$. The LASSO estimate is then defined by

$$\widehat{\boldsymbol{\beta}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{3.12}$$

where $\lambda \geqslant 0$ is still a tuning parameter that controls the amount of shrinkage. Just like ridge, we can also write the LASSO problem (3.12) as

$$\widehat{\boldsymbol{\beta}}_{\text{LASSO}} = \begin{cases} \arg\min_{\boldsymbol{\beta}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \\ \text{subject to } \ \|\boldsymbol{\beta}\|_1 \leqslant R_\lambda, \end{cases} \tag{3.13}$$

where there is a one–to–one correspondence between the shrinkage parameter λ in (3.12) and the radius R_λ in (3.13). The fundamental difference with ridge regression is that the L_1–penalty is intended to induce the sparsity of the solution. Indeed, large enough λ (or equivalently small enough R_λ) will cause some coefficients to be exactly equal to zero. LASSO is therefore widely used for variable selection purposes.

Unfortunately, unlike ridge regression, $\widehat{\boldsymbol{\beta}}_{\text{LASSO}}$ has no closed form expression due to the L_1–penalty. This explains why plethora of algorithms have been developed to implement the LASSO. Among them, the two most popular are certainly the least–angle regression (Efron et al., 2004), and the cyclic coordinate descent. The latter have been proposed a number of times for the LASSO but has only been popularised by Friedman et al. (2007) and Wu et al. (2008). Now, it seems that the cyclic coordinate descent generates consensus, thanks to its simplicity of implementation and its easy generalisation to GLMs. We will come back to this algorithm in Sections 3.4.1.3 and 3.4.2.

In short, since all coefficients are shrinked towards zero, LASSO reduces their variance, as does ridge. The main advantage of LASSO is that it is designed to eliminate nuisance variables in the model by estimating their coefficients as zeros. However, LASSO has two limitations. First, it will fail to select a whole set of highly correlated explanatory variables. Indeed, if explanatory variables are highly correlated with each other, the LASSO tends to choose only one and ignore the others. In addition, if $p \gg n$, the LASSO selects at most n explanatory variables: the number of selected variables is bounded by the number of observations. These two limitations are particularly problematic for genetic applications. The elastic net regularisation proposes to overcome them.

3.4.1.3 Elastic net regression

The elastic net method (Zou and Hastie, 2005) is a regularisation technique that combines L_1 and L_2 penalties. Thus, this method is a trade–off between LASSO and ridge regressions. The elastic net estimator is defined by

$$\widehat{\boldsymbol{\beta}}_{\text{EN}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \left[\alpha \|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2 \right] \right\}, \tag{3.14}$$

where $\lambda \geqslant 0$ is still a tuning parameter that controls the amount of shrinkage, and $\alpha \in [0, 1]$ is a parameter that tunes the trade–off between the ridge penalty ($\alpha = 0$) and the LASSO penalty ($\alpha = 1$).

To solve (3.14), the cyclic coordinate decent (Tseng, 2001; Friedman et al., 2007) optimises each parameter β_j separately holding all the others fixed, and cycle around the coefficients until they stabilise. Suppose we have estimates $\widetilde{\beta}_k$ for $k \neq j$, and we wish to partially optimise with respect to β_j. The update for β_j then writes

$$\widetilde{\beta}_j \longleftarrow \frac{\text{sT}\Big(\sum_{i=1}^{n} x_{ij} \widetilde{r}_i^{(-j)}, \lambda\alpha \Big)}{1 + \lambda(1-\alpha)}, \tag{3.15}$$

where

- $\widetilde{r}_i^{(-j)}$ is the partial residual that ignores the contribution from x_{ij}, namely

$$\widetilde{r}_i^{(-j)} = y_i - \widetilde{y}_i^{(-j)} = y_i - \sum_{k \neq j} x_{ik} \widetilde{\beta}_k,$$

- and sT is the soft–thresholding operator defined as

$$\text{sT}(z, t) = \text{sgn}(z)\big(|z| - t\big)_+ = \begin{cases} z - t & \text{if } z > 0 \text{ and } t < |z| \\ z + t & \text{if } z < 0 \text{ and } t < |z| \\ 0 & \text{if } t \geqslant |z|. \end{cases}$$

Hence, the elastic net is a powerful continuum between ridge and LASSO regressions, easy to implement via the cyclic coordinate descent procedure. The L_1 penalty generates a sparse model, while the L_2 penalty, initially introduced to overcome the limitations of LASSO described at the end of Section 3.4.1.2, stabilises the L_1 regularisation. While providing a variable selection and continuous shrinkage, the elastic net is also able to select a whole group of correlated explanatory variables instead of selecting a single one. Section 3.4.2 gives its extension to GLMs.

3.4.2 Elastic net for GLMs

All the regularisation methods presented so far take the form of penalised least squares. As a result, they correspond to the maximisation of a penalised log–likelihood of a Gaussian model. Friedman et al. (2010) derive the extension of elastic net regression to the more general framework of GLMs. The idea is to integrate the cyclic coordinate descent procedure within the IRLS algorithm. The update for β_j (3.15) becomes

$$\widetilde{\beta}_j \longleftarrow \frac{\mathrm{sT}\Big(\sum_{i=1}^{n} w_i x_{ij} \widetilde{r}_i^{(-j)}, \lambda\alpha\Big)}{\sum_{i=1}^{n} w_i x_{ij}^2 + \lambda(1-\alpha)}, \tag{3.16}$$

where w_i is the weight of the i^{th} observation inherited from the IRLS algorithm. Note that the partial residual $\widetilde{r}_i^{(-j)}$ in (3.16) involves this time the working variable z_i, also inherited from the IRLS algorithm:

$$\widetilde{r}_i^{(-j)} = z_i - \sum_{k \neq j} x_{ik} \widetilde{\beta}_k.$$

The elastic net regression for GLMs is available in the R package **glmnet**. For classification problems in particular, it has proven to reduce prediction errors compared to competing methods, with lower computational costs.

The literature on penalty–based regularisation methods is very extensive and it would be unrealistic to expect a comprehensive review of it. Besides, the framework we want to focus on is that of many explanatory variables that can be highly correlated, seen as proxies to latent phenomena to be found and interpreted. In this context, variable selection is inappropriate and ridge suffers from a lack of interpretability. It is then necessary to turn to component–based approaches.

3.5 Component–based regularisation methods

Component–based approaches assume that the information contained in the explanatory variables $\boldsymbol{X}$ can be summarised into a much lower dimensional space. The idea is to produce a small number of components (i.e. linear combinations $\{\boldsymbol{f_1}, \ldots, \boldsymbol{f_K}\}$ of the original explanatory variables $\boldsymbol{x_j}$'s), and to use the $\boldsymbol{f_k}$'s instead of the $\boldsymbol{x_j}$'s as inputs in the regression. The different approaches only differ in how the components are constructed. Section 3.5.1 provides a quick reminder on the Principal Components and Partial Least Squares

Regressions (PCR/PLSR), and Section 3.5.2 presents the extensions of the PLSR for GLMs.

3.5.1 Principal Components and Partial Least Squares Regressions

We reconsider the simple linear model $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$, such as defined at the beginning of Section 3.4.1. We still assume that the response is centred and that each explanatory variable is normalised.

3.5.1.1 Principal Components Regression

In this approach (see Jolliffe, 1982), the linear combinations $\boldsymbol{f_k}$'s used are the principal components. Suppose that the first K principal components are constructed, for some $K < p$. Since they are orthogonal, the PCR provides a predictor of the form

$$\widehat{\boldsymbol{y}}_{\text{PCR}} = \sum_{k=1}^{K} \widehat{\theta}_k \boldsymbol{f_k}, \tag{3.17}$$

where $\widehat{\theta}_k$ is the coefficient of the classical regression of $\boldsymbol{y}$ on $\boldsymbol{f_k}$, namely:

$$\widehat{\theta}_k = \frac{\langle\, \boldsymbol{y} \,|\, \boldsymbol{f_k} \,\rangle}{\langle\, \boldsymbol{f_k} \,|\, \boldsymbol{f_k} \,\rangle} = \frac{\langle\, \boldsymbol{y} \,|\, \boldsymbol{f_k} \,\rangle}{\|\boldsymbol{f_k}\|_2^2}. \tag{3.18}$$

The first principal component $\boldsymbol{f_1} = \boldsymbol{X}\boldsymbol{u_1}$ is designed to capture as much of the variability in the data as possible. It is then the solution of the maximisation program

$$\begin{aligned} \max_{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \operatorname{Var}(\boldsymbol{X}\boldsymbol{u}) \quad &\Longleftrightarrow \quad \max_{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \|\boldsymbol{X}\boldsymbol{u}\|_2^2 \quad \Longleftrightarrow \quad \max_{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \langle\, \boldsymbol{X}\boldsymbol{u} \,|\, \boldsymbol{X}\boldsymbol{u} \,\rangle \\ &\Longleftrightarrow \quad \max_{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u}. \end{aligned}$$

Each successive component in turn has the highest variance under the constraint that it is orthogonal to the preceding components. Each component $\boldsymbol{f_k}$ is a linear combination of the original explanatory variables: we indeed have $\boldsymbol{f_k} = \boldsymbol{X}\boldsymbol{u_k}$, where $\boldsymbol{u_k}$ is called the loading–vector associated with $\boldsymbol{f_k}$. Predictions (3.17) can then be expressed in terms of coefficients of the original explanatory variables:

$$\widehat{\boldsymbol{y}}_{\text{PCR}} = \sum_{k=1}^{K} \widehat{\theta}_k \boldsymbol{X}\boldsymbol{u_k} = \boldsymbol{X} \underbrace{\sum_{k=1}^{K} \widehat{\theta}_k \boldsymbol{u_k}}_{\widehat{\beta}_{\text{PCR}}}.$$

The main disadvantage of PCR is that the principal components do not take into account the response variable $\boldsymbol{y}$. In order to favour the directions of interest for modelling $\boldsymbol{y}$, regression on Partial Least Squares (PLS) components is preferable.

3.5.1.2 Partial Least Squares Regression

The PLS regression was first introduced by Wold (1966), and then had a great success, particularly in the field of chemometrics (Wold et al., 1983, 2001). Like PCR, the PLS regression constructs new components $\left\{ \boldsymbol{f}_k = \boldsymbol{X}\boldsymbol{u}_k \mid k = 1, \ldots, K \right\}$ as linear combinations of the original explanatory variables. But unlike PCR, the loading–vector $\boldsymbol{u}_k$ maximises the covariance between the component $\boldsymbol{f}_k$ and the response $\boldsymbol{y}$. The underlying maximisation program also assumes that each $\boldsymbol{u}_k$ have unit norm and that each $\boldsymbol{f}_k$ is orthogonal to the preceding components:

$$\boldsymbol{u}_k = \begin{cases} \underset{\boldsymbol{u}}{\arg\max} \ \operatorname{Cov}(\boldsymbol{X}\boldsymbol{u}, \boldsymbol{y}) \\ \|\boldsymbol{u}\|_2 = 1 \\ \boldsymbol{X}\boldsymbol{u} \perp \boldsymbol{f}_1, \ldots, \boldsymbol{f}_{k-1} \end{cases} = \begin{cases} \underset{\boldsymbol{u}}{\arg\max} \ \langle \boldsymbol{X}\boldsymbol{u} \mid \boldsymbol{y} \rangle \\ \|\boldsymbol{u}\|_2 = 1 \\ \boldsymbol{X}\boldsymbol{u} \perp \boldsymbol{f}_1, \ldots, \boldsymbol{f}_{k-1}. \end{cases} \tag{3.19}$$

A simple way to solve maximisation programs (3.19), is given in Algorithm 3.2. The orthogonality constraint of the components is ensured by deflating design matrix $\boldsymbol{X}$ at each step of the algorithm.

Algorithm 3.2: The univariate PLS (PLS1).

Input: response vector $\boldsymbol{y}$, design matrix $\boldsymbol{X}$, number of components K.

Set $\boldsymbol{X}_0 = \boldsymbol{X}$

for $h = 1$ *to* K **do**

$\quad \boldsymbol{u}_h = \dfrac{\boldsymbol{X}_{h-1}^\top \boldsymbol{y}}{\left\| \boldsymbol{X}_{h-1}^\top \boldsymbol{y} \right\|}$ `// Computing the loading-vector`

$\quad \boldsymbol{f}_h = \boldsymbol{X}_{h-1}\boldsymbol{u}_h$ `// Computing the component`

$\quad \boldsymbol{p}_h = \dfrac{\boldsymbol{X}_{h-1}^\top \boldsymbol{f}_h}{\boldsymbol{f}_h^\top \boldsymbol{f}_h}$ `// Regression coefficient of` $\boldsymbol{X}_{h-1}$ `on` $\boldsymbol{f}_h$

$\quad \boldsymbol{X}_h = \boldsymbol{X}_{h-1} - \boldsymbol{f}_h \boldsymbol{p}_h^\top$ `// Deflation of matrix` $\boldsymbol{X}_{h-1}$

end

The components constructed in Algorithm 3.2 can be expressed using the original explanatory variables. As for PCR, there is a vector of coefficients $\widehat{\boldsymbol{\beta}}_{\text{PLSR}}$, which can be recovered from the sequence of PLS transformations, such

that $\widehat{\boldsymbol{y}}_{\text{PLSR}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{PLSR}}$. It is straightforward to show that the k^{th} PLS loading–vector $\boldsymbol{u}_k$ also solves

$$\begin{cases} \max\limits_{\boldsymbol{u}} \ \text{Corr}^2\left(\boldsymbol{X}\boldsymbol{u}, \boldsymbol{y}\right) \text{Var}\left(\boldsymbol{X}\boldsymbol{u}\right) \\ \|\boldsymbol{u}\|_2 = 1 \\ \boldsymbol{X}\boldsymbol{u} \ \perp \ \boldsymbol{f_1}, \ldots, \boldsymbol{f_{k-1}}, \end{cases}$$

meaning that the direction of the first loading–vector in PLS is a trade–off between that of the ordinary regression coefficient–vector and that of the loading–vector of the first principal component.

The PLSR algorithm has been extended to the multivariate case, i.e. with several response variables $\boldsymbol{y_1}, \ldots, \boldsymbol{y_q}$. This point will be developed in more details in Section 3.5.2.3. We will now briefly discuss the few extensions of PLSR to GLMs that have been proposed in the literature.

3.5.2 Extension to GLMs

3.5.2.1 PLS for Generalised Linear Regression (PLS–GLR)

Bastien et al. (2005) note that the PLSR of a quantitative response $\boldsymbol{y}$ on $\boldsymbol{X} = \left[\boldsymbol{x_1} \mid \ldots \mid \boldsymbol{x_p}\right]$ yields a rank–1 component

$$\boldsymbol{f_1} = \frac{1}{\sqrt{\sum\limits_{j=1}^{p} \text{Cov}\left(\boldsymbol{y}, \boldsymbol{x_j}\right)^2}} \sum_{j=1}^{p} \text{Cov}\left(\boldsymbol{y}, \boldsymbol{x_j}\right) \boldsymbol{x_j},$$

where the quantity $\text{Cov}\left(\boldsymbol{y}, \boldsymbol{x_j}\right)$ is nothing but the coefficient associated with the simple OLS regression of $\boldsymbol{y}$ on $\boldsymbol{x_j}$. Rank–2 component can be obtained likewise, after replacing each $\boldsymbol{x_j}$ with its OLS regression residuals on $\boldsymbol{f_1}$, and so on. The extension of this approach to GLMs is straightforward and consists in replacing the OLS regressions of $\boldsymbol{y}$ on each $\boldsymbol{x_j}$ alone by Generalised Linear Regressions (GLR). Algorithm 3.3 summarises the strategy.

This extension is very simple but the weighting of observations seems inconsistent, since the estimated weighting matrix associated with the GLR of $\boldsymbol{y}$ on the components is not correctly used by this method. That is why we prefer the method proposed by Marx (1996), summarised in Section 3.5.2.2.

Algorithm 3.3: The PLS–GLR.

Input: response vector $\boldsymbol{y}$, design matrix $\boldsymbol{X}$, number of components K.

for $h = 1$ *to* K **do**
  for $j = 1$ *to* p **do**
    Carry out the GLR of $\boldsymbol{y}$ on $\{\boldsymbol{f_1}, \ldots, \boldsymbol{f_{h-1}}, \boldsymbol{x_j}\}$.
    Let a_{hj} be the coefficient associated with $\boldsymbol{x_j}$.
  end
  Set $\boldsymbol{u_h} = \boldsymbol{a_h} / \|\boldsymbol{a_h}\|$, where $\boldsymbol{a_h} = (a_{h1}, a_{h2}, \ldots, a_{hp})^{\mathrm{T}}$
  Set $\boldsymbol{f_h} = \boldsymbol{X}\boldsymbol{u_h}$
  $\boldsymbol{X} \leftarrow$ residual matrix of the linear regression of $\boldsymbol{X}$ on $\{\boldsymbol{f_1}, \ldots, \boldsymbol{f_h}\}$
end

3.5.2.2 Iteratively Reweighted PLS (IRPLS)

The method developed by Marx (1996), IRPLS, is another way to extend the PLSR to GLMs, but unlike PLS–GLR, it is based on the IRLS scheme (see Algorithm 3.1). IRPLS can then be viewed as an IRLS in which the weighted least squares regression used to update parameter $\boldsymbol{\beta}$ is replaced by a weighted PLS regression. More precisely, let $\boldsymbol{z}^{[t]}$ and $\boldsymbol{W}^{[t]}$ respectively be the working variable and the weight matrix at the t^{th} iteration of the IRLS. Instead of the classical update for $\boldsymbol{\beta}$, namely

$$\boldsymbol{\beta}^{[t+1]} = \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}^{[t]} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}^{[t]} \boldsymbol{z}^{[t]},$$

Marx (1996) rather suggests to set

$$\boldsymbol{\beta}^{[t+1]} \leftarrow \mathrm{PLSR}_{\boldsymbol{W}^{[t]}}\left(\boldsymbol{z}^{[t]}, \boldsymbol{X}\right),$$

where $\mathrm{PLSR}_{\boldsymbol{W}^{[t]}}\left(\boldsymbol{z}^{[t]}, \boldsymbol{X}\right)$ refers to the PLS regression of $\boldsymbol{z}^{[t]}$ on $\boldsymbol{X}$, in which the observations are weighted by $\boldsymbol{W}^{[t]}$. IRPLS seems much more consistent than PLS–GLR since at each iteration, the weighting matrix deriving from the maximum likelihood estimation is taken into account in the PLS regression. The IRPLS procedure has inspired Bry et al. (2013), who have proposed to extend it to the multivariate case.

3.5.2.3 Supervised Component Generalised Linear Regression (SCGLR)

Bry et al. (2013) consider a multivariate GLM involving several response vectors $\boldsymbol{y_1}, \ldots, \boldsymbol{y_q}$. In their rank–1 component model, the authors suggest a linear predictor associated with response variable $\boldsymbol{y_k}$ of the form

$$\boldsymbol{\eta_k} = (\boldsymbol{X}\boldsymbol{u})\,\gamma_k,$$

where the component $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$ is common to all the $\boldsymbol{y_k}$'s while the regression parameter γ_k is specific to each response. At step t of the IRLS algorithm, each $\boldsymbol{y_k}$ leads to a specific working variable $\boldsymbol{z_k^{[t]}}$ and a specific weighting matrix $\boldsymbol{W_k^{[t]}}$. The usual weighted least squares is then replaced, in Marx's wake, by a multivariate PLSR (PLS2 regression). The first unit loading–vector $\boldsymbol{u}_1$ then solves

$$\begin{aligned} & \max_{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \quad \sum_{k=1}^{q} \left\langle \boldsymbol{z_k^{[t]}} \,\middle|\, \boldsymbol{X}\boldsymbol{u} \right\rangle^2_{\boldsymbol{W_k^{[t]}}} \\ \iff & \max_{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \quad \sum_{k=1}^{q} \left\| \boldsymbol{z_k^{[t]}} \right\|^2_{\boldsymbol{W_k^{[t]}}} \left\| \boldsymbol{X}\boldsymbol{u} \right\|^2_{\boldsymbol{W_k^{[t]}}} \cos^2_{\boldsymbol{W_k^{[t]}}} \left(\boldsymbol{z_k^{[t]}}, \boldsymbol{X}\boldsymbol{u} \right), \end{aligned} \tag{3.20}$$

and higher–rank components can be constructed either by deflation or by adding an extra orthogonality constraint to the previous components.

Bry et al. (2013) also propose the possibility of modifying the geometry of the maximisation program (3.20), leading to the notion of "Supervised Components" (SC). By introducing a tuning parameter $s \in [0, +\infty)$, they suggest an alternative program which writes

$$\max_{\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X})^{-s}\boldsymbol{u}=1} \quad \sum_{k=1}^{q} \left\langle \boldsymbol{z_k^{[t]}} \,\middle|\, \boldsymbol{X}\boldsymbol{u} \right\rangle^2_{\boldsymbol{W_k^{[t]}}}, \tag{3.21}$$

where $\boldsymbol{P}$ is a weighting matrix reflecting the importance given a priori to each unit. In (3.21), parameter s allows to fine–tune the attraction of the components towards the principal components of $\boldsymbol{X}$, and thus creates a continuum between PCR and PLSR. Whenever $\boldsymbol{X}$ is not of full column rank, it should be replaced by the set of its principal components associated with non-zero eigenvalues. SCGLR was then refined by Bry et al. (2014, 2016, 2018).

3.6 A few words about hybrid methods

Recently, some methodologies involving both penalty–based regularisation and component building have been developed. For instance, Fort and Lambert-Lacroix (2005) have proposed to sequentially combine PLS regression with a ridge–penalised estimation of a logistic model (Eilers et al., 2001). Their method is a sequence of two steps.

1. A ridge–penalised logistic regression of the binary response $\boldsymbol{y}$ is carried out on the explanatory variables $\boldsymbol{X}$. At the convergence, this yields a working variable $\boldsymbol{z}^{\infty}$ and a weighting matrix $\boldsymbol{W}^{\infty}$.

2. A PLSR of $\boldsymbol{z}^{\infty}$ on $\boldsymbol{X}$ is then carried out with respect to the weighting matrix $\boldsymbol{W}^{\infty}$, yielding explanatory components.

More recently, Durif et al. (2017) have extended this technique to explanatory variable selection by replacing the PLS step by an adaptive version of the sparse PLS proposed by Chun and Keleş (2010). However, although the methods developed by Fort and Lambert-Lacroix (2005) and Durif et al. (2017) greatly stabilise the estimation process, it would appear that they suffer from a drawback similar to that of PLS–GLR. Indeed, the sequential nature of their methods does not allow to properly take into account the estimation weights since they are never updated.

Mention can also be made of the recent work by Bouveyron et al. (2018), whose purpose is to build sparse PCA components in such a way that all components have the same sparsity structure. Such sparse principal components could be conveniently used for regressing the responses.

3.7 Discussion

As suggested in this chapter, the range of methods available for the regularisation of a GLM is very wide. However in this thesis, we want to focus on situations where the many explanatory variables involved are seen as proxies to one or more phenomena that must be recovered and interpreted. In this context, the component–based regularisation is undoubtedly the most appropriate method, because the components are intended to reconstruct the latent variables that underlie the proxies. The useful part of the information contained in the explanatory variables is thus summarised in a small number of dimensions, without using a variable selection strategy that makes no sense in our context. The decomposition of the linear predictor on these interpretable orthogonal components facilitates the interpretation of the model, through the facility of such visual diagnoses as component planes.

However, as recalled in Chapter 2, PCR and PLSR do not always focus on the most important directions in $\boldsymbol{X}$ that best explain $\boldsymbol{y}$ (see for instance Frank and Friedman, 1993). We will not attempt here to improve the PCR or the PLSR by adding an extra sparsity criterion, especially since such work has already been done many times. By contrast, our work will consist in considering a more flexible version of the PCR/PLSR–like methods, which avoids the pitfall described in Chapter 2, and extending it to the broader framework of mixed models. Following the works by Marx (1996) and Bry et al. (2013), we will focus on regularisation methods based on the construction of *Supervised Components*.

IV

Classical GLMM estimation methods

Contents

4.1 Introduction

All the regularisation methods presented so far in Chapter 3 are based on a strong assumption: the independence of observations. However, many applications (e.g in biology, epidemiology, social science and economy) often have to deal with panel data, for which we must take account of the dependence induced by repeatedly measuring an outcome on each individual over time, or grouped data with an even more complex dependency structure. In such situations, as the independence assumption of observations is no longer valid, the introduction of random effects in models is widespread. Hence the need to consider Generalised Linear Mixed Models (GLMMs). For the most general distribution assumptions in such models, parameter estimation faces the intractability of the likelihood expressed as an integral with respect to the random effects. Specific approximate methods must thus be implemented.

The chapter is organised as follows. After a reminder of the random effect concept, we provide a description of GLMMs in Section 4.2. Then, a brief state–of–the–art of approximate methods for GLMM parameter estimation is presented in Sections 4.3 to 4.5. Finally, we discuss the use of the different methods in Section 4.6.

4.2 Definition, assumptions and notations

Random effect. Statistical studies are mainly motivated by the detection of data variability sources and their quantification. With its ANalysis Of VAriance (ANOVA) model, Fisher was one of the forerunners in this field. His strategy was to partition the different sources of variation using predefined "fixed effect" factors, and thus to assess the significance of observed differences between averages of data subgroups. However, this model is limited: when a factor contains a very large number of levels (or even an infinite number of levels), the statistical sampling of the experiment is unable to visit all of them.

Incorporating random effects into the model is an answer to this pitfall. More specifically, mixed models are a more elaborate way of studying data variability by considering more diverse sources of variation. Two types of factors are considered.

- Fixed effect factors, with a finite number of levels that all occur in the data. These levels are considered interesting in themselves since the goal is to quantify the effect of each level on the variable of interest.

- Random effect factors, with a usually infinite number of levels, of which only a random sample occurs in the data. In this case, it does not matter how each level affects the outcome of the experiment, but the interest is rather in the variability generated by such a sampling of levels.

It is worth noting that the type of factor to consider depends on the context, the questions of interest, and how the data is gathered (refer for example to Kreft and De Leeuw, 1998, Section 1.3.3). Let us illustrate this remark with a toy example. Consider a district with three high schools and imagine that a parents' association commissions a study to determine the effect of each high school on obtaining the A level. In this case, the high schools considered are interesting *per se*. The high school factor will then be seen as a fixed effect factor. On the other hand, if the study is conducted on a national scale — in order to know to what extent the high school attended has an influence on obtaining the A level —, it will certainly not be possible (or be too expensive) to take into account all the high schools. Only a sample of them will be selected. Here, the specific effects of the selected high schools are therefore not of interest anymore, as they are seen as representatives of a much larger set of high schools. In such a context, the high school factor is considered as a random effect factor.

This very simple example tried to clarify the notions of fixed and random effect factors, knowing that in practice, modelling is often more complicated and may involve cross effects. However, determining the nature (fixed or random) of an effect is not always easy, perhaps because conflicting definitions of "fixed effect" and "random effect" can be found in the literature (Gelman et al., 2005). That is why, to avoid any ambiguity, we will adopt the conventional rule proposed by Searle et al. (2009), Section 1.4:

> « [...] the important question is that of inference: are the levels of the factor going to be considered a random sample from a population of values?
> "Yes" — then the effects are to be considered as random effects.
> "No" — then, presumably, inferences will be made just about the levels occurring in the data and the effects are considered as fixed effects.»

That being said, the introduction of random effects in modelling allows us to be more precise on the origin of total variation compared to statistical modelling without random effects. Indeed, this variation is divided into two parts: that due to random effects and that due to errors. Mixed models are ultimately used to estimate fixed effects, but also to identify and quantify the different sources of variation, notably through the estimation of variance components.

Another important aspect of random effects is that they can be seen as a way of capturing some dependence between observations. In the case of repeated measures on several individuals for example, two observations coming

from the same individual are more likely to be similar to each other than two observations coming from two different individuals. In other words, observations from the same individual are generally correlated and non–independent. In such a situation, including an "individual" random effect in the model links the observations associated with the same realisation of the random effect, which will induce a particular covariance structure between these observations. Overall, in addition to providing fixed–effect estimates, mixed models take into account possible complex dependence structures between statistical units.

Notations for random effects. Throughout this work, the vector of random effects will be denoted $\boldsymbol{\xi}$. Generally speaking, $\boldsymbol{\xi}$ is a q–dimensional vector composed of Q subvectors so that $\boldsymbol{\xi} = \left(\boldsymbol{\xi}_1^{\mathsf{T}}, \ldots, \boldsymbol{\xi}_Q^{\mathsf{T}}\right)^{\mathsf{T}}$, where Q is the number of random effects involved in the model. Each $\boldsymbol{\xi_j}$ is a q_j–dimensional random effect vector, where q_j is the number of realisations of the j^{th} random effect observed in the data. So we have $\sum_{j=1}^{Q} q_j = q$, with q the total number of random levels.

The known design matrix of random effects will be denoted $\boldsymbol{U}$. With n the number of statistical units, we can write $\boldsymbol{U} = [\boldsymbol{U_1} \,|\, \ldots \,|\, \boldsymbol{U_Q}]$, where $\boldsymbol{U_j}$ is the $(n \times q_j)$–design matrix associated with the j^{th} random effect.

GLMM assumptions. GLMMs are based on assumptions similar to those for GLMs, except that they involve reasoning conditional on random effects. For a complete description of GLMMs, we refer the reader to McCulloch and Searle (2004). We will keep here the same notations for the vector of fixed–effect parameters $\boldsymbol{\beta}$ and its associated design matrix $\boldsymbol{X}$, both introduced in Chapter 3. We just recall that $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathsf{T}}$ is the observed response, which is a realisation of random vector $Y = (Y_1, \ldots, Y_n)^{\mathsf{T}}$. The following three assumptions are made.

($\boldsymbol{\mathcal{H}_1'}$) Conditional on random effects, the Y_i's are independent and have a distribution belonging to the exponential family. With the notations defined in Section 3.2, the conditional density function of Y_i given $\boldsymbol{\xi}$ can then be written as

$$p\left(y_i|\boldsymbol{\xi}\,;\theta_i^{\xi}\right) = \exp\left\{\frac{y_i\theta_i^{\xi} - b\left(\theta_i^{\xi}\right)}{a_i(\phi)} + c(y_i,\phi)\right\}.$$

Note that there is a link between the conditional expectation and the conditional variance of vector Y since we have

$$\mathbb{V}\left(Y|\boldsymbol{\xi}\right) = \mathbf{Diag}\Big(a_i\left(\phi\right)v\left(\mu_i^{\xi}\right)\Big)_{i=1,\ldots,n},$$

where μ_i^{ξ} is the i^{th} element of vector $\boldsymbol{\mu^{\xi}} := \mathbb{E}\left(Y|\boldsymbol{\xi}\right)$ and $v = b'' \circ b'^{-1}$.

($\mathcal{H}_2'$) Unlike GLMs, linear predictor $\boldsymbol{\eta}^{\boldsymbol{\xi}}$ of a GLMM contains a random part. It is expressed as a combination of the fixed and random effects:

$$\boldsymbol{\eta}^{\boldsymbol{\xi}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\xi}.$$

($\mathcal{H}_3'$) Link function $g(\cdot)$ relates random vector Y to the linear predictor through

$$\boldsymbol{\eta}^{\boldsymbol{\xi}} = g\left(\boldsymbol{\mu}^{\boldsymbol{\xi}}\right).$$

As with GLMs, link $g(\cdot)$ must be strictly monotonic and twice–differentiable.

Likelihood of GLMM. Even if some authors relax the assumption that random effects are normally distributed (Lee and Nelder (1996) indeed suggest other distributions such as gamma, inverse-gamma or beta), we keep here the most commonly assumed Gaussian distribution:

$$\forall j \in \{1, \ldots, Q\}, \quad \boldsymbol{\xi_j} \sim \mathcal{N}_{q_j}(\mathbf{0}, \boldsymbol{D_j}).$$

Since the present work cares about modelling variance–components, we further assume

$$\boldsymbol{D_j} = \sigma_j^2 \boldsymbol{G_j},$$

where $\boldsymbol{G_j}$ is a known $(q_j \times q_j)$–matrix and σ_j^2 the variance component relative to the j^{th} random effect, which has to be estimated. Subvectors $\boldsymbol{\xi_1}, \ldots, \boldsymbol{\xi_Q}$ being assumed independent, the distribution of random–effect vector $\boldsymbol{\xi}$ writes

$$\boldsymbol{\xi} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{D_\sigma}),$$

where variance–covariance matrix $\boldsymbol{D_\sigma}$ is a block diagonal matrix:

$$\boldsymbol{D_\sigma} = \mathbf{bDiag}\left(\boldsymbol{D_j}\right)_{j=1,\ldots,Q} = \mathbf{bDiag}\left(\sigma_j^2 \boldsymbol{G_j}\right)_{j=1,\ldots,Q}.$$

Note that the variance–covariance matrix is indexed by $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_Q)$ to stress the fact that it is parameterised by variance components only.

It is emphasised that a GLMM is correctly defined only conditional on random effects. More precisely, since we only know the response distribution conditional on random effects, the likelihood function of the parameters is expressed through the integral

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}) = \int_{\mathbb{R}^q} \prod_{i=1}^{n} p\left(y_i | \boldsymbol{\xi}; \boldsymbol{\beta}, \boldsymbol{\sigma}\right) p\left(\boldsymbol{\xi}; \boldsymbol{\sigma}\right) \mathrm{d}\boldsymbol{\xi}, \tag{4.1}$$

where $p(\boldsymbol{\xi}; \boldsymbol{\sigma})$ denotes the density function of random vector $\boldsymbol{\xi}$. Choosing the canonical link provides a more explicit form for the likelihood function:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}) = \int_{\mathbb{R}^q} \exp\left\{ \frac{\boldsymbol{y}^\intercal \boldsymbol{\eta}^{\boldsymbol{\xi}} - \mathbf{1}^\intercal b\left(\boldsymbol{\eta}^{\boldsymbol{\xi}}\right)}{a(\phi)} + \mathbf{1}^\intercal c(\boldsymbol{y}, \phi) \right\} \times \frac{\exp\left(-\frac{1}{2}\boldsymbol{\xi}^\intercal \boldsymbol{D}_{\boldsymbol{\sigma}}^{-1} \boldsymbol{\xi}\right)}{(2\pi)^{q/2} \left|\boldsymbol{D}_{\boldsymbol{\sigma}}\right|^{1/2}} \mathrm{d}\boldsymbol{\xi}. \tag{4.2}$$

However, except for a few special models such as Linear Mixed Models (LMMs, or Gaussian–identity), beta–binomial or Poisson–gamma (see for instance Lee et al. (2017) or Molenberghs et al. (2010)), Y does not have a closed–form marginal distribution. Therefore, as pointed out for instance by Trottier (1998), under the most general assumptions, there is no single reference method for estimating fixed effects and variance components of a GLMM. To deal with the integral with respect to the random effects' distribution in (4.1) or (4.2), three approaches can be mentioned:

- numerical integrations (quadrature) and analytic approximations of the likelihood,
- Monte Carlo (MC) integrations and indirect MC–based likelihood maximisations,
- linearisation methods.

As there does not appear to be a consensus on any of the approaches, Sections 4.3 to 4.5 are devoted to a brief review of them.

4.3 Numerical approximations

There are essentially two types of numerical approximations in our context: numerical integrations (Section 4.3.1) and methods based on an analytic approximation of the integrand (Sections 4.3.2 and 4.3.3).

4.3.1 Gauss–Hermite quadrature and adaptive version

Quadrature is based on the fact that an integral can be interpreted as an infinite weighted sum. It then approximates an integral by a finite weighted sum of the integrand evaluated at specified points within the domain of integration. The best choices of weights and points essentially depend on the type of integrand. The two most common types of quadrature used for GLMMs are Gauss–Hermite quadrature (GHQ) and an adaptive version of it (AGHQ).

The unidimensional case. With $t \in \mathbb{R}$, GHQ rules (Liu and Pierce, 1993, 1994) are designed to evaluate integrals of the form

$$I = \int_{\mathbb{R}} \exp\left(-t^2\right) f(t)\, \mathrm{d}t,$$

where $f(\cdot)$ is a sufficiently regular function to be well approximated by a polynomial. The approximations provided take the form

$$I \simeq \widehat{I}_M = \sum_{m=1}^{M} v_m\, f(t_m), \tag{4.3}$$

where:

- M is the number of quadrature points,
- $t_1, \ldots, t_M$ are the quadrature points (abscissas), defined as the M roots of the M^{th} Hermite polynomial which writes

$$H_M(x) = (-1)^M \exp\left(\frac{x^2}{2}\right)\left(\frac{\mathrm{d}^M}{\mathrm{d}x^M} \exp\left(\frac{-x^2}{2}\right)\right),$$

- $v_1, \ldots, v_M$ are the associated weights given by

$$v_m = \frac{2^{M-1} M! \sqrt{\pi}}{M^2 \left(H_{M-1}(t_m)\right)^2}.$$

For $M \leqslant 20$, the quadrature points and corresponding weights can be found in Abramowitz and Stegun (1964), while the algorithm described in Golub and Welsch (1969) allows to calculate them when $M > 20$. The accuracy of the method depends on the number of quadrature points: the higher M, the better the approximation of the integral. With this in mind, R package **fastGHQuad** presented in Blocker et al. (2014) allows a fast and numerically–stable computation of quadrature rules with more than 1000 points.

Unidimensional Gaussian density. In many applications, the following integral must be calculated:

$$J = \int_{\mathbb{R}} \varphi_{\mu,\sigma}(t)\, f(t)\, \mathrm{d}t,$$

where $\varphi_{\mu,\sigma}(\cdot)$ refers to the density function of the Gaussian distribution with expectation μ and standard deviation σ. As shown in Naylor and Smith (1982), the GHQ–approximation of J is given by

$$\widehat{J}_M = \sum_{m=1}^{M} v_m^*\, f(t_m^*), \tag{4.4}$$

where $t_m^* = \mu + \sigma\sqrt{2}\,t_m$ and $v_m^* = \dfrac{v_m}{\sqrt{\pi}}$. The quadrature points defined in (4.3) are therefore linearly transformed according to the values of μ and σ, and the initial weights are simply divided by $\sqrt{\pi}$. In (4.4), weights v_m^*'s and quadrature points t_m^*'s are said to be based on weight function $\varphi_{\mu,\sigma}(\cdot)$.

Application to GLMMs. For illustration, suppose that we have a simple GLMM with only one q–dimensional random effect $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_q)^{\mathrm{T}}$ such that $\boldsymbol{\xi} \sim \mathcal{N}_q\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_q\right)$. In this simplified case, the likelihood function writes

$$\mathcal{L}\left(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}\right) = \prod_{i=1}^{q} \prod_{j=1}^{n_i} \int_{\mathbb{R}} \varphi_{0,\sigma}\left(\xi_i\right) p\left(y_{ij}|\xi_i\right) \, \mathrm{d}\xi_i = \prod_{i=1}^{q} \underbrace{\int_{\mathbb{R}} \varphi_{0,\sigma}\left(\xi_i\right) \prod_{j=1}^{n_i} p\left(y_{ij}|\xi_i\right) \, \mathrm{d}\xi_i}_{K_i},$$

where n_i is the number of observations associated with the i^{th} realisation of the random effect. A variable change within the integral to a standard Gaussian variable (by defining $\widetilde{\xi}_i = \xi_i/\sigma$), leads to another expression of the likelihood contribution from the i^{th} cluster:

$$K_i = \int_{\mathbb{R}} \varphi_{0,1}\left(\widetilde{\xi}_i\right) \prod_{j=1}^{n_i} p\left(y_{ij}|\sigma\widetilde{\xi}_i\right) \, \mathrm{d}\widetilde{\xi}_i.$$

Applying the GHQ described in (4.4) provides

$$\mathcal{L}\left(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}\right) \simeq \prod_{i=1}^{q} \left\{ \sum_{m=1}^{M} v_m^* \prod_{j=1}^{n_i} p(y_{ij}|\sigma t_m^*) \right\}, \tag{4.5}$$

where $t_m^* = \sqrt{2}\,t_m$ and $v_m^* = \dfrac{v_m}{\sqrt{\pi}}$. Anderson and Aitkin (1985) were probably the first to apply the GHQ to evaluate and maximise the likelihood function in a logistic regression model with one random effect. Unfortunately, even for a large number of quadrature points, it may not be possible to approximate the GLMM likelihood accurately with GHQ, the approximation errors being amplified by large clusters and high random effects variances. This is due to the fact that in the GHQ, the quadrature points and weights are not adjusted to the shape of the integrand.

AGHQ for GLMMs. The main idea of Adaptive Gauss–Hermite quadrature (AGHQ) — see Pinheiro and Bates (1995) or Pinheiro and Chao (2006) for instance — is to shift and rescale the quadrature points to lie under the peak of the function to be integrated. To this end, instead of treating the prior density $\varphi_{0,1}\left(\widetilde{\xi}_i\right)$ as the weight function of the GHQ (as it was done in (4.5)), the

proposal is to rewrite K_i as

$$K_i = \int_{\mathbb{R}} \varphi_{\nu_i,\tau_i}\left(\widetilde{\xi}_i\right) \frac{\varphi_{0,1}\left(\widetilde{\xi}_i\right) \prod_{j=1}^{n_i} p\left(y_{ij}|\sigma\widetilde{\xi}_i\right)}{\varphi_{\nu_i,\tau_i}\left(\widetilde{\xi}_i\right)} \mathrm{d}\widetilde{\xi}_i,$$

where φ_{ν_i,τ_i} is a Gaussian density which approximates the posterior density of ξ_i given the observed data. Treating this density as the weight function leads to the definition of new quadrature points and associated weights

$$\begin{cases} t_{im}^{\mathrm{ad}} = \nu_i + \tau_i t_m^* \\ v_{im}^{\mathrm{ad}} = \tau_i \sqrt{2\pi} \exp\left(\frac{1}{2} t_m^{*2}\right) \varphi_{0,1}\left(t_{im}^{\mathrm{ad}}\right) v_m^* \end{cases}$$

and to a more accurate likelihood approximation:

$$\mathcal{L}\left(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}\right) \simeq \prod_{i=1}^{q} \left\{ \sum_{m=1}^{M} v_{im}^{\mathrm{ad}} \prod_{j=1}^{n_i} p(y_{ij}|\sigma t_{im}^{\mathrm{ad}}) \right\}. \tag{4.6}$$

Even if the AGHQ generally outperforms the GHQ with much fewer quadrature points, it is much more computationally intensive. Indeed, the AGHQ requires the computation of means ν_i's and variances τ_i^2's of the Gaussian density which approximates the posterior, resulting in time–consuming iterative calculations.

There are some extensions of the GHQ and the AGHQ to higher dimensional random effects, essentially based on the Cartesian product quadrature. Since it seems unnecessary to give a full description of these extensions here, we simply refer the reader to Pan and Thompson (2003) and Rabe-Hesketh and Skrondal (2004) for further details.

The main advantage of the AGHQ is that it can be made arbitrarily accurate by increasing the number of quadrature points. The price to pay will be computation time, higher and higher with the desired level of accuracy. The main drawback is that as soon as the likelihood is not factorisable into low–dimensional integrals, the numerical Gaussian quadrature may not be applicable. It is the case, for example, for correlated, crossed and nested random effects, greatly reducing the scope of the method.

4.3.2 Laplace approximation

Proposed by Tierney and Kadane (1986), the Laplace method is particularly adapted to approximate integrals of the form

$$I = \int_{\mathbb{R}^q} e^{S(\boldsymbol{u})} \, \mathrm{d}\boldsymbol{u}, \tag{4.7}$$

where $S(\cdot)$ is a regular function. The procedure starts with the expansion of $S(\boldsymbol{u})$ as a second–order Taylor series around its mode $\widetilde{\boldsymbol{u}}$. The fact that $\boldsymbol{\nabla} S(\widetilde{\boldsymbol{u}}) = \mathbf{0}$ provides

$$S(\boldsymbol{u}) \simeq S(\widetilde{\boldsymbol{u}}) + \frac{1}{2}(\boldsymbol{u} - \widetilde{\boldsymbol{u}})^{\intercal} \left[\mathbf{H} S(\widetilde{\boldsymbol{u}})\right](\boldsymbol{u} - \widetilde{\boldsymbol{u}}), \tag{4.8}$$

where $\boldsymbol{\nabla} S$ and $\mathbf{H} S$ refers to the gradient and the Hessian matrix of S respectively. Then, putting back (4.8) into (4.7) leads to the approximation

$$\begin{aligned} I \simeq \widehat{I}_{\text{La}} &= e^{S(\widetilde{\boldsymbol{u}})} \int_{\mathbb{R}^q} \exp\left\{\frac{1}{2}(\boldsymbol{u} - \widetilde{\boldsymbol{u}})^{\intercal} \left[\mathbf{H} S(\widetilde{\boldsymbol{u}})\right](\boldsymbol{u} - \widetilde{\boldsymbol{u}})\right\} \mathrm{d}\boldsymbol{u} \\ &= e^{S(\widetilde{\boldsymbol{u}})} \int_{\mathbb{R}^q} \exp\left\{-\frac{1}{2}(\boldsymbol{u} - \widetilde{\boldsymbol{u}})^{\intercal} \left[-\mathbf{H} S(\widetilde{\boldsymbol{u}})\right](\boldsymbol{u} - \widetilde{\boldsymbol{u}})\right\} \mathrm{d}\boldsymbol{u} \\ &= e^{S(\widetilde{\boldsymbol{u}})} (2\pi)^{q/2} \left|-\mathbf{H} S(\widetilde{\boldsymbol{u}})\right|^{-1/2}. \end{aligned}$$

Laplace approximation applied to GLMM estimation. Now, let $R(\boldsymbol{\xi}) := \log\left[p(\boldsymbol{y}|\boldsymbol{\xi})\, p(\boldsymbol{\xi})\right]$. The likelihood function of a GLMM (4.1) can be rewritten as

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}) = \int_{\mathbb{R}^q} e^{R(\boldsymbol{\xi})} \mathrm{d}\boldsymbol{\xi}.$$

Subsequently, let $\widetilde{\boldsymbol{\xi}}$ denote the conditional mode of random effects, namely $\widetilde{\boldsymbol{\xi}} = \arg\max_{\boldsymbol{\xi} \in \mathbb{R}^q} R(\boldsymbol{\xi})$. The Laplace approximation of the likelihood in then given by

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}) \simeq (2\pi)^{q/2} \left|-\mathbf{H} R\left(\widetilde{\boldsymbol{\xi}}\right)\right|^{-1/2} \exp\left[R\left(\widetilde{\boldsymbol{\xi}}\right)\right]. \tag{4.9}$$

Maximisation of (4.9) yields estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\sigma}}$ for $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$. But it should be stressed that unlike quadrature methods, the Laplace approximation can not be made arbitrarily accurate. As a result, the estimates obtained can not be made arbitrarily close to maximum likelihood estimates. That is why $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\sigma}}$ are often called "approximate maximum likelihood estimates" in the literature. GLMM estimation by both Laplace and AGHQ methods can be done using the R package (R Core Team, 2017) **lme4** (Bates et al., 2015).

Even if they are derived from an analytical approximation of the likelihood, the estimates obtained by maximising the Laplace approximation of the likelihood work reasonably well in many situations. However, they are prone to some bias towards zero when the conditional response distribution is highly non–normal and when the variances of random effects are large (McCulloch, 1997). That is why enhanced versions of the Laplace approximation have been developed, notably by Shun and McCullagh (1995) and Raudenbush et al. (2000).

Based on higher–order Taylor series expansion, they allow to reduce this bias and also to deal with high dimensional integrals.

Note that AGHQ is equivalent to Laplace approximation when only one quadrature point is used. AGHQ is therefore generally more accurate. However, Laplace approximation is still applicable for correlated, crossed and nested random effects, which could explain its prevalence. In Section 4.3.3, we present one of the most popular approximate maximum likelihood approaches related to the Laplace approximation: the Penalised Quasi–likelihood (PQL).

4.3.3 Penalised Quasi–Likelihood

The Penalised Quasi–Likelihood (PQL) method presented by Breslow and Clayton (1993) is based on a Laplace approximation (as presented in Section 4.3.2) of the marginal quasi–likelihood function defined by

$$Q\left(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}\right) = \text{cte} \times \left|\boldsymbol{D}_{\boldsymbol{\sigma}}\right|^{-\frac{1}{2}} \int_{\mathbb{R}^q} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} d_i\left(y_i, \mu_i^{\xi}\right) - \frac{1}{2}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{\xi}\right\} \mathrm{d}\boldsymbol{\xi}, \tag{4.10}$$

which is obtained by integrating the conditional quasi–likelihood of Y given $\boldsymbol{\xi}$ with respect to the distribution of $\boldsymbol{\xi}$. In (4.10), $d_i\left(y_i, \mu_i^{\xi}\right)$ is defined by

$$d_i\left(y_i, \mu_i^{\xi}\right) = -2\int_{y_i}^{\mu_i^{\xi}} \frac{y_i - t}{a_i(\phi)v(t)}\,\mathrm{d}t,$$

so that it can be interpreted as a deviance measure of fit, also referred to as "quasi–deviance". We can rewrite the marginal quasi–likelihood function Q as

$$Q\left(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}\right) = \text{cte} \times \left|\boldsymbol{D}_{\boldsymbol{\sigma}}\right|^{-\frac{1}{2}} \int_{\mathbb{R}^q} e^{-k(\boldsymbol{\xi})}\,\mathrm{d}\boldsymbol{\xi},$$

where

$$k(\boldsymbol{\xi}) = \frac{1}{2}\sum_{i=1}^{n} d_i\left(y_i, \mu_i^{\xi}\right) + \frac{1}{2}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{\xi}.$$

Then, a second–order Taylor series around $\widetilde{\boldsymbol{\xi}}$ such that $\boldsymbol{\nabla} k\left(\widetilde{\boldsymbol{\xi}}\right) = \mathbf{0}$, provides the following approximation:

$$Q\left(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}\right) \simeq \underbrace{\text{cte} \times (2\pi)^{q/2}}_{\text{constant terms}} \left|\boldsymbol{D}_{\boldsymbol{\sigma}}\right|^{-\frac{1}{2}} \left|\mathbf{H}k\left(\widetilde{\boldsymbol{\xi}}\right)\right|^{-1/2} e^{-k\left(\widetilde{\boldsymbol{\xi}}\right)}. \tag{4.11}$$

Taking the log and ignoring the multiplicative constant terms in (4.11), the approximation yields

$$q\left(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}\right) := \log Q\left(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}\right) \simeq -\frac{1}{2}\log\left|\boldsymbol{D}_{\boldsymbol{\sigma}}\right| - \frac{1}{2}\log\left|\mathbf{H}k\left(\widetilde{\boldsymbol{\xi}}\right)\right| - k\left(\widetilde{\boldsymbol{\xi}}\right). \quad (4.12)$$

Let us now recall that

$$k(\boldsymbol{\xi}) = -\sum_{i=1}^{n}\int_{y_i}^{\mu_i^{\xi}} \frac{y_i - t}{a_i(\phi)v(t)}\,\mathrm{d}t + \frac{1}{2}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{\xi}, \text{ with}$$
$$\mu_i^{\xi} = g^{-1}\left(\eta_i^{\xi}\right) = g^{-1}\left(\boldsymbol{x}_{i:}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{u}_{i:}^{\mathrm{T}}\boldsymbol{\xi}\right).$$

So, the gradient and the Hessian matrix of k respectively write

$$\begin{aligned}
\boldsymbol{\nabla}k(\boldsymbol{\xi}) &= -\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\xi}}\int_{y_i}^{\mu_i^{\xi}}\frac{y_i - t}{a_i(\phi)v(t)}\,\mathrm{d}t + \boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{\xi} \\
&= -\sum_{i=1}^{n}\frac{y_i - \mu_i^{\xi}}{a_i(\phi)v\left(\mu_i^{\xi}\right)}\frac{\partial g^{-1}\left(\boldsymbol{x}_{i:}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{u}_{i:}^{\mathrm{T}}\boldsymbol{\xi}\right)}{\partial\boldsymbol{\xi}} + \boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{\xi} \\
&= -\sum_{i=1}^{n}\frac{\left(y_i - \mu_i^{\xi}\right)\boldsymbol{u}_{i:}}{a_i(\phi)v\left(\mu_i^{\xi}\right)g'\left(\mu_i^{\xi}\right)} + \boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{\xi}, \\
\mathbf{H}k(\boldsymbol{\xi}) &= \sum_{i=1}^{n}\frac{\boldsymbol{u}_{i:}\boldsymbol{u}_{i:}^{\mathrm{T}}}{a_i(\phi)v\left(\mu_i^{\xi}\right)\left[g'\left(\mu_i^{\xi}\right)\right]^2} + \boldsymbol{D}_{\boldsymbol{\sigma}}^{-1} \\
&\quad -\sum_{i=1}^{n}\left(y_i - \mu_i^{\xi}\right)\boldsymbol{u}_{i:}\left[\frac{\partial}{\partial\boldsymbol{\xi}}\left(\frac{1}{a_i(\phi)v\left(\mu_i^{\xi}\right)g'\left(\mu_i^{\xi}\right)}\right)\right]^{\mathrm{T}}.
\end{aligned}$$

The authors then ignore the last term of $\mathbf{H}k(\boldsymbol{\xi})$, which has expectation $\mathbf{0}$ and equals $\mathbf{0}$ for canonical link functions. (4.12) then becomes

$$q\left(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}\right) \simeq -\frac{1}{2}\log\left(\left|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{W}^{\widetilde{\boldsymbol{\xi}}}\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{\sigma}} + \boldsymbol{I}_q\right|\right) - k\left(\widetilde{\boldsymbol{\xi}}\right), \quad (4.13)$$

where $\boldsymbol{W}^{\widetilde{\boldsymbol{\xi}}} = \mathbf{Diag}\left[\left(a_i(\phi)\,v\left(\mu_i^{\widetilde{\xi}}\right)\left[g'\left(\mu_i^{\widetilde{\xi}}\right)\right]^2\right)^{-1}\right]_{i=1,\ldots,n}$. Assuming that $\boldsymbol{W}^{\widetilde{\boldsymbol{\xi}}}$ varies slowly as a function of the parameters, the first term of (4.13) is ignored. Therefore, the maximisation of function q can be reduced to the joint–maximisation (i.e. with respect to $\boldsymbol{\beta}, \boldsymbol{\sigma}$ and $\boldsymbol{\xi}$) of the log–PQL function defined by Green (1987) as

$$\log \mathrm{PQL}\left(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}\right) = -k\left(\boldsymbol{\xi}\right) = -\frac{1}{2}\sum_{i=1}^{n} d_i\left(y_i, \mu_i^{\xi}\right) - \frac{1}{2}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{\xi}. \quad (4.14)$$

Note that (4.14) is known as the *penalised* log–quasi–likelihood function since a penalty term, namely $-\frac{1}{2}\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{\xi}$, has been added to the log–quasi–likelihood of a GLM defined by Wedderburn (1974) (see Section 3.3.2). Based on (4.14), the iterative process suggested for estimating fixed–effect parameters and variance components is the same as Schall's (see Section 4.5.1 for a in–depth description). This is why these two methods are sometimes considered as one and the same method.

As the PQL approach of Breslow and Clayton (1993) uses the Laplace approximation, its advantages and drawbacks are essentially those discussed at the end of Section 4.3.2. Its wide spectrum of applications explains why it is still the most widely used method for GLMM estimation, not to mention that the method is also quite computationally fast (Venables and Ripley (2002), R package **MASS**). However, since estimates from the PQL method can be biased in certain situations (e.g. for binary data or when the number of observations within each level of random effect is small), some bias correction techniques have been proposed. In a non–exhaustive manner, we refer the reader to Breslow and Lin (1995), Goldstein and Rasbash (1996), and Kuk (1995).

4.4 Stochastic approximations

Two categories of stochastic approximations can be considered in our context: indirect maximisation of the likelihood via MC–based methods (typically based on stochastic variants of the EM algorithm, see Section 4.4.1) and MC–based methods that approximate posterior distributions (Section 4.4.2) or the entire likelihood (Section 4.4.3).

4.4.1 Monte Carlo EM algorithm

The EM algorithm (see Dempster et al. (1977) and Chapter 6, Section 6.3 for a more detailed description) is a popular and often efficient approach for computing maximum likelihood estimates, in the context of missing data or latent variables. Considering random effects $\boldsymbol{\xi}$ as missing data, let ℓ^{c} denote the complete log–likelihood. Let also $\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\cdot \,|\, \boldsymbol{\theta}^{[t]}\right]$ be the expectation with respect to the conditional distribution of $\boldsymbol{\xi}$ given $\boldsymbol{y}$ at the current value $\boldsymbol{\theta}^{[t]}$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma})$ refers to the set of unknown parameters composed of fixed–effect parameters and variance components. The general formulation of the EM algorithm is recalled in Algorithm 4.1.

Algorithm 4.1: The EM algorithm.

Start with an initial guess $\boldsymbol{\theta}^{[0]}$ and set $t = 0$

while *some convergence criterion not reached* **do**

E–step: Compute $\mathcal{Q}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) = \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\ell^{\mathrm{c}}\left(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}\right) \,|\, \boldsymbol{\theta}^{[t]}\right]$

M–step: Set $\boldsymbol{\theta}^{[t+1]} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right)$

$t \leftarrow t + 1$

end

Unfortunately, since the conditional distribution of $\boldsymbol{\xi}$ given $\boldsymbol{y}$ is generally unknown in the GLMM context, alternative strategies have been developed to face the intractability of objective function $\mathcal{Q}$. One of them is the Monte Carlo EM algorithm, initially developed by Wei and Tanner (1990) and later refined by McCulloch (1994) and McCulloch (1997) in the GLMM context.

Principle of the method. The idea proposed by McCulloch (1997) is to simulate random draws from the conditional distribution of $\boldsymbol{\xi}$ given $\boldsymbol{y}$ by using a Metropolis–Hastings scheme (Hastings, 1970; Robert and Casella, 2013), and then approximate the $\mathcal{Q}$–function by Monte Carlo integration. Algorithm 4.2 summarises this strategy.

Algorithm 4.2: The MCEM algorithm.

Start with an initial guess $\boldsymbol{\theta}^{[0]}$ and set $t = 0$

while *some convergence criterion not reached* **do**

Metropolis–Hastings scheme: Generate M values, namely $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(M)}$, from the conditional distribution of $\boldsymbol{\xi}|\boldsymbol{y}$ knowing current parameter value $\boldsymbol{\theta}^{[t]}$.

E–step: Compute $\mathcal{Q}^{\mathrm{MC}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) = \dfrac{1}{M}\displaystyle\sum_{m=1}^{M} \ell^{\mathrm{c}}\left(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}^{(m)}\right)$

M–step: Set $\boldsymbol{\theta}^{[t+1]} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}^{\mathrm{MC}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right)$

$t \leftarrow t + 1$

end

Details on the Metropolis–Hastings algorithm involved. The aim is to obtain a sequence of random samples from target density $p\left(\boldsymbol{\xi}|\boldsymbol{y}\right)$. As an MCMC method, the Metropolis–Hastings algorithm produces an ergodic Markov chain $\left(\boldsymbol{\xi}^{(m)}\right)_m$ whose stationary distribution is $p\left(\boldsymbol{\xi}|\boldsymbol{y}\right)$. It requires the choice of a proposal distribution, $h\left(\boldsymbol{\xi}\right)$, from which potential new values are drawn

and an acceptance function, $A(\cdot, \cdot)$, which gives the probability of accepting the new value. Let $\boldsymbol{\xi}^{(m)}$ denote the m^{th} draw from $p(\boldsymbol{\xi}|\boldsymbol{y})$. The $(m+1)^{\text{th}}$ draw, namely $\boldsymbol{\xi}^{(m+1)}$, is then generated through the following transition kernel:

1. Generate a new value ξ_k^* for the k^{th} element of $\boldsymbol{\xi}$ as $\xi_k^* \sim h(\xi_k)$.
2. Take

$$\boldsymbol{\xi}^{(m+1)} = \begin{cases} \left(\xi_1^{(m)}, \ldots, \xi_{k-1}^{(m)}, \xi_k^*, \xi_{k+1}^{(m)}, \ldots, \xi_q^{(m)}\right)^{\intercal} & \text{with prob. } A\left(\boldsymbol{\xi}^{(m)}, \boldsymbol{\xi}^{(m+1)}\right) \\ \boldsymbol{\xi}^{(m)} & \text{with prob. } 1 - A\left(\boldsymbol{\xi}^{(m)}, \boldsymbol{\xi}^{(m+1)}\right) \end{cases},$$

where

$$A\left(\boldsymbol{\xi}^{(m)}, \boldsymbol{\xi}^{(m+1)}\right) = \min\left\{1, \frac{p\left(\boldsymbol{\xi}^{(m+1)} \,|\, \boldsymbol{y}\right) h\left(\boldsymbol{\xi}^{(m)}\right)}{p\left(\boldsymbol{\xi}^{(m)} \,|\, \boldsymbol{y}\right) h\left(\boldsymbol{\xi}^{(m+1)}\right)}\right\}. \tag{4.15}$$

Note that the ratio in (4.15) depends on the unknown conditional distribution of $\boldsymbol{\xi}$ given $\boldsymbol{y}$. McCulloch (1997) then suggests to choose the marginal distribution of the random effects, $p(\boldsymbol{\xi})$, as the proposal distribution. The ratio indeed simplifies to

$$\begin{aligned} \frac{p\left(\boldsymbol{\xi}^{(m+1)} \,|\, \boldsymbol{y}\right) p\left(\boldsymbol{\xi}^{(m)}\right)}{p\left(\boldsymbol{\xi}^{(m)} \,|\, \boldsymbol{y}\right) p\left(\boldsymbol{\xi}^{(m+1)}\right)} &= \frac{p\left(\boldsymbol{y} \,|\, \boldsymbol{\xi}^{(m+1)}\right) \cancel{p\left(\boldsymbol{\xi}^{(m+1)}\right)} \cancel{p\left(\boldsymbol{\xi}^{(m)}\right)}}{p\left(\boldsymbol{y} \,|\, \boldsymbol{\xi}^{(m)}\right) \cancel{p\left(\boldsymbol{\xi}^{(m)}\right)} \cancel{p\left(\boldsymbol{\xi}^{(m+1)}\right)}} \\ &= \frac{\prod_{i=1}^n p\left(y_i \,|\, \boldsymbol{\xi}^{(m+1)}\right)}{\prod_{i=1}^n p\left(y_i \,|\, \boldsymbol{\xi}^{(m)}\right)}, \end{aligned}$$

and only depends on the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{\xi}$. The main advantage of such an approach is that it ensures the convergence in distribution of Markov chain $\left(\boldsymbol{\xi}^{(m)}\right)_m$ to a random variable from $\boldsymbol{\xi}$ given $\boldsymbol{y}$. However, it should be stressed that the MCEM is very time–consuming because a large number of simulations must be performed at each iteration of the EM. On top of that, each Metropolis–Hastings requires a burn–in period where an initial number of samples are thrown away.

4.4.2 Gibbs sampling approach

Just like the Metropolis–Hastings algorithm discussed in Section 4.4.1, the Gibbs sampling approach can be viewed as an alternative strategy to tackle the intractability of the likelihood by sampling from simpler conditional distributions. Very often used in the Bayesian paradigm, it proves to

be a powerful tool to obtain approximations of parameter posterior distributions. Within this framework, the use of the Gibbs sampler for GLMMs has been introduced by Zeger and Karim (1991). Given value $\boldsymbol{\xi}^{(m)}$ for $\boldsymbol{\xi}$, the Gibbs sampler requires generating values according to the conditional distribution $p\big(\boldsymbol{\beta} \,|\, \boldsymbol{\xi}^{(m)}, \boldsymbol{y}\big)$. The authors then assume a flat prior for $\boldsymbol{\beta}$, which implies $p\big(\boldsymbol{\beta} \,|\, \boldsymbol{\xi}^{(m)}, \boldsymbol{y}\big) \propto \prod_{i=1}^{n} p(y_i \,|\, \boldsymbol{\xi})$, and approximate it by a particular Gaussian distribution. This approach was refined and extended by Clayton (1996), using Metropolis–Hastings algorithms. Since their work, the development of new MCMC–based methods for GLMMs has been flourishing, which gave rise to numerous software packages to implement these techniques. The most popular one is probably the R package **MCMCglmm** developed by Hadfield (2010). In this package, emphasis is placed onto reducing the computational cost, in particular through the use of the C library **CSparse** for solving sparse linear systems. Therefore, even if the full conditional density is not in a standard form, models can be fitted in a rather reasonable amount of time.

The above methods have two main disadvantages. The first one is intrinsic to Bayesian methods: they require the specification of prior distributions for the parameters — hard to determine for non–specialists —, which can greatly impact the results. The second disadvantage is intrinsic to MCMC–based statistical inference methods: they assume that the outputs obtained are from the distribution of interest, these outputs being then used to estimate the characteristics of this distribution. However, despite the numerous convergence diagnoses proposed in the literature, we can never be sure that the Markov chain generated has converged towards the target distribution.

We will not give a more in–depth description of the GLMM Bayesian approach here but we refer the reader to Hadfield (2010) and the references therein for further details. The following section focuses instead on recent developments concerning direct approximations of the likelihood by importance sampling–based Monte Carlo methods.

4.4.3 Monte Carlo likelihood approximation

In Section 4.3.1, we have presented deterministic numerical integration approaches — Gauss–Hermite quadrature and adaptive version — to approximate the likelihood of a GLMM which, it should be remembered, is expressed as an integral with respect to the random effects' distribution. Other integration techniques exist, but are based on simulations instead. The focus here will be on a recent importance sampling method proposed by Knudson (2016). Let us begin by briefly discussing the main issues of importance sampling.

Reminder on importance sampling. Importance sampling is a particularly relevant method to approach quantities of the form

$$\mathcal{I} = \int_{\mathbb{R}^d} \phi(\boldsymbol{z}) f(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z} =: \mathbb{E}_f \left[\phi(Z)\right], \tag{4.16}$$

where $Z \in \mathbb{R}^d$ is a random vector with density f, and ϕ is a function. A natural way to approach (4.16) is to generate $\boldsymbol{z_1}, \ldots, \boldsymbol{z_M}$ from f and approximate this expectation by the sample mean

$$\mathbb{E}_f \left[\phi(Z)\right] \simeq \frac{1}{M} \sum_{m=1}^{M} \phi(\boldsymbol{z_m}) =: \widehat{\mathcal{I}}_M^1.$$

Importance sampling is rather based on an alternative representation of (4.16) given by

$$\mathbb{E}_f \left[\phi(Z)\right] = \int_{\mathbb{R}^d} \phi(\boldsymbol{z}) \frac{f(\boldsymbol{z})}{h(\boldsymbol{z})} h(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z} = \mathbb{E}_h \left[\phi(Z) \frac{f(Z)}{h(Z)}\right],$$

where function h — often called importance distribution — fulfils

$$\mathrm{supp}(h) = \left\{\boldsymbol{z} \in \mathbb{R}^d \,\middle|\, h(\boldsymbol{z}) > 0\right\} \supset \left\{\boldsymbol{z} \in \mathbb{R}^d \,\middle|\, \phi(\boldsymbol{z}) f(\boldsymbol{z}) \neq 0\right\}.$$

Therefore, a second way to approximate (4.16) is to simulate $\boldsymbol{y_1}, \ldots, \boldsymbol{y_M}$ from h and take

$$\mathbb{E}_f \left[\phi(Z)\right] \simeq \frac{1}{M} \sum_{m=1}^{M} \phi(\boldsymbol{y_m}) \frac{f(\boldsymbol{y_m})}{h(\boldsymbol{y_m})} =: \widehat{\mathcal{I}}_M^2.$$

As recalled for instance by Robert and Casella (2011), we know by the strong law of large numbers that both estimators $\widehat{\mathcal{I}}_M^1$ and $\widehat{\mathcal{I}}_M^2$ almost surely converge to $\mathcal{I}$, provided that the latter is finite. The popularity of importance sampling is due to the fact that it puts very little restriction on the choice of instrumental density h, a practical requirement for h being that it should be easy to sample from. Assuming

$$\begin{cases} \mathbb{E}_h \left[\phi(Z) \dfrac{f(Z)}{h(Z)}\right] < \infty \\ \mathbb{V}_h \left[\phi(Z) \dfrac{f(Z)}{h(Z)}\right] < \infty, \end{cases}$$

a good idea may be to choose the importance distribution h^* that minimises the variance of $\widehat{\mathcal{I}}_M^2$. Unfortunately, the optimal importance distribution obtained requires the knowledge of $\mathcal{I}$ and thus can not be used in practice. However, heuristics have been developed (see Robert and Casella (2013) and Rubinstein

and Kroese (2016) for instance) whose goal is to propose an importance distribution such that

$$\mathbb{V}_h\left(\widehat{\mathcal{I}}_M^2\right) < \mathbb{V}_f\left(\widehat{\mathcal{I}}_M^1\right).$$

In most cases, the available knowledge on the target density is very limited: identifying h that is easy to sample from and that provides a good approximation of $\mathcal{I}$ remains a delicate problem. But in some situations, adaptive or sequential versions of the importance sampling can be derived (Tokdar and Kass, 2010).

Application to GLMMs. Recall that the likelihood function of a GLMM can be expressed as an integral over the random effects of the joint density of random effects and response vectors:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}) = \int_{\mathbb{R}^q} p(\boldsymbol{y}, \boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi}. \tag{4.17}$$

Considering an importance sampling distribution for the random effects, $h(\boldsymbol{\xi})$, we can rewrite (4.17) as

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}) = \int_{\mathbb{R}^q} \frac{p(\boldsymbol{y}, \boldsymbol{\xi})}{h(\boldsymbol{\xi})} h(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} = \mathbb{E}_h\left(\frac{p(\boldsymbol{y}, \boldsymbol{\xi})}{h(\boldsymbol{\xi})}\right),$$

where $h > 0$ on $\mathbb{R}^q$ in general. An approximation of the entire likelihood function is then given by

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}) \simeq \widehat{\mathcal{L}}_M(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}) = \frac{1}{M} \sum_{m=1}^{M} \frac{p(\boldsymbol{y}, \boldsymbol{\xi}_m)}{h(\boldsymbol{\xi}_m)}, \tag{4.18}$$

where $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M$ are vectors of length q drawn from a distribution with density h. It should be emphasized that the method approximates the entire likelihood function, that is why it is called "Monte Carlo Likelihood Approximation" (MCLA). Unlike MCEM which exclusively focuses on maximum likelihood, the approximation provided by MCLA — namely (4.18) — can be used for any likelihood–based inference.

General properties of MCLA for GLMMs have been studied by Sung et al. (2007). But their approach only involves importance distributions that are independent from the observed data. In theory, this assumption is not restrictive because any importance distribution such that its support contains the support of the target distribution can be chosen. In practice, however, constructing importance distributions independently of the observed data may require extremely long computing time to obtain a reasonable approximation of the likelihood. To counter this problem, the work of Knudson (2016) consists in constructing a family of sampling distributions depending on the observed data such that

(i) it performs well in practice,

(ii) it reduces the computational cost,

(iii) it has attractive theoretical properties.

The importance sampling distribution she suggests is a mixture of three distributions that can be written

$$h(\boldsymbol{\xi}) = \alpha_1 p_1(\boldsymbol{\xi}) + \alpha_2 p_2(\boldsymbol{\xi}) + \alpha_3 p_3(\boldsymbol{\xi}), \tag{4.19}$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$. More precisely in (4.19),

- $p_1(\cdot)$ is a multivariate Student distribution with expectation $\mathbf{0}$ and scale matrix determined by the Penalised Quasi–Likelihood (PQL) estimates of the variance components,
- $p_2(\cdot)$ is a normal distribution centred at the PQL estimates of the random effects and with a variance matrix depending on the PQL estimates of the variance components,
- $p_3(\cdot)$ is also a normal distribution with the same expectation, but with a variance matrix based on the Hessian of the PQL.

Note that the first component of the mixture, $p_1(\cdot)$, is essential to guarantee that the gradient of the MCLA has a central limit theorem. If $\alpha_1 > 0$, as M increases to infinity, the quantity

$$\sqrt{M}\left[\boldsymbol{\nabla}\widehat{\mathcal{L}}_M(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y}) - \boldsymbol{\nabla}\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{y})\right]$$

converges in distribution to a Gaussian distribution with mean $\mathbf{0}$ and finite variance for every $(\boldsymbol{\beta}, \boldsymbol{\sigma})$ and every $\boldsymbol{y}$. That is why in practice, α_1 must be strictly positive. The two other terms are included more for technical reasons and to reduce the computational burden.

The method is implemented in the R package **glmm** (Knudson, 2015), and tested on two real data sets: the *Salamander* data set (McCullagh and Nelder, 1989) and the *Radish* data set (Ridley and Ellstrand, 2010). In the first data set, the response is binary while it is a count in the second. As highlighted in Knudson (2015), the results produced by the R package **glmm** seem close to those produced by the MCEM and those produced by the R package **lme4**. The method seems to reduce the PQL bias but unfortunately, no simulation study is provided to confirm this.

4.5 Linearisation methods

4.5.1 Schall's estimation approach

The strategy proposed by Schall (1991) consists in mixing the estimation approaches used for GLMs on the one hand and for LMMs on the other. Its strength lies in the fact that no specification is required concerning the distribution or the model of the random effects. Initially, the GLMM is considered conditional on the random effects, thus allowing a linearisation of the model via the introduction of a working variable. The underlying linearised model is then seen as an LMM: parameter estimation is performed by solving Henderson's mixed model equations (Henderson, 1975).

Linearisation step. Given random effects $\boldsymbol{\xi}$, the first order Taylor approximation of $g(\boldsymbol{y})$ near $\boldsymbol{\mu}^{\boldsymbol{\xi}}$ is given by

$$\begin{aligned} g(\boldsymbol{y}) &\approx g\left(\boldsymbol{\mu}^{\boldsymbol{\xi}}\right) + \left(\boldsymbol{y} - \boldsymbol{\mu}^{\boldsymbol{\xi}}\right) g'\left(\boldsymbol{\mu}^{\boldsymbol{\xi}}\right) \\ &= \quad \boldsymbol{\eta}^{\boldsymbol{\xi}} \quad + \left(\boldsymbol{y} - \boldsymbol{\mu}^{\boldsymbol{\xi}}\right) g'\left(\boldsymbol{\mu}^{\boldsymbol{\xi}}\right). \end{aligned}$$

As in the GLM estimation procedure, working variable $\boldsymbol{z}^{\boldsymbol{\xi}}$ is defined by

$$\boldsymbol{z}^{\boldsymbol{\xi}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\xi} + \boldsymbol{e},$$

where $\boldsymbol{e} = \left(\boldsymbol{y} - \boldsymbol{\mu}^{\boldsymbol{\xi}}\right) g'\left(\boldsymbol{\mu}^{\boldsymbol{\xi}}\right)$. This leads Schall (1991) to consider the conditional linearised model $\mathcal{M}^{\boldsymbol{\xi}}$ given by

$$(\mathcal{M}^{\boldsymbol{\xi}}): \quad Z^{\boldsymbol{\xi}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\xi} + \boldsymbol{e},$$

for which

$$\begin{aligned} \mathbb{E}\left(Z^{\boldsymbol{\xi}} \middle| \boldsymbol{\xi}\right) &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\xi}, \\ \mathbb{V}\left(Z^{\boldsymbol{\xi}} \middle| \boldsymbol{\xi}\right) &= \mathbb{V}\left(\boldsymbol{e} | \boldsymbol{\xi}\right) \\ &= \mathbb{V}\left[\left(Y - \boldsymbol{\mu}^{\boldsymbol{\xi}}\right) g'\left(\boldsymbol{\mu}^{\boldsymbol{\xi}}\right) \middle| \boldsymbol{\xi}\right] \\ &= \mathbf{Diag}\left(\left[g'\left(\mu_i^{\boldsymbol{\xi}}\right)\right]^2 \mathbb{V}\left(Y_i | \boldsymbol{\xi}\right)\right)_{i=1,\ldots,n} \\ &=: \boldsymbol{W}^{\boldsymbol{\xi}^{-1}}. \end{aligned}$$

Estimation step. In a second step, Schall (1991) considers model $\mathcal{M}^{\boldsymbol{\xi}}$ more as an LMM, with

$$\begin{aligned} \mathbb{E}\left(Z^{\boldsymbol{\xi}}\right) &= \boldsymbol{X}\boldsymbol{\beta}, \\ \mathbb{V}\left(Z^{\boldsymbol{\xi}}\right) &= \boldsymbol{U}\boldsymbol{D}_{\boldsymbol{\sigma}}\boldsymbol{U}^{\intercal} + \boldsymbol{W}^{\boldsymbol{\xi}^{-1}}. \end{aligned} \tag{4.20}$$

At first sight, (4.20) seems inconsistent since the first term of the sum takes into account the random nature of $\boldsymbol{\xi}$ while the second is conditional on $\boldsymbol{\xi}$. The use of $\boldsymbol{W}^{\xi^{-1}}$ instead of $\mathbb{E}\left(\boldsymbol{W}^{\xi^{-1}}\right)$ can nevertheless be justified by the fact that parameter estimates can be obtained by solving the Henderson's equations associated with model $\mathcal{M}^{\xi}$. Indeed, these equations are based on the joint distribution of $\left(Z^{\xi}, \boldsymbol{\xi}\right)$, seen as the product of the conditional distribution of $Z^{\xi}|\boldsymbol{\xi}$ and the distribution of $\boldsymbol{\xi}$. And in the normal approximation of $Z^{\xi}|\boldsymbol{\xi}$, it is matrix that $\boldsymbol{W}^{\xi^{-1}}$ that is involved, not $\mathbb{E}\left(\boldsymbol{W}^{\xi^{-1}}\right)$. The Henderson's equations mentioned above take the form

$$\begin{pmatrix} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{W}^{\xi}\boldsymbol{X} & \boldsymbol{X}^{\mathsf{T}}\boldsymbol{W}^{\xi}\boldsymbol{U} \\ \boldsymbol{U}^{\mathsf{T}}\boldsymbol{W}^{\xi}\boldsymbol{X} & \boldsymbol{U}^{\mathsf{T}}\boldsymbol{W}^{\xi}\boldsymbol{U} + \boldsymbol{D}_{\sigma}^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \boldsymbol{\xi} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{W}^{\xi}\, \boldsymbol{z}^{\xi} \\ \boldsymbol{U}^{\mathsf{T}}\boldsymbol{W}^{\xi}\, \boldsymbol{z}^{\xi} \end{pmatrix}. \tag{4.21}$$

Once random effects predictions are obtained, maximum likelihood (ML) — or restricted maximum likelihood (REML) — estimates of the variance components are available. Algorithm 4.3 describes in detail the resulting iterative process.

4.5.2 Engel and Keen's method

The method proposed by Engel and Keen (1994) is also an iterative procedure alternating between a model linearisation step and an estimation step. It differs from Schall only in the estimation step of the linearised model, since they consider model $\mathcal{M}^{\xi}$ as an LMM with

$$\begin{aligned} \mathbb{E}\left(Z^{\xi}\right) &= \boldsymbol{X}\boldsymbol{\beta}, \\ \mathbb{V}\left(Z^{\xi}\right) &= \mathbb{V}\left[\mathbb{E}\left(Z^{\xi}|\boldsymbol{\xi}\right)\right] + \mathbb{E}\left[\mathbb{V}\left(Z^{\xi}|\boldsymbol{\xi}\right)\right] \\ &= \boldsymbol{U}\boldsymbol{D}_{\sigma}\boldsymbol{U}^{\mathsf{T}} + \mathbb{E}\left(\boldsymbol{W}^{\xi^{-1}}\right). \end{aligned} \tag{4.23}$$

Estimation of fixed–effect parameters and variance components can then be achieved with Algorithm 4.3, but replacing $\boldsymbol{W}^{\xi^{-1}}$ with $\mathbb{E}\left(\boldsymbol{W}^{\xi^{-1}}\right)$.

This procedure seems to be more consistent because (4.23) involves the "true" marginal variance of the working variable. In other words, the randomness of $\boldsymbol{\xi}$ is really taken into account in the marginal model structure. Unfortunately, this method is not always directly usable: although $\mathbb{E}\left(\boldsymbol{W}^{\xi^{-1}}\right)$ is easy to compute in the case of a canonical link (Lavergne and Trottier, 2000), it is not always analytically calculable (e.g. binomial distribution with probit link).

Algorithm 4.3: Schall's algorithm.

Start with an initial guess $\boldsymbol{\beta}^{[0]}$, $\boldsymbol{\xi}^{[0]}$, $\boldsymbol{\sigma}^{[0]}$ and set $t = 0$

while *some convergence criterion not reached* **do**

1. Update the working variable. Given linear predictor $\boldsymbol{\eta}^{\boldsymbol{\xi}^{[t]}} = \boldsymbol{X}\boldsymbol{\beta}^{[t]} + \boldsymbol{U}\boldsymbol{\xi}^{[t]}$, set $\boldsymbol{\mu}^{\boldsymbol{\xi}^{[t]}} = g^{-1}\left(\boldsymbol{\eta}^{\boldsymbol{\xi}^{[t]}}\right)$ and define the working variable as

$$\boldsymbol{z}^{\boldsymbol{\xi}^{[t]}} = \boldsymbol{\eta}^{\boldsymbol{\xi}^{[t]}} + \left(\boldsymbol{y} - \boldsymbol{\mu}^{\boldsymbol{\xi}^{[t]}}\right) g'\left(\boldsymbol{\mu}^{\boldsymbol{\xi}^{[t]}}\right)$$

2. Define the conditional linearised model as

$$\left(\mathcal{M}^{\boldsymbol{\xi}^{[t]}}\right): \quad Z^{\boldsymbol{\xi}^{[t]}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\xi} + \boldsymbol{e}^{[t]},$$

with residual variance matrix $\boldsymbol{W}^{\boldsymbol{\xi}^{[t]}{}^{-1}}$ defined as

$$\boldsymbol{W}^{\boldsymbol{\xi}^{[t]}{}^{-1}} = \mathbf{Diag}\left(\left[g'\left(\mu_i^{\boldsymbol{\xi}^{[t]}}\right)\right]^2 a_i(\phi)\, v\left(\mu_i^{\boldsymbol{\xi}^{[t]}}\right)\right)_{i=1,\ldots,n}$$

3. Solve the Henderson's system

$$\begin{pmatrix} \boldsymbol{X}^\intercal \boldsymbol{W}^{\boldsymbol{\xi}^{[t]}} \boldsymbol{X} & \boldsymbol{X}^\intercal \boldsymbol{W}^{\boldsymbol{\xi}^{[t]}} \boldsymbol{U} \\ \boldsymbol{U}^\intercal \boldsymbol{W}^{\boldsymbol{\xi}^{[t]}} \boldsymbol{X} & \boldsymbol{U}^\intercal \boldsymbol{W}^{\boldsymbol{\xi}^{[t]}} \boldsymbol{U} + \boldsymbol{D}_{\boldsymbol{\sigma}}^{[t]}{}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\xi} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^\intercal \boldsymbol{W}^{\boldsymbol{\xi}^{[t]}} \boldsymbol{z}^{\boldsymbol{\xi}^{[t]}} \\ \boldsymbol{U}^\intercal \boldsymbol{W}^{\boldsymbol{\xi}^{[t]}} \boldsymbol{z}^{\boldsymbol{\xi}^{[t]}} \end{pmatrix} \tag{4.22}$$

to get $\boldsymbol{\beta}^{[t+1]}$ and $\boldsymbol{\xi}^{[t+1]}$.

4. Variance components estimates. $\forall j \in \{1, \ldots, Q\}$, set

- For ML, $\sigma_j^{2[t+1]} = \left(q_j - \dfrac{\text{Trace}\left(\boldsymbol{G}_j^{-1}\boldsymbol{C}_j^{[t]}\right)}{\sigma_j^{2[t]}}\right)^{-1} \boldsymbol{\xi}_j^{[t+1]^\intercal} \boldsymbol{G}_j^{-1} \boldsymbol{\xi}_j^{[t+1]}$,

  where $\boldsymbol{C}^{[t]} = \left(\boldsymbol{U}^\intercal \boldsymbol{W}^{\boldsymbol{\xi}^{[t]}} \boldsymbol{U} + \boldsymbol{D}_{\boldsymbol{\sigma}}^{[t]}{}^{-1}\right)^{-1}$, $\boldsymbol{C}_j^{[t]}$ being the j^{th} sub–matrix of $\boldsymbol{C}^{[t]}$ associated with the j^{th} random effect.

- For REML, $\sigma_j^{2[t+1]} = \left(q_j - \dfrac{\text{Trace}\left(\boldsymbol{G}_j^{-1}\overline{\boldsymbol{C}}_j^{[t]}\right)}{\sigma_j^{2[t]}}\right)^{-1} \boldsymbol{\xi}_j^{[t+1]^\intercal} \boldsymbol{G}_j^{-1} \boldsymbol{\xi}_j^{[t+1]}$,

  where $\overline{\boldsymbol{C}}^{[t]}$ is the matrix formed by the last q rows and columns of the inverse of the Henderson system matrix in (4.22), $\overline{\boldsymbol{C}}_j^{[t]}$ being the j^{th} sub–matrix of $\overline{\boldsymbol{C}}^{[t]}$.

$t \leftarrow t + 1$

end

A brief discussion on linearisation and competing methods is given in Section 4.6.

4.6 Discussion

Because of the intractability of the likelihood, we have seen that various approximate methods have been developed in order to estimate GLMM parameters. Each of these methods has its own advantages and drawbacks.

Deterministic numerical integration methods (GHQ, AGHQ) and Monte Carlo integration methods (MCLA, MCMC) have two main advantages. First, these methods provide approximations of the entire likelihood (not just the maximum likelihood), which can be very useful for any other likelihood–based inference (confidence interval, etc). Second, the accuracy level of these methods can be arbitrarily high by increasing the number of quadrature points or the number of total simulations. Unfortunately, this comes with a price: these methods are computationally intensive, and especially cumbersome for high dimensional random effects, nested random effects, or crossed random effects. Moreover, once an approximation of the likelihood is obtained, it must be maximised. Obviously, this maximisation step is not always easy and may require many additional computational resources. MCMC–type methods have other intrinsic disadvantages, particularly in the assessment of convergence and in the choice of prior distributions. As an indirect likelihood maximisation approach which does not attempt to construct an approximation of the entire likelihood function, one would have thought that MCEM would be less time–consuming. But the method requires the use of a Metropolis–Hastings sampler at each iteration of the EM algorithm, resulting in a considerable computational cost.

Initially defined without any specification of the random effects distribution, Schall's linearisation method (Schall, 1991) proves to be an interesting alternative, because it is so much less time–consuming. This approach first appeared to be awkward but was justified by many authors, who obtained the same equations but with different reasoning. The first two examples we can think of are the method of Wolfinger and O'Connell (1993) — which can be seen as another linearisation method slightly extending Schall's —, and the PQL method of Breslow and Clayton (1993) — based on a Laplace approximation of the marginal quasi–likelihood (see Section 4.3.3). Lee and Nelder (1996) also use Laplace approximation, within the broader framework of Hierarchical Generalised Linear Models (HGLMs), which incorporates GLMMs. In the case of a GLMM, maximising the "h–likelihood" they introduce is equivalent

to maximising the penalised quasi–likelihood, and thus to solving the system (4.21) suggested by Schall.

Finally, as initially emphasised by Stiratelli et al. (1984) in the specific context of a random effect model with a binary response, Schall's method can also be justified from a Bayesian point of view. Let $p(\boldsymbol{\beta} \mid \boldsymbol{B})$ be the normal prior distribution of $\boldsymbol{\beta}$ with variance matrix $\boldsymbol{B}$ and $p(\boldsymbol{\xi} \mid \boldsymbol{D_\sigma})$ the centred normal prior distribution of $\boldsymbol{\xi}$ with dispersion matrix $\boldsymbol{D_\sigma}$. Let $p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\xi})$ be the conditional density of $\boldsymbol{y}$. The posterior density for $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ satisfies

$$p(\boldsymbol{\beta}, \boldsymbol{\xi} \mid \boldsymbol{y}, \boldsymbol{B}, \boldsymbol{D_\sigma}) \propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\xi})\, p(\boldsymbol{\beta} \mid \boldsymbol{B})\, p(\boldsymbol{\xi} \mid \boldsymbol{D_\sigma})\,. \tag{4.24}$$

Now, under a non–informative (or diffuse) prior for $\boldsymbol{\beta}$, i.e. when $\boldsymbol{B}^{-1} = \boldsymbol{0}$, (4.24) becomes

$$p(\boldsymbol{\beta}, \boldsymbol{\xi} \mid \boldsymbol{y}, \boldsymbol{B}, \boldsymbol{D_\sigma}) \propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\xi})\, p(\boldsymbol{\xi} \mid \boldsymbol{D_\sigma})\,.$$

The posterior log–likelihood is then, up to a constant, the sum of $\log[p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\xi})]$ and $\log[p(\boldsymbol{\xi} \mid \boldsymbol{D_\sigma})]$. Recall that the conditional density of $\boldsymbol{y}$ is in the exponential family by assumption. Using some results from McCullagh and Nelder (1989) relative to exponential families, Schall (1991) finally shows that mixed model equations (4.21) constitute an iteration of the Fisher's scoring method to maximise $p(\boldsymbol{\beta}, \boldsymbol{\xi} \mid \boldsymbol{y}, \boldsymbol{B}, \boldsymbol{D_\sigma})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$.

Even if the Schall's method proves to work reasonably well in many situations, it is known to exhibit downward bias when the conditional distribution of the response is highly non–normal as is the case for binary data. Other estimation strategies should then be considered, such as Monte Carlo methods. The latter require much longer computing time but in case of binary data, the estimates provided are likely to be more accurate.

Despite its disadvantages, Schall's method will be used a lot throughout this work, mainly for its ease of implementation, speed and good results in most cases. In addition, as we will see in Chapter 5, Schall's method provides us with a linear setting more suitable for the computation of components we want to implement. Estimating the variance components accurately will indeed be an issue secondary to us: even if it means refining the parameter estimation in a second step — using sophisticated simulation methods for instance — our primary and main purpose will be to investigate the explanatory structure of the GLMM's fixed design and relate it to interpretable dimensions.

V

Component–based regularisation of multivariate GLMMs

Do the best you can until you know better.
Then, when you know better, do better.
— Maya Angelou

Contents

This chapter is based on an article accepted for future publication in *Journal of Computational and Graphical Statistics*.

5.1 Introduction

In hindsight, it seems that many applications (in ecology or social sciences to cite just a few of the countless areas involved) require facing three difficulties together.

(i) First, responses may be non–Gaussian ;

(ii) second, observations may be clustered ;

(iii) last, it often happens that the true explanatory dimensions are latent and indirectly measured through highly correlated proxies.

As mentioned in Chapter 4, items (i) and (ii) involve the use of GLMMs. On the other hand, item (iii) demands the development of specific regularisation techniques. In a non–exhaustive way, essentially two penalised–based regularisation methods have been proposed in the GLMM framework. With a view towards variable selection, Schelldorfer et al. (2014) and Groll and Tutz (2014) simultaneously developed L_1–penalty approaches for fitting potentially high–dimensional GLMMs. Both are based on a Laplace approximation of the likelihood but each uses a particular optimisation algorithm. Furthermore, for Gaussian responses only, Eliot et al. (2011) extended the ridge regression to LMMs using L_2–penalised EM algorithm.

However, in the situation of interest where the explanatory variables are considered as proxies to latent dimensions which have to be recovered, none of the methods mentioned above is satisfactory. Indeed, variable selection via the LASSO is inappropriate, because it tends to select a single predictor among a set of correlated ones and ignore the others. By preserving all the explanatory variables in the model, the ridge–based approach seems more appropriate, but unfortunately leads to a linear predictor that is difficult to interpret.

In the situation described by item (iii), the most appropriate approach is undoubtedly the component–based regularisation. However, to the best of our knowledge, no component–based regularisation method for GLMMs is currently available. In addition, neither of the two penalty–based methods discussed above is designed for multivariate responses, let alone for multivariate responses of different types (e.g. one binary and another Poisson), when it is often necessary in many cases. To fill these gaps, it proved necessary to develop a method allowing modelling grouped responses through a multivariate GLMM with a large number of explanatory variables. For this purpose, we combine Schall's iterative model linearisation with regularisation at each step. However, we do not use a penalty on the coefficient vector's norm — as proposed by Zhang et al. (2017) within the framework of multi-

variate count data. We rather propose to combine dimension–reduction and predictor–regularisation using supervised components aligning on the most predictive and interpretable directions in the explanatory space. As an extension of the Supervised Component–based Generalised Linear Regression (SCGLR), the main purpose still remains to investigate the explanatory structure and link it to interpretable dimensions.

The chapter is organised as follows. In Section 5.2, we formalise the model and set the main notations used throughout the chapter. In Section 5.3, we present the key features of a slight improvement of SCGLR that includes additional explanatory variables. Section 5.4 designs an extension of this methodology to mixed models, and particularly to grouped data. In Section 5.5, our extended method, "mixed–SCGLR", is evaluated through simulations and compared to ridge– and LASSO–based regularisations. In order to highlight the power of mixed–SCGLR in terms of model interpretation, Section 5.6 presents an application to real data in the Poisson case. In this application, we aim at modelling abundances of several tree genera with multiple redundant explanatory variables. The land–plots on which the measures are collected are grouped into forest concessions. Finally, Section 5.7 discusses the positive points, limitations and possible improvements of mixed–SCGLR. In addition, Appendix 5.8 presents some technical details useful for an overall understanding of the method.

5.2 Model definition and notations

In the framework of a multivariate GLMM, we consider q response–vectors $\boldsymbol{y_1}, \ldots, \boldsymbol{y_q}$ forming matrix $\boldsymbol{Y}_{n\times q}$, to be explained by two categories of explanatory variables. The first category consists of few weakly correlated variables $\boldsymbol{A}_{n\times r} = \left[\boldsymbol{a_1} \mid \ldots \mid \boldsymbol{a_r}\right]$. These variables are assumed to be interesting per se and their marginal effects need to be precisely quantified. The second category consists of abundant and highly correlated variables $\boldsymbol{X}_{n\times p} = \left[\boldsymbol{x_1} \mid \ldots \mid \boldsymbol{x_p}\right]$ considered as proxies to latent dimensions which must be found and interpreted. Since explanatory variables in $\boldsymbol{A}$ are few, non–redundant and of interest, they are kept as such in the model. By contrast, $\boldsymbol{X}$ may contain several unknown structurally relevant dimensions $K < p$ important to model and predict $\boldsymbol{Y}$, how many we do not know. $\boldsymbol{X}$ is thus to be searched for an appropriate number of orthogonal components that both capture relevant structural information in $\boldsymbol{X}$ and contribute to model $\boldsymbol{Y}$.

This work addresses grouped data: the n observations form N groups. Within each group, observations are not assumed independent. For each re-

sponse $\boldsymbol{y}_k$, a N–level random effect $\boldsymbol{\xi}_k$ is used to model the dependence of observations within each group. Hence, each $\boldsymbol{y}_k$ is modelled with a GLMM assuming a conditional distribution from the exponential family.

For the sake of clarity, here are the notations and conventions that will be used throughout the chapter:

- As earlier, bold lowercase letters (e.g. $\boldsymbol{u}$) refer to vectors and bold capital letters (e.g. $\boldsymbol{M}$) refer to matrices. The transpose of $\boldsymbol{M}$ is noted $\boldsymbol{M}^{\mathrm{T}}$.
- All variables (namely the $\boldsymbol{a}_i$'s, $\boldsymbol{x}_j$'s and $\boldsymbol{y}_k$'s) will be identified with n–vectors.
- Let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ and $\boldsymbol{M}$ be a symmetric positive semi–definite matrix of size $d \times d$. The Euclidean norm of $\boldsymbol{u}$ with respect to metric $\boldsymbol{M}$ will be denoted $\|\boldsymbol{u}\|_{\boldsymbol{M}}$, and the Euclidean scalar product of $\boldsymbol{u}$ and $\boldsymbol{v}$ will be denoted $\langle \boldsymbol{u} \,|\, \boldsymbol{v} \rangle_{\boldsymbol{M}} = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{v}$. If $\boldsymbol{u}$ and $\boldsymbol{v}$ are non–zero vectors,

$$\cos_{\boldsymbol{M}}(\boldsymbol{u}, \boldsymbol{v}) = \frac{\langle \boldsymbol{u} \,|\, \boldsymbol{v} \rangle_{\boldsymbol{M}}}{\|\boldsymbol{u}\|_{\boldsymbol{M}} \, \|\boldsymbol{v}\|_{\boldsymbol{M}}}$$

  refers to their cosine with respect to $\boldsymbol{M}$. When $M = \boldsymbol{I}_d$, we will simply write $\|\boldsymbol{u}\|$, $\langle \boldsymbol{u} \,|\, \boldsymbol{v} \rangle$ and $\cos(\boldsymbol{u}, \boldsymbol{v})$.
- The space spanned by vectors $\boldsymbol{u_1}, \ldots, \boldsymbol{u_h}$ is denoted by $\mathrm{span}\{\boldsymbol{u_1}, \ldots, \boldsymbol{u_h}\}$. $\boldsymbol{U}$ being any matrix, $\mathrm{span}\{\boldsymbol{U}\}$ refers to the space spanned by the column–vectors of $\boldsymbol{U}$.
- Let $\mathbb{R}^n$ be endowed with metric $\boldsymbol{W}$ and let $\boldsymbol{Z}$ be a matrix of size $n \times p$. Then $\Pi^{\boldsymbol{W}}_{\mathrm{span}\{\boldsymbol{Z}\}}$ refers to the $\boldsymbol{W}$–orthogonal projector onto $\mathrm{span}\{\boldsymbol{Z}\}$. Similarly, when $\boldsymbol{W} = \boldsymbol{I}_n$, we will simply denote $\Pi_{\mathrm{span}\{\boldsymbol{Z}\}}$. Let $\boldsymbol{b}$ be a vector in $\mathbb{R}^n$. The cosine of the angle between $\boldsymbol{b}$ and $\mathrm{span}\{\boldsymbol{Z}\}$ with respect to $\boldsymbol{W}$ is defined by $\cos_{\boldsymbol{W}}(\boldsymbol{b}, \mathrm{span}\{\boldsymbol{Z}\}) = \cos_{\boldsymbol{W}}\left(\boldsymbol{b}, \Pi^{\boldsymbol{W}}_{\mathrm{span}\{\boldsymbol{Z}\}}\boldsymbol{b}\right)$.

5.3 SCGLR with additional explanatory variables

In this section, we consider the situation where each $\boldsymbol{y}_k$ is modelled with a GLM (without random effect). For the sake of simplicity, we focus on the single–component SCGLR ($K = 1$). Section 5.3.1 briefly recalls the useful standards of a univariate GLM. Section 5.3.2 defines the linear predictors considered in the SCGLR methodology, in a multivariate GLM framework with additional explanatory variables. Finally, Section 5.3.3 introduces the criterion which SCGLR maximises to compute the component.

5.3.1 Notations and main features of univariate GLMs

We refer the reader to McCullagh and Nelder (1989) for a thorough overview of GLMs. This section is only intended to recall the classical iterative scheme performing maximum likelihood (ML) estimation. Let $\boldsymbol{X}$ denote the $n \times p$ matrix of explanatory variables and $\boldsymbol{\beta}$ the p–dimensional parameter vector. At iteration $t+1$, the Fisher Scoring Algorithm (FSA) for ML estimation calculates

$$\boldsymbol{\beta}^{[t+1]} = \left(\boldsymbol{X}^{\intercal}\boldsymbol{W}^{[t]}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\intercal}\boldsymbol{W}^{[t]}\boldsymbol{z}^{[t]}, \tag{5.1}$$

where $\boldsymbol{z}^{[t]}$ and $\boldsymbol{W}^{[t]}$ respectively denote the classical working variable and the associated weight matrix at iteration t (see Sections 3.2 and 3.3.1). As pointed out by Nelder and Wedderburn (1972), update (5.1) may be interpreted as a weighted least squares step in the linearised model $\mathcal{M}^{[t]}$ defined by

$$\mathcal{M}^{[t]}: \left| \begin{array}{l} \boldsymbol{z}^{[t]} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\zeta}^{[t]}, \\ \text{with: } \begin{cases} \mathbb{E}\left(\boldsymbol{\zeta}^{[t]}\right) = \boldsymbol{0}, \\ \mathbb{V}\left(\boldsymbol{\zeta}^{[t]}\right) = {\boldsymbol{W}^{[t]}}^{-1}. \end{cases} \end{array} \right. \tag{5.2}$$

5.3.2 Linear predictors for SCGLR with multiple responses

We are now considering a multivariate GLM (Fahrmeir and Tutz, 1994). In this context, SCGLR searches for an explanatory component common to all the $\boldsymbol{y_k}$'s. This component will be denoted $\boldsymbol{f}$ and its p–dimensional loading–vector will be denoted $\boldsymbol{u}$, so that $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$. The linear predictor associated with response–vector $\boldsymbol{y_k}$ then writes

$$\boldsymbol{\eta_k} = (\boldsymbol{X}\boldsymbol{u})\,\gamma_k + \boldsymbol{A}\boldsymbol{\delta_k}, \tag{5.3}$$

where γ_k and $\boldsymbol{\delta_k}$ are the regression parameters associated respectively with component $\boldsymbol{f}$ and additional explanatory variables $\boldsymbol{A}$. $\boldsymbol{f}$ being common to all the $\boldsymbol{y_k}$'s, predictors are collinear in their $\boldsymbol{X}$–part. For identification purposes, we impose $\boldsymbol{u}^{\intercal}\boldsymbol{M}^{-1}\boldsymbol{u} = 1$, where $\boldsymbol{M}$ may so far be any $p \times p$ symmetric positive definite matrix. Let us note $y_{k,i}$ the i–th observation of the k–th response–vector and $\boldsymbol{H} = \{\eta_{k,i} \mid 1 \leqslant k \leqslant q, 1 \leqslant i \leqslant n\}$ the predictor set. We assume that the q responses are independent conditional on any linear combination of $\boldsymbol{X}$ and $\boldsymbol{A}$, and that the n observations are independent. The log–density then writes

$$\ell\left(\boldsymbol{Y}|\boldsymbol{H}\right) = \sum_{i=1}^{n}\sum_{k=1}^{q}\ell_k\left(y_{k,i}|\eta_{k,i}\right),$$

where ℓ_k is the log–density of the k–th response, conditional on its linear predictor. As a result, $\boldsymbol{z_k}$ being the working variable associated with $\boldsymbol{y_k}$ and $\boldsymbol{W_k}^{-1}$ its variance matrix, the corresponding linearised model derived from the FSA at iteration t is

$$\mathcal{M}_k^{[t]}: \left| \begin{array}{l} \boldsymbol{z_k}^{[t]} = (\boldsymbol{X}\boldsymbol{u})\,\gamma_k + \boldsymbol{A}\boldsymbol{\delta_k} + \boldsymbol{\zeta_k}^{[t]}, \\ \text{with: } \begin{cases} \mathbb{E}\left(\boldsymbol{\zeta_k}^{[t]}\right) = \mathbf{0}, \\ \mathbb{V}\left(\boldsymbol{\zeta_k}^{[t]}\right) = \boldsymbol{W_k}^{[t]^{-1}}. \end{cases} \end{array} \right. \tag{5.4}$$

Although linearised models (5.2) and (5.4) seem very similar, (5.4) is no longer linear, owing to the product $\boldsymbol{u}\gamma_k$. An alternated version of the FSA must therefore be used:

(i) Given current values of all the γ_k's and $\boldsymbol{\delta_k}$'s, a new loading–vector $\boldsymbol{u}$ is obtained by solving an SCGLR–specific program (see Section 5.3.3 for details).

(ii) Given a current value of $\boldsymbol{u}$, each $\boldsymbol{z_k}$ is regressed independently on $\left[\, \boldsymbol{X}\boldsymbol{u} \mid \boldsymbol{A} \,\right]$ with respect to weight matrix $\boldsymbol{W_k}$, yielding new regression parameter estimates γ_k and $\boldsymbol{\delta_k}$.

5.3.3 Calculating the component: an SCGLR–specific criterion

For an easier reading of this part, we omit the $[t]$ index. For each $k \in \{1, \ldots, q\}$, consider model $\mathcal{M}_k$ endowed with weight matrix $\boldsymbol{W_k}$. The best loading–vector in the weighted least–squares sense would be the solution of (Bry et al., 2013):

$$\begin{aligned} & \min_{\boldsymbol{u}:\, \boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{u}=1} \sum_{k=1}^{q} \left\| \boldsymbol{z_k} - \Pi^{\boldsymbol{W_k}}_{\mathrm{span}\{\boldsymbol{X}\boldsymbol{u},\boldsymbol{A}\}} \boldsymbol{z_k} \right\|^2_{\boldsymbol{W_k}} \\ \Longleftrightarrow \quad & \max_{\boldsymbol{u}:\, \boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{u}=1} \sum_{k=1}^{q} \left\| \Pi^{\boldsymbol{W_k}}_{\mathrm{span}\{\boldsymbol{X}\boldsymbol{u},\boldsymbol{A}\}} \boldsymbol{z_k} \right\|^2_{\boldsymbol{W_k}}. \end{aligned}$$

The maximisation program also writes $\max_{\boldsymbol{u}:\, \boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{u}=1} \psi_{\boldsymbol{A}}(\boldsymbol{u})$, where

$$\begin{aligned} \psi_{\boldsymbol{A}}(\boldsymbol{u}) &= \sum_{k=1}^{q} \left\| \boldsymbol{z_k} \right\|^2_{\boldsymbol{W_k}} \cos^2_{\boldsymbol{W_k}} \left(\boldsymbol{z_k},\, \mathrm{span}\,\{\boldsymbol{X}\boldsymbol{u}, \boldsymbol{A}\} \right) \\ &= \sum_{k=1}^{q} \left\| \boldsymbol{z_k} \right\|^2_{\boldsymbol{W_k}} \cos^2_{\boldsymbol{W_k}} \left(\boldsymbol{z_k},\, \Pi^{\boldsymbol{W_k}}_{\mathrm{span}\{\boldsymbol{X}\boldsymbol{u},\boldsymbol{A}\}} \boldsymbol{z_k} \right). \end{aligned} \tag{5.5}$$

Now, $\psi_{\boldsymbol{A}}$ is a mere goodness–of–fit (GoF) measure that does not take into account the closeness of component $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$ to interpretable directions in $\boldsymbol{X}$. The GoF measure, $\psi_{\boldsymbol{A}}$, must therefore be combined with a measure ϕ of structural relevance (SR).

Assume matrix $\boldsymbol{X}$ consists of p standardised numeric variables. Consider a weight system $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_p\}$ — e.g. $\omega_j = \frac{1}{p}\ \forall j \in \{1, \ldots, p\}$ — reflecting the a priori relative importance of variables. Also consider a weight matrix $\boldsymbol{P}$ — e.g. $\boldsymbol{P} = \frac{1}{n}\boldsymbol{I}_n$ — reflecting the a priori relative importance of observations. We define the most structurally relevant loading–vector as the solution of

$$\max_{\boldsymbol{u}:\, \boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{u}=1} \phi\left(\boldsymbol{u}\right),$$

where

$$\begin{aligned} \phi\left(\boldsymbol{u}\right) &= \left[\sum_{j=1}^{p} \omega_j \left(\langle\, \boldsymbol{X}\boldsymbol{u} \,|\, \boldsymbol{x_j} \,\rangle_{\boldsymbol{P}}^2\right)^l\right]^{\frac{1}{l}} \\ &= \left[\sum_{j=1}^{p} \omega_j \left(\boldsymbol{u}^{\mathrm{T}}\, \boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{x_j}\boldsymbol{x_j}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}\, \boldsymbol{u}\right)^l\right]^{\frac{1}{l}}, \ l \geqslant 1, \end{aligned} \tag{5.6}$$

for the scalar product is commutative. Formula (5.6) is in fact a particular case of the SR criterion proposed by Bry et al. (2016). It can be viewed as a generalised average version of the usual dual PCA criterion: $\sum_{j=1}^{p} \cos_{\boldsymbol{P}}^2\left(\boldsymbol{X}\boldsymbol{u}, \boldsymbol{x_j}\right) = \sum_{j=1}^{p} \langle\, \boldsymbol{X}\boldsymbol{u} \,|\, \boldsymbol{x_j} \,\rangle_{\boldsymbol{P}}^2$. For $\boldsymbol{M} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}\right)^{-1}$, (5.6) is called "Variable–Powered Inertia" (VPI). It should be stressed that for $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}$ to be invertible, $\boldsymbol{X}$ must be a full column rank matrix. In case of strict collinearities within $\boldsymbol{X}$, as always happens in high–dimensional settings, we replace $\boldsymbol{X}$ with the matrix $\boldsymbol{C}$ of its principal components associated with non–zero eigenvalues. Bry et al. (2018) and Appendix 5.8.1.3 provide further details on the use of principal components in VPI when $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}$ is singular.

Tuning parameter l allows to draw component towards more (greater l) or less (smaller l) local bundles of correlated variables, as depicted on Figure 5.1 in the particular instance of four coplanar variables. Informally, a bundle is a set of variables correlated "enough" to be viewed as proxies to the same latent dimension. The notion of bundle is flexible, and parameter l tunes the level of within–bundle correlation to be considered: the higher the correlation, the more local the bundle. Overall, taking $l = 1$ draws the components towards global structural directions (namely the principal components) while taking l higher leads to more local ones (ultimately, the variables themselves). The goal is to focus on the most interpretable directions.

Finally, let $s \in [0, 1]$ be a parameter tuning the importance of the SR rela-

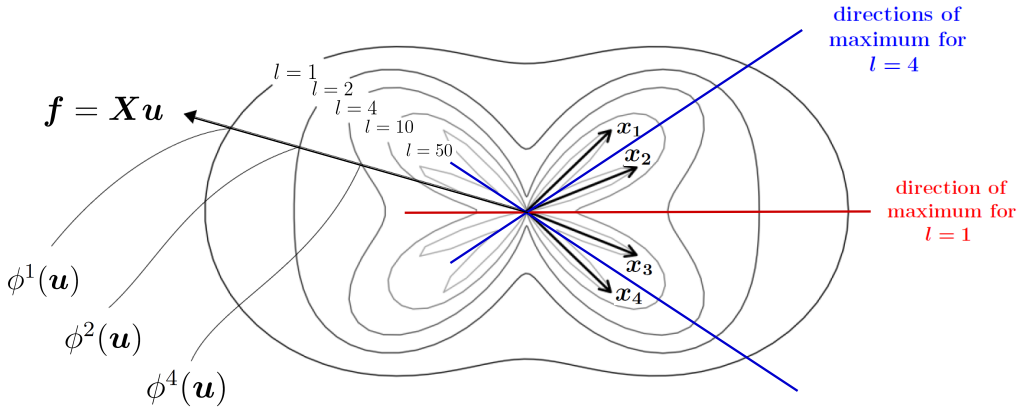


Figure 5.1 – ***Polar representation of the VPI according to the value of l in the elementary case of four coplanar variables, $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \boldsymbol{x_4}$, with $\omega_j = \frac{1}{4}\, \forall j \in \{1, 2, 3, 4\}$.*** *Loading–vector $\boldsymbol{u}$ is identified with complex number $e^{i\theta}$, where $\theta \in [0, 2\pi)$. Curves $z_l(\theta) := \left[\phi\left(e^{i\theta}\right)\right]^l e^{i\theta}$ are graphed for $l \in \{1, 2, 4, 10, 50\}$. The intersection of curve z_l with $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$ has a radius equal to $\left[\phi\left(e^{i\theta}\right)\right]^l$. The red line is the direction of maximum for $l = 1$, which is in fact the first principal component. These four variables are then regarded as a unique bundle. By contrast, the blue lines represent the two directions of maximum for $l = 4$. The variables are then seen as two bundles containing two variables each. Finally, when $l = 50$, each variable is considered a bundle in itself.*

tive to the GoF. SCGLR attempts a trade–off between (5.5) and (5.6) by solving

$$\max_{\boldsymbol{u}:\, \boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{u}=1} \left\{ \mathcal{J}(\boldsymbol{u}) := [\phi(\boldsymbol{u})]^s \; [\psi_{\boldsymbol{A}}(\boldsymbol{u})]^{1-s} \right\},$$

or equivalently

$$\max_{\boldsymbol{u}:\, \boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{u}=1} \left\{ \log \mathcal{J}(\boldsymbol{u}) = s \log[\phi(\boldsymbol{u})] + (1-s) \log[\psi_{\boldsymbol{A}}(\boldsymbol{u})] \right\}. \tag{5.7}$$

Although criterion $\mathcal{J}$ may seem unconventional, it brings out a clear interpretation of trade–off parameter s. Indeed, at the maximum, relative variations compensate since we have

$$\begin{aligned} & \boldsymbol{\nabla} \log \mathcal{J}(\boldsymbol{u}) = \boldsymbol{0} \\ \iff \quad & s\, \frac{\boldsymbol{\nabla}\phi(\boldsymbol{u})}{\phi(\boldsymbol{u})} + (1-s) \frac{\boldsymbol{\nabla}\psi_{\boldsymbol{A}}(\boldsymbol{u})}{\psi_{\boldsymbol{A}}(\boldsymbol{u})} = \boldsymbol{0} \\ \iff \quad & \frac{\boldsymbol{\nabla}\psi_{\boldsymbol{A}}(\boldsymbol{u})}{\psi_{\boldsymbol{A}}(\boldsymbol{u})} = \frac{s}{1-s} \left[-\frac{\boldsymbol{\nabla}\phi(\boldsymbol{u})}{\phi(\boldsymbol{u})} \right]. \end{aligned}$$

Ratio $s/(1-s)$ can therefore be interpreted as the marginal rate of substitution of SR to GoF. Informally, it measures locally the number of additional units

of GoF needed to compensate the loss of one unit of SR, in order to maintain criterion $\mathcal{J}$ constant. For illustrative purposes, a few examples can be given:

- If $s = 0.2$, then ratio $s/(1-s) = 0.25$. In this case, GoF is considered more important than SR, and one unit of SR can be substituted for only 0.25 unit of GoF.
- If $s = 0.5$, then $s/(1-s) = 1$ which means that GoF and SR are equally important.
- Finally, if $s = 0.9$, then $s/(1-s) = 9$. Here, SR is given more weight than GoF so that 9 units of GoF are necessary to compensate the loss of one unit of SR.

For an easier reading, this section was centred on a particular case of SR called VPI. The general formula for structural relevance and other detailed examples can be found in Appendix 5.8.1. Note also that analytical expression of $\mathcal{J}(\boldsymbol{u})$ is derived in Appendix 5.8.2.

5.4 Extension to mixed models

We now propose to extend SCGLR to mixed models. This extension will be called "mixed–SCGLR". A particular focus is placed on grouped data, for which the independence assumption of observations is no longer valid. The within–group dependence of each response is modelled with a random group–effect. Consequently, each $\boldsymbol{y_k}$ is modelled with a GLMM. As in SCGLR, the responses are assumed independent conditional on the components. Section 5.4.1 presents the single–component mixed–SCGLR method. The underlying algorithm is given in Section 5.4.2. Considering only one component is generally not enough to explain the responses making it necessary to search for K explanatory components, with $1 \leqslant K \leqslant \text{rank}(\boldsymbol{X})$. The way in which we extract higher rank components is explained in Section 5.4.3.

5.4.1 First component

The random group–effect is assumed different across responses. This leads to q random–effect vectors $\boldsymbol{\xi_1}, \ldots, \boldsymbol{\xi_q}$, which are assumed independent and normally distributed:

$$\forall k \in \{1, \ldots, q\}, \quad \boldsymbol{\xi_k} \overset{\text{ind.}}{\sim} \mathcal{N}_N(\boldsymbol{0}, \boldsymbol{D_k}),$$

where N denotes the number of groups. In this work, variance components models will be considered. We assume $\boldsymbol{D_k} = \sigma_k^2 \boldsymbol{I}_N$, where σ_k^2 is the group variance component associated with response $\boldsymbol{y_k}$. Linear predictors involved in mixed–SCGLR are expressed as

$$\forall k \in \{1, \ldots, q\}\,, \quad \boldsymbol{\eta_k^\xi} = (\boldsymbol{X u})\gamma_k + \boldsymbol{A \delta_k} + \boldsymbol{U \xi_k}, \tag{5.8}$$

where $\boldsymbol{U}$ is the known random effect design matrix. Predictor $\boldsymbol{\eta_k^\xi}$ epitomises the way we capture the dependence between outcomes. Indeed, as component $\boldsymbol{f} = \boldsymbol{X u}$ does not depend on k, it captures a structural dependence between the various $\boldsymbol{y_k}$'s. By contrast, the random effect $\boldsymbol{\xi_k}$ models the within–group stochastic dependence of outcomes forming response–vector $\boldsymbol{y_k}$.

The distribution of the responses conditional on the random effects is supposed to belong to the exponential family. The FSA was adapted by Schall (1991) to the GLMM dependence structure. The key idea is to extend Schall's algorithm to the component–based predictors in (5.8).

5.4.1.1 Linearisation step

Let g_k denote the link function for response $\boldsymbol{y_k}$, g_k' its first derivative and $\boldsymbol{\mu_k^\xi}$ the conditional expectation (i.e. $\boldsymbol{\mu_k^\xi} := \mathbb{E}\left(\boldsymbol{y_k} \,|\, \boldsymbol{\xi_k}\right)$). Working variable associated with $y_{k,i}$ is calculated through

$$\begin{aligned} z_{k,i}^\xi &= g_k\left(\mu_{k,i}^\xi\right) + \left(y_{k,i} - \mu_{k,i}^\xi\right) g_k'\left(\mu_{k,i}^\xi\right) \\ &= \eta_{k,i}^\xi + e_{k,i}, \quad \text{where} \quad e_{k,i} = \left(y_{k,i} - \mu_{k,i}^\xi\right) g_k'\left(\mu_{k,i}^\xi\right). \end{aligned}$$

In view of the conditional independence assumption, the conditional variance matrix for $\boldsymbol{z_k^\xi}$ is

$$\text{Var}\left(\boldsymbol{z_k^\xi} \,|\, \boldsymbol{\xi_k}\right) = \boldsymbol{W_k^\xi}^{-1} = \textbf{Diag}\left(\left[g_k'\left(\mu_{k,i}^\xi\right)\right]^2 a_{k,i}(\phi_k)\, v_k\left(\mu_{k,i}^\xi\right)\right)_{i=1,\ldots,n},$$

where $a_{k,i}$ and v_k are known functions, and ϕ_k is the dispersion parameter related to $\boldsymbol{y_k}$. At iteration t, the conditional linearised model for working vector $\boldsymbol{z_k^\xi}$ is then defined by

$$\mathcal{M}_k^{\xi\,[t]}: \left| \begin{array}{l} \boldsymbol{z_k^\xi}^{[t]} = (\boldsymbol{X u})\,\gamma_k + \boldsymbol{A \delta_k} + \boldsymbol{U \xi_k} + \boldsymbol{e_k^{[t]}} \\ \text{with:} \begin{cases} \mathbb{E}\left(\boldsymbol{e_k^{[t]}} \,|\, \boldsymbol{\xi_k}\right) = \boldsymbol{0}, \\ \mathbb{V}\left(\boldsymbol{e_k^{[t]}} \,|\, \boldsymbol{\xi_k}\right) = \boldsymbol{W_k^\xi}^{-1[t]}. \end{cases} \end{array} \right. \tag{5.9}$$

Besides the variance component estimation, an alternated estimation step has to be developed (as aforementioned in Section 5.3.2) to deal with the non–linearity of (5.9).

5.4.1.2 Estimation step

Calculating the component: Given current values of all the γ_k's, $\boldsymbol{\delta_k}$'s, $\boldsymbol{\xi_k}$'s and σ_k^2's, a new component $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$ is calculated by solving a (5.7)–type program. However, (5.5) has to be adapted to conditional linearised models $\mathcal{M}_k^{\xi}$'s, involving weight matrices $\boldsymbol{W}_k^{\xi}$'s. The appropriate goodness–of–fit measure is

$$\psi_{\boldsymbol{A}}(\boldsymbol{u}) = \sum_{k=1}^{q} \left\| \boldsymbol{z}_k^{\xi} \right\|_{\boldsymbol{W}_k^{\xi}}^2 \cos^2_{\boldsymbol{W}_k^{\xi}} \left(\boldsymbol{z}_k^{\xi}, \operatorname{span}\{\boldsymbol{X}\boldsymbol{u}, \boldsymbol{A}\} \right). \tag{5.10}$$

Computing regression parameters and variance–component estimates: Given a current value of component $\boldsymbol{f}$, we apply Schall's method with the linear predictors given in (5.8). New values of parameters γ_k and $\boldsymbol{\delta_k}$ as well as new prediction $\boldsymbol{\xi_k}$ are obtained by solving the following Henderson's system (Henderson, 1975):

$$\begin{pmatrix} \boldsymbol{f}^{\mathrm{T}}\boldsymbol{W}_k^{\xi}\boldsymbol{f} & \boldsymbol{f}^{\mathrm{T}}\boldsymbol{W}_k^{\xi}\boldsymbol{A} & \boldsymbol{f}^{\mathrm{T}}\boldsymbol{W}_k^{\xi}\boldsymbol{U} \\ \boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_k^{\xi}\boldsymbol{f} & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_k^{\xi}\boldsymbol{A} & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_k^{\xi}\boldsymbol{U} \\ \boldsymbol{U}^{\mathrm{T}}\boldsymbol{W}_k^{\xi}\boldsymbol{f} & \boldsymbol{U}^{\mathrm{T}}\boldsymbol{W}_k^{\xi}\boldsymbol{A} & \boldsymbol{U}^{\mathrm{T}}\boldsymbol{W}_k^{\xi}\boldsymbol{U} + \boldsymbol{D}_k^{-1} \end{pmatrix} \begin{pmatrix} \gamma_k \\ \boldsymbol{\delta_k} \\ \boldsymbol{\xi_k} \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}^{\mathrm{T}}\boldsymbol{W}_k^{\xi}\boldsymbol{z}_k^{\xi} \\ \boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_k^{\xi}\boldsymbol{z}_k^{\xi} \\ \boldsymbol{U}^{\mathrm{T}}\boldsymbol{W}_k^{\xi}\boldsymbol{z}_k^{\xi} \end{pmatrix}.$$

Finally, as mentioned by Schall (1991), given prediction $\widehat{\boldsymbol{\xi}_k}$ for $\boldsymbol{\xi_k}$, the update of the ML estimation of variance component σ_k^2 is

$$\sigma_k^2 \longleftarrow \frac{\widehat{\boldsymbol{\xi}_k}^{\mathrm{T}}\,\widehat{\boldsymbol{\xi}_k}}{N - \frac{1}{\sigma_k^2}\operatorname{Trace}\left[\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{W}_k^{\xi}\boldsymbol{U} + \boldsymbol{D}_k^{-1}\right)^{-1}\right]}.$$

5.4.2 The algorithm

The conditional linearised models considered at iteration t are given by (5.9). Algorithm 5.1 describes the $(t+1)$–th iteration of the single–component mixed–SCGLR. It is repeated until stability of parameters is reached.

Algorithm 5.1: The single component mixed–SCGLR algorithm (generic iteration).

Step 1: Computing the component

With $\psi_{\boldsymbol{A}}(\boldsymbol{u})$ given by (5.10) and $\phi(\boldsymbol{u})$ by (5.6), set

$$\boldsymbol{u}^{[t+1]} = \underset{\boldsymbol{u}:\, \boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{u}=1}{\arg\max} \; [\phi(\boldsymbol{u})]^{s} \left[\psi_{\boldsymbol{A}}^{[t]}(\boldsymbol{u})\right]^{1-s}$$
$$\boldsymbol{f}^{[t+1]} = \boldsymbol{X}\boldsymbol{u}^{[t+1]}$$

Step 2: Henderson systems

For each $k \in \{1, \ldots, q\}$, solve the system

$$\begin{pmatrix} \boldsymbol{f}^{[t+1]^{\mathrm{T}}}\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}}\boldsymbol{f}^{[t+1]} & \boldsymbol{f}^{[t+1]^{\mathrm{T}}}\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}}\boldsymbol{A} & \boldsymbol{f}^{[t+1]^{\mathrm{T}}}\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}}\boldsymbol{U} \\ \boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}}\boldsymbol{f}^{[t+1]} & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}}\boldsymbol{A} & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}}\boldsymbol{U} \\ \boldsymbol{U}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}}\boldsymbol{f}^{[t+1]} & \boldsymbol{U}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}}\boldsymbol{A} & \boldsymbol{U}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}}\boldsymbol{U} + \boldsymbol{D}_{\boldsymbol{k}}^{[t]^{-1}} \end{pmatrix} \begin{pmatrix} \gamma_k \\ \boldsymbol{\delta}_{\boldsymbol{k}} \\ \boldsymbol{\xi}_{\boldsymbol{k}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}^{[t+1]^{\mathrm{T}}}\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}}\boldsymbol{z}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}} \\ \boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}}\boldsymbol{z}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}} \\ \boldsymbol{U}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}}\boldsymbol{z}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}} \end{pmatrix}$$

Call $\gamma_k^{[t+1]}$, $\boldsymbol{\delta}_{\boldsymbol{k}}^{[t+1]}$ and $\boldsymbol{\xi}_{\boldsymbol{k}}^{[t+1]}$ the solutions.

Step 3: Updating variance–component estimates

For each $k \in \{1, \ldots, q\}$, compute

$$\sigma_k^{2[t+1]} = \frac{\boldsymbol{\xi}_{\boldsymbol{k}}^{[t+1]^{\mathrm{T}}}\boldsymbol{\xi}_{\boldsymbol{k}}^{[t+1]}}{N - \frac{1}{\sigma_k^{2[t]}}\,\mathrm{Trace}\left[\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t]}}\boldsymbol{U} + \boldsymbol{D}_{\boldsymbol{k}}^{[t]^{-1}}\right)^{-1}\right]}$$
$$\boldsymbol{D}_{\boldsymbol{k}}^{[t+1]} = \sigma_k^{2[t+1]}\boldsymbol{I}_N$$

Step 4: Updating working variables and weighting matrices

For each $k \in \{1, \ldots, q\}$, compute

$$\boldsymbol{\eta}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t+1]}} = \boldsymbol{f}^{[t+1]}\gamma_k^{[t+1]} + \boldsymbol{A}\boldsymbol{\delta}_{\boldsymbol{k}}^{[t+1]} + \boldsymbol{U}\boldsymbol{\xi}_{\boldsymbol{k}}^{[t+1]}$$
$$\mu_{k,i}^{\xi\ [t+1]} = g_k^{-1}\left(\eta_{k,i}^{\xi\ [t+1]}\right),\ i = 1, \ldots, n$$
$$z_{k,i}^{\xi\ [t+1]} = \eta_{k,i}^{\xi\ [t+1]} + \left(y_i^k - \mu_{k,i}^{\xi\ [t+1]}\right) g_k'\left(\mu_{k,i}^{\xi\ [t+1]}\right),\ i = 1, \ldots, n$$
$$\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{[t+1]}} = \mathbf{Diag}\left(\left\{\left[g_k'\left(\mu_{k,i}^{\xi\ [t+1]}\right)\right]^2 a_{k,i}(\phi_k)\, v_k\left(\mu_{k,i}^{\xi\ [t+1]}\right)\right\}^{-1}\right)_{i=1,\ldots,n}$$

Incrementing: $t \longleftarrow t+1$

5.4.3 Extracting higher rank components

Let $\boldsymbol{F_h} = \left[\, \boldsymbol{f_1} \mid \ldots \mid \boldsymbol{f_h} \,\right]$ be the matrix of the first h components, where $h < K$. An extra component $\boldsymbol{f_{h+1}}$ must best complement the existing ones plus $\boldsymbol{A}$, i.e. $\boldsymbol{A_h} = \left[\, \boldsymbol{F_h} \mid \boldsymbol{A} \,\right]$. So $\boldsymbol{f_{h+1}}$ must be calculated using $\boldsymbol{A_h}$ as additional explanatory variables. Moreover, we must impose that $\boldsymbol{f_{h+1}}$ be orthogonal to $\boldsymbol{F_h}$, i.e.

$$\boldsymbol{F_h^\top P f_{h+1}} = \boldsymbol{0}.$$

Component $\boldsymbol{f_{h+1}} = \boldsymbol{X u_{h+1}}$ is thus obtained by solving

$$\begin{cases} \max \quad s \log\left[\phi\left(\boldsymbol{u}\right)\right] + (1-s) \log\left[\psi_{\boldsymbol{A_h}}\left(\boldsymbol{u}\right)\right] \\ \text{subject to:} \quad \boldsymbol{u}^\top \boldsymbol{M}^{-1} \boldsymbol{u} = 1 \text{ and } \boldsymbol{\Delta_h^\top u} = \boldsymbol{0}, \end{cases} \tag{5.11}$$

where $\psi_{\boldsymbol{A_h}}\left(\boldsymbol{u}\right) = \sum_{k=1}^{q} \left\| \boldsymbol{z_k^\xi} \right\|_{\boldsymbol{W_k^\xi}}^2 \cos^2_{\boldsymbol{W_k^\xi}} \left(\boldsymbol{z_k^\xi}, \text{span}\left\{\boldsymbol{Xu}, \boldsymbol{A_h}\right\} \right)$ and $\boldsymbol{\Delta_h} = \boldsymbol{X}^\top \boldsymbol{P F_h}$.

In Appendix 5.8.3, we give a simple tool to maximise, at least locally, any criterion on the unit sphere: the Projected Iterated Normed Gradient (PING) algorithm. In particular, PING solves (5.11)–type programs, which give all components of rank $h > 1$. The rank–one component is computed using the same program with $\boldsymbol{A_0} = \boldsymbol{A}$ and $\boldsymbol{\Delta_0} = \boldsymbol{0}$.

5.5 Simulation studies

Four simulation studies have been implemented to assess our method. The first one (Section 5.5.1) focuses on LMMs, so with Gaussian responses. It compares the performances of mixed–SCGLR, LMM–ridge (Eliot et al., 2011) and GLMM–LASSO (Groll and Tutz, 2014; Schelldorfer et al., 2014). The second simulation (Section 5.5.2) extends the first one to binary, Binomial and Poisson outcomes, with comparison between mixed–SCGLR and GLMM–LASSO. The third one (Section 5.5.3) assesses the performance of mixed–SCGLR on a different explanatory bundle structure and presents results on variance component estimates. The last simulation study deals with high dimensional data, and is presented in Chapter 7. All simulation studies have been performed using R (R Core Team, 2017). To compute LASSO regressions, we have used the R package **glmmLasso** (Groll, 2017). The extension of SCGLR to mixed models has been implemented in the R package **mixedSCGLR** and is available at `https://github.com/SCnext/mixedSCGLR`.

5.5.1 Simulation study with Gaussian outcomes

5.5.1.1 Data generation

To generate grouped data, we consider $N = 10$ groups and $R = 10$ observations per group (i.e. a total of $n = 100$ observations). The random effect design matrix is given by $\boldsymbol{U} = \boldsymbol{I}_N \otimes \mathbf{1}_R$. Explanatory variables $\boldsymbol{X}$ consist of three independent bundles: $\boldsymbol{X_0}$ (15 variables), $\boldsymbol{X_1}$ (10 variables) and $\boldsymbol{X_2}$ (5 variables). Each explanatory variable is normally simulated with mean 0 and variance 1. Parameter $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ tunes the level of redundancy within each bundle: the correlation matrix of bundle $\boldsymbol{X_j}$ is

$$\text{cor}\left(\boldsymbol{X_j}\right) = \tau \mathbf{1}_{p_j}\mathbf{1}_{p_j}^{\intercal} + (1-\tau)\boldsymbol{I}_{p_j},$$

where p_j is the number of variables in $\boldsymbol{X_j}$. In order to enable comparison with LASSO and ridge, and to focus on regularisation, our simulations do not involve additional explanatory variables (i.e. $\boldsymbol{A} = \mathbf{0}$). Two random responses $\boldsymbol{Y} = \left[\, \boldsymbol{y_1} \mid \boldsymbol{y_2} \,\right]$ are generated as

$$\begin{cases} \boldsymbol{y_1} = \boldsymbol{X}\boldsymbol{\beta_1} + \boldsymbol{U}\boldsymbol{\xi_1} + \boldsymbol{\varepsilon_1} \\ \boldsymbol{y_2} = \boldsymbol{X}\boldsymbol{\beta_2} + \boldsymbol{U}\boldsymbol{\xi_2} + \boldsymbol{\varepsilon_2}, \end{cases} \tag{5.12}$$

such that $\boldsymbol{y_1}$ is predicted only by bundle $\boldsymbol{X_1}$, $\boldsymbol{y_2}$ only by bundle $\boldsymbol{X_2}$, while bundle $\boldsymbol{X_0}$ plays no explanatory role. Our choice for the fixed–effect parameters is

$$\boldsymbol{\beta_1} = (\, \underbrace{0, \ldots\ldots\ldots\ldots\ldots, 0}_{\text{15 times}}, \underbrace{0.3, .., 0.3}_{\text{3 times}}, \underbrace{0.4, \ldots, 0.4}_{\text{4 times}}, \underbrace{0.5, .., 0.5}_{\text{3 times}}, \underbrace{0, \ldots, 0}_{\text{5 times}} \,)^{\intercal},$$

$$\boldsymbol{\beta_2} = (\, \underbrace{0, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, 0}_{\text{25 times}}, 0.3, 0.3, 0.4, 0.5, 0.5 \,)^{\intercal}.$$

Finally, for each $k \in \{1, 2\}$, random effect and noise vectors are simulated respectively from

$$\boldsymbol{\xi_k} \sim \mathcal{N}_N\left(\mathbf{0},\ \sigma_k^2\, \boldsymbol{I}_N\right) \text{ and } \boldsymbol{\varepsilon_k} \sim \mathcal{N}_n\left(\mathbf{0},\ \omega_k^2\, \boldsymbol{I}_n\right),$$

where $\sigma_k^2 = \omega_k^2 = 1$. For each value of τ, $B = 100$ samples are generated according to Model (5.12).

5.5.1.2 Parameter calibration

In order to compare mixed–SCGLR with the ridge and LASSO regressions, we recall the regularisation parameters required by each method. For

both LMM–ridge and GLMM–LASSO methods, a unique shrinkage parameter has to be calibrated: λ_{ridge} and λ_{LASSO} respectively. For mixed–SCGLR, three tuning parameters need to be calibrated: the number of components, K, the trade–off parameter, s, which are both regularisation parameters, and the bundle–locality parameter, l. For greater clarity, the simulation focuses on the behaviour of K and s. As recommended by Bry et al. (2013), we set $l = 4$. In case–studies, l has to be tuned to maximise the interpretability of components.

For both mixed–SCGLR and GLMM–LASSO, optimal regularisation parameters are obtained through a 5–fold cross–validation, withdrawing 2 observations from each group every time. This could be termed "leave–two–observations–out per group." The data is thus divided into five parts $\mathcal{P}_1, \ldots, \mathcal{P}_5$, each $\mathcal{P}_j$ containing 20 observations, 2 for each of the 10 groups. Let $y_{k,i}^{(b)}$ be the i–th observation of the k–th response vector in the b–th sample. Let also $\widehat{y_{k,i(-j)}^{(b)}}$ be the fit for $y_{k,i}^{(b)}$ with part $\mathcal{P}_j$ removed. The cross–validation error in the b–th sample, $E^{(b)}$, is defined as

$$E^{(b)} = \frac{1}{2} \sum_{k=1}^{2} E_k^{(b)}, \tag{5.13}$$

where

$$E_k^{(b)} = \frac{1}{5} \sum_{j=1}^{5} \sqrt{\frac{1}{20} \sum_{i \in \mathcal{P}_j} \left(y_{k,i}^{(b)} - \widehat{y_{k,i(-j)}^{(b)}} \right)^2}.$$

In the b–th sample, the optimal number of components, $K^{\star(b)}$, the trade–off parameter, $s^{\star(b)}$, and the shrinkage parameter, $\lambda_{\text{LASSO}}^{\star(b)}$, are selected to minimise the cross–validation error (5.13). We then define

$$s^\star = \frac{1}{B} \sum_{b=1}^{B} s^{\star(b)}, \quad K^\star = \text{mode}\left(\left\{ K^{\star(1)}, \ldots, K^{\star(B)} \right\} \right),$$
$$\lambda_{\text{LASSO}}^\star = \frac{1}{B} \sum_{b=1}^{B} \lambda_{\text{LASSO}}^{\star(b)}.$$

By contrast, Eliot et al. (2011) suggest to calibrate the ridge parameter at each step of their EM implementation, using the generalised cross–validation. We thus define

$$\lambda_{\text{ridge}}^\star = \frac{1}{B} \sum_{b=1}^{B} \lambda_{\text{ridge}}^{\star(b)},$$

where $\lambda_{\text{ridge}}^{\star(b)}$ denotes the average of the ridge parameter values obtained over all the iterations of the EM algorithm in the b–th sample.

Table 5.1 summarises the optimal regularisation parameters selected through cross–validation. In both ridge and LASSO, the shrinkage parameter value increases with the level of redundancy τ. Whereas for mixed-SCGLR, when τ increases, $K^\star$ decreases towards the true number of predictive variable–bundles: the greater the value of τ, the better mixed–SCGLR focuses on the structures in $\boldsymbol{X}$ that contribute to model $\boldsymbol{Y}$. Moreover, when τ increases, the trade–off parameter $s^\star$ increases, meaning that regularisation requires a greater importance of the structural relevance relative to the goodness–of–fit.

Table 5.1 – *Optimal regularisation parameter values obtained through cross–validation over* 100 *simulations.*

	GLMM–LASSO	**LMM–ridge**	**mixed–SCGLR**	
	shrinkage parameter $\lambda^\star_{\text{LASSO}}$	shrinkage parameter $\lambda^\star_{\text{ridge}}$	number of components $K^\star$	trade–off parameter $s^\star$
$\tau = 0.1$	65	24	15	0.50
$\tau = 0.3$	92	54	5	0.58
$\tau = 0.5$	124	73	3	0.70
$\tau = 0.7$	163	78	3	0.73
$\tau = 0.9$	175	85	2	0.80

5.5.1.3 Comparison of estimate accuracies

Once tuning parameters are obtained, we focus on the fixed–effect estimates' accuracy. Since the response–vectors $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$ are normally distributed and have comparable orders of magnitude, the fixed–effect relative errors are on the same scale. Then we consider a risk–averse comparison criterion called "Mean Upper Relative Squared Error" (MURSE) defined as

$$\text{MURSE}\left(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}\right) = \frac{1}{B}\sum_{b=1}^{B} \max\left\{ \frac{\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}^{(b)} - \boldsymbol{\beta_1}\right\|^2}{\left\|\boldsymbol{\beta_1}\right\|^2}, \frac{\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}^{(b)} - \boldsymbol{\beta_2}\right\|^2}{\left\|\boldsymbol{\beta_2}\right\|^2} \right\},$$

where $\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}^{(b)}$ is the estimate of $\boldsymbol{\beta_k}$ associated with sample b.

The MURSE values for mixed–SCGLR, LMM–ridge and GLMM–LASSO are presented in Table 5.2. The LMM results obtained without regularisation are also presented. They were computed using the R package **lme4** (Bates et al., 2015). In the latter case, relative errors increase dramatically with τ. Those of ridge and LASSO increase less drastically (but increase anyway) because these methods suffer from the high correlations among the explanatory

variables. Except for $\tau = 0.1$, mixed–SCGLR provides the most accurate fixed effect estimates. Indeed, if there are no real bundles in $\boldsymbol{X}$ ($\tau \simeq 0$), searching for structures in $\boldsymbol{X}$ may lead mixed–SCGLR to be slightly less accurate. Conversely, mixed–SCGLR takes advantage of the high correlations among the explanatory variables: the stronger the structures (high τ), the more efficient the method.

Table 5.2 – ***Mean Upper Relative Squared Error (MURSE) values associated with the optimal parameter values.***

	LMM (no regularisation)	**GLMM–LASSO**	**LMM–ridge**	**mixed–SCGLR**
$\tau = 0.1$	0.12	0.05	0.08	0.12
$\tau = 0.3$	0.33	0.12	0.13	0.10
$\tau = 0.5$	0.61	0.20	0.16	0.07
$\tau = 0.7$	1.32	0.25	0.20	0.06
$\tau = 0.9$	4.62	0.26	0.31	0.05

5.5.1.4 Model interpretation

This section aims at highlighting the power of mixed–SCGLR for model interpretation. Figure 5.2 presents an example of the first component planes obtained for $\tau = 0.5$, with associated optimal parameter values $s^\star = 0.7$ and $K^\star = 3$. We still impose $l = 4$. The first two components obtained are the ones which explain the responses. It clearly appears that $\boldsymbol{y_1}$ is explained by bundle $\boldsymbol{X_1}$ and $\boldsymbol{y_2}$ by $\boldsymbol{X_2}$. Interestingly, although bundle $\boldsymbol{X_0}$ is the one with maximum inertia (26.83%), it appears only along the third component, for having no explanatory part.

5.5.2 Additional simulations involving non–Gaussian outcomes

5.5.2.1 Binary and Poisson outcomes

This section aims at assessing our method in the case of Bernoulli ($\mathcal{B}$) and Poisson ($\mathcal{P}$) distributions of responses. We still consider $N = 10$ groups and $R = 10$ observations per group. Design matrices $\boldsymbol{X}$ and $\boldsymbol{U}$, as well as the values of $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, σ_1^2 and σ_2^2, are still defined as in Section 5.5.1.1. The group variance components are given by $\varsigma_1^2 = 0.1\sigma_1^2$ and $\varsigma_2^2 = \sigma_2^2$ so that for each $k \in \{1, 2\}$, $\widetilde{\boldsymbol{\xi_k}} \sim \mathcal{N}_N(\boldsymbol{0},\, \varsigma_k^2\, \boldsymbol{I}_N)$. Given $\widetilde{\boldsymbol{\xi_1}}$ and $\widetilde{\boldsymbol{\xi_2}}$, $\boldsymbol{Y} = \left[\, \boldsymbol{y_1} \mid \boldsymbol{y_2} \,\right]$ is then

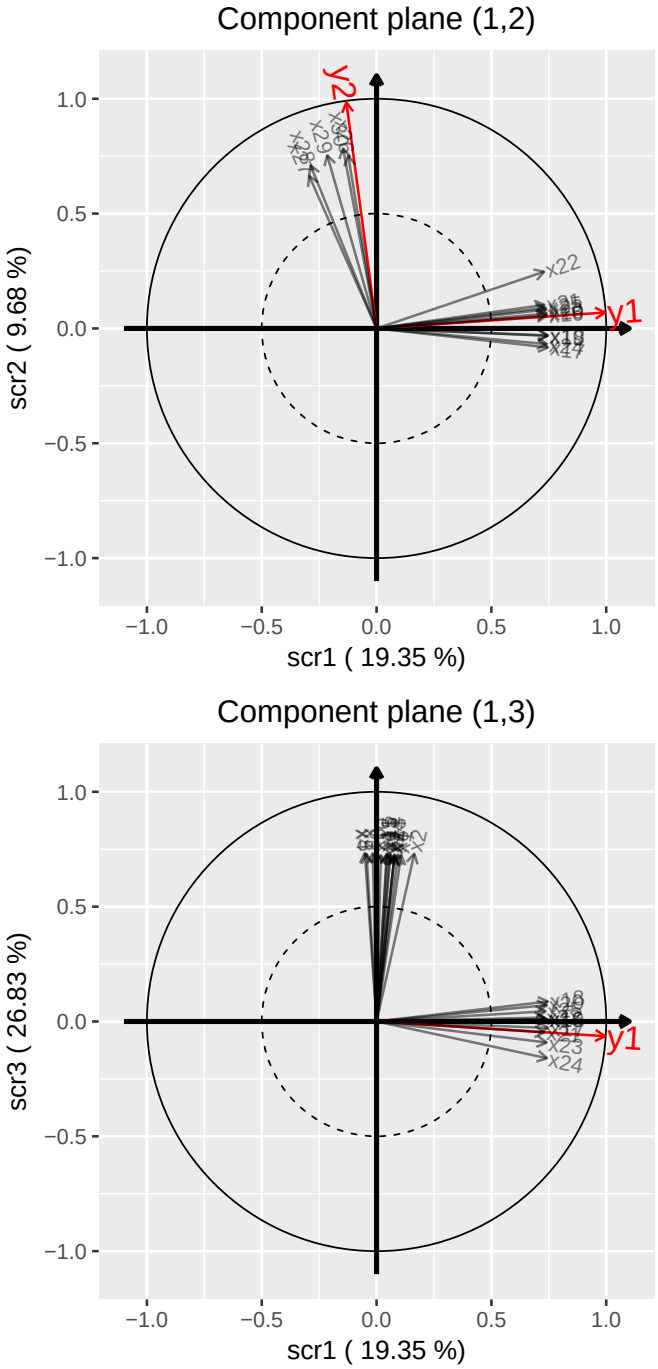


Figure 5.2 – *Component planes (1,2) and (1,3) given by mixed–SCGLR on simulated Gaussian data.* *The black arrows represent the explanatory variables. The red ones represent the projection of the $\boldsymbol{X}$–part of the linear predictors associated with $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$. The percentage of inertia captured by each component is given in parentheses. For an easier model interpretation, our method hides the least relevant explanatory variables on each component plane with a simple thresholding. Here, we hide all the predictors having cosine with the component plane lower than* 0.5.

simulated as

$$\begin{cases} \boldsymbol{y_1} \sim \mathcal{B}\left(\boldsymbol{p} = \text{logit}^{-1}\left[\boldsymbol{X\theta_1} + \boldsymbol{U\widetilde{\xi}_1}\right]\right) \\ \boldsymbol{y_2} \sim \mathcal{P}\left(\boldsymbol{\lambda} = \exp\left[\boldsymbol{X\theta_2} + \boldsymbol{U\widetilde{\xi}_2}\right]\right), \end{cases} \tag{5.14}$$

where $\boldsymbol{\theta_1} = 0.1\boldsymbol{\beta_1}$ and $\boldsymbol{\theta_2} = \boldsymbol{\beta_2}$.

Again, for each value of τ, $B = 100$ samples are generated according to Model (5.14). As in Section 5.5.1.2, tuning parameters are calibrated so as to minimise the cross–validation error (5.13). However, since $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$ do not have the same range of values, the prediction errors have to be standardised. The cross–validation error for response $\boldsymbol{y_k}$ in the b–th sample is now given by

$$E_k^{(b)} = \frac{1}{5}\sum_{j=1}^{5}\sqrt{\frac{1}{20}\sum_{i\in\mathcal{P}_j}\frac{\left(y_{k,i}^{(b)} - \widehat{y_{k,i(-j)}^{(b)}}\right)^2}{\widehat{\text{var}}\left(\widehat{y_{k,i(-j)}^{(b)}}\right)}}.$$

In the same way, as the response vectors come from different distributions and have different orders of magnitude, the fixed–effect relative errors are not comparable. Unlike in Section 5.5.1.3, to compare mixed–SCGLR with GLMM–LASSO and classical GLMM (without regularisation), we thus use the Mean Relative Squared Error (MRSE) defined as

$$\text{MRSE}\left(\boldsymbol{\theta_k}\right) = \frac{1}{B}\sum_{b=1}^{B}\frac{\left\|\widehat{\boldsymbol{\theta}}_{\boldsymbol{k}}^{(b)} - \boldsymbol{\theta_k}\right\|^2}{\left\|\boldsymbol{\theta_k}\right\|^2},\ k \in \{1,2\},$$

where $\widehat{\boldsymbol{\theta}}_{\boldsymbol{k}}^{(b)}$ is the estimate of $\boldsymbol{\theta_k}$ from the b–th sample.

MRSE values for the GLMM, mixed–SCGLR and GLMM–LASSO are presented in Table 5.3. For all methods, estimating a Bernoulli model is obviously a more challenging task than estimating a Poisson model. Regardless of the level of redundancy, τ, both mixed–SCGLR and GLMM–LASSO outperform classical GLMM estimation. Compared with the Gaussian case (Section 5.5.1.3), the results deteriorate but overall, the same behaviours are observed.

- For $\tau = 0.1$, fixed–effect estimates provided by mixed–SCGLR are less accurate than those provided by GLMM–LASSO. In this case, GLMM–LASSO has indeed a double advantage. First, many $\theta_{k,j}$'s are true zeros. Unlike mixed–SCGLR, GLMM–LASSO often shrinks their estimates to exactly zero. Second, since the level of redundancy is low, GLMM–LASSO also provides accurate coefficient estimates of active variables.

- By contrast, for $\tau \geqslant 0.3$, mixed–SCGLR takes advantage of redundancies within the explanatory variables. Thus, mixed-SCGLR outperforms GLMM–LASSO in this case, *despite the sparse structure* of the $\boldsymbol{\theta}_k$'s.

Table 5.3 – *Mean Relative Squared Error (MRSE) values obtained with Bernoulli and Poisson responses.*

	GLMM (no regularisation)		GLMM–LASSO		mixed–SCGLR	
	Bernoulli	Poisson	Bernoulli	Poisson	Bernoulli	Poisson
$\tau = 0.1$	316.48	0.54	8.61	0.30	14.71	0.46
$\tau = 0.3$	398.78	0.64	9.23	0.36	7.21	0.21
$\tau = 0.5$	576.68	0.87	14.48	0.44	2.01	0.09
$\tau = 0.7$	886.04	1.28	17.37	0.47	1.50	0.07
$\tau = 0.9$	2840.10	3.72	17.24	0.59	1.31	0.05

Even if the response variables are not Gaussian, the power of mixed–SCGLR for model interpretation is preserved. The component planes still reveal that $\boldsymbol{y_1}$ is explained by bundle $\boldsymbol{X_1}$ and $\boldsymbol{y_2}$ by $\boldsymbol{X_2}$. Figure 5.3 illustrates what may happen when the level of redundancy is very high ($\tau = 0.7$). Since the explanatory variables are highly correlated, the mixed–SCGLR regularisation requires that the structural relevance be given a heavy weight with respect to the goodness–of–fit, which leads to a trade–off parameter s close to 1 ($s = 0.9$ here). Having the greatest structural strength, the nuisance bundle is captured by the second component despite its lack of explanatory power. This is sometimes the price to pay for the trade–off. Indeed, a higher within–bundle correlation requires a stronger regularisation, hence a higher value of s. In return, dimensions with a higher SR but lower GoF may be tracked first. In our example, the second explanatory bundle is captured by the third component, so that the predictive dimensions are accurately represented in component plane $(1, 3)$.

5.5.2.2 Binomial and Poisson outcomes

In this section, we simply extend the simulation scheme presented in Section 5.5.2.1 to binomial and Poisson outcomes. We maintain design matrices $\boldsymbol{X}$ and $\boldsymbol{U}$ as defined in Section 5.5.1.1. Fixed–effect parameters $\boldsymbol{\theta}_k$'s and random–effect vectors $\widetilde{\boldsymbol{\xi}}_k$'s are defined in Section 5.5.2.1. Given $\widetilde{\boldsymbol{\xi}}_1$ and $\widetilde{\boldsymbol{\xi}}_2$, we then simulate $\boldsymbol{Y} = \left[\, \boldsymbol{y_1} \mid \boldsymbol{y_2} \,\right]$ as

$$\begin{cases} \boldsymbol{y_1} \sim \mathcal{B}\text{in}\left(\textbf{trials} = 50\,\mathbf{1}_n,\ \boldsymbol{p} = \text{logit}^{-1}\left[\boldsymbol{X\theta_1} + \boldsymbol{U}\widetilde{\boldsymbol{\xi}}_1\right]\right) \\ \boldsymbol{y_2} \sim \mathcal{P}\left(\boldsymbol{\lambda} = \exp\left[\boldsymbol{X\theta_2} + \boldsymbol{U}\widetilde{\boldsymbol{\xi}}_2\right]\right). \end{cases}$$

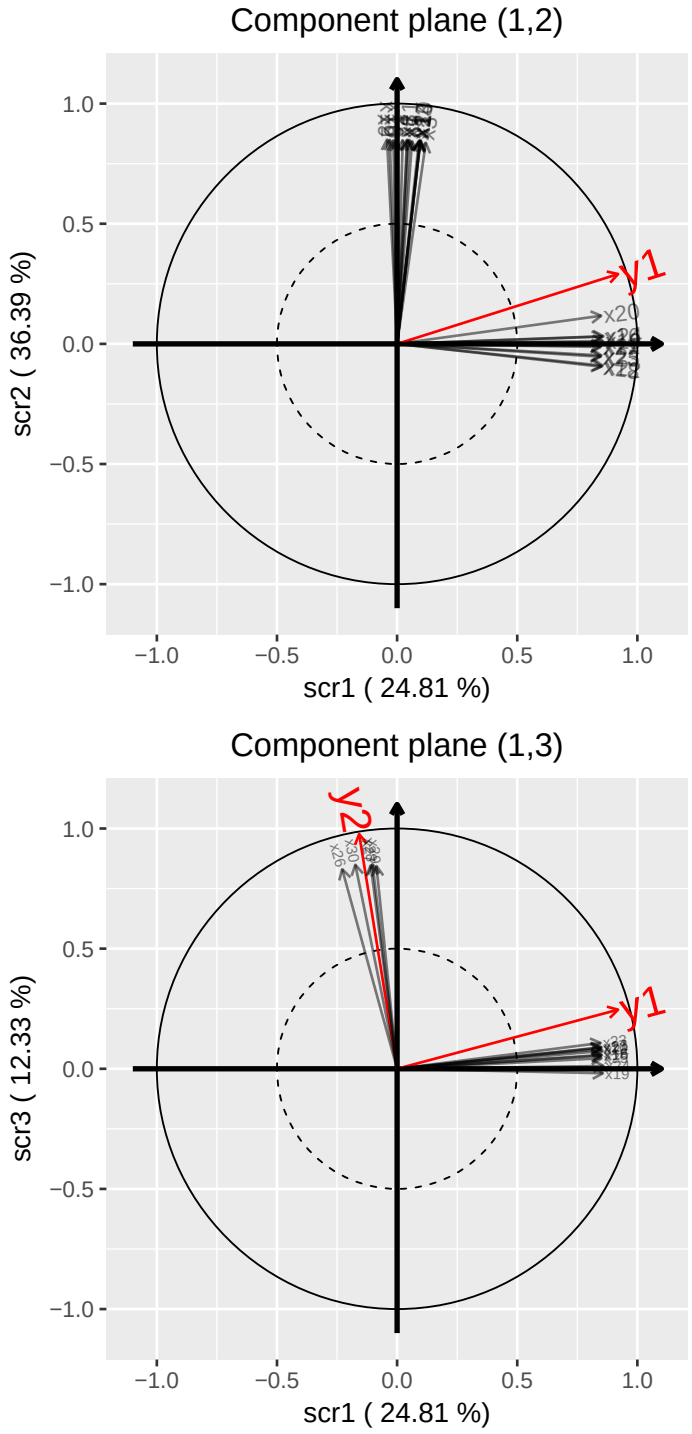


Figure 5.3 – *Example of component planes given by mixed–SCGLR in the Bernoulli–Poisson case.* *In this example, the redundancy level is set to* $\tau = 0.7$ *and the optimal parameter triplet selected through cross–validation is* $(K, s, l) = (3, 0.9, 4)$. *As previously, only the variables having cosine greater than* 0.5 *with the component plane are represented.*

Table 5.4 gives the Mean Relative Squared Error (MRSE) values for $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ obtained on 100 samples for each value of τ. For the Poisson distribution, the results are essentially identical to those in the previous simulation: mixed–SCGLR outperforms GLMM–LASSO except for $\tau = 0.1$. As for the binomial distribution, the regularisation provided by mixed–SCGLR improves the results obtained without regularisation, regardless of the level τ of redundancy within the explanatory variables. Unsurprisingly, the errors are much smaller than in the binary case.

Table 5.4 – *Mean Relative Squared Error (MRSE) values obtained with binomial and Poisson distributions.* *The* R *package* ***glmmLasso*** *(Groll, 2017) does not handle binomial outcomes but only Bernoulli ones, which precludes comparison in this case.*

	GLMM (no regularisation)		**GLMM–LASSO**		**mixed–SCGLR**	
	Binomial	Poisson	Binomial	Poisson	Binomial	Poisson
$\tau = 0.1$	2.31	0.50	NA	0.31	0.51	0.45
$\tau = 0.3$	3.07	0.60	NA	0.33	0.28	0.18
$\tau = 0.5$	3.93	0.75	NA	0.39	0.15	0.09
$\tau = 0.7$	6.50	1.07	NA	0.40	0.10	0.07
$\tau = 0.9$	19.29	2.71	NA	0.42	0.07	0.05

Figure 5.4 presents an example of the first component planes output by mixed–SCGLR in the binomial/Poisson case with a rather low level of redundancy ($\tau = 0.3$ here). The component planes clearly reveal that $\boldsymbol{y_1}$ is explained by bundle $\boldsymbol{X_1}$ and $\boldsymbol{y_2}$ by $\boldsymbol{X_2}$. As in the Gaussian simulation, predictive bundles $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are captured respectively by the first and the second components. The third component aligns on noise bundle $\boldsymbol{X_0}$, despite its high inertia.

5.5.3 Simulations with a more complex variable–structure

This simulation study tests mixed-SCGLR on a slightly more complex bundle structure. Results concerning variance component estimates are also presented.

We consider a fixed–effect design matrix $\boldsymbol{X}_{n\times p}$ partitioned into 3 blocks $\boldsymbol{\mathcal{G}_1}$, $\boldsymbol{\mathcal{G}_2}$ and $\boldsymbol{\mathcal{G}_3}$. Block $\boldsymbol{\mathcal{G}_1}$ contains 10 predictive explanatory variables structured about a latent variable $\boldsymbol{\varphi_1} \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2_{\text{LV}}\boldsymbol{I}_n\right)$. Thus for each $j \in \{1, \ldots, 10\}$, $\boldsymbol{x_j} = \boldsymbol{\varphi_1} + \boldsymbol{\varepsilon_j}$, where $\boldsymbol{\varepsilon_j} \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2_{\text{noise}}\boldsymbol{I}_n\right)$ such that $\sigma^2_{\text{LV}} + \sigma^2_{\text{noise}} = 1$. The correlation within $\boldsymbol{\mathcal{G}_1}$ is tuned by signal to noise (StN) ratio $\sigma^2_{\text{LV}}/\sigma^2_{\text{noise}}$ (chosen in $\left\{\frac{1}{3}, 1, 3\right\}$ in practice). $\boldsymbol{\mathcal{G}_2}$ contains a single predictive variable $\boldsymbol{\varphi_2} = \boldsymbol{x_{11}} \sim \mathcal{N}_n\left(\mathbf{0}, \boldsymbol{I}_n\right)$. $\boldsymbol{\mathcal{G}_3}$ contains 20 unstructured noise variables: for each

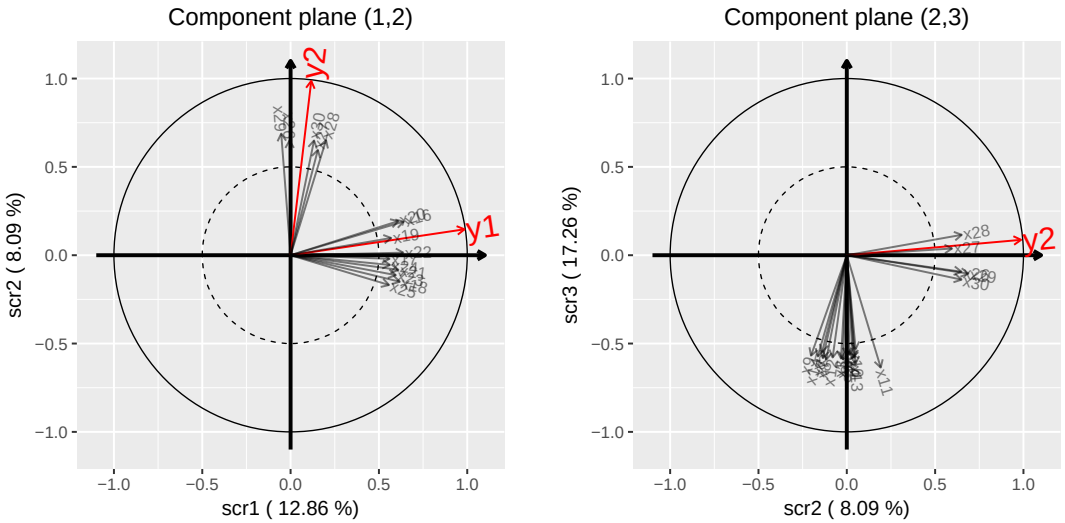


Figure 5.4 – *Example of component planes given by mixed–SCGLR in the binomial–Poisson case.* *In this example, the redundancy level is set to $\tau = 0.3$ and the optimal parameter triplet selected through cross–validation is $(K, s, l) = (3, 0.5, 2)$. Again, only the variables having cosine greater than 0.5 with the component plane are represented.*

$j \in \{12, \ldots, 31\}$, $\boldsymbol{x_j} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{I}_n)$. For each $k \in \{1, 2, 3\}$, random–effect vectors are simulated as $\boldsymbol{\xi_k} \overset{\text{ind.}}{\sim} \mathcal{N}_N(\mathbf{0}, \sigma_k^2 \boldsymbol{I}_N)$. Given $\boldsymbol{\xi_1}, \boldsymbol{\xi_2}, \boldsymbol{\xi_3}$, we simulate 3 responses having different distributions, $\boldsymbol{Y} = \left[\, \boldsymbol{y_1} \mid \boldsymbol{y_2} \mid \boldsymbol{y_3} \,\right]$, as

$$\begin{cases} \boldsymbol{y_1} \sim \mathcal{N}_n\Big(\boldsymbol{\mu} = \alpha_1 \boldsymbol{\varphi_1} + \boldsymbol{U}\boldsymbol{\xi_1},\ \boldsymbol{\Sigma} = \boldsymbol{I}_n\Big) \\ \boldsymbol{y_2} \sim \mathcal{P}\Big(\boldsymbol{\lambda} = \exp\Big[\alpha_2 \boldsymbol{\varphi_2} + \boldsymbol{U}\boldsymbol{\xi_2}\Big]\Big) \\ \boldsymbol{y_3} \sim \mathcal{B}\text{in}\Big(\textbf{trials} = 25\,\mathbf{1}_n,\ \boldsymbol{p} = \text{logit}^{-1}\Big[\alpha_3\,(\boldsymbol{\varphi_1} + \boldsymbol{\varphi_2}) + \boldsymbol{U}\boldsymbol{\xi_3}\Big]\Big). \end{cases}$$

In our simulations, we set $\alpha_1 = \sigma_1^2 = 2$, $\alpha_2 = \sigma_2^2 = 1$ and $\alpha_3 = \sigma_3^2 = 0.5$.

We consider in turn $N = 10$ and $N = 50$ groups, and $R = 10$ observations per group ($n = 100$ and $n = 500$ observations in total). $B = 100$ samples are generated for each pair of values (N, StN). The main goal of the study is to assess the ability of mixed–SCGLR to track down both latent predictive variables $\boldsymbol{\varphi_1}$ and $\boldsymbol{\varphi_2}$. For $j = 1$ and 2, we then define

$$\text{cor}_j = \frac{1}{B} \sum_{b=1}^{B} \left| \text{cor}\left(\boldsymbol{\varphi}_j, \boldsymbol{f}_j^{(b)}\right) \right|,$$

where $\boldsymbol{f}_j^{(b)}$ is the component most correlated with $\boldsymbol{\varphi}_j$ issued from mixed–SCGLR in the b–th sample. Consistency of fixed–effect estimates is assessed

through criteria err_1, err_2 and err_3 defined by

$$\text{err}_j = \frac{1}{B}\sum_{b=1}^{B} \frac{\left\|\alpha_j \boldsymbol{\varphi_j} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(b)}\right\|^2}{\left\|\alpha_j \boldsymbol{\varphi_j}\right\|^2},\ j \in \{1,2\}$$

$$\text{err}_3 = \frac{1}{B}\sum_{b=1}^{B} \frac{\left\|\alpha_3\left(\boldsymbol{\varphi_1}+\boldsymbol{\varphi_2}\right) - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}^{(b)}\right\|^2}{\left\|\alpha_3\left(\boldsymbol{\varphi_1}+\boldsymbol{\varphi_2}\right)\right\|^2},$$

where $\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(b)}$ is the fixed–effect estimate related to response $\boldsymbol{y_j}$ associated with sample b.

Table 5.5 – ***Summary of cor_j and err_j values, and presentation of biases and standard errors of estimated variance components.***

	$N=10, R=10$ $(n=100)$			$N=50, R=10$ $(n=500)$		
$\sigma^2_{\text{LV}}/\sigma^2_{\text{noise}}$	$\frac{1}{3}$	1	3	$\frac{1}{3}$	1	3
cor_1	0.71	0.91	0.96	0.75	0.92	0.96
cor_2	0.93	0.94	0.94	0.97	0.98	0.98
err_1	0.47	0.15	0.06	0.38	0.13	0.05
err_2	0.12	0.12	0.12	0.05	0.04	0.04
err_3	0.19	0.14	0.11	0.11	0.07	0.04
bias $\left(\widehat{\sigma^2_1}\right)$	−0.02	−0.01	0.00	0.02	0.00	−0.02
sd $\left(\widehat{\sigma^2_1}\right)$	1.04	1.05	1.06	0.41	0.40	0.39
bias $\left(\widehat{\sigma^2_2}\right)$	−0.11	−0.08	−0.06	−0.06	−0.06	−0.06
sd $\left(\widehat{\sigma^2_2}\right)$	0.50	0.51	0.52	0.21	0.21	0.21
bias $\left(\widehat{\sigma^2_3}\right)$	−0.03	−0.04	−0.04	−0.02	−0.02	−0.02
sd $\left(\widehat{\sigma^2_3}\right)$	0.22	0.21	0.21	0.11	0.11	0.11

Table 5.5 summarises the values of the afore–defined criteria and presents biases and standard errors of variance components estimates. For a given value of N, cor_1 increases towards 1 with ratio $\sigma^2_{\text{LV}}/\sigma^2_{\text{noise}}$: the tighter the block $\boldsymbol{\mathcal{G}_1}$ is structured about its latent variable, the better mixed–SCGLR can reconstruct it. The associated criterion err_1 then naturally decreases towards 0. On the other hand, cor_2 and err_2 are very stable, which proves that mixed–SCGLR is able to detect an isolated predictive variable among a large number of irrelevant others. As err_3 depends on how accurately mixed–SCGLR recovers $\boldsymbol{\varphi_1}$ and $\boldsymbol{\varphi_2}$, it slightly decreases when the StN ratio increases. Both variance component biases and standard errors seem rather stable regardless of the value of StN. Finally, when N increases, all the cor_j's increase towards 1 and all the err_j's decrease towards 0. As far as variance component estimates are con-

cerned, the biases are getting slightly closer to 0 and the standard errors decrease significantly.

Model interpretation is revealed by Figures 5.5 in the case of $N = 10$ groups and $R = 10$ observations per group. The first component aligns with block $\boldsymbol{\mathcal{G}_1}$ which alone explains response $\boldsymbol{y_1}$. The second aligns with $\boldsymbol{\mathcal{G}_2}$ (containing single explanatory variable $\boldsymbol{x_{11}}$) which alone explains $\boldsymbol{y_2}$. Finally, note that the projection of the $\boldsymbol{X}$–part of the linear predictor related to $\boldsymbol{y_3}$ is well represented on component plane $(1, 2)$. This indicates that $\boldsymbol{y_3}$ is explained jointly by $\boldsymbol{\mathcal{G}_1}$ and $\boldsymbol{\mathcal{G}_2}$.

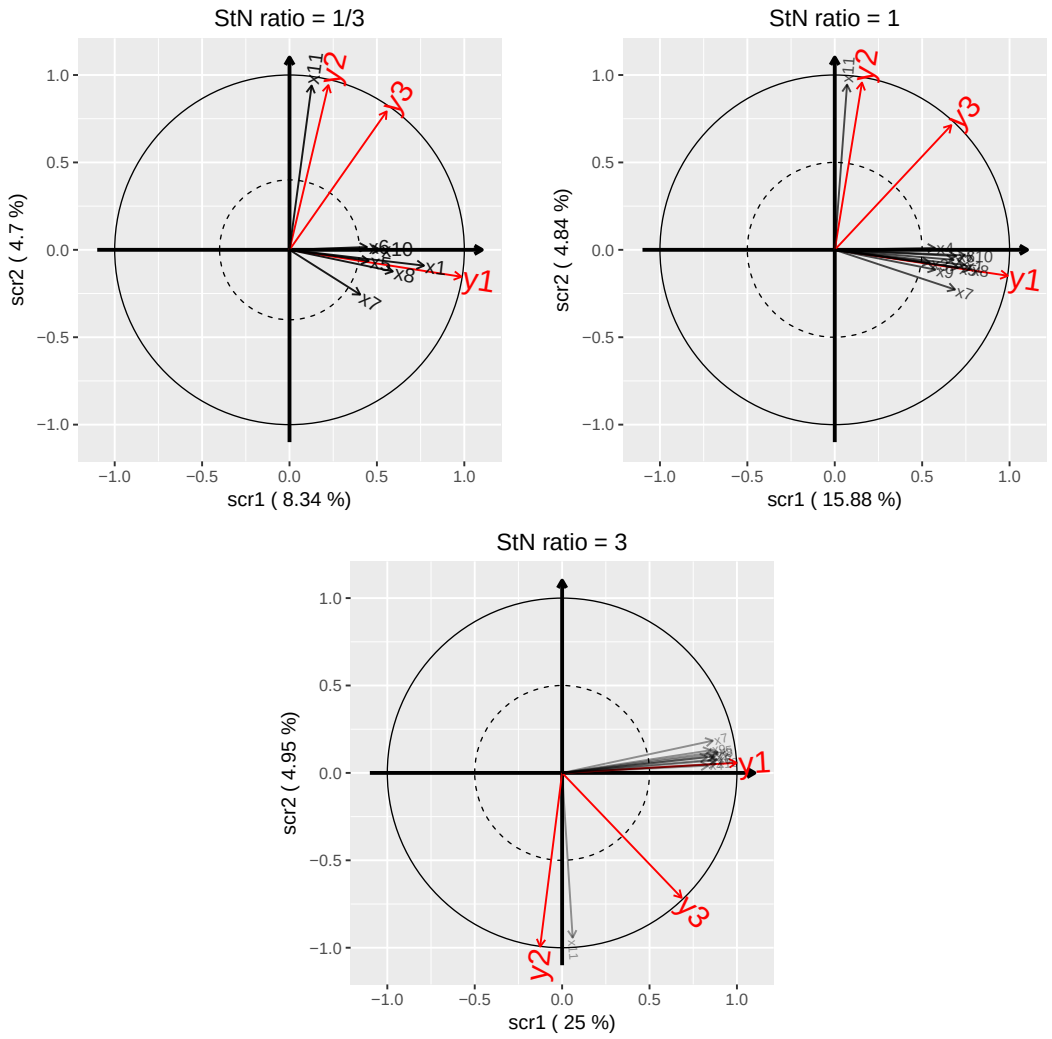


Figure 5.5 – ***Examples of the first two–component planes given by mixed–SCGLR when*** $\boldsymbol{\sigma^2_{LV}/\sigma^2_{noise} = 1/3}$ ***(top left),*** $\boldsymbol{\sigma^2_{LV}/\sigma^2_{noise} = 1}$ ***(top right), and*** $\boldsymbol{\sigma^2_{LV}/\sigma^2_{noise} = 3}$ ***(bottom).*** *When StN ratio* $= 1/3$ *(resp. StN ratio* $\in \{1, 3\}$*), only the variables having cosine greater than* 0.4 *(resp.* 0.5*) with component plane* $(1, 2)$ *are represented.*

5.6 An application to forest ecology data

5.6.1 Data description

The present study is based on the *Genus* dataset of the CoForChange project (see http://www.coforchange.eu). The subsample we consider gives the abundance of 8 common tree genera on 2615 Congo Basin land plots. These plots are grouped into 22 forest concessions. To predict abundances, we have 56 environmental variables, plus 2 explanatory variables which code geology and anthropogenic interference. $\boldsymbol{X}$ consists of all environmental variables which are:

- 29 physical factors linked to topography, rainfall or soil moisture,
- 25 photosynthesis activity indicators (the Enhanced Vegetation Indices, EVI, the Near–InfraRed indices, NIR, and the Mid–InfraRed indices, MIR),
- 2 indicators which describe the tree height.

Physical factors are many and redundant: monthly rainfalls are highly correlated, and so are photosynthesis activity indicators. The correlation heatmap indicating correlations among all environmental variables is presented on Figure 5.6. By contrast, geology and anthropogenic interference are weakly correlated and interesting per se. These variables are then considered as additional explanatory variables and included in matrix $\boldsymbol{A}$.

5.6.2 Model and parameter calibration

Abundances of species given in *Genus* are count data. For each $k \in \{1, \ldots, 8\}$, we consider a Poisson regression with $\log$ link

$$\boldsymbol{y_k} \sim \mathcal{P}\left(\boldsymbol{\lambda} = \exp\left[\sum_{j=1}^{K} (\boldsymbol{X}\boldsymbol{u_j})\,\gamma_{k,j} + \boldsymbol{A}\boldsymbol{\delta_k} + \boldsymbol{U}\boldsymbol{\xi_k}\right]\right), \tag{5.15}$$

where $\boldsymbol{\xi_k}$ is the 22–level random–effect vector used to model the dependence between the observations of $\boldsymbol{y_k}$ within concessions. The first cross–validations we performed — with different fixed values of parameters s and l — indicated that four components were sufficient to capture most of the information in $\boldsymbol{X}$ needed to model and predict responses. We therefore keep $K^\star = 4$. The optimal values of trade–off and locality parameter $s^\star$ and $l^\star$ are then determined through another cross–validation.

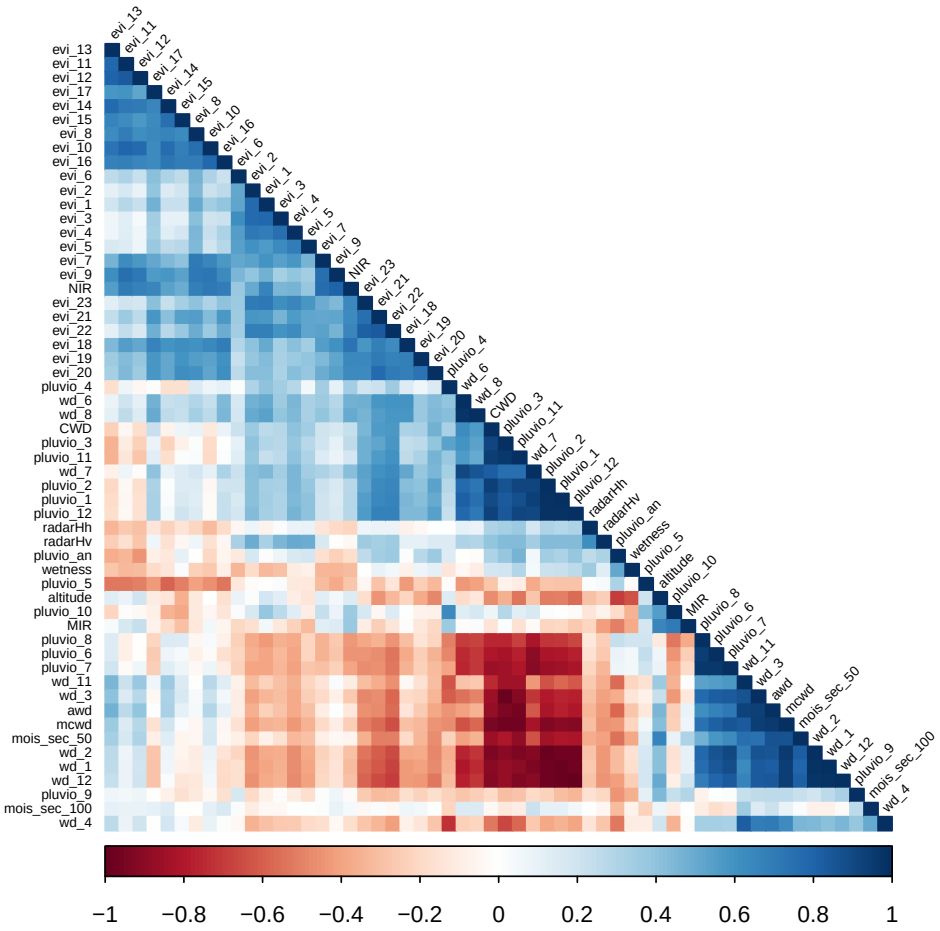


Figure 5.6 – *Correlation heatmap of Genus explanatory variables.* *The blue color corresponds to a correlation close to* 1*, the red color corresponds to a correlation close to* -1 *and the white color corresponds to a correlation close to* 0*. It clearly appears several subsets of highly positively correlated variables and a subset of highly negatively correlated variables.*

Using the same procedure and notations as in Section 5.5.1.2, the data is divided into five parts $\mathcal{P}_1, \ldots, \mathcal{P}_5$. Let n_j be the size of $\mathcal{P}_j$. The cross–validation error is defined as

$$E = \frac{1}{8} \sum_{k=1}^{8} E_k,$$

where

$$E_k = \frac{1}{5} \sum_{j=1}^{5} \sqrt{\frac{1}{n_j} \sum_{i \in \mathcal{P}_j} \frac{\left(y_{k,i} - \widehat{y_{k,i(-j)}}\right)^2}{\widehat{\text{var}}\left(\widehat{y_{k,i(-j)}}\right)}}. \tag{5.16}$$

On Figure 5.7, we plot the errors E for parameter pairs $(s, l) \in \mathcal{E}_s \times \mathcal{E}_l$,

where

$$\mathcal{E}_s = \{0.025, 0.1, 0.2, \ldots, 1\}$$
$$\mathcal{E}_l = \{1, 2, \ldots, 10, 12, 14, \ldots, 30, 35, 40, 45, 50\}\,.$$

Parameter grid $\mathcal{E}_s \times \mathcal{E}_l$ therefore contains 264 pair values. Selecting the best parameter pair from $\mathcal{E}_s \times \mathcal{E}_l$ through a 5–fold cross–validation requires a computation time of about 65 minutes (parallel computing on 6 CPU cores, Intel Core i7–6700HQ, 2.6GHz). It should be noted that there is a risk of non–convergence when the trade–off parameter s is too close to 0. Indeed, if we consider no structural information in $\boldsymbol{X}$ (s exactly equals 0), mixed–SCGLR merely performs classical GLMM estimation and does not converge for this data set. When $s = 0.025$, our algorithm converges but leads to fairly unstable estimates and high cross–validation errors because regularisation is then very weak. By contrast, the components calculated with $s \in \{0.5, 0.6, \ldots, 1\}$ are close to principal components. The associated errors are therefore stable in most cases, but rather high. Finally, $s \in \{0.1, \ldots, 0.4\}$ leads to the lowest cross–validation errors, but only for $l \leqslant 10$. Indeed, when s is not too high, mixed–SCGLR may focus on the most predictive structures of $\boldsymbol{X}$. However, parameter l must not exceed a certain value, in order to avoid being drawn towards too local variable–bundles. As can be seen, choosing $(s^\star, l^\star) = (0.1, 10)$ minimises the cross–validation error.

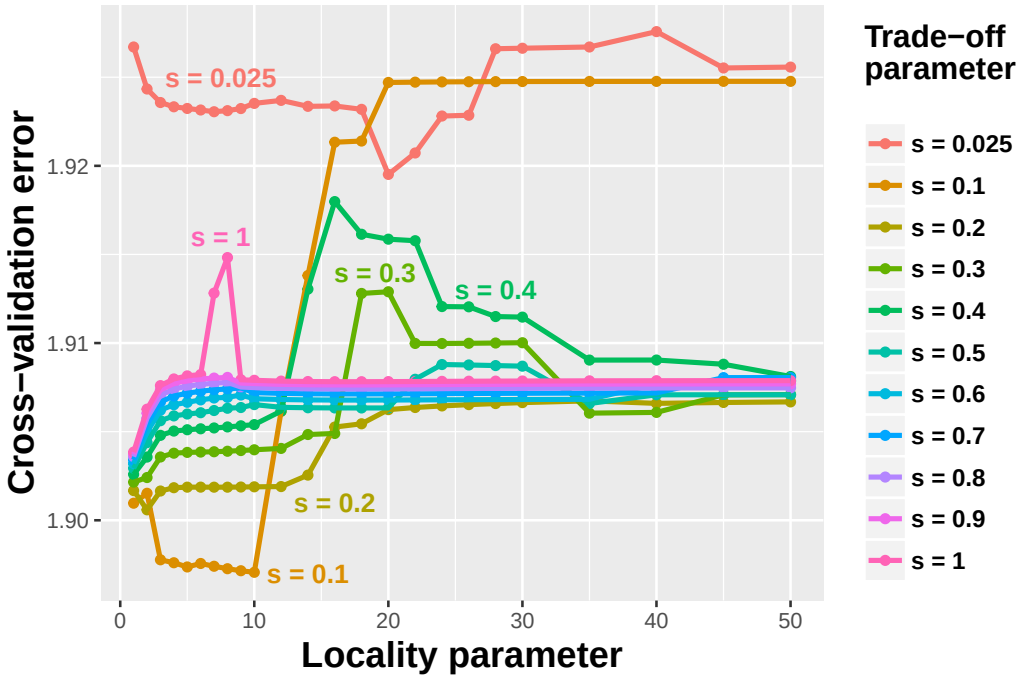


Figure 5.7 – *Behaviour of the cross–validation error E.* *For each trade–off parameter value in $\{0.025, 0.1, 0.2, \ldots, 0.9, 1\}$, we plot the cross–validation error against bundle–locality parameter $l \in [1, 50]$.*

5.6.3 Prediction quality and interpretation results

This part evaluates the benefits obtained by taking within–group dependence into account. The predictions we get with mixed–SCGLR and with the initial version of SCGLR are compared with respect to the cross–validation criterion given by (5.16). Table 5.6 summarises the E_k's for both SCGLR and mixed–SCGLR methods. Parameter value triplet $(K^\star, s^\star, l^\star) = (4, 0.1, 10)$ is optimal for both methods. For each $k \in \{1, \ldots, 8\}$, mixed–SCGLR gives a lower cross–validation error than SCGLR: taking into account the within–group dependence has clearly improved prediction performances.

Table 5.6 – *Cross–validation errors for each response variable.*

	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8
SCGLR	1.32	2.46	3.27	1.43	2.56	1.28	1.54	3.44
mixed–SCGLR	1.24	1.95	2.92	1.32	2.27	1.15	1.31	3.01

Moreover, mixed–SCGLR enables to correctly reconstitute observed abundance maps, as illustrated on Figure 5.8.

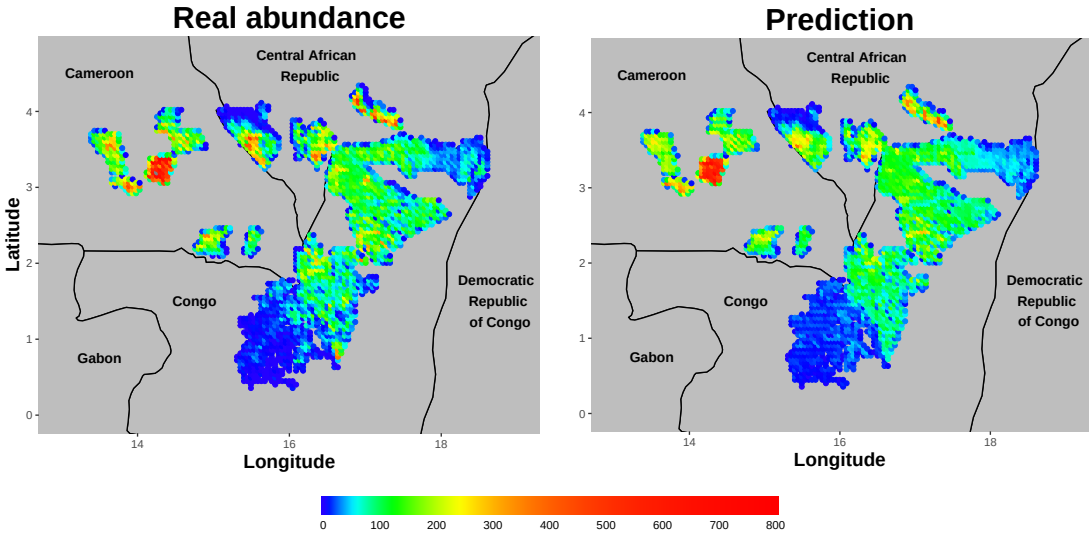


Figure 5.8 – *Abundance maps issued from mixed–SCGLR.* *The plots respectively show real abundance (left) and associated conditional predictions (right) of the tree species number 8. Each point represents a land plot (2615 in total).*

As has been seen in Section 5.5.1.4, mixed–SCGLR allows an easy interpretation of the model through the decomposition of linear predictors on interpretable components. Figure 5.9 shows the first two component planes resulting from mixed–SCGLR on real data *Genus*.

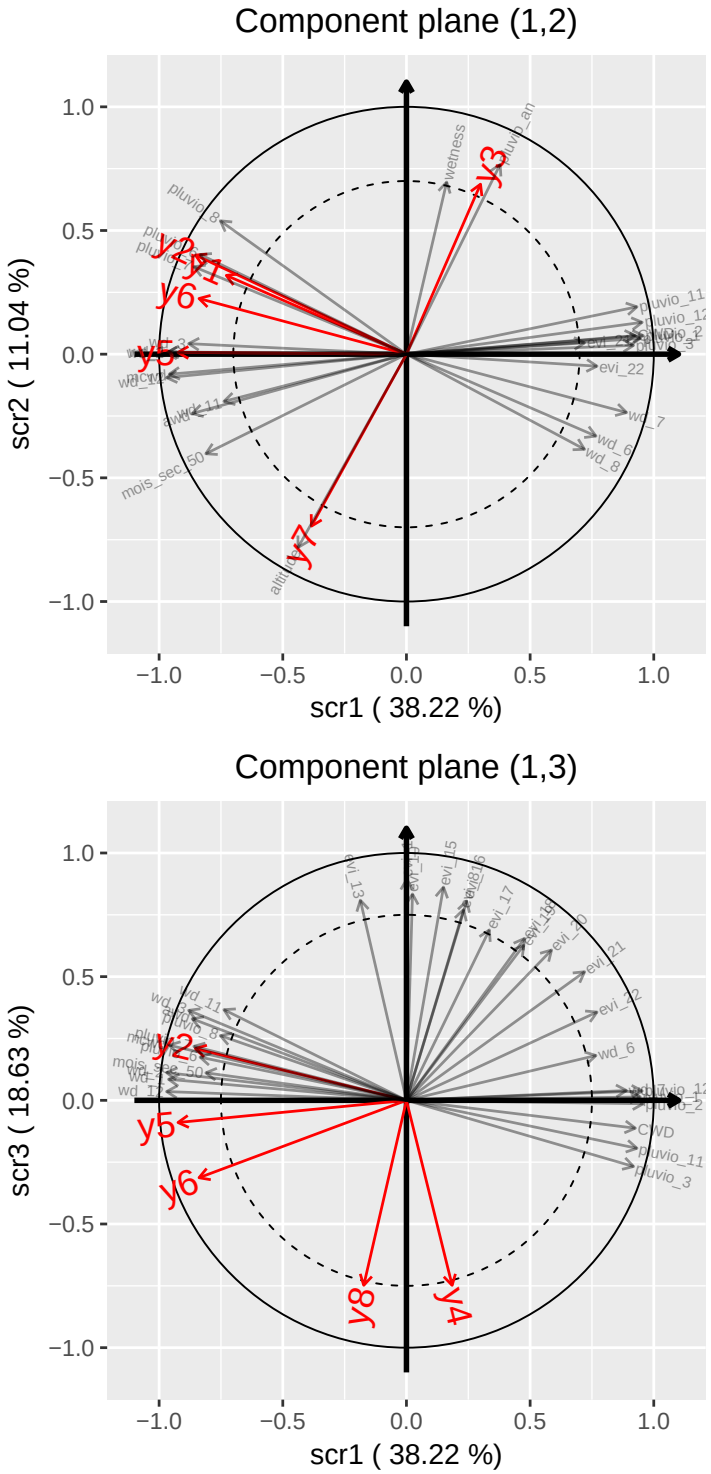


Figure 5.9 – *Component planes* $(1, 2)$ *and* $(1, 3)$ *output by mixed–SCGLR on dataset Genus, with optimal parameter triplet* $(K^\star, s^\star, l^\star) = (4, 0.1, 10)$. *The upper plot displays only variables having cosine greater than* 0.7 *with component plane* $(1, 2)$*. The lower one displays variables having cosine greater than* 0.75 *with component plane* $(1, 3)$*.*

Component plane $(1, 2)$ reveals two patterns. The first one is a global rain–wind pattern driven by the *pluvio*'s and *wd*'s variables which explain the abundances of species $1, 2, 5, 6$. The second is a rather local pattern driven by variables *altitude*, *wetness* and annual pluviometry (*pluvio_an*) which prove important to model and predict responses $\boldsymbol{y_3}$ and $\boldsymbol{y_7}$. Lastly, component 3 reveals a photosynthesis pattern driven by a part of the *Evi*'s, which seems useful to predict $\boldsymbol{y_4}$ and $\boldsymbol{y_8}$.

The decomposition of linear predictors on interpretable components allows to detect the species that tend to share common explanatory dimensions and those which are more idiosyncratic. We can then identify the variable–bundles these dimensions are related to. The underlying goal is a better understanding of the bio– and ecosystem diversity with a view to preserve them. Species 1, 2, 5 and 6 are sensitive to the same rain–wind regime, and species 4 and 8 are explained by the same photosynthetic pattern. On the contrary, species 3 and 7 are clearly separated. Species 7 grows at high altitudes where the atmosphere is rather dry while the abundance of species 3 is favoured by regular rainfall and high humidity.

Finally, one clarification deserves to be emphasised: to make this section easier to read, the model described by (5.15) does not take into account the surface area of each plot. This model then inherently assumes that the surface area is constant, which is not the case. All the results in this section were actually obtained considering the surface area of each plot as an offset (see Appendix 5.8.4 for more details).

5.7 Discussion and conclusions

Like Sufficient Dimension Reduction (SDR) methods (Li, 1991; Cook et al., 2007; Adragni and Cook, 2009), mixed–SCGLR is based on the construction of a reduction function of dimension less than p which tries to capture all the relevant information that $\boldsymbol{X}$ contains about $\boldsymbol{Y}$. However, the two approaches do not exactly pursue the same objectives. Indeed, SDR methods look for the "central subspace" containing the predictive information irrespective of the structures within $\boldsymbol{X}$ (e.g. dimensions capturing a large part of $\boldsymbol{X}$'s variance, or bundles of correlated variables). Mixed–SCGLR rather aims at basing the explanatory subspace on such structural dimensions so as to both gain interpretability and stabilise prediction. We think that extracting a hierarchy of strong and interpretable dimensions, and decomposing the linear predictor on them, is an essential asset in model–building. The difference in goals entails a difference in means: SDR is based on the sufficiency principle, which is

enough to identify a subspace but not to track strong predictive dimensions in it. By contrast, in the wake of PLS regression, mixed–SCGLR uses a criterion combining goodness–of–fit and structural relevance of components.

The supervised–component paradigm has proved effective in situations where regularisation is necessary but where variable selection is inappropriate — for instance when the true explanatory dimensions are latent and indirectly measured through highly correlated proxies.

- When $l = 1$, trade–off parameter s allows to continuously tune the attraction of components towards the principal components of explanatory variables. This results in a continuum between classical GLMM estimation ($s = 0$ is associated with no regularisation) and principal component generalised linear mixed regression (with $s = 1$).
- When $l > 1$, we take better advantage of local predictive structures in $\boldsymbol{X}$. The components we build are then usually closer to local gatherings of variables, thus easier to interpret.

Mixed–SCGLR is able to identify more or less local predictive structures common to all the $\boldsymbol{y_k}$'s and performs well on grouped data with Gaussian, Bernoulli, binomial and Poisson outcomes. Compared to penalty–based approaches as ridge or LASSO, the orthogonal components built by mixed–SCGLR reveal the multidimensional explanatory and predictive dimensions, and greatly facilitate the interpretation of the model.

However, a natural question arises as to the accuracy of our methodology under significant deviations from normality. With binary data for instance, variance components estimates are prone to some bias towards zero (McCulloch, 1997). As mentioned in Chapter 4, other estimation strategies might be considered, especially Monte Carlo integration methods such as the Monte Carlo Likelihood Approximation (MCLA) proposed by Knudson (2016). Their advantage is that they provide estimates based on direct approximations of the likelihood. Indirect maximisations of the likelihood are also available such as Monte Carlo Expectation–Maximisation (MCEM) and Monte Carlo Newton–Raphson, both introduced by McCulloch (1997). Alternatively, other methods are available within the Bayesian paradigm, such as the MCMC methods developed by Hadfield (2010) and the Sequential Monte Carlo (SMC) sampling approach proposed by Fan et al. (2008), both specifically designed for the GLMM framework. We think that these methodologies and Schall's could be combined sequentially. Indeed we could first take advantage of the linear approximation of the model in order to build the components, and then use MC–based methods to estimate both fixed–effect parameters and variance components. This would lead to replacing the current iteration of mixed–

SCGLR given by Algorithm 5.1 with the following steps (to keep things simple, we take the canonical link and we assume the dispersion parameter constant equal to 1):

Step 1. Compute components $\boldsymbol{F} = \left[\boldsymbol{f_1} \mid \ldots \mid \boldsymbol{f_K} \right]$ via the PING algorithm on Schall's linearised models.

Step 2. For each $k \in \{1, \ldots, q\}$, consider the hierarchy

$$p\left(\boldsymbol{y}^k | \boldsymbol{\xi}^k; \boldsymbol{\gamma}_k, \boldsymbol{\delta}_k\right) = \exp\left\{ \boldsymbol{y}^{k\mathrm{T}} \boldsymbol{\eta}_\xi^k - \mathbf{1}^\mathrm{T} c\left(\boldsymbol{\eta}_\xi^k\right) + \mathbf{1}^\mathrm{T} d\left(\boldsymbol{y}^k\right) \right\}$$
$$\boldsymbol{\xi}_k | \boldsymbol{D}_k \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{D}_k\right),$$

where $\boldsymbol{\eta}_k^\xi = \boldsymbol{F}\boldsymbol{\gamma}_k + \boldsymbol{A}\boldsymbol{\delta}_k + \boldsymbol{U}\boldsymbol{\xi}_k$, and c, d are the functions associated with the natural parametrisation of the GLM. For example, for the Bernoulli–logistic regression, we have: $c(\boldsymbol{x}) = \log(1 + e^{\boldsymbol{x}})$ and $d(\boldsymbol{x}) = \mathbf{0}$.

Step 3. Apply MC–based methods such as MCMC, MCLA, MCEM or MCNR to update $\boldsymbol{\gamma}_k$, $\boldsymbol{\delta}_k$, $\boldsymbol{\xi}_k$ and $\boldsymbol{D}_k$, $k \in \{1, \ldots, q\}$.

Step 4. Update working variables and weight matrices to define the new Schall's linearised models.

Even though such MC–based methods are computationally much more intensive than the "Joint–Maximisation" approaches (e.g. Schall (1991) or Breslow and Clayton (1993)) and have intrinsic disadvantages (particularly in the assessment of convergence and in the choice of prior distributions), they could give better results in case of binary data.

5.8 Appendices

5.8.1 Structural relevance: general formula and examples

For an easier reading in Section 5.3.3, we focused on a particular case of Structural Relevance (SR) measure called Variable–Powered Inertia (VPI). Indeed in practice, VPI was the measure used in the simulation schemes (Section 5.5) and real data study (Section 5.6). Introduced by Bry and Verron (2015), the notion of SR actually covers a broad spectrum of measures, according to the type of structure loading–vector $\boldsymbol{u}$ (or component $\boldsymbol{f}$) should align with. Unlike estimation methods without regularisation — in which all directions in $\operatorname{span}\{\boldsymbol{X}\}$ are considered equally important — the introduction of SR into the estimation process favours certain directions we see as stronger or more

relevant (for instance directions correlated to local gatherings of variables, or directions close to known interpretable subspaces, etc). Moreover, as we will see, this measure perfectly extends to mixtures of numerical and categorical explanatory variables. Section 5.8.1.1 presents the general formula of the SR measure and Sections 5.8.1.2, 5.8.1.3 and 5.8.1.4 detail three particular examples associated with different goals that deserve attention.

5.8.1.1 General formula

Let $\boldsymbol{X}$ be a $n \times p$ design matrix of explanatory variables, endowed with a $p \times p$ metric matrix $\boldsymbol{M}$ whose purpose is to weight $\boldsymbol{X}$'s variables appropriately. As suggested by Bry et al. (2012), $\boldsymbol{M}$ may take various forms according to the type of variables and structure of data. For identification purposes, the loading–vector of any component $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$ is constrained by: $\|\boldsymbol{u}\|^2_{\boldsymbol{M}^{-1}} = 1$.

Let us now recall the general formula of SR (Bry and Verron, 2015). Consider a set $\aleph$ of J "reference" symmetric positive semi–definite matrices, namely $\aleph = \{\boldsymbol{N_1}, \boldsymbol{N_2}, \ldots, \boldsymbol{N_J}\}$. Also consider a weight system $\boldsymbol{\omega} = \{\omega_1, \omega_2, \ldots, \omega_J\}$, and a scalar $l \in [1, +\infty[$. The associated SR measure, $\phi_{\aleph,\boldsymbol{\omega},l}$, is defined as the following two–degree homogeneous function of $\boldsymbol{u}$:

$$\phi_{\aleph,\boldsymbol{\omega},l}\left(\boldsymbol{u}\right) = \left(\sum_{j=1}^{J} \omega_j \left(\boldsymbol{u}^{\mathrm{T}} \boldsymbol{N_j} \boldsymbol{u}\right)^l\right)^{\frac{1}{l}}. \tag{5.17}$$

The $\boldsymbol{N_j}$'s are coding the directions of concern so that quadratic form $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{N_j}\boldsymbol{u}$ measures the closeness of vector $\boldsymbol{u}$ to a reference structure $\boldsymbol{S_j}$. With the additional constraint $\sum_{j=1}^{J} \omega_j = 1$, (5.17) expresses a generalised weighted average of quadratic forms $\{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{N_j}\boldsymbol{u} \,|\, j = 1, \ldots, J\}$, and thus averages the closeness measures of vector $\boldsymbol{u}$ to the associated reference structures.

As depicted in Section 5.3.3 for the VPI measure in the particular instance of four coplanar variables (see Figure 5.1), parameter l can be viewed as a "bundle locality parameter". Indeed, the value of l tunes the "width" of the bundles of directions considered relevant. The objective is then to focus on the most interpretable bundles of directions. The analysis of (5.17) for the extreme values of l allows a better understanding of the role of this parameter. The following details are taken from Bry and Verron (2015). For convenience, we consider here that $\forall j \in \{1, \ldots, J\}, \omega_j \neq 0$.

Lower extreme value. When $l = 1$, (5.17) writes

$$\phi_{\mathbf{N},\boldsymbol{\omega},1}(\boldsymbol{u}) = \sum_{j=1}^{J} \omega_j \boldsymbol{u}^{\mathrm{T}} \boldsymbol{N}_j \boldsymbol{u} = \boldsymbol{u}^{\mathrm{T}} \left(\sum_{j=1}^{J} \omega_j \boldsymbol{N}_j \right) \boldsymbol{u}. \tag{5.18}$$

As (5.18) can be interpreted as some inertia, its maximisation leads to the first eigenvector of the corresponding PCA. The directions having maximal structural relevance are thus the principal components of this PCA. The case where the $\boldsymbol{N}_j$'s are the orthogonal projectors on reference subspaces $\boldsymbol{S}_j$'s and all ω_j's are equal is of particular interest: it brings us back to the Generalised Canonical Analysis (Kettenring, 1971) of the set of subspaces $\{\boldsymbol{S}_j \mid j = 1, \ldots, J\}$.

Upper extreme value. We now consider the case where $l \to +\infty$. For that, the reasoning is based on the quantity

$$\|\boldsymbol{N}_j\| := \sup_{\boldsymbol{u}:\, \boldsymbol{u}^{\mathrm{T}} \boldsymbol{M}^{-1} \boldsymbol{u} = 1} \boldsymbol{u}^{\mathrm{T}} \boldsymbol{N}_j \boldsymbol{u}.$$

Two sub–cases deserve attention.

- If $\exists j^\star \in \{1, \ldots, J\} : \forall j \neq j^\star, \|\boldsymbol{N}_j\| < \|\boldsymbol{N}_{j^\star}\|$, then

  $$\operatorname*{arg\,max}_{\boldsymbol{u}:\, \boldsymbol{u}^{\mathrm{T}} \boldsymbol{M}^{-1} \boldsymbol{u} = 1} \phi_{\mathbf{N},\boldsymbol{\omega},\infty}(\boldsymbol{u}) = \operatorname*{arg\,max}_{\boldsymbol{u}:\, \boldsymbol{u}^{\mathrm{T}} \boldsymbol{M}^{-1} \boldsymbol{u} = 1} \boldsymbol{u}^{\mathrm{T}} \boldsymbol{N}_{j^\star} \boldsymbol{u}.$$

  In this case, the loading–vector $\boldsymbol{u}$ maximising $\phi_{\mathbf{N},\boldsymbol{\omega},\infty}(\boldsymbol{u})$ will be drawn to the subspace $\boldsymbol{S}_{j^\star}$ associated with $\boldsymbol{N}_{j^\star}$.
- If $\exists \boldsymbol{\lambda} \in \mathbb{R}^p : \forall j \in \{1, \ldots, J\}, \|\boldsymbol{N}_j\| = \boldsymbol{\lambda}$, then the first eigenvector (associated with the maximum eigenvalue) of the $\boldsymbol{N}_j$ having maximum weight ω_j maximises $\phi_{\mathbf{N},\boldsymbol{\omega},\infty}(\boldsymbol{u})$. If all the ω_j's are equal, the first eigenvectors of all the $\boldsymbol{N}_j$'s maximise it. It follows that in the particular case where the $\boldsymbol{N}_j$'s are the orthogonal projectors on reference subspaces $\boldsymbol{S}_j$'s and all ω_j's are equal, any vector $\boldsymbol{u}$ belonging to any $\boldsymbol{S}_j$ maximises $\phi_{\mathbf{N},\boldsymbol{\omega},\infty}(\boldsymbol{u})$. So, $\boldsymbol{u}$ will be drawn to the $\boldsymbol{S}_j$ closest to it.

Continuum between the two extreme values. To sum things up, what we see is that when l is minimum, the structurally relevant directions are the principal components, thus very global structural directions. By contrast, when l is maximum, the structurally relevant directions are more local (ultimately, the reference–subspaces, e.g. the variables themselves). As the appropriate bundle locality parameter depends on the data, it must be chosen by cross–validation techniques.

We will now present some usual particular cases of structural relevance measures covered by general formula (5.17). In the following, $\boldsymbol{P}$ denotes the weight matrix reflecting the a priori relative importance of observations — typically $\boldsymbol{P} = \frac{1}{n}\boldsymbol{I}_n$ for a uniform weighting.

5.8.1.2 Component Variance

Let $\boldsymbol{X}$ being composed of centred numeric variables. Suppose we want to determine the direction $\text{span}\{\boldsymbol{u}\}$ such that the inertia of observations along $\text{span}\{\boldsymbol{u}\}$ is maximum. The associated structural relevance criterion can then be defined as

$$\phi(\boldsymbol{u}) = \text{Var}(\boldsymbol{X}\boldsymbol{u}) = \|\boldsymbol{X}\boldsymbol{u}\|_{\boldsymbol{P}}^2 = \boldsymbol{u}^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{X}\boldsymbol{u}. \tag{5.19}$$

When we set

$$\begin{cases} \aleph = \{\boldsymbol{N_1}\}, \text{ where } \boldsymbol{N_1} = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{X} \\ \boldsymbol{\omega} = \{\omega_1\}, \text{ where } \omega_1 = 1 \\ l = 1, \end{cases}$$

general formula (5.17) reduces to (5.19). The latter is maximised by the first direct eigenvector in the PCA of $(\boldsymbol{X}, \boldsymbol{M}, \boldsymbol{P})$. So here, metric $\boldsymbol{M}$ must be such that PCA of $(\boldsymbol{X}, \boldsymbol{M}, \boldsymbol{P})$ is relevant. Now, in practice, explanatory variables are most often a mixture of numeric and nominal variables (see for example Lebart et al. (1995) and Bry (1994), or more recently Chavent et al. (2014) and Chavent et al. (2017)). More precisely, we often have

$$\boldsymbol{X} = \left[\, \boldsymbol{x_1} \mid \ldots \mid \boldsymbol{x_H} \mid \boldsymbol{X_1} \mid \ldots \mid \boldsymbol{X_B} \,\right], \tag{5.20}$$

where $\boldsymbol{x_1}, \ldots, \boldsymbol{x_H}$ are column–vectors coding the centred numeric variables and $\boldsymbol{X_1}, \ldots, \boldsymbol{X_B}$ are blocks of centred indicator–variables, each block coding a categorical variable. Note that if the b–th categorical variable has l_b levels, then $\boldsymbol{X_b}$ has $l_b - 1$ columns, the removed level being taken as "reference level". In such a framework, we must consider the block–diagonal metric matrix

$$\boldsymbol{M} = \textbf{bDiag}\Big[(\boldsymbol{x_1}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{x_1})^{-1}, \ldots, (\boldsymbol{x_H}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{x_H})^{-1}, \\ (\boldsymbol{X_1}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{X_1})^{-1}, \ldots, (\boldsymbol{X_B}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{X_B})^{-1} \Big].$$

Indeed, this matrix bridges ordinary PCA of numeric variables with that of Multiple Correspondence Analysis (Greenacre and Blasius, 2006).

5.8.1.3 Block variance captured by component and Variable–Powered Inertia

Suppose $\boldsymbol{X}$ consists of p standardised numeric variables. If we want to find the normalised component $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$ that captures the maximum inertia of

the variable set $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_p}$, the structural relevance should be defined as

$$\begin{aligned}\phi(\boldsymbol{u}) &= \sum_{j=1}^{p} \rho^2(\boldsymbol{f}, \boldsymbol{x_j}) = \sum_{j=1}^{p} \langle \boldsymbol{f} \,|\, \boldsymbol{x_j} \rangle_{\boldsymbol{P}}^2 \\ &= \sum_{j=1}^{p} \boldsymbol{f}^{\mathsf{T}} \boldsymbol{P} \boldsymbol{x_j} \boldsymbol{x_j}^{\mathsf{T}} \boldsymbol{P} \boldsymbol{f} \\ &= \boldsymbol{u}^{\mathsf{T}} \boldsymbol{X}^{\mathsf{T}} \boldsymbol{P} \left(\sum_{j=1}^{p} \boldsymbol{x_j} \boldsymbol{x_j}^{\mathsf{T}} \right) \boldsymbol{P} \boldsymbol{X} \boldsymbol{u} \\ &= \boldsymbol{u}^{\mathsf{T}} \left(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{P} \boldsymbol{X} \boldsymbol{X}^{\mathsf{T}} \boldsymbol{P} \boldsymbol{X} \right) \boldsymbol{u} = \boldsymbol{u}^{\mathsf{T}} \left(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{P} \boldsymbol{X} \right)^2 \boldsymbol{u}. \end{aligned} \tag{5.21}$$

This time, the particular structural relevance criterion defined by (5.21) corresponds to the general formula (5.17) after setting

$$\begin{cases} \boldsymbol{\aleph} = \{\boldsymbol{N_1}\}, \text{ where } \boldsymbol{N_1} = \left(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{P} \boldsymbol{X} \right)^2 \\ \boldsymbol{\omega} = \{\omega_1\}, \text{ where } \omega_1 = 1 \\ l = 1. \end{cases}$$

For identification purposes, we impose $\|\boldsymbol{f}\|_{\boldsymbol{P}}^2 = 1$. We then have $\|\boldsymbol{X}\boldsymbol{u}\|_{\boldsymbol{P}}^2 = \langle \boldsymbol{X}\boldsymbol{u} \,|\, \boldsymbol{X}\boldsymbol{u} \rangle_{\boldsymbol{P}} = \boldsymbol{u}^{\mathsf{T}} \left(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{P} \boldsymbol{X} \right) \boldsymbol{u} = 1$. As a result, the suitable metric matrix is here $\boldsymbol{M} = \left(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{P} \boldsymbol{X} \right)^{-1}$, so that $\boldsymbol{u}^{\mathsf{T}} \boldsymbol{M}^{-1} \boldsymbol{u} = 1 \iff \|\boldsymbol{f}\|_{\boldsymbol{P}}^2 = 1$. Since (5.21) is the inertia of variables along $\operatorname{span}\{\boldsymbol{X}\boldsymbol{u}\}$, it is maximised by the first dual eigenvector in the PCA of $(\boldsymbol{X}, \boldsymbol{M}, \boldsymbol{P})$.

Naturally, for $\boldsymbol{X}^{\mathsf{T}} \boldsymbol{P} \boldsymbol{X}$ to be regular, $\boldsymbol{X}$ has to be a column full rank matrix, which we will assume momentarily. As mentioned in Section 5.3.3, still imposing $\|\boldsymbol{f}\|_{\boldsymbol{P}}^2 = 1$ through metric $\boldsymbol{M} = \left(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{P} \boldsymbol{X} \right)^{-1}$, we can extend criterion (5.21) to the "Variable–Powered Inertia" (VPI) defined as

$$\phi(\boldsymbol{u}) = \left(\sum_{j=1}^{p} \omega_j \left[\rho^2(\boldsymbol{f}, \boldsymbol{x^j}) \right]^l \right)^{\frac{1}{l}} = \left(\sum_{j=1}^{p} \omega_j \left[\langle \boldsymbol{f} \,|\, \boldsymbol{x^j} \rangle_{\boldsymbol{P}}^2 \right]^l \right)^{\frac{1}{l}}. \tag{5.22}$$

Two particular cases can be mentioned.

- If $l = 1$ and $\omega_j = 1 \ \forall j \in \{1, \ldots, p\}$, (5.22) gives back block–variance criterion (5.21).
- If $l = 2$ and $\omega_j = 1 \ \forall j \in \{1, \ldots, p\}$, (5.22) yields a varimax–like criterion, initially introduced by Kaiser (1958).

There is a simple way to extend the VPI criterion to both quantitative and qualitative explanatory variables. To this end, let us consider the typical mixture of numeric and nominal variables defined by (5.20). In this case, the VPI

criterion is defined as

$$\phi(\boldsymbol{u}) = \left(\sum_{h=1}^{H} \omega_h \left[\rho^2 (\boldsymbol{X}\boldsymbol{u}, \boldsymbol{x}_h) \right]^l + \sum_{b=1}^{B} \omega_b \left[\mathrm{R}^2 (\boldsymbol{X}\boldsymbol{u}, \boldsymbol{X}_b) \right]^l \right)^{\frac{1}{l}}, \tag{5.23}$$

where ρ^2 is the square of the correlation coefficient and R^2 the coefficient of determination. In order to unify the two terms of the sum of formula (5.23) within the same framework, let us recall that

- first,

$$\begin{aligned} \rho^2 (\boldsymbol{X}\boldsymbol{u}, \boldsymbol{x}_h) &= \cos^2_{\boldsymbol{P}} (\boldsymbol{X}\boldsymbol{u}, \boldsymbol{x}_h) \\ &= \cos^2_{\boldsymbol{P}} \left(\boldsymbol{X}\boldsymbol{u}, \Pi^{\boldsymbol{P}}_{\mathrm{span}\{\boldsymbol{x}_h\}} \boldsymbol{X}\boldsymbol{u} \right) = \frac{\left\| \Pi^{\boldsymbol{P}}_{\mathrm{span}\{\boldsymbol{x}_h\}} \boldsymbol{X}\boldsymbol{u} \right\|^2_{\boldsymbol{P}}}{\|\boldsymbol{X}\boldsymbol{u}\|^2_{\boldsymbol{P}}}, \end{aligned}$$

- and then by definition,

$$\mathrm{R}^2 (\boldsymbol{X}\boldsymbol{u}, \boldsymbol{X}_b) = \frac{\left\| \Pi^{\boldsymbol{P}}_{\mathrm{span}\{\boldsymbol{X}_b\}} \boldsymbol{X}\boldsymbol{u} \right\|^2_{\boldsymbol{P}}}{\|\boldsymbol{X}\boldsymbol{u}\|^2_{\boldsymbol{P}}},$$

  where $\Pi^{\boldsymbol{P}}_{\mathrm{span}\{\boldsymbol{X}_b\}} = \boldsymbol{X}_b^{\mathsf{T}} (\boldsymbol{X}_b^{\mathsf{T}} \boldsymbol{P} \boldsymbol{X}_b)^{-1} \boldsymbol{X}_b^{\mathsf{T}} \boldsymbol{P}$.

In addition, as detailed in Section 5.8.1.4, we have

$$\left\| \Pi^{\boldsymbol{P}}_{\mathrm{span}\{\boldsymbol{X}_b\}} \boldsymbol{X}\boldsymbol{u} \right\|^2_{\boldsymbol{P}} = \left\langle \boldsymbol{X}\boldsymbol{u} \,\middle|\, \Pi^{\boldsymbol{P}}_{\mathrm{span}\{\boldsymbol{X}_b\}} \boldsymbol{X}\boldsymbol{u} \right\rangle_{\boldsymbol{P}}.$$

Finally, since $\|\boldsymbol{X}\boldsymbol{u}\|^2_{\boldsymbol{P}} = 1$, (5.23) writes more explicitly

$$\phi(\boldsymbol{u}) = \left[\sum_{h=1}^{H} \omega_h \left\langle \boldsymbol{X}\boldsymbol{u} \,\middle|\, \Pi^{\boldsymbol{P}}_{\mathrm{span}\{\boldsymbol{x}_h\}} \boldsymbol{X}\boldsymbol{u} \right\rangle^l_{\boldsymbol{P}} + \sum_{b=1}^{B} \omega_b \left\langle \boldsymbol{X}\boldsymbol{u} \,\middle|\, \Pi^{\boldsymbol{P}}_{\mathrm{span}\{\boldsymbol{X}_b\}} \boldsymbol{X}\boldsymbol{u} \right\rangle^l_{\boldsymbol{P}} \right]^{\frac{1}{l}}.$$

Let us now consider the case where matrix $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{X}$ is singular. Owing to collinearity, as suggested in Section 5.3.3, $\boldsymbol{X}$ should be replaced with the matrix $\boldsymbol{C}$ of its principal components associated with non–null eigenvalues. More precisely, $\boldsymbol{C} = \boldsymbol{X}\boldsymbol{V}$, where $\boldsymbol{V}$ is the matrix of corresponding unit–eigenvectors. The component is then sought as $\boldsymbol{f} = \boldsymbol{C}\boldsymbol{u} = \boldsymbol{X}\widetilde{\boldsymbol{u}}$, where $\widetilde{\boldsymbol{u}} = \boldsymbol{V}\boldsymbol{u}$. Bry et al. (2018) show that among all coefficient vectors $\boldsymbol{t}$ such that $\boldsymbol{X}\boldsymbol{t} = \boldsymbol{f}$, $\widetilde{\boldsymbol{u}}$ is that which has the minimum L_2–norm. Indeed, consider the following program:

$$\begin{cases} \min & \|\boldsymbol{t}\|^2 \\ \text{subject to} & \boldsymbol{X}\boldsymbol{t} = \boldsymbol{f} \end{cases} \iff \begin{cases} \min & \|\boldsymbol{t}\|^2 \\ \text{subject to} & \boldsymbol{X}\boldsymbol{t} = \boldsymbol{X}\widetilde{\boldsymbol{u}}. \end{cases}$$

Of course,

$$\boldsymbol{X}\boldsymbol{t} = \boldsymbol{X}\widetilde{\boldsymbol{u}} \Longleftrightarrow \boldsymbol{X}\boldsymbol{e} = \boldsymbol{0}, \text{ where } \boldsymbol{e} = \boldsymbol{t} - \widetilde{\boldsymbol{u}}.$$

According to the definition of matrix $\boldsymbol{V}$, each column–vector $\boldsymbol{v_k}$ of $\boldsymbol{V}$ satisfies

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}\boldsymbol{v_k} = \lambda_k \boldsymbol{v_k}, \text{ with } \lambda_k > 0.$$

Therefore, $\widetilde{\boldsymbol{u}} = \boldsymbol{V}\boldsymbol{u} \in \mathrm{Im}\boldsymbol{X}^{\mathrm{T}} = (\mathrm{Ker}\boldsymbol{X})^{\perp}$ and on the other hand, since $\boldsymbol{X}\boldsymbol{e} = \boldsymbol{0}$, $\boldsymbol{e} \in \mathrm{Ker}\boldsymbol{X}$. The decomposition $\boldsymbol{t} = \widetilde{\boldsymbol{u}} + \boldsymbol{e}$ is then unique. The Pythagore's theorem yields

$$\|\boldsymbol{t}\|^2 = \|\widetilde{\boldsymbol{u}}\|^2 + \|\boldsymbol{e}\|^2,$$

thus implying that $\|\boldsymbol{t}\|^2$ is minimum for $\boldsymbol{e} = \boldsymbol{0}$, i.e. for $\boldsymbol{t} = \widetilde{\boldsymbol{u}}$.

□

5.8.1.4 Closeness of the component's coefficient vector to some reference subspaces

Finally, suppose we would like to determine vector $\boldsymbol{u}$ that is simultaneously as close as possible to a family of predefined subspaces $(\boldsymbol{S_1}, \ldots, \boldsymbol{S_J})$. For the sake of simplicity, we choose $\boldsymbol{M}^{-1} = \boldsymbol{I}$, so that $\|\boldsymbol{u}\|^2 = 1$. The closeness of vector $\boldsymbol{u}$ to subspace $\boldsymbol{S_j}$ can be measured through $\cos^2(\boldsymbol{u}, \mathrm{span}\{\boldsymbol{S_j}\})$, which can be rewritten as

$$\begin{aligned}
\cos^2(\boldsymbol{u}, \mathrm{span}\{\boldsymbol{S_j}\}) &= \cos^2\left(\boldsymbol{u}, \Pi_{\mathrm{span}\{\boldsymbol{S_j}\}}\boldsymbol{u}\right) \\
&= \frac{\left\|\Pi_{\mathrm{span}\{\boldsymbol{S_j}\}}\boldsymbol{u}\right\|^2}{\|\boldsymbol{u}\|^2} = \left\|\Pi_{\mathrm{span}\{\boldsymbol{S_j}\}}\boldsymbol{u}\right\|^2 \quad \text{because } \|\boldsymbol{u}\|^2 = 1 \\
&= \left\langle \Pi_{\mathrm{span}\{\boldsymbol{S_j}\}}\boldsymbol{u} \,\middle|\, \Pi_{\mathrm{span}\{\boldsymbol{S_j}\}}\boldsymbol{u} \right\rangle \\
&= \left\langle \boldsymbol{u} \,\middle|\, \Pi^{\star}_{\mathrm{span}\{\boldsymbol{S_j}\}}\Pi_{\mathrm{span}\{\boldsymbol{S_j}\}}\boldsymbol{u} \right\rangle \quad \text{where } \Pi^{\star} \text{ is the adjoint of } \Pi \\
&= \left\langle \boldsymbol{u} \,\middle|\, \Pi^{2}_{\mathrm{span}\{\boldsymbol{S_j}\}}\boldsymbol{u} \right\rangle \quad \text{since } \Pi \text{ is self–adjoint} \\
&= \left\langle \boldsymbol{u} \,\middle|\, \Pi_{\mathrm{span}\{\boldsymbol{S_j}\}}\boldsymbol{u} \right\rangle \quad \text{since } \Pi \text{ is idempotent.}
\end{aligned}$$

In this context, we may define the structural relevance as

$$\phi(\boldsymbol{u}) = \sum_{j=1}^{J} \cos^2(\boldsymbol{u}, \boldsymbol{S_j}) = \sum_{j=1}^{J} \boldsymbol{u}^{\mathrm{T}} \Pi_{\mathrm{span}\{\boldsymbol{S_j}\}}\boldsymbol{u}, \tag{5.24}$$

which goes back to the general structural relevance formula (5.17) with

$$\begin{cases} \aleph = \left\{\Pi_{\text{span}\{\boldsymbol{S_1}\}}, \ldots, \Pi_{\text{span}\{\boldsymbol{S_J}\}}\right\} \\ \boldsymbol{\omega} = \{1, \ldots, 1\} \\ l = 1. \end{cases}$$

5.8.2 Analytical expression of the SCGLR–specific criterion

As a reminder, the specific criterion which SCGLR maximises for computing the $(h+1)$–th loading–vector, $\boldsymbol{u_{h+1}}$, writes

$$\mathcal{J}_h(\boldsymbol{u}) = [\phi(\boldsymbol{u})]^s \; [\psi_{\boldsymbol{A_h}}(\boldsymbol{u})]^{1-s},$$

with

$$\begin{cases} \phi(\boldsymbol{u}) = \left(\sum_{j=1}^{J} \omega_j \left(\boldsymbol{u}^{\mathsf{T}} \boldsymbol{N_j} \boldsymbol{u}\right)^l\right)^{\frac{1}{l}} \\ \psi_{\boldsymbol{A_h}}(\boldsymbol{u}) = \sum_{k=1}^{q} \left\|\boldsymbol{z_k^\xi}\right\|^2_{\boldsymbol{W_k^\xi}} \cos^2_{\boldsymbol{W_k^\xi}}\left(\boldsymbol{z_k^\xi}, \text{span}\{\boldsymbol{Xu}, \boldsymbol{A_h}\}\right). \end{cases}$$

To facilitate the computation of the loading–vector, a solution is to express $\mathcal{J}_h(\boldsymbol{u})$ as a function of quadratic forms. As $\phi(\boldsymbol{u})$ corresponds to a generalised weighted average of quadratic forms, it is now necessary to transform the expression of $\psi_{\boldsymbol{A_h}}(\boldsymbol{u})$. To achieve that, we decompose the projection on the regression space as follows. With $\boldsymbol{\mathcal{X}_k^h} = \Pi^{\boldsymbol{W_k^\xi}}_{\text{span}\{\boldsymbol{A_h}\}^\perp}\boldsymbol{X}$, we have

$$\text{span}\left\{\boldsymbol{Xu}, \boldsymbol{A_h}\right\} = \text{span}\left\{\boldsymbol{\mathcal{X}_k^h u}, \boldsymbol{A_h}\right\}.$$

Since $\text{span}\left\{\boldsymbol{\mathcal{X}_k^h}\right\}$ is orthogonal to $\text{span}\left\{\boldsymbol{A_h}\right\}$,

$$\Pi^{\boldsymbol{W_k^\xi}}_{\text{span}\{\boldsymbol{Xu}, \boldsymbol{A_h}\}} = \Pi^{\boldsymbol{W_k^\xi}}_{\text{span}\{\boldsymbol{\mathcal{X}_k^h u}, \boldsymbol{A_h}\}} = \Pi^{\boldsymbol{W_k^\xi}}_{\text{span}\{\boldsymbol{\mathcal{X}_k^h u}\}} + \Pi^{\boldsymbol{W_k^\xi}}_{\text{span}\{\boldsymbol{A_h}\}}.$$

In addition, classical Euclidean statistical concepts give at a time

$$\begin{cases} \cos_{\boldsymbol{W_k^\xi}}\left(\boldsymbol{z_k^\xi}, \text{span}\{\boldsymbol{Xu}, \boldsymbol{A_h}\}\right) = \dfrac{\left\|\Pi^{\boldsymbol{W_k^\xi}}_{\text{span}\{\boldsymbol{Xu}, \boldsymbol{A_h}\}} \boldsymbol{z_k^\xi}\right\|_{\boldsymbol{W_k^\xi}}}{\left\|\boldsymbol{z_k^\xi}\right\|_{\boldsymbol{W_k^\xi}}}, \\ \qquad \text{and} \\ \cos_{\boldsymbol{W_k^\xi}}\left(\boldsymbol{z_k^\xi}, \text{span}\{\boldsymbol{Xu}, \boldsymbol{A_h}\}\right) = \dfrac{\left\langle \boldsymbol{z_k^\xi} \middle| \Pi^{\boldsymbol{W_k^\xi}}_{\text{span}\{\boldsymbol{Xu}, \boldsymbol{A_h}\}} \boldsymbol{z_k^\xi} \right\rangle_{\boldsymbol{W_k^\xi}}}{\left\|\boldsymbol{z_k^\xi}\right\|_{\boldsymbol{W_k^\xi}} \left\|\Pi^{\boldsymbol{W_k^\xi}}_{\text{span}\{\boldsymbol{Xu}, \boldsymbol{A_h}\}} \boldsymbol{z_k^\xi}\right\|_{\boldsymbol{W_k^\xi}}}. \end{cases}$$

Consequently,

$$\begin{aligned}\cos^2_{\boldsymbol{W}_k^\xi}\left(\boldsymbol{z}_k^\xi,\,\mathrm{span}\,\{\boldsymbol{X}\boldsymbol{u},\boldsymbol{A}_h\}\right) &= \frac{\left\langle \boldsymbol{z}_k^\xi \middle| \Pi^{\boldsymbol{W}_k^\xi}_{\mathrm{span}\{\boldsymbol{X}\boldsymbol{u},\boldsymbol{A}_h\}}\boldsymbol{z}_k^\xi \right\rangle_{\boldsymbol{W}_k^\xi}}{\left\|\boldsymbol{z}_k^\xi\right\|^2_{\boldsymbol{W}_k^\xi}} \\ &= \frac{\left\langle \boldsymbol{z}_k^\xi \middle| \left(\Pi^{\boldsymbol{W}_k^\xi}_{\mathrm{span}\left\{\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right\}} + \Pi^{\boldsymbol{W}_k^\xi}_{\mathrm{span}\{\boldsymbol{A}_h\}}\right)\boldsymbol{z}_k^\xi \right\rangle_{\boldsymbol{W}_k^\xi}}{\left\|\boldsymbol{z}_k^\xi\right\|^2_{\boldsymbol{W}_k^\xi}} \\ &= \frac{\left\langle \boldsymbol{z}_k^\xi \middle| \Pi^{\boldsymbol{W}_k^\xi}_{\mathrm{span}\left\{\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right\}}\boldsymbol{z}_k^\xi \right\rangle_{\boldsymbol{W}_k^\xi}}{\left\|\boldsymbol{z}_k^\xi\right\|^2_{\boldsymbol{W}_k^\xi}} + \frac{\left\langle \boldsymbol{z}_k^\xi \middle| \Pi^{\boldsymbol{W}_k^\xi}_{\mathrm{span}\{\boldsymbol{A}_h\}}\boldsymbol{z}_k^\xi \right\rangle_{\boldsymbol{W}_k^\xi}}{\left\|\boldsymbol{z}_k^\xi\right\|^2_{\boldsymbol{W}_k^\xi}}.\end{aligned}$$

The Goodness–of–Fit measure $\psi_{\boldsymbol{A}_h}(\boldsymbol{u})$ then writes more explicitly

$$\begin{aligned}\psi_{\boldsymbol{A}_h}(\boldsymbol{u}) &= \sum_{k=1}^{q}\left\|\boldsymbol{z}_k^\xi\right\|^2_{\boldsymbol{W}_k^\xi}\cos^2_{\boldsymbol{W}_k^\xi}\left(\boldsymbol{z}_k^\xi,\,\mathrm{span}\,\{\boldsymbol{X}\boldsymbol{u},\boldsymbol{A}_h\}\right) \\ &= \left\langle \boldsymbol{z}_k^\xi \middle| \Pi^{\boldsymbol{W}_k^\xi}_{\mathrm{span}\left\{\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right\}}\boldsymbol{z}_k^\xi \right\rangle_{\boldsymbol{W}_k^\xi} + \left\langle \boldsymbol{z}_k^\xi \middle| \Pi^{\boldsymbol{W}_k^\xi}_{\mathrm{span}\{\boldsymbol{A}_h\}}\boldsymbol{z}_k^\xi \right\rangle_{\boldsymbol{W}_k^\xi}.\end{aligned}$$

Now,

$$\begin{aligned}\left\langle \boldsymbol{z}_k^\xi \middle| \Pi^{\boldsymbol{W}_k^\xi}_{\mathrm{span}\left\{\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right\}}\boldsymbol{z}_k^\xi \right\rangle_{\boldsymbol{W}_{\xi,k}} &= \boldsymbol{z}_k^{\xi^\mathrm{T}}\boldsymbol{W}_k^\xi\,\Pi^{\boldsymbol{W}_k^\xi}_{\mathrm{span}\left\{\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right\}}\boldsymbol{z}_k^\xi \\ &= \mathrm{Trace}\left(\boldsymbol{z}_k^{\xi^\mathrm{T}}\boldsymbol{W}_k^\xi\,\Pi^{\boldsymbol{W}_k^\xi}_{\mathrm{span}\left\{\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right\}}\boldsymbol{z}_k^\xi\right) \\ &= \mathrm{Trace}\left(\boldsymbol{z}_k^{\xi^\mathrm{T}}\boldsymbol{W}_k^\xi\,\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\left[\left(\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right)^\mathrm{T}\boldsymbol{W}_k^\xi\,\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right]^{-1}\left(\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right)^\mathrm{T}\boldsymbol{W}_k^\xi\,\boldsymbol{z}_k^\xi\right) \\ &= \mathrm{Trace}\left(\left[\left(\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right)^\mathrm{T}\boldsymbol{W}_k^\xi\,\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right]^{-1}\left(\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right)^\mathrm{T}\boldsymbol{W}_k^\xi\,\boldsymbol{z}_k^\xi\boldsymbol{z}_k^{\xi^\mathrm{T}}\boldsymbol{W}_k^\xi\,\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right) \\ &= \left[\left(\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right)^\mathrm{T}\boldsymbol{W}_k^\xi\,\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right]^{-1}\left(\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}\right)^\mathrm{T}\boldsymbol{W}_k^\xi\,\boldsymbol{z}_k^\xi\boldsymbol{z}_k^{\xi^\mathrm{T}}\boldsymbol{W}_k^\xi\,\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u} \\ &= \frac{\boldsymbol{u}^\mathrm{T}\boldsymbol{\mathcal{X}}_k^{h^\mathrm{T}}\boldsymbol{W}_k^\xi\,\boldsymbol{z}_k^\xi\boldsymbol{z}_k^{\xi^\mathrm{T}}\boldsymbol{W}_k^\xi\,\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}}{\boldsymbol{u}^\mathrm{T}\boldsymbol{\mathcal{X}}_k^{h^\mathrm{T}}\boldsymbol{W}_k^\xi\,\boldsymbol{\mathcal{X}}_k^h\boldsymbol{u}}.\end{aligned}$$

By defining

$$\begin{cases} \boldsymbol{A}_{\boldsymbol{k}}^{\boldsymbol{h}} := \boldsymbol{\mathcal{X}}_{\boldsymbol{k}}^{\boldsymbol{h}^{\mathrm{T}}} \boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}} \boldsymbol{z}_{\boldsymbol{k}}^{\boldsymbol{\xi}} \boldsymbol{z}_{\boldsymbol{k}}^{\boldsymbol{\xi}^{\mathrm{T}}} \boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}} \boldsymbol{\mathcal{X}}_{\boldsymbol{k}}^{\boldsymbol{h}} \\ \boldsymbol{B}_{\boldsymbol{k}}^{\boldsymbol{h}} := \boldsymbol{\mathcal{X}}_{\boldsymbol{k}}^{\boldsymbol{h}^{\mathrm{T}}} \boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}} \boldsymbol{\mathcal{X}}_{\boldsymbol{k}}^{\boldsymbol{h}} \\ c_k^h := \left\langle \boldsymbol{z}_{\boldsymbol{k}}^{\boldsymbol{\xi}} \middle| \Pi_{\mathrm{span}\{\boldsymbol{A_h}\}}^{\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}}} \boldsymbol{z}_{\boldsymbol{k}}^{\boldsymbol{\xi}} \right\rangle_{\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}}} \end{cases}$$

we get the general matrix form of the SCGLR–specific criterion to be maximised:

$$\mathcal{J}_h\left(\boldsymbol{u}\right) = \left[\sum_{j=1}^{J} \omega_j \left(\boldsymbol{u}^{\mathrm{T}} \boldsymbol{N_j} \boldsymbol{u}\right)^l\right]^{\frac{s}{l}} \left[\sum_{k=1}^{q} \left(\frac{\boldsymbol{u}^{\mathrm{T}} \boldsymbol{A}_{\boldsymbol{k}}^{\boldsymbol{h}} \boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}} \boldsymbol{B}_{\boldsymbol{k}}^{\boldsymbol{h}} \boldsymbol{u}} + c_k^h\right)\right]^{1-s}.$$

5.8.3 The Projected Iterated Normed Gradient (PING) algorithm

The Projected Iterated Normed Gradient (PING) is an extension of the iterated power algorithm — see Householder (2013) and Wilkinson (1965) for a complete treatment, and Golub and van der Vorst (2000) for a nice review. The purpose of the PING algorithm is to solve any program which has the form

$$\begin{cases} \max \quad \mathcal{J}_h\left(\boldsymbol{u}\right) \\ \text{subject to} \quad \boldsymbol{u}^{\mathrm{T}} \boldsymbol{M}^{-1} \boldsymbol{u} = 1 \text{ and } \boldsymbol{\Delta}_{\boldsymbol{h}}^{\mathrm{T}} \boldsymbol{u} = \boldsymbol{0}. \end{cases} \tag{5.25}$$

Note that putting $\boldsymbol{v} := \boldsymbol{M}^{-1/2}\boldsymbol{u}$, $\mathcal{G}_h\left(\boldsymbol{v}\right) := \mathcal{J}_h\left(\boldsymbol{M}^{1/2}\boldsymbol{v}\right)$ and $\boldsymbol{E_h} := \boldsymbol{M}^{1/2}\boldsymbol{\Delta_h}$, program (5.25) is strictly equivalent to program (5.26):

$$\begin{cases} \max \quad \mathcal{G}_h\left(\boldsymbol{v}\right) \\ \text{subject to} \quad \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = 1 \text{ and } \boldsymbol{E}_{\boldsymbol{h}}^{\mathrm{T}}\boldsymbol{v} = \boldsymbol{0}. \end{cases} \tag{5.26}$$

Solving (5.26) requires the definition of the following Lagrange function

$$\mathscr{L}\left(\boldsymbol{v}, \lambda, \boldsymbol{\mu}\right) = \mathcal{G}_h\left(\boldsymbol{v}\right) - \lambda\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} - 1\right) - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{E}_{\boldsymbol{h}}^{\mathrm{T}}\boldsymbol{v},$$

where λ and $\boldsymbol{\mu}$ are the Lagrange multipliers associated with constraints $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = 1$ and $\boldsymbol{E}_{\boldsymbol{h}}^{\mathrm{T}}\boldsymbol{v} = \boldsymbol{0}$ respectively. As usual, we have the following equivalence:

$$\underset{\lambda,\boldsymbol{\mu}}{\boldsymbol{\nabla}} \mathscr{L}\left(\boldsymbol{v}, \lambda, \boldsymbol{\mu}\right) = \boldsymbol{0} \Longleftrightarrow \begin{cases} \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = 1 \\ \boldsymbol{E}_{\boldsymbol{h}}^{\mathrm{T}}\boldsymbol{v} = \boldsymbol{0}, \end{cases} \tag{5.27}$$

where $\underset{\lambda,\boldsymbol{\mu}}{\boldsymbol{\nabla}}\mathscr{L}$ is the gradient of $\mathscr{L}$ with respect to $(\lambda, \boldsymbol{\mu})$. Besides, with $\Gamma_h(\boldsymbol{v}) = \boldsymbol{\nabla}\mathcal{G}_h(\boldsymbol{v})$,

$$\underset{\boldsymbol{v}}{\boldsymbol{\nabla}}\mathscr{L}(\boldsymbol{v}, \lambda, \boldsymbol{\mu}) = \mathbf{0} \iff \Gamma_h(\boldsymbol{v}) - 2\lambda\boldsymbol{v} - \boldsymbol{E_h}\boldsymbol{\mu} = \mathbf{0} \tag{5.28}$$

$$\iff \boldsymbol{v} = \frac{1}{2\lambda}\left[\Gamma_h(\boldsymbol{v}) - \boldsymbol{E_h}\boldsymbol{\mu}\right] \tag{5.29}$$

Premultiplying (5.28) by $\boldsymbol{E}_h^{\mathrm{T}}$ yields

$$\begin{aligned}
2\lambda\boldsymbol{E}_h^{\mathrm{T}}\boldsymbol{v} &= \boldsymbol{E}_h^{\mathrm{T}}\Gamma_h(\boldsymbol{v}) - \boldsymbol{E}_h^{\mathrm{T}}\boldsymbol{E_h}\boldsymbol{\mu} \\
&\iff \boldsymbol{E}_h^{\mathrm{T}}\Gamma_h(\boldsymbol{v}) - \boldsymbol{E}_h^{\mathrm{T}}\boldsymbol{E_h}\boldsymbol{\mu} = \mathbf{0} \quad \text{according to (5.27, line 2)} \\
&\iff \boldsymbol{E}_h^{\mathrm{T}}\Gamma_h(\boldsymbol{v}) = \boldsymbol{E}_h^{\mathrm{T}}\boldsymbol{E_h}\boldsymbol{\mu} \\
&\iff \boldsymbol{\mu} = \left(\boldsymbol{E}_h^{\mathrm{T}}\boldsymbol{E_h}\right)^{-1}\boldsymbol{E}_h^{\mathrm{T}}\Gamma_h(\boldsymbol{v})\,.
\end{aligned} \tag{5.30}$$

Then, putting back (5.30) into (5.29) provides

$$\begin{aligned}
\boldsymbol{v} &= \frac{1}{2\lambda}\left[\Gamma_h(\boldsymbol{v}) - \boldsymbol{E_h}\left(\boldsymbol{E}_h^{\mathrm{T}}\boldsymbol{E_h}\right)^{-1}\boldsymbol{E}_h^{\mathrm{T}}\Gamma_h(\boldsymbol{v})\right] \\
&= \frac{1}{2\lambda}\left[\boldsymbol{I} - \boldsymbol{E_h}\left(\boldsymbol{E}_h^{\mathrm{T}}\boldsymbol{E_h}\right)^{-1}\boldsymbol{E}_h^{\mathrm{T}}\right]\Gamma_h(\boldsymbol{v}) \\
&= \frac{1}{2\lambda}\Pi_{\mathrm{span}\{\boldsymbol{E_h}\}^{\perp}}\Gamma_h(\boldsymbol{v})\,,
\end{aligned} \tag{5.31}$$

where $\Pi_{\mathrm{span}\{\boldsymbol{E_h}\}^{\perp}} = \boldsymbol{I} - \boldsymbol{E_h}\left(\boldsymbol{E}_h^{\mathrm{T}}\boldsymbol{E_h}\right)^{-1}\boldsymbol{E}_h^{\mathrm{T}}$.

Finally, constraint $\|\boldsymbol{v}\|^2 = 1$ (Equation 5.27, line 1) implies, with (5.31):

$$\boldsymbol{v} = \frac{\frac{1}{2\lambda}\Pi_{\mathrm{span}\{\boldsymbol{E_h}\}^{\perp}}\Gamma_h(\boldsymbol{v})}{\left\|\frac{1}{2\lambda}\Pi_{\mathrm{span}\{\boldsymbol{E_h}\}^{\perp}}\Gamma_h(\boldsymbol{v})\right\|} = \frac{\Pi_{\mathrm{span}\{\boldsymbol{E_h}\}^{\perp}}\Gamma_h(\boldsymbol{v})}{\left\|\Pi_{\mathrm{span}\{\boldsymbol{E_h}\}^{\perp}}\Gamma_h(\boldsymbol{v})\right\|}, \tag{5.32}$$

which gives the basic iteration of the PING algorithm:

$$\boldsymbol{v}^{[t+1]} = \frac{\Pi_{\mathrm{span}\{\boldsymbol{E_h}\}^{\perp}}\Gamma_h\left(\boldsymbol{v}^{[t]}\right)}{\left\|\Pi_{\mathrm{span}\{\boldsymbol{E_h}\}^{\perp}}\Gamma_h\left(\boldsymbol{v}^{[t]}\right)\right\|}. \tag{5.33}$$

Direction of ascent. The purpose of this paragraph is to show that the iteration given by (5.33) follows a direction of ascent. One solution to do this is to show that the direction given by the arc $\left(\boldsymbol{v}^{[t]}, \boldsymbol{v}^{[t+1]}\right)$, $\boldsymbol{v}^{[t]}$ being the current starting point, is a direction of ascent, namely

$$\left\langle\, \boldsymbol{v}^{[t+1]} - \boldsymbol{v}^{[t]} \,\middle|\, \Gamma_h\left(\boldsymbol{v}^{[t]}\right) \right\rangle \geqslant 0.$$

First of all let us underline that by construction, at each step t of the process, $\boldsymbol{v}^{[t]}$ is orthogonal to $\operatorname{span}\{\boldsymbol{E_h}\}$. As a result, since

$$\forall t, \boldsymbol{v}^{[t]} = \Pi_{\operatorname{span}\{\boldsymbol{E_h}\}^\perp} \boldsymbol{v}^{[t]},$$

we have

$$\begin{aligned} \left\langle \boldsymbol{v}^{[t+1]} - \boldsymbol{v}^{[t]} \,\middle|\, \Gamma_h\left(\boldsymbol{v}^{[t]}\right) \right\rangle &= \left\langle \Pi_{\operatorname{span}\{\boldsymbol{E_h}\}^\perp}\left(\boldsymbol{v}^{[t+1]} - \boldsymbol{v}^{[t]}\right) \,\middle|\, \Gamma_h\left(\boldsymbol{v}^{[t]}\right) \right\rangle \\ &= \left\langle \boldsymbol{v}^{[t+1]} - \boldsymbol{v}^{[t]} \,\middle|\, \Pi_{\operatorname{span}\{\boldsymbol{E_h}\}^\perp}\Gamma_h\left(\boldsymbol{v}^{[t]}\right) \right\rangle. \end{aligned} \tag{5.34}$$

Now, (5.33) implies

$$\Pi_{\operatorname{span}\{\boldsymbol{E_h}\}^\perp}\Gamma_h\left(\boldsymbol{v}^{[t]}\right) = \boldsymbol{v}^{[t+1]} \left\| \Pi_{\operatorname{span}\{\boldsymbol{E_h}\}^\perp}\Gamma_h\left(\boldsymbol{v}^{[t]}\right) \right\| \tag{5.35}$$

and equations (5.34) and (5.35) lead to

$$\begin{aligned} \operatorname{sgn}\left(\left\langle \boldsymbol{v}^{[t+1]} - \boldsymbol{v}^{[t]} \,\middle|\, \Gamma_h\left(\boldsymbol{v}^{[t]}\right) \right\rangle\right) &= \operatorname{sgn}\left(\left\langle \boldsymbol{v}^{[t+1]} - \boldsymbol{v}^{[t]} \,\middle|\, \boldsymbol{v}^{[t+1]} \right\rangle\right) \\ &= \operatorname{sgn}\left(\left\|\boldsymbol{v}^{[t+1]}\right\|^2 - \left\langle \boldsymbol{v}^{[t]} \,\middle|\, \boldsymbol{v}^{[t+1]} \right\rangle\right) \\ &= \operatorname{sgn}\left(1 - \cos\left(\boldsymbol{v}^{[t]}, \boldsymbol{v}^{[t+1]}\right)\right). \end{aligned}$$

Therefore,

$$\left\langle \boldsymbol{v}^{[t+1]} - \boldsymbol{v}^{[t]} \,\middle|\, \Gamma_h\left(\boldsymbol{v}^{[t]}\right) \right\rangle \geqslant 0.$$

□

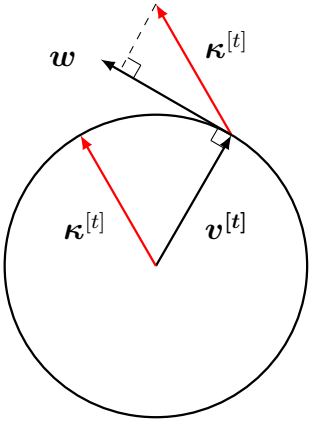
Increase of function $\mathcal{G}_h$ at each iteration. Despite the fact that iteration (5.33) follows a direction of ascent, it does not guarantee that function $\mathcal{G}_h$ actually increases on every step. Indeed, we may go too far in such a direction and overshoot the maximum. However, let us consider

$$\boldsymbol{\kappa}^{[t]} := \frac{\Pi_{\operatorname{span}\{\boldsymbol{E_h}\}^\perp}\Gamma_h\left(\boldsymbol{v}^{[t]}\right)}{\left\|\Pi_{\operatorname{span}\{\boldsymbol{E_h}\}^\perp}\Gamma_h\left(\boldsymbol{v}^{[t]}\right)\right\|}.$$

We can then show that staying close enough to the current starting point on the arc $\left(\boldsymbol{v}^{[t]}, \boldsymbol{\kappa}^{[t]}\right)$ ensures that function $\mathcal{G}_h$ increases. With this aim in mind, let ϖ be the plane tangent to the unit sphere on $\boldsymbol{v}^{[t]}$ and let $\boldsymbol{w}$ denote the unit–vector tangent to arc $\left(\boldsymbol{v}^{[t]}, \boldsymbol{\kappa}^{[t]}\right)$ on $\boldsymbol{v}^{[t]}$. Then $\exists \tau > 0 : \boldsymbol{w} = \tau \Pi_\varpi \boldsymbol{\kappa}^{[t]}$, and

$$\left\langle \boldsymbol{w} \,\middle|\, \boldsymbol{\kappa}^{[t]} \right\rangle = \tau \left\langle \Pi_\varpi \boldsymbol{\kappa}^{[t]} \,\middle|\, \boldsymbol{\kappa}^{[t]} \right\rangle = \tau \cos^2\left(\boldsymbol{\kappa}^{[t]}, \varpi\right) > 0.$$

□



Current iteration of PING. Although staying close enough to the current starting point on the arc $\left(\boldsymbol{v}^{[t]}, \boldsymbol{\kappa}^{[t]}\right)$ ensures the increase of function $\mathcal{G}_h$, staying too close can impact the convergence speed of the algorithm to reach the maximum. On the other hand, going too far from the starting point can cause the divergence of the algorithm. Bry et al. (2018) therefore propose a generic iteration of PING (Algorithm 5.2) and an alternative one (Algorithm 5.3) which take this curse into consideration.

First rank component. Component $\boldsymbol{f_1} = \boldsymbol{X}\boldsymbol{u_1}$ is obtained by solving

$$\begin{cases} \max & [\phi(\boldsymbol{u})]^s \times [\psi_{\boldsymbol{A}}(\boldsymbol{u})]^{1-s} \\ \text{subject to:} & \boldsymbol{u}^{\mathsf{T}}\boldsymbol{M}^{-1}\boldsymbol{u} = 1. \end{cases}$$

This corresponds to Program (5.25) with $h = 0$, where

- $\mathcal{J}_0(\boldsymbol{u}) = [\phi(\boldsymbol{u})]^s \times [\psi_{\boldsymbol{A_0}}(\boldsymbol{u})]^{1-s}$,
- $\boldsymbol{A_0} = \boldsymbol{A}$ (the matrix of additional explanatory variables), and
- $\boldsymbol{\Delta_0} = \boldsymbol{0}$.

In this particular case, we have $\boldsymbol{E_0} = \boldsymbol{M}^{1/2}\boldsymbol{\Delta_0} = \boldsymbol{0}$, and so:

$$\Pi_{\text{span}\{\boldsymbol{E_0}\}^\perp} = \boldsymbol{I}.$$

Higher rank components. Let $\boldsymbol{F_h} = \left[\, \boldsymbol{f_1} \mid \ldots \mid \boldsymbol{f_h} \,\right]$ be the matrix of the first h components and $\boldsymbol{A_h} = \left[\, \boldsymbol{F_h} \mid \boldsymbol{A} \,\right]$. Let $\boldsymbol{P}$ denote the weight matrix reflecting the a priori relative importance of observations ($\boldsymbol{P} = \frac{1}{n}\boldsymbol{I}_n$ if all observations are of equal importance). Component $\boldsymbol{f_{h+1}} = \boldsymbol{X}\boldsymbol{u_{h+1}}$ is obtained by solving

$$\begin{cases} \max & [\phi(\boldsymbol{u})]^s \times [\psi_{\boldsymbol{A_h}}(\boldsymbol{u})]^{1-s} \\ \text{subject to:} & \boldsymbol{u}^{\mathsf{T}}\boldsymbol{M}^{-1}\boldsymbol{u} = 1 \text{ and } \boldsymbol{F_h}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{X}\boldsymbol{u} = \boldsymbol{0}. \end{cases}$$

This corresponds to Program (5.25), where

- $\mathcal{J}_h(\boldsymbol{u}) = [\phi(\boldsymbol{u})]^s \times [\psi_{\boldsymbol{A_h}}(\boldsymbol{u})]^{1-s}$,
- $\boldsymbol{A_h} = \left[\, \boldsymbol{F_h} \mid \boldsymbol{A} \,\right]$, and
- $\boldsymbol{\Delta_h} = \boldsymbol{X}^\top \boldsymbol{P} \boldsymbol{F_h}$.

Algorithm 5.2: The PING algorithm (generic iteration).

while *convergence of* $\boldsymbol{v}$ *non reached* **do**

$$\boldsymbol{\kappa}^{[t]} \leftarrow \frac{\Pi_{\text{span}\{\boldsymbol{E_h}\}^\perp} \Gamma_h\left(\boldsymbol{v}^{[t]}\right)}{\left\| \Pi_{\text{span}\{\boldsymbol{E_h}\}^\perp} \Gamma_h\left(\boldsymbol{v}^{[t]}\right) \right\|}.$$

A Newton–Raphson unidimensional maximisation procedure is used to find the maximum of $\mathcal{G}_h(\boldsymbol{v})$ on the arc $\left(\boldsymbol{v}^{[t]}, \boldsymbol{\kappa}^{[t]}\right)$ and take it as $\boldsymbol{v}^{[t+1]}$.

$t \longleftarrow t+1$

end

Algorithm 5.3: The PING algorithm (alternative generic iteration).

while *convergence of* $\boldsymbol{v}$ *non reached* **do**

$$\boldsymbol{m} \leftarrow \frac{\Pi_{\text{span}\{\boldsymbol{E_h}\}^\perp} \Gamma_h\left(\boldsymbol{v}^{[t]}\right)}{\left\| \Pi_{\text{span}\{\boldsymbol{E_h}\}^\perp} \Gamma_h\left(\boldsymbol{v}^{[t]}\right) \right\|}$$

while $\mathcal{G}_h(\boldsymbol{m}) < \mathcal{G}_h\left(\boldsymbol{v}^{[t]}\right)$ **do**

$$\boldsymbol{m} \leftarrow \frac{\boldsymbol{v}^{[t]} + \boldsymbol{m}}{\|\boldsymbol{v}^{[t]} + \boldsymbol{m}\|}$$

end

$\boldsymbol{v}^{[t+1]} \leftarrow \boldsymbol{m}$

$t \leftarrow t+1$

end

Initialisation. To quickly find $\boldsymbol{f_1}$, algorithm PING is initialised with the first PLS component of the responses on $\boldsymbol{X}$. Likewise, for $h \geqslant 2$, PING is initialised with the first PLS component of the responses on $\boldsymbol{X}$ deflated on components $\boldsymbol{F_{h-1}} = \left[\, \boldsymbol{f_1} \mid \ldots \mid \boldsymbol{f_{h-1}} \,\right]$.

5.8.4 Models with offset

This section is inspired by Bry et al. (2016). In many situations, we know that response variable $\boldsymbol{y}$ is proportional to a certain variable $\boldsymbol{o}$, which does not have the same status as the set of explanatory variables $\boldsymbol{X}$. Indeed, the values of variable $\boldsymbol{o}$ are simply added to the linear predictor of the target, and the associated coefficient does not have to be estimated since is assumed to be 1. Such a framework is referred to as "model with offset".

The tree species abundance dataset *Genus* (Section 5.6), from which the development of the mixed–SCGLR method orginated, presents such a configuration. As a reminder, abundance $y_{k,i}$ of species k on plot i has been modelled by

$$\begin{cases} y_{k,i} \,|\, \boldsymbol{\xi_k} \sim \mathcal{P}\left(\lambda = \mu^{\xi}_{k,i}\right) \\ \mu^{\xi}_{k,i} = \mathbb{E}\left(y_{k,i} \,|\, \boldsymbol{\xi_k}\right) = \exp\left[\eta^{\xi}_{k,i}\right] = \exp\left[\sum_{j=1}^{K} \left(\boldsymbol{x}^{\mathrm{T}}_{\boldsymbol{i:}} \boldsymbol{u_j}\right) \gamma_{k,j} + \boldsymbol{a}^{\mathrm{T}}_{\boldsymbol{i:}} \boldsymbol{\delta_k} + \boldsymbol{u}^{\mathrm{T}}_{\boldsymbol{i:}} \boldsymbol{\xi_k}\right]. \end{cases} \tag{5.36}$$

$\boldsymbol{\beta_k} = \sum_{j=1}^{K} \boldsymbol{u_j} \gamma_{k,j}$ is the vector of regularised fixed effects, $\boldsymbol{\delta_k}$ the vector associated with additional explanatory variables, and $\boldsymbol{\xi_k}$ the vector of random effects. However, in practice, the plots do not have the same surface area and it is necessary to take this feature into account. The abundance of a tree species on a given plot is then assumed to be proportional to the surface area of that plot. Let o_i be the surface area of plot i. Conditional expectation $\mu^{\xi}_{k,i}$ involved in modelling (5.36) is then replaced by

$$\widetilde{\mu}^{\xi}_{k,i} = \mathbb{E}\left(y_{k,i} \,|\, \boldsymbol{\xi_k}, o_i\right) = o_i \times \exp\left(\eta^{\xi}_{k,i}\right) = \exp\left(\eta^{\xi}_{k,i} + \ln(o_i)\right),$$

so that variable $\boldsymbol{o}$ is considered as an offset.

Compared to Algorithm 5.1, the main change concerns the update of conditional expectation $\widetilde{\mu}^{\xi}_{k,i}$. Algorithm 5.4 summarises the mixed–SCGLR method with offset, in the case of a Poisson regression with log–link.

Algorithm 5.4: The single component mixed–SCGLR algorithm with an offset (The case of a Poisson regression with log–link).

while *convergence criterion not reached* **do**

Step 1: Computing component $\boldsymbol{f}^{[t+1]}$.

Steps 2 – 3: Updating parameter estimates $\gamma_k^{[t+1]}$, $\boldsymbol{\delta}_{\boldsymbol{k}}^{[t+1]}$, $\sigma_k^{2[t+1]}$, and updating predictions $\boldsymbol{\xi}_{\boldsymbol{k}}^{[t+1]}$.

Step 4: Updating the working variables and the weighting matrices. For each $k \in \{1, \ldots, q\}$, set

$$\begin{aligned}
\boldsymbol{\eta}_{\boldsymbol{k}}^{\boldsymbol{\xi}[t+1]} &= \boldsymbol{f}^{[t+1]}\gamma_k^{[t+1]} + \boldsymbol{A}\boldsymbol{\delta}_{\boldsymbol{k}}^{[t+1]} + \boldsymbol{U}\boldsymbol{\xi}_{\boldsymbol{k}}^{[t+1]} \\
\widetilde{\mu}_{k,i}^{\xi\ [t+1]} &= \exp\left(\eta_{k,i}^{\xi\ [t+1]} + \log(o_i)\right), \quad i \in \{1, \ldots, n\} \\
z_{k,i}^{\xi\ [t+1]} &= \eta_{k,i}^{\xi\ [t+1]} + \frac{y_{k,i} - \widetilde{\mu}_{k,i}^{\xi\ [t+1]}}{\widetilde{\mu}_{k,i}^{\xi\ [t+1]}}, \quad i \in \{1, \ldots, n\} \\
\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}[t+1]} &= \mathbf{Diag}\left(\widetilde{\mu}_{k,i}^{\xi\ [t+1]}\right)_{i=1,\ldots,n}
\end{aligned}$$

Incrementing: $t \longleftarrow t + 1$

end

VI

Regularisation of GLMMs with an autoregressive random time–specific effect

An individual can't be judged by his group mean.
— Stephen Jay Gould

Contents

6.1 Introduction

This chapter is essentially inspired by the reading of the articles by Eliot et al. (2011) and Karlsson and Skoglund (2004), and has resulted in an paper currently being written. The article by Eliot et al. (2011) extends the ridge regression to linear mixed models. The estimation method they suggest is based on a L_2–penalised EM algorithm, which handles the potentially high correlations between explanatory variables. In order to take into account the dependence induced by repeated measures on several individuals over time, the model includes an individual–specific random effect. Although this paper focuses on longitudinal data, the model does not consider a time–specific random effect common to all individuals. Complementarily, Karlsson and Skoglund (2004) consider both individual– and time–specific random effects. The latter are viewed as latent phenomena (not accounted for by the explanatory variables) affecting all individuals, and persistent over time. However, they do not consider any situation where it is necessary to regularise the model.

Therefore as a first step in the LMM framework, it appears necessary to adapt the L_2–penalised EM of Eliot et al. (2011) to the case of an autoregressive time–specific random effect. Besides, most of the existing regularisation methods — including those involving the use of the EM algorithm — are based on penalised likelihood. However, these methods suffer from the strong correlations among explanatory variables instead of taking advantage of them. We then propose a Supervised Component–based regularised EM (SCEM) as an alternative: instead of subtracting a penalty term to the likelihood, we rather suggest the possibility to add a bonus term favouring the alignment of components on the most interpretable directions in the explanatory subspace. Finally, an extension of the previous strategies to GLMMs with an autoregressive time–specific random effect is also proposed. Inspired by the iterative procedure proposed by Schall (1991), we keep the same linearisation step. But in order to take into account both the strong correlations in $\boldsymbol{X}$ and the autocorrelated structure of the time–specific random effect, we suggest to replace the usual estimation step (involving an Henderson's system) with an L_2–penalised or an SCEM algorithm.

The chapter is organised as follows. First, Section 6.2 formalises the model, sets the main notations used throughout the chapter, and gives some situations in which serially correlated time–specific random effects arise naturally. Since this chapter focuses on the EM algorithm, Section 6.3 recalls different ways of conceptualising it. In Section 6.4 we reinterpret the penalised EM algorithm as a double maximisation, and we introduce the alternative SC–based regularisation. Sections 6.5 and 6.6 respectively give the technical details of the ridge–

and SC–based EMs for LMMs with an autocorrelated time–specific random effect. Section 6.7 designs their extensions to GLMMs. Finally, comparative results on simulated data are presented in Section 6.8.

6.2 Model definition and notations

In this section, we present the main hypotheses of the GLMM framework considered in this work, in particular the random effects distributions. For the sake of simplicity, we consider balanced panel data with q_1 individuals, each of them being observed at the same q_2 time–points. We denote by $n = q_1 \times q_2$ the total number of observations. Let $\boldsymbol{X}$ be the $n \times p$ fixed–effect design matrix, and $\boldsymbol{U}$ the $n \times q$ random–effect design matrix, where $q = q_1 + q_2$. Let also Y be the n–dimensional random response vector, $\boldsymbol{\beta}$ the p–dimensional vector of fixed effects, and $\boldsymbol{\xi}$ the q–dimensional vector of random effects. We observe a realisation $\boldsymbol{y}$ of Y, but $\boldsymbol{\xi}$ is not observed.

We conventionally assume that:

(i) Given $\boldsymbol{\xi}$, the Y_i, $i \in \{1, \ldots, n\}$, are independent and their distribution belongs to the exponential family.

(ii) The conditional mean $\mu_i = \mathbb{E}(Y_i \,|\, \boldsymbol{\xi})$ depends on $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ through the link function, g, and the linear predictor, $\eta_i = \boldsymbol{x}_{i:}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{u}_{i:}^{\mathrm{T}}\boldsymbol{\xi}$, with $\eta_i = g(\mu_i)$.

Less conventionally, we consider a model with two random effects $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$ with different roles and distributions:

(i) $\boldsymbol{\xi_1}$ is the individual–specific random effect, which links all the observations of an individual to the same realisation of the random effect. We suppose:

$$\boldsymbol{\xi_1} \sim \mathcal{N}_{q_1}\left(0,\, \sigma_1^2 \boldsymbol{A_1}\right),$$

with $\boldsymbol{A_1}$ a known matrix (by default here $\boldsymbol{A_1} = \boldsymbol{I}_{q_1}$ since individuals are assumed independent), and σ_1^2 the unknown "individual" variance component.

(ii) $\boldsymbol{\xi_2}$ is the serially correlated time–specific random effect which links all the observations at time t to the same realisation of the random effect. It is common to all the individuals, and can be viewed as some latent

phenomenon evolving in time and not measured in the explanatory variables. As these effects tend to persist over time, we model them through an order–1 stationary Gaussian autoregressive process (AR(1)). More precisely, we assume that on the first time–point ($t = 1$),

$$\xi_{2,1} \sim \mathcal{N}_1\left(0,\, \frac{\sigma_2^2}{1-\rho^2}\right), \tag{6.1}$$

and for each $t \in \{2, \ldots, q_2\}$,

$$\begin{aligned} \xi_{2,t} &= \rho\, \xi_{2,t-1} + \nu_t, \\ \nu_t &\sim \mathcal{N}_1\left(0,\, \sigma_2^2\right). \end{aligned} \tag{6.2}$$

In (6.1) and (6.2), ρ is the unknown parameter of the AR(1) and σ_2^2 the unknown "time–specific" variance component. The order–1 autoregressive correlation structure appears explicitly in the entire distribution of vector $\boldsymbol{\xi_2}$:

$$\boldsymbol{\xi_2} \sim \mathcal{N}_{q_2}\Big(\mathbf{0},\, \sigma_2^2 \boldsymbol{A_2}(\rho)\Big), \text{ with}$$

$$\boldsymbol{A_2}(\rho) = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{q_2-1} \\ \rho & 1 & \rho & \cdots & \rho^{q_2-2} \\ \rho^2 & \rho & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \rho \\ \rho^{q_2-1} & \rho^{q_2-2} & \cdots & \rho & 1 \end{pmatrix}.$$

Finally, $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$ are assumed independent. Let us denote $\boldsymbol{\xi} = (\boldsymbol{\xi_1}^\top, \boldsymbol{\xi_2}^\top)^\top$, $\boldsymbol{U_1} = \boldsymbol{I}_{q_1} \otimes \mathbf{1}_{q_2}$, $\boldsymbol{U_2} = \mathbf{1}_{q_1} \otimes \boldsymbol{I}_{q_2}$ and $\boldsymbol{U} = [\boldsymbol{U_1} \,|\boldsymbol{U_2}]$. Linear predictor $\boldsymbol{\eta}$ can be matricially written:

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\xi}.$$

This kind of two–way random effects model arises naturally in many areas. In a non–exhaustive way, let us just give three of them.

(a) In an economic context for instance, if one studies the trends in the number of loans granted by several banks, the bank–specific effect has to be taken into account. But we must not forget that all banks share a common economic context — seen as a latent phenomenon — which tends to persist over time (e.g. a common supply–demand context, the general degree of confidence, the expected depreciation of the currency, etc).

(b) In forest ecology, if one studies the growth of many trees over time, the specific potential of each tree has to be taken into account, but also the

common ecological environment of the zone (e.g. close weather conditions such as light exposure, rainfall or moisture regime, presence of pests, etc), which is often too complex to be directly observed through the explanatory variables, and has a certain temporal inertia.

(c) Lastly in epidemiology, clinical studies often involve administering a new treatment to several patients with a certain disease. The goal is to observe the speed of the healing process by means of several visits scheduled over time. It is obvious to consider the dependence of the measures related to the same patient. But it may be necessary to take into account the latent state evolution shared by all patients. Indeed, since they participate in the same clinical study — often double–blind —, they are likely to share common conditions, including the same common standard human physiological struggle.

6.3 Brief review of the EM algorithm

Since we consider a model with non–observed random effects, the EM algorithm seems to be a good option to perform maximum likelihood estimation. Initially introduced by Dempster et al. (1977), the EM algorithm has proven to be an essential tool to find (local) maximum likelihood parameters of a statistical model in cases where the equations can not be solved directly (e.g. in a context of latent variable or missing data). In the following, the key points of the EM algorithm and other ways of conceptualising it are presented. For a comprehensive overview on the EM algorithm, we refer the reader to the nice informal tutorial by Roche (2011) and the references therein.

Generically, assume that $\boldsymbol{y}$ refers to the observed realisation of some random response Y, while on the contrary $\boldsymbol{\xi} \in \Xi$ refers to unobserved realisations (the random effects in our case). It is further assumed that given $\boldsymbol{y}$ and $\boldsymbol{\xi}$, the complete log–likelihood, ℓ^{c}, is is a function of parameter $\boldsymbol{\theta}$ (in our case $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_1^2, \sigma_2^2, \rho)$).

6.3.1 A sequence of parameters increasing the likelihood

The EM algorithm can be seen as an iterative procedure whose purpose is to construct a sequence of parameters $\left\{\boldsymbol{\theta}^{[t]}\right\}_{t\geqslant 0}$ so that the log–likelihood $\ell\left(\boldsymbol{\theta}^{[t]}; \boldsymbol{y}\right)$ increases with t. To this end, it can be noted that

$$
\begin{aligned}
\ell\left(\boldsymbol{\theta}^{[t+1]}; \boldsymbol{y}\right) - \ell\left(\boldsymbol{\theta}^{[t]}; \boldsymbol{y}\right) &= \log\left[p\left(\boldsymbol{y}; \boldsymbol{\theta}^{[t+1]}\right)\right] - \log\left[p\left(\boldsymbol{y}; \boldsymbol{\theta}^{[t]}\right)\right] \\
&= \log\left[\frac{p\left(\boldsymbol{y}; \boldsymbol{\theta}^{[t+1]}\right)}{p\left(\boldsymbol{y}; \boldsymbol{\theta}^{[t]}\right)}\right] \\
&= \log\left[\int_{\Xi} \frac{p\left(\boldsymbol{y}, \boldsymbol{\xi}; \boldsymbol{\theta}^{[t+1]}\right)}{p\left(\boldsymbol{y}; \boldsymbol{\theta}^{[t]}\right)} \, \mathrm{d}\boldsymbol{\xi}\right] \\
&= \log\left[\int_{\Xi} \frac{p\left(\boldsymbol{y}, \boldsymbol{\xi}; \boldsymbol{\theta}^{[t+1]}\right)}{p\left(\boldsymbol{y}, \boldsymbol{\xi}; \boldsymbol{\theta}^{[t]}\right)} \, p\left(\boldsymbol{\xi} | \boldsymbol{y}; \boldsymbol{\theta}^{[t]}\right) \, \mathrm{d}\boldsymbol{\xi}\right] \qquad (6.3) \\
&\geqslant \int_{\Xi} \log\left[\frac{p\left(\boldsymbol{y}, \boldsymbol{\xi}; \boldsymbol{\theta}^{[t+1]}\right)}{p\left(\boldsymbol{y}, \boldsymbol{\xi}; \boldsymbol{\theta}^{[t]}\right)}\right] \, p\left(\boldsymbol{\xi} | \boldsymbol{y}; \boldsymbol{\theta}^{[t]}\right) \, \mathrm{d}\boldsymbol{\xi} \qquad (6.4) \\
&:= \mathcal{Q}\left(\boldsymbol{\theta}^{[t+1]}, \boldsymbol{\theta}^{[t]}\right).
\end{aligned}
$$

Step (6.3) simply results from the fact that

$$
p\left(\boldsymbol{y}; \boldsymbol{\theta}^{[t]}\right) = \frac{p\left(\boldsymbol{y}, \boldsymbol{\xi}; \boldsymbol{\theta}^{[t]}\right)}{p\left(\boldsymbol{\xi} | \boldsymbol{y}; \boldsymbol{\theta}^{[t]}\right)}
$$

and step (6.4) follows from the concavity of the logarithm function and Jensen's inequality.

In a nutshell, finding $\boldsymbol{\theta}^{[t+1]}$ such that $\mathcal{Q}\left(\boldsymbol{\theta}^{[t+1]}, \boldsymbol{\theta}^{[t]}\right) > 0$ and iterating the process defines an EM algorithm. $\boldsymbol{\theta}^{[t+1]}$ is usually chosen to maximise $\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{[t]}\right)$ with respect to $\boldsymbol{\theta}$.

6.3.2 Initial formulation of the EM

Expanding (6.4) allows to rewrite the objective function, $\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{[t]}\right)$, as a difference:

$$
\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{[t]}\right) = \mathcal{Q}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) - \mathcal{Q}\left(\boldsymbol{\theta}^{[t]} \,|\, \boldsymbol{\theta}^{[t]}\right),
$$

where

$$
\begin{cases}
\mathcal{Q}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) := \displaystyle\int_{\Xi} \log\left[p\left(\boldsymbol{y}, \boldsymbol{\xi}; \boldsymbol{\theta}\right)\right] \, p\left(\boldsymbol{\xi} | \boldsymbol{y}; \boldsymbol{\theta}^{[t]}\right) \, \mathrm{d}\boldsymbol{\xi} \\
\mathcal{Q}\left(\boldsymbol{\theta}^{[t]} \,|\, \boldsymbol{\theta}^{[t]}\right) := \displaystyle\int_{\Xi} \log\left[p\left(\boldsymbol{y}, \boldsymbol{\xi}; \boldsymbol{\theta}^{[t]}\right)\right] \, p\left(\boldsymbol{\xi} | \boldsymbol{y}; \boldsymbol{\theta}^{[t]}\right) \, \mathrm{d}\boldsymbol{\xi}.
\end{cases}
$$

Now, $\mathcal{Q}\left(\boldsymbol{\theta}^{[t]} \,|\, \boldsymbol{\theta}^{[t]}\right)$ does not depend on $\boldsymbol{\theta}$. Maximising $\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{[t]}\right)$ with respect to $\boldsymbol{\theta}$ is then equivalent to maximising $\mathcal{Q}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right)$, which can be expressed as the expectation of the complete log–likelihood under the distribution of $\boldsymbol{\xi}|\boldsymbol{y}$ at the current value $\boldsymbol{\theta}^{[t]}$. Algorithm 6.1 summarises the most widespread form of the EM algorithm, as initially formulated by Dempster et al. (1977).

Algorithm 6.1: The EM algorithm (initial formulation).

Start with an initial guess $\boldsymbol{\theta}^{[0]}$, and set $t = 0$
while *some convergence criterion not reached* **do**
  E–step: Compute $\mathcal{Q}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) = \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\ell^{\mathrm{c}}\left(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}\right) \,|\, \boldsymbol{\theta}^{[t]}\right]$
  M–step: Set $\boldsymbol{\theta}^{[t+1]} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right)$
  $t \leftarrow t + 1$
end

In their paper, Dempster et al. (1977) prove that each iteration provided by Algorithm 6.1 increases the log–likelihood, namely

$$\forall t, \; \ell\left(\boldsymbol{\theta}^{[t+1]}; \boldsymbol{y}\right) \geqslant \ell\left(\boldsymbol{\theta}^{[t]}; \boldsymbol{y}\right).$$

Under mild conditions, they also prove the convergence of the sequence $\left\{\boldsymbol{\theta}^{[t]}\right\}_{t \geqslant 0}$ toward some $\boldsymbol{\theta}^{\star}$. However, $\boldsymbol{\theta}^{\star}$ may be a saddle point of the log–likelihood. Wu (1983) presents more sophisticated conditions (however difficult to verify in practice) which ensure the convergence of the sequence of parameters toward a local maximum of the likelihood.

6.3.3 EM as a proximal point algorithm

In order to maximise a concave function $\mathcal{J}(\cdot)$, the proximal point algorithm (Martinet, 1970; Rockafellar, 1976) is an iterative procedure which can be written

$$\boldsymbol{\theta}^{[t+1]} = \arg\max_{\boldsymbol{\theta}} \left\{ \mathcal{J}(\boldsymbol{\theta}) - \lambda_t \left\| \boldsymbol{\theta} - \boldsymbol{\theta}^{[t]} \right\|^2 \right\}. \tag{6.6}$$

The procedure includes a quadratic penalty, $\left\| \boldsymbol{\theta} - \boldsymbol{\theta}^{[t]} \right\|^2$, which is relaxed by a sequence of positive parameters $\{\lambda_t\}_{t \geqslant 0}$.

With a view towards maximum likelihood estimation, Chrétien and Hero (2000) propose to replace the quadratic penalty in (6.6) by a Kullback–type

information measure. Denoting $\mathcal{D}\left(\boldsymbol{\theta}^{[t]}, \boldsymbol{\theta}\right)$ the Kullback–Leibler divergence from $p\left(\boldsymbol{\xi}|\boldsymbol{y};\boldsymbol{\theta}\right)$ to $p\left(\boldsymbol{\xi}|\boldsymbol{y};\boldsymbol{\theta}^{[t]}\right)$, i.e.

$$\begin{aligned}\mathcal{D}\left(\boldsymbol{\theta}^{[t]}, \boldsymbol{\theta}\right) &= D_{\mathrm{KL}}\left[p\left(\boldsymbol{\xi}|\boldsymbol{y};\boldsymbol{\theta}^{[t]}\right) \parallel p\left(\boldsymbol{\xi}|\boldsymbol{y};\boldsymbol{\theta}\right)\right] \\ &= \int_{\Xi} \log\left[\frac{p\left(\boldsymbol{\xi}|\boldsymbol{y};\boldsymbol{\theta}^{[t]}\right)}{p\left(\boldsymbol{\xi}|\boldsymbol{y};\boldsymbol{\theta}\right)}\right] p\left(\boldsymbol{\xi}|\boldsymbol{y};\boldsymbol{\theta}^{[t]}\right) \mathrm{d}\boldsymbol{\xi},\end{aligned}$$

they suggest an iterative scheme of the form

$$\boldsymbol{\theta}^{[t+1]} = \arg\max_{\boldsymbol{\theta}} \left\{ \ell\left(\boldsymbol{\theta};\boldsymbol{y}\right) - \lambda_t \mathcal{D}\left(\boldsymbol{\theta}^{[t]}, \boldsymbol{\theta}\right) \right\}. \tag{6.7}$$

The fact is that the EM, as described by Algorithm 6.1, is equivalent to the update rule (6.7) with $\lambda_t = 1$ for all t. Indeed, the objective function $\mathcal{Q}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right)$ can be rewritten as

$$\mathcal{Q}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) = \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\ell^{\mathrm{c}}\left(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{\xi}\right) \,|\, \boldsymbol{\theta}^{[t]}\right] = \ell\left(\boldsymbol{\theta};\boldsymbol{y}\right) + \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\ell\left(\boldsymbol{\theta};\boldsymbol{\xi}|\boldsymbol{y}\right) \,|\, \boldsymbol{\theta}^{[t]}\right].$$

We then have

$$\boldsymbol{\theta}^{[t+1]} = \arg\max_{\boldsymbol{\theta}} \left\{ \ell\left(\boldsymbol{\theta};\boldsymbol{y}\right) + \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\ell\left(\boldsymbol{\theta};\boldsymbol{\xi}|\boldsymbol{y}\right) \,|\, \boldsymbol{\theta}^{[t]}\right] - \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\ell\left(\boldsymbol{\theta}^{[t]};\boldsymbol{\xi}|\boldsymbol{y}\right) \,|\, \boldsymbol{\theta}^{[t]}\right] \right\}$$

since the last term is constant in $\boldsymbol{\theta}$. We finally obtain

$$\begin{aligned}\boldsymbol{\theta}^{[t+1]} &= \arg\max_{\boldsymbol{\theta}} \left\{ \ell\left(\boldsymbol{\theta};\boldsymbol{y}\right) - \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\log\left(\frac{p\left(\boldsymbol{\xi}|\boldsymbol{y};\boldsymbol{\theta}^{[t]}\right)}{p\left(\boldsymbol{\xi}|\boldsymbol{y};\boldsymbol{\theta}\right)}\right) \,|\, \boldsymbol{\theta}^{[t]}\right] \right\} \\ &= \arg\max_{\boldsymbol{\theta}} \left\{ \ell\left(\boldsymbol{\theta};\boldsymbol{y}\right) - \mathcal{D}\left(\boldsymbol{\theta}^{[t]}, \boldsymbol{\theta}\right) \right\}.\end{aligned}$$

□

The reinterpretation of the EM as a proximal point algorithm is very powerful in that it allows to establish convergence results. Under rather restrictive regularity conditions, Chrétien and Hero (2000) show that the sequence $\left\{\boldsymbol{\theta}^{[t]}\right\}_{t\geqslant 0}$ linearly converges to the global maximum of the likelihood, provided that $\lim_{t\to+\infty} \lambda_t = \lambda^\star \in [0, +\infty)$. The convergence rate may even be superlinear if $\lambda^\star = 0$. Unfortunately, as soon as $\lambda_t \neq 1$, update rule (6.7) is generally intractable and requires approximations of the log–likelihood and the Kullback–Leibler divergence.

6.3.4 EM as a double maximisation algorithm

As suggested by Neal and Hinton (1998), another way of conceptualising EM is to reinterpret the E–step as another maximisation. Let us note B a lower bound of the log–likelihood, ℓ, defined by

$$B\left(q, \boldsymbol{\theta}\right) = \mathbb{E}_q\left[\ell^{\mathsf{c}}\left(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}\right)\right] + H(q), \tag{6.8}$$

where q denotes any probability distribution function and $H(q)$ its entropy, which writes $H(q) = -\int_{\Xi} \log\left[q\left(\boldsymbol{\xi}; \boldsymbol{\theta}\right)\right] q\left(\boldsymbol{\xi}; \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{\xi}$. Then, the iterations given by Algorithm 6.1 are equivalent to the iterations given by Algorithm 6.2.

Algorithm 6.2: EM as an iterative double–maximisation algorithm.

Start with an initial guess $\boldsymbol{\theta}^{[0]}$, and set $t = 0$
while *some convergence criterion not reached* **do**
  E–step: Set $q^{[t+1]} = \arg\max_{q} B\left(q, \boldsymbol{\theta}^{[t]}\right)$
  M–step: Set $\boldsymbol{\theta}^{[t+1]} = \arg\max_{\boldsymbol{\theta}} B\left(q^{[t+1]}, \boldsymbol{\theta}\right)$
  $t \leftarrow t + 1$
end

Section 6.4.1 shows this equivalence within the more general framework of a penalised log–likelihood.

6.3.5 Extensions of the EM algorithm

When the log–likelihood has saddle points or plateaus, it is well known that the EM algorithm may be very sensitive to the initial guess $\boldsymbol{\theta}^{[0]}$ and may exhibit slow convergence rate. In addition, there are situations where the auxiliary function $\mathcal{Q}\left(\cdot, \cdot\right)$ is intractable. So many deterministic and stochastic extensions have been proposed for circumventing these limitations that an exhaustive state of art could not be presented here.

Nevertheless, some deterministic schemes mainly motivated by convergence speed consideration can be mentioned, such as the Accelerated EM Meng and Rubin (1993) and the Expectation Conditional Maximisation (Jamshidian and Jennrich, 1993). In order to avoid convergence towards a saddle–point, stochastic extensions have been developed (e.g. the Stochastic EM by Celeux (1985), and the Stochastic Approximation type EM by Celeux et al. (1995)). The Monte Carlo EM proposed by McCulloch (1997) is designed to deal with the intractability of the auxiliary function that occurs for GLMM estimation.

In the following, we focus on the use of the EM algorithm in the context of regularised and penalised likelihood estimation.

6.4 Regularised EM algorithms

6.4.1 Penalised EM as a double maximisation

Seeing the EM algorithm as an iterative double–maximisation procedure is a good point of view for its extension to the case of a penalised log–likelihood. In the following, ℓ_{pen} denotes the penalised log–likelihood, $\text{pen}(\boldsymbol{\theta})$ the penalty term and $\lambda \geqslant 0$ its associated shrinkage parameter. We then have

$$\begin{aligned}
\ell_{\text{pen}}(\boldsymbol{\theta}; \boldsymbol{y}) &= \ell(\boldsymbol{\theta}; \boldsymbol{y}) - \lambda\,\text{pen}(\boldsymbol{\theta}) = \log\left[p(\boldsymbol{y}; \boldsymbol{\theta})\right] - \lambda\,\text{pen}(\boldsymbol{\theta}) \\
&= \log\left[\int_{\Xi} p(\boldsymbol{y}, \boldsymbol{\xi}; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\xi}\right] - \lambda\,\text{pen}(\boldsymbol{\theta}) \\
&= \log\left[\int_{\Xi} \frac{p(\boldsymbol{y}, \boldsymbol{\xi}; \boldsymbol{\theta})}{q(\boldsymbol{\xi}; \boldsymbol{\theta})}\, q(\boldsymbol{\xi}; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\xi}\right] - \lambda\,\text{pen}(\boldsymbol{\theta}) \\
&\overset{\text{(Jensen)}}{\geqslant} \int_{\Xi} \log\left[\frac{p(\boldsymbol{y}, \boldsymbol{\xi}; \boldsymbol{\theta})}{q(\boldsymbol{\xi}; \boldsymbol{\theta})}\right] q(\boldsymbol{\xi}; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\xi} - \lambda\,\text{pen}(\boldsymbol{\theta}) \\
&=: B_{\text{pen}}(q, \boldsymbol{\theta}).
\end{aligned} \tag{6.9}$$

Thus, the lower bound (6.8) writes now for the penalised log–likelihood:

$$\begin{aligned}
B_{\text{pen}}(q, \boldsymbol{\theta}) &= \int_{\Xi} \ell^{\text{c}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi})\, q(\boldsymbol{\xi}; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\xi} \\
&\qquad - \int_{\Xi} \log\left[q(\boldsymbol{\xi}; \boldsymbol{\theta})\right] q(\boldsymbol{\xi}; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\xi} - \lambda\,\text{pen}(\boldsymbol{\theta}) \\
&= \mathbb{E}_q\left[\ell^{\text{c}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi})\right] + H(q) - \lambda\,\text{pen}(\boldsymbol{\theta}) \\
&= \mathbb{E}_q\left[\ell^{\text{c}}_{\text{pen}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi})\right] + H(q),
\end{aligned} \tag{6.10}$$

with $\ell^{\text{c}}_{\text{pen}}$ the complete penalised log–likelihood. Thanks to (6.9), the direct relation between B_{pen} and ℓ_{pen} is given by

$$\begin{aligned}
B_{\text{pen}}(q, \boldsymbol{\theta}) &= \int_{\Xi} \log\left[\frac{p(\boldsymbol{\xi}|\boldsymbol{y}; \boldsymbol{\theta})\; p(\boldsymbol{y}; \boldsymbol{\theta})}{q(\boldsymbol{\xi}; \boldsymbol{\theta})}\right] q(\boldsymbol{\xi}; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\xi} - \lambda\,\text{pen}(\boldsymbol{\theta}) \\
&= \ell_{\text{pen}}(\boldsymbol{y}; \boldsymbol{\theta}) - \int_{\Xi} \log\left[\frac{q(\boldsymbol{\xi}; \boldsymbol{\theta})}{p(\boldsymbol{\xi}|\boldsymbol{y}; \boldsymbol{\theta})}\right] q(\boldsymbol{\xi}; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\xi} \\
&= \ell_{\text{pen}}(\boldsymbol{y}; \boldsymbol{\theta}) - D_{\text{KL}}\Big(q(\boldsymbol{\xi}; \boldsymbol{\theta}) \parallel p(\boldsymbol{\xi}|\boldsymbol{y}; \boldsymbol{\theta})\Big),
\end{aligned} \tag{6.11}$$

where $D_{\mathrm{KL}}(q \,\|\, p)$ is the Kullback–Leibler divergence from p to q. The procedure described by Algorithm 6.2 can then be rewritten using B_{pen} (Equation 6.10) instead of B (Equation 6.8).

Penalised E–step The penalised expectation step t is to set

$$\begin{aligned} q^{[t+1]} &= \arg\max_{q} B_{\mathrm{pen}}\left(q, \boldsymbol{\theta}^{[t]}\right) \\ &= \arg\min_{q} D_{\mathrm{KL}}\left(q\left(\boldsymbol{\xi}; \boldsymbol{\theta}^{[t]}\right) \,\|\, p\left(\boldsymbol{\xi}|\boldsymbol{y}; \boldsymbol{\theta}^{[t]}\right)\right) \\ &= \arg\min_{q} \int_{\Xi} \log\left[\frac{q\left(\boldsymbol{\xi}; \boldsymbol{\theta}^{[t]}\right)}{p\left(\boldsymbol{\xi}|\boldsymbol{y}; \boldsymbol{\theta}^{[t]}\right)}\right] q\left(\boldsymbol{\xi}; \boldsymbol{\theta}^{[t]}\right) \mathrm{d}\boldsymbol{\xi}. \end{aligned}$$

However, the Kullback–Leibler divergence is always non–negative and equals zero when the two distributions coincide. Thus, we simply set

$$q^{[t+1]}\left(\boldsymbol{\xi}; \boldsymbol{\theta}^{[t]}\right) = p\left(\boldsymbol{\xi}|\boldsymbol{y}; \boldsymbol{\theta}^{[t]}\right). \tag{6.12}$$

Penalised M–step Given $q^{[t+1]}$ defined by (6.12), the penalised maximisation step is to set

$$\begin{aligned} \boldsymbol{\theta}^{[t+1]} &= \arg\max_{\boldsymbol{\theta}} \left\{B_{\mathrm{pen}}\left(q^{[t+1]}, \boldsymbol{\theta}\right)\right\} \\ &= \arg\max_{\boldsymbol{\theta}} \left\{\int_{\Xi} \log\left[\frac{p\left(\boldsymbol{y}, \boldsymbol{\xi}; \boldsymbol{\theta}\right)}{p\left(\boldsymbol{\xi}|\boldsymbol{y}; \boldsymbol{\theta}^{[t]}\right)}\right] p\left(\boldsymbol{\xi}|\boldsymbol{y}; \boldsymbol{\theta}^{[t]}\right) \mathrm{d}\boldsymbol{\xi} - \lambda\,\mathrm{pen}\left(\boldsymbol{\theta}\right)\right\} \\ &= \arg\max_{\boldsymbol{\theta}} \left\{\int_{\Xi} \log\left[p\left(\boldsymbol{y}, \boldsymbol{\xi}; \boldsymbol{\theta}\right)\right] p\left(\boldsymbol{\xi}|\boldsymbol{y}; \boldsymbol{\theta}^{[t]}\right) \mathrm{d}\boldsymbol{\xi} - \lambda\,\mathrm{pen}\left(\boldsymbol{\theta}\right)\right\} \\ &= \arg\max_{\boldsymbol{\theta}} \left\{\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\ell^{\mathrm{c}}_{\mathrm{pen}}\left(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}\right) \,|\, \boldsymbol{\theta}^{[t]}\right]\right\}. \end{aligned}$$

Equation 6.11 ensures that on the fixed point of the algorithm, since $q^{[\infty]} = p\left(\boldsymbol{\xi}|\boldsymbol{y}; \boldsymbol{\theta}^{[\infty]}\right)$, $D_{\mathrm{KL}} = 0$, so $B_{\mathrm{pen}} = \ell_{\mathrm{pen}}$ and is being maximised on $\boldsymbol{\theta}$. The parallel between the initial formulation of the EM algorithm and the version involving a double–maximisation thus remains valid in the case of penalised likelihood. Using the same arguments as those of Dempster et al. (1977), each penalised EM iteration increases the penalised log–likelihood, namely

$$\forall t \geqslant 0,\ \ell_{\mathrm{pen}}\left(\boldsymbol{\theta}^{[t+1]}; \boldsymbol{y}\right) \geqslant \ell_{\mathrm{pen}}\left(\boldsymbol{\theta}^{[t]}; \boldsymbol{y}\right),$$

such that for most models, Algorithm 6.3 will converge to a local maximum of ℓ_{pen}.

Algorithm 6.3: The penalised EM algorithm.

Start with an initial guess $\boldsymbol{\theta}^{[0]}$, and set $t = 0$

while *some convergence criterion not reached* **do**

  E–step: Compute $\mathcal{Q}_{\text{pen}}\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{[t]}\right) = \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\ell^{\text{c}}_{\text{pen}}\left(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}\right) \mid \boldsymbol{\theta}^{[t]}\right]$

  M–step: Set $\boldsymbol{\theta}^{[t+1]} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}_{\text{pen}}\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{[t]}\right)$

  $t \leftarrow t + 1$

end

6.4.2 Supervised Component EM

The most commonly used penalty terms involve a norm — or a convex combination of norms — of the coefficient vector. In these frameworks, the penalised complete log–likelihood writes:

$$\ell^{\text{c}}_{\text{pen}}\left(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}\right) = \ell^{\text{c}}\left(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}\right) - \frac{\lambda}{2} \times \begin{cases} \|\boldsymbol{\beta}\|_1 & \text{for LASSO,} \\ \|\boldsymbol{\beta}\|_2^2 & \text{for ridge,} \\ (1-\alpha)\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|_2^2 & \text{for elastic–net.} \end{cases} \tag{6.13}$$

In (6.13), the shrinkage parameter λ is used to adjust the intensity of the penalty, while the hyperparameter $\alpha \in [0, 1]$ determines the most appropriate trade–off geometry between the L_1– and the L_2–norm with respect to the structure of the explanatory variables. In short, these methods penalise large coefficients because they consider the high correlations among the explanatory variables as a pure nuisance, favouring effect–confusion. We propose an alternative regularisation strategy, which takes advantage of the high correlations among the explanatory variables to drive the linear predictor away from the "noise–dimensions".

With this in mind, we assume that $\boldsymbol{X}$ may contain an unknown number $K < p$ of latent structurally relevant dimensions important to model and predict $\boldsymbol{y}$. We thus propose to decompose the linear predictor through K orthogonal normalised components $\boldsymbol{X}\boldsymbol{u}_1, \boldsymbol{X}\boldsymbol{u}_2, \ldots, \boldsymbol{X}\boldsymbol{u}_K$ in such a way that

$$\boldsymbol{\eta} = \sum_{k=1}^{K} \left(\boldsymbol{X}\boldsymbol{u}_k\right) \gamma_k + \boldsymbol{U}\boldsymbol{\xi},$$

where γ_k is the regression parameter associated with component $\boldsymbol{f}_k = \boldsymbol{X}\boldsymbol{u}_k$. But for easy reading in the further development, we focus on the single com-

ponent case, considering

$$\boldsymbol{\eta} = (\boldsymbol{X}\boldsymbol{u})\,\gamma + \boldsymbol{U}\boldsymbol{\xi}.$$

Note that for identification purposes, we impose $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{u} = 1$, where $\boldsymbol{M}$ may be any symmetric definite positive matrix relevant to maximise the chosen Structural Relevance measure.

The key idea consists in basing the regularisation on the Variable–Powered Inertia (VPI) criterion, already defined in Chapter 5, and more particularly developed in Section 5.3.3. As a reminder, this criterion writes

$$\phi\,(\boldsymbol{u}) = \left[\sum_{j=1}^{p} \omega_j \left(\langle \boldsymbol{X}\boldsymbol{u} | \boldsymbol{x}_j \rangle_{\boldsymbol{P}}^2\right)^l\right]^{\frac{1}{l}} = \left[\sum_{j=1}^{p} \omega_j \left(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{x}_j\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}\boldsymbol{u}\right)^l\right]^{\frac{1}{l}}, \tag{6.14}$$

where

- $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_p\}$ is a weight system reflecting the a priori relative importance of variables (e.g. $\omega_1 = \omega_2 = \ldots = \omega_p = \frac{1}{p}$ for a uniform weighting),
- $\boldsymbol{P}$ is a weight matrix reflecting the a priori relative importance of observations (e.g. $\boldsymbol{P} = \frac{1}{n}\boldsymbol{I}_n$ for a uniform weighting),
- l is a scalar that fulfils $l \geqslant 1$.

We also recall that for VPI, the metric matrix is $\boldsymbol{M} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X})^{-1}$.

Like hyperparameter α involved in the elastic–net penalty, parameter l can be seen as a tool for refining the geometry of the regularisation, but in a different way. As a matter of fact, parameter l allows to draw component towards more (greater l) or less (smaller l) local variable–bundles. The best l value must be found through cross–validation.

Now, instead of subtracting a penalty term to the likelihood, we propose to add a bonus term — involving the SR criterion (6.14) — in order to favour the alignment of component $\boldsymbol{X}\boldsymbol{u}$ on directions we see as structural. In the single component case, with $\boldsymbol{\theta} = (\boldsymbol{u}, \gamma, \sigma_1^2, \sigma_2^2, \rho)$, we define a Supervised Component–based Expectation–Maximisation algorithm (SCEM) built on the complete regularised likelihood

$$\mathcal{L}_{\mathrm{SC}}^{\mathrm{c}}\,(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) = \left[\mathcal{L}^{\mathrm{c}}\,(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi})\right]^{1-s} \left[\phi\,(\boldsymbol{u})\right]^s,$$

where $s \in [0, 1]$ is a trade–off parameter tuning the bonus intensity. This leads to the following SC complete log–likelihood

$$\ell_{\mathrm{SC}}^{\mathrm{c}}\,(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) = (1 - s)\,\ell^{\mathrm{c}}\,(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) + s\,\log\left[\phi\,(\boldsymbol{u})\right]. \tag{6.15}$$

In order to link shrinkage parameter λ in (6.13) and trade–off parameter s in (6.15), let us underline that for $s \neq 1$,

$$\ell^{\mathrm{c}}_{\mathrm{SC}}\left(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}\right) \propto \ell^{\mathrm{c}}\left(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}\right) + \frac{s}{1-s} \log\left[\phi\left(\boldsymbol{u}\right)\right].$$

As pointed out in Chapter 5, the greater the need to regularise, the closer s has to be to 1, and the larger ratio $s/(1-s)$ — as well as λ — has to be. With $\ell^{\mathrm{c}}_{\mathrm{SC}}$ as defined in (6.15), Algorithm 6.4 summarises the key steps of our SC regularised EM.

Algorithm 6.4: The SC regularised EM.

Start with an initial guess $\boldsymbol{\theta}^{[0]}$ and set $t = 0$

while *some convergence criterion not reached* **do**

  E–step: Compute $\mathcal{Q}_{\mathrm{SC}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) = \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\ell^{\mathrm{c}}_{\mathrm{SC}}\left(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}\right) \,|\, \boldsymbol{\theta}^{[t]}\right]$

  M–step: Set $\boldsymbol{\theta}^{[t+1]} = \underset{\boldsymbol{\theta}:\boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{u}=1}{\arg\max}\ \mathcal{Q}_{\mathrm{SC}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right)$

  $t \leftarrow t + 1$

end

Proposition 6.1. The supervised component log–likelihood $\ell_{\mathrm{SC}}\left(\boldsymbol{\theta}^{[t]}; \boldsymbol{y}\right)$ increases with t.

Proof. Let us begin by rewriting the auxiliary function $\mathcal{Q}_{\mathrm{SC}}$ introduced in Algorithm 6.4 as

$$\begin{aligned}
\mathcal{Q}_{\mathrm{SC}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) &= \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[(1-s)\left[\ell(\boldsymbol{\theta}; \boldsymbol{y}) + \ell(\boldsymbol{\theta}; \boldsymbol{\xi}|\boldsymbol{y})\right] + s \log\left[\phi(\boldsymbol{u})\right] \,|\, \boldsymbol{\theta}^{[t]}\right] \\
&= (1-s)\,\ell(\boldsymbol{\theta}; \boldsymbol{y}) + s \log\left[\phi(\boldsymbol{u})\right] + (1-s)\,\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\ell(\boldsymbol{\theta}; \boldsymbol{\xi}|\boldsymbol{y}) \,|\, \boldsymbol{\theta}^{[t]}\right] \\
&= \ell_{\mathrm{SC}}(\boldsymbol{\theta}; \boldsymbol{y}) + (1-s)\, R\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right),
\end{aligned}$$

where $R\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) = \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\ell(\boldsymbol{\theta}; \boldsymbol{\xi}|\boldsymbol{y}) \,|\, \boldsymbol{\theta}^{[t]}\right]$. The SC log–likelihood then writes

$$\ell_{\mathrm{SC}}\left(\boldsymbol{\theta}; \boldsymbol{y}\right) = \mathcal{Q}_{\mathrm{SC}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) - (1-s)\, R\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right).$$

Subtracting $\ell_{\mathrm{SC}}\left(\boldsymbol{\theta}^{[t]}; \boldsymbol{y}\right)$ from $\ell_{\mathrm{SC}}\left(\boldsymbol{\theta}^{[t+1]}; \boldsymbol{y}\right)$ gives

$$\begin{aligned}
&\ell_{\mathrm{SC}}\left(\boldsymbol{\theta}^{[t+1]}; \boldsymbol{y}\right) - \ell_{\mathrm{SC}}\left(\boldsymbol{\theta}^{[t]}; \boldsymbol{y}\right) = \\
&\quad \mathcal{Q}_{\mathrm{SC}}\left(\boldsymbol{\theta}^{[t+1]} \,|\, \boldsymbol{\theta}^{[t]}\right) - \mathcal{Q}_{\mathrm{SC}}\left(\boldsymbol{\theta}^{[t]} \,|\, \boldsymbol{\theta}^{[t]}\right) - (1-s)\left[R\left(\boldsymbol{\theta}^{[t+1]} \,|\, \boldsymbol{\theta}^{[t]}\right) - R\left(\boldsymbol{\theta}^{[t]} \,|\, \boldsymbol{\theta}^{[t]}\right)\right].
\end{aligned}$$

Since $\boldsymbol{\theta}^{[t+1]}$ maximises $\mathcal{Q}_{\text{SC}}\left(\cdot \,|\, \boldsymbol{\theta}^{[t]}\right)$, we have

$$\mathcal{Q}_{\text{SC}}\left(\boldsymbol{\theta}^{[t+1]} \,|\, \boldsymbol{\theta}^{[t]}\right) - \mathcal{Q}_{\text{SC}}\left(\boldsymbol{\theta}^{[t]} \,|\, \boldsymbol{\theta}^{[t]}\right) \geqslant 0.$$

In addition, using Jensen's inequality, one can show that $R\left(\cdot \,|\, \boldsymbol{\theta}^{[t]}\right)$ reaches its maximum for $\boldsymbol{\theta}^{[t]}$: for all $\boldsymbol{\theta}$,

$$\begin{aligned} R\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) - R\left(\boldsymbol{\theta}^{[t]} \,|\, \boldsymbol{\theta}^{[t]}\right) &= \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\log\left(\frac{p\left(\boldsymbol{\xi}|\boldsymbol{y};\boldsymbol{\theta}\right)}{p\left(\boldsymbol{\xi}|\boldsymbol{y};\boldsymbol{\theta}^{[t]}\right)}\right) \,|\, \boldsymbol{\theta}^{[t]}\right] \\ &\leqslant \log\left[\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\frac{p\left(\boldsymbol{\xi}|\boldsymbol{y};\boldsymbol{\theta}\right)}{p\left(\boldsymbol{\xi}|\boldsymbol{y};\boldsymbol{\theta}^{[t]}\right)} \,|\, \boldsymbol{\theta}^{[t]}\right)\right] \\ &= \log\left[\int_{\Xi} p\left(\boldsymbol{\xi}|\boldsymbol{y};\boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{\xi}\right] \\ &= 0. \end{aligned}$$

It follows from $1 - s \geqslant 0$ that

$$\forall t \geqslant 0,\ \ell_{\text{SC}}\left(\boldsymbol{\theta}^{[t+1]}; \boldsymbol{y}\right) \geqslant \ell_{\text{SC}}\left(\boldsymbol{\theta}^{[t]}; \boldsymbol{y}\right).$$

□

6.5 L_2–penalised EM for Gaussian panel data

Using the notations of Section 6.2, and denoting $\boldsymbol{\xi_0}$ the vector of random errors, we focus on the following Gaussian model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U_1}\boldsymbol{\xi_1} + \boldsymbol{U_2}\boldsymbol{\xi_2} + \boldsymbol{\xi_0}.$$

We assume $\boldsymbol{\xi_0} \sim \mathcal{N}_n\left(\boldsymbol{0},\, \sigma_0^2 \boldsymbol{A_0}\right)$, with $\boldsymbol{A_0}$ a known matrix (by default here $\boldsymbol{A_0} = \boldsymbol{I}_n$), and σ_0^2 the unknown residual variance component. It is also assumed that $\boldsymbol{\xi_0}$, $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$ are mutually independent. In this framework, with $q_0 = n$, the complete log–likelihood is given by

$$\ell^{\text{c}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) = \text{cte} - \frac{1}{2}\left\{\sum_{j=0}^{2}\left(q_j \log\left(\sigma_j^2\right) + \frac{\boldsymbol{\xi}_j^{\text{T}} \boldsymbol{A}_j^{-1} \boldsymbol{\xi}_j}{\sigma_j^2}\right) + \log\left(|\boldsymbol{A_2}(\rho)|\right)\right\}, \quad (6.16)$$

where $|\boldsymbol{A}|$ denotes the determinant of any square matrix $\boldsymbol{A}$.

In this part, we have $\boldsymbol{\theta} = (\beta, \sigma_0^2, \sigma_1^2, \sigma_2^2, \rho)$ and the complete penalised log–likelihood writes

$$\ell^{\mathrm{c}}_{\mathrm{pen}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) = \ell^{\mathrm{c}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\xi}) - \frac{\lambda}{2} \|\beta\|_2^2 ,$$

where ℓ^{c} is given by (6.16). At step t of the EM algorithm, the associated penalised objective function $\mathcal{Q}_{\mathrm{pen}}$ is

$$\begin{aligned} \mathcal{Q}_{\mathrm{pen}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) = \mathrm{cte} - \frac{1}{2} \Bigg\{ \sum_{j=0}^{2} q_j \log\left(\sigma_j^2\right) + \sum_{j=0}^{1} \frac{\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_j^{\mathsf{T}} \boldsymbol{A}_j^{-1} \boldsymbol{\xi}_j \,|\, \boldsymbol{\theta}^{[t]}\right)}{\sigma_j^2} + \\ \frac{\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_{\mathbf{2}}^{\mathsf{T}} \boldsymbol{A}_{\mathbf{2}}^{-1}(\rho) \boldsymbol{\xi}_{\mathbf{2}} \,|\, \boldsymbol{\theta}^{[t]}\right)}{\sigma_2^2} + \log\left(|\boldsymbol{A}_{\mathbf{2}}(\rho)|\right) + \lambda \beta^{\mathsf{T}} \beta \Bigg\} . \end{aligned} \tag{6.17}$$

Regarding parameters β, σ_0^2 and σ_1^2, the first–order conditions directly lead to

$$\begin{aligned} \sigma_j^{2[t+1]} &= q_j^{-1} \, \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_j^{\mathsf{T}} \boldsymbol{A}_j^{-1} \boldsymbol{\xi}_j \,|\, \boldsymbol{\theta}^{[t]}\right), \text{ for } j \in \{0, 1\}, \\ \beta^{[t+1]} &= \left(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{A}_{\mathbf{0}}^{-1} \boldsymbol{X} + \lambda \sigma_0^{2[t+1]} \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}^{\mathsf{T}} \boldsymbol{A}_{\mathbf{0}}^{-1} \, \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{y} - \sum_{j=1}^{2} \boldsymbol{U}_j \boldsymbol{\xi}_j \,|\, \boldsymbol{\theta}^{[t]}\right) \end{aligned}$$

while $\sigma_2^{2[t+1]}$ and $\rho^{[t+1]}$ are solutions of the following system (see Appendix 6.10.1.2 for details):

$$\begin{cases} \dfrac{2(q_2 - 2)\rho\, \sigma_2^2}{\rho^2 - 1} + \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_{\mathbf{2}}^{\mathsf{T}} \dfrac{\partial \boldsymbol{A}_{\mathbf{2}}^{-1}(\rho)}{\partial \rho} \boldsymbol{\xi}_{\mathbf{2}} \,|\, \boldsymbol{\theta}^{[t]}\right) = 0 \\ \sigma_2^2 = q_2^{-1} \, \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_{\mathbf{2}}^{\mathsf{T}} \boldsymbol{A}_{\mathbf{2}}^{-1}(\rho) \boldsymbol{\xi}_{\mathbf{2}} \,|\, \boldsymbol{\theta}^{[t]}\right). \end{cases} \tag{6.18}$$

Let us denote $\boldsymbol{S}_\rho = \boldsymbol{A}_{\mathbf{2}}^{-1}(\rho)$ and $\boldsymbol{S}'_\rho = \dfrac{\partial \boldsymbol{A}_{\mathbf{2}}^{-1}(\rho)}{\partial \rho}$, which respectively explicitly write

$$\begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ 0 & -\rho & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 1+\rho^2 & -\rho \\ 0 & 0 & \cdots & -\rho & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 0 & -1 & 0 & \cdots & 0 \\ -1 & 2\rho & -1 & \cdots & 0 \\ 0 & -1 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 2\rho & -1 \\ 0 & 0 & \cdots & -1 & 0 \end{pmatrix}.$$

In addition, let us define $J^{[t]}(\rho) := \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_{\mathbf{2}}^{\mathsf{T}} \boldsymbol{K}_\rho \boldsymbol{\xi}_{\mathbf{2}} \,|\, \boldsymbol{\theta}^{[t]}\right)$, with $\boldsymbol{K}_\rho = \dfrac{2(q_2 - 2)\rho}{q_2(\rho^2 - 1)} \boldsymbol{S}_\rho + \boldsymbol{S}'_\rho$. Then (6.18) admits a more concise form:

$$\begin{cases} J^{[t]}(\rho) = 0 & \text{(6.19a)} \\ \sigma_2^2 = q_2^{-1} \, \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_{\mathbf{2}}^{\mathsf{T}} \boldsymbol{S}_\rho \boldsymbol{\xi}_{\mathbf{2}} \,|\, \boldsymbol{\theta}^{[t]}\right). & \text{(6.19b)} \end{cases}$$

Finally, denoting

$$\boldsymbol{S}''_\rho = \frac{\partial^2 \boldsymbol{A}_{\boldsymbol{2}}^{-1}(\rho)}{\partial \rho^2} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 2 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 2 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix},$$

and $\boldsymbol{K}'_\rho = \dfrac{-2(q_2-2)(1+\rho^2)}{q_2(\rho^2-1)^2}\boldsymbol{S}_\rho + \dfrac{2(q_2-2)\rho}{q_2(\rho^2-1)}\boldsymbol{S}'_\rho + \boldsymbol{S}''_\rho$, we can also define the first derivative $J'^{[t]}(\rho) := \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_{\boldsymbol{2}}^\intercal \boldsymbol{K}'_\rho \boldsymbol{\xi}_{\boldsymbol{2}} \,|\, \boldsymbol{\theta}^{[t]}\right)$. A simple Newton–Raphson method can therefore be used to find the appropriate $\rho^{[t+1]}$ that solves (6.19a). Once obtained, (6.19b) becomes:

$$\sigma_2^{2[t+1]} = q_2^{-1}\,\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_{\boldsymbol{2}}^\intercal \boldsymbol{S}_{\rho^{[t+1]}} \boldsymbol{\xi}_{\boldsymbol{2}} \,|\, \boldsymbol{\theta}^{[t]}\right).$$

As we can see, especially when updating parameter $\beta^{[t+1]}$, an appropriate value of shrinkage parameter λ is necessary. Since Algorithm 6.3 is designed for a fixed value of λ, the most common solution would be to choose this parameter as the minimiser of a 5–fold (for instance) cross–validation error. The major disadvantage of the cross–validation is the high algorithmic cost. That is why some authors — including Yi and Caramanis (2015) — are increasingly considering the possibility of iteratively updating the shrinkage parameter. As encouraged by Golub et al. (1979), the Generalised Cross–Validation (GCV) appears to be efficient for choosing a good ridge parameter. As a reminder, let us note $\widehat{\boldsymbol{y}}$ the fitted values and $\boldsymbol{H}_\lambda$ the "hat" matrix (depending on λ in the case of ridge regression) that satisfies the equality $\widehat{\boldsymbol{y}} = \boldsymbol{H}_\lambda \boldsymbol{y}$. The GCV consists in choosing the shrinkage parameter that minimises

$$\text{GCV}(\lambda) = \frac{n^{-1}\left\|\boldsymbol{y} - \boldsymbol{H}_\lambda \boldsymbol{y}\right\|^2}{\left[1 - n^{-1}\,\text{Trace}\left(\boldsymbol{H}_\lambda\right)\right]^2}.$$

Greatly inspired by Eliot et al. (2011), Algorithm 6.5 presents our ridge–penalised EM algorithm for LMM with an AR(1) random time–specific effect, improved by iteratively updating the shrinkage parameter using a GCV procedure. Note that, writing

$$\begin{cases} \boldsymbol{\Gamma}^{[t+1]} := \text{Var}^{[t+1]}(Y) = \displaystyle\sum_{j=0}^{1} \sigma_j^{2[t+1]} \boldsymbol{U_j} \boldsymbol{A_j} \boldsymbol{U_j}^\intercal + \sigma_2^{2[t+1]} \boldsymbol{U_2} \boldsymbol{A_2}\left(\rho^{[t+1]}\right) \boldsymbol{U_2}^\intercal \\ \boldsymbol{D}^{[t+1]} := \text{Var}^{[t+1]}(\boldsymbol{\xi}) = \textbf{bDiag}\left(\left[\sigma_1^{2[t+1]} \boldsymbol{A_1}\right] ; \left[\sigma_2^{2[t+1]} \boldsymbol{A_2}\left(\rho^{[t+1]}\right)\right]\right), \end{cases}$$

where $\boldsymbol{U_0} = \boldsymbol{I}_n$, the "hat" matrix involved in the calibration of $\lambda^{[t+1]}$ explicitly writes

$$\begin{aligned}\boldsymbol{H}_{\lambda}^{[t+1]} &= \boldsymbol{X}\left(\boldsymbol{X}^{\intercal}{\boldsymbol{\Gamma}^{[t+1]}}^{-1}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^{\intercal}{\boldsymbol{\Gamma}^{[t+1]}}^{-1} \\ &+ \boldsymbol{U}\boldsymbol{D}^{[t+1]}\boldsymbol{U}^{\intercal}{\boldsymbol{\Gamma}^{[t+1]}}^{-1}\left[\boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}^{\intercal}{\Gamma^{[t+1]}}^{-1}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^{\intercal}{\boldsymbol{\Gamma}^{[t+1]}}^{-1}\right].\end{aligned}$$

Algorithm 6.5: The adaptive ridge–penalised EM algorithm for LMM with an AR(1) random time–specific effect.

Start with an initial guess $\boldsymbol{\theta}^{[0]} = \left(\boldsymbol{\beta}^{[0]}, \sigma_0^{2\,[0]}, \sigma_1^{2\,[0]}, \sigma_2^{2\,[0]}, \rho^{[0]}\right)$ and set $t = 0$

while *some convergence criterion not reached* **do**

E–step: Compute

$\varphi_j^{[t]} = \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_j^{\intercal}\boldsymbol{A}_j^{-1}\boldsymbol{\xi}_j \,|\, \boldsymbol{\theta}^{[t]}\right),\ j \in \{0,1\}$

$\boldsymbol{\varphi}_{\boldsymbol{\beta}}^{[t]} = \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{y} - \sum_{j=1}^{2}\boldsymbol{U}_j\boldsymbol{\xi}_j \,|\, \boldsymbol{\theta}^{[t]}\right)$

M–step: Set

$\sigma_j^{2\,[t+1]} = q_j^{-1}\,\varphi_j^{[t]},\ j \in \{0,1\}$

$\widetilde{\rho} \leftarrow \rho^{[t]}$

while $|J(\rho)|$ *large* **do**

$\widetilde{\rho} \leftarrow \widetilde{\rho} - \dfrac{J^{[t]}(\widetilde{\rho})}{J'^{[t]}(\widetilde{\rho})}$

end

Set $\rho^{[t+1]} = \widetilde{\rho}$

$\sigma_2^{2\,[t+1]} = q_2^{-1}\,\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_{\boldsymbol{2}}^{\intercal}\boldsymbol{S}_{\rho^{[t+1]}}\boldsymbol{\xi}_{\boldsymbol{2}} \,|\, \boldsymbol{\theta}^{[t]}\right)$

Update $\boldsymbol{H}_{\lambda}^{[t+1]}$ and set

$$\lambda^{[t+1]} = \arg\min_{\lambda}\left\{\text{GCV}(\lambda) = \frac{n^{-1}\left\|\boldsymbol{y} - \boldsymbol{H}_{\lambda}^{[t+1]}\boldsymbol{y}\right\|^2}{\left[1 - n^{-1}\text{Trace}\left(\boldsymbol{H}_{\lambda}^{[t+1]}\right)\right]^2}\right\}$$

$\boldsymbol{\beta}^{[t+1]} = \left(\boldsymbol{X}^{\intercal}\boldsymbol{A}_{\boldsymbol{0}}^{-1}\boldsymbol{X} + \lambda^{[t+1]}\sigma_0^{2\,[t+1]}\boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^{\intercal}\boldsymbol{A}_{\boldsymbol{0}}^{-1}\,\boldsymbol{\varphi}_{\boldsymbol{\beta}}^{[t]}$

$t \leftarrow t + 1$

end

6.6 Supervised component EM for Gaussian panel data

Especially in the case of a large number of explanatory variables with an intrinsic high level of redundancies and collinearities, the decomposition of the linear predictor on a small number of relevant dimensions can greatly facilitate model interpretation. First focusing on the single component model, the goal now is to propose an alternative to Algorithm 6.5. For that, let us consider the model

$$\boldsymbol{y} = (\boldsymbol{X}\boldsymbol{u})\,\gamma + \boldsymbol{U_1}\boldsymbol{\xi_1} + \boldsymbol{U_2}\boldsymbol{\xi_2} + \boldsymbol{\xi_0}.$$

Given the fact that the construction of component $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$ must be guided by both the structure of the explanatory space and the prediction quality of the response, the regularised EM we suggest is based on the complete regularised log–likelihood defined in (6.15). Hence, updates for variance components $\sigma_j^2, j \in \{0,1,2\}$, and for parameter ρ are the same as in Algorithm 6.5. Nevertheless, owing to the product $\boldsymbol{u}\gamma$, the maximisation step relative to these parameters at step t is divided into two parts. First, given parameter $\gamma^{[t]}$, we start by updating loading–vector $\boldsymbol{u}^{[t+1]}$. Then, once $\boldsymbol{u}^{[t+1]}$ obtained, we update $\gamma^{[t+1]}$. More specifically, the update for σ_0^2 being carried out, let us consider function $\widetilde{\mathcal{Q}}_{\text{SC}}^1$ defined by

$$\begin{aligned}\widetilde{\mathcal{Q}}_{\text{SC}}^1\left(\boldsymbol{u},\gamma \,|\, \boldsymbol{\theta}^{[t]}\right) &= s\,\log\left[\phi\left(\boldsymbol{u}\right)\right] \;- \\ &(1-s)\frac{\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\|\boldsymbol{y}-(\boldsymbol{X}\boldsymbol{u})\,\gamma-\boldsymbol{U}\boldsymbol{\xi}\|^2_{\boldsymbol{A}_0^{-1}} \,|\, \boldsymbol{\theta}^{[t]}\right)}{2\sigma_0^{2\,[t+1]}}.\end{aligned} \tag{6.20}$$

The new values of loading–vector $\boldsymbol{u}^{[t+1]}$ and parameter $\gamma^{[t+1]}$ are thus obtained by setting

$$\begin{aligned}\boldsymbol{u}^{[t+1]} &= \underset{\boldsymbol{u}:\boldsymbol{u}^{\mathsf{T}}\boldsymbol{M}^{-1}\boldsymbol{u}=1}{\arg\max}\; \widetilde{\mathcal{Q}}_{\text{SC}}^1\left(\boldsymbol{u},\gamma^{[t]} \,|\, \boldsymbol{\theta}^{[t]}\right) \\ \gamma^{[t+1]} &= \arg\max_{\gamma}\; \widetilde{\mathcal{Q}}_{\text{SC}}^1\left(\boldsymbol{u}^{[t+1]},\gamma \,|\, \boldsymbol{\theta}^{[t]}\right).\end{aligned} \tag{6.21}$$

More generally, if we consider K orthogonal components, the model then writes

$$\boldsymbol{y} = \sum_{k=1}^{K} (\boldsymbol{X}\boldsymbol{u_k})\,\gamma_k + \boldsymbol{U_1}\boldsymbol{\xi_1} + \boldsymbol{U_2}\boldsymbol{\xi_2} + \boldsymbol{\xi_0}.$$

Let us briefly explain how to adapt the definition of the objective function given by (6.20) and the previous alternated maximisation (6.21). Suppose the first $k-1$ components are built and concatenated into matrix $\boldsymbol{F_{k-1}}$. An extra

component $\boldsymbol{f_k} = \boldsymbol{X}\boldsymbol{u_k}$ has to be such that $\boldsymbol{u_k}^\top \boldsymbol{M}^{-1}\boldsymbol{u_k} = 1$, along with the extra orthogonality constraint to $\boldsymbol{F_{k-1}}$, i.e.

$$(\boldsymbol{X}\boldsymbol{u_k})^\top \boldsymbol{P}\boldsymbol{F_{k-1}} = \boldsymbol{0}.$$

At step t of the regularised EM, the objective function associated with rank $k \geqslant 1$ component is then defined by

$$\begin{aligned}\widetilde{\mathcal{Q}}^k_{\text{SC}}\left(\boldsymbol{u_k}, \gamma_k \,|\, \boldsymbol{\theta}^{[t]}\right) &= s\,\log\left[\phi\left(\boldsymbol{u_k}\right)\right] - (1-s) \\ &\times \frac{\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\left\|\boldsymbol{y} - \sum_{h=0}^{k-1}\left(\boldsymbol{X}\boldsymbol{u_h}^{[t+1]}\right)\gamma_h^{[t+1]} - \left(\boldsymbol{X}\boldsymbol{u_k}\right)\gamma_k - \boldsymbol{U}\boldsymbol{\xi}\right\|^2_{\boldsymbol{A_0}^{-1}} \,|\, \boldsymbol{\theta}^{[t]}\right)}{2\sigma_0^{2\,[t+1]}},\end{aligned} \tag{6.22}$$

where $\boldsymbol{u_0}^{[t+1]}$ and $\gamma_0^{[t+1]}$ are considered null by convention, thus recovering function $\widetilde{\mathcal{Q}}^1_{\text{SC}}$ defined in (6.20) for $k = 1$. In view of the aforementioned considerations, the new values of loading–vector $\boldsymbol{u_k}^{[t+1]}$ and associated parameter $\gamma_k^{[t+1]}$ are obtained by setting

$$\begin{aligned}\boldsymbol{u_k}^{[t+1]} &= \underset{\boldsymbol{u_k} \in \mathcal{S}_k^{[t+1]}}{\arg\max}\ \widetilde{\mathcal{Q}}^k_{\text{SC}}\left(\boldsymbol{u_k}, \gamma_k^{[t]} \,|\, \boldsymbol{\theta}^{[t]}\right) \\ \gamma_k^{[t+1]} &= \underset{\gamma_k}{\arg\max}\ \widetilde{\mathcal{Q}}^k_{\text{SC}}\left(\boldsymbol{u_k}^{[t+1]}, \gamma_k \,|\, \boldsymbol{\theta}^{[t]}\right)\end{aligned} \tag{6.23}$$

where $\mathcal{S}_k^{[t+1]} = \left\{\boldsymbol{u} \in \mathbb{R}^p \mid \boldsymbol{u}^\top \boldsymbol{M}^{-1}\boldsymbol{u} = 1 \text{ and } (\boldsymbol{X}\boldsymbol{u})^\top \boldsymbol{P}\boldsymbol{F}_{\boldsymbol{k-1}}^{[t+1]} = \boldsymbol{0}\right\}$. As recently detailed in Bry et al. (2018) and recalled in Appendix 5.8.3 (Chapter 5), the PING algorithm can be used to maximise, at least locally, any criterion on such a set as $\mathcal{S}_k$. At step t of the SC regularised EM, it allows us to build component of rank $k > 1$ by updating loading–vector $\boldsymbol{u_k}^{[t+1]}$ as in (6.23), and also first rank component by imposing $\boldsymbol{F_0}^{[t+1]} = \boldsymbol{0}$. Appendix 6.10.2.1 details how the update of parameter γ_k is obtained. Algorithm 6.6 summarises the SC regularised EM we propose, designed for LMMs with an AR(1) random time–specific effect.

6.7 Extension to the non–Gaussian case

We shall now consider the general case in which the conditional distribution of the data given the random effects is assumed to belong to the exponential family. Unlike the Gaussian case, due to the intractability of the likelihood, the direct application of the strategies presented in Sections 6.5 and 6.6 is impossible. The numerical and stochastic approximations used to handle this intractability being still computationally intensive, we preferably develop a quicker estimation technique, based on a linear approximation of the model itself.

Algorithm 6.6: The supervised component–based regularised EM algorithm for LMMs with an AR(1) random time–specific effect.

Start with an initial guess
$\boldsymbol{\theta}^{[0]} = \left(\boldsymbol{u}_{\mathbf{1}}^{[0]}, \ldots, \boldsymbol{u}_{\boldsymbol{K}}^{[0]}, \gamma_1^{[0]}, \ldots, \gamma_K^{[0]}, \sigma_0^{2\,[0]}, \sigma_1^{2\,[0]}, \sigma_2^{2\,[0]}, \rho^{[0]}\right)$ and set $t = 0$
while *some convergence criterion not reached* **do**
  Set $\sigma_0^{2\,[t+1]}, \sigma_1^{2\,[t+1]}, \sigma_2^{2\,[t+1]}, \rho^{[t+1]}$ as in Algorithm 6.5
  Compute $\boldsymbol{\varphi}_{\boldsymbol{\gamma}}^{[t]} = \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{y} - \sum_{j=1}^{2} \boldsymbol{U}_j\boldsymbol{\xi}_j \,|\, \boldsymbol{\theta}^{[t]}\right)$
  for k *from* 1 *to* K **do**
    SC E–step:
    Compute $\widetilde{\mathcal{Q}}_{\text{SC}}^{k}\left(\boldsymbol{u}_{\boldsymbol{k}}, \gamma_k \,|\, \boldsymbol{\theta}^{[t]}\right)$ as defined by (6.22)
    SC M–step: Set
    $\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]} = \underset{\boldsymbol{u}_{\boldsymbol{k}} \in \mathcal{S}_k^{[t+1]}}{\arg\max}\ \widetilde{\mathcal{Q}}_{\text{SC}}\left(\boldsymbol{u}_{\boldsymbol{k}}, \gamma_k^{[t]} \,|\, \boldsymbol{\theta}^{[t]}\right)$
    $\gamma_k^{[t+1]} = \left[\left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]}\right)^{\mathrm{T}} \boldsymbol{A}_{\mathbf{0}}^{-1}\left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]}\right)\right]^{-1}\left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]}\right)^{\mathrm{T}} \boldsymbol{A}_{\mathbf{0}}^{-1}$
    $\times \left[\boldsymbol{\varphi}_{\boldsymbol{\gamma}}^{[t]} - \sum_{h=0}^{k-1}\left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{h}}^{[t+1]}\right)\gamma_h^{[t+1]}\right]$
  end
  $t \leftarrow t + 1$
end

Owing to the GLMM dependence structure, the Fisher scoring algorithm — that performs maximum likelihood estimation in GLMs — was adapted by Schall (1991). We, in turn, adapt Schall's algorithm, still considering a Gaussian approximation of the linearised model. Our proposal is to replace the usual estimation step with a regularised EM. This makes it possible to take into account the high level of correlation in $\boldsymbol{X}$ and the particular random effects distributions. The method can be decomposed into a linearisation step and an estimation step:

Linearisation step. For each $i \in \{1, \ldots, n\}$, a classic order–1 linearisation of y_i around μ_i is given by: $g(y_i) \simeq z_i = g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$. Matricially, this approximation provides a working variable $\boldsymbol{z}$ entering the following linearised model

$$\mathcal{M}: \quad \boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\xi} + \boldsymbol{e},$$

with $\text{Var}\,(\boldsymbol{e} \,|\, \boldsymbol{\xi}) = \mathbf{Diag}\left(\left[g'(\mu_i)\right]^2 \text{Var}\,(Y_i \,|\, \boldsymbol{\xi})\right)_{i=1,\ldots,n} = \boldsymbol{R}$.

Estimation step. Schall's method involves interpreting model $\mathcal{M}$ as an LMM, and then solving Henderson's system to get current parameter estimates. Al-

ternatively, we rather propose a regularised EM step based on linearised model $\mathcal{M}$, the latter still being interpreted as an LMM.

At step t, the conditional expectation we consider is therefore $\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{z}^{[t]}}$ instead of $\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}$. In the same vein, residual variance is $\boldsymbol{R}^{[t]}$ instead of $\sigma_0^{2[t]}\boldsymbol{A_0}$. Considering this, in order to avoid ambiguity, two main refinements have to be detailed. The first one concerns the update of shrinkage parameter λ that occurs in the ridge–penalised version of the EM extended to GLMM. Contrary to the homoskedastic LMM considered in Eliot et al. (2011), $\mathcal{M}$ contains heteroskedastic errors. We will then opt for the modified GCV criterion suggested by Andrews (1991) and we set

$$\lambda^{[t+1]} = \arg\min_{\lambda} \left\{ \text{GCV}(\lambda) = \frac{n^{-1} \left\| \boldsymbol{z}^{[t]} - \boldsymbol{H}_{\lambda}^{[t+1]} \boldsymbol{z}^{[t]} \right\|^2_{\boldsymbol{R}^{[t]-1}}}{\left[1 - n^{-1} \text{Trace}\left(\boldsymbol{H}_{\lambda}^{[t+1]}\right)\right]^2} \right\}. \tag{6.24}$$

The second refinement is about the objective function specific to the construction of the loading–vector in the SC–regularised version. In the GLMM framework, the previously defined function $\widetilde{\mathcal{Q}}_{\text{SC}}^{k}$ given by (6.22) becomes

$$\begin{aligned} \widetilde{\mathcal{Q}}_{\text{SC}}^{k}\left(\boldsymbol{u_k}, \gamma_k \,|\, \boldsymbol{\theta}^{[t]}\right) &= s \, \log\left[\phi\left(\boldsymbol{u_k}\right)\right] - \frac{1-s}{2} \\ &\times \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{z}^{[t]}} \left(\left\| \boldsymbol{z}^{[t]} - \sum_{h=0}^{k-1} \left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{h}}^{[t+1]}\right) \gamma_h^{[t+1]} - \left(\boldsymbol{X}\boldsymbol{u_k}\right)\gamma_k - \boldsymbol{U}\boldsymbol{\xi} \right\|^2_{\boldsymbol{R}^{[t]-1}} \,|\, \boldsymbol{\theta}^{[t]} \right), \end{aligned} \tag{6.25}$$

still imposing both $\boldsymbol{u_0}^{[t+1]}$ and $\gamma_0^{[t+1]}$ to be null.

Algorithms 6.7 recapitulates the generic iteration of both ridge–penalised and SC–regularised EMs for GLMM with an AR(1) random time–specific effect. Steps $(\mathbf{1}) - (\mathbf{3})$ have to be repeated until stability of parameters $\boldsymbol{\theta}$ is reached.

Algorithm 6.7: The ridge–penalised and SC–regularised EMs extended to GLMM with an AR(1) random time–specific effect (generic iteration).

(1) Linearisation step: Set

$$\mathcal{M}^{[t]}: \left| \begin{array}{l} \boldsymbol{z}^{[t]} = \begin{cases} \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\xi} + \boldsymbol{e}^{[t]} & \textbf{for ridge penalisation} \\ \sum_{k=1}^{K} (\boldsymbol{X}\boldsymbol{u_k})\,\gamma_k + \boldsymbol{U}\boldsymbol{\xi} + \boldsymbol{e}^{[t]} & \textbf{for SC regularisation} \end{cases} \\ \text{with: } \mathrm{Var}\left(\boldsymbol{e}^{[t]} \,|\, \boldsymbol{\xi}\right) = \boldsymbol{R}^{[t]} \end{array} \right.$$

(2) Estimation step:

For ridge penalisation: $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_1^2, \sigma_2^2, \rho)$

Compute $\mathcal{Q}_{\text{pen}}\left(\boldsymbol{\theta}, \lambda \,|\, \boldsymbol{\theta}^{[t]}\right) = \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{z}^{[t]}}\left[\ell^{\text{c}}\left(\boldsymbol{\theta}; \boldsymbol{z}^{[t]}, \boldsymbol{\xi}\right) - \frac{\lambda}{2}\boldsymbol{\beta}^{\text{T}}\boldsymbol{\beta} \,|\, \boldsymbol{\theta}^{[t]}\right]$

Set $\left(\sigma_1^{2[t+1]}, \sigma_2^{2[t+1]}, \rho^{[t+1]}\right) = \underset{\sigma_1^2, \sigma_2^2, \rho}{\arg\max}\ \mathcal{Q}_{\text{pen}}\left(\boldsymbol{\theta}, \lambda \,|\, \boldsymbol{\theta}^{[t]}\right)$

Update $\lambda^{[t+1]}$ using formula (6.24)

Set $\boldsymbol{\beta}^{[t+1]} = \arg\ \underset{\boldsymbol{\beta}}{\max}\ \mathcal{Q}_{\text{pen}}\left(\boldsymbol{\theta}, \lambda^{[t+1]} \,|\, \boldsymbol{\theta}^{[t]}\right)$

For SC regularisation: $\boldsymbol{\theta} = (\boldsymbol{u_1}, \ldots, \boldsymbol{u_K}, \gamma_1, \ldots, \gamma_K, \sigma_1^2, \sigma_2^2, \rho)$

Compute

$\mathcal{Q}_{\text{reg}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) = \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{z}^{[t]}}\left[s\,\ell^{\text{c}}\left(\boldsymbol{\theta}; \boldsymbol{z}^{[t]}, \boldsymbol{\xi}\right) + (1-s)\log\left[\phi\left(\boldsymbol{u}\right)\right] \,|\, \boldsymbol{\theta}^{[t]}\right]$

Set $\left(\sigma_1^{2[t+1]}, \sigma_2^{2[t+1]}, \rho^{[t+1]}\right) = \underset{\sigma_1^2, \sigma_2^2, \rho}{\arg\max}\ \mathcal{Q}_{\text{reg}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right)$

$\forall k \in \{1, \ldots, K\}$,

Compute $\widetilde{\mathcal{Q}}_{\text{reg}}^{k}\left(\boldsymbol{u_k}, \gamma_k \,|\, \boldsymbol{\theta}^{[t]}\right)$ as defined by (6.25)

Set $\boldsymbol{u_k}^{[t+1]} = \underset{\boldsymbol{u_k} \in \mathcal{S}_k^{[t+1]}}{\arg\max}\ \widetilde{\mathcal{Q}}_{\text{reg}}^{k}\left(\boldsymbol{u_k}, \gamma_k^{[t]} \,|\, \boldsymbol{\theta}^{[t]}\right)$

Set $\gamma_k^{[t+1]} = \arg\ \underset{\gamma_k}{\max}\ \widetilde{\mathcal{Q}}_{\text{reg}}^{k}\left(\boldsymbol{u_k}^{[t+1]}, \gamma_k \,|\, \boldsymbol{\theta}^{[t]}\right)$

(3) Updating step:

Set $\boldsymbol{\xi}^{[t+1]} = \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{z}^{[t]}}\left(\boldsymbol{\xi} \,|\, \boldsymbol{\theta}^{[t+1]}\right)$

Update working variables $\boldsymbol{z}^{[t+1]}$ and variance matrix $\boldsymbol{R}^{[t+1]}$

6.8 Comparative results on simulated Poisson data

6.8.1 First design

The goal of this section is to characterise the relative performances of ridge– and SC–regularisations in the framework of a log–link Poisson regression, including both individual and autoregressive time–specific random effects. We thus simulate response $\boldsymbol{y}$ as

$$\boldsymbol{y} \sim \mathcal{P}\Big(\exp\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U_1}\boldsymbol{\xi_1} + \boldsymbol{U_2}\boldsymbol{\xi_2}\right) \Big). \tag{6.26}$$

Explanatory variables $\boldsymbol{X}$ consist of three independent bundles of standardised variables $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ and $\boldsymbol{X_3}$, so that $\boldsymbol{X} = [\boldsymbol{X_1}|\boldsymbol{X_2}|\boldsymbol{X_3}]$. The level of redundancy within each bundle is tuned through parameter $\tau \in [0, 1]$ such that correlations among explanatory variables within bundle $\boldsymbol{X_j}$ are

$$\text{cor}\left(\boldsymbol{X_j}\right) = \tau \mathbf{1}\mathbf{1}^{\mathrm{T}} + (1 - \tau)\boldsymbol{I}.$$

Bundle $\boldsymbol{X_1}$ contains 10 noise variables that play no explanatory role, while $\boldsymbol{X_2}$ and $\boldsymbol{X_3}$ are two bundles of 5 variables each, which homogeneously contribute to model $\boldsymbol{y}$. Hence our choice for the fixed effects parameters is

$$\boldsymbol{\beta} = (\underbrace{0, \ldots\ldots, 0}_{10 \text{ times}}, \underbrace{b, \ldots\ldots, b}_{10 \text{ times}})^{\mathrm{T}},$$

with $b = 1$ in practice. In all the simulations, we set $\sigma_1 = \sigma_2 = 0.3$, to prevent the random part of the linear predictor from being too huge.

This study aims at answering 4 questions:

1. Is the convergence assured ? (Figure 6.1)
2. Is there any sensitivity to the value of ρ ? (Figure 6.2)
3. How good are the estimations ? (Figure 6.3)
4. Does the use of SC–regularisation facilitate the interpretation of the model ? (Figure 6.4)

One of the objectives of the simulation study is to verify that the components built by SCEM align with the bundles of predictive variables, despite the presence of a large nuisance bundle.

Is the convergence assured ? 20 simulations are conducted according to model (6.26) for $q_1 = 20$ individuals, $q_2 = 20$ time–points, $\rho = 0.5$ and $\tau = 0.8$. In order to check the convergence of ridge– and SC–regularised EMs, we consider the L_2–convergence criterion between two iterations. Figure 6.1 presents the 20 trajectories of the criterion $\left\|\boldsymbol{\beta}^{[t+1]} - \boldsymbol{\beta}^{[t]}\right\|_2, t \in \{2, \ldots, 500\}$, for both ridge and SC regularisations. The SCEM requires about a hundred iterations to achieve convergence but compared to ridge, the estimations provided are much more stable.

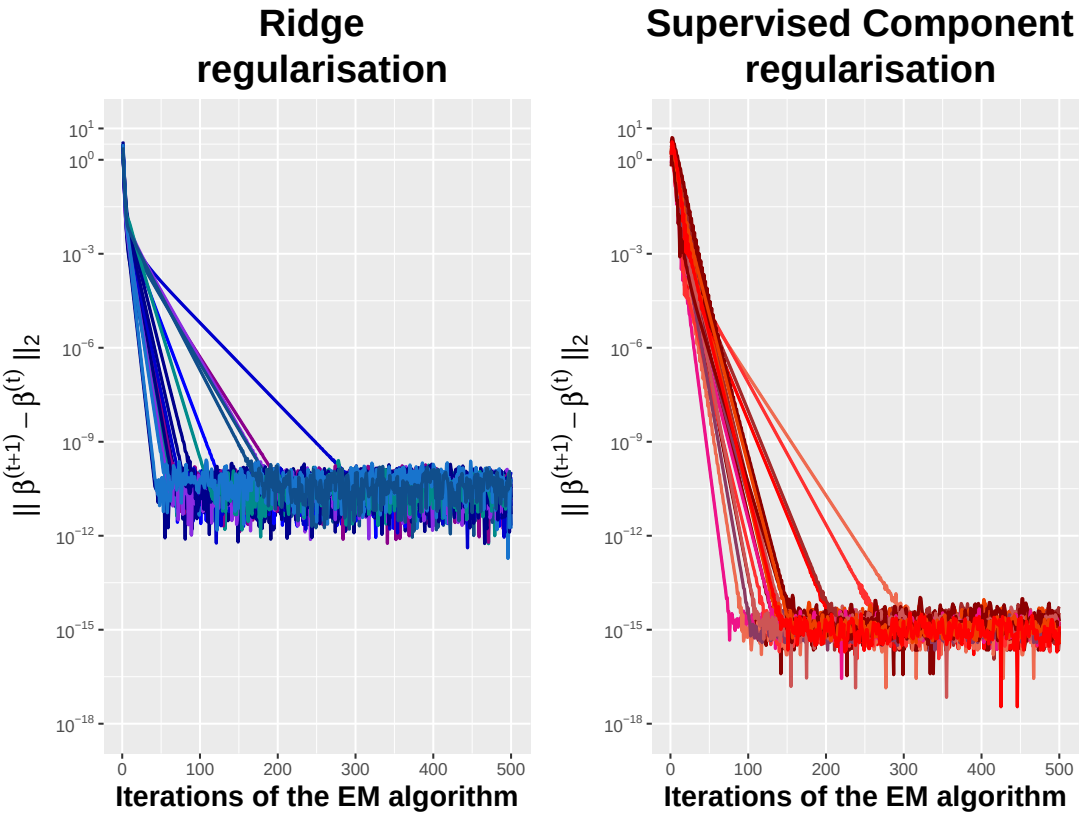


Figure 6.1 – *First design – Convergence assessment.* *20 trajectories of* $\left\|\boldsymbol{\beta}^{[t+1]} - \boldsymbol{\beta}^{[t]}\right\|_2$ *are graphed, for* $t \in \{2, \ldots, 500\}$ *and for both ridge and Supervised Component regularisations.*

Is there any sensitivity to the value of ρ ? We still consider $q_1 = q_2 = 20$, and $\tau = 0.8$. For each simulated value of $\rho \in \{-0.9, -0.8, \ldots, 0.8, 0.9\}$, 120 samples are generated, and Figure 6.2 shows the boxplots of the estimated values $\widehat{\rho}$ obtained with the SC regularisation according to their real values. The same behaviour is observed with the ridge–based regularised EM. They seem close to the first bissector, showing that the estimates are good whatever the real value of the parameter. Each estimate was obtained by choosing $\rho^{[0]} = 0$ as the starting value of the EM algorithm, For this reason, there is a slight downward bias for simulated autocorrelations close to 1 and a slight upward bias for simulated autocorrelations close to -1.

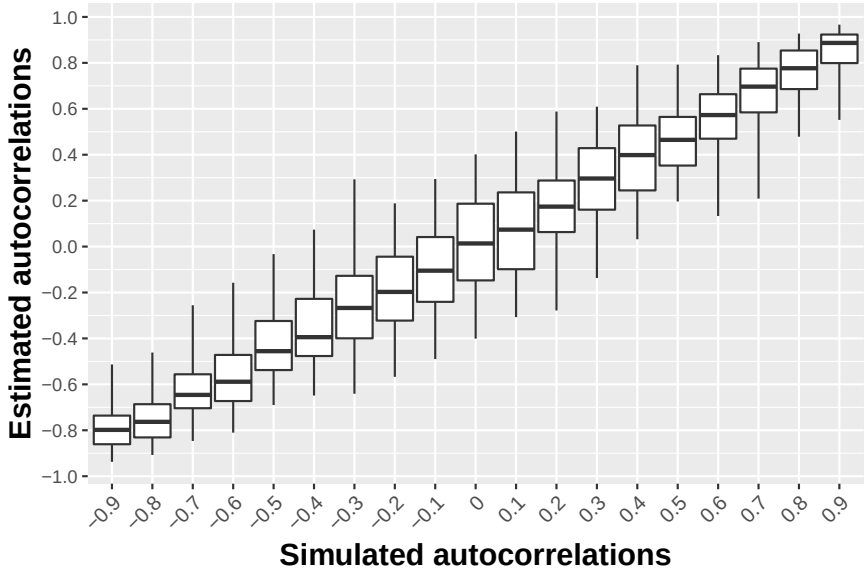


Figure 6.2 – *First design – Sensitivity to the value of ρ.* *The graph shows the boxplots of estimated autocorrelations $\widehat{\rho}$ obtained with SCEM according to real values.*

How good are the estimations ? The number of individuals is set to $q_1 = 10$ while the number of time–points q_2 varies from 10 to 100. For each value of q_2, we generate 50 samples according to (6.26), with $\rho = 0.5$ and $\tau = 0.8$. Figure 6.3 gives the RMSEs relative to fixed effects parameter β and autocorrelation parameter ρ. There are no significant differences between ridge and SC concerning parameter ρ, the same behaviour being observed for parameters σ_1^2 and σ_2^2. By contrast, compared to ridge, the SC–based regularisation performs a better estimation of the fixed effects β, very likely due to the importance given to the strong inner structures of $\boldsymbol{X}$.

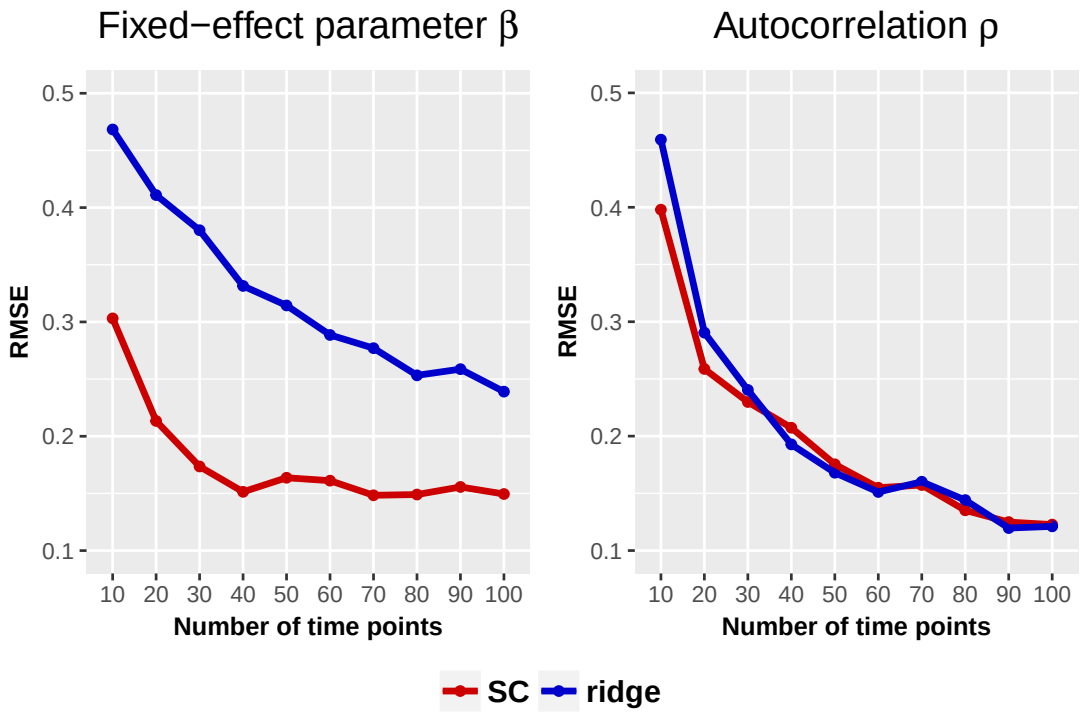


Figure 6.3 – *First design – Estimate accuracies.* *The RMSEs of fixed–effect parameter and autocorrelation parameter estimates are represented, for both ridge and Supervised Component regularisations. They are obtained over 50 samples for each number of time–points* $q_2 \in \{10, 20, \ldots, 100\}$.

Does the use of SC–regularisation facilitate the interpretation of the model ? Figure 6.4 shows an example of the first component planes output by SCEM for $\tau = 0.5$. Considering such level of redundancy, our cross–validation selects only two components (associated in most cases with a bundles–locality parameter close to 3 and a trade–off parameter close to 1), which are the ones that explain the response. Regardless of the cross–validation results, we edit component plane $(1, 3)$. Interestingly, although bundle of noise is the one with maximum inertia, it appears only along the third component. The algorithm thus detected its structural relevance, but also the fact that it did not play any explanatory role. It is essential to note that the components obtained by the SC method are much better constructed than those of classical methods such as PCA or PLS. Indeed, with such a design, the first principal component focuses on the noise bundle, and the first PLS component combines the two predictive bundles into a single one, which greatly deteriorates their predictive and interpretative power.

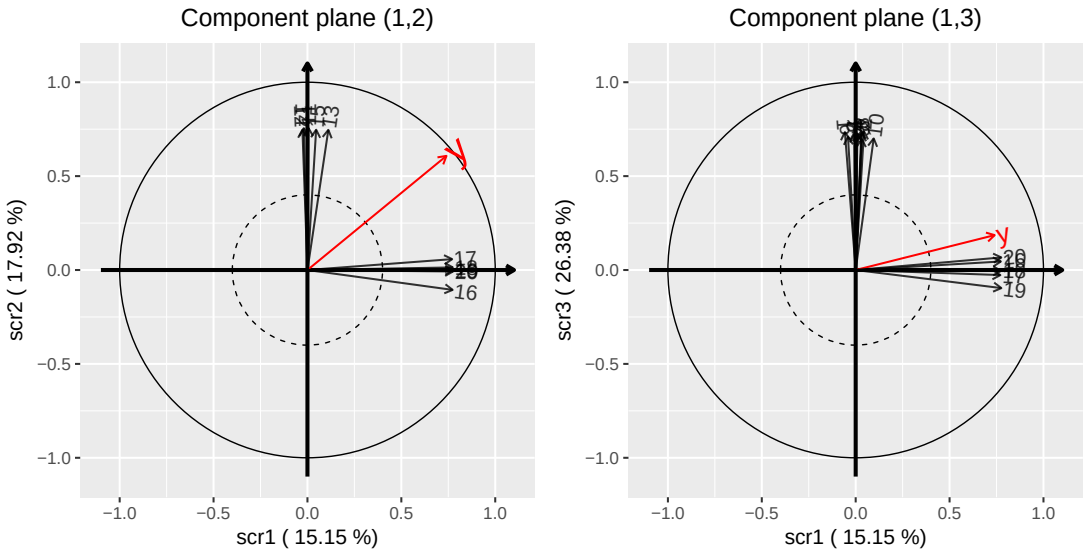


Figure 6.4 – *First design – Model interpretation.* *Component planes* $(1, 2)$ *and* $(1, 3)$ *obtained using the SC regularisation are given here. The black arrows represent the explanatory variables while the red one represent the projection of the* $\boldsymbol{X}$*–part of the linear predictor. The percentage of inertia captured by each component is given in parentheses.*

6.8.2 Second design

Given the individual random effect, $\boldsymbol{\xi_1}$, and the autoregressive time–specific random effect, $\boldsymbol{\xi_2}$, we still simulate response $\boldsymbol{y}$ as

$$\boldsymbol{y} \sim \mathcal{P}\Big(\exp\left(\boldsymbol{X\beta} + \boldsymbol{U_1\xi_1} + \boldsymbol{U_2\xi_2}\right) \Big). \tag{6.27}$$

The main change is in the structure of explanatory variables. We now consider that $\boldsymbol{X} = \big[\, \boldsymbol{X_1} \mid \boldsymbol{X_2} \mid \boldsymbol{X_3} \mid \boldsymbol{X_4} \,\big]$, where

- $\boldsymbol{X_1} = \big[\, \boldsymbol{x_1} \mid \ldots \mid \boldsymbol{x_{10}} \,\big]$ is a nuisance bundle of 10 correlated variables that play no explanatory role,
- $\boldsymbol{X_2} = \big[\, \boldsymbol{x_{11}} \mid \ldots \mid \boldsymbol{x_{14}} \,\big]$ contains 4 uncorrelated noise variables (without explanatory role either),
- $\boldsymbol{X_3} = \big[\, \boldsymbol{x_{15}} \mid \ldots \mid \boldsymbol{x_{19}} \,\big]$ is a bundle of 5 correlated variables that partially contribute to model $\boldsymbol{y}$,
- $\boldsymbol{X_4}$ contains a single predictive variable, $\boldsymbol{x_{20}}$.

The fixed–effect parameter is therefore set to

$$\boldsymbol{\beta} = \Big(\underbrace{b_1, \ldots\ldots\ldots\ldots, b_1}_{10 \text{ times}}, \underbrace{b_2, \ldots\ldots, b_2}_{4 \text{ times}}, \underbrace{b_3, \ldots\ldots, b_3}_{5 \text{ times}}, b_4 \Big)^{\mathsf{T}},$$

where $b_1 = b_2 = 0$, $b_3 = 0.2$, $b_4 = 1$ in practice. As explained in Section 6.8.1, parameter τ tunes the level of redundancy within bundles $\boldsymbol{X_1}$ and $\boldsymbol{X_3}$. Finally, we set $\sigma_1^2 = \sigma_2^2 = 0.5$.

Convergence assessment. 20 simulations are conducted according to model (6.27) for $q_1 = 10$ individuals, $q_2 = 20$ time–points, $\rho = 0.5$ and $\tau = 0.6$. Figures 6.5, 6.6 and 6.7 show the 20 trajectories of criteria $\left\|\sigma_1^{2[t+1]} - \sigma_1^{2[t]}\right\|_2$, $\left\|\sigma_2^{2[t+1]} - \sigma_2^{2[t]}\right\|_2$, and $\left\|\rho^{[t+1]} - \rho^{[t]}\right\|_2$ respectively, for both ridge and SC regularisations. The trajectories for the fixed–effect parameter are similar to those presented in Figure 6.1. The ridge procedure seems slightly faster than the Supervised Component regularisation: ridge requires about 75 iterations to achieve convergence while the SCEM requires about a hundred. Interestingly, compared to ridge, the estimations provided by SCEM are more stable, including for parameters σ_1^2, σ_2^2 and ρ. This can be explained by the fact that the shrinkage parameter of the ridge regression is autocalibrated at each iteration, while the SCEM tuning parameters, (K, s, l), are previously calibrated through a cross–validation. For the fixed–effect parameter, this stability is enhanced by the ability of SCEM to focus on the most predictive structures within $\boldsymbol{X}$.

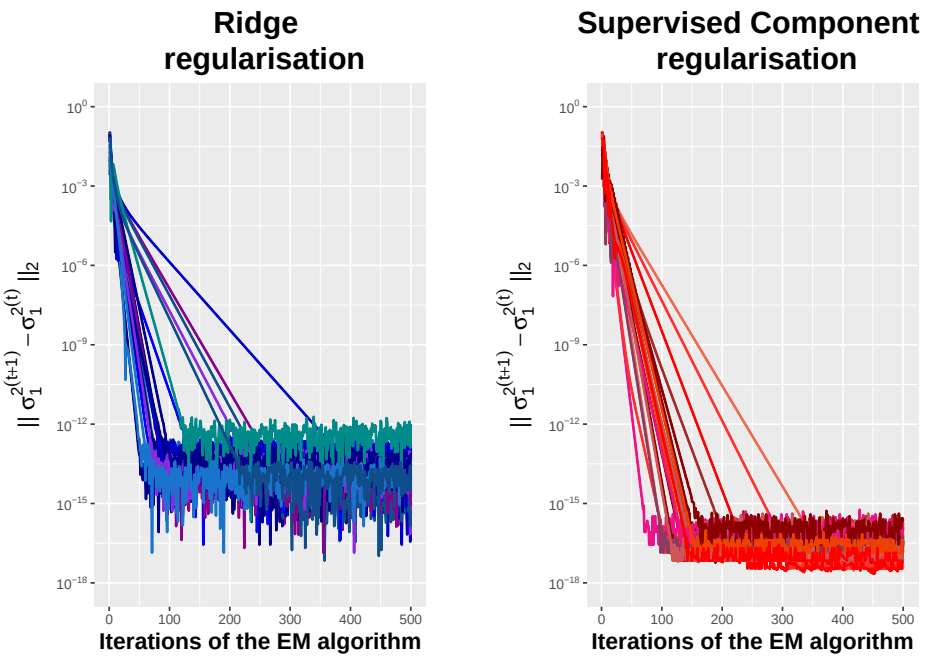


Figure 6.5 – *Second design – Convergence assessment.* *20 trajectories of* $\left\|\sigma_1^{2[t+1]} - \sigma_1^{2[t]}\right\|_2$ *are graphed, for* $t \in \{2, \ldots, 500\}$ *and for both ridge and Supervised Component regularisations.*

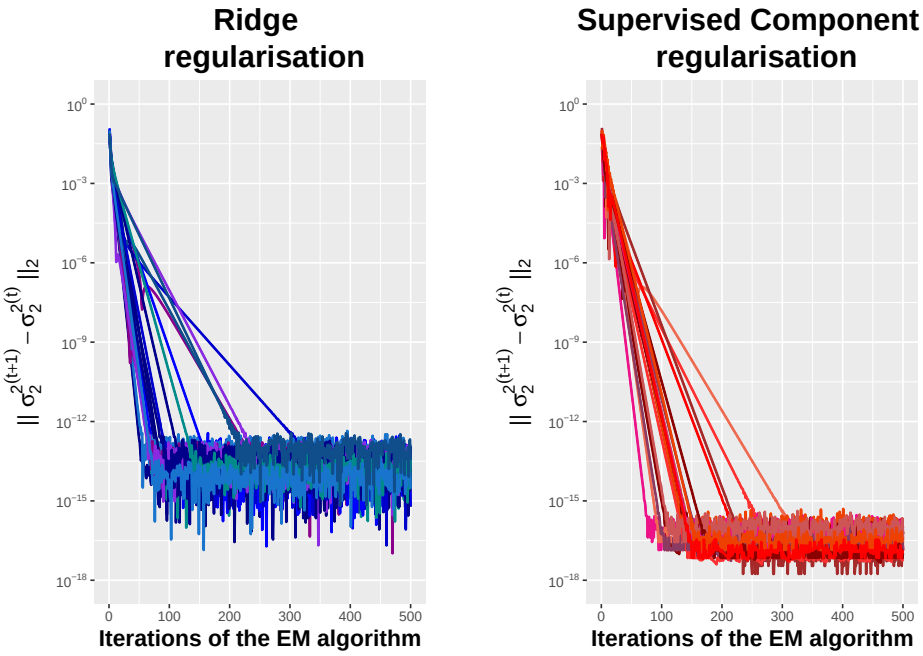


Figure 6.6 – *Second design – Convergence assessment.* *20 trajectories of* $\left\|\sigma_2^{2[t+1]} - \sigma_2^{2[t]}\right\|_2$ *are graphed, for* $t \in \{2, \ldots, 500\}$ *and for both ridge and Supervised Component regularisations.*

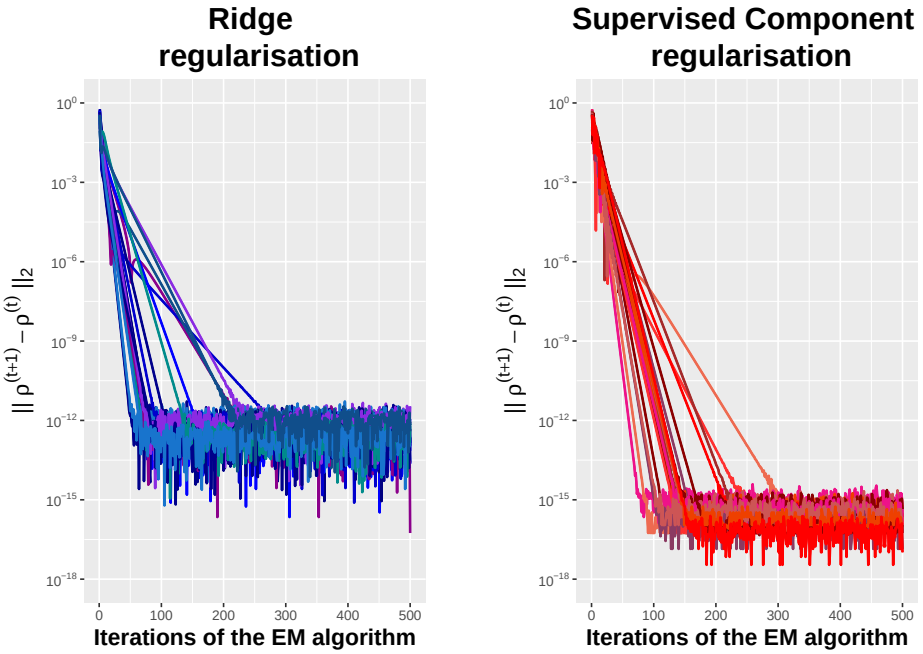


Figure 6.7 – *Second design – Convergence assessment.* *20 trajectories of* $\left\|\rho^{[t+1]} - \rho^{[t]}\right\|_2$ *are graphed, for* $t \in \{2, \ldots, 500\}$ *and for both ridge and Supervised Component regularisations.*

Sensitivity to the value of ρ. We still consider $q_1 = 10$, $q_2 = 20$, and $\tau = 0.6$. For each simulated value of $\rho \in \{-0.95, -0.9, \ldots, 0.9, 0.95\}$, 100 samples are generated. This time, Figure 6.8 shows the boxplots of the estimated values $\widehat{\rho}$ obtained with the ridge regularisation according to their real values. A similar graph is obtained with the SC–based regularised EM. The same behaviour as in the first design is observed here.

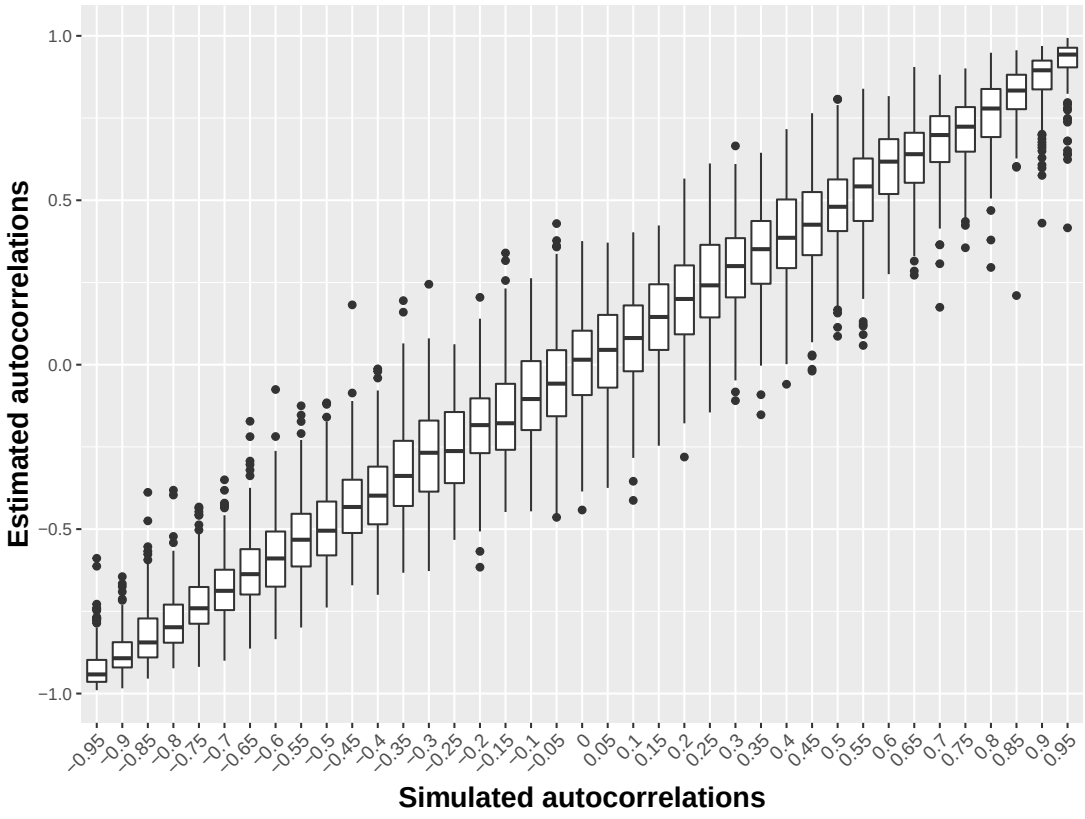


Figure 6.8 – *Second design – Sensitivity to the value of ρ.* *The graph shows the boxplots of the estimated autocorrelations $\widehat{\rho}$ obtained with the ridge regularisation according to real values.*

Comparison of estimate accuracies. The number of individuals is still set to $q_1 = 10$ while the number of time–points q_2 varies from 10 to 80. For each value of q_2, we generate 50 samples according to (6.27), with $\rho = 0.5$ and $\tau = 0.6$. To be more comprehensive, Figure 6.9 gives the RMSEs relative to fixed–effect parameter β, individual and time–specific variance components σ_1^2 and σ_2^2, and autocorrelation parameter ρ. As observed in the first design, there are no significant differences between ridge and SC concerning parameters σ_1^2, σ_2^2 and ρ. By contrast, thanks to the ability of SCEM to focus on the most predictive structures within $\boldsymbol{X}$, it performs a better estimation of the fixed effects β.

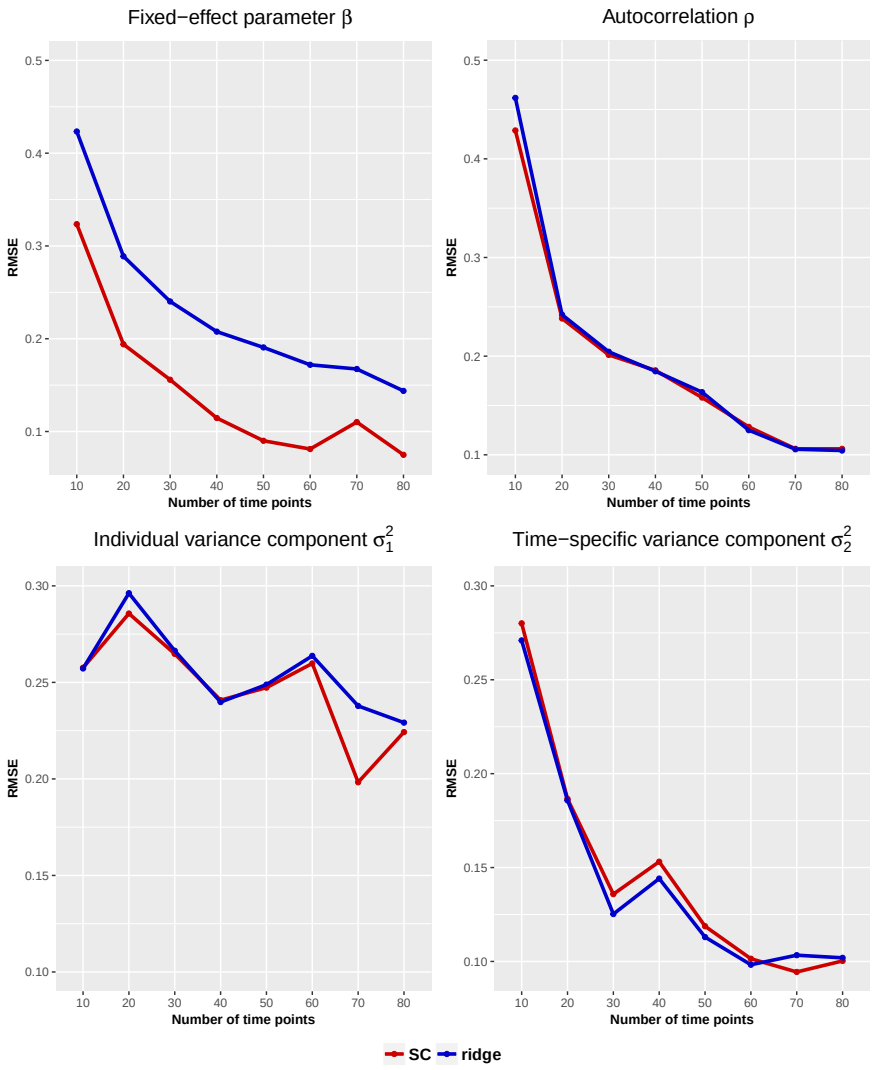


Figure 6.9 – *Second design – Estimate accuracies.* *The RMSEs of parameter estimates are represented, for both ridge and Supervised Component regularisations. They are obtained over* 50 *samples for each number of time–points* $q_2 \in \{10, 20, \ldots, 80\}$.

Model interpretation Figure 6.10 shows an example of the first component planes obtained for $q_1 = 10$, $q_2 = 20$, $\rho = 0.5$ and $\tau = 0.5$. Our cross–validation selects only two components, but component planes $(1, 3)$ and $(2, 3)$ are also represented. Due to the high level of redundancy within $\boldsymbol{X_1}$ and $\boldsymbol{X_3}$, the optimal trade–off parameter selected through cross–validation is still close to 1. Because of the presence of an isolated predictive variable, namely $\boldsymbol{x_{20}}$, the optimal bundle–locality parameter is rather close to 10. The first component aligns with bundle $\boldsymbol{X_3}$ which partially contribute to model $\boldsymbol{y}$, and the second one aligns with the single predictive variable $\boldsymbol{x_{20}}$. The nuisance bundle appears only along the third component. SCEM is therefore able to focus primarily on the predictive variable bundle, and also to detect a single predictive variable among irrelevant others.

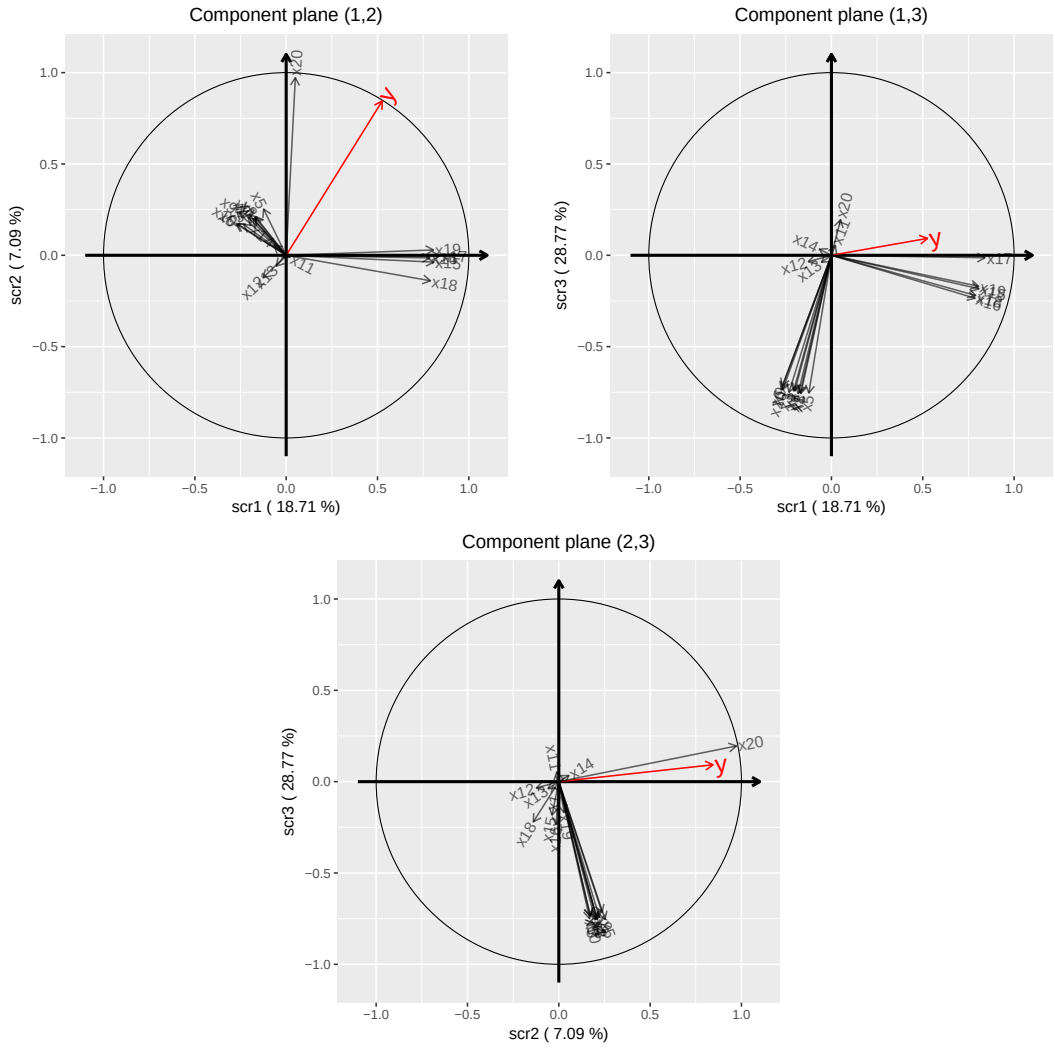


Figure 6.10 – *Second design – Model interpretation.* *We give component planes* $(1, 2)$, $(1, 3)$ *and* $(2, 3)$ *issued from SCEM. No thresholding is implemented here, so that all variables are visible on all component planes.*

6.9 Discussion and conclusions

The two methods proposed previously are intended to focus on regression modelling for panel data by means of GLMMs involving a large number of explanatory variables. They are designed to address three complementary issues.

1. The two–way random effects model allows to consider simultaneously a dependence within individuals on which data is repeatedly collected, and a time–specific effect reflecting the hidden influence of a common context (with a certain temporal inertia) shared by all the individuals.
2. GLMMs handle the wide range of response distributions considered in this work, allowing the modelling of many types of outcomes.
3. The regularisation procedures implemented address the strong correlations within the redundant explanatory variables.

There are immediate applications of the proposed methods, particularly in epidemiology. Indeed, clinical studies involve repeated observations over (long periods of) time of different patients with their own characteristics. The many variables collected are prone to high redundancies because they are often seen as proxies of latent phenomena that are difficult to measure directly (well–being for instance). It is then almost impossible to separate the contribution of each explanatory variable, which greatly complicates the model interpretation.

Our ridge–penalised EM algorithm is a simple way to solve the problem since the L_2–regularisation mechanically reduces the variance of the fixed–effect estimator. At first sight, the iterative calibration of the shrinkage parameter at each iteration of the EM may seem odd in the Gaussian case since the underlying model does not change, but it avoids the high algorithmic cost of a classical cross-validation. However, this auto–calibration becomes necessary in non–Gaussian cases, because the initial model is linearised at each iteration. The major drawback of ridge regression is that it considers the high correlations among the explanatory variables as a pure nuisance instead of a possible asset, resulting in a model that remains difficult to interpret.

The supervised component–based regularisation is designed to address this drawback: in the EM algorithm, instead of subtracting a penalty term to the likelihood, we add a bonus term to favour the alignment of components on strong directions. The theory of the penalised EM algorithm also applies in this framework. Compared to ridge, estimates provided by the SC–regularisation are generally more stable and accurate. In addition, it makes interpretation of the linear predictor easier through its decomposition on interpretable components.

6.10 Appendices

6.10.1 Calculus — Ridge EM for Gaussian panel data

6.10.1.1 Reminder of the model, notations and hypothesis

Recall that Section 6.5 focuses on the Gaussian panel data model given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U_1}\boldsymbol{\xi_1} + \boldsymbol{U_2}\boldsymbol{\xi_2} + \boldsymbol{U_0}\boldsymbol{\xi_0}. \tag{6.28}$$

q_1 being the number of individuals and q_2 the number of time–points, random effects design matrices in (6.28) write $\boldsymbol{U_1} = \boldsymbol{I}_{q_1} \otimes \boldsymbol{1}_{q_2}$, $\boldsymbol{U_2} = \boldsymbol{1}_{q_1} \otimes \boldsymbol{I}_{q_2}$, and $\boldsymbol{U_0} = \boldsymbol{I}_n$, where $n = q_0 = q_1 \times q_2$. Besides, in Sections 6.2 and 6.5, it was assumed that

- $\boldsymbol{\xi_1} \sim \mathcal{N}_{q_1}(\boldsymbol{0}, \boldsymbol{D_1})$, where $\boldsymbol{D_1} = \sigma_1^2 \boldsymbol{A_1}$,
- $\boldsymbol{\xi_2} \sim \mathcal{N}_{q_2}(\boldsymbol{0}, \boldsymbol{D_2})$, where $\boldsymbol{D_2} = \sigma_2^2 \boldsymbol{A_2}(\rho)$,
- $\boldsymbol{\xi_0} \sim \mathcal{N}_{q_0}(\boldsymbol{0}, \boldsymbol{D_0})$, where $\boldsymbol{D_0} = \sigma_0^2 \boldsymbol{A_0}$,
- $\boldsymbol{\xi_0}$, $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$ are independent.

We also recall the expression of the associated objective function already defined by (6.17):

$$\mathcal{Q}_{\text{pen}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right) = \text{cte} - \frac{1}{2}\left\{\sum_{j=0}^{2} q_j \log\left(\sigma_j^2\right) + \sum_{j=0}^{1} \frac{\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi_j}^{\mathrm{T}} \boldsymbol{A_j}^{-1} \boldsymbol{\xi_j} \,|\, \boldsymbol{\theta}^{[t]}\right)}{\sigma_j^2} + \frac{\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi_2}^{\mathrm{T}} \boldsymbol{A_2}^{-1}(\rho) \boldsymbol{\xi_2} \,|\, \boldsymbol{\theta}^{[t]}\right)}{\sigma_2^2} + \log\left(|\boldsymbol{A_2}(\rho)|\right) + \lambda \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\beta}\right\}.$$

6.10.1.2 Updating the autocorrelation and the time–specific variance component

System (6.18) follows from the following lemma:

Lemma 6.2. Let q_2 be the number of time–points. Recall that correlation matrix $\boldsymbol{A_2}(\rho) = \left(\frac{\rho^{|i-j|}}{1-\rho^2}\right)_{1 \leqslant i,j \leqslant q_2}$ is of size $(q_2 \times q_2)$, and let us note momentarily $|\boldsymbol{A_2}^{q_2}(\rho)|$ its determinant. We then have

$$\forall q_2 \in \mathbb{N}^\star,\ |\boldsymbol{A_2}^{q_2}(\rho)| = (-1)^{q_2}\left(-1+\rho^2\right)^{q_2-2}.$$

Proof. By mathematical induction. Let us define $\mathcal{P}(q_2)$ the statement

$$\mathcal{P}(q_2): \ |\boldsymbol{A}_{\mathbf{2}}^{q_2}(\rho)| = (-1)^{q_2}\left(-1+\rho^2\right)^{q_2-2}.$$

- $\mathcal{P}(1)$ is easily seen to be true since

$$\left|\boldsymbol{A}_{\mathbf{2}}^{1}(\rho)\right| = \left|\frac{1}{1-\rho^2}\right| = \frac{1}{1-\rho^2} = (-1)^{1}\left(-1+\rho^2\right)^{1-2}.$$

- Assume $\mathcal{P}(q_2)$ holds for some value of q_2. We then have

$$\begin{aligned}
\left|\boldsymbol{A}_{\mathbf{2}}^{q_2+1}(\rho)\right| &= \left|\frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{q_2-1} & \rho^{q_2} \\ \rho & 1 & \rho & \cdots & \rho^{q_2-2} & \rho^{q_2-1} \\ \rho^2 & \rho & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \rho & \rho^2 \\ \rho^{q_2-1} & \rho^{q_2-2} & \cdots & \rho & 1 & \rho \\ \rho^{q_2} & \rho^{q_2-1} & \ddots & \ddots & \rho & 1 \end{pmatrix}\right| \\
&= \left|\frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{q_2-1} & 0 \\ \rho & 1 & \rho & \cdots & \rho^{q_2-2} & 0 \\ \rho^2 & \rho & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \rho & 0 \\ \rho^{q_2-1} & \rho^{q_2-2} & \cdots & \rho & 1 & 0 \\ \rho^{q_2} & \rho^{q_2-1} & \ddots & \ddots & \rho & 1-\rho^2 \end{pmatrix}\right|
\end{aligned}$$

  by performing the column operation: $\text{col}_{q_2+1} \leftarrow \text{col}_{q_2+1} - \rho\text{col}_{q_2}$. Using the induction hypothesis that $\mathcal{P}(q_2)$ holds, we have

$$\begin{aligned}
\left|\boldsymbol{A}_{\mathbf{2}}^{q_2+1}(\rho)\right| &= \left(1-\rho^2\right) \times (-1)^{2(q_2+1)}\left|\boldsymbol{A}_{\mathbf{2}}^{q_2}(\rho)\right| \\
&= -\left(-1+\rho^2\right)(-1)^{q_2}\left(-1+\rho^2\right)^{q_2-2} \\
&= (-1)^{q_2+1}\left(-1+\rho^2\right)^{q_2-1},
\end{aligned}$$

  which concludes the proof.

□

The following equivalences hold:

$$\begin{cases} \dfrac{\partial}{\partial\rho}\left[\mathcal{Q}_{\text{pen}}\left(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{[t]}\right)\right] = 0 \\ \dfrac{\partial}{\partial\sigma_2^2}\left[\mathcal{Q}_{\text{pen}}\left(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{[t]}\right)\right] = 0 \end{cases}$$

$$\iff \begin{cases} \dfrac{\partial}{\partial\rho}\left[\log\left(|\boldsymbol{A_2}(\rho)|\right) + \dfrac{1}{\sigma_2^2}\,\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi_2}^{\mathrm{T}}\boldsymbol{A_2}^{-1}(\rho)\boldsymbol{\xi_2}\,|\,\boldsymbol{\theta}^{[t]}\right)\right] = 0 \\ \dfrac{\partial}{\partial\sigma_2^2}\left[q_2\log\left(\sigma_2^2\right) + \dfrac{1}{\sigma_2^2}\,\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi_2}^{\mathrm{T}}\boldsymbol{A_2}^{-1}(\rho)\boldsymbol{\xi_2}\,|\,\boldsymbol{\theta}^{[t]}\right)\right] = 0 \end{cases}$$

$$\iff \begin{cases} \dfrac{\partial}{\partial\rho}\left[\log\left((-1)^{q_2}\left(-1+\rho^2\right)^{q_2-2}\right)\right] + \dfrac{1}{\sigma_2^2}\,\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi_2}^{\mathrm{T}}\dfrac{\partial\boldsymbol{A_2}^{-1}(\rho)}{\partial\rho}\boldsymbol{\xi_2}\,|\,\boldsymbol{\theta}^{[t]}\right) = 0 \\ \dfrac{q_2}{\sigma_2^2} - \dfrac{\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi_2}^{\mathrm{T}}\boldsymbol{A_2}^{-1}(\rho)\boldsymbol{\xi_2}\,|\,\boldsymbol{\theta}^{[t]}\right)}{\sigma_2^4} = 0 \end{cases}$$

$$\iff \begin{cases} \dfrac{2\rho(q_2-2)}{-1+\rho^2}\,\sigma_2^2 + \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi_2}^{\mathrm{T}}\dfrac{\partial\boldsymbol{A_2}^{-1}(\rho)}{\partial\rho}\boldsymbol{\xi_2}\,|\,\boldsymbol{\theta}^{[t]}\right) = 0 & \text{(6.29a)} \\ \sigma_2^2 = q_2^{-1}\,\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi_2}^{\mathrm{T}}\boldsymbol{A_2}^{-1}(\rho)\boldsymbol{\xi_2}\,|\,\boldsymbol{\theta}^{[t]}\right) & \text{(6.29b)} \end{cases}$$

Putting back (6.29b) into (6.29a) provides

$$\frac{2(q_2-2)\rho}{q_2(-1+\rho^2)}\,\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi_2}^{\mathrm{T}}\boldsymbol{A_2}^{-1}(\rho)\boldsymbol{\xi_2}\,|\,\boldsymbol{\theta}^{[t]}\right) + \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi_2}^{\mathrm{T}}\frac{\partial\boldsymbol{A_2}^{-1}(\rho)}{\partial\rho}\boldsymbol{\xi_2}\,|\,\boldsymbol{\theta}^{[t]}\right) = 0.$$

By defining $\boldsymbol{S}_\rho := \boldsymbol{A_2}^{-1}(\rho)$, $\boldsymbol{S}'_\rho := \dfrac{\partial\boldsymbol{A_2}^{-1}(\rho)}{\partial\rho}$, and $\boldsymbol{K}_\rho := \dfrac{2(q_2-2)\rho}{q_2(-1+\rho^2)}\boldsymbol{S}_\rho + \boldsymbol{S}'_\rho$ (6.29) is equivalent to

$$\begin{cases} J^{[t]}(\rho) := \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi_2}^{\mathrm{T}}\boldsymbol{K}_\rho\boldsymbol{\xi_2}\,|\,\boldsymbol{\theta}^{[t]}\right) = 0 \\ \sigma_2^2 = q_2^{-1}\,\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi_2}^{\mathrm{T}}\boldsymbol{S}_\rho\boldsymbol{\xi_2}\,|\,\boldsymbol{\theta}^{[t]}\right). \end{cases}$$

Finally, as suggested by Section 6.5, update $\rho^{[t+1]}$ that fulfils

$$J^{[t]}\left(\rho^{[t+1]}\right) = 0$$

can be found using a Newton–Raphson method. Then, in a second time, this allows the time–specific variance component to be updated as

$$\sigma_2^{2\,[t+1]} = q_2^{-1}\,\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi_2}^{\mathrm{T}}\boldsymbol{S}_{\rho^{[t+1]}}\boldsymbol{\xi_2}\,|\,\boldsymbol{\theta}^{[t]}\right).$$

6.10.1.3 Updating the individual–specific and the residual variance components

For each $j \in \{0, 1\}$, the first–order conditions lead to

$$\begin{aligned}
& \frac{\partial}{\partial \sigma_j^2}\left[\mathcal{Q}_{\text{pen}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right)\right] = 0 \\
\iff \quad & \frac{\partial}{\partial \sigma_j^2}\left[q_j \log\left(\sigma_j^2\right) + \frac{\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_j^{\mathrm{T}} \boldsymbol{A}_j^{-1} \boldsymbol{\xi}_j \,|\, \boldsymbol{\theta}^{[t]}\right)}{\sigma_j^2}\right] = 0 \\
\iff \quad & \frac{q_j}{\sigma_j^2} - \frac{\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_j^{\mathrm{T}} \boldsymbol{A}_j^{-1} \boldsymbol{\xi}_j \,|\, \boldsymbol{\theta}^{[t]}\right)}{\sigma_j^4} = 0 \\
\iff \quad & \sigma_j^2 = q_j^{-1}\, \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_j^{\mathrm{T}} \boldsymbol{A}_j^{-1} \boldsymbol{\xi}_j \,|\, \boldsymbol{\theta}^{[t]}\right).
\end{aligned}$$

The associated updates then simply write

$$\forall j \in \{0, 1\}\,,\ \sigma_j^{2[t+1]} = q_j^{-1}\, \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_j^{\mathrm{T}} \boldsymbol{A}_j^{-1} \boldsymbol{\xi}_j \,|\, \boldsymbol{\theta}^{[t]}\right). \tag{6.30}$$

6.10.1.4 Updating the fixed–effect parameter

Concerning the fixed–effect parameter, we have

$$\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\beta}}\left[\mathcal{Q}_{\text{pen}}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]}\right)\right] = \mathbf{0} \\
\iff \quad & \frac{\partial}{\partial \boldsymbol{\beta}}\left[\frac{1}{\sigma_0^2}\, \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_{\mathbf{0}}^{\mathrm{T}} \boldsymbol{A}_{\mathbf{0}}^{-1} \boldsymbol{\xi}_{\mathbf{0}} \,|\, \boldsymbol{\theta}^{[t]}\right) + \lambda \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\beta}\right] = \mathbf{0} \\
\iff \quad & \frac{\partial}{\partial \boldsymbol{\beta}}\left\{\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j=1}^{2} \boldsymbol{U}_j \boldsymbol{\xi}_j\right)^{\mathrm{T}} \boldsymbol{A}_{\mathbf{0}}^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \sum_{j=1}^{2} \boldsymbol{U}_j \boldsymbol{\xi}_j\right) \,|\, \boldsymbol{\theta}^{[t]}\right]\right. \\
& \qquad \left. + \lambda \sigma_0^2 \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\beta}\right\} = \mathbf{0} \\
\iff \quad & \frac{\partial}{\partial \boldsymbol{\beta}}\left\{\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left[-2\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{A}_{\mathbf{0}}^{-1} \left(\boldsymbol{y} - \sum_{j=1}^{2} \boldsymbol{U}_j \boldsymbol{\xi}_j\right) \,|\, \boldsymbol{\theta}^{[t]}\right] + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{A}_{\mathbf{0}}^{-1} \boldsymbol{X} \boldsymbol{\beta}\right. \\
& \qquad \left. + \lambda \sigma_0^2 \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\beta}\right\} = \mathbf{0}.
\end{aligned}$$

As a result,

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\beta}} \left[\mathcal{Q}_{\text{pen}} \left(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{[t]} \right) \right] = \mathbf{0} \\
\iff \quad & - \cancel{2} \boldsymbol{X}^{\intercal} \boldsymbol{A}_{\mathbf{0}}^{-1} \, \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}} \left(\boldsymbol{y} - \sum_{j=1}^{2} \boldsymbol{U}_j \boldsymbol{\xi}_j \,|\, \boldsymbol{\theta}^{[t]} \right) + \cancel{2} \boldsymbol{X}^{\intercal} \boldsymbol{A}_{\mathbf{0}}^{-1} \boldsymbol{X} \boldsymbol{\beta} + \cancel{2} \lambda \sigma_0^2 \boldsymbol{\beta} = \mathbf{0} \\
\iff \quad & \left(\boldsymbol{X}^{\intercal} \boldsymbol{A}_{\mathbf{0}}^{-1} \boldsymbol{X} + \lambda \sigma_0^2 \boldsymbol{I}_p \right) \boldsymbol{\beta} = \boldsymbol{X}^{\intercal} \boldsymbol{A}_{\mathbf{0}}^{-1} \, \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}} \left(\boldsymbol{y} - \sum_{j=1}^{2} \boldsymbol{U}_j \boldsymbol{\xi}_j \,|\, \boldsymbol{\theta}^{[t]} \right) \\
\iff \quad & \boldsymbol{\beta} = \left(\boldsymbol{X}^{\intercal} \boldsymbol{A}_{\mathbf{0}}^{-1} \boldsymbol{X} + \lambda \sigma_0^2 \boldsymbol{I}_p \right)^{-1} \boldsymbol{X}^{\intercal} \boldsymbol{A}_{\mathbf{0}}^{-1} \, \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}} \left(\boldsymbol{y} - \sum_{j=1}^{2} \boldsymbol{U}_j \boldsymbol{\xi}_j \,|\, \boldsymbol{\theta}^{[t]} \right).
\end{aligned}
$$

This gives the following update:

$$
\boldsymbol{\beta}^{[t+1]} = \left(\boldsymbol{X}^{\intercal} \boldsymbol{A}_{\mathbf{0}}^{-1} \boldsymbol{X} + \lambda \sigma_0^{2[t+1]} \boldsymbol{I}_p \right)^{-1} \boldsymbol{X}^{\intercal} \boldsymbol{A}_{\mathbf{0}}^{-1} \, \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}} \left(\boldsymbol{y} - \sum_{j=1}^{2} \boldsymbol{U}_j \boldsymbol{\xi}_j \,|\, \boldsymbol{\theta}^{[t]} \right), \quad (6.31)
$$

where $\sigma_0^{2[t+1]}$ is given by (6.30). As suggested in Section 6.5, it is possible to take advantage of the updates $\rho^{[t+1]}$ and $\sigma_j^{2[t+1]}, j \in \{0, 1, 2\}$, to calibrate the shrinkage parameter λ at each iteration by generalised cross–validation.

6.10.1.5 Explicit expressions of conditional expectations

Since $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$ are independent, their joint distribution is given by

$$
\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\xi_1} \\ \boldsymbol{\xi_2} \end{pmatrix} \sim \mathcal{N}_{q_1+q_2} \left(\mathbf{0}, \boldsymbol{D} \right),
$$

where $\boldsymbol{D} = \mathbf{bDiag}\left(\boldsymbol{D_1}, \boldsymbol{D_2} \right)$, $\boldsymbol{D_1} = \sigma_1^2 \boldsymbol{A_1}$, $\boldsymbol{D_2} = \sigma_2^2 \boldsymbol{A_2}(\rho)$. In addition, the joint distribution of Y and $\boldsymbol{\xi}$ writes

$$
\begin{pmatrix} Y \\ \boldsymbol{\xi} \end{pmatrix} \sim \mathcal{N}_{q_0+q_1+q_2} \left(\begin{bmatrix} \boldsymbol{X}\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Gamma} & \boldsymbol{U}\boldsymbol{D} \\ \boldsymbol{D}\boldsymbol{U}^{\intercal} & \boldsymbol{D} \end{bmatrix} \right),
$$

where $\boldsymbol{\Gamma} = \text{Var}(Y) = \sum_{j=0}^{1} \left(\sigma_j^2 \, \boldsymbol{U}_j \boldsymbol{A}_j \boldsymbol{U}_j^{\intercal} \right) + \sigma_2^2 \, \boldsymbol{U_2} \boldsymbol{A_2}(\rho) \boldsymbol{U_2^{\intercal}}$. In what follows, we define

$$
\begin{cases} \boldsymbol{V}_j = \boldsymbol{U}_j \boldsymbol{A}_j \boldsymbol{U}_j^{\intercal}, \; j \in \{0, 1\} \\ \boldsymbol{V_2} = \boldsymbol{U_2} \boldsymbol{A_2}(\rho) \boldsymbol{U_2^{\intercal}} \end{cases}
$$

so that matrix $\boldsymbol{\Gamma}$ simply rewrites

$$\boldsymbol{\Gamma} = \sum_{j=0}^{2} \sigma_j^2 \boldsymbol{V}_j.$$

Explicit expressions of conditional expectations are based on the following proposition.

Proposition 6.3. (i) The conditioning properties of the Gaussian distribution yields the following conditional distribution:

$$\boldsymbol{\xi}|Y = \boldsymbol{y} \sim \mathcal{N}_{q_1+q_2}\left(\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right), \boldsymbol{D} - \boldsymbol{D}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\boldsymbol{U}\boldsymbol{D}\right).$$

(ii) For any symmetric matrix $\boldsymbol{M}$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\xi}\right) = &\left[\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)\right]^{\mathrm{T}}\boldsymbol{M}\left[\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)\right] \\ &+ \text{Trace}\left[\boldsymbol{M}\left(\boldsymbol{D} - \boldsymbol{D}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\boldsymbol{U}\boldsymbol{D}\right)\right].\end{aligned}$$

To avoid any ambiguity, note that at step t of the algorithm, random–effect variance matrices are given by

$$\begin{cases}\boldsymbol{D}_{\boldsymbol{j}}^{[t]} = {\sigma_j^2}^{[t]}\boldsymbol{A}_{\boldsymbol{j}},\ j \in \{0,1\} \\ \boldsymbol{D}_{\boldsymbol{2}}^{[t]} = {\sigma_2^2}^{[t]}\boldsymbol{A}_{\boldsymbol{2}}^{[t]},\ \text{where } \boldsymbol{A}_{\boldsymbol{2}}^{[t]} = \boldsymbol{A}_{\boldsymbol{2}}\left(\rho^{[t]}\right),\end{cases}$$

and

$$\text{Var}^{[t]}(Y) = \boldsymbol{\Gamma}^{[t]} = \sum_{j=0}^{1}\left({\sigma_j^2}^{[t]}\boldsymbol{V}_{\boldsymbol{j}}\right) + {\sigma_2^2}^{[t]}{\boldsymbol{V}_{\boldsymbol{2}}}^{[t]},$$

where ${\boldsymbol{V}_{\boldsymbol{2}}}^{[t]} = \boldsymbol{U}_{\boldsymbol{2}}^{\mathrm{T}}\boldsymbol{A}_{\boldsymbol{2}}^{[t]}\boldsymbol{U}_{\boldsymbol{2}}$.

Conditional expectation involved in updating individual and residual variance components: For $j \in \{0,1\}$,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_{\boldsymbol{j}}^{\mathrm{T}}\boldsymbol{A}_{\boldsymbol{j}}^{-1}\boldsymbol{\xi}_{\boldsymbol{j}} \,|\, \boldsymbol{\theta}^{[t]}\right) \\ &\quad= \left[\boldsymbol{D}_{\boldsymbol{j}}^{[t]}\boldsymbol{U}_{\boldsymbol{j}}^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{[t]}\right)\right]^{\mathrm{T}}\boldsymbol{A}_{\boldsymbol{j}}^{-1}\left[\boldsymbol{D}_{\boldsymbol{j}}^{[t]}\boldsymbol{U}_{\boldsymbol{j}}^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{[t]}\right)\right] \\ &\qquad+ \text{Trace}\left[\boldsymbol{A}_{\boldsymbol{j}}^{-1}\left(\boldsymbol{D}_{\boldsymbol{j}}^{[t]} - \boldsymbol{D}_{\boldsymbol{j}}^{[t]}\boldsymbol{U}_{\boldsymbol{j}}^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\boldsymbol{U}_{\boldsymbol{j}}\boldsymbol{D}_{\boldsymbol{j}}^{[t]}\right)\right] \\ &\quad= \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{[t]}\right)^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\boldsymbol{U}_{\boldsymbol{j}}\left({\sigma_j^2}^{[t]}\boldsymbol{A}_{\boldsymbol{j}}\right)\cancel{\boldsymbol{A}_{\boldsymbol{j}}^{-1}}\left({\sigma_j^2}^{[t]}\cancel{\boldsymbol{A}_{\boldsymbol{j}}}\right)\boldsymbol{U}_{\boldsymbol{j}}^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{[t]}\right) \\ &\qquad+ \text{Trace}\left[\boldsymbol{A}_{\boldsymbol{j}}^{-1}\left({\sigma_j^2}^{[t]}\boldsymbol{A}_{\boldsymbol{j}}\right)\right] - \text{Trace}\left[\cancel{\boldsymbol{A}_{\boldsymbol{j}}^{-1}}\left({\sigma_j^2}^{[t]}\cancel{\boldsymbol{A}_{\boldsymbol{j}}}\right)\boldsymbol{U}_{\boldsymbol{j}}^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\boldsymbol{U}_{\boldsymbol{j}}\left({\sigma_j^2}^{[t]}\boldsymbol{A}_{\boldsymbol{j}}\right)\right] \\ &\quad= {\sigma_j^4}^{[t]}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{[t]}\right)^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\boldsymbol{V}_{\boldsymbol{j}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{[t]}\right) \\ &\qquad+ q_j{\sigma_j^2}^{[t]} - {\sigma_j^4}^{[t]}\,\text{Trace}\left({\boldsymbol{\Gamma}^{[t]}}^{-1}\boldsymbol{V}_{\boldsymbol{j}}\right).\end{aligned}$$

Conditional expectation involved in updating autocorrelation and time–specific variance component:

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_{\mathbf{2}}^{\mathrm{T}}\boldsymbol{K}_{\rho}\boldsymbol{\xi}_{\mathbf{2}} \mid \boldsymbol{\theta}^{[t]}\right) \\
&\quad = \sigma_2^{4[t]}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{[t]}\right)^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\boldsymbol{U}_{\mathbf{2}}\boldsymbol{A}_{\mathbf{2}}^{[t]}\boldsymbol{K}_{\rho}\boldsymbol{A}_{\mathbf{2}}^{[t]}\boldsymbol{U}_{\mathbf{2}}^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{[t]}\right) \\
&\qquad + \sigma_2^{2[t]}\,\mathrm{Trace}\left(\boldsymbol{K}_{\rho}\boldsymbol{A}_{\mathbf{2}}^{[t]}\right) - \sigma_2^{4[t]}\,\mathrm{Trace}\left(\boldsymbol{K}_{\rho}\boldsymbol{A}_{\mathbf{2}}^{[t]}\boldsymbol{U}_{\mathbf{2}}^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\boldsymbol{U}_{\mathbf{2}}\boldsymbol{A}_{\mathbf{2}}^{[t]}\right).
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_{\mathbf{2}}^{\mathrm{T}}\boldsymbol{S}_{\rho^{[t+1]}}\boldsymbol{\xi}_{\mathbf{2}} \mid \boldsymbol{\theta}^{[t]}\right) \\
&= \sigma_2^{4[t]}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{[t]}\right)^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\boldsymbol{U}_{\mathbf{2}}\boldsymbol{A}_{\mathbf{2}}^{[t]}\boldsymbol{S}_{\rho^{[t+1]}}\boldsymbol{A}_{\mathbf{2}}^{[t]}\boldsymbol{U}_{\mathbf{2}}^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{[t]}\right) \\
&\quad + \sigma_2^{2[t]}\,\mathrm{Trace}\left(\boldsymbol{S}_{\rho^{[t+1]}}\boldsymbol{A}_{\mathbf{2}}^{[t]}\right) - \sigma_2^{4[t]}\,\mathrm{Trace}\left(\boldsymbol{S}_{\rho^{[t+1]}}\boldsymbol{A}_{\mathbf{2}}^{[t]}\boldsymbol{U}_{\mathbf{2}}^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\boldsymbol{U}_{\mathbf{2}}\boldsymbol{A}_{\mathbf{2}}^{[t]}\right) \\
&= \sigma_2^{4[t]}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{[t]}\right)^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\boldsymbol{U}_{\mathbf{2}}\,\boldsymbol{A}_{\mathbf{2}}\big(\rho^{[t]}\big)\,\boldsymbol{A}_{\mathbf{2}}^{-1}\big(\rho^{[t+1]}\big)\,\boldsymbol{A}_{\mathbf{2}}\big(\rho^{[t]}\big)\,\boldsymbol{U}_{\mathbf{2}}^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{[t]}\right) \\
&\quad + \sigma_2^{2[t]}\,\mathrm{Trace}\left[\boldsymbol{A}_{\mathbf{2}}^{-1}\big(\rho^{[t+1]}\big)\,\boldsymbol{A}_{\mathbf{2}}\big(\rho^{[t]}\big)\right] \\
&\quad - \sigma_2^{4[t]}\,\mathrm{Trace}\left[\boldsymbol{A}_{\mathbf{2}}^{-1}\big(\rho^{[t+1]}\big)\,\boldsymbol{A}_{\mathbf{2}}\big(\rho^{[t]}\big)\,\boldsymbol{U}_{\mathbf{2}}^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\boldsymbol{U}_{\mathbf{2}}\,\boldsymbol{A}_{\mathbf{2}}\big(\rho^{[t]}\big)\right].
\end{aligned}
$$

Conditional expectation involved in updating fixed–effect parameter:

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(Y-\sum_{j=1}^{2}\boldsymbol{U}_{\boldsymbol{j}}\boldsymbol{\xi}_{\boldsymbol{j}} \mid \boldsymbol{\theta}^{[t]}\right) &= \boldsymbol{y}-\sum_{j=1}^{2}\boldsymbol{U}_{\boldsymbol{j}}\,\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\boldsymbol{\xi}_{\boldsymbol{j}} \mid \boldsymbol{\theta}^{[t]}\right) \\
&= \boldsymbol{y}-\sum_{j=1}^{2}\boldsymbol{U}_{\boldsymbol{j}}\boldsymbol{D}_{\boldsymbol{j}}^{[t]}\boldsymbol{U}_{\boldsymbol{j}}^{\mathrm{T}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{[t]}\right) \\
&= \boldsymbol{y}-\sum_{j=1}^{2}\sigma_j^{2[t]}\boldsymbol{V}_{\boldsymbol{j}}^{[t]}{\boldsymbol{\Gamma}^{[t]}}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{[t]}\right) \\
&= \boldsymbol{y}-\left(\boldsymbol{\Gamma}^{[t]}-\sigma_0^{2[t]}\boldsymbol{V}_{\mathbf{0}}\right){\boldsymbol{\Gamma}^{[t]}}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{[t]}\right) \\
&= \boldsymbol{X}\boldsymbol{\beta}^{[t]}+\sigma_0^{2[t]}\boldsymbol{V}_{\mathbf{0}}{\boldsymbol{\Gamma}^{[t]}}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{[t]}\right).
\end{aligned}
$$

6.10.2 Calculus — SCEM for Gaussian panel data

6.10.2.1 Updating loading–vector $\boldsymbol{u_k}$ and its associated parameter γ_k

Recall that the updates of the loading–vector $\boldsymbol{u_k}$ and its associated parameter γ_k involve objective function $\widetilde{\mathcal{Q}}_{\text{SC}}^{k}$ defined as

$$\widetilde{\mathcal{Q}}_{\text{SC}}^{k}\left(\boldsymbol{u_k}, \gamma_k \,|\, \boldsymbol{\theta}^{[t]}\right) = s\, \log\left[\phi\left(\boldsymbol{u_k}\right)\right] - (1-s) \times \frac{\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\left\|\boldsymbol{y} - \sum_{h=0}^{k-1}\left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{h}}^{[t+1]}\right)\gamma_h^{[t+1]} - \left(\boldsymbol{X}\boldsymbol{u_k}\right)\gamma_k - \boldsymbol{U}\boldsymbol{\xi}\right\|_{\boldsymbol{A_0^{-1}}}^{2} \,|\, \boldsymbol{\theta}^{[t]}\right)}{2\sigma_0^{2\,[t+1]}}.$$

The new values of $\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]}$ and $\gamma_k^{[t+1]}$ are then obtained by setting

$$\begin{aligned} \boldsymbol{u}_{\boldsymbol{k}}^{[t+1]} &= \underset{\boldsymbol{u_k} \in \mathcal{S}_k^{[t+1]}}{\arg\max}\ \widetilde{\mathcal{Q}}_{\text{SC}}^{k}\left(\boldsymbol{u_k}, \gamma_k^{[t]} \,|\, \boldsymbol{\theta}^{[t]}\right) \\ \gamma_k^{[t+1]} &= \underset{\gamma_k}{\arg\max}\ \widetilde{\mathcal{Q}}_{\text{SC}}^{k}\left(\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]}, \gamma_k \,|\, \boldsymbol{\theta}^{[t]}\right). \end{aligned} \tag{6.32}$$

In (6.32),

$$\mathcal{S}_k^{[t+1]} = \left\{\boldsymbol{u} \in \mathbb{R}^p \mid \boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{u} = 1 \text{ and } \left(\boldsymbol{X}\boldsymbol{u}\right)^{\mathrm{T}}\boldsymbol{P}\boldsymbol{F}_{\boldsymbol{k-1}}^{[t+1]} = \boldsymbol{0}\right\},$$

where $\boldsymbol{P}$ reflects the a priori relative importance of observations (by default $\boldsymbol{P} = \frac{1}{n}\boldsymbol{I}_n$), and $\boldsymbol{F}_{\boldsymbol{k-1}}^{[t+1]}$ concatenates the first $k-1$ components computed at iteration $t+1$, namely

$$\boldsymbol{F}_{\boldsymbol{k-1}}^{[t+1]} = \left[\, \boldsymbol{X}\boldsymbol{u}_{\boldsymbol{1}}^{[t+1]} \mid \ldots \mid \boldsymbol{X}\boldsymbol{u}_{\boldsymbol{k-1}}^{[t+1]} \,\right].$$

The first maximisation is performed using the PING algorithm, detailed in Appendix 5.8.3. For the second maximisation, let us note $\boldsymbol{B}_{\boldsymbol{k-1}}^{[t+1]} = \sum_{h=0}^{k-1}\left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{h}}^{[t+1]}\right)\gamma_h^{[t+1]}$. The maximisation program

$$\max_{\gamma_k}\left\{\widetilde{\mathcal{Q}}_{\text{SC}}^{k}\left(\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]}, \gamma_k \,|\, \boldsymbol{\theta}^{[t]}\right)\right\}$$

is equivalent to the following minimisation:

$$\min_{\gamma_k}\left\{\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}}\left(\left\|\left(\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\xi} - \boldsymbol{B}_{\boldsymbol{k-1}}^{[t+1]}\right) - \left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]}\right)\gamma_k\right\|_{\boldsymbol{A_0^{-1}}}^{2} \,|\, \boldsymbol{\theta}^{[t]}\right)\right\}.$$

We then have

$$\begin{aligned}
& \frac{\partial}{\partial \gamma_k} \left\{ \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}} \left(\left\| \left(\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\xi} - \boldsymbol{B}_{\boldsymbol{k-1}}^{[t+1]} \right) - \left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]} \right) \gamma_k \right\|_{\boldsymbol{A}_{\boldsymbol{0}}^{-1}}^2 \mid \boldsymbol{\theta}^{[t]} \right) \right\} = 0 \\
\iff & \frac{\partial}{\partial \gamma_k} \Bigg\{ \gamma_k^2 \left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]} \right)^{\mathrm{T}} \boldsymbol{A}_{\boldsymbol{0}}^{-1} \left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]} \right) \\
& \qquad - 2\,\gamma_k \left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]} \right)^{\mathrm{T}} \boldsymbol{A}_{\boldsymbol{0}}^{-1} \left[\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}} \left(\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\xi} \mid \boldsymbol{\theta}^{[t]} \right) - \boldsymbol{B}_{\boldsymbol{k-1}}^{[t+1]} \right] \Bigg\} = 0 \\
\iff & \cancel{2}\,\gamma_k \left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]} \right)^{\mathrm{T}} \boldsymbol{A}_{\boldsymbol{0}}^{-1} \left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]} \right) \\
& \qquad - \cancel{2} \left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]} \right)^{\mathrm{T}} \boldsymbol{A}_{\boldsymbol{0}}^{-1} \left[\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}} \left(\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\xi} \mid \boldsymbol{\theta}^{[t]} \right) - \boldsymbol{B}_{\boldsymbol{k-1}}^{[t+1]} \right] = 0
\end{aligned}$$

This gives the following update:

$$\begin{aligned}
\gamma_k^{[t+1]} = & \left[\left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]} \right)^{\mathrm{T}} \boldsymbol{A}_{\boldsymbol{0}}^{-1} \left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]} \right) \right]^{-1} \\
& \times \left(\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{k}}^{[t+1]} \right)^{\mathrm{T}} \boldsymbol{A}_{\boldsymbol{0}}^{-1} \left[\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{y}} \left(\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\xi} \mid \boldsymbol{\theta}^{[t]} \right) - \boldsymbol{B}_{\boldsymbol{k-1}}^{[t+1]} \right]
\end{aligned}$$

6.10.2.2 Updating the other parameters

Updates of the other parameters are obtained using the same formulas as for ridge, with

$$\boldsymbol{\beta}^{[t]} = \sum_{k=1}^{K} \boldsymbol{u}_{\boldsymbol{k}}^{[t]} \gamma_{\boldsymbol{k}}^{[t]}.$$

VII

Ongoing work and perspectives

The world is full of collinearities
and non–linearities.
— Rolf Harald Baayen

Contents

In three years of PhD research, many questions still remain open and many perspectives are emerging. In conclusion of this manuscript, we propose to briefly detail some of the ongoing work and perspectives.

7.1 Mixed–SCGLR for high dimensional data

7.1.1 Key idea

High dimensional data entail that $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}$ is not invertible, causing a problem with the norm–constraint associated with the VPI measure of Structural Relevance. A first obvious solution is to replace $\boldsymbol{X}$ with the matrix $\boldsymbol{C}$ of its principal components associated with non–zero, or even non–negligible eigenvalues. Another possible solution is to replace the constraint with

$$\boldsymbol{u}^{\mathrm{T}}\left[\tau\boldsymbol{I} + (1-\tau)\,\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}\right]\boldsymbol{u} = 1,$$

where $\tau \in (0, 1]$. But for now, we are going to focus on the first solution. Let λ_j be the eigenvalue associated with the j–th eigenvector $\boldsymbol{v_j}$. The last eigenvector we consider, $\boldsymbol{v_r}$, is such that

$$\frac{\lambda_r}{\sum_{j=1}^{r}\lambda_j} > \frac{1}{p},$$

where p is the number of columns of matrix $\boldsymbol{X}$. The matrix of the corresponding unit-eigenvectors is denoted $\boldsymbol{V} = \left[\boldsymbol{v_1} \mid \ldots \mid \boldsymbol{v_r}\right]$, and $\boldsymbol{C} = \boldsymbol{X}\boldsymbol{V}$. The component $\boldsymbol{f}$ is then sought as a combination of the principal components: $\boldsymbol{f} = \boldsymbol{C}\boldsymbol{u} = \boldsymbol{X}\widetilde{\boldsymbol{u}}$, where $\widetilde{\boldsymbol{u}} = \boldsymbol{V}\boldsymbol{u}$. Mixed–SCGLR then solves

$$\begin{cases} \max & s\log\left[\phi\left(\boldsymbol{u}\right)\right] + (1-s)\log\left[\psi_{\boldsymbol{A}}\left(\boldsymbol{u}\right)\right] \\ \text{subject to} & \boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{C}\boldsymbol{u} = 1, \end{cases}$$

where the goodness–of–fit measure, $\psi_{\boldsymbol{A}}$, is given at each linearisation step by

$$\begin{aligned} \psi_{\boldsymbol{A}}\left(\boldsymbol{u}\right) &= \sum_{k=1}^{q}\left\|\boldsymbol{z}_{\boldsymbol{k}}^{\boldsymbol{\xi}}\right\|_{\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}}}^{2}\cos^{2}_{\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}}}\left(\boldsymbol{z}_{\boldsymbol{k}}^{\boldsymbol{\xi}},\,\mathrm{span}\left\{\boldsymbol{C}\boldsymbol{u},\boldsymbol{A}\right\}\right) \\ &= \sum_{k=1}^{q}\left\|\boldsymbol{z}_{\boldsymbol{k}}^{\boldsymbol{\xi}}\right\|_{\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}}}^{2}\cos^{2}_{\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}}}\left(\boldsymbol{z}_{\boldsymbol{k}}^{\boldsymbol{\xi}},\,\Pi^{\boldsymbol{W}_{\boldsymbol{k}}^{\boldsymbol{\xi}}}_{\mathrm{span}\{\boldsymbol{C}\boldsymbol{u},\boldsymbol{A}\}}\boldsymbol{z}_{\boldsymbol{k}}^{\boldsymbol{\xi}}\right), \end{aligned}$$

and the structural relevance by

$$\phi\left(\boldsymbol{u}\right) = \left[\sum_{j=1}^{p}\omega_j\Big(\left\langle\,\boldsymbol{C}\boldsymbol{u}\,\middle|\,\boldsymbol{x_j}\,\right\rangle_{\boldsymbol{P}}^{2}\Big)^{l}\right]^{\frac{1}{l}} = \left[\sum_{j=1}^{p}\omega_j\Big(\boldsymbol{u}^{\mathrm{T}}\,\boldsymbol{C}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{x_j}\boldsymbol{x_j}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{C}\,\boldsymbol{u}\Big)^{l}\right]^{\frac{1}{l}}.$$

This idea is tested on simulated data where the number of explanatory variables p exceeds the number of observations n.

7.1.2 Data generation

To generate grouped data here, we consider $N = 10$ groups, and $R = 10$ observations per group (i.e. a total of $n = 100$ observations). The random–effect design matrix is then $\boldsymbol{U} = \boldsymbol{I}_N \otimes \mathbf{1}_R$. Explanatory variables consist of four independent bundles $\boldsymbol{X_j}, j \in \{0, 1, 2, 3\}$, such as $\boldsymbol{X} = \left[\, \boldsymbol{X_0} \mid \boldsymbol{X_1} \mid \boldsymbol{X_2} \mid \boldsymbol{X_3} \,\right]$, each explanatory variable being normally simulated with mean 0 and variance 1. Parameter $\tau \in \{0.3, 0.5, 0.7\}$ tunes the level of redundancy within each bundle: the correlation matrix of bundle $\boldsymbol{X_j}$ is

$$\text{cor}\,(\boldsymbol{X_j}) = \tau \mathbf{1}_{p_j} \mathbf{1}_{p_j}^{\mathrm{T}} + (1 - \tau)\boldsymbol{I}_{p_j},$$

where p_j is the number of variables in $\boldsymbol{X_j}$. For each $k \in \{1, 2, 3, 4\}$, random–effect vectors are simulated as $\boldsymbol{\xi_k} \overset{\text{ind.}}{\sim} \mathcal{N}_N\,(\mathbf{0}, \sigma_k^2\, \boldsymbol{I}_N)$.

Given $\boldsymbol{\xi_1}$, $\boldsymbol{\xi_2}$, $\boldsymbol{\xi_3}$, $\boldsymbol{\xi_4}$, we simulate 4 response–vectors, $\boldsymbol{Y} = \left[\, \boldsymbol{y_1} \mid \boldsymbol{y_2} \mid \boldsymbol{y_3} \mid \boldsymbol{y_4} \,\right]$, having different distributions, as

$$\begin{cases} \boldsymbol{y_1} \sim \mathcal{N}_n\Big(\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta_1} + \boldsymbol{U}\boldsymbol{\xi_1},\, \boldsymbol{\Sigma} = \boldsymbol{I}_n\Big) \\ \boldsymbol{y_2} \sim \mathcal{B}\Big(\boldsymbol{p} = \text{logit}^{-1}\Big[\boldsymbol{X}\boldsymbol{\beta_2} + \boldsymbol{U}\boldsymbol{\xi_2}\Big]\Big) \\ \boldsymbol{y_3} \sim \mathcal{B}\text{in}\Big(\textbf{trials} = 30\,\mathbf{1}_n,\, \boldsymbol{p} = \text{logit}^{-1}\Big[\boldsymbol{X}\boldsymbol{\beta_3} + \boldsymbol{U}\boldsymbol{\xi_3}\Big]\Big) \\ \boldsymbol{y_4} \sim \mathcal{P}\Big(\boldsymbol{\lambda} = \exp\Big[\boldsymbol{X}\boldsymbol{\beta_4} + \boldsymbol{U}\boldsymbol{\xi_4}\Big]\Big). \end{cases} \tag{7.1}$$

The response $\boldsymbol{y_1}$ is predicted only by the $\boldsymbol{X_1}$ bundle, $\boldsymbol{y_2}$ only by the $\boldsymbol{X_2}$ bundle, $\boldsymbol{y_3}$ only by the $\boldsymbol{X_3}$ bundle, $\boldsymbol{y_4}$ by both $\boldsymbol{X_2}$ and $\boldsymbol{X_3}$, while the $\boldsymbol{X_0}$ bundle plays no explanatory role. Our choice for the fixed–effect parameters is

$$\boldsymbol{\beta_1} = (\underbrace{0, \ldots\ldots\ldots\ldots, 0}_{p_0 \text{ times}}, \underbrace{0.1, \ldots\ldots\ldots\ldots 0.1}_{p_1 \text{ times}}, \underbrace{0, \ldots\ldots\ldots\ldots, 0}_{p_2 \text{ times}}, \underbrace{0, \ldots\ldots\ldots\ldots, 0}_{p_3 \text{ times}})^{\mathrm{T}},$$

$$\boldsymbol{\beta_2} = (\underbrace{0, \ldots\ldots\ldots\ldots, 0}_{p_0 \text{ times}}, \underbrace{0, \ldots\ldots\ldots\ldots 0}_{p_1 \text{ times}}, \underbrace{0.1, \ldots\ldots\ldots, 0.1}_{p_2 \text{ times}}, \underbrace{0, \ldots\ldots\ldots\ldots, 0}_{p_3 \text{ times}})^{\mathrm{T}},$$

$$\boldsymbol{\beta_3} = (\underbrace{0, \ldots\ldots\ldots\ldots, 0}_{p_0 \text{ times}}, \underbrace{0, \ldots\ldots\ldots\ldots 0}_{p_1 \text{ times}}, \underbrace{0, \ldots\ldots\ldots\ldots, 0}_{p_2 \text{ times}}, \underbrace{0.05, \ldots.., 0.05}_{p_3 \text{ times}})^{\mathrm{T}},$$

$$\boldsymbol{\beta_4} = (\underbrace{0, \ldots\ldots\ldots\ldots, 0}_{p_0 \text{ times}}, \underbrace{0.025, \ldots\ldots\ldots 0.025}_{p_1 \text{ times}}, \underbrace{0.025, \ldots.., 0.025}_{p_2 \text{ times}}, \underbrace{0, \ldots\ldots\ldots\ldots, 0}_{p_3 \text{ times}})^{\mathrm{T}}$$

We consider in turn $p = 150$ ($p_0 = 60$, $p_1 = 45$, $p_2 = 30$, $p_3 = 15$) and $p = 200$ ($p_0 = 80$, $p_1 = 60$, $p_2 = 40$, $p_3 = 20$) explanatory variables. The variance components are set to $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 0.1$, and $\sigma_4^2 = 0.05$. For each value of p and for each value of τ, $B = 20$ samples are generated according to model (7.1).

7.1.3 Results

Table 7.1 and Table 7.2 present the results for respectively 150 and 200 explanatory variables. They give the Mean Relative Squared Error (MRSE) values for $\boldsymbol{\beta_k}, k \in \{1, \ldots, 4\}$, as well as biases and standard errors of estimated variance components, obtained on 20 samples for each value of τ. Some component planes are given on Figure 7.1 (150 explanatory variables) and Figure 7.2 (200 explanatory variables).

Table 7.1 – *Mean Relative Squared Error (MRSE) values for fixed–effect estimates, and biases and standard errors for estimated variance components, in the case of $n = 100$ observations and $p = 150$ explanatory variables.* *The results are obtained on 20 samples for each value of redundancy parameter τ.*

	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_4}$	σ_1^2	σ_2^2	σ_3^2	σ_4^2
$\tau = 0.3$	0.06	0.26	0.19	0.13	−0.01 (0.09)	−0.03 (0.09)	−0.02 (0.03)	0.02 (0.06)
$\tau = 0.5$	0.03	0.20	0.10	0.07	0.01 (0.11)	−0.03 (0.08)	0.00 (0.07)	0.01 (0.07)
$\tau = 0.7$	0.01	0.10	0.05	0.04	0.01 (0.07)	−0.05 (0.09)	0.01 (0.10)	0.02 (0.07)

Table 7.2 – *Mean Relative Squared Error (MRSE) values for fixed–effect estimates, and biases and standard errors for estimated variance components, in the case of $n = 100$ observations and $p = 200$ explanatory variables.* *The results are obtained on 20 samples for each value of redundancy parameter τ.*

	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_4}$	σ_1^2	σ_2^2	σ_3^2	σ_4^2
$\tau = 0.3$	0.06	0.15	0.18	0.10	−0.04 (0.04)	−0.05 (0.09)	0.01 (0.05)	−0.02 (0.05)
$\tau = 0.5$	0.03	0.17	0.09	0.05	−0.05 (0.06)	0.00 (0.19)	−0.02 (0.04)	−0.01 (0.04)
$\tau = 0.7$	0.01	0.15	0.04	0.03	0.03 (0.08)	0.00 (0.14)	−0.01 (0.05)	−0.02 (0.05)

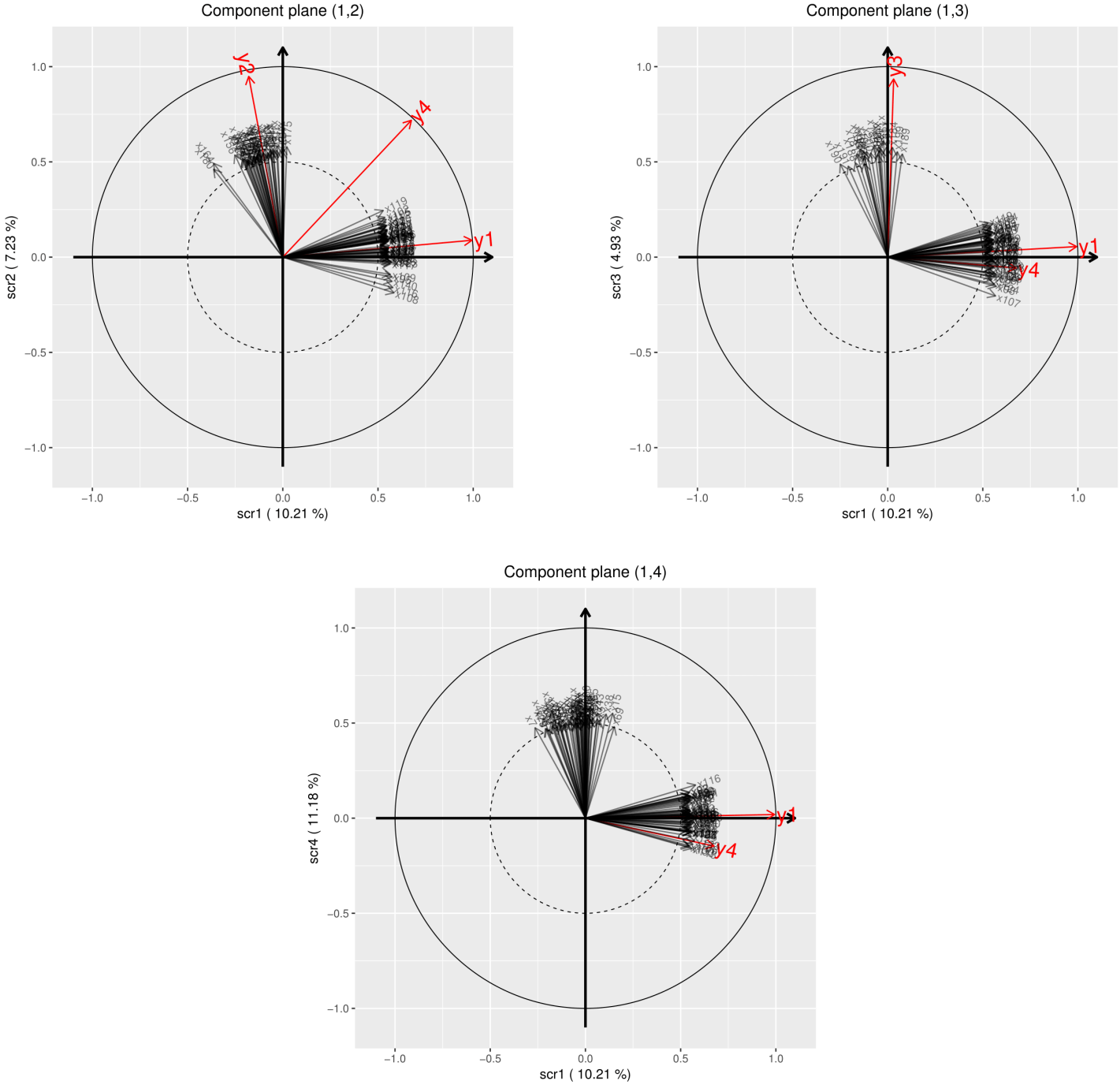


Figure 7.1 – *Component planes* $(1, 2)$*,* $(1, 3)$ *and* $(1, 4)$ *given by mixed–SCGLR for* $n = 100$ *observations and* $p = 150$ *explanatory variables. The within–bundle correlation is $\tau = 0.3$. The tuning parameter triplet selected through cross–validation is $(K^\star, s^\star, l^\star) = (3, 0.5, 4)$. Even if $K^\star = 3$, component plane $(1, 4)$ has been edited to emphasise that the fourth component aligns with the nuisance bundle. For an easier model interpretation, we hide all the explanatory variables whose cosine with the component plane is lower than 0.5.

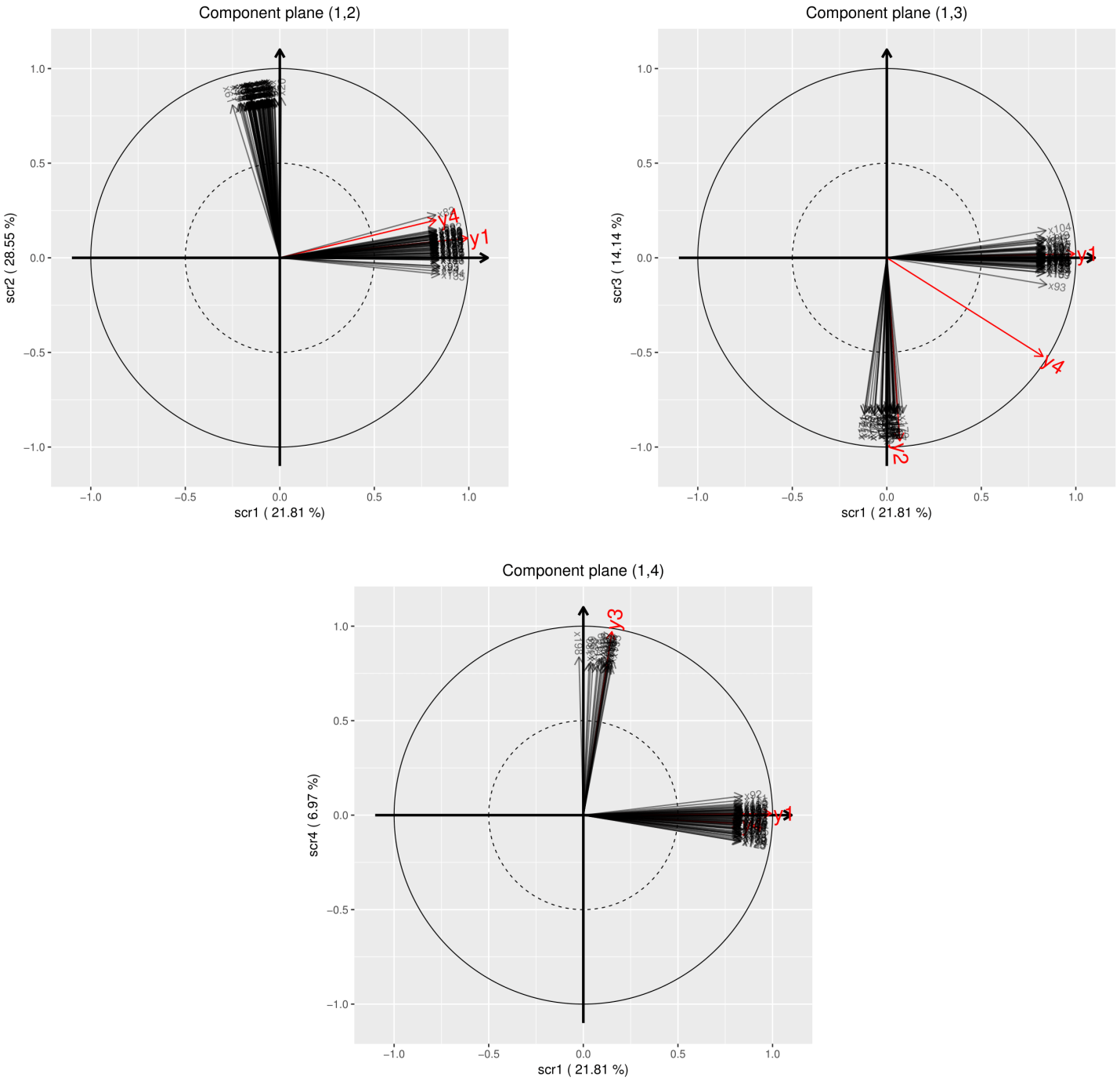


Figure 7.2 – *Component planes* $(1, 2)$*,* $(1, 3)$ *and* $(1, 4)$ *given by mixed–SCGLR for* $n = 100$ *observations and* $p = 200$ *explanatory variables. The within–bundle correlation is $\tau = 0.7$. With such a high level of redundancy, the optimal trade–off parameter selected through cross–validation is $s^\star = 0.9$. The nuisance bundle is then captured by the second component, and the optimal number of component selected is $K^\star = 4$.

7.2 Application in Psychiatry: Link between major depressive disorders and persistent grey–matter volume reduction

As suggested at the beginning of Chapter 6, the methods developed in this thesis have applications in epidemiology and psychiatric sciences. The purpose of this section is to outline one of them.

The paper by Carrière et al. (2017) addresses the analysis of depressive symptoms and their changes over time. The authors highlight different trajectories of depressive symptoms over 10 years in community–dwelling–elderly men and women. They characterise the current and life–time risk factors (demographic characteristics, level of education, mode of living, etc) associated with these trajectories, while taking into account the individual correlation between repeated measures (Proust-Lima et al., 2017, R package **lcmm**, latent class mixed models). More recently, Ancelin et al. (2019) focus on major depressive disorders (MDD). But instead of studying the clinical characteristics of MDD and the associated risk factors, they rather analyse whether lifetime episodes of MDD are associated with specific alterations in the grey–matter volume. The data set they use is derived from a longitudinal study of neuropsychiatric disorders in community-dwelling French elderly adults, called "Enquête de Santé Psychologique — Risques, Incidence et Traitemement" (*ESPRIT*, Ritchie et al., 2004). The data set also includes brain–volume measurements on 636 participants, obtained by MRI protocol and image post–processing (FreeSurfer image analysis suite, `http://surfer.nmr.mgh.harvard.edu/`).

We believe that the mixed–SCGLR method could help to understand the links between the occurrence of MDD and alterations in certain brain areas.

(i) The extended data set *ESPRIT* contains 528 potentially redundant explanatory variables that concatenate, for each participant, measurements of the thickness, area, volume and curvature of a number of brain areas. Many explanatory variables are highly correlated (see the correlation heatmap on Figure 7.3) and we believe that a model regularisation based on the construction of supervised components could be useful.

(ii) The response variable is binary (depressive or not), which can lead to the use of a GLM.

(iii) Since the study involves repeated measurements, the intra–subject correlations must be taken into account, hence the use of a GLMM.

(iv) Finally, a small number of variables does not have to be regularised: gender, age, total brain volume and consumption of antidepressants. We can meet this requirement by considering these variables as additional explanatory variables.

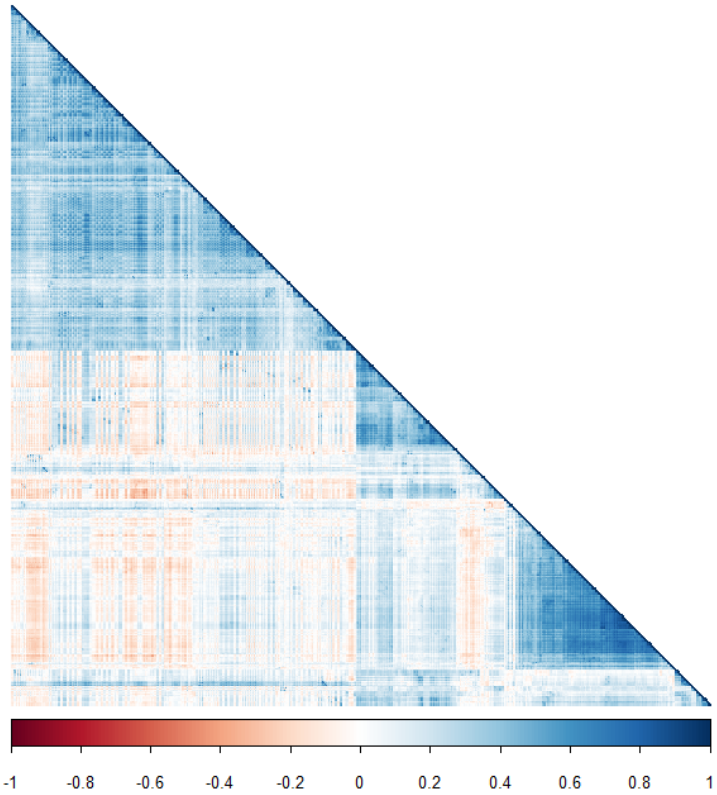


Figure 7.3 – *Correlation heatmap of the explanatory variables of the ESPRIT data set.* *The blue color corresponds to a correlation close to* 1*, the red color corresponds to a correlation close to* -1 *and the white color corresponds to a correlation close to* 0*.*

In addition, these studies often require the implementation of hypothesis testing and/or the computation of confidence intervals to determine which explanatory variables have the greatest impact on the phenomenon. The following section presents the basic ideas of bootstrap–based confidence intervals for GLMMs.

7.3 Bootstrap–based confidence intervals

Let $\widetilde{\boldsymbol{\beta}}$ be the estimator of $\boldsymbol{\beta}$ obtained from the original sample $\mathcal{S}$, say of size n. One way to obtain a confidence interval for each $\widetilde{\beta}_j$ is to use the bootstrap approach, which consists of randomly draw B datasets with replacement

from $\mathcal{S}$, each dataset having size n. Then, by refitting the model on each bootstrap sample, we can get an idea of the behaviour of the estimators over the B replications, and build confidence intervals.

Gaussian approximation method. Let $\widehat{\boldsymbol{\beta}}^{(b)}$ be the estimate of $\boldsymbol{\beta}$ associated with the b^{th} bootstrap sample. The expectation of $\widetilde{\boldsymbol{\beta}}$ is then estimated by

$$\widehat{\mathbb{E}}\left(\widetilde{\boldsymbol{\beta}}\right) = \overline{\boldsymbol{\beta}} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\boldsymbol{\beta}}^{(b)}.$$

The bias of $\widetilde{\boldsymbol{\beta}}$ is estimated by $\widehat{\text{bias}}\left(\widetilde{\boldsymbol{\beta}}\right) = \overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$, so that the bias–corrected estimator of $\widetilde{\boldsymbol{\beta}}$ is given by

$$\widehat{\boldsymbol{\beta}}_{\text{BC}} = \widetilde{\boldsymbol{\beta}} - \widehat{\text{bias}}\left(\widetilde{\boldsymbol{\beta}}\right) = 2\widetilde{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}.$$

In the same vein, the variance of $\widetilde{\beta}_j$ is estimated by

$$\widehat{\text{Var}}\left(\widetilde{\beta}_j\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\beta}_j^{(b)} - \overline{\beta}_j\right)^2.$$

If the asymptotic normality of $\widetilde{\boldsymbol{\beta}}$ is assumed, an approximate bias–corrected $(1-2\alpha)$–confidence interval for each β_j is given by

$$\begin{aligned}\text{CI}_{1-2\alpha}\left(\beta_j\right) &= \left[\widetilde{\beta}_j - \widehat{\text{bias}}\left(\widetilde{\beta}_j\right) \pm z_{1-\alpha}\sqrt{\widehat{\text{Var}}\left(\widetilde{\beta}_j\right)}\right] \\ &= \left[2\widetilde{\beta}_j - \overline{\beta}_j \pm z_{1-\alpha}\sqrt{\widehat{\text{Var}}\left(\widetilde{\beta}_j\right)}\right],\end{aligned}$$

where $z_{1-\alpha}$ denotes the $(1-\alpha)$–th percentile of the standard Gaussian distribution.

The bootstrap percentile. An alternative approach, called the bootstrap percentile interval, is to use the empirical quantiles of bootstrap replicates $\left(\beta_j^{(1)}, \ldots, \beta_j^{(B)}\right)$ to form a confidence interval for β_j. Let $\widehat{q}_{\alpha,j}$ (respectively $\widehat{q}_{1-\alpha,j}$) be the bootstrap–based estimate of the α–th (respectively the $(1-\alpha)$–th) percentile of $\widetilde{\beta}_j$. The following equivalences hold.

$$\begin{aligned}&P\left(\widetilde{\beta}_j \in [\widehat{q}_{\alpha,j}\,,\,\widehat{q}_{1-\alpha,j}]\right) = 1-2\alpha \\ \Leftrightarrow\; &P\left(\widetilde{\beta}_j \in \left[\mathbb{E}(\widetilde{\beta}_j) + \widehat{q}_{\alpha,j} - \mathbb{E}(\widetilde{\beta}_j)\,,\,\mathbb{E}(\widetilde{\beta}_j) + \widehat{q}_{1-\alpha,j} - \mathbb{E}(\widetilde{\beta}_j)\right]\right) = 1-2\alpha \\ \Leftrightarrow\; &P\left(\widetilde{\beta}_j \in \left[\beta_j + \text{bias}(\widetilde{\beta}_j) + \widehat{q}_{\alpha,j} - \mathbb{E}(\widetilde{\beta}_j)\,,\,\beta_j + \text{bias}(\widetilde{\beta}_j) + \widehat{q}_{1-\alpha,j} - \mathbb{E}(\widetilde{\beta}_j)\right]\right) = 1-2\alpha \\ \Leftrightarrow\; &P\left(\beta_j \in \left[\widetilde{\beta}_j + \mathbb{E}(\widetilde{\beta}_j) - \text{bias}(\widetilde{\beta}_j) - \widehat{q}_{1-\alpha,j}\,,\,\widetilde{\beta}_j + \mathbb{E}(\widetilde{\beta}_j) - \text{bias}(\widetilde{\beta}_j) - \widehat{q}_{\alpha,j}\right]\right) = 1-2\alpha\end{aligned}$$

Now, as $\mathbb{E}(\widetilde{\beta}_j) - \text{bias}(\widetilde{\beta}_j)$ is estimated by $\widetilde{\beta}_j$, another approximate $(1-2\alpha)$–confidence interval for β_j is given by

$$\text{CI}_{1-2\alpha}\left(\beta_j\right) = \left[2\widetilde{\beta}_j - \widehat{q}_{1-\alpha,j}\,,\, 2\widetilde{\beta}_j - \widehat{q}_{\alpha,j}\right].$$

Caution on resampling for GLMMs. A particular attention must be paid to the way in which bootstrap resampling is carried out for GLMMs. Indeed, all the bootstrap samples have to preserve the grouping structure of the original sample. Although the method allows statistical inferences to be made, its main problem is its slowness, as we have to refit the model once for every bootstrap sample. This is why a large number of variants, improvements (such as the accelerated bootstrap for instance) have been proposed (Efron and Tibshirani, 1994).

7.4 Mixed–SCGLR for spatial correlation modelling

In this thesis, individual–specific and autocorrelated time–specific random effects were considered. Now, it is often necessary to develop models with spatial correlations, which commonly occur in ecology. In particular, we could have considered it for the *Genus* data set (Section 5.6), but we did not do so since our main interest was to investigate the explanatory structure of the GLMM fixed design and relate it to interpretable dimensions. As suggested in the nice review by Sun et al. (2000), many approaches are available for modelling spatial correlations, one of which is to consider conditional autoregressive (CAR) random effects (Besag, 1974).

For instance, Rousset (2017), who developed the R package **spaMM** for "spatial Mixed Models", models spatial correlations with a particular order–1 CAR random effect whose distribution writes

$$\boldsymbol{\xi} \sim \mathcal{N}_q\left(\mathbf{0}, \varsigma^2 \boldsymbol{B}^{-1}\right). \tag{7.2}$$

In (7.2),

- ς^2 is the unknown spatial–specific variance component, and
- $\boldsymbol{B} = \boldsymbol{I} - \rho\boldsymbol{A}$, where ρ is the unknown autoregressive spatial parameter and $\boldsymbol{A} = (a_{ij})_{1\leqslant i,j\leqslant q}$ is the adjacency matrix defined by

$$a_{ij} = \begin{cases} 1 & \text{if } j \text{ is adjacent to } i \\ 0 & \text{otherwise.} \end{cases}$$

Interestingly, the R package **spaMM** is not limited to LMMs as it can also fits GLMMs with an order–1 CAR random effect. The estimation method is based on different variants of the Laplace approximation, including the PQL of Breslow and Clayton (1993) (Section 4.3.3) as also discussed by Lee and Nelder (1996), which is very close to the Schall's method (see Section 4.6). **spaMM** is therefore very inspiring to extend the mixed–SCGLR method for spatial correlation modelling.

7.5 Mixed–THEME–SCGLR

Very recently, Bry et al. (2018) have extended SCGLR to several thematic blocks $\boldsymbol{X_1}, \ldots, \boldsymbol{X_R}$ of explanatory variables, which led to the "THEME–SCGLR" method. The search for the rank–1 components $\boldsymbol{f_1^1} = \boldsymbol{X_1 u_1^1}, \ldots, \boldsymbol{f_R^1} = \boldsymbol{X_R u_R^1}$, i.e. the rank–1 component of each theme, is carried out through the program

$$\left| \begin{array}{ll} \max & \left[\psi_{\boldsymbol{A}}\left(\boldsymbol{u_1}, \ldots, \boldsymbol{u_R}\right)\right]^g \prod_{r=1}^{R} \left[\phi\left(\boldsymbol{u_r}\right)\right]^{s_r} \\ \text{subject to} & \boldsymbol{u_r}^{\intercal} \boldsymbol{M_r}^{-1} \boldsymbol{u_r} = 1, \text{ for each } r \in \{1, \ldots, R\}. \end{array} \right. \tag{7.3}$$

Concerning the tuning parameters g and s_r, a common choice is to take: $\forall r = 1, \ldots, R$, $s_r = s$ and $g = 1 - s$. Moreover, in (7.3), ϕ is one of the Structural Relevance (SR) measures introduced in Appendix 5.8.1, and $\psi_{\boldsymbol{A}}$ is a Goodness–of–Fit (GoF) measure depending on additional explanatory variables $\boldsymbol{A}$. More precisely, $\boldsymbol{z_k}$ and $\boldsymbol{W_k}$ denoting respectively the k^{th} working variable and the k^{th} weight matrix derived from the IRLS algorithm, the GoF measure writes

$$\begin{aligned} \psi_{\boldsymbol{A}}\left(\boldsymbol{u_1}, \ldots, \boldsymbol{u_R}\right) &= \sum_{k=1}^{q} \|\boldsymbol{z_k}\|^2_{\boldsymbol{W_k}} \cos^2_{\boldsymbol{W_k}} \left(\boldsymbol{z_k}, \text{ span}\left\{\boldsymbol{X_1 u_1}, \ldots, \boldsymbol{X_R u_R}, \boldsymbol{A}\right\}\right) \\ &= \sum_{k=1}^{q} \|\boldsymbol{z_k}\|^2_{\boldsymbol{W_k}} \cos^2_{\boldsymbol{W_k}} \left(\boldsymbol{z_k}, \text{ span}\left\{\boldsymbol{X_r u_r}, \widetilde{\boldsymbol{A}}_{\boldsymbol{r}}\right\}\right) \\ &=: \psi_{\widetilde{\boldsymbol{A}}_{\boldsymbol{r}}}\left(\boldsymbol{u_r}\right), \end{aligned} \tag{7.4}$$

where $\widetilde{\boldsymbol{A}}_{\boldsymbol{r}} = \boldsymbol{A} \cup \{\boldsymbol{X_j u_j} \mid j \neq r\}$. As (7.4) suggests that the GoF measure can be seen as a function of a particular $\boldsymbol{u_r}$, Bry et al. (2018) propose to solve (7.3) by iteratively solving

$$\left| \begin{array}{ll} \max & \left[\psi_{\widetilde{\boldsymbol{A}}_{\boldsymbol{r}}}\left(\boldsymbol{u_r}\right)\right]^{1-s} \left[\phi\left(\boldsymbol{u_r}\right)\right]^{s} \\ \text{subject to} & \boldsymbol{u_r}^{\intercal} \boldsymbol{M_r}^{-1} \boldsymbol{u_r} = 1. \end{array} \right.$$

The higher rank components are obtained using the same type of program, but considering appropriate extra orthogonality constraints and playing on the additional explanatory variables.

The extension of mixed–SCGLR to several blocks of thematic variables, "mixed–THEME–SCGLR", appears rather straightforward: it involves iterating mixed–SCGLR on each $\boldsymbol{X}_r$ by considering the components of all the other blocks as additional explanatory variables. By contrast, compared to mixed–SCGLR, the explanatory power of mixed–THEME–SCGLR will be much greater.

7.6 Sparse SCGLR

In all the simulations in Chapters 5 and 6, we considered structured or unstructured nuisance explanatory variables (i.e. without any explanatory role). The true coefficients associated with these variables are therefore exactly zero. However, in the methods we have developed, the explanatory variables that are not relevant to explain the responses only have near zero weights in the supervised components. In the wake of Lê Cao et al. (2008) and Chun and Keleş (2010), an interesting perspective would be to introduce an L_1–constraint to our SCGLR–specific criterion (5.7). The h^{th} component, namely $\boldsymbol{f}_h = \boldsymbol{X}\boldsymbol{u}_h$, would then be obtained by solving

$$\begin{cases} \max & s \log\left[\phi\left(\boldsymbol{u}\right)\right] + (1-s) \log\left[\psi_{\boldsymbol{A}_{h-1}}\left(\boldsymbol{u}\right)\right] - \lambda \left\|\boldsymbol{u}\right\|_1 \\ \text{subject to} & \left\|\boldsymbol{u}\right\|^2_{\boldsymbol{M}^{-1}} = 1 \text{ and } \boldsymbol{X}\boldsymbol{u} \perp \boldsymbol{f}_1, \ldots, \boldsymbol{f}_{h-1}. \end{cases} \tag{7.5}$$

Note that Chun and Keleş (2010) also proposed an alternative to simply adding an L_1–constraint, by generalising the sparse PCA of Zou et al. (2006). Then Chung and Keles (2010) have developed an extension of this technique by integrating it into the GLM framework, in particular to solve classification problems. We believe that substantial modifications of the PING algorithm are necessary to solve the maximisations given by (7.5). The approaches and algorithms developed by the above–mentioned authors could be very inspiring.

7.7 “The world is full of collinearities and non–linearities”

This thesis was devoted to the construction of so–called “supervised” components for GLMs and GLMMs. A linear relationship between the predictor, $\boldsymbol{\eta}$, and the supervised components, $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_K$, has therefore always been assumed. However, this linearity assumption may be restrictive in some situations. To overcome this hypothesis, a large number of PLSR extensions have been developed (see for instance the overview by Rosipal, 2011). These methods can be divided into two types of Nonlinear PLS methods. In the Type I method, the explanatory variables are appended with nonlinear transformations. This results in the mapping of the initial explanatory variables in a higher dimensional space, and in the implementation of the classical PLSR on these new variables. By contrast, Type II method assumes a nonlinear relationship within the latent variable structure of the model (Wold et al., 1989). An interesting perspective could be to draw inspiration from this abundant literature to develop methods for the construction of supervised components adapted to highly non–linear contexts.

Bibliography

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, volume 55. Courier Corporation. 79

Adragni, K. P. and Cook, R. D. (2009). Sufficient Dimension Reduction and Prediction in Regression. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405. 127

Ancelin, M.-L., Carrière, I., Artero, S., Maller, J., Meslin, C., Ritchie, K., Ryan, J., and Chaudieu, I. (2019). Lifetime major depression and grey-matter volume. *Journal of Psychiatry & Neuroscience: JPN*, 44(1):45. 195

Anderson, D. A. and Aitkin, M. (1985). Variance Component Models with Binary Response: Interviewer Variability. *Journal of the Royal Statistical Society, Series B (Methodological)*, 47(2):203–210. 34, 46, 80

Andrews, D. W. (1991). Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics*, 47(2-3):359–377. 166

Bastien, P., Vinzi, V. E., and Tenenhaus, M. (2005). PLS generalised linear regression. *Computational Statistics & Data Analysis*, 48(1):17–46. 69

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48. 82, 112

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 192–236. 198

Blocker, A. W., Blocker, M. A. W., Rcpp, D., and Rcpp, L. (2014). Package 'fastghquad'. 79

Bouveyron, C., Latouche, P., Mattei, P.-A., et al. (2018). Bayesian Variable Selection for Globally Sparse Probabilistic PCA. *Electronic Journal of Statistics*, 12(2):3036–3070. 72

Breslow, N. E. and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88(421):9–25. 34, 46, 83, 85, 95, 129, 199

Breslow, N. E. and Lin, X. (1995). Bias Correction in Generalised Linear Mixed Models with a Single Component of Dispersion. *Biometrika*, 82(1):81–91. 85

Bry, X. (1994). *Analyses factorielles simples*. Economica Poche. 132

Bry, X., Redont, P., Verron, T., and Cazes, P. (2012). THEME-SEER: a multidimensional exploratory technique to analyze a structural model using an extended covariance criterion. *Journal of Chemometrics*, 26(5):158–169. 130

Bry, X., Trottier, C., Mortier, F., and Cornu, G. (2018). Component-based regularization of a multivariate GLM with a thematic partitioning of the explanatory variables. *Statistical Modelling (in press)*. 30, 42, 71, 103, 134, 141, 164, 199

Bry, X., Trottier, C., Mortier, F., Cornu, G., and Verron, T. (2014). Extending SCGLR to multiple regressor-groups: The THEME-SCGLR method. In *Proceedings of the eighth International Conference on Partial Least Squares and Related Methods*, Paris, France. 30, 42, 71

Bry, X., Trottier, C., Mortier, F., Cornu, G., and Verron, T. (2016). *The Multiple Facets of Partial Least Squares and Related Methods*, chapter Supervised Component Generalized Linear Regression with Multiple Explanatory Blocks: THEME-SCGLR, pages 141–154. Springer International Publishing. 30, 42, 71, 103, 143

Bry, X., Trottier, C., Verron, T., and Mortier, F. (2013). Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, 119(4):47–60. 30, 33, 42, 45, 70, 71, 72, 102, 111

Bry, X. and Verron, T. (2015). THEME: THEmatic Model Exploration through multiple co-structure maximization. *Journal of Chemometrics*, 29(12):637–647. 129, 130

Carrière, I., Farré, A., Proust-Lima, C., Ryan, J., Ancelin, M.-L., and Ritchie, K. (2017). Chronic and remitting trajectories of depressive symptoms in the elderly. Characterisation and risk factors. *Epidemiology and Psychiatric Sciences*, 26(2):146–156. 195

Celeux, G. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2:73–82. 153

Celeux, G., Chauveau, D., and Diebolt, J. (1995). On Stochastic Versions of the EM Algorithm. Research Report RR-2514, INRIA. 153

Chavent, M., Kuentz, V., Labenne, A., Liquet, B., and Saracco, J. (2017). Package 'PCAmixdata'. 132

Chavent, M., Kuentz-Simonet, V., Labenne, A., and Saracco, J. (2014). Multivariate analysis of mixed data: The PCAmixdata R package. *arXiv preprint arXiv:1411.4911*. 132

Chrétien, S. and Hero, A. O. (2000). Kullback Proximal Algorithms for Maximum-Likelihood Estimation. *IEEE Transactions on Information Theory*, 46(5):1800–1810. 151, 152

Chun, H. and Keleş, S. (2010). Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25. 31, 43, 72, 200

Chung, D. and Keles, S. (2010). Sparse Partial Least Squares Classification for High Dimensional Data. *Statistical Applications in Genetics and Molecular Biology*, 9(1). 200

Clayton, D. G. (1996). Generalized Linear Mixed Models. In *Markov chain Monte Carlo in practice*, pages 275–301. Springer. 34, 46, 88

Cook, R. D. et al. (2007). Fisher Lecture: Dimension Reduction in Regression.

Statistical Science, 22(1):1–26. 127

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 1–38. 85, 149, 151, 155

Dey, D. K., Ghosh, S. K., and Mallick, B. K. (2000). *Generalized Linear Models: A Bayesian Perspective*. CRC Press. 60

Durif, G., Modolo, L., Michaelsson, J., Mold, J. E., Lambert-Lacroix, S., and Picard, F. (2017). High Dimensional Classification with combined Adaptive Sparse PLS and Logistic Regression. *Bioinformatics*, 34(3):485–493. 31, 42, 72

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least Angle Regression. *The Annals of Statistics*, 32(2):407–499. 64

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press. 198

Eilers, P. H., Boer, J. M., van Ommen, G.-J., and van Houwelingen, H. C. (2001). Classification of Microarray Data with Penalized Logistic Regression. In *Microarrays: Optical Technologies and Informatics*, volume 4266, pages 187–199. International Society for Optics and Photonics. 31, 42, 71

Eliot, M., Ferguson, J., Reilly, M. P., and Foulkes, A. S. (2011). Ridge Regression for Longitudinal Biomarker Data. *The International Journal of Biostatistics*, 7(1):1–11. 35, 37, 39, 46, 48, 50, 98, 109, 111, 146, 161, 166

Engel, B. and Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, 48(1):1–22. 93

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Series in Statistics. Springer-Verlag. 54, 101

Fan, Y., Leslie, D. S., Wand, M., et al. (2008). Generalised linear mixed model analysis via sequential Monte Carlo sampling. *Electronic Journal of Statistics*, 2:916–938. 128

Fort, G. and Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. *Bioinformatics*, 21(7):1104–1111.

31, 42, 71, 72

Frank, L. E. and Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, 35(2):109–135. 72

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332. 64, 65

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*, volume 1. Springer series in statistics New York. 63

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1. 30, 42, 66

Gelman, A. et al. (2005). Analysis of variance—why it is more important than ever. *The Annals of Statistics*, 33(1):1–53. 75

Godambe, V. and Thompson, M. E. (1989). An extension of quasi-likelihood estimation. *Journal of Statistical Planning and Inference*, 22(2):137–152. 60

Goldstein, H. and Rasbash, J. (1996). Improved Approximations for Multilevel Models with Binary Responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 505–513. 85

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, 21(2):215–223. 161

Golub, G. H. and van der Vorst, H. A. (2000). Eigenvalue Computation in the 20th Century. *Journal of Computational and Applied Mathematics*, 123(1):35 – 65. Numerical Analysis 2000. Vol. III: Linear Algebra. 138

Golub, G. H. and Welsch, J. H. (1969). Calculation of Gauss Quadrature Rules. *Mathematics of Computation*, 23(106):221–230. 79

Green, P. J. (1984). Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 149–192. 58

Green, P. J. (1987). Penalized Likelihood for General Semi-Parametric Regression Models. *International Statistical Review*, pages 245–259. 84

Greenacre, M. and Blasius, J. (2006). *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall/CRC. 132

Groll, A. (2017). *glmmLasso: Variable Selection for Generalized Linear Mixed Models by L_1-Penalized Estimation*. R package version 1.5.1. 25, 109, 118

Groll, A. and Tutz, G. (2014). Variable selection for generalized linear mixed models by L_1-penalized estimation. *Statistics and Computing*, 24(2):137–154. 35, 46, 98, 109

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: the MCMCglmm R Package. *Journal of Statistical Software*, 33(2):1–22. 88, 128

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109. 86

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics*, 31(2):423–447. 92, 107

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, 12(1):69–82. 30, 41, 62

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67. 30, 41, 62

Householder, A. S. (2013). *The Theory of Matrices in Numerical Analysis*. Courier Corporation. 138

Jamshidian, M. and Jennrich, R. I. (1993). Conjugate Gradient Acceleration of the EM Algorithm. *Journal of the American Statistical Association*, 88(421):221–228. 153

Jolliffe, I. T. (1982). A Note on the Use of Principal Components in Regression. *Applied Statistics*, pages 300–303. 30, 42, 67

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analy-

sis. *Psychometrika*, 23(3):187–200. 133

Karlsson, S. and Skoglund, J. (2004). Maximum-likelihood based inference in the two-way random effects model with serially correlated time effects. *Empirical Economics*, 29(1):79–88. 37, 48, 146

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451. 131

Knudson, C. (2015). glmm: Generalized Linear Mixed Models via Monte Carlo Likelihood Approximation. *R package version*, 1(2). 91

Knudson, C. (2016). *Monte Carlo Likelihood Approximation for Generalized Linear Mixed Models*. PhD thesis, University of Minnesota. 34, 46, 88, 90, 128

Kreft, I. G. and De Leeuw, J. (1998). *Introducing Multilevel Modeling*. Sage. 75

Kuk, A. Y. (1995). Asymptotically Unbiased Estimation in Generalized Linear Models with Random Effects. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 395–407. 85

Lavergne, C. and Trottier, C. (2000). Sur l'estimation dans les modèles linéaires généralisés à effets aléatoires. *Revue de Statistique Appliquée*, 48(1):49–67. 93

Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., and Besse, P. (2008). A Sparse PLS for Variable Selection when Integrating Omics Data. *Statistical Applications in Genetics and Molecular Biology*, 7(1). 200

Lebart, L., Morineau, A., and Piron, M. (1995). *Statistique exploratoire multidimensionnelle*, volume 3. Dunod Paris. 132

Lee, J. Y., Green, P. J., and Ryan, L. M. (2017). Conjugate generalized linear mixed models for clustered data. *arXiv preprint arXiv:1709.06288*. 78

Lee, Y. and Nelder, J. A. (1996). Hierarchical Generalized Linear Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 619–678. 77, 95, 199

Li, K.-I. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86(414):316–327. 127

Liu, Q. and Pierce, D. A. (1993). Heterogeneity in Mantel-Haenszel-type models. *Biometrika*, 80(3):543–556. 79

Liu, Q. and Pierce, D. A. (1994). A note on Gauss—Hermite quadrature. *Biometrika*, 81(3):624–629. 79

Martinet, B. (1970). Brève communication. Régularisation d'inéquations variationnelles par approximations successives. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 4(R3):154–158. 151

Marx, B. D. (1996). Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression. *Technometrics*, 38(4):374–381. 30, 42, 69, 70, 72

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. 54, 57, 59, 60, 91, 96, 101

McCulloch, C. E. (1994). Maximum Likelihood Variance Components Estimation for Binary Data. *Journal of the American Statistical Association*, 89(425):330–335. 86

McCulloch, C. E. (1997). Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 92(437):162–170. 34, 46, 82, 86, 87, 128, 153

McCulloch, C. E. and Searle, S. R. (2004). *Generalized, Linear, and Mixed Models*. John Wiley & Sons. 76

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278. 153

Molenberghs, G., Verbeke, G., Demétrio, C. G., Vieira, A. M., et al. (2010). A Family of Generalized Linear Models for Repeated Measures with Normal and Conjugate Random Effects. *Statistical Science*, 25(3):325–347. 78

Naylor, J. C. and Smith, A. F. (1982). Applications of a Method for the Efficient Computation of Posterior Distributions. *Applied Statistics*, pages 214–225. 79

Neal, R. M. and Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer. 153

Nelder, J. A. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, 74(2):221–232. 60

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384. 29, 41, 54, 57, 101

Pan, J. and Thompson, R. (2003). Gauss-Hermite Quadrature Approximation for Estimation in Generalised Linear Mixed Models. *Computational Statistics*, 18(1):57–78. 81

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics*, 4(1):12–35. 34, 46, 80

Pinheiro, J. C. and Chao, E. C. (2006). Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models. *Journal of Computational and Graphical Statistics*, 15(1):58–81. 80

Proust-Lima, C., Philipps, V., and Liquet, B. (2017). Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. *Journal of Statistical Software, Articles*, 78(2):1–56. 195

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 82, 109

Rabe-Hesketh, S. and Skrondal, A. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman and Hall/CRC. 81

Raudenbush, S. W., Yang, M.-L., and Yosef, M. (2000). Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation. *Journal of Computational and Graphical Statistics*, 9(1):141–157. 82

Ridley, C. E. and Ellstrand, N. C. (2010). Rapid evolution of morphology

and adaptive life history in the invasive California wild radish (Raphanus sativus) and the implications for management. *Evolutionary Applications*, 3(1):64–76. 91

Ritchie, K., Artero, S., Beluche, I., Ancelin, M.-L., Mann, A., Dupuy, A.-M., Malafosse, A., and Boulenger, J.-P. (2004). Prevalence of DSM-IV psychiatric disorder in the French elderly population. *The British Journal of Psychiatry*, 184(2):147–152. 195

Robert, C. and Casella, G. (2011). *Méthodes de Monte-Carlo avec R*. Springer Science & Business Media. 89

Robert, C. and Casella, G. (2013). *Monte Carlo Statistical Methods*. Springer Science & Business Media. 86, 89

Roche, A. (2011). EM algorithm and variants: an informal tutorial. *arXiv preprint arXiv:1105.1476*. 149

Rockafellar, R. T. (1976). Monotone Operators and the Proximal Point Algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898. 151

Rosipal, R. (2011). Nonlinear Partial Least Squares: An Overview. In *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*, pages 169–189. IGI Global. 201

Rousset, F. (2017). An introduction to the spaMM package for mixed models. 198

Rubinstein, R. Y. and Kroese, D. P. (2016). *Simulation and the Monte Carlo Method*, volume 10. John Wiley & Sons. 89

Schall, R. (1991). Estimation in Generalized Linear Models with Random Effects. *Biometrika*, 78(4):719–727. 35, 46, 92, 95, 96, 106, 107, 129, 146, 165

Schelldorfer, J., Meier, L., and Bühlmann, P. (2014). GLMMLasso: An Algorithm for High-Dimensional Generalized Linear Mixed Lodels Using ℓ_1-Penalization. *Journal of Computational and Graphical Statistics*, 23(2):460–477. 35, 46, 98, 109

Searle, S. R., Casella, G., and McCulloch, C. E. (2009). *Variance Components*,

volume 391. John Wiley & Sons. 75

Shun, Z. and McCullagh, P. (1995). Laplace Approximation of High Dimensional Integrals. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 57(4):749–760. 34, 46, 82

Stiratelli, R., Laird, N., and Ware, J. H. (1984). Random-Effects Models for Serial Observations with Binary Response. *Biometrics*, 40(4):961–971. 96

Sun, D., Speckman, P. L., and Tsutakawa, R. K. (2000). Random effects in generalized linear mixed models (glmms). *BIOSTATISTICS-BASEL-*, 5:23–40. 198

Sung, Y. J., Geyer, C. J., et al. (2007). Monte Carlo likelihood inference for missing data models. *The Annals of Statistics*, 35(3):990–1011. 90

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 58(1):267–288. 30, 42, 64

Tierney, L. and Kadane, J. B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, 81(393):82–86. 81

Tokdar, S. T. and Kass, R. E. (2010). Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60. 90

Trottier, C. (1998). *Estimation dans les modèles linéaires généralisés à effets aléatoires*. PhD thesis, Institut National Polytechnique de Grenoble-INPG. 78

Tseng, P. (2001). Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494. 65

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0. 85

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika*, 61(3):439–447. 59, 85

Wei, G. C. and Tanner, M. A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704. 86

Wilkinson, J. H. (1965). *The Algebraic Eigenvalue Problem*, volume 87. Clarendon Press Oxford. 138

Wold, H. (1966). Estimation of Principal Components and Related Models by Iterative Least Squares. *Multivariate Analysis*, pages 391–420. 30, 42, 68

Wold, S., Kettaneh-Wold, N., and Skagerberg, B. (1989). Nonlinear PLS modeling. *Chemometrics and Intelligent Laboratory Systems*, 7(1-2):53–65. 201

Wold, S., Martens, H., and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In *Matrix Pencils*, pages 286–293. Springer. 30, 42, 68

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130. 68

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48(3-4):233–243. 95

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103. 151

Wu, T. T., Lange, K., et al. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244. 64

Yi, X. and Caramanis, C. (2015). Regularized EM Algorithms: A Unified Framework and Statistical Guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575. 161

Zeger, S. L. and Karim, M. R. (1991). Generalized Linear Models With Random Effects; A Gibbs Sampling Approach. *Journal of the American Statistical Association*, 86(413):79–86. 34, 46, 88

Zhang, Y., Zhou, H., Zhou, J., and Sun, W. (2017). Regression Models for Mul-

tivariate Count Data. *Journal of Computational and Graphical Statistics*, 26(1):1–13. 98

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 67(2):301–320. 30, 42, 65

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286. 200

Résumé. Une forte redondance des variables explicatives cause de gros problèmes d'identifiabilité et d'instabilité des coefficients dans les modèles de régression. Même lorsque l'estimation est possible, l'interprétation des résultats est donc extrêmement délicate. Il est alors indispensable de combiner à leur vraisemblance un critère supplémentaire qui régularise l'estimateur. Dans le sillage de la régression PLS, la stratégie de régularisation que nous considérons dans cette thèse est fondée sur l'extraction de *composantes supervisées*. Contraintes à l'orthogonalité entre elles, ces composantes doivent non seulement capturer l'information structurelle des variables explicatives, mais aussi prédire autant que possible les variables réponses, qui peuvent être de types divers (continues ou discrètes, quantitatives, ordinales ou nominales). La régression sur composantes supervisées a été développée pour les GLMs multivariés, mais n'a jusqu'alors concerné que des modèles à observations indépendantes.
Or dans de nombreuses situations, les observations sont groupées. Nous proposons une extension de la méthode aux GLMMs multivariés, pour lesquels les corrélations intra–groupes sont modélisées au moyen d'effets aléatoires. À chaque étape de l'algorithme de Schall permettant l'estimation du GLMM, nous procédons à la régularisation du modèle par l'extraction de composantes maximisant un compromis entre qualité d'ajustement et pertinence structurelle. Comparé à la régularisation par pénalisation de type ridge ou LASSO, nous montrons sur données simulées que notre méthode non seulement permet de révéler les dimensions explicatives les plus importantes pour l'ensemble des réponses, mais fournit souvent une meilleure prédiction. La méthode est aussi évaluée sur données réelles.
Nous développons enfin des méthodes de régularisation dans le contexte spécifique des données de panel (impliquant des mesures répétées sur différents individus aux mêmes dates). Deux effets aléatoires sont introduits : le premier modélise la dépendance des mesures relatives à un même individu, tandis que le second modélise un effet propre au temps (possédant donc une certaine inertie) partagé par tous les individus. Pour des réponses Gaussiennes, nous proposons d'abord un algorithme EM pour maximiser la vraisemblance du modèle pénalisée par la norme L_2 des coefficients de régression. Puis nous proposons une alternative consistant à donner une prime aux directions les plus "fortes" de l'ensemble des prédicteurs. Une extension de ces approches est également proposée pour des données non–Gaussiennes, et des tests comparatifs sont effectués sur données Poissonniennes.

Mots clés. Composantes supervisées, pertinence structurelle, régularisation, GLMM multivarié, effets aléatoires.

Abstract. High redundancy of explanatory variables results in identification troubles and a severe lack of stability of regression model estimates. Even when estimation is possible, a consequence is the near–impossibility to interpret the results. It is then necessary to combine its likelihood with an extra–criterion regularising the estimates. In the wake of PLS regression, the regularising strategy considered in this thesis is based on extracting *supervised components*. Such orthogonal components must not only capture the structural information of the explanatory variables, but also predict as well as possible the response variables, which can be of various types (continuous or discrete, quantitative, ordinal or nominal). Regression on supervised components was developed for multivariate GLMs, but so far concerned models with independent observations.
However, in many situations, the observations are grouped. We propose an extension of the method to multivariate GLMMs, in which within–group correlations are modelled with random effects. At each step of Schall's algorithm for GLMM estimation, we regularise the model by extracting components that maximise a trade–off between goodness–of–fit and structural relevance. Compared to penalty–based regularisation methods such as ridge or LASSO, we show on simulated data that our method not only reveals the important explanatory dimensions for all responses, but often gives a better prediction too. The method is also assessed on real data.
We finally develop regularisation methods in the specific context of panel data (involving repeated measures on several individuals at the same time–points). Two random effects are introduced: the first one models the dependence of measures related to the same individual, while the second one models a time–specific effect (thus having a certain inertia) shared by all the individuals. For Gaussian responses, we first propose an EM algorithm to maximise the likelihood penalised by the L_2–norm of the regression coefficients. Then, we propose an alternative which rather gives a bonus to the "strongest" directions in the explanatory subspace. An extension of these approaches is also proposed for non–Gaussian data, and comparative tests are carried out on Poisson data.

Keywords. Supervised components, structural relevance, regularisation, multivariate GLMM, random effects.