Component-based regularisation of multivariate generalised linear mixed models

Jocelyn Chauvet, Catherine Trottier, Xavier Bry

Component–based regularisation of multivariate generalised linear mixed models

Jocelyn Chauvet *
Catherine Trottier†*
and
Xavier Bry*

Abstract

We address the component–based regularisation of a multivariate Generalised Linear Mixed Model (GLMM) in the framework of grouped data. A set $\boldsymbol{Y}$ of random responses is modelled with a multivariate GLMM, based on a set $\boldsymbol{X}$ of explanatory variables, a set $\boldsymbol{A}$ of additional explanatory variables, and random effects to introduce the within–group dependence of observations. Variables in $\boldsymbol{X}$ are assumed many and redundant so that regression demands regularisation. This is not the case for $\boldsymbol{A}$, which contains few and selected variables. Regularisation is performed building an appropriate number of orthogonal components that both contribute to model $\boldsymbol{Y}$ and capture relevant structural information in $\boldsymbol{X}$. To estimate the model, we propose to maximise a criterion specific to the Supervised Component–based Generalised Linear Regression (SCGLR) within an adaptation of Schall's algorithm. This extension of SCGLR is tested on both simulated and real grouped data, and compared to ridge and LASSO regularisations. Supplementary material for this article is available online.

Keywords: generalised linear regression, supervised components, random effects, structural relevance.

*IMAG, Univ Montpellier, CNRS, Montpellier, France.
jocelyn.chauvet@umontpellier.fr ; xavier.bry@umontpellier.fr
†Univ Paul–Valéry Montpellier 3, Montpellier, France.
catherine.trottier@univ-montp3.fr

1 Introduction

In the framework of regression models on a large number of explanatory variables with redundancies and collinearities, the search for a reduced number of relevant dimensions to model responses has been an ongoing research over the last decades. In particular, the case where the explanatory variables outnumber the observations tends to be a new standard. Generalised Linear Models (GLMs) are the most widely used regression models, because they are easy to interpret and address a very large scope of applications with a variety of response distributions. For instance, Epidemiology, Biology and Social Sciences need to model binary outcomes, count data and survival times. All these fields often have to deal with both grouped data and multivariate responses combining variables of different types (e.g. one binary and another Poisson). In this work, we particularly aim at modelling abundances of several tree genera on plots of land grouped into forest concessions, using multiple redundant explanatory variables.

As far as dimension–reduction is concerned, two main approaches have been developed. The first one is variable–selection, whereas the second one builds components, i.e. linear combinations of the explanatory variables, which synthesise the useful part of their information. As far as variable–selection is concerned, the most popular method is currently the LASSO, introduced by Tibshirani (1996), which combines the likelihood with a penalty based on the L_1–norm of the coefficient vector. LASSO is one of the penalty–based regularisation methods, as are also ridge (Hoerl and Kennard, 1970) and elastic–net (Zou and Hastie, 2005). This LASSO selection approach has proved efficient to explain the phenomenon of interest when some of the explanatory variables are the "true" ones, surrounded by a high number of irrelevant others. Nevertheless, it may be very unstable and helpless when the true explanatory dimensions are latent and indirectly measured through highly correlated proxies. This is where the component–based approach turns out to be useful. Bry et al. (2013) have developed a new methodology named Supervised Component–based Generalised Linear Regression (SCGLR), later extended and refined in Bry et al. (2014, 2016, 2018). As in any PLS–type method, the construction of components in SCGLR is guided both by the correlation–structure of variables in the explanatory space 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 components are wanted to align with in the explanatory space (e.g. variable bundles, principal components, other subspaces). Moreover, SCGLR searches for explanatory directions common to multiple responses with probability distributions in the exponential family, each response being entitled to their own distribution. The current SCGLR method is implemented in the R package `SCGLR` (Cornu et al., 2018) available at `https://CRAN.R-project.org/package=SCGLR` and `https://github.com/SCnext/SCGLR`.

In the present work, we aim at modelling responses with a repeated or grouped design. For this purpose, the use of mixed models with random effects is widespread. Research on variance–component estimation in Generalised Linear Mixed Models (GLMMs) has been very active since the 1980s. 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 numerical approximations of the integral have been proposed: Gaussian quadrature (Anderson and Aitkin, 1985) or adaptive versions of it (Pinheiro and Bates, 1995), Laplace approximation leading to the definition of the penalised quasi–likelihood (Breslow and Clayton, 1993) or modified versions of it (Shun and McCullagh, 1995). An alternative to this type of analytic approximation is a stochastic approximation of the integral calculation via MCMC techniques. In this approach, Zeger and Karim (1991) described an approximate Gibbs sampling for GLMMs, which was extended by Clayton (1996) to more general Metropolis–Hastings algorithms. In parallel, McCulloch (1997) developed the Monte Carlo EM algorithm where the expectation is computed numerically through a Monte Carlo approximation, after generating random effects with a Metropolis–Hastings sampler. Mention can also be made of the recent work by Knudson (2016): her strategy is to approximate the entire likelihood function using random effects simulated from a parametrised importance sampling distribution depending on the data. Unfortunately, these different approaches are not necessarily suitable for the same types of random effect designs (one–dimensional random effect, embedded random effects, etc). In the wake of the first type of approximations, we here adopt the "Joint–Maximisation" strategy (McCulloch, 1997), as introduced for instance by Schall (1991). The model is

iteratively linearised conditional on the random effects and variance components are then estimated using adapted linear mixed models methods. This strategy can be used for any random effect design and is less computationally intensive than Monte Carlo methods. Moreover, it provides us with a linear setting more suitable for the computation of components. Once the components calculated, model parameters can be estimated using any of the aforementioned strategies (see Section 7).

Modelling grouped responses through a GLMM with a large number of explanatory variables is the focus of this paper. The need for dimension–reduction and regularisation has to accommodate the presence of random effects in the model, but our main purpose still remains to investigate the explanatory structure and link it to interpretable dimensions. For Gaussian responses, Eliot et al. (2011) proposed to extend the ridge regression to Linear Mixed Models (LMMs). Based on a penalised complete log–likelihood, the adaptation of the Expectation–Maximisation algorithm they suggest includes a new step to find the best shrinkage parameter using a generalised cross–validation scheme at each iteration. More recently, Schelldorfer et al. (2014) — and also Groll and Tutz (2014) — proposed an L_1–penalised algorithm for fitting a high–dimensional GLMM, using Laplace approximation and an efficient coordinate gradient descent. In this work, 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 multivariate 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.

The paper is organised as follows. In Section 2, we formalise the model and set the main notations used throughout the paper. In Section 3, we present the key features of SCGLR. Section 4 designs an extension of this methodology to mixed models, and particularly to grouped data. In Section 5, our extended method "mixed–SCGLR" is evaluated on simulations and compared to ridge– and LASSO–based regularisations. Finally, in order to highlight the power of mixed–SCGLR in terms of model interpretation, Section 6 presents an application to real data in the Poisson case.

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 response $\boldsymbol{y_k}$, 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.

Notations and conventions

- All variables (namely the $\boldsymbol{a_i}$'s, $\boldsymbol{x_j}$'s and $\boldsymbol{y_k}$'s) will be identified with n–vectors.
- We will use bold lowercase letters for vectors (e.g. $\boldsymbol{u}$) and bold capital letters for matrices (e.g. $\boldsymbol{M}$).
- $\boldsymbol{M}$ being any matrix, $\boldsymbol{M}^{\mathrm{T}}$ denotes the transpose of $\boldsymbol{M}$.
- $\boldsymbol{I}_n$ denotes the identity matrix of size n.
- $\boldsymbol{1}_m$ denotes the all–ones vector of size m.
- Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be non–zero vectors in $\mathbb{R}^d$ and let $\boldsymbol{M}$ be a symmetric positive definite matrix of size $d \times d$. Then $\langle\, \boldsymbol{u} \,|\, \boldsymbol{v} \,\rangle_{\boldsymbol{M}} = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{v}$ refers to the Euclidean scalar product of $\boldsymbol{u}$ and $\boldsymbol{v}$ with respect to metric $\boldsymbol{M}$. The cosine of the angle between $\boldsymbol{u}$ and $\boldsymbol{v}$ with respect to $\boldsymbol{M}$ is given by $\cos_{\boldsymbol{M}}(\boldsymbol{u}, \boldsymbol{v}) = \dfrac{\langle\, \boldsymbol{u} \,|\, \boldsymbol{v} \,\rangle_{\boldsymbol{M}}}{\|\boldsymbol{u}\|_{\boldsymbol{M}} \, \|\boldsymbol{v}\|_{\boldsymbol{M}}}$, where $\|\boldsymbol{u}\|_{\boldsymbol{M}} = \sqrt{\langle\, \boldsymbol{u} \,|\, \boldsymbol{u} \,\rangle_{\boldsymbol{M}}}$.

- The space spanned by vectors $\boldsymbol{u_1}, \ldots, \boldsymbol{u_h}$ is denoted by $\text{span}\{\boldsymbol{u_1}, \ldots, \boldsymbol{u_h}\}$. $\boldsymbol{U}$ being any matrix, $\text{span}\{\boldsymbol{U}\}$ 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 $\Pi^{\boldsymbol{W}}_{\text{span}\{\boldsymbol{Z}\}}$ refers to the $\boldsymbol{W}$–orthogonal projector onto $\text{span}\{\boldsymbol{Z}\}$. Let $\boldsymbol{b}$ be a vector in $\mathbb{R}^n$. The cosine of the angle between $\boldsymbol{b}$ and $\text{span}\{\boldsymbol{Z}\}$ with respect to $\boldsymbol{W}$ is defined by $\cos_{\boldsymbol{W}}(\boldsymbol{b}, \text{span}\{\boldsymbol{Z}\}) = \cos_{\boldsymbol{W}}\left(\boldsymbol{b}, \Pi^{\boldsymbol{W}}_{\text{span}\{\boldsymbol{Z}\}}\boldsymbol{b}\right)$.

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 3.1 briefly recalls some standards for univariate GLMs. Section 3.2 defines the linear predictors considered in the SCGLR methodology, in a multivariate GLM framework with additional explanatory variables. Finally, Section 3.3 introduces the criterion SCGLR maximises to compute the component.

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

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

where $\boldsymbol{z}^{[t]}$ and $\boldsymbol{W}^{[t]}$ respectively denote the classical working variable and the associated weight matrix at iteration t. As pointed out by Nelder and Wedderburn (1972), update (1) may be interpreted as a weighted least squares step in the linearised model $\mathcal{M}^{[t]}$ defined by

$$\mathcal{M}^{[t]} : \left| \begin{array}{l} \boldsymbol{z}^{[t]} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\zeta}^{[t]} \\ \text{with: } \mathbb{E}\left(\boldsymbol{\zeta}^{[t]}\right) = \boldsymbol{0} \text{ and } \mathbb{V}\left(\boldsymbol{\zeta}^{[t]}\right) = \boldsymbol{W}^{[t]^{-1}}. \end{array} \right. \tag{2}$$

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 a 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

$$\boldsymbol{\eta_k} = (\boldsymbol{X}\boldsymbol{u})\,\gamma_k + \boldsymbol{A}\boldsymbol{\delta_k}, \tag{3}$$

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 $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{u} = 1$, 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 $\boldsymbol{f}$, 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

$$\mathcal{M}_k^{[t]}: \left| \begin{array}{l} \boldsymbol{z_k}^{[t]} = (\boldsymbol{X}\boldsymbol{u})\,\gamma_k + \boldsymbol{A}\boldsymbol{\delta_k} + \boldsymbol{\zeta_k}^{[t]} \\ \text{with: } \mathbb{E}\left(\boldsymbol{\zeta_k}^{[t]}\right) = 0 \text{ and } \mathbb{V}\left(\boldsymbol{\zeta_k}^{[t]}\right) = {\boldsymbol{W_k}^{[t]}}^{-1}. \end{array} \right. \tag{4}$$

Although linearised models (2) and (4) seem very similar, (4) is no longer linear, owing to the product $\boldsymbol{u}\gamma_k$. An alternate 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 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 parameters γ_k and $\boldsymbol{\delta_k}$.

3.3 Calculating the component maximising 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 $\boldsymbol{W_k}$. As suggested in Bry et al. (2013), the best loading–vector in the weighted least–squares sense would be the solution of

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

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

$$\begin{aligned} \psi_{\boldsymbol{A}}(\boldsymbol{u}) &= \sum_{k=1}^{q} \left\| \boldsymbol{z_k} \right\|^2_{\boldsymbol{W_k}} \cos^2_{\boldsymbol{W_k}} \left(\boldsymbol{z_k},\, \mathrm{span}\left\{ \boldsymbol{Xu}, \boldsymbol{A} \right\} \right) \\ &= \sum_{k=1}^{q} \left\| \boldsymbol{z_k} \right\|^2_{\boldsymbol{W_k}} \cos^2_{\boldsymbol{W_k}} \left(\boldsymbol{z_k},\, \Pi^{\boldsymbol{W_k}}_{\mathrm{span}\{\boldsymbol{Xu},\boldsymbol{A}\}} \boldsymbol{z_k} \right). \end{aligned} \tag{5}$$

Now, $\psi_{\boldsymbol{A}}$ is a mere goodness–of–fit (GoF) measure that does not take into account the closeness of component $\boldsymbol{f} = \boldsymbol{Xu}$ 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

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

where

$$\phi(\boldsymbol{u}) = \left[\sum_{j=1}^{p} \omega_j \left(\langle\, \boldsymbol{Xu} \,|\, \boldsymbol{x_j} \,\rangle^2_{\boldsymbol{P}} \right)^{l} \right]^{\frac{1}{l}} = \left[\sum_{j=1}^{p} \omega_j \left(\boldsymbol{u}^{\mathrm{T}}\, \boldsymbol{X}^{\mathrm{T}} \boldsymbol{P} \boldsymbol{x_j} \boldsymbol{x_j}^{\mathrm{T}} \boldsymbol{P} \boldsymbol{X}\, \boldsymbol{u} \right)^{l} \right]^{\frac{1}{l}}, \ l \geqslant 1, \tag{6}$$

for the scalar product is commutative. Formula (6) is in fact a particular case of the SR criterion proposed by Bry and Verron (2015); Bry et al. (2016). It can be viewed as a generalised average version of the usual dual PCA criterion: $\sum_{j=1}^{p} \cos^2_{\boldsymbol{P}}(\boldsymbol{Xu}, \boldsymbol{x_j}) =$

$\sum_{j=1}^{p} \langle \boldsymbol{X}\boldsymbol{u} \,|\, \boldsymbol{x_j} \rangle_{\boldsymbol{P}}^2$. For $\boldsymbol{M} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X})^{-1}$, (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 column full rank matrix. In case of strict collinearities within $\boldsymbol{X}$, as it always happens in high–dimensional settings, we replace $\boldsymbol{X}$ with the matrix $\boldsymbol{C}$ of its principal components associated with non–zero eigenvalues. The component is then sought as $\boldsymbol{f} = \boldsymbol{C}\boldsymbol{u}$. We have $\boldsymbol{C} = \boldsymbol{X}\boldsymbol{V}$, where $\boldsymbol{V}$ is the matrix of corresponding unit-eigenvectors. Then, $\boldsymbol{f} = \boldsymbol{C}\boldsymbol{u} = \boldsymbol{X}\widetilde{\boldsymbol{u}}$ with $\widetilde{\boldsymbol{u}} = \boldsymbol{V}\boldsymbol{u}$. Bry et al. (2018) show that among all loading–vectors $\boldsymbol{t}$ such that $\boldsymbol{X}\boldsymbol{t} = \boldsymbol{f}$, $\widetilde{\boldsymbol{u}}$ is that which has the minimum L_2–norm.

Tuning parameter l allows to draw component towards more (greater l) or less (smaller l) local bundles of correlated variables, as depicted on Figure 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 relative to the GoF. SCGLR attempts a trade–off between (5) and (6) by solving

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

or equivalently

$$\max_{\boldsymbol{u}:\, \boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{u}=1} s \log\left[\phi\left(\boldsymbol{u}\right)\right] + (1-s) \log\left[\psi_{\boldsymbol{A}}\left(\boldsymbol{u}\right)\right]. \tag{7}$$

More detail can be found in Bry et al. (2018).

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 as-

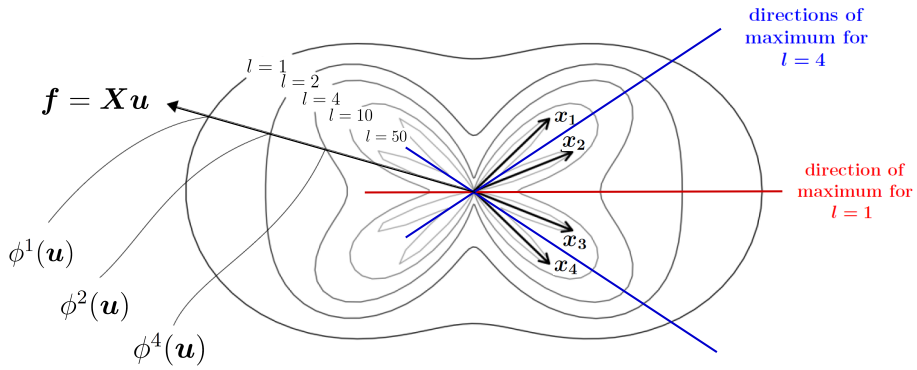

Figure 1: Polar representation of the VPI according to the value of l in the elementary case of four coplanar variables, $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \boldsymbol{x_4}$, with $\omega_j = \frac{1}{4} \forall j \in \{1, 2, 3, 4\}$. *Loading–vector* $\boldsymbol{u}$ *is identified 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.*

sumption 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 to be independent conditional on the components. Section 4.1 presents the single–component mixed–SCGLR method. The underlying algorithm is given in Section 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 4.3.

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:

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

where N denotes the number of groups. In this paper, 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

$$\forall k \in \{1, \ldots, q\}\,, \quad \boldsymbol{\eta_k^{\xi}} = (\boldsymbol{X}\boldsymbol{u})\gamma_k + \boldsymbol{A}\boldsymbol{\delta_k} + \boldsymbol{U}\boldsymbol{\xi_k}, \tag{8}$$

where $\boldsymbol{U}$ is the known random effects' design matrix. Predictor $\boldsymbol{\eta_k^{\xi}}$ epitomises the way we capture the dependence between outcomes. Indeed, as component $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$ 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 $\boldsymbol{y_k}$.

Recall that the distribution of the data 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 (8).

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. $\boldsymbol{\mu_k^{\xi}} := \mathbb{E}\left(\boldsymbol{y_k} \,|\, \boldsymbol{\xi_k}\right)$). The 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

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

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

$$\mathcal{M}_k^{\xi\,[t]} : \left| \begin{array}{l} \boldsymbol{z_k^{\xi}}^{[t]} = (\boldsymbol{X}\boldsymbol{u})\,\gamma_k + \boldsymbol{A}\boldsymbol{\delta_k} + \boldsymbol{U}\boldsymbol{\xi_k} + \boldsymbol{e_k}^{[t]} \\ \text{with: } \mathbb{E}\left(\boldsymbol{e_k}^{[t]} \,|\, \boldsymbol{\xi_k}\right) = 0 \text{ and } \mathbb{V}\left(\boldsymbol{e_k}^{[t]} \,|\, \boldsymbol{\xi_k}\right) = \boldsymbol{W_k^{\xi}}^{-1[t]}. \end{array} \right. \tag{9}$$

Besides the variance component estimation, an alternated estimation step has to be developed (as aforementioned in Section 3.2) to deal with the non–linearity of (9).

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 (7)–type program. However, (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

$$\psi_{\boldsymbol{A}}(\boldsymbol{u}) = \sum_{k=1}^{q} \left\| \boldsymbol{z_k^{\xi}} \right\|^2_{\boldsymbol{W_k^{\xi}}} \cos^2_{\boldsymbol{W_k^{\xi}}} \left(\boldsymbol{z_k^{\xi}},\, \text{span}\left\{\boldsymbol{X}\boldsymbol{u}, \boldsymbol{A}\right\} \right). \tag{10}$$

Estimating the regression parameters and variance–components: Given a current value of component $\boldsymbol{f}$, we apply Schall's method with the linear predictors given in (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 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

$$\sigma_k^2 \longleftarrow \frac{\widehat{\boldsymbol{\xi_k}}^{\mathrm{T}}\,\widehat{\boldsymbol{\xi_k}}}{N - \frac{1}{\sigma_k^2}\,\text{Trace}\left[\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{W_k^{\xi}}\boldsymbol{U} + \boldsymbol{D_k^{-1}}\right)^{-1}\right]}.$$

4.2 The algorithm

The conditional linearised models considered at iteration t are given by (9). Algorithm 1 describes the $(t+1)$–th iteration of the single–component mixed–SCGLR. It is repeated until stability of parameters is reached.

Step 1: Computing the component. Set

$$\boldsymbol{u}^{[t+1]} = \underset{\boldsymbol{u}:\,\boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{u}=1}{\arg\max} \left[\phi\left(\boldsymbol{u}\right)\right]^{s} \left[\psi_{\boldsymbol{A}}^{[t]}\left(\boldsymbol{u}\right)\right]^{1-s}, \text{ where } \psi_{\boldsymbol{A}}\left(\boldsymbol{u}\right) \text{ is given by (10) and } \phi\left(\boldsymbol{u}\right) \text{ by (6)}$$

$$\boldsymbol{f}^{[t+1]} = \boldsymbol{X}\boldsymbol{u}^{[t+1]}$$

Step 2: Henderson systems. For each $k \in \{1,\ldots,q\}$, solve the system

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

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

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

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: $t \longleftarrow t+1$

Algorithm 1: Iteration of the single–component mixed–SCGLR

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

$$\begin{cases} \max \quad s \log \left[\phi \left(\boldsymbol{u} \right) \right] + (1 - s) \log \left[\psi_{\boldsymbol{A_h}} \left(\boldsymbol{u} \right) \right] \\ \text{subject to:} \quad \boldsymbol{u}^\top \boldsymbol{M}^{-1} \boldsymbol{u} = 1 \text{ and } \boldsymbol{D_h^\top u} = \boldsymbol{0}, \end{cases} \tag{11}$$

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

In the online Supplementary Material, 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 (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 $\boldsymbol{D_0} = \boldsymbol{0}$.

5 Comparative results on simulated data

Five simulation studies have been implemented to assess our method. The first one (discussed in Sections 5.1 – 5.4) focuses on LMMs. 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) extends the first one to binary and Poisson outcomes. 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 is available at `https://github.com/SCnext/mixedSCGLR`. Three additional simulations are presented in the online Supplementary Material. The first one reproduces the simulation scheme of Section 5.5 with binomial and Poisson outcomes. The second one assesses the performance of mixed–SCGLR on a different bundle structure and presents results concerning variance component estimates. The third one deals with high dimensional data.

5.1 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 effects' design matrix is then $\boldsymbol{U} = \boldsymbol{I}_N \otimes \mathbf{1}_R$. 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

$$\operatorname{cor}(\boldsymbol{X_j}) = \tau \mathbf{1}_{p_j} \mathbf{1}_{p_j}^{\mathrm{T}} + (1-\tau)\boldsymbol{I}_{p_j},$$

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

$$\begin{cases} \boldsymbol{y_1} = \boldsymbol{X}\boldsymbol{\beta_1} + \boldsymbol{U}\boldsymbol{\xi_1} + \boldsymbol{\varepsilon_1} \\ \boldsymbol{y_2} = \boldsymbol{X}\boldsymbol{\beta_2} + \boldsymbol{U}\boldsymbol{\xi_2} + \boldsymbol{\varepsilon_2}, \end{cases} \tag{12}$$

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

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

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

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 (12).

5.2 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 and 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 are 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

$$E^{(b)} = \frac{1}{2} \sum_{k=1}^{2} E_k^{(b)}, \tag{13}$$

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 (13). We then define

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

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 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 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

5.3 Comparison of the 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 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 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

5.4 Model interpretation

This section aims at highlighting the power of mixed–SCGLR for model interpretation. Figure 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.

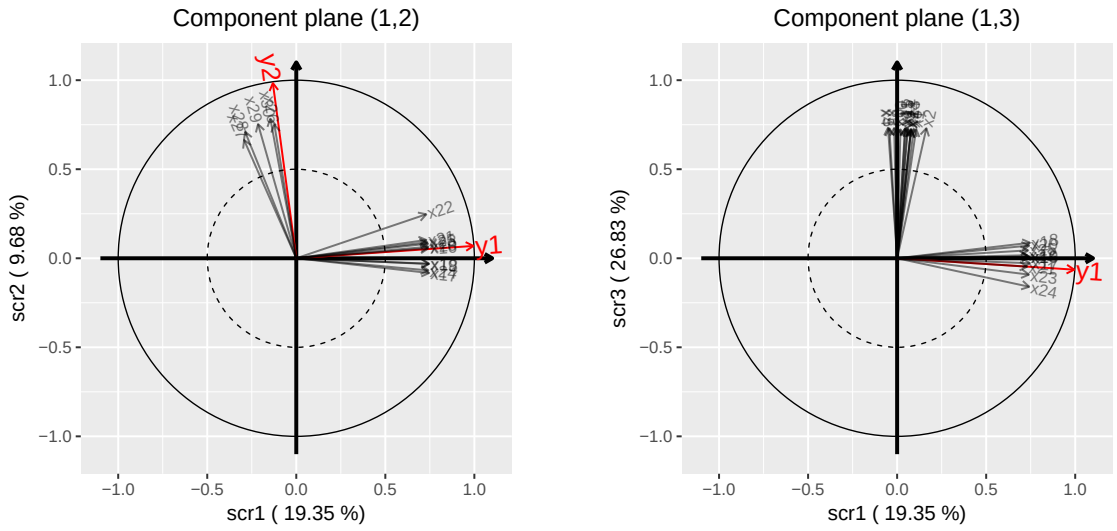

Figure 2: Component planes $(1, 2)$ and $(1, 3)$ given by mixed–SCGLR on simulated data. *The black arrows represent the explanatory variables. The red ones represent the projection of the* $\boldsymbol{X}$*–part of the linear predictors associated with* $\boldsymbol{y_1}$ *and* $\boldsymbol{y_2}$*. The percentage of inertia captured by each component is given in parentheses.*

5.5 Additional simulations involving non–Gaussian 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. We keep design matrices $\boldsymbol{X}$ and $\boldsymbol{U}$ defined in Section 5.1, as well as the values of $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, σ_1^2 and σ_2^2. 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 $k \in \{1, 2\}$, $\widetilde{\boldsymbol{\xi_k}} \sim \mathcal{N}_N(\boldsymbol{0},\, \varsigma_k^2\, \boldsymbol{I}_N)$. Then given $\widetilde{\boldsymbol{\xi_1}}$ and $\widetilde{\boldsymbol{\xi_2}}$, we simulate $\boldsymbol{Y} = \left[\, \boldsymbol{y_1} \mid \boldsymbol{y_2} \,\right]$ as

$$\begin{cases} \boldsymbol{y_1} \sim \mathcal{B}\left(\boldsymbol{p} = \text{logit}^{-1}\left[\boldsymbol{X\theta_1} + \boldsymbol{U}\widetilde{\boldsymbol{\xi_1}}\right]\right) \\ \boldsymbol{y_2} \sim \mathcal{P}\left(\boldsymbol{\lambda} = \exp\left[\boldsymbol{X\theta_2} + \boldsymbol{U}\widetilde{\boldsymbol{\xi_2}}\right]\right), \end{cases} \tag{14}$$

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 (14). As in Section 5.2, tuning parameters are calibrated so as to minimise the cross–validation error (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

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

Unlike in Section 5.3, the response–vectors do not come from the same distribution and have different orders of magnitude. The fixed–effect relative errors are thus not comparable. 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 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 5.3), 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.

Even if the response variables are not Gaussian, the power of mixed–SCGLR for model interpretation is preserved. Graphical diagnoses similar to those provided in Section 5.4 are available in the Supplementary Material.

Table 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

6 An application to forest ecology data

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. 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 $\boldsymbol{A}$.

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

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

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. Using the same procedure and notations as in Section 5.2, the data are divided into five parts $\mathcal{P}_1, \ldots, \mathcal{P}_5$. Let n_j be the size of $\mathcal{P}_j$.

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

where

$$E_k = \frac{1}{5}\sum_{j=1}^{5}\sqrt{\frac{1}{n_j}\sum_{i\in\mathcal{P}_j}\frac{\left(y_{k,i} - \widehat{y_{k,i(-j)}}\right)^2}{\widehat{\text{var}}\left(\widehat{y_{k,i(-j)}}\right)}}. \tag{15}$$

On Figure 3, 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 (s exactly equal to 0) in $\boldsymbol{X}$, mixed–SCGLR merely performs classical GLMM estimation and does not converge with this data. 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.

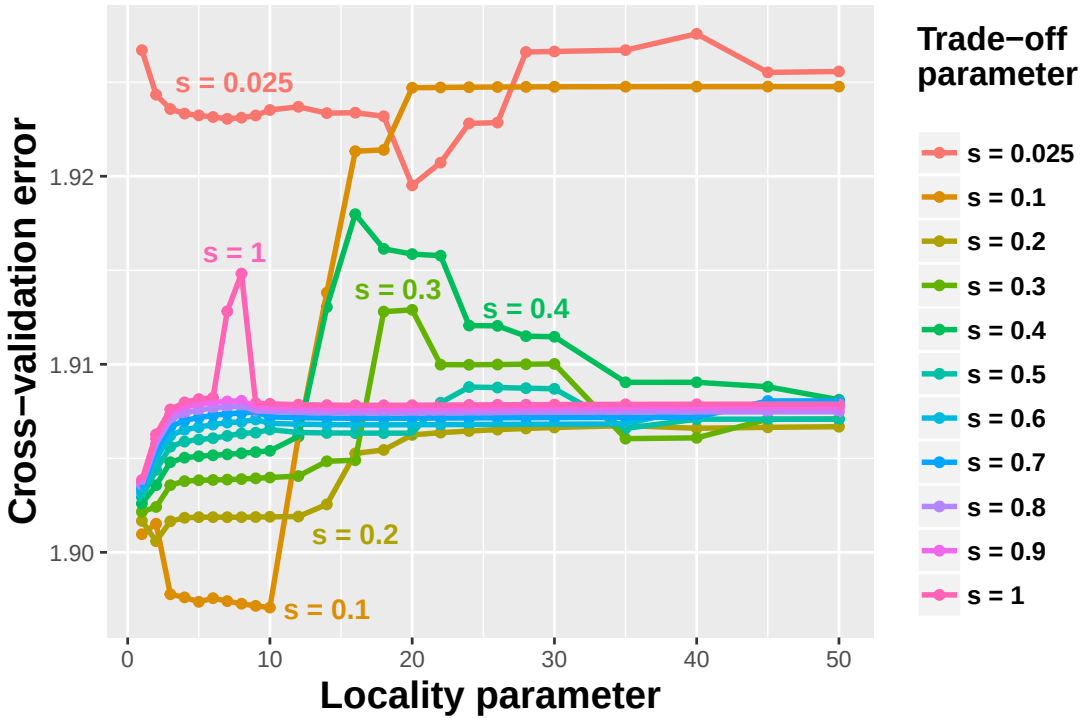

Figure 3: Behaviour of the cross–validation error E for trade–off parameter $s \in \{0.025, 0.1, 0.2, \ldots, 1\}$, as a function of locality parameter $l \in [1, 50]$.

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 initial version of SCGLR are compared with respect to the cross–validation criterion given by (15). Table 4 summarises the E_k's for both SCGLR and mixed–SCGLR methods. Optimal parameter value triplet $(K^\star, s^\star, l^\star) =$

$(4, 0.1, 10)$ is selected 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 4: Cross–validation errors for each response variable.

	E_{cv}^1	E_{cv}^2	E_{cv}^3	E_{cv}^4	E_{cv}^5	E_{cv}^6	E_{cv}^7	E_{cv}^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 4.

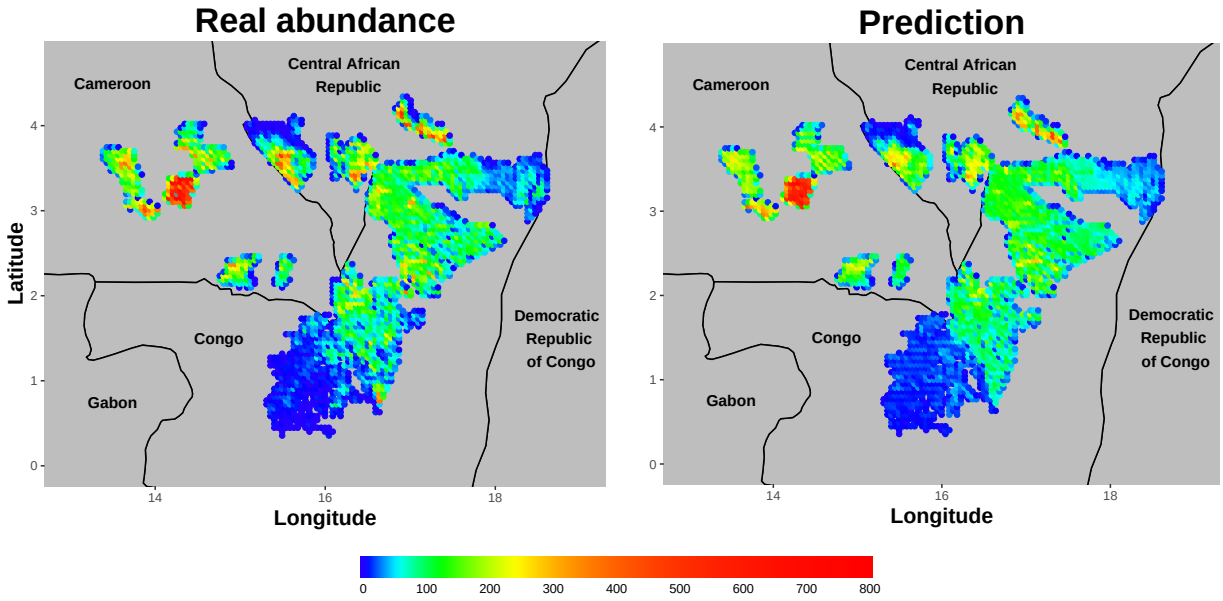

Figure 4: Abundance maps issued from mixed–SCGLR. *The plots respectively show real abundance (left) and associated conditional predictions (right) of the tree species number* 8. *Each point represents a land plot (2615 in total).*

As has been seen in Section 5.4, mixed–SCGLR allows an easy interpretation of the model through the decomposition of linear predictors on interpretable components. Figure 5 shows the first two component planes resulting from mixed–SCGLR on real data *Genus*. 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 $\boldsymbol{y_8}$.

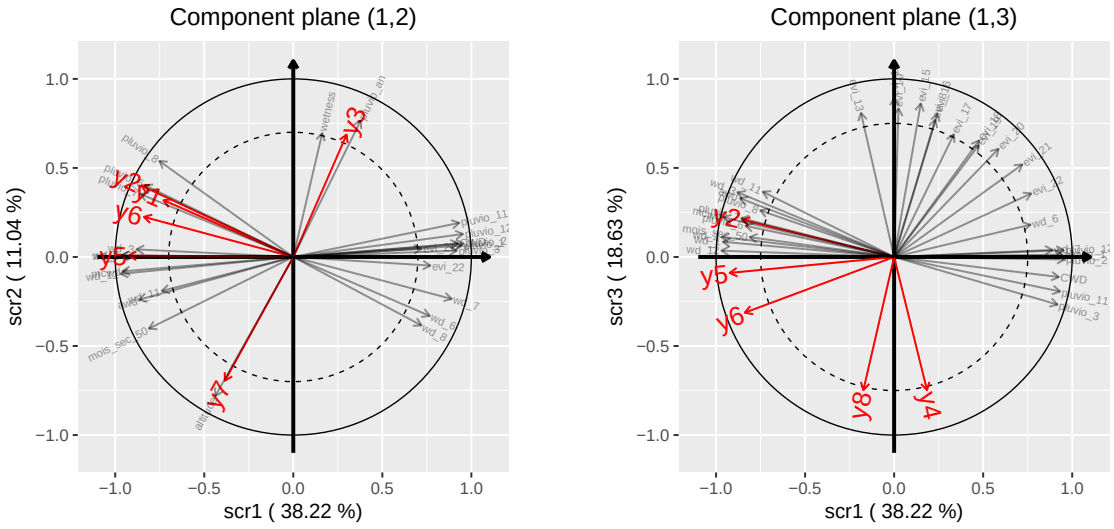

Figure 5: Component planes $(1, 2)$ and $(1, 3)$ output by mixed–SCGLR on dataset *Genus*, with optimal parameter triplet $(K^\star, s^\star, l^\star) = (4, 0.1, 10)$. *The left–hand side plot displays only variables having cosine greater than* 0.7 *with component plane* $(1, 2)$. *The right–hand side plots variables having cosine greater than* 0.75 *with component plane* $(1, 3)$.

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.

7 Discussion and Conclusions

Like Sufficient Dimension Reduction (SDR) methods, 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 component estimates are prone to some bias towards zero (McCulloch, 1997). That is why other estimation strategies might be considered, especially Monte Carlo integration methods which have the advantage of being based on direct approximations of the likelihood. Some examples are the MCMC methods developed by Hadfield (2010) in the GLMM framework, and the Monte Carlo Likelihood Approximation (MCLA) proposed by Knudson (2016). Indirect maximisations of the likelihood are also available such as Monte Carlo Expectation–Maximisation (MCEM) and Monte Carlo Newton–Raphson (McCulloch, 1997). 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 1 with the following steps (to keep things simple, we take the canonical link):

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

$$f_{\boldsymbol{y_k}|\boldsymbol{\xi_k},\boldsymbol{\gamma_k},\boldsymbol{\delta_k}}\left(\boldsymbol{y_k}|\boldsymbol{\xi_k},\boldsymbol{\gamma_k},\boldsymbol{\delta_k}\right) = \exp\left\{\boldsymbol{y_k}^{\mathrm{T}}\boldsymbol{\eta_k^{\xi}} - \mathbf{1}^{\mathrm{T}} c\left(\boldsymbol{\eta_k^{\xi}}\right) + \mathbf{1}^{\mathrm{T}} d\left(\boldsymbol{y_k}\right)\right\}$$
$$\boldsymbol{\xi_k}|\boldsymbol{D_k} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{D_k}\right),$$

   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: $c(x) = \log(1 + e^x)$ and $d(x) = 0$.
3. Apply MC–based methods such as MCMC, MCLA, MCEM or MCNR to update $\boldsymbol{\gamma_k}$, $\boldsymbol{\delta_k}$, $\boldsymbol{\xi_k}$ and $\boldsymbol{D_k}$, $k \in \{1, \ldots, q\}$.
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" 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.

SUPPLEMENTARY MATERIAL

Additional simulations: The first simulation reproduces that of Section 5.5 in the case of binomial and Poisson outcomes. The second simulation explores a different structure of variable–bundles, considers Gaussian, binomial and Poisson outcomes, and presents results concerning variance component estimates. The third one involves high dimensional data. (pdf file)

Projected Iterated Normed Gradient (PING) algorithm: We give some technical details about the PING algorithm, which maximises, at least locally, any criterion on the unit sphere. (pdf file)

R package `mixedSCGLR`: We provide an R package to perform mixed–SCGLR, also available at `https://github.com/SCnext/mixedSCGLR`. It contains the dataset *Genus* used in Section 6. The package also provides demo codes, in particular for visualising the component planes (mixedSCGLR.tar.gz).

Code for running simulations: We also provide the R codes required to reproduce most of the simulation results (R and Rdata files).

ACKNOWLEDGMENTS

The extended data *Genus* required the arrangement and the inventory of 140.000 developed plots across four countries : Central African Republic, Gabon, Cameroon and Democratic Republic of Congo. The authors thank the members of the CoForTips project for allowing the use of this data. We are also grateful to the editor, the associate editor and to the referees for their thorough and constructive review of this work.

Supplementary file to "Component–based regularisation of multivariate generalised linear mixed models": THE PING ALGORITHM

Jocelyn Chauvet[1] **Catherine Trottier**[1,2] **Xavier Bry**[1]

[1] IMAG, Univ Montpellier, CNRS, Montpellier, France.
jocelyn.chauvet@umontpellier.fr ; xavier.bry@umontpellier.fr
[2] Univ Paul-Valéry Montpellier 3, Montpellier, France.
catherine.trottier@univ-montp3.fr

The Projected Iterated Normed Gradient (PING) is a basic extension of the iterated power algorithm, for solving any program of the form

$$\begin{cases} \max \quad \mathcal{J}_h\left(\boldsymbol{u}\right), \\ \text{subject to:} \quad \boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{u} = 1 \text{ and } \boldsymbol{\Delta}_{\boldsymbol{h}}^{\mathrm{T}}\boldsymbol{u} = \boldsymbol{0}. \end{cases} \tag{16}$$

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 $\boldsymbol{B_h} = \boldsymbol{M}^{1/2}\boldsymbol{\Delta_h}$, Program (16) is strictly equivalent to Program (17):

$$\begin{cases} \max \quad \mathcal{G}_h\left(\boldsymbol{v}\right), \\ \text{subject to:} \quad \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = 1 \text{ and } \boldsymbol{B}_{\boldsymbol{h}}^{\mathrm{T}}\boldsymbol{v} = \boldsymbol{0}. \end{cases} \tag{17}$$

Denoting

$$\Pi_{\mathrm{span}\{\boldsymbol{B_h}\}^{\perp}} = \boldsymbol{I} - \boldsymbol{B_h}\left(\boldsymbol{B}_{\boldsymbol{h}}^{\mathrm{T}}\boldsymbol{B_h}\right)^{-1}\boldsymbol{B}_{\boldsymbol{h}}^{\mathrm{T}} \quad \text{and}$$
$$\Gamma_h\left(\boldsymbol{v}\right) = \underset{\boldsymbol{v}}{\nabla}\,\mathcal{G}_h\left(\boldsymbol{v}\right),$$

a Lagrange multiplier-based reasoning gives the basic iteration of the PING algorithm:

$$\boldsymbol{v}^{[t+1]} = \frac{\Pi_{\mathrm{span}\{\boldsymbol{B_h}\}^{\perp}}\,\Gamma_h\left(\boldsymbol{v}^{[t]}\right)}{\left\|\Pi_{\mathrm{span}\{\boldsymbol{B_h}\}^{\perp}}\,\Gamma_h\left(\boldsymbol{v}^{[t]}\right)\right\|}. \tag{18}$$

Although Iteration (18) follows a direction of ascent, it does not guarantee that $\mathcal{G}_h$ actually increases on every step. We therefore propose a generic iteration of PING (Algorithm 2) and an alternative one (Algorithm 3), which both ensure that the criterion increases.

while *convergence of* $\boldsymbol{v}$ *non reached* **do**

$$\boldsymbol{\kappa}^{[t]} = \frac{\Pi_{\text{span}\{\boldsymbol{B_h}\}^\perp} \Gamma_h\left(\boldsymbol{v}^{[t]}\right)}{\left\|\Pi_{\text{span}\{\boldsymbol{B_h}\}^\perp} \Gamma_h\left(\boldsymbol{v}^{[t]}\right)\right\|}$$

A unidimensional Newton–Raphson 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 2: Generic iteration of the PING algorithm

while *convergence of* $\boldsymbol{v}$ *non reached* **do**

$$\boldsymbol{m} \longleftarrow \frac{\Pi_{\text{span}\{\boldsymbol{B_h}\}^\perp} \Gamma_h\left(\boldsymbol{v}^{[t]}\right)}{\left\|\Pi_{\text{span}\{\boldsymbol{B_h}\}^\perp} \Gamma_h\left(\boldsymbol{v}^{[t]}\right)\right\|}$$

while $\mathcal{G}_h(\boldsymbol{m}) < \mathcal{G}_h\left(\boldsymbol{v}^{[t]}\right)$ **do**

$$\boldsymbol{m} \longleftarrow \frac{\boldsymbol{v}^{[t]} + \boldsymbol{m}}{\|\boldsymbol{v}^{[t]} + \boldsymbol{m}\|}$$

end

$\boldsymbol{v}^{[t+1]} = \boldsymbol{m}$

$t \longleftarrow t+1$

end

Algorithm 3: Alternative generic iteration of the PING algorithm

First rank component. Component $\boldsymbol{f_1} = \boldsymbol{X}\boldsymbol{u_1}$ is obtained by solving

$$\begin{cases} \max & s \log\left[\phi\left(\boldsymbol{u}\right)\right] + (1-s) \log\left[\psi_{\boldsymbol{A}}\left(\boldsymbol{u}\right)\right] \\ \text{subject to:} & \boldsymbol{u}^{\mathrm{T}} \boldsymbol{M}^{-1} \boldsymbol{u} = 1. \end{cases}$$

This corresponds to Program (16) with $h = 0$, where

- $\mathcal{J}_0\left(\boldsymbol{u}\right) = s \log\left[\phi\left(\boldsymbol{u}\right)\right] + (1-s) \log\left[\psi_{\boldsymbol{A_0}}\left(\boldsymbol{u}\right)\right]$,
- $\boldsymbol{A_0} = \boldsymbol{A}$ (the matrix of additional explanatory variables), and
- $\boldsymbol{\Delta_0} = \boldsymbol{0}$.

In this particular case, we have $\boldsymbol{B_0} = \boldsymbol{M}^{1/2}\boldsymbol{\Delta_0} = \boldsymbol{0}$, and so:

$$\Pi_{\mathrm{span}\{\boldsymbol{B_0}\}^{\perp}} = \boldsymbol{I}.$$

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

$$\begin{cases} \max & s \log\left[\phi\left(\boldsymbol{u}\right)\right] + (1-s) \log\left[\psi_{\boldsymbol{A_h}}\left(\boldsymbol{u}\right)\right] \\ \text{subject to:} & \boldsymbol{u}^{\mathrm{T}} \boldsymbol{M}^{-1} \boldsymbol{u} = 1 \text{ and } \boldsymbol{F_h}^{\mathrm{T}} \boldsymbol{P} \boldsymbol{X} \boldsymbol{u} = \boldsymbol{0}. \end{cases}$$

This corresponds to Program (16), where

- $\mathcal{J}_h\left(\boldsymbol{u}\right) = s \log\left[\phi\left(\boldsymbol{u}\right)\right] + (1-s) \log\left[\psi_{\boldsymbol{A_h}}\left(\boldsymbol{u}\right)\right]$,
- $\boldsymbol{A_h} = \left[\, \boldsymbol{F_h} \mid \boldsymbol{A} \,\right]$, and
- $\boldsymbol{\Delta_h} = \boldsymbol{X}^{\mathrm{T}} \boldsymbol{P} \boldsymbol{F_h}$.

Initialisation. To quickly find $\boldsymbol{f_1}$, algorithm PING is initialised with the first PLS component of the responses on $\boldsymbol{X}$. In like manner, for $h \geqslant 2$, PING is initialised with the first PLS component of the responses on $\boldsymbol{X}$ deflated on components $\boldsymbol{F_{h-1}} = \left[\, \boldsymbol{f_1} \mid \ldots \mid \boldsymbol{f_{h-1}} \,\right]$.

Supplementary file to "Component–based regularisation of multivariate generalised linear mixed models": ADDITIONAL SIMULATIONS

Jocelyn Chauvet[1] **Catherine Trottier**[1,2] **Xavier Bry**[1]

[1] IMAG, Univ Montpellier, CNRS, Montpellier, France.
jocelyn.chauvet@umontpellier.fr ; xavier.bry@umontpellier.fr
[2] Univ Paul-Valéry Montpellier 3, Montpellier, France.
catherine.trottier@univ-montp3.fr

8 Comparative results with binomial and Poisson outcomes

In this section, we simply extend the simulation scheme presented in Section 5.5 to binomial and Poisson outcomes. We maintain design matrices $\boldsymbol{X}$ and $\boldsymbol{U}$ as defined in Section 5.1. Fixed–effect parameters $\boldsymbol{\theta_k}$'s and random–effect vectors $\widetilde{\boldsymbol{\xi}}_{\boldsymbol{k}}$'s are defined in Section 5.5. Given $\widetilde{\boldsymbol{\xi}}_{\mathbf{1}}$ and $\widetilde{\boldsymbol{\xi}}_{\mathbf{2}}$, we then simulate $\boldsymbol{Y} = \left[\, \boldsymbol{y_1} \mid \boldsymbol{y_2} \,\right]$ as

$$\begin{cases} \boldsymbol{y_1} \sim \mathcal{B}\text{in}\left(\textbf{trials} = 50\,\mathbf{1}_n,\ \boldsymbol{p} = \text{logit}^{-1}\left[\boldsymbol{X\theta_1} + \boldsymbol{U}\widetilde{\boldsymbol{\xi}}_{\mathbf{1}}\right]\right) \\ \boldsymbol{y_2} \sim \mathcal{P}\left(\boldsymbol{\lambda} = \exp\left[\boldsymbol{X\theta_2} + \boldsymbol{U}\widetilde{\boldsymbol{\xi}}_{\mathbf{2}}\right]\right). \end{cases}$$

Table 5 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 in Table 5 are essentially identical to those in the article: mixed–SCGLR outperforms GLMM–LASSO (Groll, 2017, R package `glmmLasso`) except for $\tau = 0.1$. As for the binomial distribution, the regularisation provided by mixed–SCGLR improves the results obtained without regularisation (Bates et al., 2015, R package `lme4`), regardless of the level of redundancy within the explanatory variables. Unsurprisingly, the errors are much smaller than in the binary case.

Table 5: Mean Relative Squared Error (MRSE) values obtained with binomial and Poisson distributions. The R package `glmmLasso` does not handle binomial outcomes but only Bernoulli ones, which precludes comparison in this case.

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

The power of mixed–SCGLR in terms of model interpretation remains the same for non–Gaussian outcomes. Figure 6 (respectively Figure 7) presents an example of the first component planes output by mixed–SCGLR in the binomial/Poisson (respectively Bernoulli/Poisson) case. As for Gaussian outcomes, the component planes reveal that $\boldsymbol{y_1}$ is explained by bundle $\boldsymbol{X_1}$ and $\boldsymbol{y_2}$ by $\boldsymbol{X_2}$. In the binomial/Poisson case with $\tau = 0.3$ (Figure 6), predictive bundles $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are captured respectively by the first and the second components. The third component aligns on nuisance bundle $\boldsymbol{X_0}$, despite its high inertia. Figure 7 illustrates what may happen when the level of redundancy is very high ($\tau = 0.7$ here). Since the explanatory variables are highly correlated, 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 be paid for the trade–off. In our example, the second explanatory bundle is captured by the third component, so that the predictive dimensions are accurately represented in component plane $(1, 3)$.

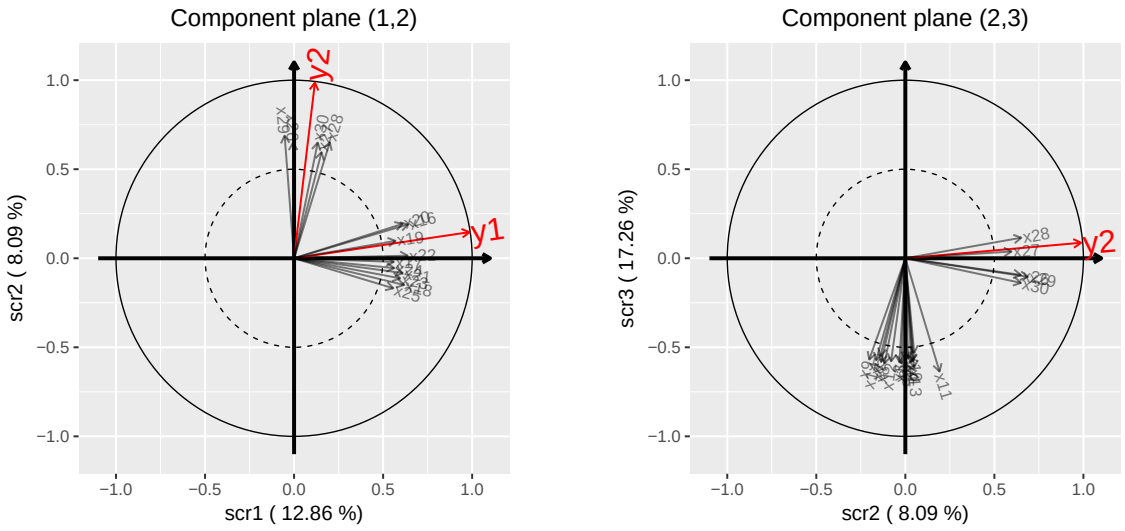

Figure 6: Example of component planes given by mixed–SCGLR in the binomial/Poisson case for $\tau = 0.3$, with parameter triplet $(K, s, l) = (3, 0.5, 2)$.

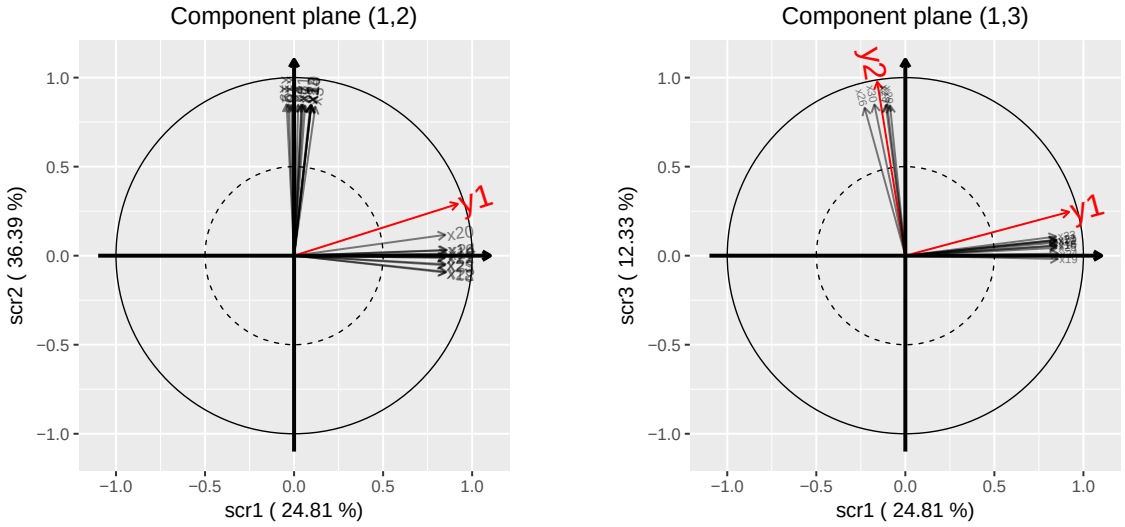

Figure 7: Example of component planes given by mixed–SCGLR in the Bernoulli/Poisson case for $\tau = 0.7$, with parameter triplet $(K, s, l) = (3, 0.9, 4)$.

9 A new structure for the bundles — Gaussian, Poisson and binomial distributions

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 $j \in \{12,\ldots,31\}$, $\boldsymbol{x_j} \sim \mathcal{N}_n\left(\mathbf{0}, \boldsymbol{I}_n\right)$. For each $k \in \{1,2,3\}$, random–effect vectors are simulated as $\boldsymbol{\xi_k} \overset{\text{ind.}}{\sim} \mathcal{N}_N\left(\mathbf{0}, \sigma^2_k\, \boldsymbol{I}_N\right)$. 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

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

In our simulations, we set $\alpha_1 = \sigma^2_1 = 2$, $\alpha_2 = \sigma^2_2 = 1$ and $\alpha_3 = \sigma^2_3 = 0.5$.

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 variable $\boldsymbol{\varphi_1}$ and predictive variable $\boldsymbol{\varphi_2}$. For $j = 1$ and 2, we then define

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

where $\boldsymbol{f}_{\boldsymbol{j}}^{(b)}$ 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 6 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 concerned, the biases are getting slightly closer to 0 and the standard errors decrease significantly.

Model interpretation is revealed by Figure 8 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 $\boldsymbol{\mathcal{G}_2}$.

Table 6: 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
$\text{bias}\left(\widehat{\sigma^2_1}\right)$	−0.02	−0.01	0.00	0.02	0.00	−0.02
$\text{sd}\left(\widehat{\sigma^2_1}\right)$	1.04	1.05	1.06	0.41	0.40	0.39
$\text{bias}\left(\widehat{\sigma^2_2}\right)$	−0.11	−0.08	−0.06	−0.06	−0.06	−0.06
$\text{sd}\left(\widehat{\sigma^2_2}\right)$	0.50	0.51	0.52	0.21	0.21	0.21
$\text{bias}\left(\widehat{\sigma^2_3}\right)$	−0.03	−0.04	−0.04	−0.02	−0.02	−0.02
$\text{sd}\left(\widehat{\sigma^2_3}\right)$	0.22	0.21	0.21	0.11	0.11	0.11

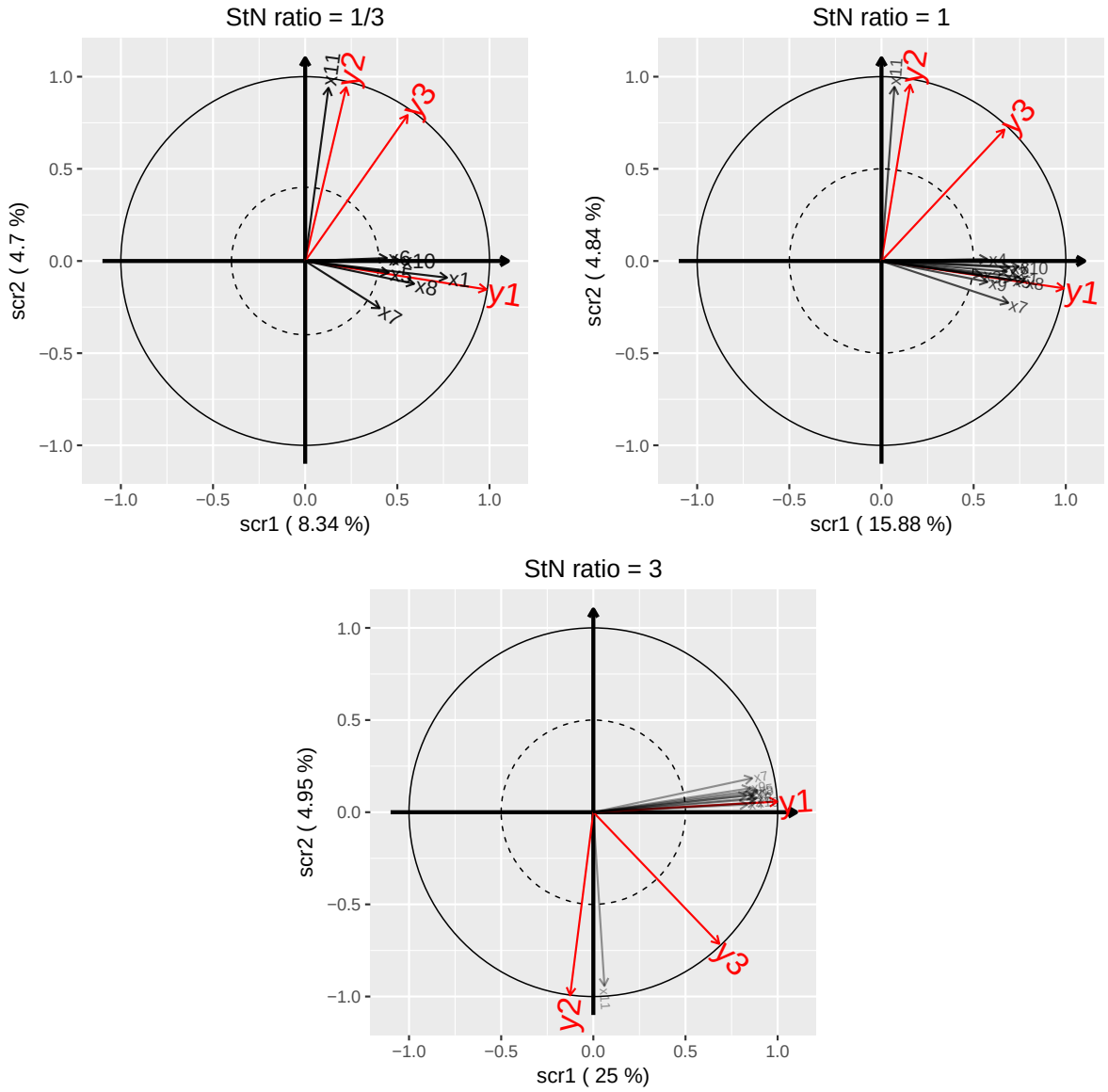

Figure 8: Examples of the first two–component planes given by mixed–SCGLR when $\sigma^2_{\text{LV}}/\sigma^2_{\text{noise}} = 1/3$ (top left), $\sigma^2_{\text{LV}}/\sigma^2_{\text{noise}} = 1$ (top right), and $\sigma^2_{\text{LV}}/\sigma^2_{\text{noise}} = 3$ (bottom). When StN ratio $= 1/3$ (resp. StN ratio $\in \{1, 3\}$), only the variables having cosine greater than 0.4 (resp. 0.5) with component plane $(1, 2)$ are represented.

10 High dimensional data

10.1 Key idea

To cope with high dimensional data, the key idea is to replace the fixed–effect design matrix, $\boldsymbol{X}$, with the matrix $\boldsymbol{C}$ of its principal components associated with non–zero eigenvalues. More precisely, λ_j being 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 $\boldsymbol{C} = \boldsymbol{XV}$. The component $\boldsymbol{f}$ is then sought as a combination of the principal components: $\boldsymbol{f} = \boldsymbol{Cu} = \boldsymbol{X}\widetilde{\boldsymbol{u}}$, where $\widetilde{\boldsymbol{u}} = \boldsymbol{Vu}$. 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 by

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

and the structural relevance by

$$\phi \left(\boldsymbol{u} \right) = \left[\sum_{j=1}^{p} \omega_j \left(\left\langle \, \boldsymbol{Cu} \, \middle| \, \boldsymbol{x^j} \, \right\rangle^2_{\boldsymbol{P}} \right)^l \right]^{\frac{1}{l}} = \left[\sum_{j=1}^{p} \omega_j \Big(\boldsymbol{u}^{\mathrm{T}} \boldsymbol{C}^{\mathrm{T}} \boldsymbol{P} \boldsymbol{x_j} \boldsymbol{x_j}^{\mathrm{T}} \boldsymbol{P} \boldsymbol{C} \, \boldsymbol{u} \Big)^l \right]^{\frac{1}{l}}.$$

This idea is tested on simulated data where the number of explanatory variables p exceeds the number of observations n.

10.2 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 effects' design matrix is then $\boldsymbol{U} =$

$\boldsymbol{I}_N \otimes \boldsymbol{1}_R$. Explanatory variables consist of four independent bundles $\boldsymbol{X_j}, j \in \{1,2,3,4\}$, such as $\boldsymbol{X} = \left[\, \boldsymbol{X_0} \mid \boldsymbol{X_1} \mid \boldsymbol{X_2} \mid \boldsymbol{X_3} \,\right]$. Each explanatory variable is 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

$$\operatorname{cor}(\boldsymbol{X_j}) = \tau \boldsymbol{1}_{p_j} \boldsymbol{1}_{p_j}^{\mathrm{T}} + (1-\tau)\boldsymbol{I}_{p_j},$$

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 $\boldsymbol{\xi_k} \overset{\text{ind.}}{\sim} \mathcal{N}_N(\boldsymbol{0}, \sigma_k^2 \boldsymbol{I}_N)$. Given $\boldsymbol{\xi_1}$, $\boldsymbol{\xi_2}$, $\boldsymbol{\xi_3}$, $\boldsymbol{\xi_4}$, we simulate 4 responses having different distributions, $\boldsymbol{Y} = \left[\, \boldsymbol{y_1} \mid \boldsymbol{y_2} \mid \boldsymbol{y_3} \mid \boldsymbol{y_4} \,\right]$, as

$$\begin{cases} \boldsymbol{y_1} \sim \mathcal{N}_n\Big(\boldsymbol{\mu} = \boldsymbol{X\beta_1} + \boldsymbol{U\xi_1},\, \boldsymbol{\Sigma} = \boldsymbol{I}_n\Big) \\ \boldsymbol{y_3} \sim \mathcal{B}\Big(\boldsymbol{p} = \operatorname{logit}^{-1}\Big[\boldsymbol{X\beta_2} + \boldsymbol{U\xi_2}\Big]\Big) \\ \boldsymbol{y_3} \sim \mathcal{B}\mathrm{in}\Big(\mathbf{trials} = 30\,\boldsymbol{1}_n,\, \boldsymbol{p} = \operatorname{logit}^{-1}\Big[\boldsymbol{X\beta_3} + \boldsymbol{U\xi_3}\Big]\Big) \\ \boldsymbol{y_4} \sim \mathcal{P}\Big(\boldsymbol{\lambda} = \exp\Big[\boldsymbol{X\beta_4} + \boldsymbol{U\xi_4}\Big]\Big). \end{cases} \tag{19}$$

Response $\boldsymbol{y_1}$ is predicted only by bundle $\boldsymbol{X_1}$, $\boldsymbol{y_2}$ only by bundle $\boldsymbol{X_2}$, $\boldsymbol{y_3}$ only by bundle $\boldsymbol{X_3}$, $\boldsymbol{y_4}$ by both bundles $\boldsymbol{X_2}$ and $\boldsymbol{X_3}$, and bundle $\boldsymbol{X_0}$ plays no explanatory role. Our choice for the fixed–effect parameters is

$$\boldsymbol{\beta_1} = (\underbrace{0,\ldots,0}_{p_0 \text{ times}}, \underbrace{0.1,\ldots 0.1}_{p_1 \text{ times}}, \underbrace{0,\ldots,0}_{p_2 \text{ times}}, \underbrace{0,\ldots,0}_{p_3 \text{ times}})^{\mathrm{T}},$$

$$\boldsymbol{\beta_2} = (\underbrace{0,\ldots,0}_{p_0 \text{ times}}, \underbrace{0,\ldots 0}_{p_1 \text{ times}}, \underbrace{0.1,\ldots,0.1}_{p_2 \text{ times}}, \underbrace{0,\ldots,0}_{p_3 \text{ times}})^{\mathrm{T}},$$

$$\boldsymbol{\beta_3} = (\underbrace{0,\ldots,0}_{p_0 \text{ times}}, \underbrace{0,\ldots 0}_{p_1 \text{ times}}, \underbrace{0,\ldots,0}_{p_2 \text{ times}}, \underbrace{0.05,\ldots,0.05}_{p_3 \text{ times}})^{\mathrm{T}},$$

$$\boldsymbol{\beta_4} = (\underbrace{0,\ldots,0}_{p_0 \text{ times}}, \underbrace{0.025,\ldots 0.025}_{p_1 \text{ times}}, \underbrace{0.025,\ldots,0.025}_{p_2 \text{ times}}, \underbrace{0,\ldots,0}_{p_3 \text{ times}})^{\mathrm{T}}$$

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. 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 (19).

10.3 Results

Table 7 and Table 8 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.

Table 7: Mean Relative Squared Error (MRSE) values for fixed–effect estimates, and biases and standard errors for estimated variance components, obtained with 100 observations and 150 explanatory variables.

	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_4}$	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	−0.03	−0.02	0.02
					0.09	0.09	0.03	0.06
$\tau = 0.5$	0.03	0.20	0.10	0.07	0.01	−0.03	0.00	0.01
					0.11	0.08	0.07	0.07
$\tau = 0.7$	0.01	0.10	0.05	0.04	0.01	−0.05	0.01	0.02
					0.07	0.09	0.10	0.07

Table 8: Mean Relative Squared Error (MRSE) values for fixed–effect estimates, and biases and standard errors for estimated variance components, obtained with 100 observations and 200 explanatory variables.

	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_4}$	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	−0.05	0.01	−0.02
					0.04	0.09	0.05	0.05
$\tau = 0.5$	0.03	0.17	0.09	0.05	−0.05	0.00	−0.02	−0.01
					0.06	0.19	0.04	0.04
$\tau = 0.7$	0.01	0.15	0.04	0.03	0.03	0.00	−0.01	−0.02
					0.08	0.14	0.05	0.05

Some component planes are given on Figure 9 (150 explanatory variables) and Figure 10 (200 explanatory variables).

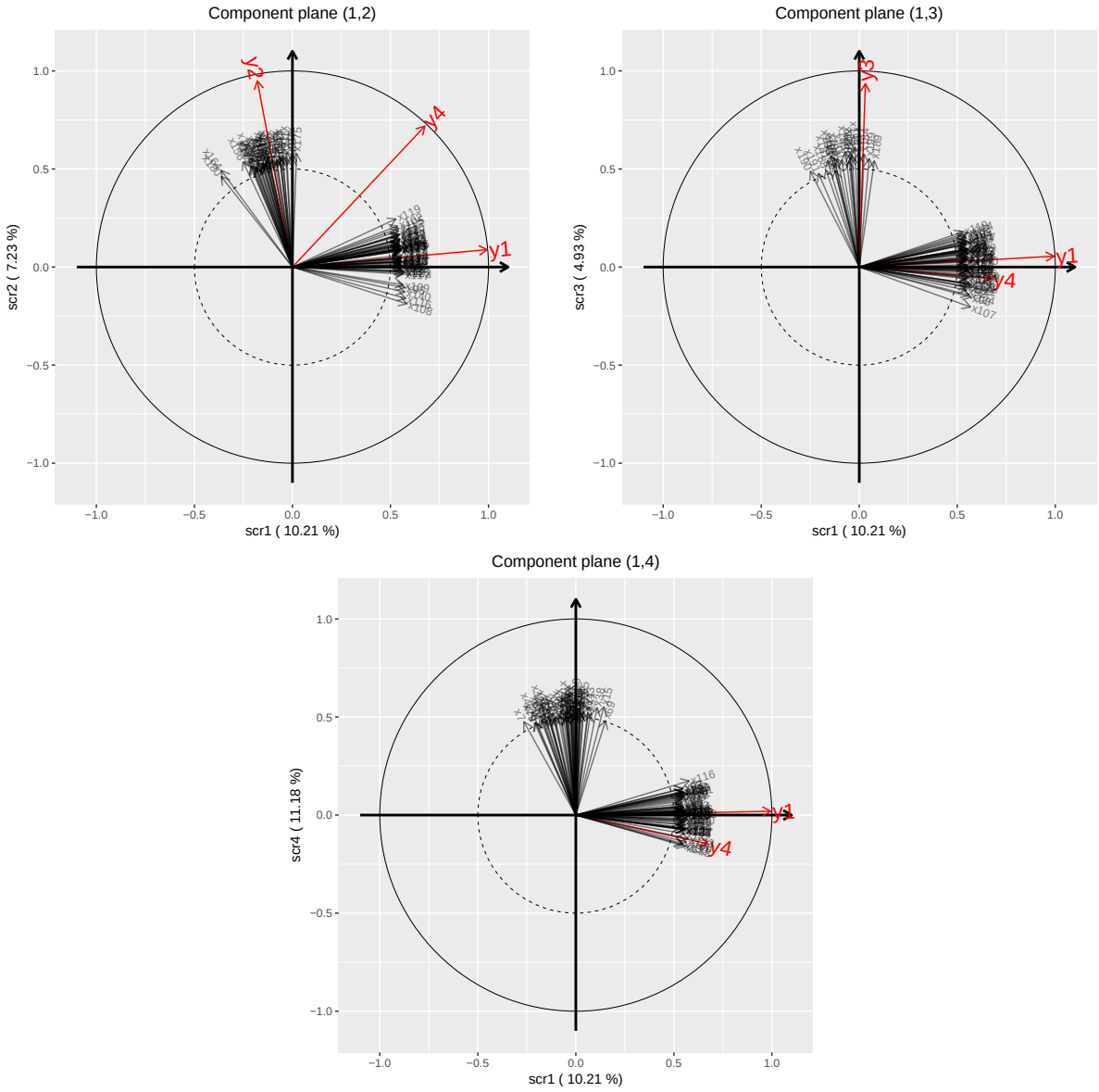

Figure 9: Component planes $(1, 2)$, $(1, 3)$ and $(1, 4)$ given by mixed–SCGLR for 100 observations, 150 explanatory variables and $\tau = 0.3$. The tuning parameter triplet (K, s, l) is set to $(4, 0.5, 4)$.

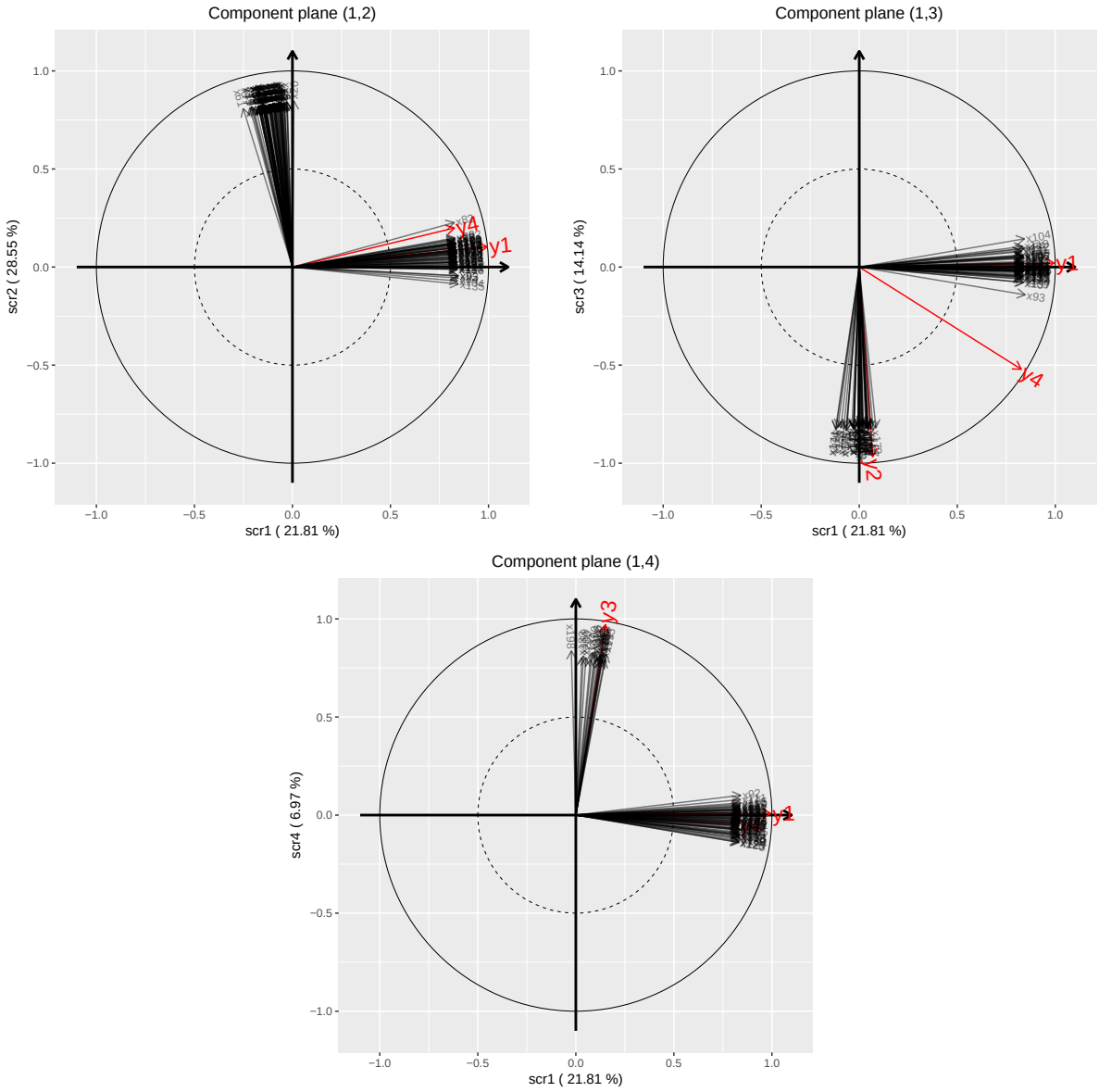

Figure 10: Component planes $(1, 2)$, $(1, 3)$ and $(1, 4)$ given by mixed–SCGLR for 100 observations, 200 explanatory variables, and $\tau = 0.9$. The tuning parameter triplet (K, s, l) is set to $(4, 0.9, 4)$.

References

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.

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.

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.

Bry, X., Trottier, C., Mortier, F., and Cornu, G. (2018). Component-based regularisation of a multivariate GLM with a thematic partitioning of the explanatory variables. *Statistical Modelling*. In press.

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.

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.

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.

Bry, X. and Verron, T. (2015). THEME: THEmatic Model Exploration through multiple co-structure maximization. *Journal of Chemometrics*, 29(12):637–647.

Clayton, D. G. (1996). Generalized linear mixed models. In *Markov chain Monte Carlo in practice*, pages 275–301. Springer.

Cornu, G., Mortier, F., Trottier, C., and Bry, X. (2018). *SCGLR: Supervised Component Generalized Linear Regression*. R package version 3.0.

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.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Series in Statistics. Springer-Verlag.

Groll, A. (2017). *glmmLasso: Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation*. R package version 1.5.1.

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.

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.

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics*, 31(2):423–447.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67.

Knudson, C. (2016). *Monte Carlo Likelihood Approximation for Generalized Linear Mixed Models*. PhD thesis, University of Minnesota.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

McCulloch, C. E. (1997). Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 92(437):162–170.

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.

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.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schall, R. (1991). Estimation in Generalized Linear Models with Random Effects. *Biometrika*, 78(4):719–727.

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.

Shun, Z. and McCullagh, P. (1995). Laplace Approximation of High Dimensional Integrals. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(4):749–760.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1):267–288.

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.

Zhang, Y., Zhou, H., Zhou, J., and Sun, W. (2017). Regression Models for Multivariate Count Data. *Journal of Computational and Graphical Statistics*, 26(1):1–13.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B (Methodological)*, 67(2):301–320.