Component-based regularisation of a multivariate GLM with a thematic partitioning of the explanatory variables

Xavier Bry, Catherine Trottier, Frédéric Mortier, Guillaume Cornu

Component-based regularisation of a multivariate GLM with a thematic partitioning of the explanatory variables

Xavier Bry [1], Catherine Trottier [1, 2], Frédéric Mortier [3], and Guillaume Cornu [3]

[1] Institut Montpelliérain Alexander Grothendieck, Université de Montpellier, Place Eugène Bataillon CC 051 - 34095, Montpellier, France

[2] Université Paul-Valéry Montpellier 3, Route de Mende - 34199, Montpellier, France

[3] Cirad, UPR F&S, Forests and Societies, Campus International de Baillarguet, TA C-105 / D - 34398, Montpellier, France.

Address for correspondence: Catherine Trottier, Institut Montpelliérain Alexander Grothendieck, Université Montpellier, Place Eugène Bataillon CC 051 - 34095, Montpellier, France.

E-mail: catherine.trottier@univ-montp2.fr.

Phone: (+33) 467 144 164.

Fax: (+33) 467 143 558.

Abstract: We address component-based regularisation of a multivariate Generalised Linear Model (GLM). A set of random responses Y is assumed to depend, through a GLM, on a set X of explanatory variables, as well as on a set A of addi-

tional covariates. X is partitioned into R conceptually homogenous variable groups $X_1, \ldots, X_R$, viewed as explanatory themes. Variables in each X_r are assumed many and redundant. Thus, generalised linear regression demands dimension-reduction and regularisation with respect to each X_r. By contrast, variables in A are assumed few and selected so as to demand no regularisation. Regularisation is performed searching each X_r for an appropriate number of orthogonal components that both contribute to model Y and capture relevant structural information in X_r. To estimate a single-theme model, we first propose an enhanced version of Supervised Component Generalised Linear Regression (SCGLR), based on a flexible measure of structural relevance of components, and able to deal with mixed-type explanatory variables. Then, to estimate the multiple-theme model, we develop an algorithm encapsulating this enhanced SCGLR: THEME-SCGLR. The method is tested on simulated data, and then applied to rainforest data in order to model the abundance of tree-species.

Key words: Components; Multivariate Generalised Linear Model; Regularisation; SCGLR; Dimension reduction.

1 Introduction

1.1 Framework

Our framework is that of a multivariate GLM (Fahrmeir and Tutz, 1994) with multiple responses and a high number of covariates partitioned into several thematic groups, hence referred to as *themes*. This high number, with possibly high collinearities within

each theme, demands that dimension-reduction and regularisation (i.e. stabilisation of regressor-coefficients) be performed during the GLM estimation. Indeed, standard GLM estimation keeping correlated covariates is bound to lead to overfitting and yield a highly unstable (if identified) linear predictor. Here, dimension-reduction is to be carried out within each theme, so that each extracted dimension refers to a specific theme.

The data we want to both explore and model is floristic data consisting of the abundance (measured through counts) of $q = 27$ tree genera on $n = 1000$ plots in the tropical forest of the Congo-Basin (see subsection 5.2 for details). We want to model and explain these counts using several thematic blocks of variables, namely: one containing $p_1 = 23$ enhanced vegetation indices (EVI), one containing $p_2 = 13$ pluviometric indicators, and finally, a categorical variable describing the geology of plots. The pluviometric indicators exhibit high-correlation patterns, and so do the EVI's. So, dimension reduction and regularisation are required in the first two themes. By contrast, we want to keep the geological type variable as such in the model. Ultimately, we want to extract from pluviometric indicators and EVI's respectively, few reliable and interpretable dimensions which best complement geology in modelling and explaining the abundances of the 27 tree-genera. In a first stage, these 27 response variables can be assumed to be Poisson random variables, independent conditional on the explanatory variables.

The approach we propose is component-based: through components (i.e. linear combinations of the covariates), we want to explore, decompose in a synthetic way, and interpret the multidimensional explanatory and predictive potential of each explanatory theme. In order to be clearly interpretable, a component has to align with at

least some variables in its theme. The more variables it aligns with, the "stronger" the component, in that: 1) it is corroborated by more measures, so captures a "well-documented information", and 2) through component-based regularisation, many correlated variables end up stabilising the part of the linear predictor which is related to them.

Within each theme, components are wanted to extract the information that is useful to predict the responses, when associated with the components of the other themes. Orthogonal components within a theme make it easy to graph the covariates through correlation scatterplots, reveal the multidimensional structure of the explanatory information they contain, and facilitate its interpretation. This is one reason why we prefer this approach to classical penalty-based approaches as Ridge, LASSO or Elastic net (Tibshirani (1996), Tikhonov and Arsenin (1977), Zou and Hastie (2005)), which do provide a regularised linear predictor, but no decomposition of it on a reduced set of interpretable dimensions. Yet, the penalty-coefficient of the latter methods is a continuous tuning-parameter, and this is a facility we would like to have.

The first work dealing with component-based regularisation of a GLM was that of Marx (1996), who introduced the PLS mechanism into the Fisher Scoring Algorithm (FSA) of a univariate GLM, leading to the Iteratively Re-weighted Partial Least Squares (IRPLS) method. In his wake, Bry et al. (2013) have introduced a component-based technique named Supervised Component-based Generalised Linear Regression (SCGLR), which extends both the work by Marx and multivariate PLS Regression, and contains an additional continuous regularisation-tuning parameter. Like IRPLS, SCGLR performs component-based regularisation of the model within the FSA. The interest of operating at FSA level is that estimation weights keep consistent with the

component-model being estimated. Fort and Lambert-Lacroix (2005) have proposed to sequentially (as opposed to iteratively) combine PLS regression with a penalised estimation of a logistic model, in two steps: on step 1, a ridge-penalised logistic regression of the binary response y is carried out on a set of covariates X so as to yield a pseudo-response z (aka "working variable"). On step 2, PLS regression of z on X is carried out, yielding explanatory components. More recently, Durif et al. (2015) have extended this technique to explanatory variable selection by replacing the PLS step with a Sparse PLS one. To us, this 2-step approach has two assets and one drawback. The first asset is that it goes beyond mere regularisation, which yields a "good" linear predictor, but fails to decompose it on interpretable directions such as strong components. The second asset is that limiting the model estimation to a 2-step-sequence clearly limits convergence problems to those of ridge GLM estimation (since PLS has none). Yet, we think that this is paid for by a notable theoretical drawback: the aim being the estimation of a model, it seems to us that once explanatory directions (be they components or selected original variables) are taken as input, estimation demands that the pseudo-responses be recalculated accordingly. This is what every 2-step approach (performing regularisation of GLM estimation prior to component-search) fails to do, by calculating pseudo-responses once and for all on the first step. By contrast, SCGLR, just as IRPLS did, recalculates the pseudo-responses and the estimation weights each time a component has been updated, thus keeping model estimation, i.e. estimation weights, consistent with the current model inputs. Practically, of course, the estimation weights may have a rather marginal impact, and the 2-step approach may be advocated for its simplicity. Nevertheless, we must add that we did not encounter convergence problems with SCGLR in our numeric experiments on gaussian and Poisson data (provided zero counts are not too may, in

the latter case).

Now, SCGLR, in its original version, does have limitations which the present work aims at overcoming. Firstly, we want to be allowed to specify the type of relevant structure we would like our components to align on: principal components, variable-bundles, original variables, subspaces based on interpretable constraints, etc. To that aim, we propose to include in a new SCGLR algorithm the Structural Relevance (SR) measure proposed by Bry and Verron (2015), which extends the component's variance used in IRPLS and the original SCGLR in a general and flexible way. Secondly, we also want to address mixed-type (i.e. quantitative and categorical) covariates. Thirdly, in many models, some covariates, which we shall term "additional covariates", have been included after some pre-processing precluding redundancy with the others, and owing to the particular attention they should get. Such covariates are bound to be few, their regression coefficients demand no regularisation, and dimension reduction does not concern them. Thus, they should be dealt with accordingly in the model-building. Finally, SCGLR only considered one explanatory theme. The situation we want to be able to deal with is of a much greater practical interest to the modeller, as sketched in the description of the floristic data, the covariates have been partitioned into several themes, each of them being conceptually homogenous. By looking for components within themes, the modeller makes their interpretation conceptually easier and more natural. In the multiple-theme situation, components are wanted to *separate* the effects of the explanatory themes on the responses, revealing and synthesizing each theme's *partial* effect, if any. So that, between themes, components must remain unconstrained. That way, every theme is quite free to express its predictive potential along the others, like variables do in a classical regression model. This, of course, precludes collinearity between themes, i.e. between strong dimensions

pertaining to distinct themes. Besides, every theme must be given the same a-priori importance. Indeed, the "thematic model" is similar to a classical regression model, where every variable would be replaced with a thematic set of variables to be searched for interesting latent dimensions. In our example, we want to extract the most predictive pluviometric components (if any), and also vegetation index dimensions (if any), which can enter a model together, along with geology, to explain and predict tree abundances.

1.2 Model and issues

A set of q random variables $Y = \{y^1, \ldots, y^q\}$, referred to as *responses*, is assumed to be dependent on p common explanatory column-vectors which can be either numeric variables or indicator-variables of factor-levels, partitioned in R themes $X_1, \ldots, X_R$, with: $\forall r,\ X_r = \{x_r^1, ..., x_r^{p_r}\}$, plus one matrix A of additional covariates (which may also contain indicator variables). Every X_r may contain several unknown structurally relevant dimensions important to predict Y, how many we do not a priori know. Variables in A are assumed to have been selected in some way, so as to preclude instability of their estimated coefficients, while variables in X_r's have not. In other words, matrix A gathers all explanatory variables that we wish to keep as such: no dimension-reduction is to be performed within A, whereas dimension reduction is needed in the X_r's. Each X_r is thus to be searched for an appropriate number of orthogonal components that both capture relevant structural information in it and contribute to model Y.

Let $X := [X_1, ..., X_R]$. Each y^k is modelled through a GLM (McCullagh and Nelder, 1989), taking $X \cup A$ as covariate set. Moreover, $\{y^1, \ldots, y^q\}$ are assumed inde-

pendent conditional on X. All variables are measured on the same n statistical units. The conceptual model stating that variables in Y are dependent on variables in $\{X_1, ..., X_R\ ;\ A\}$, and that structurally relevant dimensions should be explicitly identified in the X_r's will be referred to as *thematic model* and denoted by symbolic equation: $Y = F(X_1) + ... + F(X_R) + A$.

We are currently interested in the following issues:

- Tracking down *structurally relevant* dimensions in spaces $Vec(X_r)$, that can ground a good explanatory and predictive model of Y. For the time being, let us informally say that a "structurally relevant" dimension (or, more shortly, *structural dimension*) in a theme X_r is one that accounts for "enough" variance in X_r.
- Regularisation of the out-coming model with respect to each X_r.

In this work, we propose to overcome the limitations of SCGLR by:

- introducing matrix A of additional covariates into SCGLR's model;
- extending the measure of structural strength of a component (as measured by its variance) to one of *structural relevance*, so as to track various kinds of structures;
- dealing with mixed-type covariates (i.e. numeric or categorical);
- extending SCGLR to the multiple-theme situation.

In the ordinary linear framework for thematic model $Y = F(X_1) + ... + F(X_R)$, Bry et al. (2009) have introduced a multi-theme extension of PLS regression named

Structural Equation Exploratory Regression, which was extended by Bry et al. (2012) and then to THEME by Bry and Verron (2015).

1.3 Plan of the paper and notations

Section 2 first adapts the Fisher Scoring Algorithm to the situation of multivariate responses when all responses depend on the same component in X. Then, it recalls the notion of Structural Relevance, which measures the ability of a component in a theme to capture strong and relevant structures within it. Eventually, it introduces the Structural Relevance into the FSA and designs the new SCGLR algorithm. Section 3 builds up THEME-SCGLR from this algorithm. Section 4 addresses model-assessment and component selection. Finally, in section 5, we study the performance of THEME-SCGLR on various simulated data structures, and then apply it to forest data.

For the sake of clarity, let us define some notations :

• u, v being two vectors pertaining to the same euclidean space endowed with metric M, their scalar product will be denoted $\langle u|v\rangle_M$, and their cosine, $\cos_M(u,v)$. The euclidean norm of u will be denoted $||u||_M$.

• $U = \{u^1, .., u^k\}, .., T = \{t^1, .., t^k\}$ being sets of vectors pertaining to the same space, the sub-space they span will be denoted $Vec(U, .., T)$.

2 An enhanced SCGLR

In this section, we consider the thematic equation with only one theme.

2.1 Multivariate GLM with partially common predictors

Ultimately, we want to find components common to all y's, i.e. on which to perform regression of every y^k. We thus consider a multivariate GLM where all responses have linear predictors collinear in their X-parts :

$$\forall k = 1, ..., q : \ \eta_k \ = \ Xu\gamma_k + A\delta_k \ .$$

For identification, as well as later taking the SR into account, we impose $u'M^{-1}u = 1$.

This model calls for adapting the FSA. Indeed, the FSA can be viewed as an iterated weighted regression on a linearised model, which reads on iteration $[t]$:

$$\forall k = 1, ..., q : \quad z_k^{[t]} = Xu\gamma_k + A\delta_k + \zeta_k^{[t]} \ , \tag{2.1}$$

where $z_k^{[t]}$ are the working variables and where the associated errors' variance matrix is denoted $V(\zeta_k^{[t]}) = W_k^{-1[t]}$.

In our context, model (2.1) is not linear, owing to the product $u\gamma_k$. So, it must be estimated through an alternated least squares step, estimating in turn γ_k and u. Let $f = Xu$ and $\Pi^k_{Vec(f,A)}$ be the projector onto $Vec(f, A)$ with respect to W_k . In view of the independence of responses conditional on predictors, the estimation of model (2.1) may be viewed as the solution of the following equivalent programs:

$$Q : \min_{f \in Vec(X)} \sum_k \|z_k - \Pi^k_{Vec(f,A)} z_k\|^2_{W_k} \quad \Longleftrightarrow \quad Q' : \max_{u \in \mathbb{R}^p \, , u'M^{-1}u=1} \psi(u) \ ,$$

where $\psi(u) = \sum_k \|z_k\|^2_{W_k} \cos^2_{W_k}(z_k; Vec(Xu, A))$. Note that program Q' extends that of a weighted Instrumental Variable PCA of Z on matrix $[X, A]$, the program of which we get by taking $\forall k, W_k = I$. ψ is a pure goodness-of-fit (GoF) measure, which is now to be aptly combined with some structural relevance measure to get regularisation.

2.2 Structural Relevance

Estimating the multivariate GLM with common predictors did not involve any notion of structural strength of the component: all directions in $Vec(X)$ were a priori equally important. But we would like to favour directions correlated to many variables as being "stronger", or directions close to known interpretable sub-spaces, as being "more relevant". The notion of Structural Relevance has been introduced by Bry and Verron (2015). Various measures of SR may be considered, according to the type of structure u (or f) should align with. Let weight-matrix W reflect the a priori relative importance of units (typically: $W = n^{-1}I_n$ for a uniform weighting). Let X be a $n \times p$ matrix associated with a theme, and M an associated $p \times p$ metric matrix in $\mathbb{R}^p$, the purpose of which is to "weight" X's variables appropriately. M may take various forms according to the type of variables and structure of data (Bry et al., 2012). Finally consider component $f = Xu$, where u is constrained by: $\|u\|^2_{M^{-1}} = 1$.

Let us recall the general formula of structural relevance. Given a set of J "reference" symmetric positive semi-definite matrices $N = \{N_j \ ; \ j = 1, ..., J\}$, a weight system $\Omega = \{\omega_j \ ; \ j = 1, ..., J\}$, and a scalar $l \geq 1$, the associated Structural Relevance (SR) measure is defined as the following function of u:

$$\phi(u) := \left(\sum_{j=1}^{J} \omega_j (u' N_j u)^l \right)^{\frac{1}{l}} . \tag{2.2}$$

Two particular examples of SR deserve mentioning.

- **Component Variance**

  X being composed of centred numeric variables, take:

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

This is the inertia of units along u, and is maximised by the first (direct) eigenvector in the PCA of (X, M, W). So here, M must be such that PCA of (X, M, W) is relevant. Consider a typical mixture of numeric and categorical variables: $X = \left[x^1, ..., x^K, X^1, ..., X^L\right]$, where: $x^1, ..., x^K$ are column-vectors coding the numeric regressors, and $X^1, ..., X^L$ are blocks of centred indicator variables, each block coding a categorical regressor (X^l has $q_l - 1$ columns if the corresponding variable has q_l levels, the removed level being taken as "reference level"). In order to get a relevant PCA of (X, M, W), we must consider the metric block-diagonal matrix:

$$M := \text{diag}\left\{(x^{1\prime}Wx^1)^{-1}, ..., (x^{K\prime}Wx^K)^{-1}, (X^{1\prime}WX^1)^{-1}, ..., (X^{L\prime}WX^L)^{-1}\right\} .$$

It is well known that this matrix bridges ordinary PCA of numeric variables with that of Multiple Correspondence Analysis.

- **Variable Powered Inertia (VPI)**

  Contrary to the component's variance, VPI involves parameter l. This parameter tunes the locality of the variable-bundles components should align on, in that increasing l results in raising the bonus given in the SR to directions close to local gatherings of variables (cf figure 1). Having components aligning better on a small number of highly correlated variables makes their interpretation easier, at the possible cost of some loss in generality, i.e. synthetic power of the component.

  We impose $\|f\|_W^2 = 1$ through $M = (X'WX)^{-1}$. Note that in order for $X'WX$ to be regular, X has to have full rank in column, which we will assume momentarily. For X consisting of p standardised numeric variables x^j, the VPI is

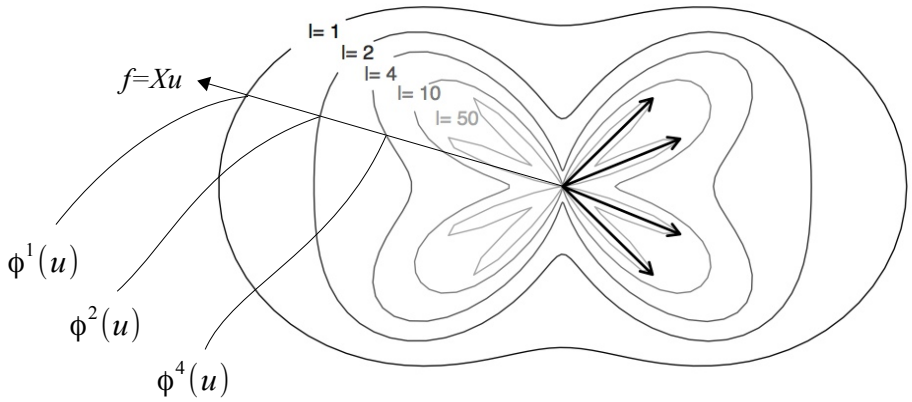


Figure 1: Polar representation of the Variable Powered Inertia according to the value of l

defined as:

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

In the elementary case of 4 coplanar variables x with $\forall j, \omega_j = 1$, Figure 1 graphs $\phi^l(u)$ in polar coordinates $(z(\theta) = \phi^l(e^{i\theta})e^{i\theta}\ ;\ \theta \in [0, 2\pi[)$ for various values of l. Note that $\phi^l(u)$ was graphed instead of $\phi(u)$ so that curves would be easier to read. One can see how the value of l tunes the locality of bundles considered: the greater the l, the more local the bundle.

For X consisting of p categorical variables X^j, each of which is coded through the set of its centred indicator variables (one being removed to avoid singularity

of $X^{j\prime}WX^j$), we will take:

$$\phi(u) = \left(\sum_{j=1}^{p} \omega_j \cos^{2l}_W(Xu, Vec(X^j))\right)^{\frac{1}{l}} = \left(\sum_{j=1}^{p} \omega_j \langle Xu|\Pi_{X^j} Xu\rangle^l_W\right)^{\frac{1}{l}} ,$$

where: $\Pi_{X^j} = X^j(X^{j\prime}WX^j)^{-1}X^{j\prime}W$.

Now, when $X'WX$ is singular, owing to collinearity, X should be replaced with the matrix C of its principal components associated with non-null eigenvalues, and the component sought as $f = Cu$. We have: $C = XV$, where V is the matrix of corresponding unit-eigenvectors. Then, $f = Cu = Xw$ with $w = Vu$. Supplemental material 7.5 shows that among all coefficient vectors t such that $Xt = f$, w is that which has the minimum L2-norm.

2.3 The enhanced SCGLR: new criterion and program

Introducing the SR into the current step of the algorithm given in Section 2.1, we now consider the program:

$$R: \max_{u'M^{-1}u=1} S(u), \quad \text{with } S(u) = \psi(u)^{1-s}\phi^s(u) , \tag{2.3}$$

where s is a parameter tuning the relative importance of the SR with respect to the GoF. Taking $s = 0$ equates the criterion with the GoF, while at the other end, taking $s = 1$ equates it with the mere SR. Thus, increasing s increases regularisation of the model ($s = 0$ is associated with no regularisation at all). This role is similar to that of the penalty-coefficient in penalty-based methods such as ridge and LASSO.

The product form in (2.3) may be advocated as follows:

$$\max_{u'M^{-1}u=1} \psi(u)^{1-s}\phi^s(u) \;\Leftrightarrow\; \max_{u'M^{-1}u=1} \left((1-s)\ln\psi(u) + s\ln\phi(u)\right) .$$

This implies that, at the maximum: $\langle du|(1-s)\nabla\ln\psi(u) + s\nabla\ln\phi(u)\rangle_{M^{-1}} = 0$, $\forall du$ tangent to the unit-sphere. So, at the maximum, *relative* variations of GoF and SR

compensate, with elasticity of ψ relative to ϕ being equal to $s/(1-s)$, which gives s a precise interpretation. Note that the product form of the criterion is a straightforward way to make the solution insensitive to "size effects" of $\phi(u)$ and $\psi(u)$. An analytical expression of $S(u)$ is derived in supplemental material 7.1.

2.4 Rank 1 component

Theme-SCGLR's rank 1 component is obtained as a solution of program R (2.3). In supplemental material 7.2, we give an algorithm to maximise, at least locally, any criterion on the unit-sphere: the Projected Iterated Normed Gradient (PING) algorithm. PING may be used to solve program R, with function $h(u) = S(u)$ or $\ln(S(u))$, and $D = 0$.

2.5 Rank $h > 1$ component

Components in X are to span, as extensively as possible, a subspace leaning on strong structures of X which is useful for predicting Y, when associated with the additional covariates A. This is what component 1 does when this subspace is constrained to have dimension 1. Because of this optimality, we choose to consider component 1 given, and are not going to reconsider it. So, component 2 will have to span, together with component 1, a 2-dimensional subspace leaning on strong structures of X and useful for predicting Y. Moreover, we want component 2 to be orthogonal to component 1 for graphing purposes. So, we shall look for a component 2 leaning on strong structures of X, orthogonal to component 1, and giving the best possible fit to the model under such constraints. And so on with higher-rank components. We thus adopt the local nesting principle (LocNes) presented in Bry et al. (2009) and extended in Bry et al.

(2012), which we will now express formally. Let $F^h := \{f^1, ..., f^h\}$ denote the set of the first h components. According to the LocNes principle, extra component f^{h+1} must best complement the existing ones and A, i.e. $A^h := F^h \cup A$. So f^{h+1} must be calculated using A^h as additional covariates. Moreover, we must impose that f^{h+1} be orthogonal to F^h:

$$F^{h\prime} W f^{h+1} = 0 \, . \tag{2.4}$$

This principle allows to build up an increasing sequence of components, which can eventually span the whole $Vec(X)$ space (when $h = p$ or sooner). But then, the noise-dimensions in $Vec(X)$ (those irrelevant to model Y, or strongly backed by no regressor) would be used in the linear predictor, producing overfitting and jeopardizing prediction quality. So, when they start producing overfitting, components will have to be discarded. This will be assessed later by cross-validation trials.

There are two ways of ensuring orthogonality of every extra-component with the previous ones, expressed in (2.4): deflation of X, or extra orthogonality constraint.

Deflation

This method, classical in PLS, consists in currently replacing X with $X^h := \Pi_{Vec(F^h)^\perp} X$ in program R to calculate $f^{h+1} = X^h u$. This technique has two drawbacks: 1) losing the original variables, it sometimes requires adapting metric M to the new covariate matrix X^h ; 2) as one ultimately needs to get an expression of the linear predictors as a function of the original variables, one has to be able to get back from every X^h to X, which requires recursive calculations (Bry et al., 2013).

Extra orthogonality constraint

This method consists in adding constraint (2.4) to program R. To calculate compo-

nent $f^{h+1} = Xu$, we would now solve:

$$R : \max_{u'M^{-1}u=1\,,D^{h}{}'u=0} S(u) \quad \text{where} \quad D^h := X'WF^h \ .$$

Again, the PING algorithm given in supplemental material 7.2 allows to solve this program. Supplemental material 7.4 proves that deflation and orthogonality constraints are equivalent for SR measures based on the component's variance, and how to make them equivalent for every SR measure based on closeness to subspaces. As far as calculation is concerned, using an extra orthogonality-constraint is much handier, since it directly provides the coefficient-vector of each component on the variables, which is needed for predictions.

3 THEME-SCGLR

Consider now the complete thematic equation: $Y = F(X_1)+...+F(X_R)+A$. In order to deal with multiple themes, we must replace Q' by another equivalent program:

$$Q'' : \max_{\forall r, u_r \in \mathbb{R}^{p_r}\,, u_r' M_r^{-1} u_r = 1} \psi(u_1, ..., u_R) \ ,$$

where $\psi(u_1, ..., u_R) = \sum_k \|z_k\|^2_{W_k} \cos^2_{W_k}(z_k; Vec(X_1u_1, ..., X_Ru_R, A))$. Q' and Q'' are obviously equivalent because $Vec(X_1, ..., X_R, A) = Vec(X, A)$ and the criterion ψ to be maximised is 0-degree-homogenous in its arguments: the direction of $Vec(X)$ maximising ψ is the same in either case. But Q'' opens the way to theme-specific regularisation.

3.1 Rank 1 components

Let us introduce the SR of components into program Q'' by solving instead:

$$R'' : \max_{\forall r, u_r \in \mathbb{R}^{p_r}, u_r' M_r^{-1} u_r = 1} \psi(u_1, ..., u_R)^g \prod_{r=1}^{R} \phi^{s_r}(u_r) \ . \tag{3.1}$$

The product form of (3.1) can be justified in much the same way as for (2.3). To make things simpler, one may take: $\forall r = 1, ..., R$: $s_r = s$ and $g = 1 - s$. This choice tunes each theme's SR with respect to the model's GoF. It is not the only possible choice, of course: one could instead want to tune the importance of the overall SR with respect to the GoF by taking $\forall r \in \{1, ..., R\}$, $s_r = \frac{s}{R}$ and $g = 1 - s$. In both cases, every theme is given the same a-priori importance. In particular, the criterion is insensitive to the number of variables in each theme. Such would not be the case if all themes were pooled into one. R'' can be solved by iteratively maximising in turn the criterion on every u_r. Now, we have:

$$\forall r \ : \ \cos^2_{W_k}(z_k \ ; \ Vec(X_1 u_1, ..., X_R u_R, A)) = \cos^2_{W_k}(z_k \ ; \ Vec(X_r u_r, \tilde{A}_r)) \ ,$$

where

$$\tilde{A}_r = A \cup \{f_s; s \neq r\} \ .$$

For convenience, when seeing $\psi(u_1, ..., u_R)$ as a function of a particular u_r, we will write it $\psi(u_r)$. The additional covariates will be specified on every such occasion. So, R'' can be solved by iteratively solving:

$$R''_r : \max_{u_r \in \mathbb{R}^{p_r}, u_r' M_r^{-1} u_r = 1} \psi(u_r)^{(1-s)} \phi^s(u_r) \ ,$$

with $\tilde{A}_r$ as additional covariates. Section 2.4 already showed how to solve this program.

3.2 Rank $h > 1$ components

Components in a theme (e.g. X_r) are to span, as extensively as possible, a subspace leaning on strong structures of X_r which is useful for predicting Y when associated, not only with A, but also the components of the other themes. So, momentarily considering these given, we can add them to A and proceed as in SCGLR (section **??**) to calculate a sequence of components in X_r. We can do that in every theme, and of course, iterate. Let us write this more formally. Suppose we want to calculate H_r components in theme X_r. $\forall h < H_r$, let $F_r^h := \{f_r^\ell; \ell = 1, ..., h\}$. The LocNes principle states that component f_r^{h+1} must best complement the existing ones, i.e. F_r^h and all components of all other themes, plus A, i.e. : $A_r^h := F_r^h \underset{s\neq r}{\cup} F_s^{H_s} \cup A$. Taking A_r^h as additional covariates, the current value of f_r^{h+1} is calculated solving:

$$R_r'' : \max_{u_r\in\mathbb{R}^{pr},\, u_r' M_r^{-1} u_r=1,\, D_r^{h\prime} u_r=0} \psi(u_r)^{(1-s)}\phi^s(u_r) \quad \text{where} \quad D_r^h := X_r' W F_r^h \, .$$

Informally, the algorithm consists in currently calculating all H_r components in X_r - in the way given in 2.5, taking $A \underset{s\neq r}{\cup} F_s^{H_s}$ as additional covariates - and then loop on r until overall convergence of the component-system.

Note that the LocNes principle induces a partial order. For $h_1 \leq H_1, ..., h_R \leq H_R$, let $M(h_1, ..., h_R)$ denote the component-model based on components $\{F_r^{h_r}\}_{1\leq r\leq R}$. Model $M(H_1, ..., H_R)$, produced by our algorithm, contains sub-models. A sub-model is defined by any ordered pair (r, h_r) where $h_r \leq H_r$, as:

$$\text{SM}(r, h_r) \;=\; \text{M}(H_1, ..., H_{r-1}, h_r, H_{r+1}..., H_R) \, .$$

The set of all sub-models is not totally ordered. But we have the following theme-local nesting property:

Every sequence of sub-models defined by SM$(r, .)$ $=$ $(SM(r, h_r))_{0\leq h_r\leq H_r}$ *is totally*

ordered through the relation: $SM(r,h'_r) \leq SM(r,h_r) \Leftrightarrow h'_r \leq h_r$.

This order is easy to interpret, considering that the component f_r^h making the difference between $\mathrm{SM}(r,h-1)$ and its successor $\mathrm{SM}(r,h)$ is the X_r -component orthogonal to F_r^{h-1} that best completes model $\mathrm{SM}(r,h-1)$ controlling for all other components in $\mathrm{SM}(r,h-1)$.

On these principles, we have built the algorithm given in supplemental material 7.3, that calculates H_r components in each theme X_r. This algorithm has been implemented in an R-package named SCGLR. If we suppose we retain enough components in every theme to exhaust their own predictive capacity, then, components in theme X_r, being determined through their partial effect, will focus on the specific role of X_r in the GLM. Of course, as in SCGLR, they will be too many as soon as they produce overfitting, so their sequence will have to be pruned according to cross-validation performance. Cross-validation trials require the coefficients of original variables in linear predictors to be calculated.

3.3 Coefficients of original variables in linear predictors

Let $U = [u_1^1, \ldots, u_1^{H_1}, \ldots, u_R^1, \ldots, u_R^{H_R}]$. Once the components $\left\{f_r^h = X_r u_r^h\right\}_{1\leq r\leq R, 1\leq h\leq H_r}$ have been calculated, a generalised linear regression of each y^k is performed on $[F,A]$, where $F := \left\{F_r^{H_r}\right\}_{1\leq r\leq R}$, yielding linear predictor:

$$\eta_k = \theta_k + A\delta_k + F\gamma_k = \theta_k + A\delta_k + X\beta_k \quad \text{where} \quad \beta_k = U\gamma_k \ .$$

4 Model Assessment

4.1 Principle

Model assessment is based on the predictive capacity of models. Prediction is assessed using cross-validation. Consider a given model $\mathrm{M} = \mathrm{M}(h_1, ..., h_R)$. The sample is first divided into two subsamples: CT (for calibration and testing) and V (for validation). Then, CT is subdivided a given number of times into two subsamples: C (calibration sample) and T (test sample). For every observation in T and every y, we calculate some prediction-error indicator e fitting y's nature. An average error rate $\mathrm{AER}(\mathrm{M}, \mathrm{C}, \mathrm{T})$ has to be defined over all dependent variables, and a cross-validation error rate CVER(M), as the average of $\mathrm{AER}(\mathrm{M}, \mathrm{C}, \mathrm{T})$ over (C,T) pairs. Models are then compared with respect to their CVER, the best ones must be validated on V.

4.2 Prediction-error indicators

To every type of y may correspond one (or more) appropriate error indicators. For instance, for a binary output $y \sim B(p(x,t))$, AUC denoting the corresponding area under ROC curve, we would advise to take:

$$e = 2(1 - \mathrm{AUC}) \; .$$

Whereas for a quantitative variable, we ought to consider indicators based on the mean quadratic error, such as:

$$e = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{E}(y_i|x_i, t_i))^2}{\hat{V}(y_i|x_i, t_i)} \; .$$

For instance, for a Poisson-distributed variable $y \sim P(\lambda(x,t))$, we would get:

$$e = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\lambda}(x_i,t_i))^2}{\hat{\lambda}(x_i,t_i)} \ .$$

But these error indicators are not necessarily comparable, and must yet be pooled into an overall indicator. We propose to use geometric averaging, since it allows relative compensations of indicators.

4.3 Backward component selection for a given structural relevance

The LocNes principle calls for backward component selection. Indeed, starting with "large enough" numbers of components in every theme allows to capture most of the theme's proper predictive power, and thus, minimise the risk of confusing effects between themes. But, to ensure having "large enough" such numbers, one should start with "too large" ones, hence an over-fitting model. As every component has been calculated given the former-rank ones in its theme, only higher-rank components should then be removed, starting with the highest. In the logic of LocNes, every extra component should allow to improve the overall quality of prediction of the responses, unless it contributes to over-fitting. So, we consider the loss in CVER(M$(h_1,...,h_R)$) brought by the highest rank component $f_r^{h_r}$ in X_r.

The backward selection algorithm consists in comparing the loss in CVER values associated with each highest rank component, removing the one which causes the highest CVER increase (or lowest decrease), recalculating the whole set of components in all themes with the updated numbers of components, and resuming so long as the model is not empty.

Such a procedure being rather costly, we considered in practice the following one, less accurate but considerably faster. The model with the maximum number of components is first calculated. The components are then considered given. Then, we remove in turn the higher-rank component in every theme, estimate the regression coefficients of the pruned model and calculate its CVER. The one with the least CVER is chosen, and the pruning procedure goes on from it.

5 Applications to data

5.1 Testing the enhanced SCGLR on simulated data

To illustrate the behaviour of the SR criterion, we restrict this simulation study to a single explanatory theme X, with empty A. We consider $n = 100$ units, $p = 200$ covariates, and $q = 50$ responses. X is structured around four variable bundles: B_1 (20 variables), B_2 (30 v.), B_3 (40 v.) and B_4 (50 v.), plus a cloud of 60 uncorrelated noise-variables. Each bundle B_k is structured about a latent variable ϕ_k . The correlation between the ϕ_k's is tuned through: $\text{corr}(\phi_k, \phi_m) = \cos^2(\alpha)$, $\alpha \in [0; \pi/2]$, and the width of bundles through parameter $\nu \in [0; 1]$ (see supplemental material 7.6 for the detailed simulation scheme). Only bundles B_1 to B_3 are really explanatory, in decreasing order of importance, B_4 being a heavy junk bundle. The idea behind this scheme is that ideally, a good method should identify central directions ϕ_1, ϕ_2, and ϕ_3 in the proper order, i.e. $Vec(\phi_1)$, then $Vec(\phi_1, \phi_2)$ and finally $Vec(\phi_1, \phi_2, \phi_3)$ without being fooled by B_4. Now, B_4 being so heavy, it is likely to mislead any method that uses a simply quadratic PCA-type measure of structural relevance. Such is the case

of PCGLR (GLR on PC's), and the basic version of SCGLR. By contrast, varying parameter l (and possibly s) should help track down the true directions in the right order.

We generate data for the following experimental design:

$$\alpha \in \{\frac{\pi}{2}, \frac{\pi}{4}\} \ \times \ \nu \in \{0.1, 0.5, 1\} \ .$$

On every dataset simulated according to an ordered pair (α, ν), we use the VPI criterion as SR measure, with different values of l and s, trying to find out which choice of (l, s) performs best according to the situation.

To judge the quality of the estimation, we calculate the following indicators:

- Quality of capture of ϕ_k by $Vec(F^h)$:

$$\mathrm{QLT}_h(\phi_k) = \sum_{j=1}^{h} \rho^2(f^j, \phi_k) \ .$$

- Hence, average quality of capture of the true explanatory directions by $Vec(F^h)$:

$$\overline{\mathrm{QLT}}_h = \frac{1}{3} \sum_{k=1}^{3} QLT_h(\phi_k) \ .$$

- Propensity of $Vec(F^h)$ to capture the wrong direction ϕ_4 :

$$\mathrm{QLT}_h(\phi_4) \ .$$

Results are presented in Figure 2. They show that VPI with $l = 1$ has been all the more cheated as s is high, while VPI with $l > 1$ has not, and all the less as l is high. For $\alpha = \frac{\pi}{2}$, for instance, and when $s = 0.5$, i.e. SR is taken into account equally with GoF, $\mathrm{QLT}_3(\phi_1)$ is poor, but $\mathrm{QLT}_3(\phi_4)$ is high (see Figure 3), which we want to avoid.

As s decreases, things improve, because the GoF has then more weight than the SR. But the key-parameter is still l: whatever the (non-zero) value of s, $l = 8$ allows all $\phi_k, k = 1..., 3$ to be captured by $Vec(F^3)$, and discards ϕ_4 (cf. Figure 3).

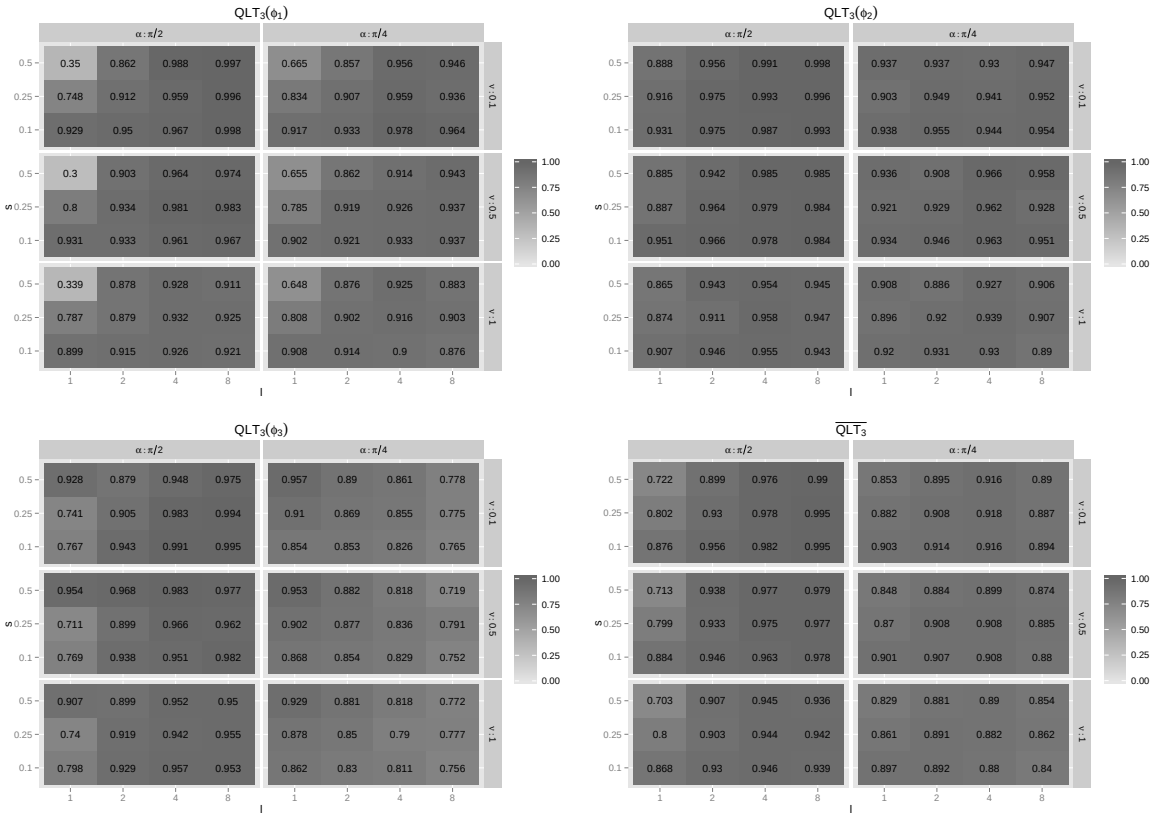


Figure 2: Quality of capture of the simulated true latent variables by the space spanned by components 1 to 3. Value 1 indicates perfect capture (wanted here), and value 0, non-capture. Parameters α and ν tune the design of the simulated bundles, while s and l tune the component-extraction algorithm.

5.2 Applying THEME-SCGLR to the tropical tree species distribution

There is a crucial need to improve our current knowledge of the tree species distribution of central African forests in order to define sustainable management policies

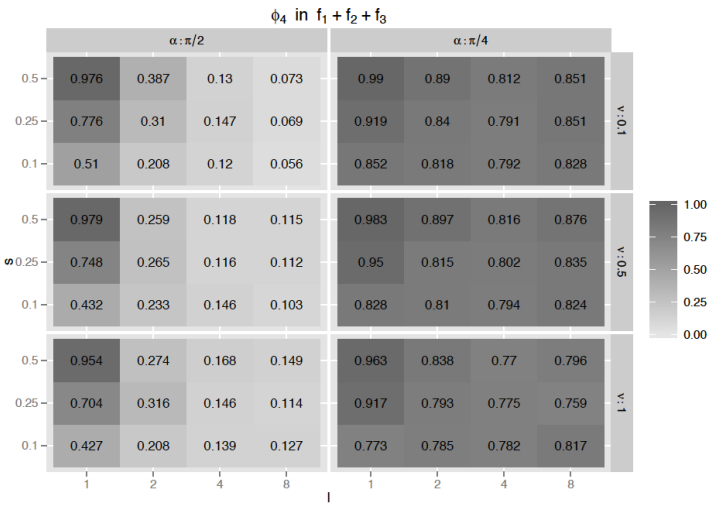


Figure 3: Quality of capture of the simulated nuisance latent variables by the space spanned by components 1 to 3. Value 1 indicates perfect capture, and value 0, non-capture (wanted here). Parameters α and ν tune the design of the simulated bundles, while s and l tune the component-extraction algorithm.

and better predict their future in the face of global changes. We analysed floristic data consisting of the abundance of 27 common tree genera, provided by logging concessions, from 1 000 plots in the tropical moist forest of the Congo-Basin covering an area of more than six Mha over four countries: Cameroon, Gabon, Central African Republic and the Republic of Congo. 40 geo-referenced environmental variables were monitored, which include 13 pluviometric indicators, geology of each plot and 23 EVI indices measuring photosynthetic activity. The environmental variables were divided into two clearly defined themes: (i) the EVI indices (X_1), and (ii) pluviometry (X_2). The geological factor was used as an additional covariate (A). Thus, 27 response variables, assumed to be Poisson random variables, were available, two themes were defined and one additional covariate was used on 1 000 plots (observation units).

We have tried all combinations of $(l, s) \in \{1, 2, 3, 4\} \times \{0.15, 0.25, 0.5\}$ and calculated the CVER along the backward component-selection path for each (l, s). A model with k components in X_1 and h in X_2 will be shorthanded k_h. In most cases, the best-predictive model was 4_4, but the model 0_4 having discarded the EVI theme was not bad either. Indeed, the model-choice issue is not as straightforward as it seems, since at least three concerns are involved, which may fail to coincide: prediction quality, parsimony, and interpretability. Prediction quality is measured by the CVER only. Parsimony is twofold: dimensional parsimony is measured by the overall number of components involved in the final model, but thematic parsimony is measured by the number of themes involved. Finally, interpretability of a component-model is that of its components, i.e. the height of their correlations with some of the variables in their themes. In our application, we have chosen the model 0_4 obtained with $l = 2, s = 0.5$ since it has one of the lowest CVER's (1.8366), together with the highest parsimony (the EVI theme playing no role in it, and 4 components being sufficient - and easily interpretable - in the pluviometry theme). Note, though, that for the same (l, s), model 4_4 led to a slightly lower CVER (1.8291), and that model 5_4 with $l = 3, s = 0.15$ still improved the performance (CVER = 1.8254), but negligibly and at a high cost in terms of parsimony.

Figure 4 presents the average CVER over the 10 replications for all k_h, k and h ranging from 1 to 6. On this figure, the path found by the backward selection procedure is plotted as a continuous black line. Note that it crosses the very best combination (4_4) and, provided we do not stop at it, leads to an almost equivalently predictive but much more parsimonious solution (4_0), which we will retain. This first result demonstrates that EVI has far less importance for predicting tree genus abundances than pluviometry.

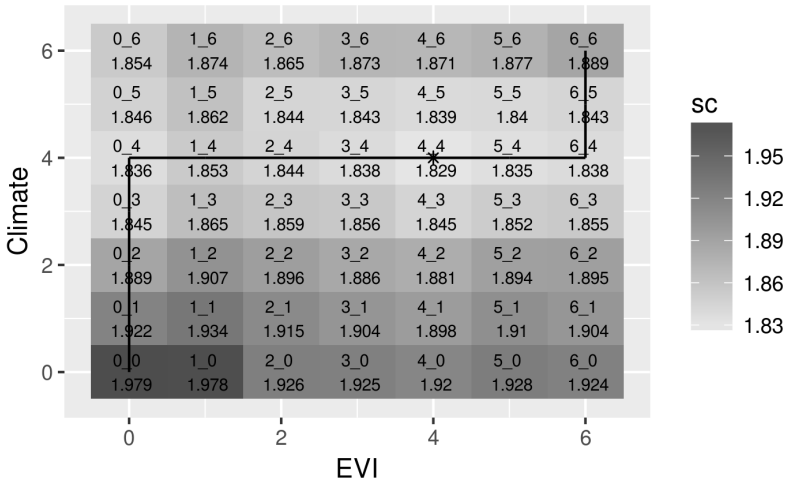


Figure 4: Average CVER values for all combinations of component numbers in the two themes.

It must be noted at this stage that if we mix up all explanatory variables in a single theme and perform SCGLR, the lowest CVER is never obtained with less than 12 components. Besides, with $l = 2, s = 0.5$ and retaining 8 components, it performs worse, with CVER=1.8366 , than the two-theme model 4_4 (CVER=1.8291). Finally, with only 4 components, it yields a CVER = 1.8883, which is significantly larger than that of model 4_0 (CVER=1.8366).

These results highlight how THEME-SCGLR may improve and refine the initial SCGLR model by dividing the set of covariates into different well-defined themes. Indeed, using SCGLR with a single theme mixing all covariates, components 1 and 2 are dominated by an EVI bundle (see fig. 5), which is strong enough to attract them, although it has but an insignificant role in prediction: SCGLR is trapped by the strong structure within the EVI covariates, whereas, by contrast, separating the EVI theme from the pluviometric one in the thematic model reveals that the EVI's have but little predictive importance and allows to discard it from the conceptual

model of the data.

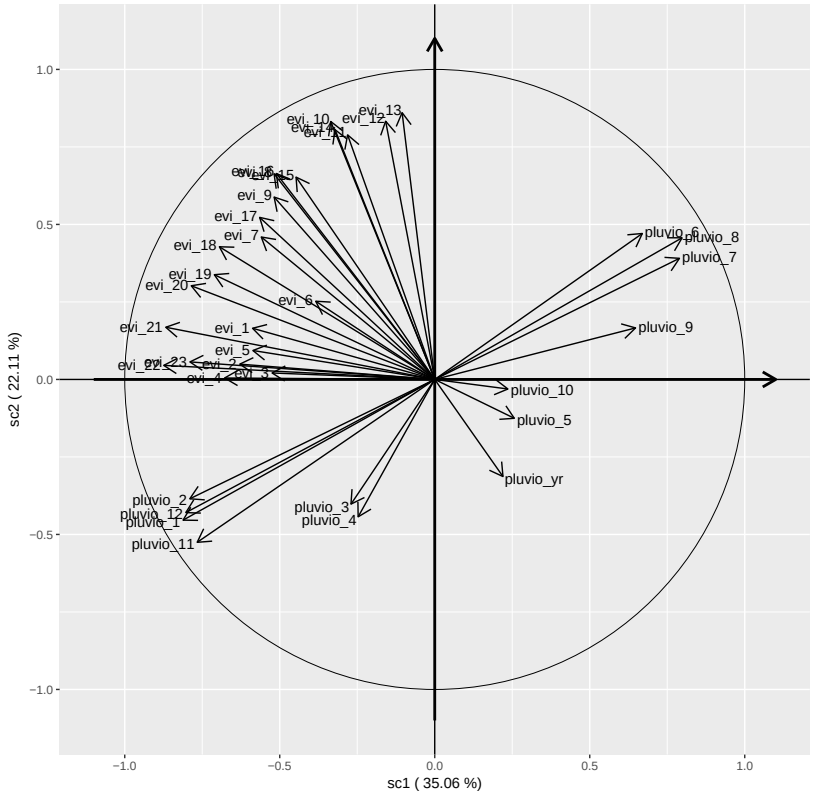
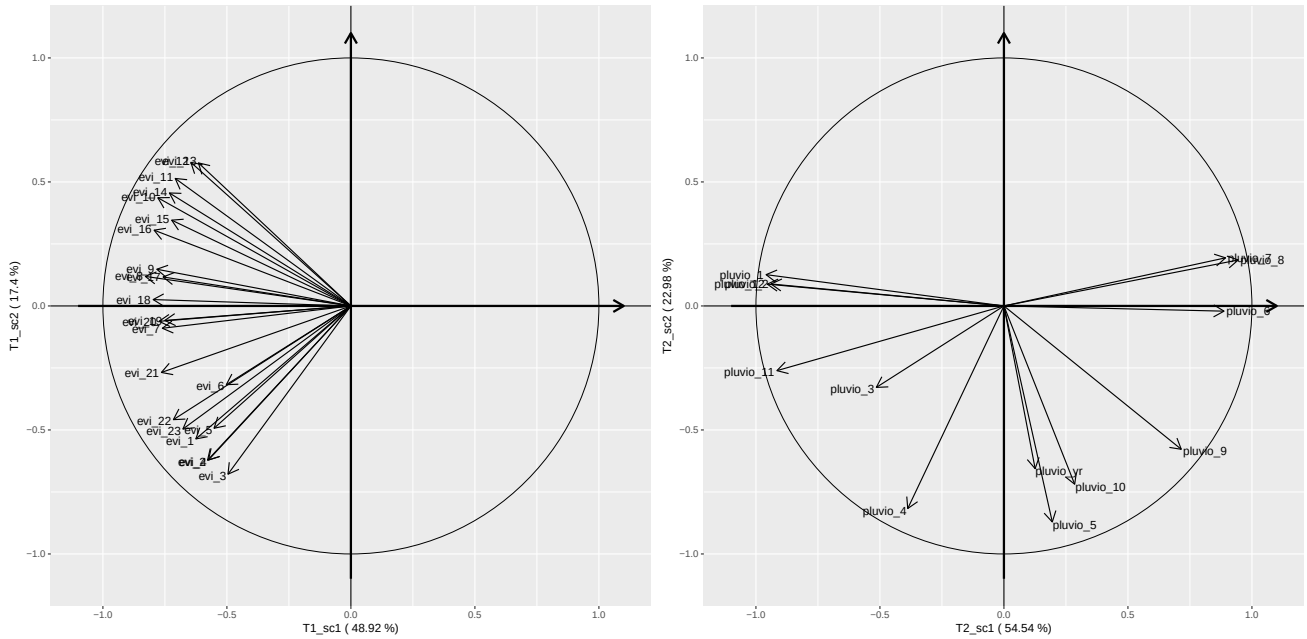


Figure 5: Correlation scatterplot of covariates in the plane spanned by the first two components obtained using the single-theme SCGLR approach.

The correlation scatterplot of Figure 6b shows that the first pluviometric component is highly linked to a rain-pattern opposing winter rainfalls to spring ones, the second pluviometric component being characterised by rainfalls on intermediate months.



(a) (b)

Figure 6: Correlation scatterplot of covariates in the planes spanned by the first two components obtained by THEME-SCGLR using EVI (a) and pluviometry (b) themes.

6 Conclusion

The original SCGLR was a PLS-type tradeoff between Multivariate GLM estimation (which cannot afford many and redundant covariates) and PCA-like dimension reduction methods (which take no explanatory model into account). It allowed both to regularise GLM estimation and to decompose the linear predictors on strong common components, which methods as LASSO or Ridge, merely penalising the norm of the coefficient-vector, do not do. THEME-SCGLR extends SCGLR in two major ways:

- It replaces the component's variance, used in SCGLR as a measure of structural strength, with a more general and flexible measure of Structural Relevance. This measure allows to better align components on more local variable structures, as bundles or theory-based subspaces, enhancing their interpretability.
- By extending SCGLR to a thematic partition of the explanatory variables, it allows to make better use of the complementarity between the explanatory themes, both statistically when fitting the model, and conceptually when interpreting the components. Moreover, through an unambiguous component selection mechanism, it allows to find the useful dimensionality of each theme and explore it hierarchically.

These features enable testing thematic models of the phenomenon under attention through the interpretability of the components and the prediction quality that the corresponding estimated models offer. Thus, THEME-SCGLR allows to explore the multiple competitive conceptual hypotheses that can be considered when modelling a phenomenon in high dimension, and gradually refine and adjust the design of the thematic model according to previous findings.

In our simulations, THEME-SCGLR proved to have the expected behaviour regarding bundles. On the environmental data set, using a 2-theme model separating EVI and pluviometry allowed to conclude to a negligible role of the EVI's, which a single-theme model could not do.

Acknowledgements

This research was partially supported by ITG-SEITA, the CoForTips project and the "Forêts et Changements Climatiques au Congo" project, funded by the European Union. The CoForTips project was funded by the ERA-Net BiodivERsA, with national funders FWF (Austria), BelSPO (Belgium) and ANR (France), and was part of the 2011-2012 BiodivERsA call for research proposals (http://www.fordev.ethz.ch/research/active/CoForTips). We also wish to thank the MEDDEFCP and ICRA of the Central African Republic, Gembloux Agro-Bio Tech (Liège University), the timber companies, the FRM and Nature+ consulting firms who participated in data collection, data compiling, and provided inventory data files.

References

Bry, X. and Verron, T. (2015). THEME: THEmatic Model Exploration through Multiple Co-Structure maximization. *Journal of Chemometrics*, **29**, 637–647.

Bry, X., Verron, T., and Cazes, P. (2009). Exploring a physico-chemical multi-array explanatory model with a new multiple covariance-based technique: Structural equation exploratory regression. *Anal. Chim. Acta*, **642**, 45–58.

Bry, X., Verron, T., Redont, P., and Cazes, P. (2012). THEME-SEER: a multidimensional exploratory technique to analyze a structural model using an extended covariance criterion. *Journal of Chemometrics*, **26**, 158–169.

Bry, X., Trottier, C., Verron, T., and Mortier, F. (2013). Supervised Component-based Generalized Linear Regression using a PLS-extension of the Fisher scoring

algorithm. *Journal of Multivariate Analysis*, **119**, 47–60.

Durif, G., Picard, F., and Lambert-Lacroix, S. (2015). Adaptive Sparse PLS for Logistic Regression. **ArXiv**.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modeling Based on Generalized Linear Models.* Springer-Verlag, New York.

Fort, G. and Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. *Bioinformatics*, **21**, 1104–1111.

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

McCullagh, P. and Nelder, J. (1989). *Generalized linear models.* Chapman and Hall, New York.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B.*, **58**, 267–288.

Tikhonov, A. and Arsenin, V. (1977). *Solution of Ill-posed Problems.* Winston and Sons, Washington.

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

7 Supplemental material

7.1 Analytical expression of $S(u)$

We aim at expressing $S(u)$ as a function of quadratic forms. To achieve that, we decompose the projection on the regression space as follows:

$$Vec(Xu, A) = Vec(\tilde{X}u, A) \text{ with } \tilde{X} := \Pi_{A^\perp} X$$

$$Vec(\tilde{X}) \perp Vec(A) \Rightarrow \Pi_{Vec(Xu,A)} = \Pi_{Vec(\tilde{X}u,A)} = \Pi_{Vec(\tilde{X}u)} + \Pi_{Vec(A)}$$

$$\Rightarrow \|z_k\|^2_{W_k} \cos^2_{W_k}(z_k; Vec(Xu, A)) = \langle z_k|\Pi_{Vec(Xu,A)} z_k\rangle_{W_k} = \langle z_k|(\Pi_{Vec(\tilde{X}u)} + \Pi_{Vec(A)})z_k\rangle_{W_k}$$

$$= \left(\langle z_k|\Pi_{Vec(\tilde{X}u)} z_k\rangle_{W_k} + \langle z_k|\Pi_{Vec(A)} z_k\rangle_{W_k}\right)$$

Now:

$$\langle z_k|\Pi_{Vec(\tilde{X}u)} z_k\rangle_{W_k} = z_k' W_k \Pi_{Vec(\tilde{X}u)} z_k = \frac{u'\tilde{X}' W_k z_k z_k' W_k \tilde{X} u}{u'\tilde{X}' W_k \tilde{X} u}$$

Let:

$$A_k := \tilde{X}' W_k z_k z_k' W_k \tilde{X} \ ; \ B_k := \tilde{X}' W_k \tilde{X} \ ; \ c_k := \langle z_k|\Pi_{Vec(A)} z_k\rangle_{W_k}$$

We have:

$$\psi(u) = \sum_k \left(\frac{u' A_k u}{u' B_k u} + c_k\right) \tag{7.1}$$

Now, from (2.2) and (7.1), we get the general matrix form of criterion $S(u)$ to be maximised:

$$S(u) = \left(\sum_k \left(\frac{u' A_k u}{u' B_k u} + c_k\right)\right) \left(\sum_{j=1}^{J} \omega_j (u' N_j u)^l\right)^{\frac{s}{l}}$$

7.2 The Projected Iterated Normed Gradient (PING) algorithm

The current value of any quantity a on iteration t is denoted: $a^{[t]}$.

Consider program:

$$\max_{u\in\mathbb{R}^p\,,\ u'M^{-1}u=1,\ D'u=0} h(u)$$

First note that, putting $v = M^{-1/2}u$, $g(x) = h(M^{1/2}x)$ and $C = M^{1/2}D$, this is strictly equivalent to:

$$R_C : \max_{v\in\mathbb{R}^p\,,\ v'v=1,\ C'v=0} g(v)$$

Also note that $C = 0$ corresponds to the case of no orthogonality constraint.

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

$$\nabla_{\lambda,\mu} L(v,\lambda,\mu) = 0 \ \Leftrightarrow \begin{cases} v'v = 1 & (7.2) \\ C'v = 0 & (7.3) \end{cases}$$

$$\nabla_v L(v,\lambda,\mu) = 0 \ \Leftrightarrow\ \Gamma(v) - 2\lambda v - C\mu = 0 \quad \text{with} \quad \Gamma(v) := \nabla_v g(v) \tag{7.4}$$

$$\Leftrightarrow\ v = \frac{1}{2\lambda}(\Gamma(v) - C\mu) \tag{7.5}$$

Premultiplying (7.4) by C', with (7.5), yields

$$C'\Gamma(v) = C'C\mu \quad \Leftrightarrow \quad \mu = (C'C)^{-1}C'\Gamma(v)$$

Put back into (7.5), this yields:

$$v = \frac{1}{2\lambda}\Pi_{C^\perp}\Gamma(v) \quad \text{where} \quad \Pi_{C^\perp} := I - C(C'C)^{-1}C' \tag{7.6}$$

Note that in the particular case where $C = 0$, we shall take: $\Pi_{C^\perp} = I$.

Finally, (7.6) and (7.2) imply:

$$v = \frac{\Pi_{C^\perp}\Gamma(v)}{\|\Pi_{C^\perp}\Gamma(v)\|} \tag{7.7}$$

This gives the basic iteration of the PING algorithm:

$$v^{[t+1]} = \frac{\Pi_{C^\perp}\Gamma(v^{[t]})}{\|\Pi_{C^\perp}\Gamma(v^{[t]})\|} \tag{7.8}$$

Let us show that this iteration follows a direction of ascent. Since, by construction:

$$\forall s\text{: } v^{[s]} \perp C$$

we have:

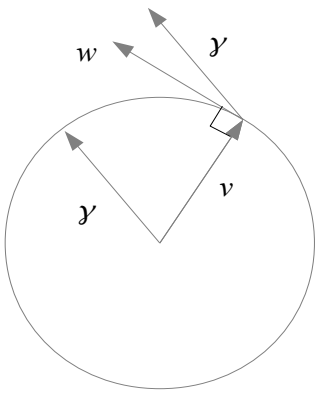
$$\begin{aligned} \forall s\text{: } v^{[s]} = \Pi_{C^\perp} v^{[s]} \;\Rightarrow\quad \langle v^{[t+1]} - v^{[t]} | \Gamma(v^{[t]})\rangle &= \langle \Pi_{C^\perp}(v^{[t+1]} - v^{[t]}) | \Gamma(v^{[t]})\rangle \\ &= \langle v^{[t+1]} - v^{[t]} | \Pi_{C^\perp}\Gamma(v^{[t]})\rangle \end{aligned}$$

which has the sign of:

$$\langle v^{[t+1]} - v^{[t]} | v^{[t+1]}\rangle = 1 - \langle v^{[t]} | v^{[t+1]}\rangle = 1 - \cos(v^{[t]}, v^{[t+1]}) \geq 0$$

Of course, picking a point on a direction of ascent does not guarantee that g actually increases, since we may "go too far" in this direction. Let $\gamma^{[t]} := \frac{\Pi_{C^\perp}\Gamma(v^{[t]})}{\|\Pi_{C^\perp}\Gamma(v^{[t]})\|}$. If we stay "close enough" to the current starting point on the arc $(v^{[t]}, \gamma^{[t]})$, we can guarantee that g increases. Indeed, let ϖ be the plane tangent to the sphere on $v^{[t]}$ and let w denote the vector tangent to arc $(v^{[t]}, \gamma^{[t]})$ on $v^{[t]}$. Then:

$$\exists \tau > 0, w = \tau \Pi_\varpi \gamma^{[t]} \quad\Rightarrow\quad \langle w | \gamma^{[t]}\rangle = \tau \langle \Pi_\varpi \gamma^{[t]} | \gamma^{[t]}\rangle = \tau \cos^2(\gamma^{[t]}, \varpi) > 0$$



Yet, if we stay "too close" to the current starting point on the arc $(v^{[t]}, \gamma^{[t]})$, the algorithm may get too slow to reach the maximum. We avoid that by using a one-dimensional maximisation function (e.g. Gauss-Newton type) to find the maximum of $g(v)$ on the arc $(v^{[t]}, \gamma^{[t]})$, and take it as $v^{[t+1]}$. The fixed point of the resulting algorithm is a critical point of (7.2), hence a local maximum of g s.t. $C'v = 0$.

7.3 The THEME-SCGLR algorithm

Additional notations: $\forall r \in \{1, ..., R\},\ F_r := F_r^{H_r}\ \ ;\ \ F := \underset{r=1,...,R}{\cup} F_r$.

Initialisation:

Let $\forall r = 1, \ldots, R$; $\forall k = 1, ..., K_r$: $f_r^k = k^{th}$PC of X_r.

Initialise $Z = [z^1|...|z^q]$ to $Z^{[0]}$ with:

$\forall i = 1, \ldots, n\ ,\ \forall j = 1, \ldots, q:\quad z_i^{j[0]} = g(\alpha y_i^j + (1-\alpha)\bar{y}^j)$, where $\alpha = 0.95$,

and $\{W_k\}_{k=1,\,q}$ to $\{W_k^{[0]}\}_{k=1,\,q} = \{\frac{1}{n} I_n\}_{k=1,\,q}$

Current iteration:

Iterate from $s = 0$ until convergence of $F^{[s]}$:

1) *GLS Regression step*: for $l = 1$ to q,

   - Carry out GLS regression of each $z^{l[s]}$ on $F^{[s]}$ with respect to weighs $W_l^{[s]}$:

   - Update each $z^{l[s]}$ and $W_l^{[s]}$ using the regression coefficients.

2) *Updating components*: Iterate from $t = 1$ until $F^{[s,\infty]}$,

- Set $F^{[s,0]} = F^{[s-1,\infty]}$

- For r= 1 to R:

For $h = 1$ to H_r:

Solve program: R''_r on $Z^{[s]}$,

with additional covariates $A_r^{h-1} := F_r^{h-1} \underset{s \neq r}{\cup} F_s^{H_s} \cup A$

and with orthogonality constraint matrix $D_r^{h-1} := X_r' W F_r^{h-1}$.

Call $u_r^{h[s,t]}$ the solution.

Set $f_r^{h[s,t]} = X_r u_r^{h[s,t]}$;

Set $F_r^{h[s,t]} = [F_r^{h-1[s,t]}, f_r^{h[s,t]}]$

End for h

End for r

7.4 Equivalence between deflation and orthogonality constraint for some measures of SR

1) **SR = Component's variance**

We show the following equivalence:

$$\max_{u'u=1,\ C'u=0} \|Xu\|_W^2, \ \text{ where } \ C' = F'WX \Leftrightarrow \max_{u'u=1} \|\Pi_{F^\perp} Xu\|_W^2 \ .$$

Indeed:

$$F'WXu = 0 \ \Rightarrow \ \Pi_{F^\perp} Xu = (I - F(F'WF)^{-1}F'W)Xu = Xu \ .$$

2) **SR = Closeness of component to a given subspace**

Consider a unit-variance component f constrained to W-orthogonality with respect to subspace F (e.g. the subspace spanned by lower-rank components), and $S \subset \mathbb{R}^n$ a given reference-subspace. Consider any SR such as:

$$g(\cos^2(f, S)) = g(\langle f | \Pi_S f \rangle_W) \quad \text{with } g \text{ increasing.}$$

Let $\{s_1, ..., s_d\}$ be an orthonormal basis of S. We have:

$$\Pi_S = \sum_{m=1}^{d} \Pi_{s_m} \quad \Rightarrow \quad \langle f|\Pi_S f\rangle_W = \sum_{m=1}^{d} \langle f|\Pi_{s_m} f\rangle_W = \sum_{m=1}^{d} \langle f|s_m\rangle_W^2 \ .$$

Now, as $\Pi_{F^\perp} f = f$, we have:

$$\forall m, \ \ \langle f|s_m\rangle_W = \langle \Pi_{F^\perp} f|s_m\rangle_W = \langle f|\Pi_{F^\perp} s_m\rangle_W \ .$$

So, $\langle f|\Pi_S f\rangle_W = \sum_{m=1}^{d} \langle f|\Pi_{F^\perp} s_m\rangle_W^2$. As a consequence, if the deflation procedure is based on deflated vectors $s_m^F := \Pi_{F^\perp} s_m$, and SR measure $\phi = g(\sum_{m=1}^{d} \langle f|s_m^F\rangle_W^2)$, deflation is equivalent to using the orthogonality-to-F constraint.

As immediate consequences:

- Deflation and orthogonality constraint are equivalent for VPI on standardised numerical variables.
- Deflation and orthogonality constraint can be made equivalent for VPI on categorical variables X^j, by considering for each the subspace spanned by its centred indicator variables.

7.5 Using principal components in VPI when $X'WX$ is singular

Let C denote the block of X's PC's associated with non-null eigenvalues: $C = XV$, where V is the matrix of corresponding unit-eigenvectors. Let $f = Cu$ be the component calculated after replacing X with C : $f = Xw$ with $w = Vu$.
Consider the following program:

$$\min_{t\in\mathbb{R}^p, \ Xt=f} \|t\|^2 \quad \Leftrightarrow \quad \min_{t\in\mathbb{R}^p, \ Xt=Xw} \|t\|^2 \ .$$

Of course,

$$Xt = Xw \quad \Leftrightarrow \quad Xe = 0 \quad \text{where } e = t - w \ .$$

Now:

1. $\forall v_k \in V,\ X'WXv_k = \lambda_k v_k,$ with $\lambda_k > 0 \quad \Rightarrow \quad V \in Vec(X') \quad \Rightarrow \quad w = Vu \in Vec(X')$.

2. $Xe = 0 \quad \Leftrightarrow \quad e \in Vec(X')^{\perp}$.

This implies that decomposition $t = w + e$ is unique.

Now, Pythagore's theorem yields $\|t\|^2 = \|w\|^2 + \|e\|^2$; by which $\|t\|^2$ is minimum for $e = 0$, i.e. $t = w$.

7.6 Simulation scheme

X is generated as follows:

1. Simulate the four latent variables directing bundles

   • Simulate five independent normal random vectors:

   $$\psi = (N(0;1))^{\otimes n} \ ; \ \{\psi^m = (N(0;1))^{\otimes n} \ ; \ m = 1, ..., 4\} \ .$$

   • Generate the exogenous latent variables as follows:

   $$\forall m = 1, ..., 4 : \quad \phi^m := \text{standardize}(\psi \cos\alpha + \psi^m \sin\alpha) \ .$$

   The correlation between the ϕ's can be tuned through angle α. The smaller the α, the more confusion between the ϕ's.

2. Simulate bundles of observed regressors about the latent variables:

   • Simulate 200 independent random vectors:

$$\{\eta_t = (N(0;1))^{\otimes n} \; ; \; t = 1, ..., 200\} \; .$$

   • Let $\nu > 0$ be a noise parameter tuning the width of variable bundles about their central direction. The larger the ν, the bigger the noise, so, the more confusion within the bundles:

   B1: $\forall j = 1, ..., 20 : x^j = \phi^1 + \nu\eta_j$; B1 is to be the most explanatory bundle (20 variables).

   B2 : $\forall j = 21, ..., 50 : x^j = \phi^2 + \nu\eta_j$; B2 is to be the second most explanatory bundle (30 variables).

   B3: $\forall j = 51, ..., 90 : x^j = \phi^3 + \nu\eta_j$; B3 is to be the third most explanatory bundle (40 variables).

   B4: $\forall j = 91, ..., 140 : x^j = \phi^4 + \nu\eta_j$; B4 is to be a non-explanatory bundle (50 variables).

   Finally, pure noise: $\forall j = 141, ..., 200 : x^j = \eta_j$ (60 variables).

3. Simulate observed responses: The y's are Poisson-distributed. They were generated according to the following scheme:

$$\forall 1 \leq k \leq q, \forall 1 \leq i \leq n : y_i^k \sim P(e^{\eta_i^k}), \quad \text{where } \eta_i^k = a_0^k + \sum_{h=1}^{3} a_h^k \phi_i^k$$

   with, independently:

$$\forall k = 1, ..., 25 : \; a_1^k \sim U[-4, 4] \; ; \; a_2^k \sim U[-2, 2] \; ; \; a_3^k \sim U[-1, 1]$$
$$\forall k = 26, ..., 40 : \; a_1^k \sim U[-2, 2] \; ; \; a_2^k \sim U[-4, 4] \; ; \; a_3^k \sim U[-1, 1]$$
$$\forall k = 41, ..., 50 : \; a_1^k \sim U[-1, 1] \; ; \; a_2^k \sim U[-1, 1] \; ; \; a_3^k \sim U[-2, 2]$$