Component-based Generalized Linear Regression using a PLS-extended variant of the Fisher scoring algorithm

X. Bry[a], C. Trottier[a,b,*], T. Verron[c]

[a] *Institut de Mathématiques et de Modélisation de Montpellier, Université Montpellier II, Place Eugène Bataillon CC 051 - 34095, Montpellier, France*

[b] *Université Montpellier III, route de Mende - 34095, Montpellier, France*

[c] *ALTADIS, Centre de recherche SCR, 4 rue André Dessaux, 45404 Fleury les Aubrais, France*

Abstract

In the current estimation of a GLM model, the correlation structure of regressors is not used as the basis on which to lean strong predictive dimensions. Looking for linear combinations of regressors that merely maximize the likelihood of the GLM has two major consequences: 1) collinearity of regressors is a factor of estimation instability, and 2) as predictive dimensions may lean on noise, both predictive and explanatory powers of the model are jeopardized. For a single dependent variable, attempts have been made to adapt PLS Regression, which solves this problem in the classical Linear Model, to GLM estimation. In this paper, we first discuss the methods thus developed, and then propose an algorithm that combines PLS regression with GLM estimation in the multivariate context, under a conditional independence assumption. Our algorithm is tested on simulated data.

[*] Corresponding author - Tel : +33 (0)467 144 164, Fax : +33 (0)467 143 558

Email addresses: bry@math.univ-montp2.fr (X. Bry), trottier@math.univ-montp2.fr (C. Trottier), thomas.verron@fr.imptob.com (T. Verron)

Preprint submitted to Elsevier *September 30, 2011*

Keywords: Multivariate Generalized Linear Models; PLS Regression; Fisher Scoring Algorithm.

1. Introduction

The framework is that of a multivariate Generalized Linear Model (GLM): a set of q random variables $Y = \{y_1, \ldots, y_q\}$ is assumed to be dependent on p common explanatory variables, $\{x_1, \ldots, x_p\}$. Each y_k is modeled through a GLM taking $X = \{x_1, \ldots, x_p\}$ as regressors. Moreover, $\{y_1, \ldots, y_q\}$ are assumed independent conditional to X. All variables are measured on the same n statistical units.

The standard estimation of a GLM model maximizes the model fit on all linear combinations of regressors. Doing so, it attaches the same importance a priori to linear combinations close to many observed variables (i.e. dimensions that focussed a lot of the attention and measuring effort) than to linear combinations far from any of them (i.e. related to weak dimensions of measurement, not to say noise). Take the extreme case where all regressors are highly correlated because they reflect the same latent variable with independent error terms and suppose this latent variable is rather poorly related to the dependent ones. Combining the regressors, one may generate as many noise dimensions. These dimensions may even span a space large enough to provide a model with an excellent fit, although there is but one poorly explanatory structural dimension in regressors. Another way of looking at the contradiction is as follows. On the one hand, such a situation as previously described is known to cause instability of coefficient estimation. On the other hand, the presence of such correlated regressors indicates a major concern as to measuring a single predictive dimension; so, if this dimension were directly observed, and the model were based on it, there would be a single precisely estimated coefficient.

In order to remedy this contradiction between measurement concerns and estimation in the classical linear model, PLS Regression (PLSR) currently maximizes a covariance criterion that combines the model's goodness of fit index (R^2) with the variance of the linear combination of regressors, that measures its structural strength. Doing so, PLSR draws this combination towards strong measurement dimensions, i.e. away from structurally weak ones.

The PLSR criterion is naturally adapted to the linear context, but not to the GLM one.

There has been attempts to combine PLSR with a GLM. Let us briefly review three of them.

- When there is but one dependent variable y to be modeled, Marx (1996) has proposed an Iteratively Reweighted Partial Least Squares (IRPLS) estimation for Generalized Linear Regression. The principle is based on the fact that the maximum likelihood estimation of a GLM can be carried out by an iterative re-weighted least square (IRLS) procedure McCullagh and Nelder (1989), derived from the Fisher Scoring Algorithm (FSA). Each iteration of it performs Generalized Least Squares (GLS) using a weighting matrix, the design of which derives from the model's hypotheses, and, as such, depends on the model parameters. Therefore, this weighting matrix has to be updated on every GLS step using the current estimated value of these parameters. Now, the GLS step can be straightforwardly replaced by a PLSR step using the current weighting matrix.

This method is consistent both with the linear aspect of PLSR and with likelihood estimation of the GLM, because the weighting matrix deriving from the GLM's likelihood is taken into account in the local PLSR estimation. But this method has not yet been extended to multiple dependent variables.

- In the same context, Bastien et al. (2005) have proposed a different way to extend PLSR to GLM: PLS Generalized Linear Regression (PLS-GLR). Only regression of a single dependent variable (PLS1) has been dealt with. PLSGLR is based on the fact that, while modeling a quantitative variable z with a linear combination of $\{x_1, \ldots,\ x_p\}$, PLSR of z on X yields a rank 1 component f^1 collinear to: $\sum_{j=1}^{p} \hat{z}_j$, where $\hat{z}_j = \frac{cov(z,x_j)}{var(x_j)} x_j$ is the predictor given by OLS regression of z on x_i alone. To obtain rank 2 component, one performs the same calculus replacing each x_i with its OLS regression residuals on f^1, and so on.

  Hence an apparently straightforward extension of PLS1: given dependent variable y modeled through a GLM using explanatory variables $\{x_1, \ldots,\ x_p\}$, rank 1 component f^1 of PLSGLR is defined as the standardized version of : $\sum_{j=1}^{p} \hat{z}_j$, where $\hat{z}_j = \hat{b}_j x_j$ is the predictor given by Generalized Linear Regression (GLR) of y on x_j alone. Rank 2 component f^2 is obtained in the same manner replacing every x_j with its OLS regression residuals on f_1, and so on.

  What may be criticized in this extension is the inconsistency in the (implicit) weighting of observations. Indeed, each GLR uses its own weighting system, linked to the estimated variance matrix of observations. As a consequence, GLR of y on x_j alone implicitly uses a weighting matrix W_j specific to $(y,\ x_j)$. Not only are matrices $\{W_j; j = 1,\ p\}$ different from one another, but they are different from the weighting matrix corresponding to GLR of y on X, i.e. that of the model to be regularized. Moreover, when calculating rank 2 component (and further ones), OLS regression is being used, i.e. uniform weighting of observations.

Thus, the estimated variance structure of observations according to the model based on X is never used by this method, as we might expect.

- Bry (2006) has proposed an extension to GLM of a PLSR-related multiblock technique: Thematic Component Analysis (TCA). As PLSR is a particular instance of TCA, this method also extends PLSR to GLM. Both univariate (PLS1) and multivariate (PLS2) PLSR are extended using Generalized Linear Thematic Component Analysis (GLTCA). In the particular case of a single group of explanatory variables X predicting Y, the principle of the method can be described easily. Rank 1 explanatory component f^1 is obtained as follows: GLR of each y_k is performed on X separately, yielding predictor $\hat{z}_k = X\hat{b}_k$. Note that each GLM of y_k on X is entitled to its own weighting matrix, this matrix corresponding to the variance structure of y_k as modeled through X. Let $\hat{Z} = \{\hat{z}_k; k = 1, q\}$. Then, PLS2 of $\hat{Z}$ on X is performed, yielding f^1.

To obtain rank 2 component, one first replaces each x_j with its OLS regression residuals on f^1, which gives a new explanatory group X^1. But then, one performs GLR of each y_k on $\tilde{X}^1 = [X^1, f^1]$ in order to completely deflate the effect of component f^1 in calculating f^2. And so on.

This method has assets and drawbacks.

The first asset is that step 1 calculates correct predictors of each y_k based on the complete X, according to GLM theory. Each predictor is calculated using the correct variance structure. In step 2, the set of these predictors is used through PLS2 with uniform weighting structure to find a common structural predictor. Step 2 may be considered a separate empirical regularization step involving $\hat{Z}$ in itself, and not Y. The uniform weighting it uses may then be advocated easily: weighting in this step does not derive from likelihood maximization, but only reflects a default balance of observations in the regularization process. The second asset is that, if regressors have no structure with respect to this weighting system (for instance if they are uncorrelated), then the method obviously yields the GLR predictor as rank 1 component. Such is not the case for PLSGLR. The third asset is that, in the estimation step, the calculation of the current component correctly deflates the effect of the former ones.

A first drawback of this method is that estimation and regularization remain separated, so that the estimated variance of the regularized GLM does not intervene in its estimation. A second drawback is that collinearity in X leads to technical problems in step 1, since direct GLR becomes then impossible. It is easy to get round this technical problem: a preliminary PCA step can be taken on X, giving component set C, on which to perform GLR of Y.

- In this paper, we propose an extension of Marx's method to the multivariate case. Integrating regularization into the estimation algorithm ensures that on each step, the estimated variance structure of the regularized model is used to estimate it.

Plan of the paper: In section 2, we recall the PLSR mechanism. In section 3, we recall the FSA. In section 4, we show how to nest PLSR within the FSA, and show how it takes the GLM variance structure into account. Finally, in section 5, we study the performance of our algorithm on various simulated data structures.

2. Multivariate PLS Regression

2.1. Rank 1 problem and solution

Let $X = \{x_1, \ldots,\ x^p\}$, $Y = \{y^1, \ldots,\ y^q\}$, $f = Xu$, $g = Yv$, with $u'u = v'v = 1$. Let W be the weighting matrix of observations. The classical rank 1 program of PLSR is:

$$P(X,Y) : \max_{u'u=1\ ;\ v'v=1} \langle Xu|\, Yv\rangle_W$$

It is easy to show that the f solution of P is the same as that of:

$$P'(X,Y) :\ \max_{u'u=1} \sum_{k=1}^{q} \langle Xu|y_k\rangle_W^2$$

Proof:

- Solution of P:

$$L = \langle Xu | Yv \rangle_W - \lambda(u'u - 1) - \mu(v'v - 1)$$

$$\nabla_u L = 0 \Leftrightarrow X'WYv = 2\lambda u \quad (1)\ ; \qquad \nabla_v L = 0 \Leftrightarrow Y'WXu = 2\mu v \quad (1')$$

$(1, 1') \Rightarrow X'WYY'WXu = \eta u \quad (2)$ and $Y'WXX'WYv = \eta v \quad (2')$ with $\eta = 4\lambda\mu$

Besides: $u'(1) \Leftrightarrow 2\lambda = u'X'WYv$, $v'(1') \Leftrightarrow 2\mu = u'X'WYv = 2\lambda = \sqrt{\eta} = \langle Xu | Yv \rangle_W$

which implies that η be maximum. So, solution u is the unit eigenvector u_1 of $X'WYY'WX$ associated with the largest eigenvalue.

- Solution of P':

$$\sum_{k=1}^{q} \langle Xu | y_k \rangle_W^2 = \sum_{k=1}^{q} u'X'Wy_k y_k' WXu = u'X'W \left(\sum_{k=1}^{q} y_k y_k' \right) WXu = u'X'WYY'WXu$$

$$P' : \max_{u'u=1} u'X'WYY'WXu$$

The solution of P' is given by the unit eigenvector u_1 of $X'WYY'WX$ associated with the largest eigenvalue.

2.2. Rank 2 and over

Let $X_0 = X$, and let $f^r = Xu_r$ be rank r component. On step $r + 1$, X_{r-1} is regressed on f^r, with respect to weighting W, leading to residuals:

$$X_r = X_{r-1} - \frac{1}{\|f^r\|_W^2} f^r f^{r\prime} W X_{r-1}$$

Rank $r + 1$ component, f^{r+1}, is found solving $P(X_r, Y)$ or $P'(X_r, Y)$.

2.3. An extended problem

- $P'(X,Y)$ clearly offers the opportunity to extend the program to the case where weighting is different for each variable y_k. Let W_k be a weighting matrix associated with y_k, and consider the program:

  $$P''(X,Y) : \max_{u'u=1} \sum_{k=1}^{q} \langle Xu|y_k\rangle^2_{W_k}$$

  Since:

  $$\sum_{k=1}^{q} \langle Xu|y_k\rangle^2_{W_k} = u'X'\Omega Xu \text{ with } \Omega = \sum_{k=1}^{q} W_k y_k y_k' W_k$$

  The solution of P'' is given by the unit eigenvector u_1 of $X'\Omega X$ associated with the largest eigenvalue.

- Some statistical interpretation remains to be given for program P''. For all k, y_k will be taken W_k-centered, which means:

  $$\forall k : \; y_k = \Pi_{e^{\perp_k}} y_k,$$

  where $e \in \mathbb{R}^n$ has all components equal to 1 and $\perp_k$ refers to orthogonality with respect to metric W_k. As a consequence, observations in X may be centered on any $a \in \mathbb{R}^p$:

  $$\forall k : \; \langle Xu|y_k\rangle^2_{W_k} \; = \; \langle (X - ea')u|y_k\rangle^2_{W_k} \; \forall a \in \mathbb{R}^p$$

  Proof:

  $$\begin{aligned} \forall j : \; \langle Xu|y_k\rangle^2_{W_k} &= \langle Xu|\Pi_{e^{\perp_k}} y_k\rangle^2_{W_k} = \langle \Pi_{e^{\perp_k}} Xu|y_k\rangle^2_{W_k} \\ &= \langle (\Pi_{e^{\perp_k}}(X - ea'))u|y_k\rangle^2_{W_k} = \langle (X - ea')u|\Pi_{e^{\perp_k}} y_k\rangle^2_{W_k} \\ &= \langle (X - ea')u|y_k\rangle^2_{W_k} \end{aligned}$$

Then:

$$\begin{aligned}\forall k:\ \langle Xu|y_k\rangle^2_{W_k} &= \langle (X-e\bar{x}^{k\prime})u|y_k\rangle^2_{W_k} \text{ where } \bar{x}^k = \frac{1}{e'W_k e}X'W_k e \\ &= \|(X-e\bar{x}^{k\prime})u\|^2_{W_k}\|y_k\|^2_{W_k}\cos^2_{W_k}\Big((X-e\bar{x}^{k\prime})u, y_k\Big)\end{aligned}$$

Thus, we find back the classical interpretation of the covariance criterion used by PLS1, as compounding interpretable terms:

- $\|(X-e\bar{x}^{k\prime})u\|^2_{W_k}$ is the variance of the component. Under constraint $u'u=1$, it measures the component's structural strength.
- $\cos^2_{W_k}\Big((X-e\bar{x}^{k\prime})u, y_k\Big)$ measures the goodness of fit of the regression model of y_k on X.

- A problem arises for rank 2 (and higher) components, since we no longer have a single weighting matrix with respect to which component orthogonality could be imposed. So, an extra weighting matrix W has to be chosen for that purpose.

  We can chose uniform weighting $W = I$ to reflect a priori balance of observations, or alternatively chose to keep closer to weights derived from estimation, taking some weighted average of matrices W_k as W.

3. Structure and estimation of the Generalized Linear Model

3.1. Model of Y conditional to X

- Let $y_i = (y_{ki})_{k=1,\ q}$ and $x_i = (x_{ji})_{j=1,\ p}$ respectively be the vector of dependent and explanatory variables for unit i. Conditional to x_i,

$(y_{ki})_{k=1,\ q}$ are assumed independently distributed according to a model having an exponential structure Nelder and Wedderburn (1972):

$$l_k(y_{ki}|\delta_{ki},\phi) = \exp\left(\frac{y_{kt}\delta_{ki} - b_k(\delta_{ki})}{a_{ki}(\phi)} + c_k(y_{ki},\phi)\right)$$

$(\delta_{ki})_{k,i}$ are called *canonical parameters.*

- Let us recall classical results for this structure:

  $\mu_{ki} = E(y_{ki}) = {b_k}'(\delta_{ki})\ \ ;\ \ Var(y_{ki}) = a_{ki}(\phi){b_k}''(\delta_{ki}) = a_{ki}(\phi){b_k}''({b_k}'^{-1}(\mu_{ki}))$

  Let $v_k(\mu_{ki}) = {b_k}''({b_k}'^{-1}(\mu_{ki}))$. Independence of $(y_{ki})_{k=1,\ q}$ conditional to x_i implies that they have conditional variance matrix:

  $$Var(\mathrm{y}_i) = diag\{a_{ki}(\phi)v_k(\mu_{ki})\}_{k=1,q}$$

- We assume that, attached to variables $\{y_1,\ldots,\ y_q\}$, are predictors $\{\eta_1,\ldots,\ \eta_q\}$ that are linear combinations of the x_j's. Then, $\forall k = 1,\ q$, $\eta_k \in \mathbb{R}^n$ can be written:

  $$\eta_k = \alpha_k e + X\beta_k \text{ where } \beta_k \text{ is a } p\text{-coefficient vector.}$$

  Let $B = (\beta_1,\ldots,\beta_q)$ be the $(p,\ q)$ coefficient matrix, $\alpha = (\alpha_1,\ \ldots,\ \alpha_q)'$, and $\eta = [\eta_1,\ldots,\eta_q]$. In matrix form, we have:

  $$\eta = e\alpha' + XB$$

- As usual in GLM, linear predictors and expectations of dependent variables are linked through a *link function* g_k:

  $$\forall k, i:\ \ \eta_{ki} = g_k(\mu_{ki})$$

  The canonical link function corresponding to y_k is: $g_k = b_k'^{-1}$.

3.2. Estimating a GLM through the Fisher scoring algorithm

3.2.1. Univariate GLM

Recall that for the GLM of some variable y, with expectation μ, link function g, and linear predictor $\eta = X\beta$, $\beta \in \mathbb{R}^p$, the log-likelihood of the model is given by:

$$L(\delta; y) = \sum_{i=1}^{n} \left(\frac{y_i \delta_i - b(\delta_i)}{a_i(\phi)} + c(y_i, \phi) \right)$$

Derivation with respect to β yields:

$$\nabla_{\beta} L = 0 \;\Leftrightarrow\; X' W_{\beta}^{-1} \frac{\partial \eta}{\partial \mu}(y - \mu) = 0 \tag{1}$$

with:

$$W_{\beta} = diag\left(g'(\mu_i)^2 a_i(\phi) v(\mu_i)\right)_{i=1,n},$$

and:

$$\frac{\partial \eta}{\partial \mu} = diag\left(\frac{\partial \eta_i}{\partial \mu_i}\right)_{i=1,n} = diag\left(g'(\mu_i)\right)_{i=1,n}.$$

Equation system (1), not linear in β, is solved using the iterative *Fisher scoring algorithm*. On iteration $t+1$:

$$\begin{aligned}
\beta^{[t+1]} &= \beta^{[t]} - \left(E\left[\frac{\partial^2 L}{\partial \beta \partial \beta'}\right]^{[t]} \right)^{-1} \left(\frac{\partial L}{\partial \beta}\right)^{[t]} \\
&= \beta^{[t]} - \left(X' W_{\beta^{[t]}}^{-1} X\right)^{-1} X' W_{\beta^{[t]}}^{-1} \left(\frac{\partial \eta}{\partial \mu}\right)^{[t]} (y - \mu^{[t]}) \\
&= \left(X' W_{\beta^{[t]}}^{-1} X\right)^{-1} X' W_{\beta^{[t]}}^{-1} z_{\beta^{[t]}} \qquad (2)
\end{aligned}$$

where:

$$z_{\beta^{[t]}} = X\beta^{[t]} + \left(\frac{\partial \eta}{\partial \mu}\right)^{[t]} (y - \mu^{[t]})$$

Equation (2) may be interpreted as the GLS normal equations of the following linear model, on iteration t:

$$M^{[t]}: \ z_{\beta^{[t]}} = X\beta + \zeta^{[t]}$$

where: $E(\zeta^{[t]}) = 0 \quad ; \quad V(\zeta^{[t]}) = W_\beta^{[t]} = g'^2(\mu_t)V(y_t)$.
We shall refer to $M^{[t]}$ as the (current) *linearized model*.

One important point is that GLS estimation of this model is nothing but a Quasi-Likelihood Estimation (QLE). This estimation by maximum of QL mimics MLE on each step, under a normality and independence assumption of the $z_{\beta^{[t]}}$'s with a fixed covariance structure.
Note: as the 1^{st} order development of g at point μ yields:

$$g(y) \approx g(\mu) + g'(\mu)(y-\mu) = z,$$

we may perform OLSR of $g(y)$ on X, in order to get an initial value $\beta^{[0]}$. When $g(y)$ is not defined owing to zero-values in data, we have to mix y up with some relevant quantity. We propose to take:

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

3.2.2. Multivariate GLM with common predictor (MGLMCP)

We are now considering a multivariate approach to GLM (for an overview, see Fahrmeir and Tutz (1994)). Let variables $y_1, \ldots, y_q$ depend on the "same" linear predictor (in fact predictors collinear to one another), conditional to which they are all independent. To be more precise:

$$\forall k = 1, q: \ \eta_k \ = \ X\beta_k \ = \ X\gamma_k u$$

For obvious identification purposes, we impose constraint $u'u = 1$.

In view of the conditional independence assumption, and of the independence of units:

$$l(y|\eta) = \prod_{i=1}^{n} \prod_{k=1}^{q} l_k(y_{ki}|\eta_{ki})$$

As a result, the corresponding linearized model in the FSA is the following:

$$\forall k = 1, q : \quad z_{k\beta_k} = X\gamma_k u + \zeta_k,$$

where the ${\zeta_k}'$s are independent and $\forall k :\ E(\zeta_k) = 0$; $V(\zeta_k) = W_{k\beta_k}$.

The FSA is used to estimate this model, with some modification, owing to u and $\gamma = (\gamma_k)_{k=1,q}$. Indeed, estimation of model $M^{[t]}$ is carried out iterating the following alternated least squares two steps sequence:

(i) Given γ, vector $\left(z_{k\beta_k}\right)_k \in \mathbb{R}^{nq}$ is regressed on matrix $\gamma \otimes X$, with respect to variance matrix $W_\beta = diag\left(W_{k\beta_k}\right)_k$. The resulting coefficient vector $\hat{u}$ is made unit-norm, yielding new u.

(ii) Given Xu, each $z_{k\beta_k}$ is regressed independently on Xu, with respect to variance matrix $W_{k\beta_k}$, yielding new β_k.

The fixed point values of u and g of these iterations are taken as $u^{[t]}$ and $\gamma^{[t]}$.

4. Component-based Generalized Linear Regression: principle and basic algorithm

The above-mentioned mechanisms can now be inter-woven to form a CGLR algorithm.

4.1. Rank 1 component f^1

The basic principle of the method we propose is quite simple: on each step of the FSA in the estimation of the MGLMCP, we replace the GLS regression step with a multivariate PLS regression one.

Thus, at step k of the FSA, the regression of the z's in the MGLMCP is the solution, as far as u is concerned, of several equivalent programs (for simplicity's sake, let us write z_k for $z_{k\beta_k}$, and W_k for $W_{k\beta_k}$):

$$Q1: \underset{\gamma,u:u'u=1}{Min} \sum_k \|z_k - X\gamma_k u\|^2_{W_k} \Leftrightarrow Q2: \underset{u:u'u=1}{Min} \sum_k \|z_k - \Pi_{Xu} z_k\|^2_{W_k}$$

$$\|z_k - \Pi_{Xu} z_k\|^2_{W_k} = \|z_k\|^2_{W_k} \sin^2_{W_k}(z_k, Xu) = \|z_k\|^2_{W_k}(1 - \cos^2_{W_k}(z_k, Xu))$$

So:

$$Q2 \Leftrightarrow Q3: \max_{u:u'u=1} \sum_k \|z_k\|^2_{W_k} \cos^2_{W_k}(z_k, Xu)$$

We propose to currently replace Q3, in the MGLMCP estimation algorithm, by:

$$R: \max_{u:u'u=1} \sum_k \|z_k\|^2_{W_k} \cos^2_{W_k}(z_k, Xu) \|Xu\|^2_{W_k} \Leftrightarrow \max_{u:u'u=1} \sum_k \langle z_k | Xu \rangle^2_{W_k}$$

$$\Rightarrow \quad R = P''(Z, X)$$

where P'' is the extended PLSR program studied in section 2.3.

As a consequence of §2.3, the current solution $u^{[t]}$ is the unit eigenvector associated with the largest eigenvalue of the following matrix:

$$X'\Omega^{[t]}X \text{ with } \Omega^{[t]} = \sum_{k=1}^{q} W_k^{[t]} z_k^{[t]} z_k^{[t]\prime} W_k^{[t]}$$

where the z_k's have been W_k-centred.

N.B. Program R's criterion not being zero-degree homogenous in $W_k^{[t]}$, it is important that all $W_k^{[t]}$ be currently normalized to unit-sum.

4.2. Rank $\geq$ 2 components

4.2.1. Orthogonality of components

We shall ensure zero-correlation of components with respect to a given fixed weighting W (i.e. the f^k will form an W-orthogonal system). Indeed, weighting is not linked here to the variance of the dependent variables, since it does not derive from estimation optimality concerns. If all observations are considered equally important, we must take $W = \frac{1}{n}I_n$.

So, let:

$$f^r = X^{r-1}u^r \text{ with } X^0 = X \text{ and } X^r = \Pi_{\langle f^r \rangle^{W\text{-}\perp}} X^{r-1} \tag{3}$$

4.2.2. Role of every extra component

Every extra component f^r must complement the existing ones $F^{r-1} = \{f^1, \ldots, f^{r-1}\}$ as well as possible. So, as far as f^r is concerned, F^{r-1} must be viewed as a group of covariates. Now:

$$\cos^2_W(z, \langle F^{r-1}, f^r \rangle) = \cos^2_W(z, \langle F^{r-1} \rangle) + \cos^2_W(z, \langle \Pi_{\langle F^{r-1} \rangle^{W-\perp}} f^r \rangle) \tag{4}$$

$$\text{where } \Pi_{\langle F^{r-1} \rangle^{W-\perp}} f^r = \Pi_{\langle F^{r-1} \rangle^{W-\perp}} X^{r-1} u^r = \tilde{X}_W^{r-1} u^r \tag{5}$$

$$\text{with } \tilde{X}_W^{r-1} = X^{r-1} - F^{r-1}(F^{r-1\prime} W F^{r-1})^{-1} F^{r-1\prime} W X^{r-1}$$

Let us take a look back at the MGLMCP. Supposing we already have $r-1$ available components for $\{z_k\}_{k=1,q}$ and we want to look for the best possible r^{th} common component. This component should be the solution of the following program (having the form of $Q3$):

$$\max_{f^r \in \langle X^{r-1} \rangle} \sum_k \ \|z_k\|^2_{W_k} \cos^2_{W_k}(z_k, \langle F^{r-1}, f^r \rangle)$$

According to (4) and (5), this is equivalent to:

$$\max_{u^r} \sum_k \ \|z_k\|^2_{W_k} \cos^2_{W_k}(z_k, \tilde{X}^{r-1}_{W_k} u^r)$$

To account for residual variance structure in X^{r-1}, we propose to replace this latter program by:

$$\max_{u^r : u^{r\prime} u^r = 1} \sum_k \ \langle z_k | \tilde{X}^{r-1}_{W_k} u^r \rangle^2_{W_k}$$

So, the solution is the unit-norm eigenvector associated with the largest eigenvalue of matrix:

$$\left[\sum_k \widetilde{X}^{r-1\prime}_{W_k} W_k z^k z^{k\prime} W_k \tilde{X}^{r-1}_{W_k} \right]$$

4.3. Algorithm

The complete algorithm used to calculate a set of R components according to these principles may be found in appendix A (algorithm $\mathbf{A}_0$).

4.4. Predictive model

Once calculated the components, they are used to produce a set of coefficients of the original explanatory variables in a predictive model of Y. We want to write an expression of the form:

$$X^r = X\pi_r \tag{6}$$

From (3) and (6), we get:

$$f^r = X\pi_{r-1}u^r \ = Xv^r \ \text{ with } \ v^r = \pi_{r-1}u^r \tag{7}$$

which leads to:

$$\begin{aligned} X^r &= X\pi_{r-1} - \frac{1}{f^{r\prime}Wf^r} f^r f^{r\prime} WX\pi_{r-1} = X\pi_{r-1} - \frac{1}{f^{r\prime}Wf^r} X\pi_{r-1}u^r f^{r\prime}WX\pi_{r-1} \\ &= X\left[Id_p - \frac{1}{f^{r\prime}Wf^r}\pi_{r-1}u^r f^{r\prime}WX\right]\pi_{r-1} \end{aligned}$$

Hence the recurrence formula:

$$\pi_r = \left[Id_p - \frac{1}{f^{r\prime}Wf^r}\pi_{r-1}u^r f^{r\prime}WX\right]\pi_{r-1}$$

from which we draw $V = [v^1|\ldots|v^R]$ in view of (7).

Then, estimating the GLM of Y on $F = XV$ along with the constant e yields the predictor matrix H:

$$H = ea + FC = ea + XB \quad \text{with} \quad B = VC \tag{8}$$

N.B. If X has been standardized prior to GLPLS/CGLR and one wants the coefficients of the unstandardized X in the model, then:

Let X_o denote the original unstandardized explanatory variable matrix, and X the standardized one. We have:

$X = (X_o - e(e'We)^{-1}e'WX_o)\Lambda^{-1},$ where $\Lambda = diag(\sigma_k),\ \sigma_k^2 = V(x_k)\ \forall k = 1,\ q.$

So, we have:

$$\begin{aligned} H &= ea + (X_o - e(e'We)^{-1}e'WX_o)\Lambda^{-1}B \\ &= e(a - (e'We)^{-1}e'WX_o\Lambda^{-1}B) + X_o\Lambda^{-1}B \end{aligned}$$

Hence the model constants: $a - (e'We)^{-1}e'WX_o\Lambda^{-1}B$ and coefficients of variables: $\Lambda^{-1}B$.

5. CGLR: an enhanced algorithm

In order to solve convergence problems of the above-given algorithm in some situations, we have added two tuning parameters giving flexibility to the combination of PLS and GLM estimation.

5.1. Tuning the attraction of predictors towards principal components

Recall that the solution of P'' is given by the unit eigenvector u_1 of $X'\Omega X$ associated with the largest eigenvalue, where:

$$\Omega = \sum_{k=1}^{q} W_k y_k {y_k}' W_k$$

It is possible to fine-tune the attraction of the corresponding component towards X's principal components by taking instead the unit eigenvector u_1 associated with the largest eigenvalue, of the following matrix:

$$A_s = (X'WX)^s X'\Omega X$$

where s is a tuning parameter indicating the strength we intend to give to this attraction, and W is the weighting matrix with which we intend to perform PCA (typically, $W = \frac{1}{n} I_n$). To understand this, let us review two particularly important values :

- $s = 0$ gives back the solution of P''.
- $s \to \infty$:
  then, u_1 is the unit eigenvector of $X'WX$ associated with its largest eigenvalue, so $f_1 = Xu_1$ is precisely $X's$ first CP in the PCA of X

weighted by W: with infinite attraction, the dependent y_k's no longer play any role in component extraction.

In the sequel of this section, we shall refer to extracting the first eigenvector of A_s as to performing a "tuned" PLS step.

5.2. Tuning the rate of the FSA steps with respect to the PLS steps in the combination

We may chose the number of steps of the FSA to be performed in between each tuned PLS step. Informally: given components F, a certain number of FSA steps of Y on F are performed, possibly until convergence, yielding variables z_k and corresponding W_k. Then, the tuned PLS step of z_k's on X updates the components, and so on.

This enables to eventually get a converging algorithm. Indeed, pushing s far enough, we get components that weakly vary about PC's. Operating on thus "stabilized" and uncorrelated components, the FSA itself is most likely to converge. Such convergence is of course paid for with less freedom for components to adjust the explanatory model.

5.3. Algorithm

The enhanced algorithm may be found in appendix B (algorithm $\mathbf{A}_1$).

6. Numerical results on simulated data

6.1. Data generation

The less easy type data to deal with is binary variables, for their values are usually never close to their expectation. So, we chose to use binary dependent variables in our simulations. Our simulation scheme is as follows. We consider $n = 100$ units.

- 100 explanatory variables X are simulated so as to be structured around two uncorrelated factors $\{\phi^1, \phi^2\}$.

  - Simulation of $\{\phi^1, \phi^2\}$:

    - Simulate a vector γ of 100 random numbers uniformly distributed on [0;1] (abbrev. 100 r.n. $U_{[0;1]}$), and take $\phi^1 =$ standardized γ.

    - Simulate a vector δ_1 of 100 r.n. $U_{[0;1]}$, take $\delta_2 = (\delta_1 - \frac{1}{n}\phi^1\phi^{1\prime}\delta_1)$ and finally $\phi^2 =$ standardized δ_2.

    Thus, we get two uncorrelated standardized factors.

    Let now a be a parameter tuning noise about factors (roughly, the tangent of the semi-angles of the bundles), and ranging from 1/5 (reduced noise) to 2 (important noise).

  - Simulation of a first bundle of variables, X_1, structured around ϕ^1. For $j = 1$ to $p_1 = 70$:

    - Simulate a vector κ^j of 100 r.n. $U_{[0;1]}$, and take $\epsilon^j =$ standardized κ^j.

    - Let $\lambda^j = \epsilon^j + \alpha_j\phi^2$ where $\alpha_j =$ r.n. $U_{[-1/5;+1/5]}$, and $\gamma^j =$ standardized λ^j.

      N.B. This step is necessary to inject a bit of ϕ^2 into the variables. Indeed, if their deviations from ϕ^1 were obtained as vectors of random numbers, they would be almost systematically orthogonal to ϕ^2.

    - ∗ Let $\xi^j = \phi^1 + a\gamma^j$, and x^j = standardized ξ^j.
  - Simulation of a second bundle of variables, X_2, structured around ϕ^2. For $j = 1$ to $p_2 = 30$:
    - ∗ Simulate a vector κ^j of 100 r.n. $U_{[0;1]}$, and take ϵ^j = standardized κ^j.
    - ∗ Let $\lambda^j = \epsilon^j + \alpha_j \phi^1$ where α_j = r.n. $U_{[-1/5;+1/5]}$, and γ^j = standardized λ^j.
    - ∗ Let $\xi^j = \phi^2 + a\gamma^j$, and x^j = standardized ξ^j.
  - $X = [X_1, X_2]$

- Dependent variables Y are simulated as follows:
  - For $k = 1$ to $q = 10$, simulate $y^k \sim B(1, p_k(\phi^1, \phi^2))$ with:
$$\ln\left[\frac{p_k(\phi^1, \phi^2)}{1 - p_k(\phi^1, \phi^2)}\right] = a_{k1}\phi^1 + a_{k2}\phi^2$$
    where, for $\quad h = 1,\ 2 : a_{kh} \sim U_{[-\frac{2}{3};+\frac{2}{3}]}$

For each value of a, we used the simulation scheme 100 times, each time yielding a pair (X, Y), on which we ran the estimation procedure asking for 3 components. For each such simulated (X, Y), estimation was carried out using algorithm $\mathbf{A}_0$ as follows: starting with $s = 0$, if convergence threshold $(sin^2(f^{k[m]}, f^{k[m+1]}){<}10^{-8})$ could not be reached in less then 100 iterations (most of the time, less than ten were enough), then increment s by 1 and try again. Convergent estimation giving components denoted (f^1, f^2, f^3) , we calculated all square correlations $\{\rho^2(\phi^k, f^l);\ k = 1, 2;\ l = 1, 2, 3\}$.
Then, for each value of a, we calculated the mean, over all simulations, of each $\rho^2(\phi^k, f^l)$ and also that of the smallest s leading to converging estimation.

6.2. Results

The distribution of the smallest required s according to the value of noise parameter a may be found in table 2.

Unsurprisingly, the mean value of s mostly increases with noise for each of the first two components (which are supposedly the two relevant ones, in view of the model). But even with the highest level of noise, the algorithm never had to go up to $s = 3$ in order to converge. The square correlations of the components with the factors are satisfactory (see table 1). In fact, components f are more or less drawn towards principal components of X, and these have generally no reason to be individually very close to the factors underlying the bundles, especially as soon as these factors are somewhat correlated (which was not the case here). So, these square correlations matter less than their sums: $R^2_{K,L} = \frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{L}\rho^2(\phi^k, f^l)$. Indeed, $R^2_{K,L}$ close to 1 means that estimation has captured explanatory space $\langle\{\phi^k\}_{k=1,K}\rangle$ with component space $\langle\{f^l\}_{l=1,L}\rangle$. What is important is to check that, K being the true number of underlying factors, $\frac{1}{K}R^2_{K,K} \approx 1$. This was clearly the case in our simulation, even with the highest degree of noise about factors (see table 1).

When component space $\langle\{f^1, f^2\}\rangle$ almost perfectly captures explanatory space $\langle\{\phi^1, \phi^2\}\rangle$, component f^3 can have no explanatory role. As a consequence, its direction can only depend on the structural strength it captures in array X^2 it is extracted from, so it should be close to X^2's first PC. It then stands to reason that the more important the noise gets, the less of $\langle\{\phi^1, \phi^2\}\rangle$ is captured by $\langle\{f^1, f^2\}\rangle$; an increasing part of $\langle\{\phi^1, \phi^2\}\rangle$ is then left in X^2, which may at times help the algorithm focus on f^3.

Table 1: Factor-component square correlations according to the amount of model simulation noise

a= (Noise increasing)	$1/5$		$1/4$		$1/3$		$1/2$		1		2	
Average Square Correlations	ϕ^1	ϕ^2	ϕ^1	ϕ^2	ϕ^1	ϕ^2	ϕ^1	ϕ^2	ϕ^1	ϕ^2	ϕ^1	ϕ^2
f^1	.94	.06	.95	.05	.93	.06	.94	.06	.92	.06	.88	.05
f^2	.06	.94	.05	.95	.06	.93	.06	.93	.07	.83	.06	.61
f^3	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	2.10^{-3}	10^{-4}	6.10^{-3}	8.10^{-4}	7.10^{-2}	.01	.21
$\frac{1}{2}R^2_{2,2}$	1		1		.99		1		.98		.91	

Table 2: Distribution of required structural strength parameter according to maount of model simulation noise

a= (Noise increasing)		1/5			1/4			1/3			1/2			1			2		
f^1	Average s	.10			.13			.15			.17			.13			.33		
	s=	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
	Distribution of s	.90	.10	0	.88	.11	.01	.87	.11	.02	.83	.17	0	.87	.13	0	.70	.27	.03
f^2	Average s	0			.05			.05			.19			.45			.60		
	s=	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
	Distribution of s	1	0	0	.95	.05	0	.95	.05	0	.81	.19	0	.55	.45	0	.43	.54	.03
f^3	Average s	1.01			1.01			1.03			.99			.95			.93		
	s=	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
	Distribution of s	0	.99	.01	0	.99	.01	.01	.95	.04	.01	.99	0	.06	.93	.01	.02	.85	.04

7. Conclusion

IRPLS being, according to us, the only extension of PLS Regression to GLM that respects the variance structure of that model, we have tried to extend it to multivariate dependent variables. In the current GLS step of the Fisher scoring algorithm, we have introduced some multivariate PLS-type regularization. As the algorithm it leads to may encounter convergence problems in some cases, we have introduced a numeric parameter that allows to continuously tune the attraction of explanatory components towards the principal components of explanatory variables. In view of this flexibility bonus, the algorithm proved to always converge, and yield very satisfactory results on simulated data.

References

Bastien, P., Esposito Vinzi, V., Tenenhaus, M.. Pls generalized linear regression. CSDA 2005;48(1):17–46.

Bry, X.. Extension de l'analyse en composantes thématiques univariée au modèle linéaire généralisé. RSA 2006;54(3).

Fahrmeir, L., Tutz, G.. Multivariate Statistical Modeling Based on Generalized Linear Models. New York, USA: Springer-Verlag, 1994.

Marx, D.. Iteratively reweighted partial least squares estimation for generalized linear regression. Technometrics 1996;34(4):374–381.

McCullagh, P., Nelder, J.. Generalized linear models. New York, USA: Chapman and Hall, 1989.

Nelder, J., Wedderburn, R.. Generalized linear models. Journal of the Royal Statistical Society: Series A 1972;135:370–384.

Appendix A

We get the following algorithm, denoted $\mathbf{A}_0$:

```
Initialization
 Let: $X^0 = X$ ; $\forall k = 1, q$ : $\tilde{X}^0_{W_k} = X$  and $F^0 = \emptyset$
Component iteration
 For $r = 1$ to $R$:
    Calculate $f^r$ as follows:
       Initialize $Z = [z_1|\ldots|z_q]$ to $Z^{[0]}$ and $\{W_k\}_{k=1,q}$ to $\{W_k^{[0]}\}_{k=1,q} = \{\frac{1}{n} Id_n\}_{k=1,q}$
       Iterate from $m = 0$, until convergence:
          For $k = 1$ to $q$:
             Standardize every $z_k^{[m]}$ with respect to $W_k^{[m]}$
             If $r > 1$, set: $\tilde{X}^{r-1}_{W_k^{[m]}} = X^{r-1} - F^{r-1}(F^{r-1\prime} W_k^{[m]} F^{r-1})^{-1} F^{r-1\prime} W_k^{[m]} X^{r-1}$
          Define $u_r^{[m]}$ as the unit-norm eigenvector associated with the largest eigenvalue
          of matrix:
                $\left[\sum_k \tilde{X}^{r-1}_{W_k^{[m]}}{}' W_k^{[m]} z_k^{[m]} z_k^{[m]\prime} W_k^{[m]} \tilde{X}^{r-1}_{W_k^{[m]}}\right]$
          Set $f^{r[m]} = X^{r-1} u_r^{[m]}$
          For $k = 1$ to $q$:
             Carry out GLS regression with respect to weighting $W_k^{[m]}$ of each model:
                $z_k^{[m]} = \gamma_{k,0} + F^{r-1}[\gamma_{k,1}, \ldots, \gamma_{k,r-1}]' + f^{r[m]} \gamma_{k,r} + \zeta_k$
             thus getting coefficient vector $\gamma_k^{[m]} = (\gamma_{k,0}^{[m]}, \ldots, \gamma_{k,r}^{[m]})$
             Update $z_k^{[m]}$ and $W_k^{[m]}$ using $\gamma_k^{[m]}$
    Set $F^r = [F^{r-1}, f^r]$
    Calculate next current $X$ array:
          $X^r = \Pi_{\langle f^r \rangle^{W-\perp}} X^{r-1}$
```

Appendix B

We get the following algorithm, denoted $\mathbf{A}_1$:

```
Initialization
 Let: $X^0 = X$ ; $\forall k = 1, q$ : $\tilde{X}^0_{W_k} = X$ , $F^0 = \emptyset$
Component iteration
 For $r = 1$ to $R$:
   Calculate $f^r$ as follows:
     Initialize $Z = [z_1|\ldots|z_q]$ to $Z^{[0]}$ and $\{W_k\}_{k=1,q}$ to $\{W_k^{[0]}\}_{k=1,\ q} = \{\frac{1}{n}Id_n\}_{k=1,\ q}$
     Iterate from $m = 0$, until convergence:
       For $k = 1$ to $q$:
         Standardize every $z_k^{[m]}$ with respect to $W_k^{[m]}$
         If $r$>1, set: $\tilde{X}^{r-1}_{W_k^{[m]}} = X^{r-1} - F^{r-1}(F^{r-1\prime}W_k^{[m]}F^{r-1})^{-1}F^{r-1\prime}W_k^{[m]}X^{r-1}$
       Define $u_r^{[m]}$ as the unit-norm eigenvector associated with the largest eigenvalue
       of matrix:
           $(X^{r-1\prime}WX^{r-1})^s \left[\sum_k \tilde{X}^{r-1}_{W_k^{[m]}}{}'W_k^{[m]}z_k^{[m]}z_k^{[m]\prime}W_k^{[m]}\tilde{X}^{r-1}_{W_k^{[m]}}\right]$
       Set: $f^{r[m]} = X^{r-1}u_r^{[m]}$
       For $k = 1$ to $q$:
         Set $z_k^{[m,1]} = z_k^{[m]}, W_k^{m,1]} = W_k^{[m]}$ and $f^{r[m,1]} = f^{r[m]}$
         and from $l = 1$ until some convergence precision is reached:
           Carry out the current step of the FSA, i.e. GLS regression with respect to
           weighting $W^{k[m,l]}$ of each model:
               $z_k^{[m,l]} = \gamma_{k,0} + F^{r-1}[\gamma_{k,1}, \ldots, \gamma_{k,r-1}]' + f^{r[m,l]}\gamma_{k,r} + \zeta_k$
           thus getting coefficient vector $\gamma_k^{[m,l+1]} = (\gamma_{k,0}^{[m,l+1]}, \ldots, \gamma_{k,r}^{[m,l+1]})$
           Update $z_k^{[m,l+1]}$ and $W_k^{[m,l+1]}$ using $\gamma_k^{[m,l+1]}$
         Update $z_k^{[m]} = z_k^{[m,\infty]}, W_k^{[m]} = W_k^{[m,\infty]}$ and $\gamma_k^{[m]} = \gamma_k^{[m,\infty]}$
```

Set $F^r = [F^{r-1}, f^r]$

Calculate next current X array:

$$X^r = \Pi_{\langle f^r \rangle^{W\text{-}\perp}} X^{r-1}$$