Journal of Multivariate Analysis 119 (2013) 47–60

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm

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

[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

[d] Cirad, UR B&SEF, Biens et Services des Ecosystèmes Forestiers Tropicaux, Campus International de Baillarguet, TA C-105/D - 34398, Montpellier, France

ARTICLE INFO

Article history:
Received 30 September 2011
Available online 12 April 2013

AMS subject classifications:
62-07
62H25
62J12

Keywords:
Supervised component generalized linear regression
Generalized linear models
PLS regression
Fisher scoring algorithm

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 a technique, Supervised Component Generalized Linear Regression (SCGLR), that combines PLS regression with GLM estimation in the multivariate context. SCGLR is tested on both simulated and real data.

1. Introduction

1.1. Framework

The framework is that of a multivariate Generalized Linear Model (GLM): 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 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 on X. All variables are measured on the same n statistical units. The assumption of conditional independence means that the statistical link between the responses is due to their common explanatory variables *only*. In our application on real data (cf. Section 7), we aim at predicting the presence/absence of $q = 10$ common tree species of the Congo Basin rainforests measured on $n = 3000$ plots in the Central African Republic. Y is thus a set of 10 binary variables. We use $p = 46$ environmental regressors reflecting the climate, topography, location, stand structure and photosynthetic activity of each plot. One key point is that we are interested in explanatory structures common to part or all of the y^k's. Another key point is that we

* Corresponding author at: Institut de Mathématiques et de Modélisation de Montpellier, Université Montpellier II, Place Eugène Bataillon CC 051 - 34095, Montpellier, France.
E-mail addresses: bry@math.univ-montp2.fr (X. Bry), catherine.trottier@univ-montp2.fr, trottier@math.univ-montp2.fr (C. Trottier), thomas.verron@fr.imptob.com (T. Verron), frederic.mortier@cirad.fr (F. Mortier).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.03.013

want to be able to deal with many and possibly correlated regressors, so that efficient dimension reduction is needed in the regressor space. We may think of other typical problems: modeling q Poisson-distributed event counts (e.g. failures by type of failure) in a complex system as a function of structural characteristics of the system; modeling q random survival times per unit (e.g. lags between stages of a disease in epidemiology) as a function of the unit's characteristics, etc.

The standard estimation of a GLM 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 focused 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 most practical situations, explanatory dimensions are not identified well enough to be measured each through a single variable. So, several indirect measures have to be included into the regressors for each such dimension. This yields many and highly correlated regressors. It is possible to perform some PCA on regressors in order to capture a few uncorrelated principal components (PC's) accounting for a sufficient part of the regressors' information, and use these components as new regressors for the GLM estimation. This Principal Components Generalized Linear Regression (PCGLR) has one possible drawback: PC's optimally capture the information of X *per se*, but not chiefly the information most useful to predict Y.

In order to direct the calculation of components towards the prediction of Y 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. In the classical linear model framework, PLSR is a successful alternative to PCR (Principal Component Regression). In PLSR just as in PCR, components are definite linear combinations of the x's. Regressing Y on components yields a prediction formula that can then be expressed in terms of the x's. Both methods are a way of regularizing regression, in that they drastically limit the transfer of effects between the x's. PLSR performs better than PCR because it takes Y into account when calculating components. But the PLSR criterion is naturally adapted to the linear context, and not to the GLM one.

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

When there is but one response y to be modeled, Marx [4] 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 reweighted least squares (IRLS) procedure [5], 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 responses.

Following that line and for want of any better method, it could seem handy to deal with multiple responses $\{y^1, \ldots, y^q\}$ by first performing IRPLS with each y^k separately, getting a specific predictor component g^k, then performing PCA on $\{g^1, \ldots, g^q\}$ and taking their first PC f^1 as the overall first predictor component. However, this f^1 would be more of a structure common to separate predictor components than a common predictor component and, even if they may not be far apart in many cases, there is some difference between the two. On the one hand, there clearly is a difference in the variance structure used for estimation: when determining separately the predictor component g^k of y^k, the variance matrix W_k used iteratively is determined by this component which is unconstrained by the other y^k's. By contrast, calculating a common predictor component should use variance matrices determined by this component, which is constrained by *all* y^k's. On the other hand, it can be shown, in the classical context of linear modeling, that PCA on multiple separate univariate PLS regressions (PLS1) does not lead to multivariate PLS regression (PLS2). As a consequence, the question of a genuine GLM extension of PLS2 has to be dealt with.

Still in the single y context, Bastien et al. [1] have proposed a different way to extend PLS1 to GLM: PLS Generalized Linear Regression (PLSGLR). PLSGLR is based on the following property: PLS1 of a quantitative variable z on $X = \{x^1, \ldots, x^p\}$ yields a rank 1 component f^1 collinear to the sum of the predictors given by OLS regression of z on each x^j alone. Rank 2 component is obtained likewise after replacing each x^j with its OLS regression residuals on f^1, and so on. Hence an apparently straightforward GLM extension of PLS1: given response y, f^1 of PLSGLR is defined as the standardized sum of predictors given by Generalized Linear Regression (GLR) of y on each x^j alone. What may seem awkward in this extension is the inconsistency in the weighting of observations. Indeed, GLR of y on x^j alone implicitly uses a weighting matrix W_j specific to $(y,\ x^j)$, which is different from the weighting matrix associated with GLR of y on components.