Using principal components for estimating logistic regression with high-dimensional multicollinear data

- Ana M. Aguilera, Manuel Escabias, Mariano J. Valderrama

- Using principal components for estimating logistic regression with high-dimensional multicollinear data
- *Computational Statistics & Data Analysis, Volume 50, Issue 8, 2006, Pages 1905-1924*
- DOI: https://doi.org/10.1016/j.csda.2005.03.011

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 50 (2006) 1905–1924
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Using principal components for estimating logistic regression with high-dimensional multicollinear data

Ana M. Aguilera*, Manuel Escabias, Mariano J. Valderrama

Department of Statistics and O.R., University of Granada, Spain

Received 3 December 2004; received in revised form 31 March 2005; accepted 31 March 2005
Available online 25 April 2005

Abstract

The logistic regression model is used to predict a binary response variable in terms of a set of explicative ones. The estimation of the model parameters is not too accurate and their interpretation in terms of odds ratios may be erroneous, when there is multicollinearity (high dependence) among the predictors. Other important problem is the great number of explicative variables usually needed to explain the response. In order to improve the estimation of the logistic model parameters under multicollinearity and to reduce the dimension of the problem with continuous covariates, it is proposed to use as covariates of the logistic model a reduced set of optimum principal components of the original predictors. Finally, the performance of the proposed principal component logistic regression model is analyzed by developing a simulation study where different methods for selecting the optimum principal components are compared.

Keywords: Logistic regression; Multicollinearity; Principal components

1. Introduction

There are many fields of study such as medicine and epidemiology, where it is very important to predict a binary response variable, or equivalently the probability of occurrence of an event (success), in terms of the values of a set of explicative variables related to it. That

* Corresponding author. Tel.: +34 958243270; fax: +34 958243267.
E-mail address: aaguiler@ugr.es (A.M. Aguilera).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.03.011

is the case of predicting, for example, the probability of suffering a heart attack in terms of the levels of a set of risk factors such as cholesterol and blood pressure. The logistic regression model serves admirably this purpose and is the most used for these cases as we can see, for example, in Prentice and Pyke (1979).

As many authors have stated (Hosmer and Lemeshow (1989) and Ryan (1997), among others), the logistic model becomes unstable when there exists strong dependence among predictors so that it seems that no one variable is important when all the others are in the model (multicollinearity). In this case the estimation of the model parameters given by most statistical packages becomes too inaccurate because of the need to invert near-singular and ill-conditioned information matrices. As a consequence, the interpretation of the relationship between the response and each explicative variable in terms of odds ratios may be erroneous. In spite of this the usual goodness-of-fit measures show that in these cases the estimated probabilities of success are good enough. In the general context of generalized linear models, Marx and Smith (1990) and Marx (1992) solve this problem by introducing a class of estimators based on the spectral decomposition of the information matrix defined by a scaling parameter.

As in many other regression methods, in logistic regression it is usual to have a very high number of predictor variables so that a reduction dimension method is needed. Principal component analysis (PCA) is a multivariate technique introduced by Hötelling that explains the variability of a set of variables in terms of a reduced set of uncorrelated linear spans of such variables with maximum variance, known as principal components (pc's). The purpose of this paper is to reduce the dimension of a logistic regression model with continuous covariates and to provide an accurate estimation of the parameters of the model avoiding multicollinearity. In order to solve these problems we propose to use as covariates of the logistic model a reduced number of pc's of the predictor variables.

The paper is divided into four sections. Section 1 is an introduction. Section 2 gives an overview of logistic regression. Section 3 introduces the principal component logistic regression (PCLR) model as an extension of the principal component regression (PCR) model introduced by Massy (1965) in the linear case. It also proposes two different methods to solve the problem of choosing the optimum pc's to be included in the logit model. One is based on including pc's in the natural order given by their explained variances, and in the other pc's are entered in the model by a stepwise method based on conditional likelihood-ratio-tests that take into account their ability to explain the response variable. The optimum number of pc's needed in each method (stopping rule) is also boarded in Section 3 where we propose and discuss several criteria based on minimizing the error with respect to the estimated parameters. Finally, accuracy of estimations provided by the proposed PCLR models and performance of different methods for choosing the optimum models will be tested on a simulation study in Section 4. The results will also be compared with those provided by the partial least-squares logit regression (PLS-LR) algorithm proposed by Bastien et al. (2005) for estimating the logistic regression model.

2. Basic theory on logistic regression

In order to establish the theoretical framework about logistic regression we will begin by formulating the model, estimating its parameters and testing its goodness of fit.

Let $\mathscr{X}_1, \mathscr{X}_2, \ldots, \mathscr{X}_p$ be a set of continuous variables observed without error and let us consider n observations of such variables that will be resumed in the matrix $\mathscr{X} = (x_{ij})_{n\times p}$. Let $Y = (y_1, \ldots, y_n)'$ be a random sample of a binary response variable $\mathscr{Y}$ associated with the observations in $\mathscr{X}$, that is, $y_i \in \{0, 1\},\ i = 1, \ldots, n$. Then, the logistic regression model is given by

$$y_i = \pi_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where π_i is the expectation of $\mathscr{Y}$ given $(\mathscr{X}_1 = x_{i1}, \mathscr{X}_2 = x_{i2}, \ldots, \mathscr{X}_p = x_{ip})$ that is modelized as

$$\pi_i = P\left\{\mathscr{Y} = 1 \,|\mathscr{X}_1 = x_{i1}, \ldots, \mathscr{X}_p = x_{ip}\right\} = \frac{\exp\left\{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right\}}{1 + \exp\left\{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right\}}, \tag{2}$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are the parameters of the model and ε_i are zero mean independent errors whose variances are given by $Var[\varepsilon_i] = \pi_i\,(1 - \pi_i)\,,\ i = 1, \ldots, n$.

Let us define the logit transformations $l_i = \ln\left(\pi_i \,/(1 - \pi_i)\right),\ i = 1, \ldots, n$, where $\pi_i \,/(1 - \pi_i)$ represents the odds of response $\mathscr{Y}=1$ for the observed value $x_i = (x_{i1}, \ldots, x_{ip})$. Then, the logistic regression model can be seen as a generalized linear model with the logit transformation as link function (McCullagh and Nelder, 1983), so that it can be equivalently expressed in matrix form as $L = X\beta$, where $L = (l_1, \ldots, l_n)'$ is the vector of logit transformations previously defined, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ the vector of parameters and $X = (\mathbf{1} \mid \mathscr{X})$ the design matrix, with $\mathbf{1} = (1, \ldots, 1)'$ being the n-dimensional vector of ones.

Before estimating the logistic model, let us remember that the relationship between the response variable and each predictor can be interpreted in terms of odds ratios from the parameters of the model. From expression (2) we have that the exponential of the jth parameter ($j = 1, \ldots, p$) is the odds ratio of success ($\mathscr{Y} = 1$) when the jth predictor variable is increased by one unit and the other predictors are controlled (fixed as constant). That is,

$$\theta\left(\Delta\mathscr{X}_j = 1 \,|\mathscr{X}_k = x_k,\ \forall k \neq j\right) = \frac{\dfrac{\pi\left(x_1, \ldots, x_j + 1, \ldots, x_p\right)}{1 - \pi\left(x_1, \ldots, x_j + 1, \ldots, x_p\right)}}{\dfrac{\pi\left(x_1, \ldots, x_j, \ldots, x_p\right)}{1 - \pi\left(x_1, \ldots, x_j, \ldots, x_p\right)}} = \exp\left\{\beta_j\right\}$$

with $x_1, x_2, \ldots, x_p$ being a single observation of the explicative variables. Then, the exponential of the jth parameter of the logistic regression model gives the multiplicative change in the odds of success so that when its associated predictor increases, the probability of success increases if the parameter is positive and decreases in the opposite case. This odds ratio can help us to measure, for example, the relative change in the probability of recovery in a patient when we increase the dose of certain medicine. This parameter interpretation states the need for an accurate estimation of the parameters of the logistic model.

The most used method for estimating the logistic model is maximum likelihood as can be seen, for example, in Hosmer and Lemeshow (1989) and Ryan (1997). Let $L(Y; \beta)$ be

the likelihood given by

$$L(Y;\beta)=\prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i}. \tag{3}$$

Then, the likelihood equations are given in matrix form by $X'(Y-\Pi)=0$, with $\Pi=(\pi_1,\ldots,\pi_n)'$. These equations are not linear in the parameters so that they must be solved by using an approximating procedure as the Newton–Raphson method.

As it has been stated by several authors, and will be corroborated in the simulation study developed at the end of this paper, the maximum-likelihood estimation is not too accurate in the case of multicollinearity. As indicated in Ryan (1997), we must first select an indicator of multicollinearity in logistic regression. If the regressors are all continuous, then pairwise correlations and variance inflation factors might be used. The problem increases when some of the predictors are not continuous. One possible check of multicollinearity for qualitative variables would be to use kappa measure of agreement. The detection and diagnosis of collinearity in logistic regression in a similar way to linear regression are also discussed in Hosmer and Lemeshow (1989). They indicate that large standard errors could be a collinearity warning. However, in Ryan (1997), an example can be seen where harmful collinearity can exist without large standard errors. The impact and diagnosis of multicollinearity and ill-conditioned information in generalized linear models are also analyzed in Marx (1992), where the effects of severe ill-conditioning information on the maximum-likelihood estimation of a Poisson response model with the natural log link function are analyzed on real data.

Once the model has been estimated, its goodness of fit must be tested. The most usual method to solve the test

$$\mathrm{H}_0: l_i=\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j \quad (i=1,\ldots,n)$$

and

$$\mathrm{H}_1: l_i\neq\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j \quad (\text{some } i)$$

is based on the Wilks statistic (deviance) defined as $-2\ln\Lambda$, with Λ been the usual likelihood-ratio statistic. In the case of the logit model (1), the deviance is given by

$$G^2(M)=2\sum_{i=1}^{n}\left[y_i\ln\left(\frac{y_i}{\widehat{\pi}_i}\right)+(1-y_i)\ln\left(\frac{1-y_i}{1-\widehat{\pi}_i}\right)\right]\underset{n\to\infty}{\overset{\mathrm{H}_0}{\rightsquigarrow}}\chi^2_{n-p-1} \tag{4}$$

and can be equivalently expressed as $G^2(M)=2(\mathscr{L}_S-\mathscr{L}_M)$ where $\mathscr{L}_S$ and $\mathscr{L}_M$ are the maximized log-likelihood values for the saturated model (the most complex model which has a separate parameter for each logit) and the model of interest M with all predictor variables, respectively (see, for example, Hosmer and Lemeshow (1989) for a detailed study).

Other measure of the validity of the model, potentially more informative and meaningful than the p-value of a goodness-of-fit statistic, is the correct classification rate (CCR) defined as the percentage of subjects in the data set that are correctly classified. In order to classify an observation according to the estimated model it is usual to fix a cutpoint between 0 and 1, and to assign value 0 to a predicted probability lesser than this cutpoint and value 1 in an other case. Then a subject is correctly classified when the observed and predicted values agree (both zero or one). The most used cutpoint is 0.5 but, as we can see in Ryan (1997), we could use another cutpoint as, for example, the percentage of responses $Y = 1$ in the data set.

Another important task in logistic regression is the selection of the best covariates to explain the response. Several model selection procedures exist, no one of which is "the best". Usual cautions applied in ordinary regression hold for logistic regression. For example, a model with several predictors has the potential for multicollinearity so that a variable may seem to have little effect simply because it overlaps considerably with other regressors in the model. The most usual model selection methods are based on stepwise selection of regressors (forward or backward). Various statistics have been suggested in literature for assessing in each step the validity of one or several predictor variables. Among the best known, Wald test, scores test and conditional likelihood-ratio test, we present hereafter the last one that will be used for selecting the best principal component logistic regression model in the next section.

Let us denote by M_P the particular logit model obtained by setting equal zero certain number l of parameters, $\beta_{(1)}, \ldots, \beta_{(l)}$, selected among the $p + 1$ ones of model M. The likelihood statistic to compare model M_P to model M tests the hypothesis that all parameters in model M but not in model M_P equal zero. Then the conditional likelihood-ratio statistics for testing model M_P, given that M holds, is given by the difference in the G^2 goodness-of-fit statistics (4) for the two compared models

$$G^2\left(M_P \,/\, M\right) = 2\left(\mathscr{L}_M - \mathscr{L}_{M_P}\right) = G^2\left(M_P\right) - G^2(M),$$

with $\mathscr{L}_{M_P}$ being the maximized log-likelihood for the simpler model M_P that deletes those l parameters. It is a large-sample chi-squared statistic, with *df* equal to the difference between the residual *df* values for the two compared models (number l of parameters equal to zero in the fitted model M).

3. Solving the multicollinearity problem

In order to reduce the dimension of a linear model and to improve the estimation of its parameters in the case of multicollinearity, different techniques have been developed, as PCR and partial least-squares linear regression. In this paper, we propose to generalize PCR by using a reduced set of pc's of the predictor variables as covariates of the logistic regression model. The performance of the proposed PCLR models will be checked on a simulation study developed in the next section, where the estimated parameters provided by PCLR will be compared with those given by PLS-LR. Because of this, at the end of this section, we also present a brief summary on PLS-LR introduced by Bastien et al. (2005) in the general context of generalized linear regression.

3.1. PCLR model

First, let us briefly present sample PCA of a set of variables and its main properties (more details in Basilevsky (1994) or Jackson (1991)). Second, we will formulate the PCLR model taking as covariates a set of pc's of the predictors. Finally, different methods to select the significant pc's explanatory variables in the PCLR model will be considered and discussed in this paper.

3.1.1. Principal component analysis

Multivariate PCA, introduced by Karl Pearson at the beginning of the 20th century and developed by Harold Hötteling in 1933, is a multivariate technique based on explaining a set of correlated variables by a reduced number of uncorrelated ones with maximum variance, called pc's.

We are going to define PCA from the sample point of view, that is, PCA of a sample of observations of a set of variables. Then, let us consider a set of p continuous variables and n observations of such variables that we will resume in the matrix $\mathcal{X} = \left(x_{ij}\right)_{n\times p}$. The column vectors of such a matrix will be denoted by $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_p$.

Let us denoted by $S = \left(s_{jk}\right)_{p\times p}$ the sample covariance matrix whose elements are defined by $s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)(x_{ik} - \overline{x}_k)$, where the sample means are given by $\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ $(j = 1, \ldots, p)$. In order to simplify we will consider, without loss of generality, that the observations are centered, so that $\overline{x}_1 = \cdots = \overline{x}_p = 0$, and the sample covariance matrix is $S = \frac{1}{n-1}\mathcal{X}'\mathcal{X}$.

The sample pc's are defined as orthogonal linear spans with maximum variance of the columns of the matrix $\mathcal{X}$, denoted by $\mathcal{Z}_j = \mathcal{X}\mathcal{V}_j (j=1, \ldots, p)$. Then, the coefficient vectors that define the pc's, $\mathcal{V}_1, \ldots, \mathcal{V}_p$, are the eigenvectors of the sample covariance matrix S associated to their corresponding eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_p \geqslant 0$ that are the variances of the corresponding pc's. If we denote by $\mathcal{Z}$ the matrix whose columns are the sample pc's previously defined, it can be expressed as $\mathcal{Z} = \mathcal{X}\mathcal{V}$, with $\mathcal{V} = \left(v_{jk}\right)_{p\times p}$ being the matrix that has as columns the eigenvectors of the sample covariance matrix.

Let us remember that the sample covariance matrix is decomposed as $S = \mathcal{V}\Delta\mathcal{V}'$, with $\mathcal{V}$ being orthogonal and $\Delta = diag\left(\lambda_1, \ldots, \lambda_p\right)$, so that the matrix of observations is given by $\mathcal{X} = \mathcal{Z}\mathcal{V}'$.

This pc decomposition led us to obtain an approximated reconstruction of each original observation in terms of a reduced number of pc's

$$\mathcal{X}_j = \sum_{k=1}^{s} \mathcal{Z}_k v_{jk}, \quad j = 1, \ldots, p,$$

that accounts a high percentage of the total variability given by

$$\left[\frac{\sum_{j=1}^{s}\lambda_j}{\sum_{j=1}^{p}\lambda_j} \times 100\right], \quad s \leqslant p.$$

3.1.2. Model formulation

As a previous step to define the PCLR model we are going to formulate the logit model in terms of all the pc's associated to the matrix $\mathscr{X}$ of observations of the continuous predictor variables. We will assume without loss of generality that the regressors are centered. The probabilities of success of the logit model given by (2) can be equivalently expressed in terms of all pc's as

$$\pi_i = \frac{\exp\left\{\beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{p} z_{ik} v_{jk}\beta_j\right\}}{1+\exp\left\{\beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{p} z_{ik} v_{jk}\beta_j\right\}} = \frac{\exp\left\{\beta_0 + \sum_{k=1}^{p} z_{ik}\gamma_k\right\}}{1+\exp\left\{\beta_0 + \sum_{k=1}^{p} z_{ik}\gamma_k\right\}}$$

with z_{ik}, $(i=1,\ldots,n; k=1,\ldots,p)$ being the elements of the pc's matrix $\mathscr{Z}=\mathscr{X}\mathscr{V}$ and $\gamma_k = \sum_{j=1}^{p} v_{jk}\beta_j$, $k=1,\ldots,p$. The logistic model can be equivalently expressed in matrix form in terms of the logit transformations and the pc's as

$$L = X\beta = ZV'\beta = Z\gamma, \tag{5}$$

where

$$Z = (\mathbf{1} \mid \mathscr{Z}), \quad \left(\begin{array}{c|c} 1 & \mathbf{0}' \\ \hline \mathbf{0} & \mathcal{V} \end{array}\right), \quad \mathbf{0} = (0,\ldots,0)', \quad \mathbf{1} = (1,\ldots,1)'.$$

Therefore, the parameters of the logit model can be obtained as follows in terms of those of the model that has as covariates all the pc's: $\beta = V\gamma$. As a consequence of the invariance property of maximum-likelihood estimates we have $\widehat{\beta} = V\widehat{\gamma}$, and the prediction equation $\widehat{Y} = \widehat{\pi}$.

In order to improve the estimation of the original parameters in the case of collinearity, we will next introduce the PCLR model that is obtained by taking as covariates of the logit model a reduced set of pc's of the original predictors.

Let us split matrices Z and V in boxes as

$$Z = \left(\begin{array}{cccc|ccc} 1 & z_{11} & \cdots & z_{1s} & z_{1s+1} & \cdots & z_{1p} \\ 1 & z_{21} & \cdots & z_{2s} & z_{2s+1} & \cdots & z_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & z_{n1} & \cdots & z_{ns} & z_{ns+1} & \cdots & z_{np} \end{array}\right) = \left(Z_{(s)} \,\middle|\, Z_{(r)}\right), \quad (r = p-s)$$

and

$$V = \left(\begin{array}{cccc|ccc} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & v_{11} & \cdots & v_{1s} & v_{1s+1} & \cdots & v_{1p} \\ 0 & v_{21} & \cdots & v_{2s} & v_{2s+1} & \cdots & v_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & v_{p1} & \cdots & v_{ps} & v_{ps+1} & \cdots & v_{pp} \end{array}\right) = \left(V_{(s)} \,\middle|\, V_{(r)}\right).$$

Then, $Z_{(s)} = XV_{(s)}$ and $Z_{(r)} = XV_{(r)}$, so that the original parameters can be expressed as

$$\beta = V\gamma = V_{(s)}\gamma_{(s)} + V_{(r)}\gamma_{(r)},$$

where $\gamma = \left(\gamma_0\ \gamma_1\ \cdots\ \gamma_s \,|\, \gamma_{s+1}\ \cdots\ \gamma_p\right)' = \left(\gamma'_{(s)} \,\middle|\, \gamma'_{(r)}\right)'$.

Taking into account that the logit model in terms of all the pc's given by Eq. (5), can be decomposed as $L = Z\gamma = Z_{(s)}\gamma_{(s)} + Z_{(r)}\gamma_{(r)}$, the PCLR model in terms of s pc's (PCLR(s)) is obtained by removing in the last equation the r last pc's, so that we have

$$y_i = \pi_{i(s)} + \varepsilon_{i(s)},$$

where

$$\pi_{i(s)} = \frac{\exp\left\{\gamma_0 + \sum_{j=1}^{s} z_{ij}\gamma_j\right\}}{1 + \exp\left\{\gamma_0 + \sum_{j=1}^{s} z_{ij}\gamma_j\right\}}, \quad i = 1, \ldots, n.$$

This model can be equivalently formulated in matrix form in terms of the vector of logit transformations $L_{(s)} = \left(l_{1(s)}, \ldots, l_{n(s)}\right)$ with components $l_{i(s)} = \ln\left(\pi_{i(s)} \big/ \left(1 - \pi_{i(s)}\right)\right)$ as follows:

$$L_{(s)} = Z_{(s)}\gamma_{(s)} = XV_{(s)}\gamma_{(s)} = X\beta_{(s)}.$$

Therefore, we have obtained a reconstruction of the original parameters given by $\beta_{(s)} = V_{(s)}\gamma_{(s)}$, in terms of the parameters of the *PCLR* model that has as covariates the first s pc's. The maximum-likelihood estimation of this PCLR model will provide an estimation of the original parameters β given by

$$\widehat{\beta}_{(s)} = V_{(s)}\widehat{\gamma}_{(s)}, \tag{6}$$

that will improve the estimation $\widehat{\beta}$ obtained with the original variables in the case of multicollinearity.

The main difference between PCR and PCLR is that in the last one the estimator $\widehat{\gamma}_{(s)}$ in terms of the first s pc's is not the vector of the first s components of the estimator $\widehat{\gamma}$ in terms of all the pc's. That is, $\widehat{\gamma}_{(s)} = \left(\widehat{\gamma}_{0(s)}, \widehat{\gamma}_{1(s)}, \ldots, \widehat{\gamma}_{s(s)}\right)' \neq \left(\widehat{\gamma}_0, \widehat{\gamma}_1, \ldots, \widehat{\gamma}_s\right)'$. As a consequence the probabilities $\widehat{\pi}_{i(s)}$ estimated by the PCLR(s) model are different to the ones obtained by truncating the maximum-likelihood estimated probabilities of the model that has as regressors all the pc's. That is,

$$\widehat{\pi}_{i(s)} = \frac{\exp\left\{\widehat{\gamma}_{0(s)} + \sum_{j=1}^{s} z_{ij}\widehat{\gamma}_{j(s)}\right\}}{1 + \exp\left\{\widehat{\gamma}_{0(s)} + \sum_{j=1}^{s} z_{ij}\widehat{\gamma}_{j(s)}\right\}} \neq \frac{\exp\left\{\widehat{\gamma}_0 + \sum_{j=1}^{s} z_{ij}\widehat{\gamma}_j\right\}}{1 + \exp\left\{\widehat{\gamma}_0 + \sum_{j=1}^{s} z_{ij}\widehat{\gamma}_j\right\}}.$$

This means a considerable increment in computational effort because PCLR model has to be readjusted each time we enter or remove a new principal component in the model.

3.1.3. Model selection

Let us observe that we have used the first s pc's (the most explicative ones) to formulate the *PCLR* model. However, in PCR it is known that pc's with the largest variances are not necessarily the best predictors because minor pc's with small variances could be highly correlated with the response variable so that they must be considered as explicative variables in the optimum model. This means that pc's might be included in the model according to their predictive ability. Different methods for including pc's in the linear regression model

have been developed for example, in Hocking (1976), Mansfield et al. (1977) and Gunst and Mason (1977).

In order to obtain an optimum reconstruction of the original parameters with a small number of pc's, we have considered in this paper two different methods for including pc's in the PCLR model. The first (Method I) does not take into account the response variable and consists of including pc's in the natural order given by explained variability. In the second (Method II) pc's are selected based on their relationship to the response by using a forward stepwise method based on the conditional likelihood-ratio tests seen in Section 2.

Many authors have criticized PCR because pc's are obtained without taking into account the dependence between response and predictor variables. Let us observe that the optimum PCLR model provided by Method II solves this problem by including pc's in the model according to their ability to explain the response. In this sense PCLR with Method II can be seen as an alternative to PLS-LR (see next subsection), where linear combinations of the original variables, that are obtained by taking into account the relationship between covariates and response, are used as regressors in the model.

In order to select a PCLR model by using Method II, we use a forward stepwise procedure, starting with the simplest model without pc's and successively adding pc's sequentially to the model until further additions do not improve the fit. At each step, we enter the pc giving the greatest improvement. If we use conditional likelihood-ratio tests to select one between two nested models, the pc added at each step has the minimum significant p-value, when we test that its associated parameter equals zero in the model that results by entering this pc in the one selected in previous step. In our case, the first step of this procedure consists of carrying out p conditional likelihood-ratio tests, each one of which tests the simple model without variables given that the model obtained by introducing each of the pc's holds. Then, the pc with the minimum p-value lesser than the fixed significance level is entered in the model. In the jth step we carry out $(p-(j-1))$ likelihood-ratio tests for testing the model with the $(j-1)$ pc's selected until the previous step given each of the models obtained by entering in that model each of the remaining pc's. The procedure finishes when all tests of a step provide p-values larger than the fixed significance level and the model selected in the previous step fits well. Let us observe that pc's are added to the models one by one so that at each step the deviance statistic (4) will only have one degree of freedom.

Another important question to solve is how to select the optimum number of pc's that we have to retain in the model. We can find in literature different criteria as, for example, the one considered in Aucott et al. (1984) for the linear case that is based on the variability of the estimators. In order to obtain the best possible estimation of the parameters of a logit model, we will propose different criteria for selecting the optimum PCLR model based on different accuracy measures of the estimated parameters.

First, we define the *mean squared error of the beta parameter vector* (*MSEB*)

$$MSEB_{(s)} = \frac{1}{p+1}\sum_{j=0}^{p}\left(\widehat{\beta}_{j(s)} - \beta_j\right)^2, \quad s = 1, \ldots, p.$$

Second, we define the *maximum of the absolute differences of the beta parameters* as

$$Max_{(s)} = Max_j\left\{\left|\beta_{j(s)} - \beta_j\right|\right\}, \quad s = 1, \ldots, p.$$

On the other hand, we can expect that the best estimation of the original parameters provides the best estimation of the probabilities of the model. Therefore, we have also defined the *mean-squared error of probabilities* as

$$MSEP_{(s)} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\pi}_{i(s)} - \pi_i\right)^2, \quad s = 1, \ldots, p.$$

Let us observe that small values of *MSEB*, *Max*, and *MSEP* will indicate better estimation of the parameters. In the simulation study developed at the end of the paper we will select as optimum model for each method of entering pc's the one with the smallest *MSEB*.

The comparison between estimated and real parameters is not possible when we are analyzing real data, so that *MSEB*, *Max*, and *MSEP* cannot be computed and we need to define another measure of the accuracy of the estimations that does not take into account the unknown real parameters. Several authors, as Aucott et al. (1984), among others, have noted that in the linear case the variance of estimated parameters is very sensitive to a bad estimation. Therefore, we have considered the estimated variance of the estimated parameters of the logistic model defined by

$$Var_{(s)} = \widehat{Var}\left[\widehat{\beta}_{(s)}\right] = V'_{(s)}\left(Z'_{(s)}\widehat{W}_{(s)}Z_{(s)}\right)^{-1}V_{(s)},$$

where $\widehat{W}_{(s)} = diag\left(\widehat{\pi}_{i(s)}\left(1-\widehat{\pi}_{i(s)}\right)\right)$. In the simulated examples developed at the end of this paper we will observe that generally the best estimations of the parameter vector (smallest *MSEB*) are followed by a great increase on its estimated variance. Then, in practice with real data, we propose to select as optimum model the PCLR model previous to a significant increment in the estimated variance of the beta parameters.

3.2. PLS-LR model

PLS regression is a technique much used by chemometricians which solves the problem of ill-conditioned design matrices in regression methods. This problem appears when the number of explanatory variables is bigger than the number of sample observations or there is a high-dependence framework among predictors. PLS regression was introduced by Wold (1973) and later developed by many authors in recent years (see, for example, Dayal and MacGregor, 1997). In order to overcome these problems, PLS regression defines latent incorrelated variables (PLS components) given by linear spans of the original predictors, and uses them as covariates of the regression model. These linear spans take into account the relationship between the original explanatory variables and the response, and are usually obtained by NIPALS algorithm (Wold, 1984).

The algorithm for computing a PLS linear regression model consists of the following steps:

(1) Computation of a set of PLS components.
(2) Linear regression of the response variable on the retained PLS components.
(3) Formulation of the PLS regression model in terms of the original predictor variables.

In order to describe in detail the first step (NIPALS algorithm), let us consider without loss of generality a centered response variable Y, and p centered predictor

variables, $\mathscr{X}_1, \mathscr{X}_2, \ldots, \mathscr{X}_p$. Then, the algorithm for computing a set of PLS components follows the next steps:

(1) *Computation of the first PLS component*: In order to obtain the first PLS component T_1, we will fit the p linear models *linfit* $\left(Y \,/\, \mathscr{X}_j\right)$ $(j = 1, \ldots, p)$, formulated with Y as response variable and each $\mathscr{X}_j$ as predictor. Then, the slope estimated parameters of these linear models, δ_{1j} $(j=1, \ldots, p)$, are normalized in the usual way, providing the coefficients a_{1j} $\left(\sum_{j=1}^{p} a_{1j}^2 = 1\right)$ of the linear span of the original variables that defines the first PLS component, $T_1 = \sum_{j=1}^{p} a_{1j}\mathscr{X}_j$.

(2) *Computation of the kth PLS component*: In order to obtain the kth PLS component T_k, given a set of $k-1$ PLS components, $T_1, \ldots, T_{k-1}$, yielded in previous steps, we will fit the p linear models *linfit* $\left(Y \,/\, \left(\mathscr{X}_j, T_1, \ldots, T_{k-1}\right)\right)$ $(j = 1, \ldots, p)$, formulated with Y as response variable and $\mathscr{X}_j, T_1, \ldots, T_{k-1}$ as predictor variables. Then, the normalized slope estimated parameters corresponding to each $\mathscr{X}_j$, δ_{kj} $(j = 1, \ldots, p)$, are considered as the coefficients of the linear span that defines the kth PLS component. The variables of this linear span will be the residuals of the linear models *linfit* $\left(\mathscr{X}_j \,/\, (T_1, \ldots, T_{k-1})\right)$ $(j = 1, \ldots, p)$, formulated with $\mathscr{X}_j$ as response variable and $T_1, \ldots, T_{k-1}$ as explanatory variables.

Let us observe that the computation of each PLS component T_k is simplified by setting to zero in its linear span those coefficients a_{kj} that are not significant. This means that only significant variables will contribute to the computation of PLS components. The algorithm stops when computing a PLS component none of its coefficients is significantly different from zero. The statistical significance of the parameters δ_{kj} associated to the variable $\mathscr{X}_j$ in each linear fit *linfit* $\left(Y \,/\, \left(\mathscr{X}_j, T_1, \ldots, T_{k-1}\right)\right)$ will be tested by using the classical statistical tests associated with linear regression.

The philosophy of PLS is based on two important questions. First, the estimated parameters of the linear fits represent the correlation between the response and the corresponding covariate. Second, the residuals of the models *linfit* $\left(\mathscr{X}_j \,/\, (T_1, \ldots, T_{k-1})\right)$ are orthogonal to each original variable $\mathscr{X}_j$, so that the latent variable defined by them is orthogonal to the PLS components previously computed.

PLS regression has been recently adapted to generalized linear models (PLS-GLR) and the particular case of logistic regression (Bastien et al., 2005). This is an *ad hoc* adaptation where each one of the linear models *linfit* $\left(Y \,/\, \left(\mathscr{X}_j, T_1, \ldots, T_{k-1}\right)\right)$ is changed by the corresponding logit model meanwhile the linear fits *linfit* $\left(\mathscr{X}_j \,/\, (T_1, \ldots, T_{k-1})\right)$ are kept. Although the parameters of a logit model are not a measure of correlation between the response and each of the predictors, in PLS-LR it is assumed that they represent in some sense the dependence framework between each predictor and the response variable. The Wald statistic (see, for example, Hosmer and Lemeshow, 1989) is generally used to test the statistical significance of the parameters δ_{kj} associated to the variable $\mathscr{X}_j$ in each logit fit of Y on the explanatory variables $\mathscr{X}_j, T_1, \ldots, T_{k-1}$.

4. Simulation study

In order to illustrate the performance of the proposed PCLR model, we developed a simulation study to show how the estimation of the parameters of a logit model with collinear regressors can be improved by using pc's.

In the simulation study we carried out a sensitivity analysis in terms of different distributions for simulating the regressors, different number of predictors and different number of sample sizes. In each simulation the binary response was simulated by following the scheme of the simulation studies developed in Hosmer et al. (1997) and Pulkstenis and Robinson (2002).

4.1. Simulation scheme

The first step in the simulation process was to obtain a sample of observations of p explicative variables with a known dependence framework. In order to have high correlation among the explicative variables, we considered as design matrix of the simulated logistic model $X = (\mathbf{1} \mid NA)$, where A is a fixed $p \times p$ matrix and N is a $n \times p$ matrix of n simulated values of p independent variables.

The second step was to fix a vector of real parameters $\beta = \left(\beta_0, \beta_1, \ldots, \beta_p\right)'$ and to compute the real probabilities by the model

$$\pi_i = \frac{\exp\left\{x_i'\beta\right\}}{1 + \exp\left\{x_i'\beta\right\}}, \quad i = 1, \ldots, n,$$

with x_i' being the ith row of the design matrix. Finally, each sample value of the binary response was simulated from a Bernouilli distribution with parameter π_i, $y_i \rightarrow B(\pi_i)$ $(i = 1, \ldots, n)$.

After each data simulation, we fitted the logistic model. As we will see the estimated parameters $\widehat{\beta} = \left(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p\right)'$ were always very different to the real ones due to multicollinearity. Despite this the logit model fitted well providing a high CCR and a goodness-of-fit statistic G^2 with high p-value. As we stated in previous sections, PCA of the regressors helps to improve this inaccurate estimation of the parameters. Once the pc's of the simulated covariates were computed, we fitted the PCLR(s) models with different sets of s pc's included by using Method I and II. Then, we computed for all the fitted PCLR(s) models the estimated parameters $\hat{\beta}_{(s)}$ in terms of the original variables, and the accuracy measures defined in previous sections for testing the improvement in the parameter estimation.

In order to compare PCLR and other techniques with similar objectives such as PLS-LR, we also obtained the PLS components associated to the simulated data, fitted the PLS-LR model and estimated the β parameters according to the computational algorithm presented in Section 3.2. Finally, we computed the same accuracy measures that in the case of PCLR(s) models.

This simulation and analysis procedure was carried out for three different number of regressors ($p = 6$, 10 and 15), three different sample sizes ($n = 100$, 200 and 300) and two different distributions of the collinear regressors (Cases I and II).

In order to validate the results obtained from each simulation we used sample replication, so that for each different case and fixed values of p and n, we repeated each simulation a great number of times by using the same beta parameters, the same A matrix, and by simulating new N and Y matrices in each replication. Then, we selected as the optimum PCLR(s) model in each simulation the one with the smallest $MSEB_{(s)}$. Finally, in order to summarize we computed the mean and the standard deviation of the different accuracy

Table 1
Simulated and estimated parameters for a specific simulation with Case I, $n = 100$ and $p = 10$

Parameters	Real	All pc's	PCLR(7)	PCLR(3)	PLS-LR
β_0	−0.67	−0.81	−0.40	−0.69	−0.31
β_1	−0.95	−14.84	−0.40	−0.96	−1.86
β_2	−0.95	4.96	−0.31	−1.24	−0.59
β_3	−0.97	4.35	0.49	−1.96	−1.02
β_4	1.40	−0.89	0.33	1.94	0.11
β_5	1.12	8.66	0.62	1.80	1.41
β_6	0.61	−2.70	0.41	0.98	0.15
β_7	−0.24	0.79	0.05	−0.60	0.12
β_8	−0.71	7.61	−0.83	0.85	−1.18
β_9	1.21	4.66	0.09	1.13	3.53
β_{10}	0.93	−5.48	0.34	0.29	0.15
MSEB		40.70	0.52	0.49	0.84

From left to right: simulated parameters (real), estimation without using pc's (all pc's), estimation with the optimum PCLR model provided by Method I (model PCLR(7) with the first seven pc's as covariates), estimation with the optimum PCLR model provided by Method II (model PCLR(3) with the first, second and ninth pc's as covariates), and estimation given by PLS-LR. The last row shows the MSEB associated to each one of the estimated models.

measures associated to the optimum PCLR(s) models and the PLS-LR models. The results are shown by using error bar plots which consist of displaying by a vertical bar centered on its mean, the variability of each specific measure ($\mu \pm 1.96\sigma$). All calculations were performed by using the SPLUS-2000 package.

4.2. Simulation results

In the first case (Case I), we simulated a sample of n values for each one of p independent variables (N matrix) with standard normal distribution. The fixed matrix A had as elements $p \times p$ values simulated from an uniform distribution in the interval [0, 1]. Then, we computed the correlation matrix of the simulated explicative variables (columns of matrix NA). All the correlations were always very high, so that there was high degree of multicollinearity among regressors. Then, a vector β of parameters of the model was fixed for each p.

In the second case (Case II) we considered a different distribution for simulating the predictors. We simulated n values of p independent variables with chi-square distribution with one degree of freedom (N matrix). The matrix A was then an upper triangular matrix with ones as upper triangular entries. A vector of fixed beta parameters was considered again for each different number of predictors p.

In order to better understand the performance of PCLR model we show the results for a specific situation of Case I with $n = 100$ and $p = 10$. In this situation correlations among predictors were all bigger than 0.5 and most of them bigger than 0.8, so that there was a high degree of correlation among regressors. The parameters considered for simulating the response appear in Table 1 next to those estimated after fitting the corresponding logit model. We can observe that the estimated parameters $\widehat{\beta}$ were mostly very different to the real ones. As we have previously stated, such erroneous estimation must be caused by multicollinearity.

Table 2
Goodness of fit and accuracy measures of PCLR(s) models for a particular simulation with Case I, $n=100$, $p=10$, and Method I of inclusion of pc's

s	PCLR(s)						
	$MSEP_{(s)}$	$CCR_{(s)}$	$MSEB_{(s)}$	$Max_{(s)}$	$Var_{(s)}$	$G^2_{(s)}$	p-value
1	0.076	66	0.776	1.34	0.060	112.35	0.153
2	0.025	78	0.628	1.37	0.149	84.77	0.808
3	0.017	79	0.594	1.41	0.214	82.05	0.844
4	0.012	80	0.567	1.42	0.339	79.84	0.868
5	0.013	78	0.567	1.28	0.490	79.42	0.859
6	0.012	78	0.563	1.26	0.687	79.40	0.842
7	0.010	80	0.519	1.47	1.003	78.51	0.841
8	0.010	82	0.541	1.32	2.098	74.28	0.899
9	0.010	86	0.603	1.56	4.672	61.43	0.991
10	0.015	87	40.70	13.89	158.949	58.44	0.995

Table 3
Goodness of fit and accuracy measures of PCLR(s) models and PLS-LR for a particular simulation with Case I, $n = 100$, $p = 10$, and Method II of inclusion of pc's

s	pc	PCLR(s)						
		$MSEP_{(s)}$	$CCR_{(s)}$	$MSEB_{(s)}$	$Max_{(s)}$	$Var_{(s)}$	$G^2_{(s)}$	p-value
1	2	0.076	72	0.658	1.45	0.09	111.45	0.17
2	1	0.025	78	0.628	1.37	0.15	84.77	0.81
3	9	0.025	85	0.487	1.56	1.75	72.74	0.96
4	3	0.016	87	0.659	1.48	2.16	68.19	0.98
		PLS-LR						
		0.163	67	0.842	2.32	105.76	69.75	0.98

This poor estimation leads to an erroneous interpretation of the parameters in terms of odds ratios. In spite of this these logit models fitted well, providing high CCR and p-values of the associated goodness-of-fit statistic (see last row of Table 2).

In order to improve these estimations, we computed the pc's of the simulated covariates. The first three pc's explained more than 90% of the total variability (93.7%) and the first six explained almost a 99% (98.81%). Then, we fitted the $PCLR(s)$ models with different number of pc's included by Methods I and II, and computed the reconstructed parameters $\widehat{\beta}_{(s)}$ from Eq. (6).

It is difficult to find the best estimation looking at the single components of the vectors, then we calculated the different accuracy measures previously defined. Tables 2 and 3 provide by rows the measures obtained for each $PCLR(s)$ model by including pc's by Method I and II, respectively. Looking at the goodness of fit of the adjusted PCLR(s)

Table 4
Estimated gamma parameters in terms of pc's for the logit model with all pc's and the optimum models provided by Methods I and II, in a particular simulation with Case I, $n = 100$ and $p = 10$

Parameters	All pc's	PCLR(7)	PCLR(3)
$\widehat{\gamma}_0$	−1.43	−0.92	−0.98
$\widehat{\gamma}_1$	0.45	0.32	0.34
$\widehat{\gamma}_2$	1.70	1.09	0.21
$\widehat{\gamma}_3$	−0.65	−0.38	—
$\widehat{\gamma}_4$	0.57	0.45	—
$\widehat{\gamma}_5$	0.20	0.24	—
$\widehat{\gamma}_6$	−0.05	−0.11	—
$\widehat{\gamma}_7$	−0.65	−0.51	—
$\widehat{\gamma}_8$	−1.66	—	—
$\widehat{\gamma}_9$	4.69	—	3.89
$\widehat{\gamma}_{10}$	20.69	—	—

models, we can see that all of them fitted well with both methods (in most cases, p-value of G^2 statistic rounded 0.9, *CCR* was around 80% and *MSEP* was very small). As was expected the G^2 statistic increased in all cases when we included more pc's in the model.

Searching the best possible estimation of the original parameters, we can observe that with both methods *MSEB* decreased until that a specific number s of pc's was included in the model, and after that, it began to increase. Let us observe from Table 2 that *MSEB* reached its minimum value of 0.519 for the model PCLR(7) with the first seven pc's as covariates, when we used Method I (pc's entered in the model by variability order). By using Method II (Table 3), (pc's entered in the model by using stepwise forward selection), the minimum *MSEB* was 0.487 and corresponded to the model PCLR(3) that had as covariates the second, first and ninth pc's. Let us observe that the stepwise procedure based on conditional likelihood tests entered in first place the second pc, in second place the first one, in third place the ninth one and in fourth place the third one.

So, by selecting as optimum the model that minimized *MSEB*, we conclude that the best possible estimation was provided by the model that had as covariates the first seven pc's selected with Method I, and the one that had as covariates the first, second and ninth pc's included by using Method II. The estimated beta parameters provided by these two optimum PCLR models appear in the fourth and fifth columns of Table 1 for Methods I and II, respectively. Let us observe that the optimum PCLR model selected by Method II is better than the one selected by Method I because it provided a larger dimension reduction (only three pc's instead of seven) and its associated MSEB was also smaller. On the other hand, looking at the minimum $Max_{(s)}$ we obtained similar conclusions with a little difference of one pc in or out in the resulting *PCLR* models.

In order to make the differences between PCR and PCLR clear, in Table 4 we can see, as an example, the estimated gamma parameters of the model with all the pc's as covariates and the optimum PCLR models previously selected by Methods I and II. As was stated when the PCLR model was formulated, the estimated gamma parameters of each particular model are different from the same parameters of the model with all pc's opposite to the case of PCR.

After the PCLR model was tested, we fitted the PLS-LR model. Four PLS components were retained and their estimated β parameters also appear in Table 1 next to those estimated by using the optimum PCLR models. From the last row of Table 3, we can observe that the accuracy measures with respect to the original β parameters, provided by PLS-LR were bigger than the ones associated to the optimum PCLR models ($MSEB$=0.842, Max=2.317). Moreover, the CCR of this model was smaller than those provided by the optimum PCLR models. All these measures lead us to conclude that in this particular case PLS-LR performed worse than PCLR.

In relation to the estimated variance of the estimated parameters $\left(Var_{(s)}\right)$, we observed that it increased as we included pc's in the PCLR model with Methods I and II. We can see from Table 2 (Method I) that the variance increased smoothly until the ninth pc was included in the model and it had a great increase when we included the tenth pc in the model (went from 4.672 to 158.949). From Table 3 (Method II) it can be observed that there was no large increase in the estimated variances of the estimated parameters and the one associated to PLS-LR model was extremely big ($Var = 105.76$).

In a study with real data it is not known the real value of the parameters so that $MSEB$ cannot be computed. In these cases we propose to choose as the optimum PCLR model the one with a number of pc's previous to a significant increase in the estimated variance of the estimated parameters because it usually increases when parameters are badly estimated. In this particular simulation, the selected optimum model would be the one with the first nine pc's as covariates by using Method I, and the one with four pc's (the second, first, ninth and third ones) by using Method II, that provide in both cases values of $MSEB$, $MSEP$ and Max very similar to the optimum models selected by the criterion of minimizing the $MSEB$.

In order to validate these results we repeated the data simulation and the fits 500 times for Cases I and II with each different p (6, 10 and 15) and each different n (100, 200 and 300). That is, we carried out a replication of size 500 of each one of the 18 different combinations of p, n and $Case$. Then, the means and standard deviations of the accuracy measures of the optimum PCLR and PLS-LR models selected in each replication were computed and plotted by error bar plots. In most cases these means were quite representative with small variability. The error bar plots are shown in Fig. 1 for Case I (normal distribution) and in Fig. 2 for Case II (chi-square distribution). In these figures we have to note that some bars are out of the plot limits because their standard deviations were too big. The first of these bars ($0.684 \pm 1.96 \times 0.922$) appears in Fig. 1 for the $MSEP$ of PCLR model with Method I and (n, p)=(100, 10). The second bar ($0.905 \pm 1.96 \times 1.454$) also appears in Fig. 1 for the $MSEP$ of PCLR model with Method II and $(n, p) = (100, 10)$. The third bar ($4.76 \pm 1.96 \times 36.09$) appears in Fig. 2 for the $MSEB$ of PLS-LR model and $(n, p) = (300, 15)$.

The results of the sensitivity analysis from the two cases (normal and chi-square distributions), the three number of regressors p (6, 10 and 15) and the three sample sizes n (100, 200 and 300) (see Figs. 1 and 2) showed that we could not find significant differences between the distribution selected to simulate the regressors (normal and chi-square). The sample size only had some influence in the $MSEB$ and $MSEP$ making that these measures globally decreased and became stabilized around small values as the sample size n increased. The number of regressors only had influence on the number of components of the optimum models so that, as it was expected, the mean number of pc's in the PCLR models increased as the number of original covariates p increased. That was not the case in PLS-LR mod-

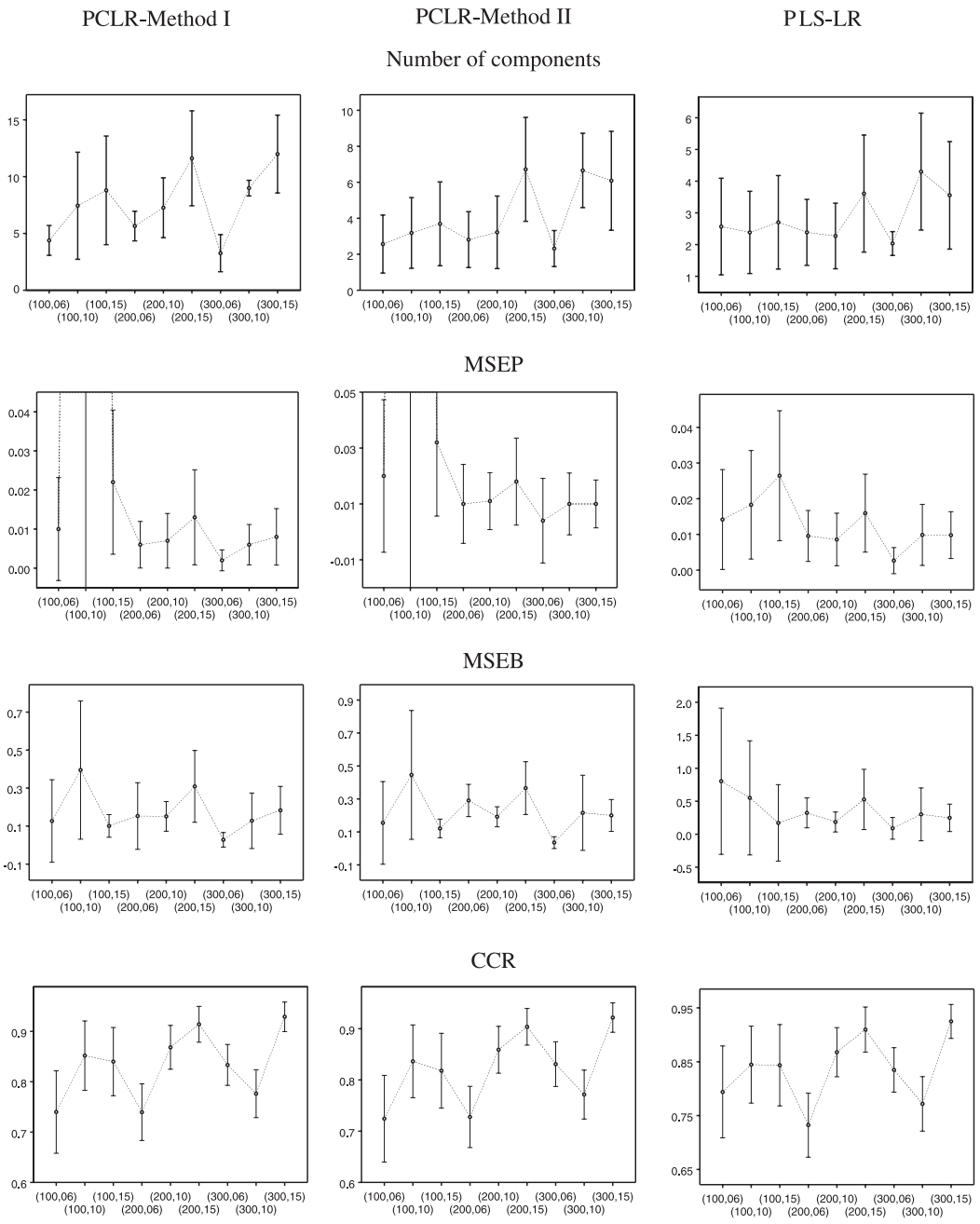

Fig. 1. Error bar plots of the goodness-of-fit measures of the optimum PCLR models provided by Methods I and II and PLS-LR, for 500 replications of the simulation with Case I (normal distribution) and different values of n and p represented in the plot by (n, p).

els. In fact we found several replications without PLS components. This may be caused because the model with all the original variables fitted well but it is possible that each specific variable did not have any influence in the response.

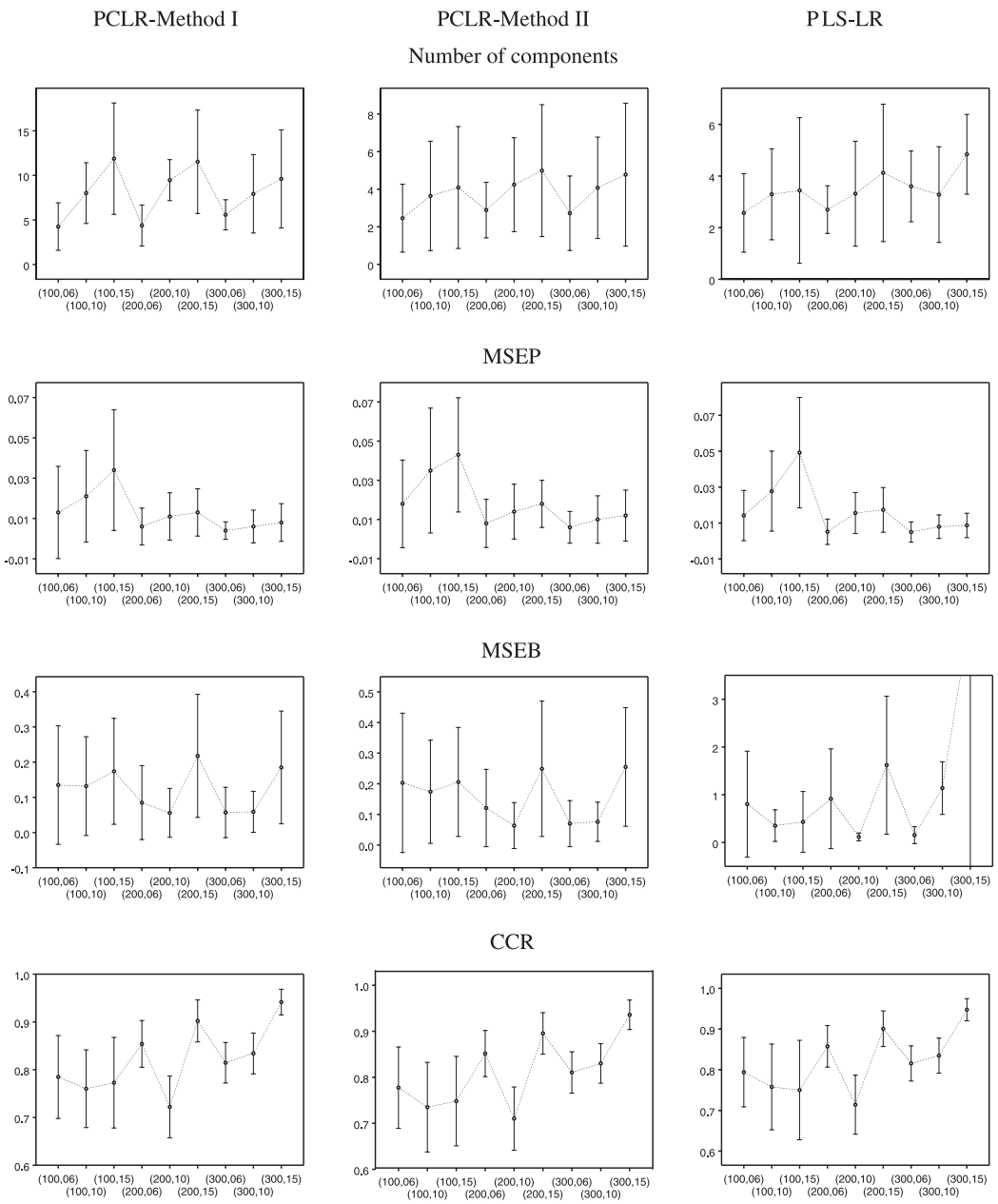

Fig. 2. Error bar plots of the goodness-of-fit measures of the optimum PCLR models provided by Methods I and II and PLS-LR, for 500 replications of the simulation with Case II (chi-square distribution) and different values of n and p represented in the plot by (n, p).

As concluded from the results of the specific situation (normal distribution, $n = 100$ and $p = 10$), this sensibility analysis corroborates that, in PCLR, Method II of inclusion of pc's provides a bigger dimension reduction than Method I with similar accuracy in the estimation of the β parameters.

Finally, with respect to the comparison between PCLR and PLS-LR, it can be observed from Figs. 1 and 2 that the mean number of PLS components is similar to the mean number of optimum pc's introduced in PCLR by using Method II. However, the mean *MSEBs* are smaller for PCLR and Method II with similar mean *CCRs* and *MSEPs*.

5. Conclusions

This paper is focused on solving the problem of high-dimensional multicollinear data in the logit model which explains a binary response variable from a set of continuous predictor variables. In order to solve this problem and to obtain an accurate estimation of the parameters in this case, a pc-based solution has been proposed.

In base to the simulation study developed in this work, where different sample sizes, number of predictors and distribution schemes have been considered, it can be concluded that the proposed PCLR models provide an accurate estimation of the parameters of a logit model in the case of multicollinearity, by using as covariates a reduced set of the pc's of the original variables.

In order to select the optimum PCLR model two different methods for including pc's in the model have been considered and compared. On the one hand, Method I includes pc's in the model according to their explained variances. On the other hand, Method II considers a stepwise procedure for selecting pc's based on conditional likelihood-ratio tests. Different accuracy measures with respect to the estimated parameters have been also introduced for selecting the optimum number of pc's. Finally, Method II, which takes into account the relationship among response and predictor pc's, has been chosen as the best because it provides better parameters estimation with smaller number of pc's (bigger reduction of dimension).

Finally, with respect to the comparison with PLS-LR, the PCLR model provides better estimation of the logit model parameters (less *MSEB*) with similar goodness-of-fit measures (*MSEP* and *CCR*) and needs less components so that the interpretation of the model parameters is more accurate.

Acknowledgements

This research has been supported by Project MTM2004-5992 from *Dirección General de Investigación, Ministerio de Ciencia y Tecnología*.

References

Aucott, L.S., Garthwaite, P.H., Currall, J., 1984. Regression methods for high dimensional multicollinear data. Comm. Statist.: Comput. Simulation 29 (4), 1021–1037.

Basilevsky, A., 1994. Statistical Factor Analysis and Related Methods: Theory and Applications. Wiley, New York.

Bastien, P., Esposito, V., Tenenhaus, M., 2005. PLS generalised linear regression. Comput. Statist. Data Anal. 48 (1), 17–46.

Dayal, B.S., MacGregor, J.F., 1997. Improved PLS algorithms. J. Chemometrics 11, 73–85.

Gunst, R.F., Mason, R.L., 1977. Biased estimation in regression: an evaluation using mean squared error. J. Amer. Statist. Assoc.: Theory Methods 359, 616–627.

Hocking, R.R., 1976. The analysis and selection of variables in linear regression. Biometrics 32, 1–49.

Hosmer, D.W., Lemeshow, S., 1989. Applied Logistic Regression. Wiley, New York.

Hosmer, D.W., Hosmer, T., Le Cessie, S., Lemeshow, S., 1997. A comparison of goodness-of-fit tests for the logistic regression model. Statist. Med. 16, 965–980.

Jackson, J.E., 1991. A User's Guide to Principal Components. Wiley, New York.

Mansfield, E.R., Webster, J.T., Gunst, R.F., 1977. An analytic selection technique for principal component regression. Appl. Statist. 26, 34–40.

Marx, B.D., 1992. A continuum of principal component generalized linear regression. Comput. Statist. Data Anal. 13, 385–393.

Marx, B.D., Smith, E.P., 1990. Principal component estimators for generalized linear regression. Biometrika 77 (1), 23–31.

Massy, W.F., 1965. Principal component regression in exploratory statistical research. J. Amer. Statist. Assoc. 60, 234–246.

McCullagh, P.Y., Nelder, J.A., 1983. Generalized Linear Models. Chapman & Hall, New York.

Prentice, R.L., Pyke, R., 1979. Logistic disease incidence models and case-control studies. Biometrika 66, 403–411.

Pulkstenis, E., Robinson, T.J., 2002. Two goodness-of-fit tests for logistic regression models with continuous covariates. Statist. Med. 21, 79–93.

Ryan, T.P., 1997. Modern Regression Methods. Wiley, New York.

Wold, H., 1984. PLS Regression. Encyclopaedia of Statistical Sciences, vol. 6. Academic Press, New York, pp. 5810–591.

Wold, S., 1973. Nonlinear iterative partial least squares (NIPALS) modelling: some current developments. In: Krishnaiah, P.R. (Ed.), Multivariate Analysis, vol. 3. Academic Press, New York, pp. 383–407.