VARIABLE CLUSTERING IN HIGH DIMENSIONAL PROBIT REGRESSION

Loïc Yengo [1,2,3], Julien Jacques [1,2] & Christophe Biernacki [1,2]

[1] *University Lille 1, CNRS UMR 8524* - [2] *Inria Lille, équipe MODAL*
[3] *CNRS UMR 8199* - *loic.yengo@good.ibl.fr*

Résumé. La réduction de la dimension est une des problématiques majeures de la régression en grande dimension. Nous avons récemment introduit, dans le cadre de régression linéaire, une nouvelle approche visant la réduction de la dimension par la classification des covariables en groupes de mêmes effets. Nous proposons ici une extension de ce dernier modèle à la régression Probit pour données binaires. Les qualités de prédiction du modèle proposé sont comparées à la régression logistique pénalisée en norme L^1 (LASSO) et L^2 (ridge). Cette comparaison effectuée à la fois sur des données réelles et simulées, révèle les bonnes qualités prédictives de notre approche.

Mots-clés. Regression Probit, Réduction de la dimension, Echantillonnage de Gibbs, Classification.

Abstract. Dimension reduction is a major issue in high-dimensional regression models. We recently introduced the CLusterwise Effect REgression (CLERE) methodology [1] in the context of linear regression for variable clustering as a way of reducing the dimensionality. We propose in this paper, an extension of the CLERE methodology to high dimensional Probit regression. The proposed extension was compared to LASSO and Ridge logistic regressions. This comparison achieved on both simulated and real data, revealed the good predictive performances of our method.

Keywords. Probit Regression, Dimension reduction, Gibbs sampling, Clustering.

1 Introduction

Binary response data are encountered numerous in scientific fields including econometrics (wealthy households versus poor households) or medical sciences (diseased versus healthy), as specific examples. The main regression-based approaches to handle these data are the Generalized Linear Models (GLM). In GLMs, the binary response c_i for an individual i ($i = 1, \ldots, n$) is assumed to take values into $\{0, 1\}$ and to follow a Bernoulli distribution of probability p_i. Probability p_i is defined as $p_i = F(\boldsymbol{x}_i\boldsymbol{\beta})$, where $\boldsymbol{x}_i$ is a vector of p covariates, $\boldsymbol{\beta}$ the vector of associated regression coefficients and F some link function mapping $\mathbb{R}$ to $]0, 1[$. The choice of F is very critical as it influences the interpretation and the tractability of the inference. Classical choices for F are the so-called *logit* link function, defined as

$$logit : x \mapsto F(x) = \sigma(x) = \frac{1}{1+e^{-x}},$$

and the the *probit* link function defined as

$$probit : x \mapsto F(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{t^2}{2}\right] dt.$$

Beyond choosing F, another difficulty arises when the number covariates in the model exceeds the sample size, i.e. $p > n$. This difficulty concerns both the hability to interpretate the model and to yield reliable predictions. We recently introduced the CLusterwise Effect REgression (CLERE) methodology [1] in the context of linear regression for variable clustering as a way of reducing the dimensionality. This methodolgy assumes that regression coefficients are independant random variables drawn from a mixture of Gaussian distributions. Recovering the latent mixture produces a clustering of the covariates.

This paper presents an extension of the CLERE methodology to handle binary response data in a high dimensional setting. This extension is considered under the GLM framework using the *probit* link function.

The present article is organized as follows. In Section 2 the extended model is presented as well as an algorithm for estimating the model parameters. Section 3 illustrates the predictive performances of the model in both simulated and real data. Finally Section 4 proposes possible improvements for our model.

2 Model definition and parameters estimation

2.1 Model definition

The standard probit regression model is defined by the following equations

$$\begin{cases} c_i = \mathbf{1}_{\{y_i > 0\}} \\ y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \\ \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1) \end{cases} \tag{1}$$

where for an individual i, c_i is an observed binary response, x_{ij} is the value of the j-th covariate and β_j is its associated regression coefficient. To extend the CLERE methodology to models such as model (1), we have to additionally assume that

$$\begin{cases} \beta_j | \boldsymbol{z}_j \overset{iid}{\sim} \mathcal{N}\left(\sum_{k=1}^{g} b_k z_{jk}, \gamma^2\right) \\ \boldsymbol{z}_j = \left(z_{j1}, \ldots, z_{jg}\right) \overset{iid}{\sim} \mathcal{M}\left(\pi_1, \ldots, \pi_g\right). \end{cases} \tag{2}$$

In model (2), z_{jk} (for $k = 1, \ldots, g$) is the cluster membership indicator that variable j belongs to group k; and $\mathcal{M}\left(\pi_1, \ldots, \pi_g\right)$ stands for the multinomial distribution.
We uses subsequently the following notations $\boldsymbol{c} = (c_1, \ldots, c_n)$, $\boldsymbol{y} = (y_1, \ldots, y_n)$, $\boldsymbol{y}^{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$, $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)$, $\boldsymbol{X}$ the $n \times p$ matrix which (i, j)-th term equals x_{ij}, $\boldsymbol{Z}$ the $p \times g$ matrix which (j, k)-th term equals z_{jk}, $\boldsymbol{Z}^{-j}$ is obtained from $\boldsymbol{Z}$ by withdrawing its j-th row and $\boldsymbol{\theta} = \left(\beta_0, b_1, \ldots, b_g, \pi_1, \ldots, \pi_g, \sigma^2, \gamma^2\right) \in \mathbb{R}^{(g+1)} \times [0; 1]^g \times \mathbb{R}_+^2$.

The model parameter $\boldsymbol{\theta}$ is estimated by maximizing the likelihood of the observed data $p(\boldsymbol{c}|\boldsymbol{X}; \boldsymbol{\theta})$. This likelihood is obtained by integrating the complete data likelihood $p(\boldsymbol{c}, \boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{\beta}|\boldsymbol{X}; \boldsymbol{\theta})$ over $(\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{\beta})$. Although an analytical integration can only be performed over $\boldsymbol{\beta}$, the maximum likelihood estimation is still achievable using the EM algorithm [2]. An exact EM algorithm is not feasible in this context since the *E step* is intractable. In turn, we propose to estimate the model parameters using the SEM algorithm. This algorithm replaces the *E step* with the simulation of the unobserved data $(\boldsymbol{y}, \boldsymbol{Z})$, using the conditional distribution of the latter given the observed data, $p(\boldsymbol{y}, \boldsymbol{Z}|\boldsymbol{c}, \boldsymbol{X}; \boldsymbol{\theta})$. Such simulation cannot be performed straightforwardly. We used therefore a Gibbs sampler to generate unobserved data. After the simulation step (*S step*) comes maximization step (*M step*) which consists in updating the model parameters with values that maximize the complete data likelihood (integrated over $\boldsymbol{\beta}$) given in Equation (3):

$$\begin{aligned}\log p\left(\mathbf{c}, \mathbf{y}, \mathbf{Z}|\mathbf{X}; \boldsymbol{\theta}\right) = \log p(\mathbf{c}|\mathbf{y}) &- \frac{n}{2}\log\left(2\pi\right) \\ &- \frac{1}{2}\left(\mathbf{y} - \beta_0\mathbf{1} - \mathbf{XZb}\right)'\left[\gamma^2\mathbf{XX}' + \mathrm{I}_p\right]^{-1}\mathbf{X}'\left(\mathbf{y} - \beta_0\mathbf{1} - \mathbf{XZb}\right) \\ &+ \sum_{j=1}^{p}\sum_{k=1}^{g} z_{jk}\log(\pi_k). \qquad (3)\end{aligned}$$

2.2 Simulation step by Gibbs sampling

The simulation step consists in sampling $(\boldsymbol{y}, \boldsymbol{Z})$ from $p(\boldsymbol{y}, \boldsymbol{Z}|\boldsymbol{c}, \boldsymbol{X}; \boldsymbol{\theta})$. As stated above, this simulation is performed using a Gibbs sampler. The necessary conditional distributions are briefly detailed in Sections 2.2.1 and 2.2.2.

2.2.1 Sampling from $p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{c}, \boldsymbol{X}; \boldsymbol{\theta})$

The distribution $(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{c}, \boldsymbol{X}; \boldsymbol{\theta})$ is a multivariate truncated Gaussian distribution. Sampling from a multivariate truncated Gaussian distribution is known to be difficult. However, using the conditional distributions $p(y_i|\boldsymbol{y}^{-i}, \boldsymbol{Z}, \boldsymbol{c}, \boldsymbol{X}; \boldsymbol{\theta})$, a Gibbs sampler can easily be applied. The latter distributions are univariate truncated Gaussian distributions.

2.2.2 Sampling from $p(\boldsymbol{Z}|\boldsymbol{y}, \boldsymbol{c}, \boldsymbol{X}; \boldsymbol{\theta})$

Sampling the whole matrix $\boldsymbol{Z}$ is challenging too. However, each of its rows follows a multinomial distribution. We can therefore also apply a Gibbs sampler to sample the rows using the conditional distributions $p(\boldsymbol{z}_j|\boldsymbol{Z}^{-j}, \boldsymbol{y}, \boldsymbol{c}, \boldsymbol{X}; \boldsymbol{\theta})$.

2.3 Maximization step

The *M step* consists in maximizing the function $\boldsymbol{\theta} \mapsto \log p(\boldsymbol{y}, \boldsymbol{Z}|\boldsymbol{c}, \boldsymbol{X}; \boldsymbol{\theta})$. This function corresponds to the marginal log-likelihood of a linear mixed model, which fixed effect parameters are $\beta_0, b_1, \ldots, b_g$ and random effect parameters are σ^2 and γ^2. We then used the EM algorithm proposed in Searle *et al.* (1992) [3] to perform that maximization.

3 Numerical experiments

This section presents numerical experiments on simulated and real data. These experiments aim at comparing our extended CLERE- probit model to classical models for high-dimensional binary regression models. The methods selected for comparison were the LASSO logistic regression and the ridge logistic regression.

3.1 Simulated data

The comparison is performed in terms of classification error on a simulated validation set of 500 individuals. The sample size in the training set was equal to 100, while the number p of covariates 200. All covariates were simulated independently according to the standard Gaussian distribution. The latent data y_i was simulated as follows:

$$y_i \overset{iid}{\sim} \mathcal{N}\left(\sum_{j=1}^{p} \beta_j x_{ij}, 1\right)$$

whith $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is defined as

$$\boldsymbol{\beta} = (\underbrace{-1, \ldots, -1}_{60}, \underbrace{0, \ldots, 0}_{80}, \underbrace{+1, \ldots, +1}_{60}).$$

The CLERE-probit model was fitted using 20 different random starting points. We used for the LASSO logistic regression and the ridge logistic regression the implementation proposed in the R package glmnet with their default parameters.

Figure 1 shows the classication errors associated with the three compared methods. CLERE-probit showed the lowest classication error compared to its competitors in addition to require a quite low number of parameters.

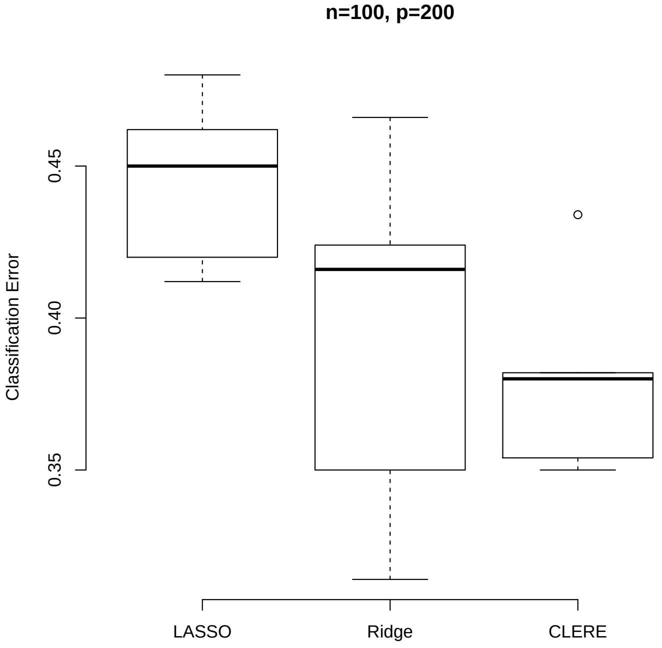

Figure 1: Classification errors associated with CLERE-probit, Logistic LASSO and Logistic Ridge. 50 replications were considered. The median number of parameters for logistic LASSO, logistic ridge and CLERE-probit was respectively, 14.2±1.3, 200±0.0 and 3±0.1.

3.2 Leukemia dataset

We used in this section the dataset leukemia from the R package spikeslab. This dataset involves gene expression measured in samples from human acute myeloid (coded as 0) and acute lymphoblastic leukemias (coded as 1). 3571 expression values were measured on 72 individuals. We primarily reduced the number of variables to 1412 by only including variables showing a nominal association with the response (p-value<0.05). The same three methods were compared using 3-fold cross-validation. Table 1, summarizes the results of that comparison.

Datasets	Methods	Averaged classication Error (Std. Error)	Number of parameters (Std. Error)
`leukemia`	LASSO	0.055 (0.03)	20.0 (1.52)
	RIDGE	0.013 (0.01)	1412 (0.00)
	CLERE-probit	0.013 (0.01)	3.00 (0.00)

Table 1: Out of sample classication error estimated via 3-fold cross-validation. The number of parameters reported for CLERE-probit was selected using AIC.

All methods were comparable in terms of classification error. However, it is noteworthy that logistic ridge and CLERE-probit showed the lowest classification errors (both yieled an error equal to 0.013), CLERE-probit being much more parsimonious.

4 Discussion

We presented in this paper an extension of the CLERE methodology for handling binary response data. This extension was shown to be more challenging than the linear regression case. The main challenge relates to the simulation of the auxiliary variable $\boldsymbol{y}$. This necessary step significantly increases the computational complexity of the inference.

Despite this computational limitation, the extension of CLERE for binary response data showed very competitive prediction performances compared to known approaches for dimension reduction, with a highly reduced number of parameters. The numerical experiments using both simulated and real datasets were really encouraging.

Finally, the extension presented here can be considred as a first step to generalize the CLERE methodology to multi-class response data and ordinal response data. The method presented in this paper is implented in `R` package `clere`.

Bibliographie

[1] Yengo L., Jacques J. and Biernacki C. (2013), Variable clustering in high dimensional linear regression models, Journal de la Société francaise de Statistiques, In press.
[2] Dempster A.P. Laird M.N and Rubin D.B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 39,1-22.
[3] S.R. Searle, G. Casella and C.E. McCullogh (1992), Variance components, Wiley series.