On the estimation of high-dimensional regression models with binary covariates

Sulla stima dei modelli di regressione ad alta dimensionalità con covariate binarie

Valentina Mameli[1], Debora Slanzi[1,2] and Irene Poli[1,3]

Abstract In this paper we address the problem of estimating the parameters of high dimensional regression models characterized by binary covariates. We suggest a new procedure which combines particular clustering for the binary covariates and group penalized regression for estimating the model parameters. The good performance of the methodology is shown in a simulation study.

Abstract *Questo lavoro affronta il tema della stima di modelli di regressione ad alta dimensionalità con covariate binarie. In particolare, si propone un procedura di stima per questa classe di modelli che combina tecniche di cluster analisi e modelli di regressione penalizzata di gruppo. La metodologia proposta viene valutata con uno studio di simulazione.*

Key words: Binary covariates, Clustering techniques, High-dimensional regression models.

1 Introduction

In many scientific fields of research, recent advances in technology have allowed to gather data sets characterized by a very high number of variables. The sample size of these data can be small compared to the number of variables and only a small number of these variables can be relevant to the study. Moreover, in several contexts binary variables are present to express the presence or absence of particular

[1] European Centre for Living Technology, Ca' Foscari University of Venice, Italy.
[2]Department of Management, Ca' Foscari University of Venice, Italy.
[3]Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Italy.
e-mail: valentina.mameli@unive.it, e-mail: debora.slanzi@unive.it, e-mail: irenpoli@unive.it

elements or features. For this structure of the problem we would like to provide a contribution in developing a procedure to select influential variables and estimate model parameters. Several different methodologies have been suggested in literature for variable selection in high-dimensional models, among these penalized regression models have gained popularity over the last few decades; see among others [4, 7, 1, 11, 6, 2]. Penalized procedures are designed with the aim of both selecting the most relevant variables and estimating the parameters of the models. In a general regression setting, models with continuous explanatory variables have been extensively studied, while models with binary explanatory variables have received much less attention.

The aim of this paper is to derive a new procedure to estimate high-dimensional models by combining the class of penalized regression models with binary variables clustering techniques. We propose to estimate penalized regressions based on the information obtained from the introduction of a grouping structure of covariates. More specifically, we propose a two-step procedure: in the first step, we group the covariates into non-overlapping clusters (or groups) using an approach able to deal with the binary nature of the covariates; in the second step, we select the most relevant clusters and covariates by using a penalized regression procedure in which the information obtained in the clustering phase is embedded.

The paper is organized as follows. In Section 2 we review some variable selection procedures, for individual, for group and for bi-level selection; we then present a new inferential procedure for high-dimensional regression models with binary covariates based on penalized regression models and clustering techniques. In Section 3 we conduct a simulation study to evaluate the performance of the new procedure.

2 Methodology

2.1 Model set-up

Le us consider the multiple linear regression model

$$y_i = X_i\beta + \varepsilon_i, \quad i = 1,\dots,n \tag{1}$$

where $X_i = (x_{i1},\dots,x_{ip})^T$ is a p-dimensional vector of explanatory variables (or covariates), y_i is the response value for the i-th observation, ε_i is the error term, n is the sample size and $\beta = (\beta_1,\dots,\beta_p)$ is the vector of parameters. The vector β is unknown and has to be inferred from the data. When the number of variables, p, is much larger than the sample size n the model is usually referred as a high-dimensional regression model. If only a small number of variables affects the response, the model results to be characterized by the *sparsity* condition [7]. To estimate the vector of regression coefficients β we consider penalized regression models and minimize the following function

$$Q(\beta) = \frac{1}{2n}(y - X\beta)^T(y - X\beta) + P(\beta|\lambda), \tag{2}$$

where $y = (y_1, \ldots, y_n)^T$ is the $n \times 1$ vector of response values and $X = (X_1, \ldots, X_n)^T$ is the $n \times p$ design matrix. The function $P(\cdot)$ is defined as a penalty on the regression coefficient parameters β and λ is a tuning parameter. The most used methods for choosing λ are cross-validation criteria or information criteria (see [9] among others). A number of penalized regression methods have been proposed in literature; see among others [6]. They include procedures for individual variable selection, group variable selection and bi-level variable selection. The Least Absolute Shrinkage Selection Operator (LASSO) proposed by [7] is one of the most famous procedure for individual variable selection. If the main interest is in selecting relevant groups of covariates and not individual ones, it is possible to take account of a grouping structure among the covariates as in group penalized regression procedures which include the group LASSO ([10]), the group Minimax Concave Penalty method [6] and the group Smoothly Clipped Absolute Deviation [6]. If, on the other hand, the focus is on selecting both the important groups of covariates as well as variables within these groups, bi-level selection procedures can be considered as the composite Minimax Concave Penalty ([1]), and the group exponential LASSO ([2]). These selection procedures have been introduced to overcome some limitations of the LASSO estimator and present a number of appealing properties in terms of both estimation accuracy as well as variable selection properties. Although a large amount of work has been done in the literature on the selection of continuous covariates in the high-dimensional framework, less attention has been given to the binary covariates case. To address the problem of estimating high-dimensional regression models with binary covariates we develop a methodology based on combining clustering techniques with penalized regression models. In this procedure a crucial step is the selection of binary variables groups to be embedded into the penalized regression model. The group selection can lead to a more effective variable selection and accurate predictions.

2.2 Estimating the parameters of the clustering structure regression

For a given clustering structure, the model (1) can be specified as

$$y = \sum_{k=1}^{K} \tilde{X}_k \tilde{\beta}_k + \varepsilon,$$

where $\tilde{X}_k$ is the $n \times d_k$ design matrix representing the d_k covariates belonging to the k-th cluster, $\tilde{\beta}_k = (\beta_{k1}, \ldots, \beta_{kd_k}) \in \mathbb{R}^{d_k}$ is the vector of regression coefficients of the k-th cluster and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is the error vector. Let x_{ij} be the value of X_j at the i-th observation. Assume that $x_{ij} = 1$ if X_j is present in the i-th observation

and $x_{ij} = 0$ otherwise, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Let c_j denote a latent cluster label for X_j, with $c_j = k$ if X_j is allocated to the k-th cluster, $k = 1, \ldots, K$. The scope of the clustering analysis in this context is to associate to each covariate X_j a unique label c_j with $j = 1, \ldots, p$.

Among the clustering techniques the most suitable for binary data include the agglomerative hierarchical clustering methods with binary dissimilarity matrices, the Bayesian non-parametric approach for binary data and the K-modes. Our purpose is to exploit the use of clustering techniques to identify non-overlapping groups of covariates. In order to estimate the grouping structure, in this paper we consider the following clustering methods:

- The standard agglomerative hierarchical clustering algorithms produce a nested sequence of clusters, which initially considers each observation as a single cluster, then at each stage the two least dissimilar clusters are combined. The process is repeated until only one cluster will contain all the observations. The dissimilarity between clusters can be measured by linkage methods: average, complete and single. Moreover, among the main distance measures between objects proposed for binary data we consider the Jaccard and the Tanimoto distances; see [3].
- The Bayesian non-parametric approach, recently proposed by [8], assumes that the data x_{ij} are independent draws from a mixture of infinite Bernoulli distributions whose parameters are distributed according to a Beta distribution. Clustering of data is obtained by calculating the posterior probability of the latent clusters labels c_j for $j = 1, \ldots, p$.
- The K-modes approach is a generalization of the K-means procedure suitable for categorical data; see [5]. This approach has two key differences with the classical K-means. First, it assumes that the representative point of the clusters (also known as centroid) is the modal value of a cluster. Second, the distance between objects is the Hamming distance. K-modes tries to find a partition of the objects into K groups by minimizing the distance between each observation and the group centroid.

We propose to estimate the model parameters by clustering the covariates according to a procedure above described and then introducing a group penalty in the estimating function (2). More specifically, the procedure involves the following two-steps:

1. cluster the covariates into non-overlapping groups by using a clustering technique suitable for binary-data;
2. regress the response variable y on the set of grouped covariates using a penalized regression procedure based on embedding the information gained from the preliminary clustering.

3 A simulation study

We conduct a simulation study to evaluate the performance of the clustering effects on the variable selection procedures. Among the several penalties that can be used

here we focus on the composite Maximum Concave Penalty (cMCP), the group exponential LASSO (gel), the group Maximum Concave Penalty (gMCP), the group LASSO (gLASSO) and the group Smoothly Clipped Absolute Deviation (gSCAD). For each of these group penalties, we consider different clustering methods in order to introduce a grouping structure into the model. For the purpose of this analysis, we consider the hierarchical clustering with three linkage methods: complete, single, average links, and three dissimilarity measures: Tanimoto and Jaccard distances. Moreover, we consider the K-modes and the Bayesian non-parametric methods. We compared group penalized regression models with the LASSO model. The simulation is based on the linear regression model $y_i = X_i\beta + \varepsilon_i$, $i = 1,\ldots,n$, where $\varepsilon_i \sim N(0,\sigma^2)$ as introduced in Equation 1. The standard deviation σ is assumed to be 1 and covariates were generated from a Bernoulli distribution. We consider the following setup: $p = 200$ covariates but just 4 of these covariates have non-zero coefficients. We randomly split the data into training and testing datasets. In this simulation the size of the data set is $n = 100$ and the size of the training set is 80. The number of clusters has been fixed to 10 for all the clustering methods considered. To evaluate the performance of the various group penalization methods combined with different clustering procedures, we calculate some measures of prediction accuracy and variable selection efficiency. In the simulation study, 1000 replicated data sets were generated from the model. For each of these datasets, we compute the Predictive Mean Square Error (PMSE), the Sensitivity (the ratio between the number of selected important variables and the number of important variables), and the Specificity (the ratio between the number of removed unimportant variables and the number of unimportant variables). The results of this simulation are presented in Tables 1 and 2. From Table 1 we can notice that all the penalized regression procedures considered yield satisfactory results in terms of PMSE for all the clustering techniques. In particular, the cMCP penalty provides almost the same good results regardless of the clustering algorithm considered. Moreover, we can notice the very good results achieved in terms of PMSE for the K-modes clustering and the Bayesian non parametric clustering for all penalties considered. Among group selection approaches, the gLASSO achieves the uppermost Sensitivity, especially when we consider the hierarchical approach with Tanimoto distance combined with average and single links. Both the bi-level selection penalties achieve low levels of Sensitivity, but high levels of Specificity, as we expected. In Table 2 we report the results for the LASSO model to allow a comparison with the results of group penalized regressions. We notice that LASSO shows good PMSE and Specificity values but a low value for Sensitivity. From this simulation study we can notice the good performance of this approach which embedded clusters of binary covariates in a group penalized regression model. Further simulation studies and analysis will be developed to evaluate conditions for better performances of this new approach.

Acknowledgements We would like to acknowledge the European Centre for Living Technology (`http://www.unive.it/pag/23664/`) for useful discussions of the research.

Table 1 Simulation results on the performance of the clustering effects on the variable selection procedures over 1000 replicates: Sensitivity, Specificity, PMSE (standard errors between brackets).

Clustering methods	Measures	Penalties: Bi-level cMCP	Bi-level gel	Group gLASSO	Group gMCP	Group gSCAD
Hierarchical average Jaccard	Sensitivity	0.210 (0.002)	0.140 (0.002)	0.275 (0.009)	0.064 (0.003)	0.255 (0.008)
	Specificity	0.978 (0.001)	0.997 (0.000)	0.650 (0.010)	0.892 (0.004)	0.671 (0.009)
	PMSE	1.659 (0.017)	1.855 (0.018)	2.046 (0.017)	2.058 (0.017)	2.035 (0.017)
Hierarchical complete Jaccard	Sensitivity	0.212 (0.002)	0.104 (0.005)	0.404 (0.016)	0.002 (0.001)	0.008 (0.003)
	Specificity	0.978 (0.001)	0.978 (0.004)	0.591 (0.014)	0.965 (0.002)	0.945 (0.003)
	PMSE	1.662 (0.017)	2.045 (0.022)	2.121 (0.017)	2.083 (0.019)	2.060 (0.018)
Hierarchical single Jaccard	Sensitivity	0.211 (0.002)	0.097 (0.004)	0.443 (0.015)	0.145 (0.011)	0.167 (0.011)
	Specificity	0.977 (0.001)	0.974 (0.004)	0.546 (0.015)	0.848 (0.011)	0.829 (0.012)
	PMSE	1.671 (0.018)	2.061 (0.022)	2.152 (0.018)	2.143 (0.019)	2.131 (0.019)
K-modes	Sensitivity	0.187 (0.002)	0.129 (0.002)	0.238 (0.008)	0.151 (0.000)	0.155 (0.001)
	Specificity	0.985 (0.001)	0.999 (0.000)	0.888 (0.010)	0.997 (0.000)	0.994 (0.000)
	PMSE	1.518 (0.017)	1.811 (0.017)	1.439 (0.015)	1.371 (0.015)	1.390 (0.015)
Hierarchical average Tanimoto	Sensitivity	0.209 (0.002)	0.081 (0.002)	0.611 (0.015)	0.278 (0.013)	0.311 (0.013)
	Specificity	0.977 (0.001)	0.994 (0.002)	0.422 (0.015)	0.766 (0.013)	0.737 (0.013)
	PMSE	1.698 (0.018)	1.974 (0.019)	2.056 (0.017)	2.122 (0.017)	2.087 (0.017)
Hierarchical complete Tanimoto	Sensitivity	0.211 (0.002)	0.101 (0.002)	0.203 (0.007)	0.087 (0.002)	0.145 (0.003)
	Specificity	0.977 (0.001)	0.999 (0.001)	0.840 (0.007)	0.946 (0.001)	0.897 (0.003)
	PMSE	1.664 (0.017)	1.893 (0.017)	1.957 (0.017)	1.898 (0.016)	1.916 (0.016)
Hierarchical single Tanimoto	Sensitivity	0.208 (0.002)	0.090 (0.004)	0.696 (0.015)	0.456 (0.016)	0.448 (0.016)
	Specificity	0.980 (0.001)	0.984 (0.003)	0.315 (0.014)	0.552 (0.015)	0.558 (0.015)
	PMSE	1.672 (0.017)	2.033 (0.021)	2.101 (0.018)	2.163 (0.018)	2.147 (0.018)
BNP	Sensitivity	0.201 (0.002)	0.165 (0.002)	0.504 (0.008)	0.213 (0.005)	0.462 (0.008)
	Specificity	0.979 (0.001)	0.998 (0.000)	0.642 (0.007)	0.877 (0.003)	0.674 (0.007)
	PMSE	1.614 (0.017)	1.796 (0.017)	1.668 (0.016)	1.700 (0.016)	1.670 (0.016)

Table 2 Simulation results on the performance of the LASSO procedure over 1000 replicates: Sensitivity, Specificity, PMSE (standard errors between brackets).

LASSO		
Sensitivity	Specificity	PMSE
0.236	0.963	1.698
(0.002)	(0.001)	(0.017)

References

1. Breheny, P., Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, **2**, 369–380.
2. Breheny, P. (2015). The Group Exponential Lasso for Bi-Level Variable Selection. *Biometrics*, **71**, 731–740.
3. Everitt, B., Landau, S., Leese, M., Stahl, D. (2011). Cluster analysis. 5th edn, Wiley, Chichester.
4. Galimberti, G., Montanari, A., Viroli, C. (2009). Penalized factor mixture analysis for variable selection in clustered data, *Computational statistics & data analysis*, **53**, 4301–4310.
5. Huang, Z (1998). Extensions to the v-means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, **2**, 28—304.
6. Huang, J., Breheny, P., Ma, S. (2012) A Selective Review of Group Selection in High-Dimensional Models. *Statistical Sciences*, **27**, 481–499.
7. Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.
8. Santra, T. (2016) A Bayesian non-parametric method for clustering high-dimensional binary data. `https://arxiv.org/pdf/1603.02494`.
9. Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461–464.
10. Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, **68**, 49–67.
11. Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, **38**, 894–942.