Regression modeling via latent predictors

Regressione basata su predittori latenti

Francesca Martella and Donatella Vicari

Abstract A proposal for multivariate regression modeling based on latent predictors (LPs) is presented. The idea of the proposed model is to predict the responses on LPs which, in turn, are built as linear combinations of disjoint groups of observed covariates. The formulation naturally allows to identify LPs that best predict the responses by jointly clustering the covariates and estimating the regression coefficients of the LPs. Clearly, in this way the LP interpretation is greatly simplified since LPs are exactly represented by a subset of covariates only. The model is formalized in a maximum likelihood framework which is intuitively appealing for comparisons with other methodologies, for allowing inference on the model parameters and for choosing the number of subsets leading to LPs. An Expectation Conditional Maximization (ECM) algorithm is proposed for parameter estimation and experiments on simulated and real data show the performance of our proposal.

Abstract *In questo lavoro si propone un nuovo modello di regressione basato su predittori latenti (PL) che, a loro volta, sono modellizzati come combinazioni lineari di gruppi disgiunti di covariate osservate. Tale formulazione permette di identificare direttamente i migliori PL che predicono le risposte attraverso: la classificazione delle covariate e la stima dei coefficienti di regressione. In questo modo l'interpretazione dei PL é notevolmente semplificata poiché i PL sono rappresentati solamente da sottoinsiemi di covariate. Il modello é formalizzato in un contesto di massima verosimiglianza e viene presentato un algoritmo Expectation Conditional Maximization (ECM) per la stima dei parametri. Esperimenti su dati simulati e reali mostrano l'utilitá e la validitá della proposta.*

Francesca Martella
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5 - 00185 Rome, e-mail: francesca.martella@uniroma1.it

Donatella Vicari
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5 - 00185 Rome e-mail: donatella.vicari@uniroma1.it

Key words: Regression model, Clustering, Latent Predictors, Maximum Likelihood

1 Introduction

Considering ordinary repression modeling of several dependent variables (responses) on a large set of covariates is not always the best choice. In fact, in such a case, difficulties in interpretation of the (many) regression coefficients and the presence of multicollinearity among predictors may arise. Many strategies can be adopted in order to reduce such problems such as standard variable selection methods, penalized (or shrinkage) techniques and dimensionality reduction methods (DRMs). The latter attempt to build small set of linear combinations of the predictors, used as input to the regression model, and differ in how the linear combinations are built. See among others, principal component regression (PCR, [7]), factor analysis regression (FAR, [2]), canonical correlation regression (CCR, [5]), partial least squares regression (PLSR, [12]), [10] for the continuum regression (unified regression technique embracing OLS, PLSR and PCR), reduced rank regression (RRR, [1], [6]), redundancy analysis (RA, [11]). [13] proposed a general formulation for dimensionality reduction and coefficient estimation in multivariate linear regression, which includes many existing DRMs as specific cases. Finally, [3] proposed a new formulation to the multiblock setting of latent root regression applied to epidemiological data and [4] investigated a continuum approach between MR and PLS. [8] proposed a multivariate regression model based on the optimal partition of predictors (MRBOP). A drawback of DRMs is that they may generally suffer from a possible difficulty of interpretability of the resulting linear combinations which are often overcome through rotation methods. Here, we propose to build latent predictors (LPs) as linear combinations of disjoint groups of covariates that best predict the responses where such groups identify block-correlated covariates. Actually, we simultaneously perform clustering of the covariates and estimation of the regression coefficients of the LPs. This turns out to be a relevant gain in the interpretation of the regression analysis, since LPs are formed by disjoint groups of covariates and, therefore easily interpretable. Clearly, in this way the LP interpretation is greatly simplified since LPs are exactly represented by a subset of covariates only. An Expectation Conditional Maximization algorithm (ECM, [9]), for maximum likelihood estimation of the model parameters is described. The performance of the proposed model is confirmed by the application on simulated and real data sets. The results are encouraging and would deserve further discussion.

2 Model

Consider $\mathbf{x}_i$ be a J-dimensional data vector representing the covariates and $\mathbf{y}_i$ be a M-dimensional data vector of the responses observed on the i-th unit in a sample of size n. Without loss of generality, $\mathbf{x}_i$ and $\mathbf{y}_i$ are assumed to be centered to zero mean vector. Our proposal can be summarized by two models: the first is a regression model formalizing the relations between responses and latent predictors, while the second one represents a dimensional reduction model where the latent predictors synthesize the relations among the covariates. In formula, we have

$$\mathbf{y}_i = \mathbf{C}'\mathbf{f}_i + \mathbf{e}_i \tag{1}$$

and

$$\mathbf{f}_i = \mathbf{V}'\mathbf{W}\mathbf{x}_i + \xi_i \tag{2}$$

where $\mathbf{C}$ is the $(Q \times M)$ regression coefficient matrix, $\mathbf{f}_i$ is the Q-th dimensional LP vector, $\mathbf{e}_i$ is the M-th dimensional noise term vector, $\mathbf{V}$ is the $(J \times Q)$ binary membership matrix defining a partition of the covariates in Q non-empty groups ($Q \leq J$), $\mathbf{W}$ is the $(J \times J)$ diagonal matrix which gives weights to the J covariates, ξ_i is the Q-th dimensional noise term vector ($i = 1, \ldots, n$).
Moreover, we assume that the noise terms are independent and follow multivariate Normal distributions: $\mathbf{e}_i \sim \text{MVN}(\mathbf{0}, \Sigma_\mathbf{e})$ with $\Sigma_\mathbf{e}$ diagonal matrix and $\xi_i \sim \text{MVN}(\mathbf{0}, \mathbf{I}_Q)$. Then, we derive that $\mathbf{f}_i \sim \text{MVN}(\mathbf{V}'\mathbf{W}\mathbf{x}_i, \mathbf{I}_Q)$, $\mathbf{y}_i \sim \text{MVN}(\mathbf{C}'\mathbf{V}'\mathbf{W}\mathbf{x}_i, \mathbf{C}'\mathbf{C} + \Sigma_\mathbf{e})$ and, conditional on $\mathbf{f}_i$, results in $\mathbf{y}_i|\mathbf{f}_i \sim \text{MVN}(\mathbf{C}'\mathbf{f}_i, \Sigma_\mathbf{e})$. Thus, the log-likelihood function $l(\Theta)$, being $\Theta = \{\mathbf{C}, \mathbf{V}, \mathbf{W}, \Sigma_\mathbf{e}\}$, is given by

$$\begin{aligned} l(\Theta) = &- \sum_{i=1}^{n} \left[(2\pi)^{M/2} |\mathbf{C}'\mathbf{C} + \Sigma_\mathbf{e}|^{1/2} \right] \\ &+ \sum_{i=1}^{n} \left\{ \frac{1}{2} \left(\mathbf{y}_i - \mathbf{C}'\mathbf{V}'\mathbf{W}\mathbf{x}_i \right)' \left(\mathbf{C}'\mathbf{C} + \Sigma_\mathbf{e} \right)^{-1} \left(\mathbf{y}_i - \mathbf{C}'\mathbf{V}'\mathbf{W}\mathbf{x}_i \right) \right\}. \end{aligned} \tag{3}$$

Finding the maximum likelihood estimates for $\mathbf{C}, \mathbf{W}, \Sigma_\mathbf{e}, \mathbf{V}$ is more problematic. We propose an ECM algorithm which iteratively computes the expected value of the complete-data log-likelihood and maximizes the expected complete-data log-likelihood over one of the parameters while holding the other fixed until convergence is achieved. Similarly to the factor analysis context, we take $\mathbf{y}$ as the observed data and $\mathbf{f}$ as the missing data, by assuming, therefore, that the complete data vector consist of $\mathbf{z}_i = (\mathbf{y}_i', \mathbf{f}_i')'$ ($i = 1, \ldots, n$). Therefore, the complete-data log-likelihood is given by:

$$\begin{aligned} \ell_C(\Theta) &= \sum_{i=1}^{n} \log \left[\phi(\mathbf{y}_i | \mathbf{f}_i, \Theta) \right] + \sum_{i=1}^{n} \log \left[\phi(\mathbf{f}_i | \Theta) \right] \\ &= \sum_{i=1}^{n} \log \left[\phi(\mathbf{y}_i | \mathbf{f}_i, \Theta) \right] + \sum_{i=1}^{n} \log \left[\phi(\mathbf{f}_i | \mathbf{W}, \mathbf{V}) \right] \end{aligned} \tag{4}$$

since the distribution of $\mathbf{f}_i$ is independent of $\mathbf{C}$ and $\Sigma_{\mathbf{e}}$. The expected value of $\mathbf{f}_i$ conditional on $\mathbf{y}_i$ and the current model parameters is

$$\mathrm{E}(\mathbf{f}_i|\mathbf{y}_i,\Theta) = \mathbf{V}'\mathbf{W}\mathbf{x}_i + \beta(\mathbf{y}_i - \mathbf{C}'\mathbf{V}'\mathbf{W}\mathbf{x}_i) \tag{5}$$

and

$$\begin{aligned}\mathrm{E}(\mathbf{f}_i\mathbf{f}_i'|\mathbf{y}_i,\Theta) = {} & (\mathbf{I}_Q - \beta\mathbf{C}') + [\mathbf{V}'\mathbf{W}\mathbf{x}_i + \beta(\mathbf{y}_i - \mathbf{C}'\mathbf{V}'\mathbf{W}\mathbf{x}_i)] \\ & [\mathbf{V}'\mathbf{W}\mathbf{x}_i + \beta(\mathbf{y}_i - \mathbf{C}'\mathbf{V}'\mathbf{W}\mathbf{x}_i)]'\end{aligned} \tag{6}$$

where $\beta = \mathbf{C}(\mathbf{C}'\mathbf{C} + \Sigma_{\mathbf{e}})^{-1}$. Therefore the expected complete-data log-likelihood Q is

$$\begin{aligned}Q = {} & K - \frac{n}{2}\log|\Sigma_{\mathbf{e}}| - \frac{1}{2}\sum_{i=1}^{n}\{\mathbf{y}_i'\Sigma_{\mathbf{e}}^{-1}\mathbf{y}_i - 2\mathbf{y}_i'\Sigma_{\mathbf{e}}^{-1}\mathbf{C}'\mathrm{E}(\mathbf{f}_i|\mathbf{y}_i,\Theta) \\ & + \mathrm{Tr}\left[\mathbf{C}\Sigma_{\mathbf{e}}^{-1}\mathbf{C}'\mathrm{E}(\mathbf{f}_i\mathbf{f}_i'|\mathbf{y}_i,\Theta)\right] + \mathrm{Tr}\left[\mathrm{E}(\mathbf{f}_i\mathbf{f}_i'|\mathbf{y}_i,\Theta)\right] \\ & - 2\mathrm{E}(\mathbf{f}_i|\mathbf{y}_i,\Theta)'\mathbf{B}_i\mathbf{w} + \mathrm{Tr}\left[\mathbf{w}'\mathbf{B}_i'\mathbf{B}_i\mathbf{w}\right]\}.\end{aligned} \tag{7}$$

where K is a constant, $\mathbf{w}$ is the J-dimensional vector of the diagonal elements of $\mathbf{W}$ (i.e. $\hat{\mathbf{W}} = \mathrm{diag}(\hat{\mathbf{w}})$), and $\mathbf{B}_i$ is the $(Q \times J)$ matrix having the j-th column equal to $\mathbf{v}_j x_{ij}$ with $\mathbf{v}_j$ being the j-th row of $\mathbf{V}$ ($i = 1,\ldots,n$).
Differentiating Q with respect to each parameter in Θ and setting to zero the corresponding score functions, we obtain

$$\hat{\mathbf{C}} = \left[\sum_{i=1}^{n}\mathrm{E}(\mathbf{f}_i\mathbf{f}_i'|\mathbf{y}_i,\Theta)\right]^{-1}\left[\sum_{i=1}^{n}\mathrm{E}(\mathbf{f}_i|\mathbf{y}_i,\Theta)\mathbf{y}_i'\right], \tag{8}$$

$$\hat{\Sigma}_{\mathbf{e}} = \frac{1}{n}\mathrm{diag}\left[\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i' - \sum_{i=1}^{n}\mathbf{y}_i\mathrm{E}(\mathbf{f}_i|\mathbf{y}_i,\Theta)'\mathbf{C}\right], \tag{9}$$

$$\hat{\mathbf{w}} = \left[\sum_{i=1}^{n}\mathbf{B}_i'\mathbf{B}_i\right]^{-1}\left[\sum_{i=1}^{n}\mathbf{B}_i'\mathrm{E}(\mathbf{f}_i|\mathbf{y}_i,\Theta)\right]. \tag{10}$$

In order to estimate the membership matrix of the covariates $\hat{\mathbf{V}}$, we proceed as follows:

- For each covariate j and group q, compute the log-likelihood values

$$l_{jq} = l(\cdot, v_{jq} = 1|\mathbf{C},\Sigma_{\mathbf{e}},\mathbf{W},\{v_{hs}\}_{h=1,\ldots,J,h\neq j;s=1,\ldots,Q,s\neq q});$$

- Fix j and compute the maximum of this set $\{l_{jq}\}$ over $q = 1,\ldots,Q$; denote this term by l_j^{max};
- Allocate the j-th covariate to the q-th group ($\hat{v}_{jq} = 1$) iff $l_{jq} = l_j^{max}$ $q = 1,\ldots,Q$.

The ECM algorithm for the proposed model therefore becomes:

- E-step: Compute the expected values $\mathrm{E}(\mathbf{f}_i|\mathbf{y}_i,\Theta)$ and $\mathrm{E}(\mathbf{f}_i\mathbf{f}_i'|\mathbf{y}_i,\Theta)$ for all data ($i = 1,\ldots,n$).
- CM-steps: Maximize Q over one of the parameters Θ while holding the other fixed.

The log-likelihood function, l, is computed for the current parameter values. The two steps are repeatedly alternated until convergence, which is reached when:

$$l_{(r)} - l_{(r-1)} < \varepsilon, \qquad \varepsilon > 0 \tag{11}$$

where r is the current iteration and ε is a small tolerance value.

3 Conclusions

A new mutivariate regression model based on latent predictors is presented. The latter are built as linear combinations of disjoint groups of observed covariates which best predict the responses. In this way, we jointly cluster covariates and estimate regression coefficients of the LPs. The model is particularly appropriate in a regression context where the reduction of the number of covariates is required for interpretability reasons or multicollinearity problems. In fact, in situations where the covariates are block-correlated, the assumptions on the covariances of the error terms, which are supposed diagonal, are fulfilled and lead to a gain in terms of parsimony and interpretability. We describe an EM algorithm for estimating model parameters and we will discuss the performance of the proposed approach on both simulated and real datasets. The results are encouraging and would deserve further discussion.

References

1. Anderson, T.W.: Estimating linear restrictions on regression coefficients for multivariate distributions. Ann. Math. Stat. **22**, 327–351 (1951)
2. Basilevsky, A.: Factor analysis regression. The Canadian Journal od Statistics, **9(1)**, 109–117 (1981)
3. Bougeard, S., Hanafi, M., Qannari, E.M.: Multiblock latent root regression: application to epidemiological data. Comput. Stat. **22(2)**, 209–222 (2007)
4. Bougeard, S., Hanafi, M., Qannari, E.M.: Continuum redundancy-PLS regression: a simple continuum approach. Comput. Stat. Data Anal. **52(7)**, 3686–3696 (2008)
5. Hotelling, H.: The most predictable criterion. J. Educ. Psychol. **25**, 139–142 (1935)
6. Izenman, A.J.: Reduced-rank regression for the multivariate linear model. J. Multivar. Anal. **5**, 248–262 (1975)
7. Kendall, M.G.: A Course in Multivaraite Analysis. Griffin, London (1957)
8. Martella, F., Vicari D., Vichi, M.: Partitioning predictors in multivariate regression models. Stat. Comput. **25(2)**, 261–272 (2013)
9. Meng, Xiao-Li, Rubin, D,: Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika. **80(2)**, 267-278 (1993).
10. Stone, M., Brooks, R.J.: Continuum regression: cross-validated squares partial least squares and principal components regression. J. R. Stat. Soc. b **52(2)**, 237–269 (1990)

11. Van Den Wollenberg, A.L.: Redundancy analysis an alternative for canonical analysis. Psychometrika **42(2)**, 207–219 (1977)
12. Wold, H.: Estimation of principal components and relates models by iterative least squares. in Krishnaiaah, P.R. (ed.) Multivariate Analysis, 391–420. Academic Press, New York (1966)
13. Yuan, M., Ekici, A., Lu, Z., Monteiro, R.: Dimension reduction and coefficient estimation in multivariate linear regression. J. R. Stat. Soc., Ser. B, Stat. Methodol. **69**, 329–346 (2007)