Journal of Statistical Planning and Inference 149 (2014) 125–139

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Regularized multivariate regression models with skew-*t* error distributions

Lianfu Chen [a], Mohsen Pourahmadi [a,*], Mehdi Maadooliat [b]

[a] *Department of Statistics, Texas A&M University, United States*
[b] *Department of Mathematics, Statistics & Computer Science, Marquette University, United States*

ARTICLE INFO

Article history:
Received 8 February 2013
Received in revised form
22 November 2013
Accepted 3 February 2014
Available online 14 February 2014

Keywords:
Cross-validation
ECM algorithm
Lasso regression
Likelihood function
Multivariate skew-*t*
Penalty

ABSTRACT

We consider regularization of the parameters in multivariate linear regression models with the errors having a multivariate skew-*t* distribution. An iterative penalized likelihood procedure is proposed for constructing sparse estimators of both the regression coefficient and inverse scale matrices simultaneously. The sparsity is introduced through penalizing the negative log-likelihood by adding L_1-penalties on the entries of the two matrices. Taking advantage of the hierarchical representation of skew-*t* distributions, and using the expectation conditional maximization (ECM) algorithm, we reduce the problem to penalized normal likelihood and develop a procedure to minimize the ensuing objective function. Using a simulation study the performance of the method is assessed, and the methodology is illustrated using a real data set with a 24-dimensional response vector.

1. Introduction

Multivariate linear regression analysis is concerned with linear relationships among q response variables $Y_1, Y_2, \ldots, Y_q$ and a single set of p predictor variables $x_1, x_2, \ldots, x_p$:

$$Y_k = b_{1k}x_1 + \cdots + b_{pk}x_p + \varepsilon_k, \quad 1 \le k \le q.$$

Suppose that $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iq})^\top$ is the ith observation of the response variables, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ is the corresponding values of the predictor variables and $\boldsymbol{\epsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iq})^\top$ is the vector of errors. Then, the multivariate linear regression can be simply expressed in matrix form as

$$\boldsymbol{Y} = \mathbf{X}\mathbf{B} + \boldsymbol{E}, \tag{1}$$

where $\boldsymbol{Y}_{n\times q} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^\top$, $\mathbf{X}_{n\times p} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, $\mathbf{B}_{p\times q} = (b_{jk})$ and $\boldsymbol{E}_{n\times q} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)^\top$ are the response, predictor, regression coefficient and error matrices, respectively.

The multivariate linear regression has been widely applied in many areas, such as chemometrics, econometrics and social sciences. The errors $\boldsymbol{\epsilon}_i$'s are commonly assumed to be independent and normally distributed (see Anderson, 2003), i.e., $\boldsymbol{\epsilon}_i \overset{i.i.d.}{\sim} \mathrm{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$ and the regression coefficient matrix $\mathbf{B}$ and covariance matrix $\boldsymbol{\Sigma}$ are estimated via their maximum likelihood

* Corresponding author.
E-mail addresses: lianfuch@gmail.com (L. Chen), pourahm@stat.tamu.edu (M. Pourahmadi), mehdi@mscs.mu.edu (M. Maadooliat).

http://dx.doi.org/10.1016/j.jspi.2014.02.001
0378-3758

estimates (MLEs):

$$\hat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{Y}, \quad \hat{\boldsymbol{\Sigma}} = \boldsymbol{S} = \frac{1}{n}(\boldsymbol{Y}-\mathbf{X}\hat{\mathbf{B}})^\top(\boldsymbol{Y}-\mathbf{X}\hat{\mathbf{B}}).$$

The MLE of $\mathbf{B}$ turns out to be equivalent to regressing each response variable on the same set of predictors *separately*, so that it does not account for the correlations among the response variables. Moreover, for high dimensional data, particularly when p and q are larger than n, the regression coefficient matrix $\mathbf{B}$ cannot be computed using the above formula since $\mathbf{X}$ is not of full rank, and the sample covariance matrix is known to be a highly unstable estimator of $\boldsymbol{\Sigma}$ (Ledoit and Wolf, 2004). In these situations, the traditional estimators for $\mathbf{B}$ and $\boldsymbol{\Sigma}$ with pq and $q(q+1)/2$ parameters, respectively, have rather poor performances and are not suitable for prediction and other applications. In recent years, various alternatives involving dimension-reduction and regularization have been proposed in the literature where the focus is on estimating either $\mathbf{B}$ or $\boldsymbol{\Sigma}$ alone. In the context of regularization, these methods are referred to under the headings of regularized multivariate regression (Peng et al., 2010) and regularized covariance estimation (Bickel and Levina, 2008), respectively.

Reduction of the pq parameters in the regression coefficient matrix $\mathbf{B}$ is usually done through the classical dimension reduction techniques such as the reduced-rank regression (Reinsel and Velu, 1998), criterion-based model selection methods (Bedrick and Tsai, 1994), principal component regression, partial least squares (Vinzi et al., 2010) and linear factor regression (Carvalho et al., 2005). The more modern approach is to reduce the number of regression parameters through regularization which may force some entries of $\mathbf{B}$ to zero (Yuan and Lin, 2006). These two broad approaches can be unified and viewed as estimating $\mathbf{B}$ by solving the following constrained optimization problem:

$$\hat{\mathbf{B}} = \arg\min_{\mathbf{B}}\{\mathrm{tr}[(\boldsymbol{Y}-\mathbf{X}\mathbf{B})^\top(\boldsymbol{Y}-\mathbf{X}\mathbf{B})]\} \quad \text{subject to}: \quad C(\mathbf{B}) \le t, \tag{2}$$

where $C(\mathbf{B})$ is a scalar function, and t is a nonnegative number. An early and natural constraint is $C(\mathbf{B}) = \sum_{j,k} b_{jk}^2$ so that (2) reduces to solving a ridge regression problem. The L_1-norm constraint or $C(\mathbf{B}) = \sum_{j,k}|b_{jk}|$ leads to the Lasso (Tibshirani, 1996) estimate of $\mathbf{B}$. Using the Lagrangian form of the Lasso penalty, this optimization problem takes the form

$$\hat{\mathbf{B}} = \arg\min_{\mathbf{B}}\left\{\mathrm{tr}[(\boldsymbol{Y}-\mathbf{X}\mathbf{B})^\top(\boldsymbol{Y}-\mathbf{X}\mathbf{B})] + \lambda\sum_{j,k}|b_{jk}|\right\}, \tag{3}$$

where λ is the tuning parameter.

Covariance estimation is an important problem in many areas of statistics dealing with correlated data (Pourahmadi, 2011). A wide range of alternatives to the sample covariance matrix has been developed in the last decade or so involving regularization of large covariance matrices. A common approach is the ridge regularization which estimates the covariance matrix as an optimal linear combination of the sample covariance matrix and the identity matrix (Ledoit and Wolf, 2004; Warton, 2008). Recently, sparse estimators of the covariance matrix $\boldsymbol{\Sigma}$ and the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ are proposed by adding to the normal likelihood a Lasso penalty on their off-diagonal entries (Bickel and Levina, 2008; Friedman et al., 2008; Rothman et al., 2008; Mazumder and Hastie, 2012). For normally distributed data, a penalized likelihood approach for joint estimation of $(\mathbf{B}, \boldsymbol{\Omega})$ has been proposed in Rothman et al. (2010) and further studied by Lee and Liu (2012). The associated optimization problem is not convex in $(\mathbf{B}, \boldsymbol{\Omega})$, and is known to be computationally demanding and unstable when p and q are large relative to n. In particular, for $p > n$ the MLE of the precision matrix can diverge to infinity (Lee and Liu, 2012, p. 245).

In practice, the normality assumption is usually violated because of the presence of skewness and kurtosis in real data (Hill and Dixon, 1982). Thus, one may seek more flexible parametric families of multivariate distributions capable of modeling such features of the data. The family of skew-normal distributions with a vector parameter to capture the skewness in the data has been widely studied due to its mathematical tractability and appealing probabilistic properties (Azzalini and Dalla-Valle, 1996; Azzalini and Capitanio, 1999; Azzalini, 2005). An extension of the skew-normal distribution is the multivariate skew-t distribution which allows for both nonzero skewness and heavy tails in the distribution (Branco and Dey, 2001). Some of the probabilistic properties of the skew-t distributions and their applications were investigated by Azzalini and Capitanio (2003). For a general background on the skew-normal and related distributions, see Genton (2004). A practical problem when fitting such models to data is the divergence of the maximum likelihood estimator of the skewness parameter in finite samples (see Section 5 of Azzalini and Capitanio, 1999).

In this paper, we assume that the errors $\boldsymbol{\epsilon}_i$'s in the model (1) have a multivariate skew-t distribution and consider regularizing the two matrices jointly. Our approach relies on and is closely related to that of Rothman et al. (2010) in which sparse estimators for both $\mathbf{B}$ and $\boldsymbol{\Sigma}$ are constructed simultaneously by minimizing the penalized negative normal log-likelihood:

$$\begin{aligned} L_p(\mathbf{B}, \boldsymbol{\Omega}) \propto{} & \frac{1}{n}\,\mathrm{tr}\{\boldsymbol{\Omega}(\boldsymbol{Y}-\mathbf{X}\mathbf{B})^\top(\boldsymbol{Y}-\mathbf{X}\mathbf{B})\} - \log|\boldsymbol{\Omega}| \\ & + \lambda_1 \sum_{k' \neq k}|\omega_{k'k}| + \lambda_2\sum_{j,k}|b_{jk}|, \end{aligned} \tag{4}$$

where $\boldsymbol{\Omega} = (\omega_{k'k}) = \boldsymbol{\Sigma}^{-1}$ and λ_1, λ_2 are the two tuning parameters to be determined from the data. The analogue of (4) for multivariate skew-t is quite complicated. However, since a multivariate skew-t distribution is conditionally normal given the relevant latent variables we use the expectation conditional maximization (ECM) algorithm (Meng and Rubin, 1993) to