Marquette University
e-Publications@Marquette

Mathematics, Statistics and Computer Science Faculty Research and Publications

Mathematics, Statistics and Computer Science, Department of

6-1-2014

Regularized Multivariate Regression Models with Skew-*t* Error Distributions

Lianfu Chen
Texas A & M University

Mohsen Pourahmadi
Texas A & M University

Mehdi Maadooliat
Marquette University, mehdi.maadooliat@marquette.edu

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.*
[b]*Department of Mathematics, Statistics & Computer Science, Marquette University*

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 we assess the performance of the method, and the methodology is illustrated using a real data set with a 24-dimensional response vector.

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

---

1. Introduction

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

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

Email addresses: `lianfuch@gmail.com` (Lianfu Chen), `pourahm@stat.tamu.edu` (Mohsen Pourahmadi), `mehdi.maadooliat@marquette.edu` (Mehdi Maadooliat)

Suppose that $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \cdots, y_{iq})^\top$ is the ith observation of the response variables, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^\top$ is the corresponding values of the predictor variables and $\boldsymbol{\epsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{iq})^\top$ 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 $\boldsymbol{Y}_{n \times q} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)^\top$, $\mathbf{X}_{n \times p} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^\top$, $\mathbf{B}_{p \times q} = (b_{jk})$ and $\boldsymbol{E}_{n \times q} = (\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_n)^\top$ 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., $\boldsymbol{\epsilon}_i \overset{i.i.d.}{\sim} N_q(\mathbf{0}, \boldsymbol{\Sigma})$ and the regression coefficient matrix $\mathbf{B}$ and covariance matrix $\boldsymbol{\Sigma}$ are estimated via their maximum likelihood estimates (MLEs):

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

The MLE for $\mathbf{B}$ is equivalent to regressing each response variable on the same set of predictors *separately* so that the estimation method 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 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, so that one must seek workable alternatives by regularizing these parameters. Traditionally, this has been done individually by focusing on regularizing $\mathbf{B}$ or $\boldsymbol{\Sigma}$ alone under the headings of regularized multivariate regression (Peng et al., 2010) or regularized covariance estimation (Bickel and Levina, 2008), respectively.

The standard way to reduce the pq parameters in the regression coefficient matrix $\mathbf{B}$ is through 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 modern approach is to reduce the number of regression parameters through regularization which may force some entries of $\mathbf{B}$ to be zero (Yuan and Lin, 2006). These two broad approaches can be unified and viewed as estimating $\mathbf{B}$ by solving the following constraint optimization problem:

$$\hat{\mathbf{B}} = \arg\min_{\mathbf{B}} \left\{ \operatorname{tr}\left[(\boldsymbol{Y} - \mathbf{XB})^\top (\boldsymbol{Y} - \mathbf{XB}) \right] \right\} \quad \text{subject to:} \quad C(\mathbf{B}) \leq t, \quad (2)$$

where $C(\mathbf{B})$ is a scalar function of $\mathbf{B}$, 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 well-known L_1-norm constraint, i.e., $C(\mathbf{B}) = \sum_{j,k} |b_{jk}|$ leads to the Lasso (Tibshirani, 1996) estimate of $\mathbf{B}$. Using the Lagrangian form of the Lasso penalty, this optimization problem takes the form

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

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 which involve regularizing large covariance matrices. For unordered multivariate data, an early and common approach is the ridge regularization which estimates the covariance matrix by an optimal linear combination of the sample covariance matrix and the identity matrix (Ledoit and Wolf, 2004; Warton, 2008). Recently, sparse estimators of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ are proposed by adding to the normal likelihood a Lasso penalty on its off-diagonal entries (Bickel and Levina, 2008; Friedman et al., 2008; Rothman et al., 2008; Mazumder and Hastie, 2012), see Pourahmadi (2011) for a recent review. A penalized likelihood approach for joint estimation of $(\mathbf{B}, \boldsymbol{\Omega})$ for Gaussian data 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), so one

may seek more flexible parametric families of multivariate distributions to model such features of the data. The family of the skew-normal distributions which generalizes the multivariate normal distributions with a vector parameter to model the skewness 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 the 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 $\boldsymbol{\epsilon}_i's$ have a multivariate skew-t distribution and consider the "small n, large p and q" case and regularize the two matrices jointly. Our approach relies on and is closely related to the recent work 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:

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

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. Exploiting the fact that a multivariate skew-t distribution is conditionally Gaussian given the relevant latent variables and using the ECM algorithm our problem is reduced to minimizing an objective function $L_c(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \nu)$ which is closely related to but more complicated than (4).

Compared with the normal log-likelihood in (4), $L_c(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \nu)$ has two additional terms involving the degrees of freedom ν and the skewness parameter $\boldsymbol{\eta}$. As expected when the skewness parameter is zero, $L_c(\mathbf{B}, \boldsymbol{\Omega}, 0, \nu)$ reduces to the complete-data log-likelihood for a symmetric multivariate t-distribution where regularizing $\mathbf{B}$ and $\boldsymbol{\Omega}$ are studied in Chen (2012). Otherwise, for a given $(\mathbf{B}, \boldsymbol{\Omega})$, $\boldsymbol{\eta}$ is estimated using a least-square method. Though, one has the option of penalizing the skewness vector $\boldsymbol{\eta}$ when its dimension q is

large, our experience and numerical results so far show that imposing such a penalty leads to estimating the skewness vector by zero. We have postponed the more elaborate study of regularizing $\boldsymbol{\eta}$ in conjunction with estimation of $(\mathbf{B}, \boldsymbol{\Omega})$ to a future publication. For a recent and possibly relevant study of penalizing the skewness parameter, see Azzalini and Arellano-Valle (2013).

The remainder of this paper is organized as follows. In Section 2, we introduce the theoretical underpinnings of our methodology for estimating multivariate regression via penalized skew-t likelihood using the ECM algorithm. We reduce the key computational component for the skew-t distribution to that of applying the ECM algorithm for skew normal distributions to a transformed set of responses and covariates. The selection of the tuning parameters is discussed in Section 2.5. We conduct a simulation study and investigate the performance of the method in terms of the prediction error in Section 3. In Section 4, we apply our methodology to the electricity wholesale spot prices in Australia. Some basic properties of skew-normal and skew-t distributions are reviewed in Appendix A and proofs of the results are given in Appendices B-E.

2. Penalized Skew-normal and Skew-t log-Likelihoods

In this section, we provide an algorithm called the MRST algorithm, for the joint estimation of the parameters $(\mathbf{B}, \boldsymbol{\Omega})$ of a multivariate regression model in (1) using penalized skew-t log-likelihoods, and the expectation conditional maximization (ECM) algorithm. We find it instructive to present the ECM algorithm for the skew-normal errors first, since the key computational component for the skew-t model is reduced to applying ECM of skew-normal model to the transformed data, see (16) and Section 2.4.

2.1. The Penalized Skew-Normal log-Likelihood

Assuming that the errors $\boldsymbol{\epsilon}_i's \overset{i.i.d.}{\sim} \mathrm{SN}_q(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ for $1 \leq i \leq n$, the negative log-likelihood for the observations $\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n$ is (Genton, 2004)

$$L(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}) \propto \mathrm{tr}\left\{\boldsymbol{\Omega}\boldsymbol{S}\right\} - \log|\boldsymbol{\Omega}| - \frac{2}{n}\sum_{i=1}^{n}\log\left[\Phi\{\boldsymbol{\eta}^\top(\boldsymbol{y}_i - \mathbf{B}^\top\boldsymbol{x}_i)\}\right], \qquad (5)$$

where $\boldsymbol{\Omega} = (\omega_{k'k}) = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{S} = \dfrac{1}{n}(\boldsymbol{Y} - \mathbf{X}\mathbf{B})^\top(\boldsymbol{Y} - \mathbf{X}\mathbf{B})$.

As in Rothman et al. (2010), we regularize the entries of $\mathbf{B}$ and $\boldsymbol{\Omega}$ using L_1 penalties, and estimate them by minimizing the penalized negative log-likelihood:

$$L(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}) + \lambda_1 \sum_{k' \neq k} |\omega_{k'k}| + \lambda_2 \sum_{j,k} |b_{jk}|. \tag{6}$$

Compared with the penalized normal likelihood in Rothman et al. (2010), because of the additional third term in (5) involving the skewness parameters, numerical optimization of (6) would be more challenging. To overcome the difficulty, we rely on the conditional Gaussianity of skew-normal distributions (Appendix A.1) and propose an iterative procedure to minimize (6) using an extension of EM, called Expectation Conditional Maximization (ECM) algorithm (see Meng and Rubin, 1993).

2.2. An Optimization Algorithm via ECM

As described in Appendix A.1, suppose that $u_1, \cdots, u_n$ are the positive latent variables associated with the observations $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$, respectively. Treating $(\boldsymbol{y}_i, u_i)$ for $i = 1, \cdots, n$ as the complete data, let $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)^\top$ and $\mathbf{U} = (u_1, \cdots, u_n)^\top$, then the negative complete-data log-likelihood is

$$L_c(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}) \propto \mathrm{tr}\{\boldsymbol{\Omega}\boldsymbol{S}\} - \log |\boldsymbol{\Omega}| + \frac{1}{n}\|\mathbf{U} - (\boldsymbol{Y} - \mathbf{X}\mathbf{B})\boldsymbol{\eta}\|^2,$$

where $\|\cdot\|$ is the L_2 norm. Adding two penalty terms on the entries of $\boldsymbol{\Omega}$ and $\mathbf{B}$ yields the penalized complete-data likelihood which is

$$L_p(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}) = L_c(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}) + \lambda_1 \sum_{k' \neq k} |\omega_{k'k}| + \lambda_2 \sum_{j,k} |b_{jk}|. \tag{7}$$

Minimizing the function in (6) is equivalent to minimizing $L_p(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta})$ using the EM algorithm (Dempster et al., 1977) which performs an expectation (E) step and a maximization (M) step alternately until convergence. In the E-step, the expectation of $L_p(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta})$ conditional on the observed data $\boldsymbol{Y}$ is evaluated using the current estimate of the parameters $\boldsymbol{\Theta} = \{\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}\}$; in the M-step, then the expected log-likelihood function is minimized over the parameter space. We describe the details as follows:

E-step: On the $(m+1)$th iteration, compute the conditional expectation of $L_p(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta})$ given the current estimate of the parameters $\boldsymbol{\Theta}^{(m)} = \{\mathbf{B}^{(m)}, \boldsymbol{\Omega}^{(m)}, \boldsymbol{\eta}^{(m)}\}$ and the observation matrix $\boldsymbol{Y}$.

In the E-step, we only have to calculate the conditional expectation of u_i given $\boldsymbol{\Theta}^{(m)}$ and $\boldsymbol{Y}$. Using the formula in (A.7), we denote this conditional expectation by $u_i^{(m)} = E(u_i|\boldsymbol{Y}, \boldsymbol{\Theta}^{(m)})$. Therefore, the expected log-likelihood, denoted by $Q(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}; \boldsymbol{\Theta}^{(m)})$, is

$$
\begin{aligned}
Q(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}; \boldsymbol{\Theta}^{(m)}) \quad \propto \quad & \mathrm{tr}\{\boldsymbol{\Omega S}\} - \log|\boldsymbol{\Omega}| + \frac{1}{n}\|\mathbf{U}^{(m)} - (\boldsymbol{Y} - \mathbf{XB})\boldsymbol{\eta}\|^2 \\
+ \quad & \mathrm{Var}(u_i|\boldsymbol{Y}, \boldsymbol{\Theta}^{(m)}) + \lambda_1\sum_{k'\neq k}|\omega_{k'k}| + \lambda_2\sum_{j,k}|b_{jk}|, \quad (8)
\end{aligned}
$$

where $\mathbf{U}^{(m)} = (u_1^{(m)}, \cdots, u_n^{(m)})^\top$ and since $\mathrm{Var}(u_i|\boldsymbol{Y}, \boldsymbol{\Theta}^{(m)})$ is a constant and is ignored in the minimization step. Note that the function in (8) is not convex in $(\mathbf{B}, \boldsymbol{\Omega})$, but it is convex in one argument when the other is fixed. This suggests an iterative algorithm alternating between estimation of $\mathbf{B}$ and $\boldsymbol{\Omega}$.

M-step: Minimizing $Q(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}; \boldsymbol{\Theta}^{(m)})$ over the whole parameter space $\boldsymbol{\Theta}$ is complicated, so we replace the M-step with the following three computationally simpler conditional minimization (CM) steps in which each block of parameters in $\boldsymbol{\Theta}$ is minimized while the other blocks are fixed (see Meng and Rubin, 1993).

CM1: Given $\boldsymbol{\eta} = \boldsymbol{\eta}^{(m)}$ and $\mathbf{B} = \mathbf{B}^{(m)}$, minimizing $Q(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}; \boldsymbol{\Theta}^{(m)})$ with respect to $\boldsymbol{\Omega}$ is equivalent to solving

$$
\boldsymbol{\Omega}^{(m+1)} = \arg\min_{\boldsymbol{\Omega}} \left\{ -\log|\boldsymbol{\Omega}| + \mathrm{tr}\{\boldsymbol{\Omega S}^{(m)}\} + \lambda_1\sum_{k'\neq k}|\omega_{k'k}| \right\}, \quad (9)
$$

where $\boldsymbol{S}^{(m)} = \frac{1}{n}\left(\boldsymbol{Y} - \mathbf{XB}^{(m)}\right)^\top \left(\boldsymbol{Y} - \mathbf{XB}^{(m)}\right)$.

This is the L_1 penalized covariance estimation problem and the fast DP-GLasso algorithm (Mazumder and Hastie, 2012) is adopted to solve (9). The estimate $\boldsymbol{\Omega}^{(m+1)}$ would remain positive definite as long as $S^{(m)}$ is positive definite.

CM2: Given $\boldsymbol{\Omega} = \boldsymbol{\Omega}^{(m+1)}$ and $\mathbf{B} = \mathbf{B}^{(m)}$, $\boldsymbol{\eta}$ can be simply updated by the ordinary least-squares estimate:

$$
\boldsymbol{\eta}^{(m+1)} = \left\{ \left(\mathbf{R}^{(m)}\right)^\top \mathbf{R}^{(m)} \right\}^{-1} \left(\mathbf{R}^{(m)}\right)^\top \mathbf{U}^{(m)}, \quad (10)
$$

where $\mathbf{R}^{(m)} = \boldsymbol{Y} - \mathbf{XB}^{(m)}$.

CM3: Given $\boldsymbol{\Omega} = \boldsymbol{\Omega}^{(m+1)}$ and $\boldsymbol{\eta} = \boldsymbol{\eta}^{(m+1)}$, finding the minimizer of $Q(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}; \boldsymbol{\Theta}^{(m)})$ with respect to $\mathbf{B}$ is equivalent to minimizing, (after some algebra, see the Appendix C):

$$\frac{1}{n}\mathrm{tr}\left\{\left(\tilde{\boldsymbol{Y}}^{(m+1)} - \mathbf{XB}\right)\tilde{\boldsymbol{\Omega}}^{(m+1)}\left(\tilde{\boldsymbol{Y}}^{(m+1)} - \mathbf{XB}\right)^{\top}\right\} + \lambda_2\sum_{j,k}|b_{jk}|, \quad (11)$$

where $\tilde{\boldsymbol{\Omega}}^{(m+1)} = \boldsymbol{\Omega}^{(m+1)} + \boldsymbol{\eta}^{(m+1)}\left(\boldsymbol{\eta}^{(m+1)}\right)^{\top}$ and $\tilde{\boldsymbol{Y}}^{(m+1)} = \boldsymbol{Y} - \mathbf{U}^{(m)}\left(\boldsymbol{\eta}^{(m+1)}\right)^{\top}(\tilde{\boldsymbol{\Omega}}^{(m+1)})^{-1}$.

We can write (11) as a Lasso regression and use the coordinate descent (Cod) algorithm to solve it.

2.3. The MRSN Algorithm for Skew-Normal Distributions

We summarize the preceding ECM algorithm for minimizing (7) and refer to it as the MRSN algorithm.

MRSN Algorithm:

With λ_1 and λ_2 fixed, initialize the parameters $\boldsymbol{\Theta} = \boldsymbol{\Theta}^{(0)} = (\mathbf{B}^{(0)}, \boldsymbol{\Omega}^{(0)})$ where $\mathbf{B}^{(0)}$ is the ridge estimate corresponding to λ_2 and $\boldsymbol{\Omega}^{(0)} = (\boldsymbol{S}^{(0)} + \lambda_1\mathbf{I})^{-1}$ with $\boldsymbol{S}^{(0)}$ the sample covariance matrix of the ridge regression residuals. On the $(m+1)^{\mathrm{th}}$ iteration,

---

E-step: Estimate the latent variables u_i by their conditional expectations as in (A.7).

CM1: Given $\mathbf{B} = \mathbf{B}^{(m)}$ and $\boldsymbol{\eta} = \boldsymbol{\eta}^{(m)}$, update $\boldsymbol{\Omega} = \boldsymbol{\Omega}^{(m+1)}$ in (9) using the DP-GLasso algorithm.

CM2: Given $\mathbf{B} = \mathbf{B}^{(m)}$ and $\boldsymbol{\Omega} = \boldsymbol{\Omega}^{(m+1)}$, update $\boldsymbol{\eta} = \boldsymbol{\eta}^{(m+1)}$ with the least square estimate in (10).

CM3: Given $\boldsymbol{\eta} = \boldsymbol{\eta}^{(m+1)}$ and $\boldsymbol{\Omega} = \boldsymbol{\Omega}^{(m+1)}$, update $\mathbf{B} = \mathbf{B}^{(m+1)}$ in (11) using the coordinate descent (Cod) algorithm.

Repeat the E- and three CM-steps until the estimates of the parameters converge, that is, $\sum_{j,k}|\hat{b}_{jk}^{(m+1)} - \hat{b}_{jk}^{(m)}| \leq \epsilon_1\sum_{jk}|\hat{b}_{jk}^{ridge}|$, where $\hat{\mathbf{B}}^{ridge} = (\hat{b}_{jk}^{ridge}) = (\mathbf{X}^{\top}\mathbf{X} + \lambda_2\mathbf{I})^{-1}\mathbf{X}^{\top}\boldsymbol{Y}$ is the ridge estimate. The analogous check for the convergence of the precision matrix is, $|\boldsymbol{\Omega}^{(m+1)} - \boldsymbol{\Omega}^{(m)}|_1 \leq \epsilon_2|(\boldsymbol{S}^{(0)} + \lambda_1\mathbf{I})^{-1}|_1$. The tolerance parameters ϵ_1 and ϵ_2 are set at 10^{-4} by default.

The MRSN algorithm is similar to the MRCE method for the normal data, but here we need an E-step for the estimation of the latent variables and an extra CM step for estimation of the skewness parameter. Consequently, the MRSN would be slower than the MRCE. Note that in the absence of asymmetry where $\boldsymbol{\alpha} = 0$, the E and CM2 steps are not needed, and the MRSN method would reduce to the MRCE algorithm for the normal data which is known to be slow for large p and q (see Rothman et al., 2010).

2.4. The MRST Algorithm for Skew-t Errors

In this section, we present the MRST algorithm when the errors $\boldsymbol{\epsilon}'_i s$ in (1) have a multivariate skew-t distribution with ν degrees of freedom so that the parameters are $\boldsymbol{\Theta} = \{\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \nu\}$. It will be shown that the three CM steps here are precisely the same as those in the MRSN algorithm applied to a modified data.

Let u_i and w_i be the corresponding latent variables for $\boldsymbol{y}_i$ with the negative complete-data log-likelihood given by

$$
\begin{aligned}
L_c(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \nu) \;\;\propto\;\; & \frac{1}{n}\mathrm{tr}\left\{\boldsymbol{\Omega}(\boldsymbol{Y} - \mathbf{X}\mathbf{B})^\top \boldsymbol{W}(\boldsymbol{Y} - \mathbf{X}\mathbf{B})\right\} - \log|\boldsymbol{\Omega}| \\
& + \frac{1}{n}\|\mathbf{U} - \sqrt{\boldsymbol{W}}(\boldsymbol{Y} - \mathbf{X}\mathbf{B})\boldsymbol{\eta}\|^2 + g(\nu),
\end{aligned} \tag{12}
$$

where $\boldsymbol{\Omega}, \boldsymbol{\eta}$, $\boldsymbol{W} = \mathrm{diag}\{\mathrm{w}_1, \cdots, \mathrm{w}_n\}$ and $\mathbf{U} = (\mathrm{u}_1, \cdots, \mathrm{u}_n)^\top$ with $(\mathrm{w}_i, \mathrm{u}_i)$ being the two positive latent variables associated with $\boldsymbol{y}_i$ are defined in Appendix A.2, and

$$
g(\nu) = 2\log\Gamma\left(\frac{\nu}{2}\right) - \nu\log\left(\frac{\nu}{2}\right) - \frac{1}{n}(\nu + q - 2)\sum_{i=1}^{n}\log\mathrm{w}_i + \frac{\nu}{n}\sum_{i=1}^{n}\mathrm{w}_i, \tag{13}
$$

involves the degrees of freedom ν; see Appendix A.2 for more details.

The complete-data sufficient statistics are

$$
\boldsymbol{S}_{YWY} = \boldsymbol{Y}^\top \boldsymbol{W}\boldsymbol{Y}, \quad \boldsymbol{S}_{XWX} = \mathbf{X}^\top \boldsymbol{W}\mathbf{X}, \quad \boldsymbol{S}_{YWX} = \boldsymbol{Y}^\top \boldsymbol{W}\mathbf{X}, \quad S_W = \sum_{i=1}^{n}\mathrm{w}_i,
$$

$$
\boldsymbol{S}_{U\sqrt{W}X} = \mathbf{U}^\top\sqrt{\boldsymbol{W}}\mathbf{X}, \quad \boldsymbol{S}_{U\sqrt{W}Y} = \mathbf{U}^\top\sqrt{\boldsymbol{W}}\boldsymbol{Y}, \quad S_{\log W} = \sum_{i=1}^{n}\log\mathrm{w}_i. \tag{14}
$$

Similar to the skew-normal case, we construct sparse estimates for both $\mathbf{B}$ and $\boldsymbol{\Omega}$ by minimizing the penalized negative log-likelihood:

$$
L_p(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \nu) = L_c(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \nu) + \lambda_1 \sum_{k' \neq k}|\omega_{k'k}| + \lambda_2 \sum_{j,k}|b_{jk}|, \tag{15}
$$

via the following ECM algorithm.

E-step: On the $(m+1)$th iteration, calculate the conditional expectation of $L_p(\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta}, \nu)$ given the current estimates of the parameters $\mathbf{\Theta}^{(m)}$ and the data $\boldsymbol{Y}, \mathbf{X}$.

From the forms of the sufficient statistics given in (14), the three expectations $a_i^{(m)} = E(\log \mathrm{w}_i | \mathbf{X}, \boldsymbol{Y}, \mathbf{\Theta}^{(m)})$, $b_i^{(m)} = E(\mathrm{w}_i | \mathbf{X}, \boldsymbol{Y}, \mathbf{\Theta}^{(m)})$ and $c_i^{(m)} = E(\mathrm{u}_i \sqrt{\mathrm{w}_i} | \mathbf{X}, \boldsymbol{Y}, \mathbf{\Theta}^{(m)})$ are needed and can be evaluated using Proposition 1 in Appendix A.2. Note that $E(\mathrm{w}_i | \mathbf{X}, \boldsymbol{Y}, \mathbf{\Theta}^{(m)})$ and $E(\mathrm{u}_i \sqrt{\mathrm{w}_i} | \mathbf{X}, \boldsymbol{Y}, \mathbf{\Theta}^{(m)})$ have closed forms, but $E(\log \mathrm{w}_i | \mathbf{X}, \boldsymbol{Y}, \mathbf{\Theta}^{(m)})$ does not, so we compute the latter numerically using the QUADPACK package developed in Piessen et al. (1983).

Set $\tilde{\boldsymbol{y}}_i^{(m)} = \sqrt{b_i^{(m)}} \boldsymbol{y}_i$, $\tilde{\boldsymbol{x}}_i^{(m)} = \sqrt{b_i^{(m)}} \boldsymbol{x}_i$, $\tilde{\mathrm{u}}_i^{(m)} = c_i^{(m)} / \sqrt{b_i^{(m)}}$, $\tilde{\boldsymbol{Y}}^{(m)} = \left(\tilde{\boldsymbol{y}}_1^{(m)}, \cdots, \tilde{\boldsymbol{y}}_n^{(m)} \right)^\top$, $\tilde{\mathbf{U}}^{(m)} = \left(\tilde{\mathrm{u}}_1^{(m)}, \cdots, \tilde{\mathrm{u}}_n^{(m)} \right)^\top$ and $\tilde{\mathbf{X}}^{(m)} = \left(\tilde{\boldsymbol{x}}_1^{(m)}, \cdots, \tilde{\boldsymbol{x}}_n^{(m)} \right)^\top$, then the expected log-likelihood function $Q(\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta}, \nu; \mathbf{\Theta}^{(m)})$ can be written (see the Appendix E) as

$$Q(\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta}, \nu; \mathbf{\Theta}^{(m)}) \propto \mathrm{tr}\left\{ \mathbf{\Omega} \tilde{\boldsymbol{S}}^{(m)} \right\} - \log |\mathbf{\Omega}| + \frac{1}{n} \| \tilde{\mathbf{U}}^{(m)} - \left(\tilde{\boldsymbol{Y}}^{(m)} - \tilde{\mathbf{X}}^{(m)} \mathbf{B} \right) \boldsymbol{\eta} \|^2$$
$$+ g^{(m)}(\nu) + \lambda_1 \sum_{k' \neq k} |\omega_{k'k}| + \lambda_2 \sum_{j,k} |b_{jk}|, \tag{16}$$

where $\tilde{\boldsymbol{S}}^{(m)} = \frac{1}{n} \left(\tilde{\boldsymbol{Y}}^{(m)} - \tilde{\mathbf{X}}^{(m)} \mathbf{B} \right)^\top \left(\tilde{\boldsymbol{Y}}^{(m)} - \tilde{\mathbf{X}}^{(m)} \mathbf{B} \right)$ and

$$g^{(m)}(\nu) = 2 \log \Gamma \left(\frac{\nu}{2} \right) - \nu \log \left(\frac{\nu}{2} \right) - \frac{1}{n}(\nu + q - 2) \sum_{i=1}^n a_i^{(m)} + \frac{\nu}{n} \sum_{i=1}^n b_i^{(m)}. \tag{17}$$

Since the degrees of freedom ν is separated from the other three blocks of parameters, the M-step for the skew-t distribution proceeds as follows:

CM1: Given the first three blocks of parameters in $\mathbf{\Theta} = (\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta}, \nu)$, update ν as $\nu^{(m+1)}$ by minimizing the function $g^{(m)}(\nu)$ in (17).

CM2: Given $\nu = \nu^{(m+1)}$, the three blocks of parameters will be estimated using exactly the same three CM steps as in the MRSN algorithm with the modified data matrices $\tilde{\mathbf{X}}^{(m)}$, $\tilde{\boldsymbol{Y}}^{(m)}$ and $\tilde{\mathbf{U}}^{(m)}$.

The preceding ECM algorithm for minimizing (15) is referred to as the MRST algorithm.

Remark 1: We point out a curious challenge when estimating the degrees of freedom ν. In practice, we have noticed that the sequence $\{\nu^{(m)}\}$ usually converges to a small positive number less than 2, whereas $\nu > 2$ is required for the existence of the covariance matrix of the skew-t distribution. Thus, the estimate of the degrees of freedom using the MRST algorithm is not satisfactory; a phenomenon which also occurs when $\epsilon_i's$ have a symmetric multivariate t-distribution (Chen, 2012). In most of what follows, we discard the CM1 step in the MRST algorithm and estimate ν separately via the maximum likelihood method (Azzalini and Capitanio, 2003).

Remark 2: When $\boldsymbol{\alpha} = \mathbf{0}$, the MRST algorithm reduces to the MRMT algorithm (Chen, 2012) developed to regularize parameters in the general linear model when the errors have a symmetric multivariate t-distribution. Moreover, if $\boldsymbol{\alpha} = \mathbf{0}$ and ν goes to infinity, the MRST algorithm reduces to the MRCE method (Rothman et al., 2010).

2.5. Tuning Parameters and Performance Measures

We use the K-fold cross-validation to select the tuning parameters over a grid of values of (λ_1, λ_2). In the cross-validation, the dataset $S = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) : 1 \leq i \leq n\}$ is randomly partitioned into K groups of roughly equal size, denoted by $S_k, k = 1, 2, \cdots, K$. For each k, we use $S \backslash S_k$ as the training data to estimate the parameters and S_k as the test set to evaluate the prediction error. The tuning parameter (λ_1, λ_2) is chosen as the minimizer of the mean squared prediction error over all q variables of the response, that is,

$$(\hat{\lambda}_1, \hat{\lambda}_2) = \arg \min_{(\lambda_1, \lambda_2)} \frac{1}{Kq} \left\{ \sum_{k=1}^{K} \| \boldsymbol{Y}_{(k)} - \mathbf{X}_{(k)} \hat{\mathbf{B}}_{(-k)}^{\lambda_1, \lambda_2} - \hat{\boldsymbol{\mu}}_{(-k)}^{\lambda_1, \lambda_2} \|^2 \right\}, \qquad (18)$$

where $\boldsymbol{Y}_{(k)}, \mathbf{X}_{(k)}$ are the validation response matrix and the predictor matrix formed from the subset S_k, $\hat{\mathbf{B}}_{(-k)}^{\lambda_1, \lambda_2}$ and $\hat{\boldsymbol{\mu}}_{(-k)}^{\lambda_1, \lambda_2}$ are the corresponding estimates of $\mathbf{B}$ and the mean vector of the errors with the training data $S \backslash S_k$.

The cross-validation based on the log-likelihood function is commonly used in the literature for selecting the tuning parameter and estimating the covariance matrix (Huang et al., 2006; Cai et al., 2012; Lee and Liu, 2012). In our simulation study, the estimated inverse scale matrix ($\boldsymbol{\Omega}$) and the associated tuning parameter (λ_1) chosen by the log-likelihood cross-validation outperforms the one selected by cross-validation based on the prediction error.

3. A Simulation Study

In this section, through a simulation study we assess and compare the performance of our method for multivariate regression having skew-t errors with that of MRCE (Rothman et al., 2010) for normal and MRMT (Chen, 2012) for symmetric multivariate t-distributions, respectively. The MRMT relies on the coordinate descent algorithm to solve the Lasso regression problems encountered in the iterations.

In the simulation study and real data analysis, we measure the overall performance of these methods in terms of the mean squared prediction error (PE). While L_2 loss is used to evaluate the performance of $\hat{\mathbf{B}}$, the sparsity recognition performance of $\hat{\mathbf{B}}$ is measured by the true positive rate (TPR) as well as the true negative rate (TNR) defined as

$$\text{TPR}(\hat{\mathbf{B}}, \mathbf{B}) = \frac{\#\{(i,j) : \hat{b}_{ij} \neq 0 \quad \text{and} \quad b_{ij} \neq 0\}}{\#\{(i,j) : b_{ij} \neq 0\}}, \tag{19}$$

$$\text{TNR}(\hat{\mathbf{B}}, \mathbf{B}) = \frac{\#\{(i,j) : \hat{b}_{ij} = 0 \quad \text{and} \quad b_{ij} = 0\}}{\#\{(i,j) : b_{ij} = 0\}}. \tag{20}$$

The TPR is the proportion of nonzero elements in $\mathbf{B}$ that $\hat{\mathbf{B}}$ identifies correctly, while the TNR measures the proportion of zero elements recognized correctly. They should be considered simultaneously since $\hat{\mathbf{B}} = 0$ always has perfect TNR and the OLS estimate always has perfect TPR.

The performance of estimators of the inverse scale matrix ($\mathbf{\Omega}$) is assessed using the following two standard loss functions:

$$\Delta_1(\mathbf{\Omega}, G) = \text{tr}\mathbf{\Omega}^{-1}G - \log|\mathbf{\Omega}^{-1}G| - n \qquad \text{and} \qquad \Delta_2(\mathbf{\Omega}, G) = \text{tr}(\mathbf{\Omega}^{-1}G - I)^2,$$

where $\mathbf{\Omega}$ is the true inverse covariance matrix and G is a positive-definite matrix with the same size. $\Delta_1(\mathbf{\Omega}, G)$ is known as the entropy loss and $\Delta_2(\mathbf{\Omega}, G)$ is called the quadratic loss. Both of these loss functions are 0, when $G = \mathbf{\Omega}$ and positive, when $G \neq \mathbf{\Omega}$. The corresponding risk functions are defined as:

$$R_i(\mathbf{\Omega}, G) = E_{\mathbf{\Omega}}\{\Delta_i(\mathbf{\Omega}, G)\}, \qquad i = 1, 2.$$

An estimator $\hat{\mathbf{\Omega}}$ is better than $\tilde{\mathbf{\Omega}}$, if its associated risk function is smaller, that is, $R_i(\mathbf{\Omega}, \hat{\mathbf{\Omega}}) < R_i(\mathbf{\Omega}, \tilde{\mathbf{\Omega}})$.

3.1. Model Design

Throughout this section we use 80 replications of the multivariate regression with $n = 50, 150$, $p = 22$ and $q = 24$, where the p, q are chosen to match the dimensions of the regression models fitted to the electricity data analyzed in the next section.

In each replication a sparse matrix $\mathbf{B}$ is generated as the elementwise product of the following three matrices:

$$\mathbf{B} = \boldsymbol{W} * \boldsymbol{K} * \boldsymbol{Q},$$

where $(\boldsymbol{W})_{ij} \overset{i.i.d.}{\sim} N(0,1), (\boldsymbol{K})_{ij} \overset{i.i.d.}{\sim}$ Bernoulli(s_1) and each row of $\boldsymbol{Q}$ is either a vector of 1's or 0's with a success probability of 1's equal to s_2. Generating $\mathbf{B}$ in this way, we expect $(1-s_2)p$ predictors to be irrelevant for all q responses, and each predictor to be relevant for $s_1 q$ of all the response variables. An $n \times p$ predictor matrix $\mathbf{X}$ is also generated with rows drawn independently from $N(0, \boldsymbol{\Sigma}_X)$, where $(\boldsymbol{\Sigma}_X)_{ij} = 0.7^{|i-j|}$, as in Yuan and Lin (2007) and Peng et al. (2010). We consider the AR(1) covariance model as the scale matrix of the errors with $(\boldsymbol{\Sigma})_{ij} = \rho^{|i-j|}$.

Finally, each row of the error matrix $\boldsymbol{E}$ is independently drawn from a multivariate skew-t distribution $\text{St}_q(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \nu)$ and the response matrix $\boldsymbol{Y}$ is constructed using $\boldsymbol{Y} = \mathbf{XB} + \boldsymbol{E}$. To save computation time, we independently generate a validation data of the same sample size $n = 50$ within each replication to estimate the prediction error for the algorithms as in Rothman et al. (2010). This is similar to performing a K-fold cross-validation for the algorithm.

We consider different combinations of $\nu, \boldsymbol{\alpha}, \rho, s_1$ and s_2 from the following ranges: (1) $\nu = 5, 15, 40$, (2) $\rho = 0, 0.4, 0.8$, (3) $\boldsymbol{\alpha} = (-1, 1, -1, \cdots, 1)^\top$ or $\mathbf{1}_q$, where $\mathbf{1}_q$ is a column vector of ones, (4) $s_1 = 0.1, 0.5$, and (5) $s_2 = 1$. The tuning parameters λ_1 and λ_2 are selected from the set $\Lambda = \{2^x : x = 0, \pm 1, \cdots, \pm 5\}$ such that the selected λ_1 and λ_2 maximize the log-likelihood function based on the validation data. Since the conclusions drawn for the two skewness vectors of $\boldsymbol{\alpha}$ are nearly the same, we only present the results for $\boldsymbol{\alpha} = \mathbf{1}_q$ here.

3.2. Results and Discussion

We obtained the prediction errors (PE), values of the L_2 loss functions of estimators of $\mathbf{B}$, and the quadratic and entropy loss functions of estimators

of $\mathbf{\Omega}$ for the simulated datasets using MRCE, MRSN, MRMT and MRST algorithms (see the Supplemantary Material). Given that, we work with three choices of *df* (ν={5,15,40}), two choices of sample size ($n = \{50, 150\}$), two choices of sparsity ($s_1 = \{0.1, 0.5\}$) and three choices of AR(1) parameter ($\rho = \{0, 0.4, 0.8\}$), there are 36 possible combinations indicated by the four-tuple index $[\nu, n, s_1, \rho]$ in Figure 1. The vertical lines correspond to an increase (change) in one of the components of the four-tuple. We also compare the numerical performance of these four algorithms and the Lasso obtained from (4) by setting $\lambda_1 = 0$ and $\mathbf{\Omega} = \mathbf{I}$ for the 36 combinations in the Supplemantary Material (Tables S.1 & S.2).

The following summary of results can be extracted from Figure 1:

- Increasing the sample size (n) and *df* (ν) leads to smaller values of the PE and other loss functions.

- Increasing the ρ leads to smaller values of PE and the L_2 loss function for $\mathbf{B}$, but higher loss for the matrix $\mathbf{\Omega}$. This might be due to $\mathbf{\Sigma}$ being ill-conditioned for ρ near one.

- Smaller s_1 or more sparse $\mathbf{B}$ results in smaller L_2 loss for $\mathbf{B}$ and PE.

- The plot of PE's in the bottom panel suggests that there is a clear distinction among the performances of the four algorithms for the lower *df* ($\nu = 5$), where MRST outperforms the others. This distinction vanishes for larger *df*s, where MRST converges to MRSN, and MRMT converges to MRCE. In spite of this, the former pair outperforms the latter.

The corresponding true positive rates (TPR) and true negative rates (TNR) for $\mathbf{B}$ are reported in Tables 1 and 2. From these tables and the Figure S.1 in the Supplementary Material, it is evident that a slightly higher TPR is accompanied by a lower TNR for the MRST model. Finally, the computational times of the four algorithms are computed and presented in the Supplementary Material. On average, the ratios of the CPU times of MRMT, MRSN and MRST to MRCE are 1.84, 4.79 and 5.72 respectively. The computations were done on "AMD Opteron 2350 clusters with 2.5GHz processors".

4. Real Data Analysis

In this section, we apply the MRST method to the hourly average electricity spot prices collected in the Australian state of New South Wales (NSW)

Figure 1: Comparison of MRST, MRMT, MRSN and MRCE for 36 different combinations of $[\nu, n, s_1, \rho]$, based on 80 simulation replicates. (A) median L_2 loss for estimating the $\mathbf{B}$; (B) median entropy loss for the matrix $\mathbf{\Omega}$; and (C) mean squared prediction error.
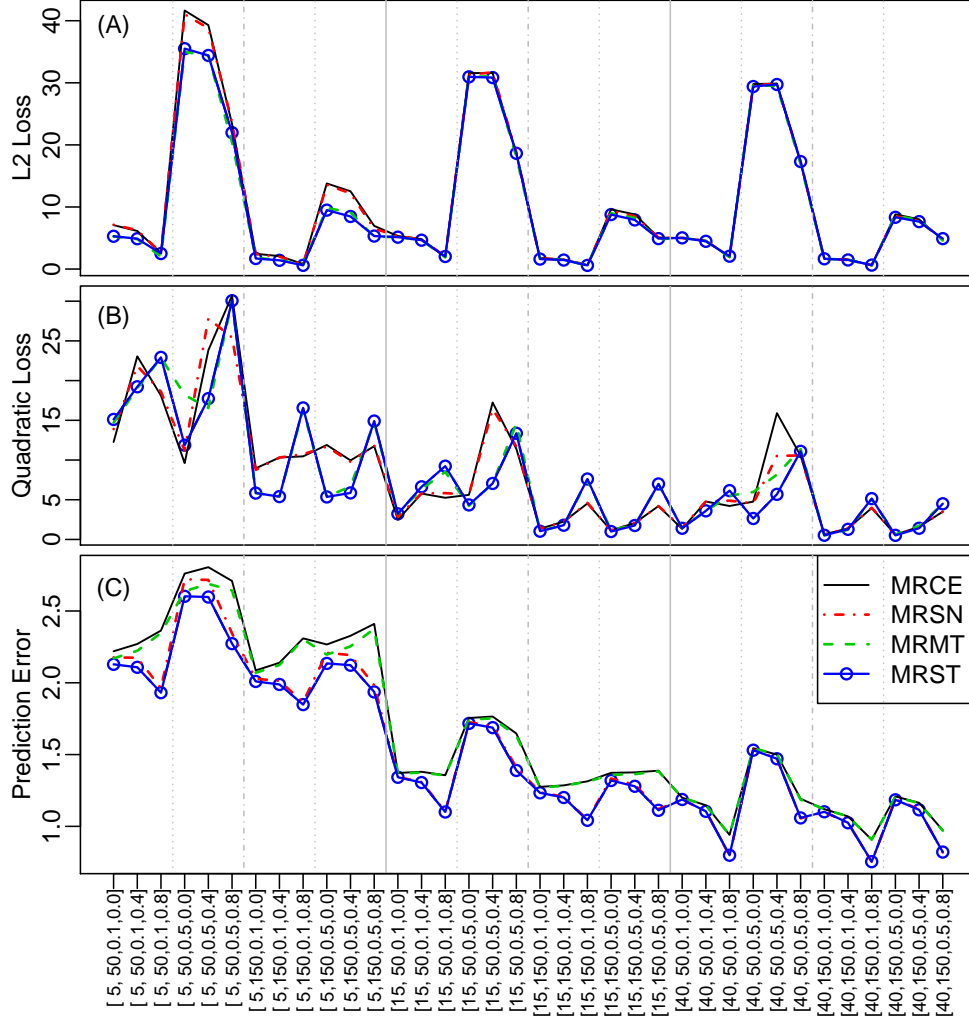
from July 2, 2003 to June 30, 2006, starting at 04:00 and ending at 03:00 each day. The dataset consists of $26,352$ observations during a period of $n = 1,098$ days and was previously analyzed in Panagiotelis and Smith (2008) using a Bayesian method and skew-t distribution (Sahu et al., 2003) for the data. Unlike other commodity prices, most electricity spot prices exhibit trend,

Table 1: TPR/TNR of MRST, MRMT, MRSN and MRCE for the matrix $\mathbf{B}$ averaged over 80 replications with $s_1 = 0.1$, $s_2 = 1$ and $\boldsymbol{\alpha} = (1, 1, 1, \cdots, 1)^\top$.

			MRCE	MRSN	MRMT	MRST
$\nu = 5$	$n = 50$	$\rho = 0$	78.24 / 78.64	78.52 / 77.82	80.43 / 80.20	80.88 / 79.75
		$\rho = 0.4$	79.26 / 79.39	78.98 / 79.78	82.10 / 78.60	82.02 / 78.69
		$\rho = 0.8$	86.88 / 76.79	87.13 / 75.72	88.27 / 76.04	88.49 / 73.19
	$n = 150$	$\rho = 0$	86.73 / 77.68	87.33 / 76.14	88.55 / 76.72	89.40 / 74.45
		$\rho = 0.4$	88.41 / 74.60	88.49 / 75.20	89.66 / 74.93	89.46 / 75.27
		$\rho = 0.8$	92.16 / 74.27	92.70 / 73.24	93.35 / 73.43	93.30 / 74.99
$\nu = 15$	$n = 50$	$\rho = 0$	81.68 / 76.34	81.62 / 75.56	81.79 / 75.71	82.19 / 76.40
		$\rho = 0.4$	82.70 / 76.26	82.53 / 77.12	82.81 / 75.99	83.07 / 76.71
		$\rho = 0.8$	88.75 / 76.14	88.35 / 75.64	88.47 / 76.33	88.38 / 76.64
	$n = 150$	$\rho = 0$	88.81 / 75.35	89.06 / 75.43	89.09 / 75.43	89.66 / 74.59
		$\rho = 0.4$	90.14 / 73.47	90.40 / 72.37	90.71 / 73.53	90.43 / 73.03
		$\rho = 0.8$	92.76 / 73.36	92.95 / 72.53	92.81 / 74.04	92.98 / 74.93
$\nu = 40$	$n = 50$	$\rho = 0$	81.88 / 79.84	81.62 / 79.98	81.28 / 80.64	81.70 / 79.91
		$\rho = 0.4$	82.64 / 79.34	82.44 / 79.43	82.39 / 79.27	82.30 / 79.47
		$\rho = 0.8$	87.81 / 79.25	87.84 / 79.89	87.78 / 79.97	87.53 / 81.50
	$n = 150$	$\rho = 0$	88.24 / 82.02	88.98 / 79.27	88.12 / 82.06	88.95 / 78.55
		$\rho = 0.4$	88.95 / 81.43	89.66 / 79.94	89.32 / 81.22	89.66 / 79.68
		$\rho = 0.8$	92.59 / 81.94	92.50 / 81.38	92.56 / 81.86	92.56 / 80.69

strong periodicity, intra-day and inter-day serial correlations, heavy tails, skewness and so on; see Panagiotelis and Smith (2008); Diongue et al. (2009) for some empirical evidence.

We consider the vector of the log spot prices at hourly intervals during a day as the response vector with $q = 24$. The profile plot of the observations in the first month (Figure 2) appears to be symmetric around the mean except that some skewness is observed at the times 08:00, 17:00-19:00 when the demand and electricity prices are highly volatile.

According to Panagiotelis and Smith (2008), the variables which may have effects on the spot prices as the predictors include a simple linear trend, dummy variables for day types (in total 13 dummy variables, representing the seven days of the week and some idiosyncratic public holidays) and eight seasonal polynomials (high order Fourier terms) for a smooth seasonal effect. We fit a multivariate linear regression of the hourly observations during a day on the covariates as follows

$$\boldsymbol{y}_i = \mathbf{B}^\top \boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \quad 1 \le i \le n, \tag{21}$$

where $\boldsymbol{y}_i$ is a 24×1 vector of log electricity prices on day i and $\boldsymbol{x}_i$ is the

Table 2: TPR/TNR of MRST, MRMT, MRSN and MRCE for the matrix $\mathbf{B}$ averaged over 80 replications with $s_1 = 0.5$, $s_2 = 1$ and $\boldsymbol{\alpha} = (1, 1, 1, \cdots, 1)^\top$.

			MRCE	MRSN	MRMT	MRST
$\nu = 5$	$n = 50$	$\rho = 0$	85.07 / 42.68	85.68 / 41.42	88.16 / 37.44	88.89 / 35.43
		$\rho = 0.4$	86.12 / 40.71	86.96 / 38.42	87.99 / 38.92	88.52 / 37.42
		$\rho = 0.8$	89.05 / 39.99	89.11 / 39.81	90.16 / 39.12	91.36 / 34.34
	$n = 150$	$\rho = 0$	90.66 / 46.44	91.06 / 45.39	92.88 / 43.17	92.97 / 43.58
		$\rho = 0.4$	91.29 / 44.28	92.11 / 41.92	93.21 / 42.20	93.27 / 41.60
		$\rho = 0.8$	94.13 / 40.86	94.60 / 39.02	95.76 / 36.93	95.97 / 36.81
$\nu = 15$	$n = 50$	$\rho = 0$	85.99 / 46.41	87.00 / 43.61	87.16 / 42.68	88.02 / 39.55
		$\rho = 0.4$	86.47 / 44.85	86.75 / 43.43	86.91 / 43.32	87.22 / 42.45
		$\rho = 0.8$	90.12 / 42.23	89.93 / 42.32	90.43 / 41.19	90.46 / 41.01
	$n = 150$	$\rho = 0$	91.87 / 48.35	91.72 / 48.25	92.01 / 47.50	92.35 / 46.54
		$\rho = 0.4$	92.49 / 45.72	92.60 / 45.45	92.70 / 45.25	92.92 / 44.44
		$\rho = 0.8$	95.19 / 39.45	95.68 / 36.44	95.51 / 38.40	95.84 / 37.00
$\nu = 40$	$n = 50$	$\rho = 0$	86.05 / 47.35	86.85 / 45.16	86.70 / 45.69	87.24 / 43.75
		$\rho = 0.4$	86.47 / 47.05	86.97 / 45.15	86.60 / 46.08	87.27 / 43.95
		$\rho = 0.8$	90.19 / 43.67	90.46 / 42.76	90.58 / 42.57	90.57 / 42.32
	$n = 150$	$\rho = 0$	92.22 / 47.45	92.52 / 46.80	92.19 / 47.48	92.60 / 46.30
		$\rho = 0.4$	92.96 / 45.36	93.10 / 45.30	92.96 / 45.16	93.21 / 44.91
		$\rho = 0.8$	95.37 / 40.70	95.14 / 42.16	95.40 / 39.96	95.26 / 40.85

corresponding vector of $p = 22$ covariates. Because of the apparent skewness in the profile plot in Figure 2, we model the error as $\boldsymbol{\epsilon}_i \sim \mathrm{St}_{24}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \nu)$.

The MLE of the degrees of freedom ν is $\hat{\nu} = 5.04$ using the whole dataset. With ν fixed at $\hat{\nu}$, we then apply the MRST method to the model (21). To access the predictive performance via the mean squared prediction error, we retain the observations from the last 100 days as the test set, while estimating the parameters using the rest of the observed spot prices. We select the tuning parameters (λ_1, λ_2) via a 5-fold cross-validation from the set $\Lambda = \left\{ 2^{-10+20(x-1)/39} : x = 1, 2, \cdots, 40 \right\}$.

The average squared prediction errors for each hour in a day for the last 100 days are plotted in Figure 3. While the overall average prediction errors are similar around the hour of 18 pm which is the most skewed or volatile period, the real differences emerge away from this time. In fact, the overall average prediction error using the MRST method turns out to be 0.075 which is the smallest among all the methods considered.

Figure 2: The profile plot of the hourly electricity wholesale prices. The dashed curve is the mean profile.
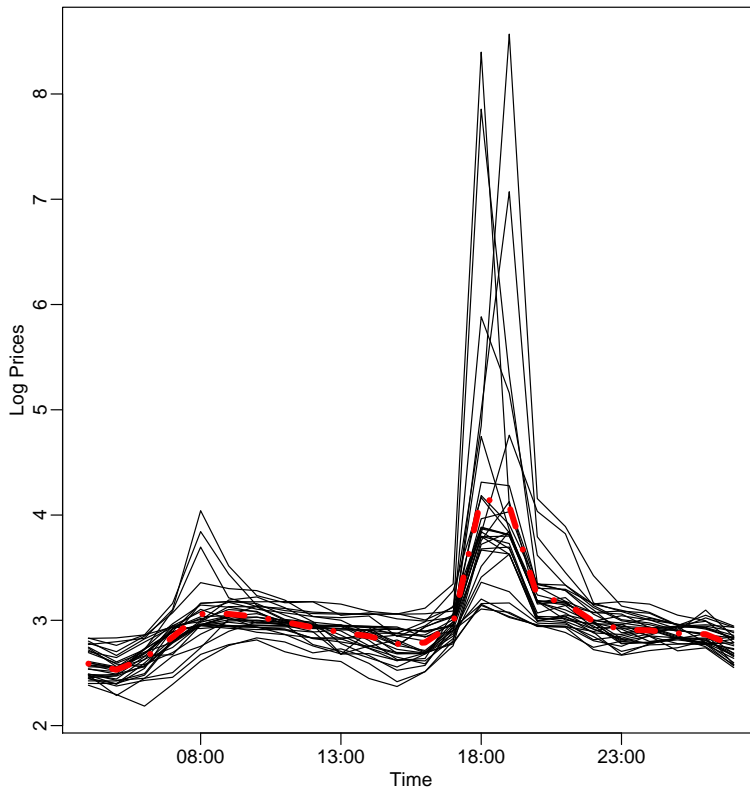
5. Summary and Conclusion

We have proposed an iterative procedure to construct sparse estimates for the regression coefficient and precision matrices simultaneously when the errors in the general linear model are skewed. The skew-t distribution is flexible for accounting for the possible skewness in the data. Two algorithms, namely MRSN for skew-normal and MRST for skew-t, which extend the MRCE (Rothman et al., 2010) and MRMT (Chen, 2012) algorithms are developed. When estimating the degrees of freedom for skew-t distribution, we encountered the same numerical problem as in Chen (2012) for symmetric t-distributions and our recommendation is to estimate it outside the ECM-

18

Figure 3: The average squared prediction error for each hour on a day based on 100 points.
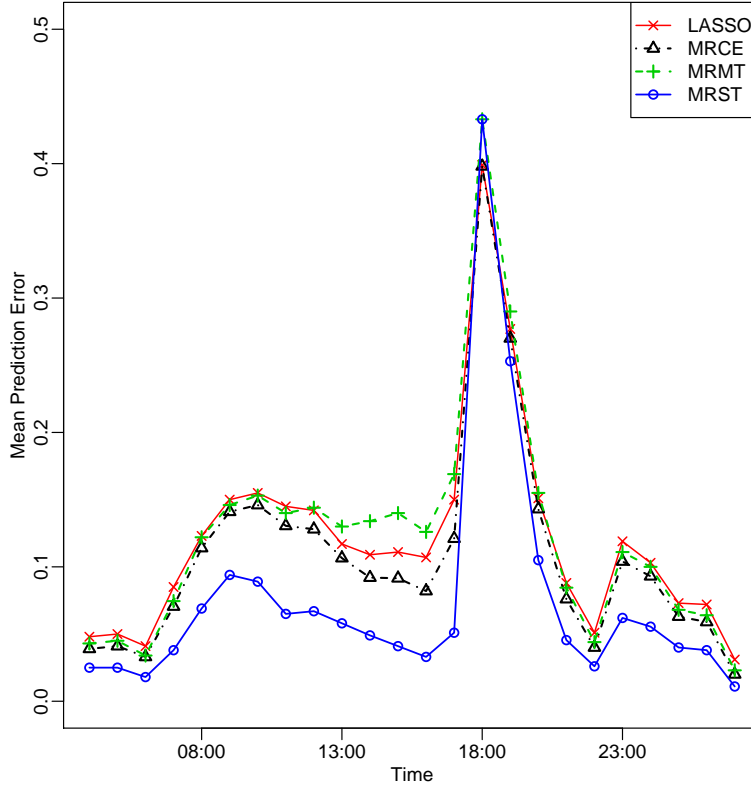
iterations. We have shown that the MRST outperforms the MRCE and MRMT in terms of prediction error when (1) $\mathbf{B}$ is less sparse or (2) $\mathbf{B}$ is sparse but $\mathbf{\Sigma}$ is highly correlated. However, the MRST and MRSN seem to be conservative in that the estimate of $\mathbf{B}$ is less sparse than that using MRCE and MRMT.

6. Acknowledgment

Appendix

Appendix A. Multivariate Skew-normal and -t Distributions

In this section, we briefly review the families of multivariate skew-normal (Azzalini and Dalla-Valle, 1996) and skew-t (Branco and Dey, 2001) distributions as well as some of their properties that would be used in developing the EM-type algorithms.

Appendix A.1. The Multivariate Skew-normal Distribution

A random vector Z is said to have a q-variate skew-normal distribution (Azzalini and Dalla-Valle, 1996) if its probability density function takes the form

$$f(\mathbf{z};\boldsymbol{\xi},\boldsymbol{\Sigma},\boldsymbol{\alpha}) = 2\phi_q(\mathbf{z};\boldsymbol{\xi},\boldsymbol{\Sigma})\Phi\{\boldsymbol{\alpha}^\top\boldsymbol{\sigma}^{-1}(\mathbf{z}-\boldsymbol{\xi})\}, \tag{A.1}$$

where $\phi_q(\cdot;\boldsymbol{\xi},\boldsymbol{\Sigma})$ is the *pdf* of the q-dimensional normal distribution with mean $\boldsymbol{\xi}$ and covariance matrix $\boldsymbol{\Sigma}$, $\Phi(\cdot)$ is the *cdf* of the univariate standard normal distribution, the vector $\boldsymbol{\alpha}$ plays the role of the skewness parameter where for $\boldsymbol{\alpha}=\mathbf{0}$ the above density reduces to the multivariate normal, and $\boldsymbol{\sigma}=\mathrm{diag}\{\sigma_{11}^{\frac{1}{2}},\cdots,\sigma_{qq}^{\frac{1}{2}}\}$ is a diagonal matrix of square roots of the diagonal elements of $\boldsymbol{\Sigma}$. We denote this distribution by $\mathsf{Z}\sim\mathrm{SN}_q(\boldsymbol{\xi},\boldsymbol{\Sigma},\boldsymbol{\alpha})$ where the parameters $\boldsymbol{\xi}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}$ shall be referred to as the location parameter, scale matrix and skewness parameter, respectively. Unlike the multivariate normal densities which are symmetric about the location parameter, the skew-normal densities in (A.1) are not symmetric. Its mean and covariance matrix are

$$\boldsymbol{\mu}=\boldsymbol{\xi}+\sqrt{\frac{2}{\pi}}\frac{\boldsymbol{\Sigma}\boldsymbol{\eta}}{(1+\boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta})^{1/2}}, \qquad \mathrm{Var}(\mathsf{Z})=\boldsymbol{\Sigma}-\boldsymbol{\mu}\boldsymbol{\mu}^\top, \tag{A.2}$$

which are different from $\boldsymbol{\xi}$ and $\boldsymbol{\Sigma}$, and $\boldsymbol{\eta}=\boldsymbol{\sigma}^{-1}\boldsymbol{\alpha}$ is a standardized version of the skewness vector.

Writing the <u>cdf</u> $\Phi(\cdot)$ in (A.1) as the integral of its <u>pdf</u> $\phi(\cdot)$ leads to the following integral representation of the $\mathrm{SN}_q(\boldsymbol{\xi},\boldsymbol{\Sigma},\boldsymbol{\alpha})$ density:

$$f(\mathbf{z};\boldsymbol{\xi},\boldsymbol{\Sigma},\boldsymbol{\alpha}) = \int_0^\infty 2\phi_q(\mathbf{z};\boldsymbol{\xi},\boldsymbol{\Sigma})\phi\{\mathrm{u}-\boldsymbol{\alpha}^\top\boldsymbol{\sigma}^{-1}(\mathbf{z}-\boldsymbol{\xi})\}d\mathrm{u}. \tag{A.3}$$

This suggests introducing a nonnegative random variable or a latent variable U such that the joint density function for (Z, U) is just the integrand in (A.3), i.e.,

$$f(\mathbf{z}, \mathrm{u}) = 2\phi_q(\mathbf{z}; \boldsymbol{\xi}, \boldsymbol{\Sigma})\phi\left\{\mathrm{u} - \boldsymbol{\alpha}^\top\boldsymbol{\sigma}^{-1}(\mathbf{z} - \boldsymbol{\xi})\right\}I(\mathrm{u} > 0). \qquad (A.4)$$

The marginal density of U turns out to be a truncated normal distribution, denoted by $U \sim \mathrm{TN}(0, 1 + \boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta})$, with the density

$$f(\mathrm{u}) = 2(1 + \boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta})^{-\frac{1}{2}}\phi\left\{(1 + \boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta})^{-\frac{1}{2}}\mathrm{u}\right\}I(\mathrm{u} > 0); \qquad (A.5)$$

see the Appendix B for more details. By definition, the conditional distribution of U given $\mathsf{Z} = \mathbf{z}$ is

$$f(\mathrm{u}|\mathbf{z}) = \frac{f(\mathbf{z}, \mathrm{u})}{f(\mathbf{z})} = \frac{\phi\left\{\mathrm{u} - \boldsymbol{\eta}^\top(\mathbf{z} - \boldsymbol{\xi})\right\}}{\Phi\left\{\boldsymbol{\eta}^\top(\mathbf{z} - \boldsymbol{\xi})\right\}}I(\mathrm{u} > 0), \qquad (A.6)$$

which is also a truncated normal distribution with the conditional mean

$$\hat{\mathrm{u}} = E(U|\mathbf{z}) = \boldsymbol{\eta}^\top(\mathbf{z} - \boldsymbol{\xi}) + \frac{\phi\left\{\boldsymbol{\eta}^\top(\mathbf{z} - \boldsymbol{\xi})\right\}}{\Phi\left\{\boldsymbol{\eta}^\top(\mathbf{z} - \boldsymbol{\xi})\right\}}. \qquad (A.7)$$

In the ECM algorithm for the skew-normal family, the formula (A.7) would be used to estimate the latent variable in the E-step; see Section 2.2.

Appendix A.2. The Multivariate Skew-t Distribution

Branco and Dey (2001) defined a new class of multivariate distributions via

$$\mathsf{Y} = \boldsymbol{\xi} + W^{-1/2}\mathsf{Z}, \qquad (A.8)$$

where $\mathsf{Z} \sim \mathrm{SN}_q(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ and $W \sim \chi_\nu^2/\nu$, independent of Z. The random vector Y is said to have a multivariate skew-t distribution, denoted by $\mathsf{Y} \sim \mathrm{St}_q(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \nu)$, with the density function

$$f(\boldsymbol{y}; \boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \nu) = 2t_q(\boldsymbol{y}; \nu)T_1\left\{\boldsymbol{\alpha}^\top\boldsymbol{\sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\xi})\left(\frac{\nu + q}{Q_{\boldsymbol{y}} + \nu}\right)^{1/2}; \nu + q\right\}, (A.9)$$

where $\boldsymbol{\sigma}$ is as in Appendix A.1,

$$Q_{\boldsymbol{y}} = (\boldsymbol{y} - \boldsymbol{\xi})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\xi}),$$

21

and

$$t_q(\boldsymbol{y}; \nu) = \frac{\Gamma\{(\nu + q)/2\}}{|\boldsymbol{\Sigma}|^{1/2}(\pi\nu)^{q/2}\Gamma(\nu/2)}\left(1 + \frac{Q\boldsymbol{y}}{\nu}\right)^{-(\nu+q)/2},$$

is the density function of a q-dimensional t-variate with degrees of freedom ν and $T_1(\cdot; \nu + q)$ denotes the cdf of a scalar t distribution with degrees of freedom $\nu + q$. The mean and covariance of Y are

$$\begin{aligned}
\boldsymbol{\mu} &= \boldsymbol{\xi} + b_\nu\boldsymbol{\sigma\delta}, \qquad \nu > 1, \\
\mathrm{Var}(\mathsf{Y}) &= \frac{\nu}{\nu - 2}\boldsymbol{\Sigma} - \boldsymbol{\sigma\mu\mu}^\top\boldsymbol{\sigma},
\end{aligned}$$

where

$$b_\nu = \left(\frac{\nu}{\pi}\right)^{1/2}\frac{\Gamma\left\{\frac{1}{2}(\nu - 1)\right\}}{\Gamma\left(\frac{1}{2}\nu\right)} \quad \text{and} \quad \boldsymbol{\delta} = \frac{\boldsymbol{\sigma}^{-1}\boldsymbol{\Sigma\eta}}{\sqrt{1 + \boldsymbol{\eta}^\top\boldsymbol{\Sigma\eta}}},$$

and $\nu > 2$ for the covariance matrix to exist.

It is known that if $\mathsf{Y} \sim \mathrm{St}_q(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \nu)$ is partitioned as

$$\mathsf{Y} = \begin{pmatrix}\mathsf{Y}_1 \\ \mathsf{Y}_2\end{pmatrix}, \quad \boldsymbol{\xi} = \begin{pmatrix}\boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2\end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}, \quad \boldsymbol{\alpha} = \begin{pmatrix}\boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2\end{pmatrix},$$

where Y_1 has dimension m, then the marginal distribution of Y_1 still belongs to the family of multivariate skew-t distributions, i.e., $\mathsf{Y}_1 \sim \mathrm{St}_m(\boldsymbol{\xi}_1, \boldsymbol{\Sigma}_{11}, \tilde{\boldsymbol{\alpha}}, \nu)$ (see Azzalini and Capitanio, 2003). However, the skewness parameter $\tilde{\boldsymbol{\alpha}}$ is a complicated function of $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$:

$$\tilde{\boldsymbol{\alpha}} = \frac{\boldsymbol{\alpha}_1 + \tilde{\boldsymbol{\Sigma}}_{11}^{-1}\tilde{\boldsymbol{\Sigma}}_{12}\boldsymbol{\alpha}_2}{(1 + \boldsymbol{\alpha}_2^\top\tilde{\boldsymbol{\Sigma}}_{22\cdot1}\boldsymbol{\alpha}_2)^{1/2}}, \tag{A.10}$$

where $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\sigma}^{-1}\boldsymbol{\Sigma\sigma}^{-1}$ is partitioned the same way as $\boldsymbol{\Sigma}$, and $\tilde{\boldsymbol{\Sigma}}_{22\cdot1} = \tilde{\boldsymbol{\Sigma}}_{22} - \tilde{\boldsymbol{\Sigma}}_{21}\tilde{\boldsymbol{\Sigma}}_{11}^{-1}\tilde{\boldsymbol{\Sigma}}_{12}$. This implies, in particular, that the ith component of Y has a univariate skew-t distribution, whose skewness parameter, denoted by $\tilde{\alpha}_i$, is different from α_i, the ith entry of $\boldsymbol{\alpha}$. In short, the vector $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, \cdots, \tilde{\alpha}_q)^\top$ of the marginal skewness parameters is quite different from $\boldsymbol{\alpha}$, and this might be the reason for the challenges in its estimation and interpretation.

From the definition of a multivariate skew-t in (A.8) as a scale mixture of multivariate skew-normals, $\mathsf{Y}|W = \mathrm{w} \sim \mathrm{SN}_q(\boldsymbol{\xi}, \mathrm{w}^{-1}\boldsymbol{\Sigma}, \sqrt{\mathrm{w}}\boldsymbol{\alpha})$ it follows that similar to the skew-normal case when developing the EM-type algorithm, it is natural to augment the observed data $\boldsymbol{y}$ by including the two latent variables

$$W \sim \chi_\nu^2/\nu \quad \text{and} \quad U \sim \mathrm{TN}(0, 1 + \boldsymbol{\eta}^\top\boldsymbol{\Sigma\eta}),$$

such that

$$f(\boldsymbol{y}, \mathrm{u}|W = \mathrm{w}) = 2\phi_q(\boldsymbol{y}; \boldsymbol{\xi}, \mathrm{w}^{-1}\boldsymbol{\Sigma})\phi(\mathrm{u} - \sqrt{\mathrm{w}}\boldsymbol{\sigma}^{-1}\boldsymbol{\alpha}^\top(\boldsymbol{y} - \boldsymbol{\xi}))I(\mathrm{u} > 0).$$

For convenience, we denote the pdf for Gamma(a, b) with mean a/b and variance a/b^2 by $h(\mathrm{w}; a, b)$, then the density function of W is $h(\mathrm{w}; \nu/2, \nu/2)$. The joint density of the complete-data (Y, U, W) at $(\boldsymbol{y}, \mathrm{u}, \mathrm{w})$ is

$$f(\boldsymbol{y}, \mathrm{u}, \mathrm{w}) = 2\phi_q(\boldsymbol{y}; \boldsymbol{\xi}, \mathrm{w}^{-1}\boldsymbol{\Sigma})\phi\left\{\mathrm{u} - \sqrt{\mathrm{w}}\boldsymbol{\eta}^\top(\boldsymbol{y} - \boldsymbol{\xi})\right\} \cdot h(\mathrm{w}; \nu/2, \nu/2)I(\mathrm{u} > 0).$$

The distributions of W and (W, U) conditional on $\mathsf{Y} = \boldsymbol{y}$ are given as follows

$$f(\mathrm{w}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}, \mathrm{w})}{f(\boldsymbol{y})} = \frac{2}{f(\boldsymbol{y})}\phi_q(\boldsymbol{y}; \boldsymbol{\xi}, \mathrm{w}^{-1}\boldsymbol{\Sigma})\Phi\{\sqrt{\mathrm{w}}\boldsymbol{\eta}^\top(\boldsymbol{y} - \boldsymbol{\xi})\}h(\mathrm{w}; \nu/2, \nu/2),\text{(A.11)}$$

and

$$f(\mathrm{w}, \mathrm{u}|\boldsymbol{y}) = \frac{2}{f(\boldsymbol{y})}\phi_q(\boldsymbol{y}; \boldsymbol{\xi}, \mathrm{w}^{-1}\boldsymbol{\Sigma})\phi\{\mathrm{u} - \sqrt{\mathrm{w}}\boldsymbol{\eta}^\top(\boldsymbol{y} - \boldsymbol{\xi})\}h\left(\mathrm{w}; \frac{\nu}{2}, \frac{\nu}{2}\right)I(\mathrm{u} > 0).\text{(A.12)}$$

The relevant conditional expectations needed for the EM algorithm are given next:

Proposition 1. Suppose that $\mathsf{Y} \sim \mathrm{St}_q(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \nu)$ with the associated latent variables $W \sim \chi_\nu^2/\nu$ and $U \sim \mathrm{TN}(0, 1 + \boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta})$. Then, for any $m > 0$, we have

$$E\left\{W^m|\boldsymbol{y}\right\} = C(\theta_1, r_1)T_1\left(\boldsymbol{\eta}^\top(\boldsymbol{y} - \boldsymbol{\xi})\sqrt{\frac{r_1}{\theta_1}}; 2r_1\right),$$

$$E\left\{UW^m|\boldsymbol{y}\right\} = \frac{1}{\sqrt{2\pi}}C(\theta_2, r_1) + \boldsymbol{\eta}^\top(\boldsymbol{y} - \boldsymbol{\xi}) \cdot E\left\{W^{m+\frac{1}{2}}|\boldsymbol{y}\right\},$$

where

$$r_1 = \frac{q + \nu}{2} + m, \qquad \theta_1 = \frac{(\boldsymbol{y} - \boldsymbol{\xi})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\xi}) + \nu}{2},$$

$$\theta_2 = \frac{(\boldsymbol{y} - \boldsymbol{\xi})^\top(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)(\boldsymbol{y} - \boldsymbol{\xi}) + \nu}{2},$$

$$C(x, y) = \left(\frac{1}{\sqrt{2\pi}}\right)^q \cdot \frac{2}{f(\boldsymbol{y})} \cdot |\boldsymbol{\Sigma}|^{-1/2}\frac{(\nu/2)^{\frac{\nu}{2}}}{\Gamma(\nu/2)} \cdot \frac{\Gamma(y)}{x^y}.$$

The formula for computing $E\{W^m|\boldsymbol{y}\}$ can also be found in the Corollary 1 of Lachos et al. (2010) where $\mathrm{E}\left\{W^{m/2}\phi[\sqrt{W}\boldsymbol{\eta}^\top(\boldsymbol{y}-\boldsymbol{\xi})]/\Phi[\sqrt{W}\boldsymbol{\eta}^\top(\boldsymbol{y}-\boldsymbol{\xi})]|\boldsymbol{y}\right\}$ is also calculated instead of $E\{UW^m|\boldsymbol{y}\}$.

Proof: For ease of notation, let $\theta = r = \frac{\nu}{2}$. The computation of the conditional expectations relies on the following result in Azzalini and Capitanio (2003):

Lemma. If $W \sim \mathrm{Gamma}(r, \theta)$, then for any $b \in R$,

$$E\left\{\Phi(b\sqrt{W})\right\} = T_1\left(b\sqrt{\frac{r}{\theta}}; 2r\right).$$

Using the density functions in (A.11) and (A.12), we have

$$
\begin{aligned}
E\{W^m|Y\} &= \frac{2}{f(\boldsymbol{y})}\int_0^\infty \mathrm{w}^m \Phi\{\sqrt{\mathrm{w}}\boldsymbol{\eta}^\top(\boldsymbol{y}-\boldsymbol{\xi})\}\phi_q(\boldsymbol{y};\boldsymbol{\xi},\mathrm{w}^{-1}\boldsymbol{\Sigma})\cdot h(\mathrm{w};\nu/2,\nu/2)d\mathrm{w} \\
&= \frac{2}{f(\boldsymbol{y})}\int_0^\infty \mathrm{w}^m \Phi\{\sqrt{\mathrm{w}}\boldsymbol{\eta}^\top(\boldsymbol{y}-\boldsymbol{\xi})\}(2\pi)^{-q/2}\exp\left\{-\frac{\mathrm{w}}{2}(\boldsymbol{y}-\boldsymbol{\xi})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\xi})\right\} \\
&\quad\times\left|\mathrm{w}^{-1}\boldsymbol{\Sigma}\right|^{-1/2}\frac{1}{\Gamma(r)}\mathrm{w}^{r-1}\exp\{-\mathrm{w}\theta\}\theta^r d\mathrm{w} \\
&= (2\pi)^{-q/2}\frac{2}{f(\boldsymbol{y})}|\boldsymbol{\Sigma}|^{-1/2}\frac{\theta^r}{\Gamma(r)}\int_0^\infty \Phi\{\sqrt{\mathrm{w}}\boldsymbol{\eta}^\top(\boldsymbol{y}-\boldsymbol{\xi})\}\mathrm{w}^{m+r+\frac{q}{2}-1}\exp(-\mathrm{w}\theta_1)d\mathrm{w} \\
&= C(\theta_1,r_1)\int_0^\infty \Phi\{\sqrt{\mathrm{w}}\boldsymbol{\eta}^\top(\boldsymbol{y}-\boldsymbol{\xi})\}h(\mathrm{w};\theta_1,r_1)d\mathrm{w} \\
&= C(\theta_1,r_1)T_1\left(\boldsymbol{\eta}^\top(\boldsymbol{y}-\boldsymbol{\xi})\sqrt{\frac{r_1}{\theta_1}};2r_1\right),
\end{aligned}
$$

and

$$E\{UW^m|Y\}$$

$$= \frac{2}{f(\boldsymbol{y})} \int_0^\infty \int_0^\infty \mathrm{u}\mathrm{w}^m \phi\left\{\mathrm{u} - \sqrt{\mathrm{w}}\boldsymbol{\eta}^\top(\boldsymbol{y} - \boldsymbol{\xi})\right\} \phi_q(\boldsymbol{y}; \boldsymbol{\xi}, \mathrm{w}^{-1}\boldsymbol{\Sigma})h(\mathrm{w}; \nu/2, \nu/2)d\mathrm{u}d\mathrm{w}$$

$$= \frac{2}{f(\boldsymbol{y})} \int_0^\infty \mathrm{w}^m \phi_q(\boldsymbol{y}; \boldsymbol{\xi}, \mathrm{w}^{-1}\boldsymbol{\Sigma})h(\mathrm{w}; \nu/2, \nu/2) \int_0^\infty \mathrm{u}\phi\left\{\mathrm{u} - \sqrt{\mathrm{w}}\boldsymbol{\eta}^\top(\boldsymbol{y} - \boldsymbol{\xi})\right\} d\mathrm{u}d\mathrm{w}$$

$$= \frac{2}{f(\boldsymbol{y})} \int_0^\infty \mathrm{w}^m \phi_q(\boldsymbol{y}; \boldsymbol{\xi}, \mathrm{w}^{-1}\boldsymbol{\Sigma})h(\mathrm{w}; \nu/2, \nu/2)\left[\sqrt{\mathrm{w}}\boldsymbol{\eta}^\top(\boldsymbol{y} - \boldsymbol{\xi})\Phi\{\sqrt{\mathrm{w}}\boldsymbol{\eta}^\top(\boldsymbol{y} - \boldsymbol{\xi})\}\right.$$

$$\left. + \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\mathrm{w}(\boldsymbol{y} - \boldsymbol{\xi})^\top\boldsymbol{\eta}\boldsymbol{\eta}^\top(\boldsymbol{y} - \boldsymbol{\xi})\right\}\right]d\mathrm{w}$$

$$= \boldsymbol{\eta}^\top(\boldsymbol{y} - \boldsymbol{\xi}) \cdot E\left\{W^{m+\frac{1}{2}}|Y\right\} + \frac{1}{\sqrt{2\pi}}C(\theta_2, r_1).$$

Appendix B. Computing the Marginal Distribution of U in (A.5)

For subsequent use, define $\mathsf{V} = \boldsymbol{\sigma}^{-1}(\mathsf{Z} - \boldsymbol{\xi})$ and let $\boldsymbol{\Sigma}_\mathbf{v} = \boldsymbol{\sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\sigma}^{-1}$. Using the standard methods for transformations of random variables, the density of U at the point $\mathrm{u} > 0$ is

$$f(\mathrm{u}) = \int_{-\infty}^\infty 2\phi_q(\mathbf{z}; \boldsymbol{\xi}, \boldsymbol{\Sigma})\phi\left\{\mathrm{u} - \boldsymbol{\alpha}^\top\boldsymbol{\sigma}^{-1}(\mathbf{z} - \boldsymbol{\xi})\right\} d\mathbf{z}$$

$$= 2(2\pi)^{-\frac{q+1}{2}}|\boldsymbol{\sigma}||\boldsymbol{\Sigma}|^{-\frac{1}{2}}\int_{-\infty}^\infty \exp\left\{-\frac{1}{2}\left[\mathbf{v}^\top\boldsymbol{\Sigma}_\mathbf{v}^{-1}\mathbf{v} + (\mathrm{u} - \boldsymbol{\alpha}^\top\mathbf{v})^2\right]\right\} d\mathbf{v}$$

$$= 2(2\pi)^{-\frac{q+1}{2}}|\boldsymbol{\Sigma}_\mathbf{v}|^{-\frac{1}{2}}\int_{-\infty}^\infty \exp\left\{-\frac{1}{2}\left[\mathbf{v}^\top(\boldsymbol{\Sigma}_\mathbf{v}^{-1} + \boldsymbol{\alpha}\boldsymbol{\alpha}^\top)\mathbf{v} - 2\mathrm{u}\boldsymbol{\alpha}^\top\mathbf{v} + \mathrm{u}^2\right]\right\} d\mathbf{v}$$

$$= 2(2\pi)^{-\frac{1}{2}}|\mathbf{A}|^{\frac{1}{2}}|\boldsymbol{\Sigma}_\mathbf{v}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(1 - \boldsymbol{\alpha}^\top\mathbf{A}\boldsymbol{\alpha})\mathrm{u}^2\right\}$$

$$\times \int_{-\infty}^\infty (2\pi)^{-\frac{q}{2}}|\mathbf{A}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{v} - \mathrm{u}\mathbf{A}\boldsymbol{\alpha})^\top\mathbf{A}^{-1}(\mathbf{v} - \mathrm{u}\mathbf{A}\boldsymbol{\alpha})\right\} d\mathbf{v}$$

$$= 2(2\pi)^{-\frac{1}{2}}|\mathbf{A}|^{\frac{1}{2}}|\boldsymbol{\Sigma}_\mathbf{v}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(1 - \boldsymbol{\alpha}^\top\mathbf{A}\boldsymbol{\alpha})\mathrm{u}^2\right\}, \tag{B.1}$$

where $\mathbf{A} = (\boldsymbol{\Sigma}_\mathbf{v}^{-1} + \boldsymbol{\alpha}\boldsymbol{\alpha}^\top)^{-1}$. Recall the binomial inverse theorem which states

$$(\mathbf{C} + \mathbf{U}\mathbf{B}\boldsymbol{V})^{-1} = \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{U}\mathbf{B}(\mathbf{B} + \mathbf{B}\boldsymbol{V}\mathbf{C}^{-1}\mathbf{U}\mathbf{B})^{-1}\mathbf{B}\boldsymbol{V}\mathbf{C}^{-1},$$

provided $\mathbf{C}$ and $\mathbf{B}+\mathbf{B}\boldsymbol{V}\mathbf{C}^{-1}\mathbf{U}\mathbf{B}$ are nonsingular and the matrix determinant lemma which states

$$|\mathbf{C} + \mathbf{u}\mathbf{v}^\top| = (1 + \mathbf{v}^\top \mathbf{C}^{-1}\mathbf{u})|\mathbf{C}|,$$

for any nonsingular matrix $\mathbf{C}$; (see Mardia et al., 1979, pp. 457-459). Using these results, we obtain

$$\mathbf{A} = (\boldsymbol{\Sigma}_\mathsf{v}^{-1} + \boldsymbol{\alpha}\boldsymbol{\alpha}^\top)^{-1} = \boldsymbol{\Sigma}_\mathsf{v} - \frac{\boldsymbol{\Sigma}_\mathsf{v}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_\mathsf{v}}{1 + \boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_\mathsf{v}\boldsymbol{\alpha}}.$$

Therefore,

$$1 - \boldsymbol{\alpha}^\top\mathbf{A}\boldsymbol{\alpha} = (1 + \boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_\mathsf{v}\boldsymbol{\alpha})^{-1} = (1 + \boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta})^{-1}, \tag{B.2}$$

$$|\mathbf{A}|^{\frac{1}{2}}|\boldsymbol{\Sigma}_\mathsf{v}|^{-\frac{1}{2}} = (1 + \boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_\mathsf{v}\boldsymbol{\alpha})^{-\frac{1}{2}} = (1 + \boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta})^{-\frac{1}{2}}. \tag{B.3}$$

Replacing (B.2) and (B.3) in (B.1), we obtain the density function in (A.5).

Appendix C. Computation of (11)

Expanding $Q(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}; \boldsymbol{\Theta}^{(m)})$ in (8) and ignoring the terms unrelated to $\mathbf{B}$ yields

$$
\begin{aligned}
Q(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}; \boldsymbol{\Theta}^{(m)}) \quad \propto \quad & \frac{1}{n}\mathrm{tr}\{(\boldsymbol{Y} - \mathbf{XB})\boldsymbol{\Omega}(\boldsymbol{Y} - \mathbf{XB})^\top\} \\
& + \frac{1}{n}\mathrm{tr}\{(\boldsymbol{Y} - \mathbf{XB})\boldsymbol{\eta}\boldsymbol{\eta}^\top(\boldsymbol{Y} - \mathbf{XB})^\top\} + \lambda_2 \sum_{i,j}|b_{ij}| \\
& - \frac{1}{n}\mathrm{tr}\left\{(\boldsymbol{Y} - \mathbf{XB})\boldsymbol{\eta}\left(\mathbf{U}^{(m)}\right)^\top + \mathbf{U}^{(m)}\boldsymbol{\eta}^\top(\boldsymbol{Y} - \mathbf{XB})^\top\right\} \\
= \quad & \frac{1}{n}\mathrm{tr}\{(\boldsymbol{Y} - \mathbf{XB})(\boldsymbol{\Omega} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)(\boldsymbol{Y} - \mathbf{XB})^\top \\
& - 2(\boldsymbol{Y} - \mathbf{XB})\boldsymbol{\eta}\left(\mathbf{U}^{(m)}\right)^\top\} + \lambda_2 \sum_{i,j}|b_{ij}|.
\end{aligned}
$$

Completing the square for $(\boldsymbol{Y} - \mathbf{XB})$, ignoring the terms not involving $\mathbf{B}$ and replacing $\boldsymbol{\Omega}$, $\boldsymbol{\eta}$ with $\boldsymbol{\Omega}^{(m+1)}$, $\boldsymbol{\eta}^{(m+1)}$, one can get the formula in (11).

26

Appendix D. Computation of the complete-data sufficient statistics for the likelihood in (12)

Expanding $L_c(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \nu)$ and ignoring the constants yields

$$
\begin{aligned}
L_c(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \nu) \;\propto\; & \frac{1}{n}\mathrm{tr}\left\{(\boldsymbol{\Omega} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)(\boldsymbol{Y} - \mathbf{X}\mathbf{B})^\top \boldsymbol{W}(\boldsymbol{Y} - \mathbf{X}\mathbf{B})\right\} - \log|\boldsymbol{\Omega}| \\
& - \frac{2}{n}\mathbf{U}^\top\sqrt{\boldsymbol{W}}(\boldsymbol{Y} - \mathbf{X}\mathbf{B})\boldsymbol{\eta} + g(\nu) \\
\propto\; & \frac{1}{n}\mathrm{tr}\left\{(\boldsymbol{\Omega} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)\boldsymbol{Y}^\top\boldsymbol{W}\boldsymbol{Y}\right\} + \frac{1}{n}\mathrm{tr}\left\{\mathbf{B}(\boldsymbol{\Omega} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)\mathbf{B}^\top\mathbf{X}^\top\boldsymbol{W}\mathbf{X}\right\} \\
& - \frac{2}{n}\mathrm{tr}\left\{\mathbf{B}(\boldsymbol{\Omega} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)\boldsymbol{Y}^\top\boldsymbol{W}\mathbf{X}\right\} - \log|\boldsymbol{\Omega}| - \frac{2}{n}\mathbf{U}^\top\sqrt{\boldsymbol{W}}\boldsymbol{Y}\boldsymbol{\eta} \\
& + \frac{2}{n}\mathbf{U}^\top\sqrt{\boldsymbol{W}}\mathbf{X}\mathbf{B}\boldsymbol{\eta} + g(\nu).
\end{aligned}
$$

Thus, the statistics in (14) are sufficient.

Appendix E. Computation of $Q(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\nu}; \boldsymbol{\Theta}^{(m)})$ in (16)

Taking the expectation of $L_p(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \nu)$ in (15) conditional on $\boldsymbol{Y}, \mathbf{X}, \boldsymbol{\Theta}^{(m)}$ and plugging the associated values in $Q(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \nu; \boldsymbol{\Theta}^{(m)})$, we obtain

$$
\begin{aligned}
Q_c(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \nu; \boldsymbol{\Theta}^{(m)}) \propto\; & \frac{1}{n}\sum_{i=1}^n b_i^{(m)}(\boldsymbol{y}_i - \mathbf{B}^\top\boldsymbol{x}_i)^\top\boldsymbol{\Omega}(\boldsymbol{y}_i - \mathbf{B}^\top\boldsymbol{x}_i) - \log|\boldsymbol{\Omega}| + g^{(m)}(\nu) \\
& + \frac{1}{n}\sum_{i=1}^n\left\{\frac{c_i^{(m)}}{\sqrt{b_i^{(m)}}} - \sqrt{b_i^{(m)}}\boldsymbol{\eta}^\top(\boldsymbol{y}_i - \mathbf{B}^\top\boldsymbol{x}_i)\right\}^2 + \lambda_1\sum_{i\neq j}|\omega_{ij}| + \lambda_2\sum_{i,j}|b_{ij}| \\
& - \frac{1}{n}\sum_{i=1}^n\frac{\left(c_i^{(m)}\right)^2}{b_i^{(m)}} + \frac{1}{n}\sum_{i=1}^n E(\mathrm{u}_i^2|\mathbf{X}, \boldsymbol{Y}, \hat{\boldsymbol{\Theta}}^{(m)}) \\
\propto\; & \frac{1}{n}\sum_{i=1}^n\left(\tilde{\boldsymbol{y}}_i^{(m)} - \mathbf{B}^\top\tilde{\boldsymbol{x}}_i^{(m)}\right)^\top\boldsymbol{\Omega}\left(\tilde{\boldsymbol{y}}_i^{(m)} - \mathbf{B}^\top\tilde{\boldsymbol{x}}_i^{(m)}\right) - \log|\boldsymbol{\Omega}| + g^{(m)}(\nu) \\
& + \frac{1}{n}\sum_{i=1}^n\left\{\tilde{\mathrm{u}}_i^{(m)} - \boldsymbol{\eta}^\top\left(\tilde{\boldsymbol{y}}_i^{(m)} - \mathbf{B}^\top\tilde{\boldsymbol{x}}_i^{(m)}\right)\right\}^2 + \lambda_1\sum_{i\neq j}|\omega_{ij}| + \lambda_2\sum_{i,j}|b_{ij}| \\
=\; & \mathrm{tr}\left\{\boldsymbol{\Omega}\tilde{\boldsymbol{S}}^{(m)}\right\} - \log|\boldsymbol{\Omega}| + \frac{1}{n}\left\|\tilde{\mathbf{U}}^{(m)} - \left(\tilde{\boldsymbol{Y}}^{(m)} - \tilde{\mathbf{X}}^{(m)}\mathbf{B}\right)\boldsymbol{\eta}\right\|^2 + g^{(m)}(\nu) \\
& + \lambda_1\sum_{i\neq j}|\omega_{ij}| + \lambda_2\sum_{i,j}|b_{ij}|. \tag{E.1}
\end{aligned}
$$

References

Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis, Third Edition. Wiley, New York.

Azzalini, A., 2005. The skew-normal distribution and related multivariate families. Scandinavian Journal of Statistics 32, 159–188.

Azzalini, A., Arellano-Valle, R.B., 2013. Maximum penalized likelihood estimation for skew-normal and skew-t distributions. Journal of Statistical Planning and Inference 143, 419 – 433.

Azzalini, A., Capitanio, A., 1999. Statistical applications of the multivariate skew normal distribution. Journal of the Royal Statistical Society Series B 61, 579–602.

Azzalini, A., Capitanio, A., 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. Journal of the Royal Statistical Society Series B 65, 367–389.

Azzalini, A., Dalla-Valle, A., 1996. The multivariate skew-normal distribution. Biometrika 83, 715–726.

Bedrick, E., Tsai, C., 1994. Model selection for multivariate regression in small samples. Biometrics 50, 226–231.

Bickel, P., Levina, E., 2008. Regularized estimation of large covariance matrices. The Annals of Statistics 36, 199–227.

Branco, M.D., Dey, D.K., 2001. A general class of multivariate skew elliptical distributions. Journal of Multivariate Analysis 79, 99–113.

Cai, T.T., Li, H., Liu, W., Xie, J., 2012. Covariate-adjusted precision matrix estimation with an application in genetical genomics. Biometrika .

Carvalho, C., Chang, J., Lucas, J., Nevins, J.R., Wang, Q., West, M., 2005. High-dimensional sparse factor models and latent factor regression. Journal of the American Statistical Association 103, 1438–1456.

Chen, L., 2012. Topics on Regularization of Parameters in Multivariate Linear Regression. Ph.D. thesis. Department of Statistics, Texas A&M University.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B 39(1), 1–38.

Diongue, A.K., Guegan, D., Vignal, B., 2009. Forecasting electricity spot market prices with a k-factor gigarch process. Applied Energy 86, 505–510.

Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9, 432–441.

Genton, M., 2004. Skew-elliptical Distributions and Their Applications: A Journey Beyond Normality. Chapman & Hall/CRC, Boca Raton.

Hill, M.A., Dixon, W.J., 1982. Roubstness in real life: A study of clinical laboratory data. Biometrics 38, 377–396.

Huang, J., Liu, N., Pourahmadi, M., Liu, L., 2006. Covariance matrix selection and estimation via penalized normal likelihood. Biometrika 93, 85–98.

Lachos, V.H., Ghosh, P., Arellano-Valle, R.B., 2010. Likelihood based inference for skew-normal independent linear mixed models. Statistica Sinica 20, 303–322.

Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88, 365–411.

Lee, W., Liu, Y., 2012. Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. J. Multivar. Anal. 111, 241–255.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, London, UK.

Mazumder, R., Hastie, T., 2012. The graphical lasso: New insights and alternatives. Electronic Journal of Statistics 6, 2125–2149.

Meng, X.L., Rubin, D.B., 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80, 267–278.

Panagiotelis, A., Smith, M., 2008. Bayesian density forecasting of intraday electricity prices using multivariate skew t distributions. International Journal of Forecasting 24, 710–727.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.Y., Pollack, J.R., Wang, P., 2010. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. The Annals of Applied Statistics 4(1), 53–77.

Piessen, R., deDoncker Kapenga, E., Uberhueber, C.W., Kahaner, D.K., 1983. QUADPACK, A Subroutine Package for Automatic Integration. Springer, Berlin.

Pourahmadi, M., 2011. Covariance estimation: The glm and regularization perspectives. Stat. Sciences 26, 369–387.

Reinsel, G., Velu, R.P., 1998. Reduced-Rank Regression: Theory and Applications. Springer: Lecture Notes in Statistics, 136, New York.

Rothman, A., Bickel, P., Levina, E., Zhu, J., 2008. Sparse permutation invariant covariance estimation. Electronic Journal of Staistics 2, 494–515.

Rothman, A., Levina, E., Zhu, J., 2010. Sparse multivariate regression with covariance estimation. Journal of Computational and Graphical Statistics 19, 947–962.

Sahu, S.K., Dey, D.K., Branco, M.D., 2003. A new class of multivariate skew distributions with applications to bayesian regression models. Canadian Journal of Statistics 31, 129–150.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B 58, 267–288.

Vinzi, V., Chin, W.W., Henseler, J., Wang, H., 2010. Handbook of Partial Least Squares. Springer, New York.

Warton, D., 2008. Penalized normal likelihood and ridge regularization of correlation and covariance matrices. Journal of the American Statistical Association 103, 340–349.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of Royal Statistical Society Serie B 68(1), 49–67.

Yuan, M., Lin, Y., 2007. Model selection and estimation in the gaussian graphical model. Biometrika 94, 19–35.