TOPICS ON REGULARIZATION OF PARAMETERS IN MULTIVARIATE LINEAR REGRESSION

A Dissertation

by

LIANFU CHEN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2011

Major Subject: Statistics

TOPICS ON REGULARIZATION OF PARAMETERS IN MULTIVARIATE LINEAR REGRESSION

A Dissertation

by

LIANFU CHEN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Mohsen Pourahmadi
Committee Members,	Daren B.H. Cline
	Suhasini Subba Rao
	Joel Zinn
Head of Department,	Simon J. Sheather

December 2011

Major Subject: Statistics

ABSTRACT

Topics on Regularization of Parameters in Multivariate Linear Regression.

(December 2011)

Lianfu Chen, B.S., University of Science & Technology of China;

M.S., Texas A&M University

Chair of Advisory Committee: Dr. Mohsen Pourahmadi

My dissertation mainly focuses on the regularization of parameters in the multivariate linear regression under different assumptions on the distribution of the errors. It consists of two topics where we develop iterative procedures to construct sparse estimators for both the regression coefficient and scale matrices simultaneously, and a third topic where we develop a method for testing if the skewness parameter in the skew-normal distribution is parallel to one of the eigenvectors of the scale matrix.

In the first project, we propose a robust procedure for constructing a sparse estimator of a multivariate regression coefficient matrix that accounts for the correlations of the response variables. Robustness to outliers is achieved using heavy-tailed t distributions for the multivariate response, and shrinkage is introduced by adding to the negative log-likelihood ℓ_1 penalties on the entries of both the regression coefficient matrix and the precision matrix of the responses. Taking advantage of the hierarchical representation of a multivariate t distribution as the scale mixture of normal distributions and the EM algorithm, the optimization problem is solved iteratively where at each EM iteration suitably modified *multivariate regression with covariance estimation* (MRCE) algorithms proposed by Rothman, Levina and Zhu are used. We propose two new optimization algorithms for the penalized likelihood, called MRCEI and MRCEII, which differ from MRCE in the way that the tuning parameters for the two matrices are selected. Estimating the degrees of freedom when penalizing the en-

tries of the matrices presents new computational challenges. A simulation study and real data analysis demonstrate that the MRCEII, which selects the tuning parameter of the precision matrix of the multiple response using the C_p criterion, generally does the best among all methods considered in terms of the prediction error, and MRCEI outperforms the MRCE methods when the regression coefficient matrix is less sparse.

The second project is motivated by the existence of the skewness in the data for which the symmetric distribution assumption on the errors does not hold. We extend the procedure we have proposed to the case where the errors in the multivariate linear regression follow a multivariate skew-normal or skew-t distribution. Based on the convenient representation of skew-normal and skew-t as well as the EM algorithm, we develop an optimization algorithm, called MRST, to iteratively minimize the negative penalized log-likelihood. We also carry out a simulation study to assess the performance of the method and illustrate its application with one real data example.

In the third project, we discuss the asymptotic distributions of the eigenvalues and eigenvectors for the MLE of the scale matrix in a multivariate skew-normal distribution. We propose a statistic for testing whether the skewness vector is proportional to one of the eigenvectors of the scale matrix based on the likelihood ratio. Under the alternative, the likelihood is maximized numerically with two different ways of parametrization for the scale matrix: Modified Cholesky Decomposition (MCD) and Givens Angle. We conduct a simulation study and show that the statistic obtained using Givens Angle parametrization performs well and is more reliable than that obtained using MCD.

To my parents

ACKNOWLEDGMENTS

I would like to gratefully and sincerely thank my advisor, Dr. Mohsen Pourahmadi, for his excellent guidance, understanding, patience and providing me with an excellent atmosphere for doing research during my graduate studies at Texas A&M University. You opened the door for me to this wonderful statistical world and helped me to build a balanced knowledge in both theory and methodology. The experience to work with you is something that I will be proud of and cherish for the rest of my life. Without your help, I would never have accomplished as much as I have achieved.

I would like to thank Dr. Daren B.H. Cline, Dr. Suhasini Subba Rao and Dr. Joel Zinn for their valuable discussions and serving on my committee. A special thanks is owed to my master advisor, Dr. Ruzong Fan, for his encouragement and consistent support.

I would like to thank my parents for their faith in me and allowing me to be as ambitious as I wanted. It was under their watchful eyes that I gained so much drive and ability to tackle challenges head on.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

In Statistics, particularly in the fields of machine learning and inverse problems, regularization involves introducing additional information about the parameters in order to solve an ill-posed problem or prevent overfitting. The information usually takes the form of a penalty for complexity such as bounds on the vector norm of the parameters. From the Bayesian point of view, regularization corresponds to imposing prior distributions on the parameters. In this dissertation, we consider the regularization of parameters in the context of multivariate linear regression where the ℓ_1-norm of the parameters is adopted as the penalty.

1.1 Multivariate Linear Regression

The *multivariate* linear regression is concerned with regressing simultaneously several response variables on the same set of predictor variables. It is commonly used in chemometrics, econometrics, biological and social sciences [1], [2, chap.6] and in the analysis of longitudinal and panel data [3, chap.10]. Specifically, let $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iq})^T$ be a q-dimensional response vector and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ be the p predictors for the ith unit. Then, the multivariate linear regression of $\boldsymbol{y}_i$ on the covariates $\boldsymbol{x}_i$ is of the form

$$\boldsymbol{y}_i = \mathbf{B}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \cdots, n \tag{1.1}$$

where $\mathbf{B}$ is the $p \times q$ regression coefficient matrix and the errors $\boldsymbol{\epsilon}_i$ of dimension q are independent of each other. Let $\mathbf{X}$ be the $n \times p$ predictor matrix with $\boldsymbol{x}_i^T$ in its

---

The journal model is *IEEE Transactions on Automatic Control.*

ith row, $\boldsymbol{Y}$ be the $n \times q$ response matrix with $\boldsymbol{y}_i^T$ in its ith row and $\boldsymbol{E}$ be the $n \times q$ random error matrix with ϵ_i^T in its ith row. Writing the regression model (1.1) into the matrix form yields the following general linear model:

$$\boldsymbol{Y} = \mathbf{X}\mathbf{B} + \boldsymbol{E}. \tag{1.2}$$

As an example, consider a biochemical data which contain chemical measurements on several characteristics of $n = 33$ individual samples of men's urine specimens. There are $q = 5$ response variables: pigment creatinine, concentrations of phosphate, phosphorus, creatinine and choline. The goal was to relate these responses to $p = 3$ predictors: the weight of the subject, volume and specific gravity. Postulating a multivariate linear regression seems to be a good starting point to analyze the data; see [4] for a recent analysis of the data and suitability of the linearity assumption.

In the multivariate regression, the errors in (1.1) are usually assumed to be independent with mean 0 and covariance matrix $\mathbf{\Sigma}$. Then, the parameters $\mathbf{B}$ and $\mathbf{\Sigma}$ can be simply estimated by the ordinary least square estimate and the sample covariance matrix of the residuals, respectively, i.e.,

$$\hat{\mathbf{B}}^{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y} \qquad S = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T \tag{1.3}$$

which are the same as their maximum likelihood estimates (MLEs) when $\boldsymbol{\epsilon}_i \sim N(0, \mathbf{\Sigma})$ [5]. However, there are some drawbacks for the estimators in (1.3):

(a) The estimators are equivalent to regressing each response on the predictors variables *separately* [6], so that the estimates may perform suboptimally since they do not utilize the information that the responses are correlated. It is also the case that this type of estimate performs poorly in the presence of outliers, highly correlated response/predictor variables.

(b) For high-dimensional data, particularly when p and q are larger than n, the regression coefficient matrix $\mathbf{B}$ can not be estimated using the above formula, since $\mathbf{X}$ is not of full column rank. Furthermore, it is known that in this case the sample covariance matrix is a highly unstable estimator of $\mathbf{\Sigma}$ [7], [8].

In these situations, the traditional estimators for $\mathbf{B}$ and $\mathbf{\Sigma}$ with pq and $q(q+1)/2$ parameters, respectively, have rather poor performances and are not suitable for prediction and other purposes, so that one must seek workable alternatives based on the idea of regularizing these parameters. Historically, this has been done either individually focusing on $\mathbf{B}/\mathbf{\Sigma}$ alone or simultaneously, depending on whether the dependence between the multivariate responses is ignored or not. We briefly review some of these developments in the next three sections.

1.2 Estimating $\mathbf{B}$ While Ignoring Correlations

A way to fix some of the pitfalls of the ordinary least squares estimator is to reduce its pq parameters in the regression coefficient matrix $\mathbf{B}$. This can be done either through dimension-reduction techniques such as reduced-rank regression [9], [10], [11], criterion-based model selection methods [12], [13], [14], Bayesian model selection [15], [16], principal components, partial least squares [17], [18] and linear factor regression [19], [20].

Another approach reduces the number of parameters through regularization which may force some entries of $\mathbf{B}$ towards zero; see [4] for a review. This approach can be unified and viewed as estimating $\mathbf{B}$ by solving the following constraint optimization problem:

$$\hat{\mathbf{B}} = \arg\min_{\mathbf{B}} \left\{ \mathrm{tr}\left[(\boldsymbol{Y} - \mathbf{XB})^T (\boldsymbol{Y} - \mathbf{XB}) \right] \right\} \quad \text{subject to:} \quad C(\mathbf{B}) \leq t, \tag{1.4}$$

where $C(\mathbf{B})$ is a scalar function of $\mathbf{B}$.

Of course, different constraints will lead to different estimates for $\mathbf{B}$. An early and natural constraint is $C(\mathbf{B}) = \sum_{j,k} b_{jk}^2$ so that (1.4) 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 estimate of $\mathbf{B}$ proposed by [21]. Using the Lagrangian form, this optimization problem takes the form

$$\hat{\mathbf{B}} = \arg\min_{\mathbf{B}} \left\{ \text{tr}\left[(\boldsymbol{Y} - \mathbf{XB})^T(\boldsymbol{Y} - \mathbf{XB})\right] + \lambda \sum_{j,k} |b_{jk}| \right\}. \tag{1.5}$$

Also, one may assign different weights to different parameters or use the adaptive Lasso [22] which amounts to setting $C(\mathbf{B}) = \sum_{j,k} w_{jk}|b_{jk}|$, where $w'_{jk}s$ are chosen adaptively using the data. Some other forms of the constraint function $C(\mathbf{B})$ which seem to make a compromise between the Lasso and the ridge regression are: the Bridge regression [23] taking $C(\mathbf{B}) = \sum_{j,k} |b_{jk}|^\gamma$ where $1 \leq \gamma \leq 2$; the elastic-net [24] with $C(\mathbf{B}) = \alpha \sum_{j,k} |b_{jk}| + \frac{(1-\alpha)}{2} \sum_{j,k} b_{jk}^2$ for $\alpha \in [0, 1]$.

Group-wise penalty functions are perhaps more suitable for regularizing the multivariate regression parameters. The first example of its kind is the grouped lasso [4] with $C(\mathbf{B}) = \sum_{j=1}^{p} (b_{j1}^2 + \cdots + b_{jq}^2)^{0.5}$. One could also combine the ℓ_1 and ℓ_2 penalties to form the constraint function $C(\mathbf{B}) = \alpha C_1(\mathbf{B}) + (1 - \alpha)C_2(\mathbf{B})$ for $\alpha \in [0, 1]$ where $C_1(\mathbf{B}) = \sum_{j,k} |b|_{jk}$ and $C_2(\mathbf{B}) = \sum_{j=1}^{p} (b_{j1}^2 + \cdots + b_{jq}^2)^{0.5}$. The first constraint controls the overall sparsity of the coefficient matrix $\mathbf{B}$ and the second imposes a group-wise penalty on the rows of $\mathbf{B}$ which controls the number of predictors entering into the multivariate regression model [25].

We note that the constraints mentioned so far introduce sparsity only into the regression coefficient matrix $\mathbf{B}$ without accounting for the covariance structure of the multivariate responses. In other words, they ignore the $q(q + 1)/2$ parameters in $\boldsymbol{\Sigma}$ whose estimation is a problem of great interest in statistics on its own right.

1.3 Covariance Matrix Regularization

Covariance estimation is an important problem in many areas of statistics dealing with correlated data. It is well-known that the sample covariance matrix performs poorly when the number of variables is large relative to the sample size [7], [8]. 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 [8], [26]. Such a regularization ends up shrinking the eigenvalues of the sample covariance matrix, and provides more accurate and well-conditioned covariance estimators. Recently, fast alternative methods have been proposed to construct sparse estimates of the precision matrix by adding to the normal likelihood a lasso penalty on its off-diagonal entries [27], [28], [29], [30], [31]. Other approaches include thresholding [32], [33], SPLICE method [34] and SPACE method [35].

For (time-) ordered data, the regularization usually relies on the modified Cholesky decomposition of the precision matrix $\mathbf{\Sigma}^{-1}$. It is known that [36] the entries of the Cholesky factor are unconstrained and have interpretation as regression coefficients when a variable is regressed on its predecessors. [37] uses a nonparametric method to smooth the Cholesky factor of the inverse covariance along its subdiagonals, and [38], [39] regularize the precision matrix by applying a lasso and adaptive lasso penalty to the Cholesky factor, respectively. However, imposing sparsity on the Cholesky factor does not necessarily imply sparsity of the precision matrix and the sparsity structure in the Cholesky factor could be sensitive to the order of the response variables. Other approaches that require a sort of time-order on the variables

are tapering [40] and banding [41].

1.4 Estimating **B** While Accounting for Correlations

The aforementioned methods consider either the regularized estimation of the regression coefficient matrix or that of the covariance matrix. In these situations, the two matrices are usually estimated separately, and the covariance matrix does not contribute much to the prediction accuracy. To improve the predictive power, one must take advantage of the correlations among the multivariate response. However, research in this area is rather scarce and there are only a few papers devoted to this important area. The authors in [1] proposed the Curds and Whey (CW) method which predicts a multivariate response vector with $\tilde{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}^{OLS}\boldsymbol{M}$ where $\hat{\boldsymbol{Y}}^{OLS}$ is the ordinary least square prediction and $\boldsymbol{M}$ is a $q \times q$ shrinkage matrix estimated from the data in a manner which exploits the correlation in the responses. [26] relies on the idea of ridge regression, and the authors of [42] present a procedure called scout under the multivariate normal assumption on the response and the predictors, and apply regularization to the inverse covariance of the joint distribution.

Rothman et al.'s multivariate regression with covariance estimation (MRCE) method [43] seems to be the first bona fide regularization approach which constructs sparse estimates for both matrices simultaneously. They add two separate lasso penalties to the negative normal log-likelihood and minimize the ensuing objective function which, up to a constant, is proportional to

$$g(\mathbf{B}, \mathbf{\Omega}) = \text{tr}\left[\frac{1}{n}(\boldsymbol{Y} - \mathbf{XB})'(\boldsymbol{Y} - \mathbf{XB})\mathbf{\Omega}\right] - \log|\mathbf{\Omega}| + \lambda_1 \sum_{j' \neq j} |\omega_{j'j}| + \lambda_2 \sum_{j,k} |b_{jk}|, \quad (1.6)$$

where $\mathbf{\Omega} = (\omega_{jj'}) = \mathbf{\Sigma}^{-1}$ and λ_1, λ_2 are the two tuning parameters to be determined from the data.

1.5 Overview Structure

It is well known that the normality assumption is too restrictive as it suffers from the lack of robustness against departures from the normal distribution, particularly when data shows multi-modality and skewness. Therefore, in this dissertation, we assume that the errors in (1.1) have a more general distribution. Following [43], the objective is to construct sparse estimators for the regression coefficient matrix and the scale matrix simultaneously in this setup. In Chapter II, we extend the MRCE method to the case where the errors in (1.1) follow a multivariate t distribution for accommodating possible outliers. We construct sparse estimators for both regression coefficient and precision matrices simultaneously by minimizing the resulting penalized likelihood for which two algorithms are developed. We conduct a simulation study to assess the performance of the proposed method and illustrate its application with two real data analysis. In Chapter III, the MRCE is further extended to the cases where the errors have a skew distribution for accommodating for the skewness in the data. In Chapter IV, we focus on the direction of the skew vector and its connection with the principal components of a skew-normal variate. We study the asymptotic distributions for the MLEs of the eigenvalues and eigenvectors of the scale matrix. We also propose a statistic for testing if the skewness parameter is proportional to an0 eigenvector of the scale matrix. In Chapter V, I will discuss some possible extensions and my future work.

CHAPTER II

SPARSE MULTIVARIATE REGRESSION AND COVARIANCE ESTIMATION

2.1 Introduction

Compared to the classical data analysis where the errors in (1.2) are assumed to be normal, handling outliers seems to be a more important problem in the high-dimensional data setup that needs special attention, since in the high-dimensional spaces the data tends to be more sparse which implies that every observation can appear as an outlier. Furthermore, the notion of which observations are outliers typically varies between users and problem domains. Thus, the traditional approach of detection and removal of outliers is not a feasible option and the idea of robust data analysis might be more suitable alternative. For handling outliers in high dimensions, one could rely on variety of robust methods such as the M-estimators [44], but we use the family of multivariate t distributions for robust estimation of the regression parameters [45], [46]. This approach is of great practical interest since it allows accommodating possible outliers by suitably choosing the tail parameter or the degrees of freedom. An important advantage of this approach to robustness is its explicit statement of the probabilistic setting, leading to a clearer interpretation of the results compared to the less explicit, say, M-estimators. The need for robust procedures is also motivated by the fact that data from heavy-tailed distributions are bound to have some extreme observations, so that the assumption of normality may not be plausible or cannot cope with outliers. Important examples of such phenomenon occur in finance, economics, data network and risk analysis [47], [48]. In such cases, the multivariate t distribution would give a more robust inference and allows one to control aspects of the impact of outliers [46], [49].

In this project, our objective is to construct robust and sparse estimates for the regression coefficient matrix while discounting the outliers and accounting for the dependence structure of the responses simultaneously. To this end, we develop robust versions of the MRCE algorithms when the error vector ϵ_i in (1.1) follows a multivariate t distribution. This provides an extension of the MRCE method in [43] since the multivariate t distribution approaches the normal distribution as the degrees of freedom goes to infinity.

Using the hierarchical representation of a multivariate t distribution as the scale mixture of normal distributions and the EM algorithm, the optimization problem is solved iteratively where a central role is played by the MRCE algorithms proposed by [43]. We propose two new optimization algorithms for the penalized likelihood, called MRCEI and MRCEII, which differ in the way that the two tuning parameters for the two matrices are selected. Estimating the degrees of freedom when penalizing the entries of the two matrices presents new computational challenges. The simulation study and real data analysis demonstrate that the MRCEII, which selects the tuning parameter of the precision matrix of the multiple response using the C_p criterion, generally does the best among all methods considered in terms of the prediction error, and MRCEI outperforms the MRCE algorithms when the regression coefficient matrix is less sparse.

The remainder of this chapter is organized as follows. We introduce our methodology for estimating multivariate regression via penalized t-likelihood in Section 2.2, and present two MRCE-type algorithms to implement it. In Section 2.3, we conduct a simulation study and compare the performance of our method to the MRCE algorithms. In Section 2.4, we apply our methodology to the datasets of weekly log-returns of nine US stocks, and the electricity spot prices from Australia. A summary and discussion of the results are given in Section 2.5.

2.2 Parameter Estimation via Penalized t-likelihood

In this section, we extend the MRCE algorithms in [43] to the setting where the errors in the multivariate regression have a multivariate t distribution. We provide the details for joint estimation of the regression coefficient and precision matrices of a multivariate regression model using a penalized t-likelihood with unknown degrees of freedom.

2.2.1 The multivariate t distribution

A $q-$dimensional random vector $\boldsymbol{Y} = (Y_1, \cdots, Y_q)^T$ has a multivariate t distribution, denoted by $t_\nu(\mu, \boldsymbol{\Sigma})$, if its probability density function is

$$f(\boldsymbol{y}; \nu, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\Gamma\left(\frac{\nu+q}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{\frac{q}{2}}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \left[1 + \frac{(\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}-\boldsymbol{\mu})}{\nu}\right]^{-\frac{\nu+q}{2}}, \tag{2.1}$$

where $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and ν are called its location, scale matrix and degrees of freedom, respectively. The mean and covariance matrix of the multivariate t distribution are

$$E(\boldsymbol{Y}) = \boldsymbol{\mu} \quad \text{and} \quad \text{Cov}(\boldsymbol{Y}) = \frac{\nu}{\nu-2}\boldsymbol{\Sigma}. \tag{2.2}$$

where ν should be greater than two for the existence of the covariance matrix.

In this project, we rely extensively on the fact that a multivariate t distribution can be represented as a scale mixture of normals with the mixing variable having a Gamma distribution [46]. Specifically, our estimation procedure exploits its hierarchical representation that if

$$\boldsymbol{Y}|W = w \sim N\left(\boldsymbol{\mu}, \frac{1}{w}\boldsymbol{\Sigma}\right) \quad \text{and} \quad W \sim \text{Gamma}(\nu/2, \nu/2), \tag{2.3}$$

then, the marginal distribution of $\boldsymbol{Y}$ is the multivariate t distribution defined in (2.1) [50].

Unlike the estimates of the parameters of multivariate normal distribution which are vulnerable to the outliers, those of the multivariate t are robust and can handle the outliers or atypical observations, without the need to detecting or removing them. The degrees of freedom ν controls the kurtosis or heaviness of the tail of the distribution. When $\nu = 1$, the distribution corresponds to the q-variate Cauchy distribution which has heavy tails; when ν goes to infinity, the multivariate t distribution approaches the normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. See [46] for more discussions on the properties of multivariate t distributions and their roles in robust estimation in variety of situations including the multivariate regression.

2.2.2 The Penalized t-likelihood

We extend the model in (1.1) by assuming that the error $\boldsymbol{\epsilon}_i$ has a multivariate t distribution with mean $\boldsymbol{\mu} = \mathbf{0}$, degrees of freedom $\nu > 2$ and scale matrix $\boldsymbol{\Sigma}$. In the following, we also assume that the columns of $\mathbf{X}$ and $\boldsymbol{Y}$ are centered so that the intercept term can be omitted.

Given the covariate matrix $\mathbf{X}$ and the response matrix $\boldsymbol{Y}$, the negative log-likelihood is proportional to

$$\begin{aligned} L(\mathbf{B}, \boldsymbol{\Omega}, \nu) &= -2 \log \Gamma\left(\frac{\nu+q}{2}\right) + 2 \log \Gamma\left(\frac{\nu}{2}\right) + q \log \nu - \log |\boldsymbol{\Omega}| \\ &+ \frac{\nu+q}{n} \sum_{i=1}^{n} \log\left(1 + \frac{1}{\nu}(\boldsymbol{y}_i - \mathbf{B}^T \boldsymbol{x}_i)^T \boldsymbol{\Omega} (\boldsymbol{y}_i - \mathbf{B}^T \boldsymbol{x}_i)\right) \end{aligned} \tag{2.4}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is the inverse covariance or precision matrix. We add two ℓ_1 penalty terms on the entries of $\mathbf{B}$ and $\boldsymbol{\Omega}$ to the negative log-likelihood, and estimate both matrices simultaneously by minimizing the penalized log-likelihood:

$$g(\mathbf{B}, \boldsymbol{\Omega}, \nu) = L(\mathbf{B}, \boldsymbol{\Omega}, \nu) + \lambda_1 \sum_{j' \neq j} |\omega_{j'j}| + \lambda_2 \sum_{j,k} |b_{jk}|. \tag{2.5}$$

The lasso penalties on $\mathbf{B}$ and $\mathbf{\Omega}$ encourage sparsity in their estimates and hence can reduce the number of parameters. When the number of predictors is large, such a lasso penalty on the regression coefficient matrix would zero out the irrelevant or redundant predictors and could improve the prediction accuracy. Moreover, in the high-dimensional situations where the empirical sample covariance is singular like when $q > n$, the lasso penalty on the precision matrix forces the covariance estimate to be nonsingular and well-conditioned.

Compared to the MRCE algorithms [43], minimization of the penalized negative likelihood $g(\mathbf{B}, \mathbf{\Omega}, \nu)$ is expected to be more complicated. Note that unlike the normal error case, even when $\lambda_1 = \lambda_2 = 0$ the maximum likelihood estimates of $\mathbf{B}$ and $\mathbf{\Omega}$ do not have closed forms [46]. A fast method for optimization of lasso-type problems is the coordinate descent algorithm [51], but this cannot be applied directly to our problem since the objective function $g(\mathbf{B}, \mathbf{\Omega}, \nu)$ is not convex in either $\mathbf{B}$ or $\mathbf{\Omega}$.

In this section, we propose iterative methods to find the minimizer of the objective function through a sequence of estimators using an Expectation Conditional Maximization (ECM) algorithm [52].

2.2.3 Iterative Optimization Algorithms via ECM and MRCE

Using the conditional Gaussian representation of the multivariate t distribution in (2.3) and the EM algorithm [53], we solve the optimization problem in (2.5) via iterative applications of the MRCE algorithms [43].

2.2.3.1 The EM Algorithm and Penalized t-likelihood

The EM algorithm is an iterative procedure for finding the MLE's of the parameters in situations where the model depends on some missing or latent variables so that computing the MLE is not straightforward. The EM algorithm alternates between an

expectation (E) step and a maximization (M) step [53]. In the E-step, it computes the expectation of the log-likelihood by replacing the unobservables with their conditional expectations given the current estimates of the parameters and the data; in the M-step, it maximizes the expected log-likelihood calculated in the E-step.

We illustrate the EM algorithms by writing the multivariate t distribution as a scale mixture of normals. Let $W_1, W_2, \cdots, W_n$ be the missing variables such that

$$\boldsymbol{\epsilon}_i | W_i = w_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}/w_i), \tag{2.6}$$

are independent for $i = 1, \cdots, n$, and

$$W_1, W_2, \cdots, W_n \; i.i.d \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right). \tag{2.7}$$

We augment the data by including the latent variables $W_i's$ and treat $(\boldsymbol{y}_i, w_i), 1 \leq i \leq n$ as the complete data. Hence in this context, the original observations $\boldsymbol{y}_i's$ are regarded as being incomplete and (2.4) is the negative incomplete-data log-likelihood. The joint distribution of $(\boldsymbol{y}_i, w_i)$, $1 \leq i \leq n$, is called the complete-data likelihood and the negative penalized complete-data log-likelihood is proportional to

$$\begin{aligned} g_c(B, \boldsymbol{\Omega}, \nu) &= -\log|\boldsymbol{\Omega}| + \frac{1}{n}\sum_{i=1}^{n} w_i(\boldsymbol{y}_i - \mathbf{B}^T\boldsymbol{x}_i)^T\boldsymbol{\Omega}(\boldsymbol{y}_i - \mathbf{B}^T\boldsymbol{x}_i) + a(\nu) \\ &\quad + \lambda_1 \sum_{j' \neq j} |\omega_{j'j}| + \lambda_2 \sum_{j,k} |b_{jk}|, \end{aligned} \tag{2.8}$$

where

$$a(\nu) = 2\log\Gamma\left(\frac{\nu}{2}\right) - \nu\log\left(\frac{\nu}{2}\right) - \frac{1}{n}(\nu + q - 2)\sum_{j=1}^{n}\log w_j + \frac{\nu}{n}\sum_{j=1}^{n} w_j. \tag{2.9}$$

The optimization problem of $g(\mathbf{B}, \boldsymbol{\Omega}, \nu)$ in (2.5) can be solved by iteratively computing the minimizer of $g_c(\mathbf{B}, \boldsymbol{\Omega}, \nu)$ in (2.8) via an EM algorithm implemented as follows:

E-step: On the $(k+1)$th iteration, calculate the conditional expectation of

the negative penalized complete-data log-likelihood function in (2.8) given the observed data matrix $\boldsymbol{Y}$ and $\mathbf{X}$ with the current estimate of the parameters $\hat{\Theta}^{(k)} = (\hat{\mathbf{B}}^{(k)}, \hat{\boldsymbol{\Omega}}^{(k)}, \hat{\nu}^{(k)})$.

Since $g_c(\mathbf{B}, \boldsymbol{\Omega}, \nu)$ is linear in both w_i and $\log w_i$, the E-step amounts to simply replacing these by their corresponding conditional expectations $E(W_j|\boldsymbol{Y}, \mathbf{X}, \hat{\Theta}^{(k)})$ and $E(\log W_j|\boldsymbol{Y}, \mathbf{X}, \hat{\Theta}^{(k)})$. Recalling that the gamma distribution is the conjugate prior distribution for W_j, then it is not difficult to show that the conditional distribution of W_j given the current estimate $\hat{\Theta}^{(k)}$ and the data $(\mathbf{X}, \boldsymbol{Y})$ is also a Gamma distribution [52], namely,

$$W_j|\boldsymbol{Y}, \mathbf{X}, \hat{\Theta}^{(k)} \sim \text{Gamma}\left(\frac{\nu^{(k)} + q}{2}, \frac{\nu^{(k)} + \delta(\boldsymbol{y}_j, \boldsymbol{x}_j; \hat{\Theta}^{(k)})}{2}\right), \tag{2.10}$$

where

$$\delta(\boldsymbol{y}_j, \boldsymbol{x}_j; \hat{\Theta}^{(k)}) = \left[\boldsymbol{y}_j - (\hat{\mathbf{B}}^{(k)})^T \boldsymbol{x}_j\right]^T \hat{\boldsymbol{\Omega}}^{(k)} \left[\boldsymbol{y}_j - (\hat{\mathbf{B}}^{(k)})^T \boldsymbol{x}_j\right], \tag{2.11}$$

is the Mahalanobis distance between $\boldsymbol{y}_j$ and $(\hat{\mathbf{B}}^{(k)})^T \boldsymbol{x}_j$. Therefore, from (2.10), we have that

$$u_j^{(k)} = E(W_j|\boldsymbol{Y}, \mathbf{X}, \hat{\Theta}^{(k)}) = \frac{\hat{\nu}^{(k)} + q}{\hat{\nu}^{(k)} + \delta(\boldsymbol{y}_j, \boldsymbol{x}_j; \hat{\Theta}^{(k)})}. \tag{2.12}$$

To calculate the conditional expectation of $\log W_i$, we rely on the fact that if W has a Gamma(α, γ) distribution, then

$$E(\log W) = \psi(\alpha) + \log \gamma,$$

where $\psi(s) = [\partial \Gamma(s)/\partial s]/\Gamma(s)$ is the digamma function. Applying this result to the

E-step yields

$$E(\log W_j|\boldsymbol{Y},\mathbf{X},\hat{\Theta}^{(k)}) = \psi\left(\frac{\hat{\nu}^{(k)}+q}{2}\right)-\log\left(\frac{\hat{\nu}^{(k)}+\delta(\boldsymbol{y}_j,\boldsymbol{x}_j;\hat{\Theta}^{(k)})}{2}\right)$$
$$= \psi\left(\frac{\nu^{(k)}+q}{2}\right)-\log\left(\frac{\nu^{(k)}+q}{2}\right)+\log\left(u_j^{(k)}\right). \quad (2.13)$$

See [52] for more details on computing these conditional expectations.

M-step: When all the latent variables are known, the regularization problem in (2.8) is similar to that considered in [43], except for optimization with respect to ν. However, since the minimization of (2.8) over the whole parameter space is challenging, we replace the M-step with a few Conditional-Maximization (CM) steps listed below.

CM1: Since the degrees of freedom ν is separated from the other parameters, we update it numerically by

$$\hat{\nu}^{(k+1)} = \arg\min_{\nu}\{a(\nu)\}. \quad (2.14)$$

CM2: Given $\mathbf{B}=\hat{\mathbf{B}}^{(k)}$, solving the optimization problem for $\boldsymbol{\Omega}$ in (2.8) is equivalent to computing

$$\hat{\boldsymbol{\Omega}}^{(k+1)} = \arg\min_{\boldsymbol{\Omega}}\left\{-\log|\boldsymbol{\Omega}|+\mathrm{tr}\{\boldsymbol{\Omega}S^{(k)}\}+\ \lambda_1\sum_{j\neq j'}|\omega_{jj'}|\right\}, \quad (2.15)$$

where $\boldsymbol{S}^{(k)}=\frac{1}{n}\sum_{i=1}^n w_i\left[\boldsymbol{y}_i-(\hat{\mathbf{B}}^{(k)})^T\boldsymbol{x}_i\right]\left[\boldsymbol{y}_i-(\hat{\mathbf{B}}^{(k)})^T\boldsymbol{x}_i\right]^T$. This is the ℓ_1 penalized covariance estimation problem considered in [27], [29], [30] and [31]. We use the fast graphical lasso algorithm in [29] to solve (2.15).

CM3: Given $\boldsymbol{\Omega}=\hat{\boldsymbol{\Omega}}^{(k+1)}$, finding the minimizer of $g_c(\mathbf{B},\boldsymbol{\Omega},\nu)$ with respect to $\mathbf{B}$ is equivalent to minimizing

$$\tilde{g}(\mathbf{B}) = \frac{1}{n}\sum_{i=1}^n w_i(\boldsymbol{y}_i-\mathbf{B}^T\boldsymbol{x}_i)^T\hat{\boldsymbol{\Omega}}^{(k+1)}(\boldsymbol{y}_i-\mathbf{B}^T\boldsymbol{x}_i)+\lambda_2\sum_{j,k}|b_{jk}|, \quad (2.16)$$

which can be solved using a lasso-type algorithm described next.

Define a long vector $\boldsymbol{\beta}$ of length pq as $\boldsymbol{\beta} = (b_{11}, b_{21}, \cdots, b_{p1}, \cdots, b_{1q}, b_{2q}, \cdots, b_{pq})^T$ and $\mathbf{X}_i = I_{q\times q} \otimes \boldsymbol{x}_i^T$, where $'\otimes'$ is the Kronecker product and $I_{q\times q}$ is the identity matrix. Consider the Cholesky decomposition of $\boldsymbol{\Omega}$ as $\boldsymbol{\Omega} = L^T L$, where L is a $q \times q$ upper triangular matrix. Let $\tilde{\boldsymbol{y}} = \frac{1}{\sqrt{n}}(\sqrt{w_1}\boldsymbol{y}_1^T\boldsymbol{L}^T, \sqrt{w_2}\boldsymbol{y}_2^T\boldsymbol{L}^T, \cdots, \sqrt{w_n}\boldsymbol{y}_n^T\boldsymbol{L}^T)^T$ which is of length qn and $\tilde{\mathbf{X}} = \frac{1}{\sqrt{n}}(\sqrt{w_1}\mathbf{X}_1^T L^T, \cdots, \sqrt{w_n}\mathbf{X}_n^T\boldsymbol{L}^T)^T$. Then, (2.16) can be rewritten more compactly as

$$\tilde{g}(\boldsymbol{\beta}) = \|\tilde{\boldsymbol{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|^2 + \lambda_2 \sum_{j=1}^{pq} |\beta_j|. \tag{2.17}$$

This is a quadratic minimization problem subject to a linear constraint on the parameters which is exactly the lasso problem. There are efficient algorithms for solving this problem for all values of λ; see the homotopy algorithm of [54] and the Lars-lasso algorithm of [55]. Another simpler algorithm for solving this problem for a fixed λ is the coordinate descent algorithm. This algorithm finds the minimizer of (2.17), say $\tilde{\boldsymbol{\beta}}$, by updating each of its coordinates $\tilde{\beta}_j, j = 1, \cdots, pq$, given the others, using

$$\tilde{\beta}_j = T\left(\sum_{i=1}^{nq} \tilde{x}_{ij}(\tilde{y}_i - \tilde{y}_i^{(j)}), 2\lambda_2\right),$$

where $\tilde{\mathbf{X}} = (\tilde{x}_{ij})$, $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \cdots, \tilde{y}_{nq})^T$, $\tilde{y}_i^{(j)} = \sum_{k\neq j} \tilde{x}_{ij}\tilde{\beta}_k$ and $T(x, \lambda) = \text{sgn}(x)(|x| - \lambda)_+$. Then it cycles through all $\tilde{\beta}_j's$ until convergence.

2.2.3.2 Two MRCE Algorithms with t-errors

In this section, first we summarize the EM algorithm for minimizing(2.8) and refer to it as the MRCEI algorithm. We use the coordinate descent algorithm to solve the lasso regression problem in (2.17). As in [43], $\sum_{j,k} |\hat{b}_{jk}^{ridge}|$ is used to scale the test of convergence in the MRCEI algorithm, where $\hat{\mathbf{B}}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{Y}$, and ϵ is

the tolerance parameter , set at 10^{-4} by default.

MRCEI algorithm: With λ_1 and λ_2 fixed, initialize the parameters $\Theta = \Theta^{(0)}$. On the (k+1)th iteration,

E-step: Estimate the latent variables W_i and $\log W_i$ by their conditional expectations as in (2.12) and (2.13).

CM1: Estimate $\nu = \hat{\nu}^{(k+1)}$ by numerically minimizing the $a(\nu)$ in (3.3).

CM2: Update $\mathbf{\Omega} = \hat{\mathbf{\Omega}}^{(k+1)}$ in (2.15) using the graphical lasso algorithm.

CM3: Update $\mathbf{B} = \hat{\mathbf{B}}^{(k+1)}$ in (2.17) using the coordinate descent algorithm.

Repeat the E- and CM-steps until the estimates of the parameters converge, that is, $\sum_{j,k} |\hat{b}_{jk}^{(k+1)} - \hat{b}_{jk}^{(k)}| \leq \epsilon \sum_{jk} |\hat{b}_{jk}^{ridge}|$.

The MRCEI is an iterative version of the MRCE method of [43], in the sense that it repeats CM2 and CM3 steps until convergence. Compared with the MRCE method, MRCEI is expected to take longer time to converge due to the iterations in the EM algorithm. This means that, just like the MRCE method, applying MRCEI to high dimensional data would be computationally expensive or intractable. In practice, even for smaller p and q, hundreds of iterations for some values of (λ_1, λ_2) might be needed for the MRCEI algorithm to converge.

As discussed in Section 2.2.4 below, the tuning parameters λ_1 and λ_2 in MRCEI would be selected via K-fold cross-validation over a grid of values of (λ_1, λ_2). To reduce the computational cost for choosing the two tuning parameters, we make two modifications in the above algorithm and propose the faster MRCEII algorithm. The key and primary modification is to keep λ_1 fixed and λ_2 variable. The secondary modification is to replace the coordinate descent algorithm in the CM3 step by the Lars-lasso algorithm.

MRCEII algorithm: For a fixed value of λ_1, initialize the parameters $\Theta = \Theta^{(0)}$. On the (k+1)th iteration,

E-step: Estimate the latent variables W_i and $\log W_i$ by their conditional expectations as in (2.12) and (2.13).

CM1: Estimate $\nu = \hat{\nu}^{(k+1)}$ by numerically minimizing $a(\nu)$ in (3.3).

CM2: Update $\boldsymbol{\Omega} = \hat{\boldsymbol{\Omega}}^{(k+1)}$ in (2.15) using the graphical lasso algorithm.

CM3: Update $B = \hat{B}^{(k+1)}$ and the value of λ_2 in (2.17) using the Lars-lasso algorithm. Repeat the E- and CM-steps until the estimates of the parameters converge, that is, $\sum_{j,k} |\hat{b}_{jk}^{(k+1)} - \hat{b}_{jk}^{(k)}| \leq \epsilon \sum_{jk} |\tilde{b}_{jk}^{ridge}|$, where $\tilde{\mathbf{B}}^{ridge} = (\mathbf{X}'\mathbf{X} + \lambda_1 \mathbf{I})^{-1}\mathbf{X}'\boldsymbol{Y}$.

In MRCEII, for each value of λ_1, an estimate of B with a corresponding value of λ_2 will be obtained in the CM3 step. When choosing the tuning parameter, one has only to consider a few selected values of λ_1, rather than a grid of values of (λ_1, λ_2). This results in a great reduction of the computational cost so far as iterations are concerned.

2.2.4 Tuning Parameters Selection

For the MRCEI, we consider a grid of values of (λ_1, λ_2) and choose the tuning parameters (λ_1, λ_2) via K-fold cross-validation as the minimizer of an unbiased estimate of the expected prediction error variance described next.

To start, we randomly split the full dataset $S = \{(\boldsymbol{x}_i, \boldsymbol{y}_i), i = 1, 2, \cdots, n\}$ into K subsets of about the same size, denoted by S_k, $k = 1, 2, \cdots, K$. For each k, we use $S - S_k$ as the training set to estimate the parameters and S_k as the test set to validate. Then, we select the tuning parameters (λ_1, λ_2) that minimizes the criterion of mean squared prediction error over all q variables of the response, that is,

$$(\hat{\lambda}_1, \hat{\lambda}_2) = \arg \min_{(\lambda_1, \lambda_2)} \frac{1}{Kq} \left\{ \sum_{k=1}^{K} \| \boldsymbol{Y}_{(k)} - \mathbf{X}_{(k)} \hat{\mathbf{B}}_{(-k)}^{\lambda_1, \lambda_2} \|_{L_2}^2 \right\}, \tag{2.18}$$

where $\boldsymbol{Y}_{(k)}, \mathbf{X}_{(k)}$ are the validation response matrix and the predictor matrix formed by the subset S_k, respectively, and $\hat{\mathbf{B}}_{(-k)}^{\lambda_1, \lambda_2}$ is the corresponding estimate of $\mathbf{B}$ using

MRCEI for the training data $S - S_k$.

For the MRCEII, we randomly partition the full dataset S into two subsets, the training set S_1 and the validation set S_2, and then select the tuning parameters in two steps. In the first step, for each value of λ_1, we follow Efron et al. (2004, p. 17) and simply choose λ_2 by the C_p criterion using the training data. That is, λ_2 is chosen as the minimizer of the function

$$\lambda_2 = \arg\min_{\lambda_2 > 0} \left\{ \frac{\text{RSS}}{\hat{\sigma}^2} - n + 2d \right\}, \tag{2.19}$$

where d is the number of nonzero elements in the estimate of $\boldsymbol{\beta}$, RSS is the residual sum of squares of model (2.17) and $\hat{\sigma}^2$ is the corresponding estimated variance of the model. At the second step, for each pair of (λ_1, λ_2) obtained in the first step, we select the one with minimum mean prediction error over all q responses:

$$(\hat{\lambda}_1, \hat{\lambda}_2) = \arg\min_{(\lambda_1, \lambda_2)} \frac{1}{q} \left\{ \|\boldsymbol{Y}^* - \mathbf{X}^* \hat{\mathbf{B}}^{\lambda_1, \lambda_2}\|_{L_2}^2 \right\}, \tag{2.20}$$

where $\hat{\mathbf{B}}^{\lambda_1, \lambda_2}$ is the sparse estimate of $\mathbf{B}$ in (2.17) using the Lars-lasso algorithm for the training set, and $\boldsymbol{Y}^*$, $\mathbf{X}^*$ are the validation matrices for the responses and predictors, respectively.

In the simulation study and the real data analysis, we select λ_1 from some predefined set Λ for both MRCEI and II, and λ_2 from the same set Λ for MRCEI.

2.2.5 Estimation of The Degrees of Freedom

If the degrees of freedom is known, the CM1 step in the algorithms can be ignored. Otherwise, one should update the estimate of ν via the CM1 step at each iteration. However, in our simulations we have noticed that the estimated sequence $\{\hat{\nu}^{(k)}\}$ using the EM algorithm usually decrease monotonically towards a small positive number less than 2 which is not compatible with the existence of the covariance matrix of a

multivariate t distribution. This phenomenon which is mostly due to the monotonicity of the likelihood function in the EM algorithm [53] is explained in more details next.

Taking the derivative of $a(\nu)$ with respect to ν, the optimization problem in (3.3) is equivalent to solving the equation

$$\psi\left(\frac{\nu}{2}\right) - \log\left(\frac{\nu}{2}\right) + \frac{1}{n}\sum_{j=1}^{n}\left(u_j^{(k)} - \log(u_j^{(k)})\right) - \psi\left(\frac{\nu^{(k)}+q}{2}\right) + \log\left(\frac{\nu^{(k)}+q}{2}\right) = (2.21)$$

At each iteration, by the monotonicity of the likelihood function in the EM algorithm, the negative penalized likelihood function will decrease and some entries in the two matrices of parameters are forced to be zero. After a few warm-up iterations, for most j's the sequences formed by $\left\{\delta(\boldsymbol{y}_j, \boldsymbol{x}_j; \hat{\Theta}^{(k)})\right\}_{k\geq 1}$ will decrease. Consequently, the corresponding sequence $\left\{u_j^{(k)}\right\}_{k\geq 1}$ will increase and be greater than 1 which makes the third term in (2.21) to increase, since the function $f(x) = x - \log(x)$ is increasing for $x \geq 1$. As shown in [52], the function $h(x) = \psi(x) - \log(x)$ is strictly increasing over $(0, \infty)$, hence the sequence $\left\{\nu^{(k)}\right\}_{k\geq 1}$ obtained from (2.21) will decrease. Finally, the decrease in the third term of (2.21) due to the shrinkage of the parameters makes the estimated degrees of freedom to be a very small number.

Thus, to obtain feasible estimates of the degrees of freedom, in what follows we ignore the CM1 step in our algorithms, and estimate ν separately using a one-dimensional search. Estimation of the unknown degrees of freedom of the t distributions, in general, is an important problem and has been studied by many authors: [46], [52] and [56] consider estimation of ν in an EM framework, while [49] and [57] utilize method of moments estimators for ν.

2.3 A Simulation Study

2.3.1 Simulation Design and Models

In this section, we compare the performance of our two algorithms with the MRCE and the approximate MRCE (ap. MRCE) algorithms in [43] using a simulation study with a design similar to theirs.

Throughout this section we will have 50 replications, and in each replication a sparse matrix B is generated by the elementwise product of three matrices:

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

where $(\boldsymbol{W})_{ij} \sim i.i.d\, N(0,1), (\boldsymbol{K})_{ij} \sim i.i.d\, Bernoulli(s_1)$ 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 we expect each predictor to be relevant for $s_1 q$ of all the response variables. An $n \times p$ predictor matrix $\mathbf{X}$ with $n = 50$ is also generated with rows drawn independently from $N(0, \mathbf{\Sigma}_x)$, where $(\mathbf{\Sigma}_x)_{ij} = 0.7^{|i-j|}$, as in Yuan et al. (2007) and Peng et al. (2009b). We consider two models for the scale matrix of the errors as follows,

- AR(1) covariance model with $(\mathbf{\Sigma}_{\boldsymbol{E}})_{ij} = \rho_{\boldsymbol{E}}^{|i-j|}$ for $\rho_{\boldsymbol{E}} = 0, 0.5, 0.7$ and 0.9.
- Fractional Gaussian Noise (FGN) error covariance model with

$$(\mathbf{\Sigma}_{\boldsymbol{E}})_{ij} = 0.5[(|i-j|+1)^{2H} - 2|i-j|^{2H} + (|i-j|-1)^{2H}]$$

  for $H = 0.90$ and 0.95.

Then each row of the error matrix $\boldsymbol{E}$ is independently drawn from a multivariate t distribution $t_\nu(\mu, \mathbf{\Sigma}_{\boldsymbol{E}})$ 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 [43]. This is similar to performing a K-fold cross-validation as in (3.26) for the MRCEI.

2.3.2 Performance Measures

We measure the performance of various methods in terms of the model error as in [27] and [43]. For an estimate of regression coefficient matrix $\hat{\mathbf{B}}$, the model error is defined as

$$ME(\hat{\mathbf{B}}) = \mathrm{tr}\left\{(\hat{\mathbf{B}} - \mathbf{B})'\mathbf{\Sigma}_x(\hat{\mathbf{B}} - \mathbf{B})\right\}. \tag{2.22}$$

The sparsity recognition performance of $\hat{\mathbf{B}}$ is measured by the true positive rate (TPR) as well as the true negatvie rate (TNR) which are defined as

$$\begin{aligned} \mathrm{TPR}(\hat{\mathbf{B}}, \mathbf{B}) &= \frac{\#\{(i,j) : \hat{b}_{ij} \neq 0 \quad \text{and} \quad b_{ij} \neq 0\}}{\#\{(i,j) : b_{ij} \neq 0\}}, \\ \mathrm{TNR}(\hat{\mathbf{B}}, \mathbf{B}) &= \frac{\#\{(i,j) : \hat{b}_{ij} = 0 \quad \text{and} \quad b_{ij} = 0\}}{\#\{(i,j) : b_{ij} = 0\}}. \end{aligned} \tag{2.23}$$

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. One should consider them simultaneously since $\hat{\mathbf{B}} = 0$ always has perfect TNR and the OLS estimate always has perfect TPR.

2.3.3 Results and Discussions

For the AR(1) error covariance model, we consider different combinations of ν, ρ_E, s_1 and s_2 from the following ranges: (1) $\nu = 10, 20, 40, 100$, (2) $\rho_{\boldsymbol{E}} = 0, 0.5, 0.7, 0.9$, (3) $s_1 = 0.1, 0.5$, and (4) $s_2 = 1$; for the FGN model, we have the same design except that $\rho_{\boldsymbol{E}}$ is replaced by the corresponding FGN error covariance model with $H = 0.90$ and

0.95. Additionally, the tuning parameters for both error covariance models would be selected from the set $\Lambda = \{10^x : x = 0, \pm 1, \cdots, \pm 5\}$. Since the conclusions drawn from these two error models are similar, we only report the results for the AR(1) error covariance model here.

Table 1. Model error for the AR(1) error covariance models for $p = q = 20$, $s_1 = 0.1$ and $s_2 = 1$. Average and standard errors in parenthesis are based on 50 replications with $n = 50$. Tuning parameters were selected using a 10^x resolution.

	$\rho_{\boldsymbol{E}}$	OLS	MRCE	ap. MRCE	MRCEI	MRCEII
$\nu = 10$	0.9	16.20	1.01	1.25	1.04	0.74
		(0.62)	(0.03)	(0.10)	(0.02)	(0.01)
	0.7	15.92	2.21	2.30	2.18	1.92
		(0.41)	(0.06)	(0.07)	(0.06)	(0.01)
	0.5	15.38	2.99	3.08	2.92	2.83
		(0.31)	(0.08)	(0.09)	(0.07)	(0.09)
	0.0	15.51	3.60	3.74	3.49	3.32
		(0.30)	(0.09)	(0.09)	(0.08)	(0.10)
$\nu = 20$	0.9	15.57	0.96	1.05	1.06	0.70
		(0.76)	(0.04)	(0.40)	(0.05)	(0.03)
	0.7	16.30	2.04	2.22	2.12	1.76
		(0.47)	(0.05)	(0.07)	(0.06)	(0.06)
	0.5	16.00	2.78	2.96	2.85	2.61
		(0.32)	(0.06)	(0.08)	(0.08)	(0.08)
	0.0	15.47	3.30	3.50	3.38	3.10
		(0.28)	(0.07)	(0.08)	(0.08)	(0.09)
$\nu = 40$	0.9	16.46	0.91	1.03	1.02	0.73
		(0.56)	(0.05)	(0.04)	(0.06)	(0.03)
	0.7	15.74	1.96	2.02	2.14	1.82
		(0.39)	(0.06)	(0.06)	(0.06)	(0.07)
	0.5	15.48	2.71	2.67	2.91	2.64
		(0.39)	(0.07)	(0.07)	(0.07)	(0.08)
	0.0	15.70	3.25	3.31	3.53	3.18
		(0.32)	(0.08)	(0.07)	(0.08)	(0.09)
$\nu = 100$	0.9	16.44	0.88	0.91	0.90	0.68
		(0.69)	(0.04)	(0.03)	(0.04)	(0.02)
	0.7	15.47	1.87	1.93	1.95	1.75
		(0.44)	(0.05)	(0.06)	(0.05)	(0.06)
	0.5	16.29	2.57	2.57	2.68	2.50
		(0.33)	(0.06)	(0.07)	(0.07)	(0.08)
	0.0	15.49	3.10	3.07	3.29	3.01
		(0.30)	(0.06)	(0.08)	(0.08)	(0.08)

Table 2. Model error for the AR(1) error covariance models for $p = q = 20, s_1 = 0.5$ and $s_2 = 1$. Average and standard errors in parenthesis are based on 50 replications with $n = 50$. Tuning parameters were selected using a 10^x resolution.

	$\rho_{\boldsymbol{E}}$	OLS	MRCE	ap. MRCE	MRCEI	MRCEII
	0.9	15.33	4.78	6.76	4.80	4.48
		(0.57)	(0.22)	(0.29)	(0.26)	(0.16)
	0.7	15.71	9.70	10.35	9.47	7.77
$\nu = 10$		(0.40)	(0.29)	(0.31)	(0.33)	(0.19)
	0.5	16.22	12.48	12.02	12.01	9.55
		(0.34)	(0.26)	(0.31)	(0.35)	(0.21)
	0.0	15.15	13.14	12.81	12.61	10.20
		(0.24)	(0.31)	(0.32)	(0.32)	(0.22)
	0.9	15.05	4.39	5.73	4.07	4.13
		(0.53)	(0.16)	(0.20)	(0.16)	(0.14)
	0.7	16.19	8.68	9.03	8.52	7.32
$\nu = 20$		(0.41)	(0.26)	(0.23)	(0.24)	(0.17)
	0.5	16.00	10.86	10.67	10.83	8.82
		(0.39)	(0.25)	(0.20)	(0.24)	(0.16)
	0.0	15.85	11.34	11.21	11.55	9.51
		(0.36)	(0.26)	(0.23)	(0.26)	(0.14)
	0.9	15.35	4.11	5.74	4.11	4.19
		(0.53)	(0.15)	(0.22)	(0.14)	(0.10)
	0.7	15.35	8.66	9.11	8.28	7.30
$\nu = 40$		(0.51)	(0.25)	(0.25)	(0.24)	(0.18)
	0.5	16.12	11.17	10.61	10.29	8.83
		(0.38)	(0.25)	(0.24)	(0.25)	(0.20)
	0.0	16.17	11.74	11.07	10.88	9.56
		(0.35)	(0.30)	(0.24)	(0.22)	(0.21)
	0.9	15.21	4.49	5.46	3.93	3.94
		(0.47)	(0.23)	(0.23)	(0.13)	(0.14)
	0.7	15.56	8.41	8.38	7.97	6.95
$\nu = 100$		(0.37)	(0.21)	(0.23)	(0.21)	(0.17)
	0.5	15.33	10.50	9.87	10.22	8.54
		(0.31)	(0.20)	(0.21)	(0.21)	(0.18)
	0.0	16.00	10.76	10.15	10.74	9.26
		(0.33)	(0.21)	(0.19)	(0.21)	(0.15)

Tables 1 and 2 present the results of the simulation study for $p = q = 20$. We note that, with ν fixed, the model errors increase as $\rho_{\boldsymbol{E}}$ decreases , except for the OLS method. The OLS has by far the largest model errors, indeed, it does the worst among the methods considered. In addition, the MRCEII algorithm generally outperforms the other methods in terms of the model error. This seems to be mostly due to

the alternative method of selecting its tuning parameters. In the MRCEI algorithm, cross-validation is carried out over a grid of points of (λ_1, λ_2). Therefore, the selected tuning parameter $(\hat{\lambda}_1, \hat{\lambda}_2)$ is usually a vertex of the rectangle that contains the optimal value. In the MRCEII algorithm, we fix λ_1 at some pre-defined points and for each value of λ_1, λ_2 is selected using the C_p criterion. The tuning parameters selected in this way allow λ_2 to move on the edge of the rectangles so that $(\hat{\lambda}_1, \hat{\lambda}_2)$ is more likely to be closer to the optimal value in (2.20), leading to smaller model errors.

Table 3. True Positive Rate/True Negative Rate for the AR(1) error covariance models averaged over 50 replications; $n = 50, p = q = 20$, $s_1 = 0.1$ and $s_2 = 1$. Tuning parameters were selected using a 10^x resolution.

	$\rho_{\boldsymbol{E}}$	MRCE	ap. MRCE	MRCEI	MRCEII
	0.9	0.92/0.59	0.92/0.61	0.94/0.52	0.92/0.74
$\nu = 10$	0.7	0.87/0.63	0.88/0.64	0.88/0.59	0.85/0.74
	0.5	0.84/0.66	0.85/0.65	0.85/0.63	0.82/0.75
	0.0	0.82/0.68	0.83/0.66	0.85/0.64	0.81/0.76
	0.9	0.93/0.58	0.93/0.61	0.94/0.53	0.92/0.75
$\nu = 20$	0.7	0.90/0.61	0.89/0.62	0.89/0.60	0.86/0.76
	0.5	0.87/0.64	0.86/0.64	0.86/0.63	0.83/0.76
	0.0	0.85/0.65	0.84/0.66	0.84/0.63	0.80/0.77
	0.9	0.94/0.58	0.94/0.62	0.93/0.55	0.91/0.75
$\nu = 40$	0.7	0.90/0.61	0.87/0.63	0.88/0.60	0.86/0.74
	0.5	0.87/0.63	0.89/0.63	0.85/0.63	0.82/0.75
	0.0	0.84/0.64	0.85/0.65	0.83/0.64	0.79/0.78
	0.9	0.93/0.58	0.92/0.60	0.92/0.56	0.91/0.73
$\nu = 100$	0.7	0.89/0.61	0.87/0.63	0.88/0.60	0.87/0.75
	0.5	0.87/0.63	0.85/0.63	0.85/0.62	0.84/0.75
	0.0	0.85/0.64	0.82/0.65	0.83/0.63	0.82/0.77

When $s_1 = 0.5$, the MRCEI does better than the MRCE and ap.MRCE. In

particular, the MRCEI has comparable performance to MRCEII when $\nu = 20, 40$ or 100 with highly correlated errors ($\rho_{\boldsymbol{E}} = 0.9$). For the more sparse coefficient matrix $\mathbf{B}$ ($s_1 = 0.1$), when ν is small, the MRCEI tends to have smaller model errors than the MRCE and ap. MRCE. However, when ν becomes large, they outperform the MRCEI even though the fitted model is incorrect. This is because as ν goes to infinity the multivariate t approaches the normal distribution for which the MRCE and ap. MRCE have outstanding performance for a sparser $\mathbf{B}$.

Table 4. True Positive Rate/True Negative Rate for the AR(1) error covariance models averaged over 50 replications; $n = 50, p = q = 20$, $s_1 = 0.5$ and $s_2 = 1$. Tuning parameters were selected using a 10^x resolution.

	$\rho_{\boldsymbol{E}}$	MRCE	ap. MRCE	MRCEI	MRCEII
	0.9	0.92/0.41	0.90/0.43	0.92/0.42	0.93/0.38
$\nu = 10$	0.7	0.87/0.45	0.86/0.44	0.85/0.52	0.90/0.41
	0.5	0.87/0.37	0.84/0.46	0.83/0.49	0.88/0.43
	0.0	0.85/0.41	0.84/0.45	0.81/0.55	0.87/0.45
	0.9	0.93/0.41	0.91/0.45	0.92/0.42	0.93/0.38
$\nu = 20$	0.7	0.87/0.48	0.87/0.44	0.86/0.52	0.90/0.41
	0.5	0.86/0.41	0.87/0.42	0.84/0.49	0.89/0.42
	0.0	0.84/0.48	0.84/0.48	0.84/0.48	0.86/0.46
	0.9	0.93/0.43	0.90/0.49	0.92/0.42	0.94/0.36
$\nu = 40$	0.7	0.87/0.46	0.86/0.47	0.87/0.46	0.90/0.39
	0.5	0.87/0.39	0.85/0.47	0.84/0.49	0.88/0.43
	0.0	0.85/0.43	0.84/0.48	0.82/0.52	0.87/0.45
	0.9	0.93/0.37	0.91/0.42	0.93/0.42	0.94/0.37
$\nu = 100$	0.7	0.87/0.48	0.88/0.45	0.86/0.51	0.91/0.40
	0.5	0.85/0.49	0.85/0.48	0.85/0.50	0.89/0.41
	0.0	0.83/0.53	0.83/0.53	0.84/0.50	0.88/0.43

The corresponding true positive and negative rates for the AR(1) covariance

error model are also reported in Tables 3 and 4. We note that, with the degrees of freedom ν fixed, as $\rho_{\boldsymbol{E}}$ decreases, the true positive rates tend to decrease while the true negative rates tend to increase. Moreover, the MRCE methods and MRCEI have comparable true positive and negative rates, so the comparison among them should be based on the model errors. The MRCEII also has comparable true positive rates with the other methods, but its true negative rates are substantially greater when $\mathbf{B}$ is sparser. Along with the substantially smaller prediction errors, the MRCEII has an excellent performance when $\mathbf{B}$ is sparser. When the coefficient matrix $\mathbf{B}$ is not so sparse, MRCEII seems to be conservative in the sense that it gives a slightly less parsimonious estimate of $\mathbf{B}$ than other methods.

We report the average CPU times in Table 5 over 50 replications when $p = q = 20$, $s_1 = 0.5$, $\rho_E = 0.9$ and $s_2 = 1$ with ν varying from 10 to 100. All computations were carried out on a quad-core Intel Xeon 2.5 GHz processor with 10GB of RAM. The MRCEI algorithm is faster than the MRCEII for larger ν, because in this situation the MRCEI algorithm takes fewer EM iterations to converge.

Table 5. The average CPU times (in minutes) over 50 replications when $p = q = 20$, $s_1 = 0.5$, $\rho_E = 0.9$ and $s_2 = 1$.

ν	MRCE	ap. MRCE	MRCEI	MRCEII
10	0.32	15.63	20.61	25.12
20	0.29	15.66	3.75	20.29
40	0.27	15.65	1.86	16.79
100	0.22	15.64	1.18	13.92

2.4 Real Data Analysis

In this section, we illustrate our methods by applying them to two real financial datasets and compare the results with those using the two MRCE methods.

2.4.1 Predicting Asset Returns

The first real data example we consider is the weekly log-returns of stocks of 9 large American companies in 2004, which was also analyzed in [27] and [43]. Following their approaches, we fit a VAR(1) (vector autoregression of order 1) model to the data:

$$\boldsymbol{y}_t = \mathbf{B}^T \boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t, 1 \leq t \leq T, \tag{2.24}$$

where $\boldsymbol{y}_t$ is the vector of log-returns of the stocks in week t. Writing (2.24) into the matrix form as

$$\boldsymbol{Y}_T = \boldsymbol{Y}_{T-1}\mathbf{B} + \boldsymbol{E}, \tag{2.25}$$

where $\boldsymbol{Y}_T = (\boldsymbol{y}_2^T, \boldsymbol{y}_3^T, \cdots, \boldsymbol{y}_T^T)^T$ and $\boldsymbol{Y}_{T-1} = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, \cdots, \boldsymbol{y}_{T-1}^T)^T$ makes it a special case of the multivariate linear regression model (1.2). In [43], it is assumed that the error in (2.25) has a multivariate normal distribution and apply the MRCE method, but there is ample empirical evidence in the finance literature that the asset returns often exhibit heavy-tails.

We model the asset returns data using the multivariate t distribution which has proved successful in handling the heavy-tailed data in such applications [58], [59]. The MLE of the degrees of freedom is $\hat{\nu} = 10.15$ using the log-returns data for the whole year. The other parameters in the model are estimated using the log-returns of the stocks for the first half of the year (T=26) as the training set, and the rest

as the test set. The tuning parameters are selected from the set $\Lambda = \{2^x : x = -25, -24, \cdots, 10\}$. For the MRCEI algorithm, we select the tuning parameters via a 10-fold cross-validation; for the MRCEII, we use the last 10% of the log-returns of the first half year as the validation data and the remaining 90% as the training data.

The estimated coefficient matrix $\hat{\mathbf{B}}$ using MRCEI turns out to be zero or a fully sparse estimate, compared with the MRCE and ap. MRCE estimates which have 4/81 and 12/81 nonzero coefficients, respectively. However, the MRCEII estimate of $\mathbf{B}$ reported in Table 6, has 19 nonzero entries, and there are 22 zeros in the estimate of $\mathbf{\Omega}$. The MRCEII, MRCE and ap. MRCE have four common nonzero entries at the positions $(1, 7), (4, 1), (4, 2)$ and $(4, 8)$ of $\mathbf{B}$. This suggests, for example, the log-returns for Walmart at week t-1 as a relevant predictor for the Citigroup at week t, and the log-returns for Ford at week t-1 as a relevant predictor of Walmart, Exxon and GM at week t.

We evaluate the predictive performance by the average squared prediction error for each company over the data from the second half of the year, the result is reported in Table 7. Except for OLS, other methods have comparable performance in terms of the prediction error, though still the MRCEII is slightly better than the other methods. This finding is consistent with the results of our simulation study. In addition, the MRCEI estimating a null model for the data indicates that the signal from the predictors in this example is relatively weak.

2.4.2 Intraday Electricity Prices

Next, we apply our method to the hourly average electricity spot prices collected in the Australian state of New South Wales (NSW) from July 2, 2003 to June 30, 2006, starting at 04:00 and ending at 03:00 each day. The dataset consists of 26352 observations during a period of $T = 1098$ days and was previously analyzed in [60] using

Table 6. Estimated coefficient matrix **B** using MRCEII.

	Wal	Exx	GM	Ford	GE	CPhi	Citi	IBM	AIG
Wal	0	0	0	0	0	0	0.2289	0.2287	0
Exx	0	0	0	0	0	0	0	-0.1168	0
GM	0	0	0.0237	0	0	0	0	0	-0.0574
Ford	-0.1639	0.0336	0	0	0.0092	0	0	-0.0834	0
GE	0	0	0	0	0	0.133	-0.0125	0.0662	0
CPhi	0	0.0505	0	0	0.0597	0	-0.0458	0	0
Citi	0	-0.0101	0.0923	0	0	0	0	0	0
IBM	0	0	0	0	0	0	0	0	0
AIG	0	0	0.0306	0	0	0	-0.0564	0	0

a Bayesian method and skew-t distribution for the data. Unlike other commodity prices, most electricity spot prices exhibit trend, strong periodicity, intra-day and inter-day serial correlations, heavy tails, skewness and so on; see [60], [61], [62], [63] for some empirical evidence. As in [60], we consider the vector of the log spot prices at hourly intervals during a day as the response vector with $q = 24$. The exogenous 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.

Instead of assuming that the covariate effects are the same at all hours within a day as in [60], we fit a multivariate regression model to the log electricity prices by

Table 7. Average squared prediction error for each company $\times 10^3$ based on 26 points. Standard errors are reported in parenthesis.

	OLS	MRCE	ap. MRCE	MRCEI	MRCEII
Walmart	0.98(0.27)	0.41(0.11)	0.41(0.11)	0.42(0.12)	0.42(0.11)
Exxon	0.39(0.08)	0.31(0.07)	0.31(0.07)	0.31(0.07)	0.32(0.07)
GM	1.68(0.42)	0.71(0.17)	0.69(0.17)	0.71(0.17)	0.68(0.19)
Ford	2.15(0.61)	0.77(0.25)	0.77(0.25)	0.77(0.25)	0.77(0.25)
GE	0.58(0.15)	0.45(0.09)	0.45(0.09)	0.45(0.09)	0.46(0.09)
ConocoPhillips	0.98(0.24)	0.79(0.22)	0.78(0.22)	0.79(0.22)	0.77(0.21)
Citigroup	0.65(0.17)	0.62(0.13)	0.62(0.13)	0.66(0.14)	0.62(0.13)
IBM	0.62(0.14)	0.49(0.10)	0.47(0.09)	0.49(0.10)	0.43(0.09)
AIG	1.93(0.93)	1.88(1.02)	1.88(1.02)	1.88(1.02)	1.90(1.03)
AVE	1.14(0.14)	0.71(0.12)	0.71(0.12)	0.72(0.12)	0.71(0.13)

regressing the hourly observations during a day on the same covariates:

$$\boldsymbol{y}_i = \mathbf{B}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i \quad 1 \leq i \leq T \tag{2.26}$$

where $\boldsymbol{y}_i$ is a 24×1 vector of log electricity prices on day i and $\boldsymbol{x}_i$ is the corresponding vector of the covariates. We assume $\boldsymbol{\epsilon}_i \sim t_\nu(0, \boldsymbol{\Sigma})$.

The MLE of the degrees of freedom ν is $\hat{\nu} = 3.51$ using the whole dataset. With ν fixed at $\hat{\nu}$, we then apply our methods to the model (2.26). 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 other parameters using the rest of the observed spot prices. The set from which we choose the tuning parameters is $\Lambda = \left\{2^{-10+20(x-1)/39} : x = 1, 2, \cdots, 40\right\}$. For the MRCEI algorithm, we select the

tuning parameters via a 10-fold cross-validation; while for MRCEII algorithm, we use 90% of the observations in the first 998 days as the training data and the remaining 10% as the validation data.

Table 8. Average squared prediction error for each hour on a day based on 100 points. Standard errors are reported in parenthesis.

Time	OLS	MRCE	ap. MRCE	MRCEI	MRCEII
04:00	0.040(0.006)	0.039(0.006)	0.039(0.006)	0.043(0.006)	0.035(0.005)
05:00	0.042(0.005)	0.041(0.005)	0.041(0.005)	0.045(0.006)	0.036(0.005)
06:00	0.033(0.003)	0.033(0.003)	0.033(0.003)	0.034(0.003)	0.027(0.003)
07:00	0.072(0.008)	0.072(0.008)	0.072(0.008)	0.075(0.008)	0.057(0.006)
08:00	0.115(0.012)	0.114(0.012)	0.114(0.012)	0.122(0.012)	0.099(0.010)
09:00	0.142(0.017)	0.141(0.017)	0.141(0.017)	0.145(0.017)	0.123(0.017)
10:00	0.146(0.015)	0.145(0.015)	0.145(0.015)	0.152(0.014)	0.121(0.013)
11:00	0.132(0.012)	0.130(0.012)	0.129(0.012)	0.139(0.012)	0.104(0.010)
12:00	0.131(0.012)	0.128(0.011)	0.126(0.011)	0.143(0.012)	0.108(0.010)
13:00	0.108(0.010)	0.106(0.010)	0.103(0.009)	0.131(0.012)	0.097(0.009)
14:00	0.100(0.009)	0.095(0.009)	0.089(0.009)	0.136(0.012)	0.094(0.009)
15:00	0.096(0.010)	0.092(0.010)	0.083(0.009)	0.140(0.013)	0.090(0.009)
16:00	0.088(0.009)	0.081(0.008)	0.072(0.007)	0.126(0.012)	0.080(0.008)
17:00	0.129(0.013)	0.118(0.012)	0.107(0.010)	0.167(0.015)	0.107(0.010)
18:00	0.393(0.115)	0.398(0.117)	0.399(0.118)	0.432(0.112)	0.425(0.120)
19:00	0.270(0.036)	0.270(0.036)	0.270(0.037)	0.291(0.035)	0.270(0.038)
20:00	0.143(0.015)	0.142(0.015)	0.143(0.015)	0.154(0.015)	0.130(0.015)
21:00	0.076(0.008)	0.076(0.008)	0.076(0.008)	0.084(0.009)	0.065(0.007)
22:00	0.040(0.004)	0.040(0.004)	0.040(0.004)	0.043(0.004)	0.035(0.004)
23:00	0.104(0.010)	0.104(0.009)	0.104(0.009)	0.111(0.010)	0.091(0.009)
00:00	0.093(0.008)	0.092(0.008)	0.093(0.008)	0.099(0.009)	0.081(0.007)
01:00	0.064(0.005)	0.063(0.005)	0.063(0.005)	0.068(0.006)	0.057(0.005)
02:00	0.060(0.005)	0.060(0.005)	0.060(0.005)	0.063(0.005)	0.052(0.004)
03:00	0.021(0.002)	0.021(0.002)	0.021(0.002)	0.023(0.003)	0.018(0.002)
AVE	0.110(0.006)	0.108(0.006)	0.107(0.006)	0.124(0.006)	0.100(0.006)

The average squared prediction errors based on the observations in the last 100 days are reported in Table 8, where the results using the MRCE methods are also included for comparison. We see that the ordinary least square estimate performs

better in this example, and MRCEII still does the best with the smallest overall prediction error. In addition, MRCEII has the smallest individual mean prediction errors at most of the times except at the times between 17:00 and 19:00 which is the evening hours and highly volatile. Moreover, MRCEI does not perform well in this case, it always has the largest prediction errors.

Table 9. Proportions of zeros in the estimate of the parameters

	MRCE	ap. MRCE	MRCEI	MRCEII
$\mathbf{B}$	74/528	70/528	32/528	120/528
$\mathbf{\Omega}$	484/576	506/576	296/576	312/576

The proportions of zeros in the estimated regression coefficient matrix as well as the regularized inverse covariance matrix are presented in Table 9. We see that the estimated coefficient matrices for all methods are fairly sparse, implying that most of the covariates do have impact on the hourly spot price. In addition, MRCEII (I) has the most (fewest) zero entries in the estimated $\mathbf{B}$, while the MRCE and ap. MRCE give more sparse estimates of the inverse covariance. We report the positions of nonzero entries in the estimated inverse covariance in Figure 1, we note that both MRCE methods give block diagonal estimates for the inverse covariance matrix with nonzero entries concentrated in the middle of the matrix corresponding to the evening hours. Due to the intraday serial correlations, we expect more correlations in the precision matrix, so the estimates of $\mathbf{\Omega}$ using the MRCE methods might be too simple to capture the conditional dependency structure of the electricity prices. For MRCEI and II, most of the nonzero entries in the precision matrix are confined to a diagonal band and the others roughly lie in the upper right (lower) corner. This can be interpreted as the model trying to relate a spot price to several preceding spot

prices and similar prices from the same time in the previous day, the kind of model that was postulated in [60] on *ad hoc* basis.

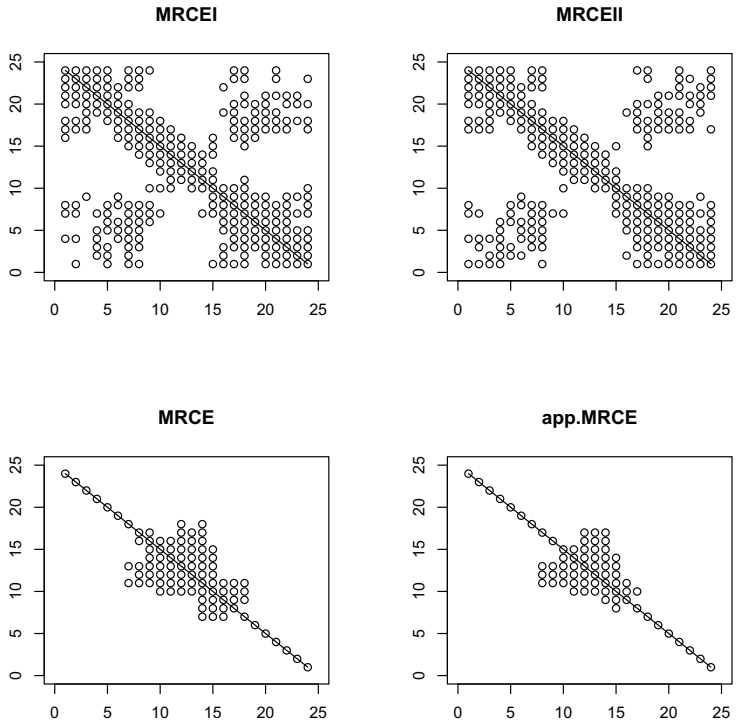

Fig. 1. Positions of nonzero entries in $\hat{\mathbf{\Omega}}$ for different methods applied to the electricity prices; The straight line indicates the diagonal of $\hat{\mathbf{\Omega}}$.

2.5 Summary

We have proposed a procedure to construct robust and sparse estimates for the regression coefficient matrix and the inverse covariance matrix in the context of multivariate regression with multivariate t distributed errors. This assumption on the errors enables us: (i) to handle the outliers and thus give robust estimates of the parameters without identifying and removing the outliers, (ii) to embed the recent MRCE algorithms [43] within the EM iterations to optimize a complicated, nonstandard objective

function . The two optimization algorithms MRCEI and II iteratively compute the estimates of the parameters for moderate size p, q, and as pointed out in Section 2.2.5, the unknown degrees of freedom is estimated first and outside these iterations. We have shown that MRCEII outperforms all the competing methods in terms of the prediction error and MRCEI outperforms the MRCE and ap. MRCE when the regression coefficient matrix is less sparse. More empirical and theoretical work remains to be done to compare these methods and to improve the numerical efficiency and speed of the MRCEI and II algorithms for larger p, q like 100 or so. At present, these methods including the MRCE are computationally intractable for large numbers of responses and covariates [43, Section 3.4]. Motivated by the skewness and high correlations in the Australian electricity spot prices, we are currently studying extensions of MRCE-type algorithms to multivariate skew-normal and skew-t distributions.

CHAPTER III

REGULARIZATION OF MULTIVARIATE REGRESSION WITH SKEW ERRORS

3.1 Introduction

Both the multivariate normal and t distributions are symmetric about the mean, but, in practice, the normality assumption is usually violated because of the presence of skewness and kurtosis in real data [64], so one may seek more flexible parametric families of multivariate distributions to represent these features of the data as adequately as possible. Among them, the family of the skew-normal distributions which generalize the multivariate normal distributions with an extra parameter to regulate skewness has been widely adopted due to its mathematical tractability and appealing probabilistic properties similar to those of the normal distributions [65], [66] and [67]. A further extension of the skew-normal distribution is the multivariate skew-t distribution [68] which allows for both nonzero skewness and heavy tails in the distribution. Some of the probabilistic properties of the skew-t distributions as well as the applications were investigatedin [69]. For the general background on the skew-normal and other skew distributions, see [70] and the survey papers [67], [71].

In this project, we assume that the errors $\boldsymbol{\epsilon}_i's$ have a multivariate skew-normal distribution and consider the "small n, large p and q" problem. Since the MLEs of the regression coefficient matrix $\mathbf{B}$ and the scale matrix $\boldsymbol{\Sigma}$ of the errors perform poorly when p and q are large relative to n, it is prudent to regularize the two matrices jointly. Following [43] and [72], we construct sparse estimators simultaneously for both the regression coefficient matrix and the inverse scale matrix by adding ℓ_1 penalties to the negative log-likelihood on the entries of these matrices. Taking advantage of

the stochastic representation of the skew-normal distribution and the ensuing latent variables, we develop an EM-type algorithm, called MRSN, to iteratively optimize the resulting penalized likelihood function. Then we extend our method to the case where the errors follow a multivariate skew-t distribution [69], and a similar algorithm called MRST is also developed.

Our approach relies on and is closely related to the recent work in [43] in which sparse estimators for both $\mathbf{B}$ and $\mathbf{\Sigma}$ are constructed simultaneously by minimizing the penalized normal log-likelihood:

$$\begin{aligned} g(\mathbf{B},\mathbf{\Omega}) &= \operatorname{tr}\left[\frac{1}{n}(\boldsymbol{Y}-\mathbf{XB})^T(\boldsymbol{Y}-\mathbf{XB})\mathbf{\Omega}\right]-\log|\mathbf{\Omega}| \\ &\quad +\lambda_1\sum_{j'\neq j}|\omega_{j'j}|+\lambda_2\sum_{j,k}|b_{jk}|, \end{aligned} \tag{3.1}$$

where $\mathbf{\Omega}=(\omega_{jj'})=\mathbf{\Sigma}^{-1}$ and λ_1,λ_2 are the two tuning parameters to be determined from the data. For example, when $\boldsymbol{\epsilon}_i$ has a multivariate skew-t distribution, our EM algorithm leads to minimizing the penalized version of the negative complete-data log-likelihood: (See Section 3.3 for more details)

$$\begin{aligned} L_c(\mathbf{B},\mathbf{\Omega},\boldsymbol{\eta},\nu) &= \frac{1}{n}\sum_{i=1}^n w_i(\boldsymbol{y}_i-\mathbf{B}^T\boldsymbol{x}_i)^T\mathbf{\Omega}(\boldsymbol{y}_i-\mathbf{B}^T\boldsymbol{x}_i)-\log|\mathbf{\Omega}| \\ &\quad +\frac{1}{n}\sum_{i=1}^n\left(z_i-\sqrt{w_i}\boldsymbol{\eta}^T(\boldsymbol{y}_i-\mathbf{B}^T\boldsymbol{x}_i)\right)^2+a(\nu), \end{aligned} \tag{3.2}$$

where $\mathbf{\Omega},\boldsymbol{\eta}$ are defined in Section 3.2.2, (w_i,z_i) are the two latent variables associated with $\boldsymbol{y}_i$ and

$$a(\nu)=2\log\Gamma\left(\frac{\nu}{2}\right)-\nu\log\left(\frac{\nu}{2}\right)-\frac{1}{n}(\nu+q-2)\sum_{j=1}^n\log w_j+\frac{\nu}{n}\sum_{j=1}^n w_j. \tag{3.3}$$

Compared with the normal log-likelihood in (3.1), $L_c(\mathbf{B},\mathbf{\Omega},\boldsymbol{\eta},\nu)$ has two additional terms involving the degrees of freedom ν and the skewness parameter $\boldsymbol{\eta}$. When the

skewness parameter is zero, $L_c(\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta}, \nu)$ reduces to the complete-data log-likelihood for a multivariate t distribution where regularizing $\mathbf{B}$ and $\mathbf{\Omega}$ is studied in [72].

The remainder of this chapter is organized as follows. Some basic properties of skew-normal and skew-t distributions are reviewed in Section 3.2. In Section 3.3, we introduce our methodology for estimating multivariate regression via penalized skew-normal and skew t likelihoods. The selection of tuning parameters is discussed in Section 3.4. We conduct a simulation study and investigate the performance of the method in terms of the prediction error (PE) in Section 3.5. In Section 3.6, we apply our methodology to the electricity wholesale spot prices in Australia.

3.2 Multivariate Skew-normal and -t Distributions

In this section, we briefly review the families of multivariate skew-normal and skew-t distributions [65], [68] as well as some of their properties that would be used in developing the EM-type algorithms.

3.2.1 The Multivariate Skew-normal Distribution

A random vector $\boldsymbol{Y}$ is said to have a q-variate skew-normal distribution if its probability density function takes the form

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \mathbf{\Sigma}, \boldsymbol{\alpha}) = 2\phi_q(\boldsymbol{y}; \boldsymbol{\mu}, \Sigma)\Phi\{\boldsymbol{\alpha}^T\omega^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\}, \tag{3.4}$$

where $\phi_q(\cdot; \boldsymbol{\mu}, \mathbf{\Sigma})$ is the *pdf* of the q-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$ and $\Phi(\cdot)$ 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 $\omega = \text{diag}\{\sigma_{11}^{\frac{1}{2}}, \cdots, \sigma_{qq}^{\frac{1}{2}}\}$ is a diagonal matrix equal to the square root of the diagonal elements of $\mathbf{\Sigma}$. We denote this

distribution by $\boldsymbol{y} \sim \mathrm{SN}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ where the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ 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 (3.4) are not symmetric, and have the mean and covariance matrix as

$$\boldsymbol{\mu}_{\boldsymbol{y}} = \boldsymbol{\mu} + \frac{(2/\pi)^{1/2}}{(1+\boldsymbol{\eta}^T\boldsymbol{\Sigma}\boldsymbol{\eta})^{1/2}}\boldsymbol{\Sigma}\boldsymbol{\eta}, \qquad \mathrm{Var}(\boldsymbol{y}) = \boldsymbol{\Sigma} - \boldsymbol{\mu}_{\boldsymbol{y}}\boldsymbol{\mu}_{\boldsymbol{y}}^T, \tag{3.5}$$

which are different from the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ where $\boldsymbol{\eta} = \omega^{-1}\boldsymbol{\alpha}$.

The family of multivariate skew-normal distributions can be obtained from the multivariate normal using a conditioning method [65]. Specifically, let $(V_0, V_1, \cdots, V_q)^T$ be a $(q+1)$-dimensional normal random vector with mean $\mathbf{0}$ and correlation matrix

$$R^* = \begin{pmatrix} 1 & \boldsymbol{\delta}^T \\ \boldsymbol{\delta} & R \end{pmatrix}.$$

where $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_p)^T$ and R is a correlation matrix. It can be shown that $\boldsymbol{V} = (V_1, \cdots, V_q)^T | V_0 > 0 \sim \mathrm{SN}_q(\mathbf{0}, R, \boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = (1 - \boldsymbol{\delta}^T R \boldsymbol{\delta})^{-1/2} R\boldsymbol{\delta}$, and the multivariate skew-normal family in (3.4) can be generated by the transformation $\boldsymbol{y} = \boldsymbol{\mu} + \omega \boldsymbol{V}$.

Note that the skew-normal density in (3.4) can be expressed as the form of the integral, i.e.,

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) = 2 \int_0^\infty \phi_q(\boldsymbol{y}; \boldsymbol{\mu}, \Sigma) \phi\{[z - \boldsymbol{\alpha}^T \omega^{-1}(\boldsymbol{y} - \boldsymbol{\mu})]\} dz$$

This representation of the skew-normal density suggests using $Z|\boldsymbol{y} \sim N(\boldsymbol{\alpha}^T\omega^{-1}(\boldsymbol{y} - \boldsymbol{\mu}), 1) I(z > 0)$ as a latent variable when developing the EM algorithm for estimating the parameters. Generally, for a skew-normal random vector $\boldsymbol{y} \sim \mathrm{SN}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$, the

joint distribution for $(\boldsymbol{y}, Z)$ is

$$f(\boldsymbol{y}, z) = 2\phi_q(\boldsymbol{y}; \boldsymbol{\mu}, \Sigma)\phi\left\{z - \boldsymbol{\alpha}^T\omega^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right\};$$

see [65, p. 718]. Therefore, the conditional distribution of Z given $\boldsymbol{y}$ is

$$f(z|\boldsymbol{y}) = \frac{f(\boldsymbol{y}, z)}{f(\boldsymbol{y})} = \frac{\phi\left\{z - \boldsymbol{\alpha}^T\omega^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right\}}{\Phi\left\{\boldsymbol{\alpha}^T\omega^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right\}} I(z > 0), \tag{3.6}$$

which is a truncated normal distribution with the mean equal to

$$\hat{Z} = E(Z|\boldsymbol{y}) = \boldsymbol{\eta}^T(\boldsymbol{y} - \boldsymbol{\mu}) + \frac{\phi\left\{\boldsymbol{\eta}^T(\boldsymbol{y} - \boldsymbol{\mu})\right\}}{\Phi\left\{\boldsymbol{\eta}^T(\boldsymbol{y} - \boldsymbol{\mu})\right\}}. \tag{3.7}$$

In the EM algorithm for the skew-normal family, the formula (3.7) would be used to estimate the latent variable in the E-step; see Section 3.3.2.

3.2.2 The Multivariate Skew-t Distribution

[68] defined a new class of multivariate distributions by the transformation

$$\boldsymbol{y} = \boldsymbol{\mu} + W^{-1/2}\boldsymbol{y}^\star \tag{3.8}$$

where $\boldsymbol{y}^\star \sim \mathrm{SN}_q(\boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ and $W \sim \chi^2_\nu/\nu$, independent of $\boldsymbol{y}^\star$. Then the random vector $\boldsymbol{y}$ has a multivariate skew-t distribution, denoted by $\boldsymbol{y} \sim \mathrm{St}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \nu)$, with the density function

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \nu) = 2t_q(\boldsymbol{y}; \nu)T_1\left\{\boldsymbol{\alpha}^T\omega^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\left(\frac{\nu + q}{Q\boldsymbol{y} + \nu}\right)^{1/2}; \nu + q\right\}, \tag{3.9}$$

where ω is as in Section 3.2.1,

$$Q\boldsymbol{y} = (\boldsymbol{y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}),$$

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

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 scalar t distribution with degrees of freedom $\nu+q$. The mean and variance of $\boldsymbol{y}$ are

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{y}} &= \boldsymbol{\mu} + b_\nu \omega \boldsymbol{\delta}, \qquad \nu > 1, \\ \mathrm{Var}(\boldsymbol{y}) &= \frac{\nu}{\nu-2}\boldsymbol{\Sigma} - \omega \boldsymbol{\mu}_{\boldsymbol{y}} \boldsymbol{\mu}_{\boldsymbol{y}}^T \omega, \end{aligned}$$

where

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

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

It is known [69] that if $\boldsymbol{y} \sim \mathrm{St}_q(\boldsymbol{\mu}, \Sigma, \boldsymbol{\alpha}, \nu)$ is partitioned as

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

where $\boldsymbol{y}_1$ has the size h, then the marginal distribuion of $\boldsymbol{y}_1$ still belongs to the family of multivariate skew-t distributions, i.e., $\boldsymbol{y}_1 \sim \mathrm{St}_h(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \tilde{\boldsymbol{\alpha}}, \nu)$. However, the skewness parameter $\tilde{\boldsymbol{\alpha}}$ is more complicated than $\boldsymbol{\alpha}_1$ and takes the form of

$$\tilde{\boldsymbol{\alpha}} = \frac{\boldsymbol{\alpha}_1 + \tilde{\boldsymbol{\Sigma}}_{11}^{-1}\tilde{\boldsymbol{\Sigma}}_{12}\boldsymbol{\alpha}_2}{(1+\boldsymbol{\alpha}_2^T\tilde{\boldsymbol{\Sigma}}_{22\cdot1}\boldsymbol{\alpha}_2)^{1/2}}, \tag{3.10}$$

where $\tilde{\boldsymbol{\Sigma}} = \omega^{-1}\boldsymbol{\Sigma}\omega^{-1}$ has the same partition 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 $\boldsymbol{y}$ has a univariate skew-t distribution, whose skewness parameter, denoted by $\tilde{\alpha}_i$, is quite different from α_i, the ith entry of $\boldsymbol{\alpha}$.

As a scale mixture of multivariate skew-normal distributions, the multivariate skew-t family provides a wider parametric family encompassing the families of multivariate normal, skew-normal and t. On the one hand, the skewness parameter in the

density allows the multivariate skew-t distribution to deal with the asymmetry of the data; on the other hand, the multivariate skew-t distribution is more robust than the skew-normal in the sense that it has heavy tails so that it can handle the outliers or atypical observations, without the need to detecting or removing them. The degrees of freedom ν controls the kurtosis or heaviness of the tail of the distribution. As ν goes to infinity, the multivariate skew-t distribution tends to the multivariate skew-normal distribution; for $\nu = 1$ and $\mathbf{\Sigma} = \mathbf{I}_{q\times q}$, a $q \times q$ identity matrix, it becomes the multivariate skew-Cauchy distribution.

Using the definition of a multivariate skew-t in (3.8) as a scale mixture of multivariate skew-normals, it is natural to augment the observed data by including the two latent variables

$$W \sim \chi^2_\nu/\nu \quad \text{and} \quad Z|\boldsymbol{y} \sim N\left(\boldsymbol{\alpha}^T\omega^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), 1\right) I(z>0),$$

when developing the EM-type algorithm. For convenience, we denote the *pdf* for Gamma(a, b) with mean a/b and variance a/b^2 by $h(w; a, b)$, so the density function for W would be $h(w; \nu/2, \nu/2)$. Therefore, the joint density for the complete-data $(\boldsymbol{y}, Z, W)$ is

$$f(\boldsymbol{y}, z, w) = 2\phi_q(\boldsymbol{y}; \boldsymbol{\mu}, \mathbf{\Sigma}/w)\phi\left\{z - \sqrt{w}\boldsymbol{\eta}^T(\boldsymbol{y}-\boldsymbol{\mu})\right\} \cdot h(w; \nu/2, \nu/2).$$

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

$$f(w|\boldsymbol{y}) = \frac{f(\boldsymbol{y}, w)}{f(\boldsymbol{y})} = \frac{2}{f(\boldsymbol{y})}\Phi\left\{\sqrt{w}\boldsymbol{\eta}^T(\boldsymbol{y}-\boldsymbol{\mu})\right\}\phi_q(\mathbf{z}; \boldsymbol{\mu}, \mathbf{\Sigma}/w) \cdot h(w; \nu/2, \nu/2), \quad (3.11)$$

and

$$f(w, z|\boldsymbol{y}) = \frac{2}{f(\boldsymbol{y})}\phi\left\{z - \sqrt{w}\boldsymbol{\eta}^T(\boldsymbol{y}-\boldsymbol{\mu})\right\}\phi_q(\boldsymbol{y}; \boldsymbol{\mu}, \mathbf{\Sigma}/w) \cdot h(w; \nu/2, \nu/2). \quad (3.12)$$

The relevant conditional expectations needed for the EM algorithm are given in the

following proposition:

Proposition 1. Suppose that $\boldsymbol{y} \sim \mathrm{St}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ with the associated latent variables $W \sim \chi^2_\nu/\nu$ and $Z \sim N(0,1)I(z>0)$. Then, for any $m>0$, we have

$$\begin{aligned} E\{W^m|\boldsymbol{y}\} &= C(\theta_1, r_1) T_1\left(\boldsymbol{\eta}^T(\boldsymbol{y}-\boldsymbol{\mu})\sqrt{\frac{r_1}{\theta_1}}; 2r_1\right) \\ E\{ZW^m|\boldsymbol{y}\} &= \frac{1}{\sqrt{2\pi}} C(\theta_2, r_1) + \boldsymbol{\eta}^T(\boldsymbol{y}-\boldsymbol{\mu}) \cdot E\{W^m|\boldsymbol{y}\} \end{aligned}$$

where

$$\begin{aligned} r_1 &= \frac{q+\nu}{2} + m, \qquad \theta_1 = \frac{(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})+\nu}{2}, \\ \theta_2 &= \frac{(\boldsymbol{y}-\boldsymbol{\mu})^T(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\eta}\boldsymbol{\eta}^T)(\boldsymbol{y}-\boldsymbol{\mu})+\nu}{2}, \\ C(x,y) &= \left(\frac{1}{\sqrt{2\pi}}\right)^q \cdot \frac{2}{f(\boldsymbol{y})} \cdot |\boldsymbol{\Sigma}|^{-1/2} \frac{\theta^r}{\Gamma(r)} \cdot \frac{\Gamma(y)}{x^y}. \end{aligned}$$

The formula for computing $E\{W^m|\boldsymbol{y}\}$ can also be found in the Corollary 1 of [73] where $E\left\{W^{m/2}\phi[\sqrt{W}\boldsymbol{\eta}^T(\boldsymbol{y}-\boldsymbol{\mu})]/\Phi[\sqrt{W}\boldsymbol{\eta}^T(\boldsymbol{y}-\boldsymbol{\mu})]|\boldsymbol{y}\right\}$ is also calculated instead of $E\{ZW^m|\boldsymbol{y}\}$. For completeness, all the details for deriving these two conditional expectations are given in the Appendix.

3.3 Penalized Skew-normal and Skew-t Log-likelihoods

In this section, we provide the details for joint estimation of the parameters $(\mathbf{B}, \boldsymbol{\Omega})$ of a multivariate regression model in (1.2) using penalized skew-normal and skew-t log-likelihoods, and the expectation conditional maximization (ECM) algorithm.

3.3.1 The Penalized Skew-normal Log-likelihood

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

$$\begin{aligned} L(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}) \;\; \propto \;\; & \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \mathbf{B}^T\boldsymbol{x}_i)^T\boldsymbol{\Omega}(\boldsymbol{y}_i - \mathbf{B}^T\boldsymbol{x}_i) - \log|\boldsymbol{\Omega}| \\ & -\frac{2}{n}\sum_{i=1}^{n}\log\left[\Phi\{\boldsymbol{\eta}^T(\boldsymbol{y}_i - \mathbf{B}^T\boldsymbol{x}_i)\}\right]. \end{aligned} \tag{3.13}$$

where $\boldsymbol{\Omega} = (\omega_{ij}) = \Sigma^{-1}$ and $\boldsymbol{\eta} = \omega^{-1}\boldsymbol{\alpha}$.

As in [43], we regularize the entries of $\mathbf{B}$ and $\boldsymbol{\Omega}$ by imposing ℓ_1 penalties on them, and estimate them by minimizing the penalized log-likelihood:

$$g(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}) = L(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}) + \lambda_1\sum_{j'\neq j}|\omega_{j'j}| + \lambda_2\sum_{j,k}|b_{jk}|. \tag{3.14}$$

When the sample size n is smaller than p and q, the parameters are not estimable through the maximum likelihood in (3.13) since we do not have enough observations. By adding the ℓ_1 penalties, the parameters that are less important or irrelevant would be forced to be zero resulting in the reduction of the number of parameters and the improvement of prediction accuracy. Moreover, the ℓ_1 penalty on the precision matrix ensures the estimate is nonsingular and well-conditioned.

Compared with the penalized normal likelihood in [43], the penalized skew-normal likelihood function has the additional third term in (3.13) involving the skewness and other parameters. As a result, the optimization of $g(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta})$ would be more challenging. To overcome the difficulty, we rely on the stochastic representation of skew-normal distributions in Section 3.2.1 and propose an iterative procedure to minimize $g(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta})$ using an extension of EM, called Expectation Conditional Maximization (ECM) algorithm [74].

3.3.2 An Optimization Algorithm via ECM

Suppose that $Z_1, \cdots, Z_n$ are the latent variables associated with the observations $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ such that $Z_i|\boldsymbol{y}_i \sim N\left(\boldsymbol{\alpha}^T\omega^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}), 1\right) I(z > 0)$. We treat $(\boldsymbol{y}_i, Z_i), 1 \leq i \leq n$, as the complete data, while the original observations $\boldsymbol{y}_i's$ are viewed as the incomplete data. The likelihood functions not depending on the latent variables would be referred as the incomplete-data likelihood. By contrast, the likelihood function of $(\boldsymbol{y}_i, Z_i)$ are the complete-data likelihood and we use a subscript c to distinguish them from the incomplete-data likelihood. Denote $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)^T$ and $\mathbf{Z} = (Z_1, \cdots, Z_n)^T$, so the negative complete-data log-likelihood is proportional to

$$\begin{aligned} L_c(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}) &= \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{y}_i - \mathbf{B}^T\boldsymbol{x}_i)^T\boldsymbol{\Omega}(\boldsymbol{y}_i - \mathbf{B}^T\boldsymbol{x}_i) - \log|\boldsymbol{\Omega}| \\ &\quad + \frac{1}{n}\sum_{j=1}^{n}\left[Z_i - \boldsymbol{\eta}^T(\boldsymbol{y}_i - \mathbf{B}^T\boldsymbol{x}_i)\right]^2 \\ &= \operatorname{tr}\{\boldsymbol{\Omega}\boldsymbol{S}\} - \log|\boldsymbol{\Omega}| + \frac{1}{n}\|\mathbf{Z} - (\boldsymbol{Y} - \mathbf{X}\mathbf{B})\boldsymbol{\eta}\|^2, \end{aligned} \tag{3.15}$$

where $\boldsymbol{S} = \frac{1}{n}(\boldsymbol{Y} - \mathbf{X}\mathbf{B})^T(\boldsymbol{Y} - \mathbf{X}\mathbf{B})$ and $\|\cdot\|$ 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

$$g_c(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}) = L_c(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}) + \lambda_1\sum_{i\neq j}|\omega_{ij}| + \lambda_2\sum_{i,j}|b_{ij}|. \tag{3.16}$$

Minimizing the function $g(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta})$ in (3.14) is equivalent to minimizing $g_c(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta})$ using the EM algorithm [53] which performs an expectation (E) step and a maximization (M) step alternately until convergence. In the E-step, the expectation of $g_c(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta})$ conditional on the observed data $\boldsymbol{Y}$ is evaluated using the current estimates of the parameters $\Theta = \{\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta}\}$; in the M-step, the expected log-likelihood

function is minimized over the parameter space. We describe the details as follows:

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

In the E-step, we only have to calculate the conditional expectation of Z_i given $\hat{\Theta}^{(k)}$ and $\boldsymbol{Y}$. Using the formula in (3.7), we denote this conditional expectation by $\hat{Z}_i = E(Z_i|\boldsymbol{Y}, \hat{\Theta}^{(k)})$. Therefore, the expected log-likelihood, denoted by $Q_c(\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta})$, is

$$\begin{aligned} Q_c(\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta}) &= \operatorname{tr}\{\mathbf{\Omega}\boldsymbol{S}\} - \log|\mathbf{\Omega}| + \frac{1}{n}\|\hat{\mathbf{Z}} - (\boldsymbol{Y} - \mathbf{X}\mathbf{B})\boldsymbol{\eta}\|^2 + \operatorname{Var}(Z_i|\boldsymbol{Y}, \hat{\Theta}^{(k)}) \\ &\quad + \lambda_1 \sum_{i \neq j} |\omega_{ij}| + \lambda_2 \sum_{i,j} |b_{ij}| \end{aligned} \tag{3.17}$$

where $\hat{\mathbf{Z}} = (\hat{Z}_1, \cdots, \hat{Z}_n)^T$ and $\operatorname{Var}(Z_i|\boldsymbol{Y}, \hat{\Theta}^{(k)})$ is a constant which can be ignored in the minimization step. Note that the function in (3.17) is not convex in $(\mathbf{B}, \mathbf{\Omega})$, but it is convex in one argument when the other is fixed. This suggests an iterative algorithm alternating between estimation of $\mathbf{B}$ and $\mathbf{\Omega}$ for minimizing it.

M-step: Minimizing $Q_c(\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta})$ over the whole parameter space Θ 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 Θ is minimized while the other blocks are fixed [74].

CM1: Given $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}^{(k)}$ and $\mathbf{B} = \hat{\mathbf{B}}^{(k)}$, minimizing $Q_c(\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta})$ with respect to $\mathbf{\Omega}$ is equivalent to solving

$$\hat{\mathbf{\Omega}}^{(k+1)} = \arg\min_{\mathbf{\Omega}} \left\{ -\log|\mathbf{\Omega}| + \operatorname{tr}\{\mathbf{\Omega}S^{(k)}\} + \lambda_1 \sum_{j \neq j'} |\omega_{jj'}| \right\}, \tag{3.18}$$

where $S^{(k)} = \frac{1}{n}\left(\boldsymbol{Y} - \mathbf{X}\hat{\mathbf{B}}^{(k)}\right)\left(\boldsymbol{Y} - \mathbf{X}\hat{\mathbf{B}}^{(k)}\right)^T$.

This is the ℓ_1 penalized covariance estimation problem considered by many au-

thors including [28], [29], [30] and [31]. The fast graphical lasso algorithm of Friedman et al. (2008) is adopted to solve (3.18). As shown in [28], the estimate $\hat{\mathbf{\Omega}}^{(k+1)}$ would remain positive definite as long as $S^{(k)}$ is positive definite.

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

$$\hat{\boldsymbol{\eta}}^{(k+1)} = \left(\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}}\right)^{-1}\tilde{\boldsymbol{Y}}^T\mathbf{Z}, \tag{3.19}$$

where $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \mathbf{X}\hat{\mathbf{B}}^{(k)}$.

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

$$\frac{1}{n}\text{tr}\left\{(\boldsymbol{Y}_0 - \mathbf{XB})^T\mathbf{\Omega}_0(\boldsymbol{Y}_0 - \mathbf{XB})\right\} + \lambda_2\sum_{i,j}|b_{ij}|, \tag{3.20}$$

where $\mathbf{\Omega}_0 = \hat{\mathbf{\Omega}}^{(k+1)} + \hat{\boldsymbol{\eta}}^{(k+1)}\left(\hat{\boldsymbol{\eta}}^{(k+1)}\right)^T$ and $\boldsymbol{Y}_0 = \boldsymbol{Y} - \hat{\mathbf{Z}}\left(\hat{\boldsymbol{\eta}}^{(k+1)}\right)^T\mathbf{\Omega}_0^{-1}$.

As in [72], the equation (3.20) can be rewritten into the form of lasso regression. Following [43] and [72], we use the coordinate descent (Cod) algorithm [29] for solving this problem. Other efficient algorithms such as the homotopy algorithm [54] and the Lars-lasso algorithm [55] can also be applied.

3.3.3 The MRSN Algorithm for Skew-Normal Errors

We summarize the preceding ECM algorithm for minimizing (3.16) and refer to it as MRSN.

MRSN Algorithm: With λ_1 and λ_2 fixed, initialize the parameters $\Theta = \Theta^{(0)}$. On the (k+1)th iteration,

E-step: Estimate the latent variables Z_i by their conditional expectations as in (3.7).

CM1: Given $\mathbf{B} = \hat{\mathbf{B}}^{(k)}$ and $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}^{(k)}$, update $\mathbf{\Omega} = \hat{\mathbf{\Omega}}^{(k+1)}$ in (3.18) using the graphical

lasso algorithm.

CM2: Given $\mathbf{B} = \hat{\mathbf{B}}^{(k)}$ and $\mathbf{\Omega} = \hat{\mathbf{\Omega}}^{(k+1)}$, update $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}^{(k+1)}$ with the least square estimate in (3.19).

CM3: Given $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}^{(k+1)}$ and $\mathbf{\Omega} = \hat{\mathbf{\Omega}}^{(k+1)}$, update $\mathbf{B} = \hat{\mathbf{B}}^{(k+1)}$ in (3.20) using the coordinate descent (Cod) algorithm.

Repeat the E- and three CM-steps until the estimates of the parameters converge, that is, $\sum_{j,k} |\hat{b}_{jk}^{(k+1)} - \hat{b}_{jk}^{(k)}| \leq \epsilon \sum_{jk} |\hat{b}_{jk}^{ridge}|$, where $\hat{\mathbf{B}}^{ridge} = (\hat{b}_{jk}^{ridge}) = (\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{Y}$ is the ridge estimate and the tolerance parameter ϵ is set at 10^{-4} by default.

The MRSN algorithm is similar to the MRCE method for the normal data, but the latter only consists of analogues of the CM1 and CM3 steps. In the MRSN algorithm, 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 take more time to converge 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.

3.3.4 The MRST Algorithm for Skew-t Errors

Now we assume that the errors $\boldsymbol{\epsilon}_i's$ in (1.2) have a multivariate skew-t distribution with the location $\boldsymbol{\mu} = 0$. Let Z_i and W_i be the corresponding latent variables for $\boldsymbol{y}_i$ with the negative complete-data log-likelihood as in (3.2). Similar to the skew-normal case, we construct sparse estimates for both $\mathbf{B}$ and $\mathbf{\Omega}$ by minimizing the penalized log-likelihood:

$$g_c(\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta}, \nu) = L_c(\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta}, \nu) + \lambda_1 \sum_{i \neq j} |\omega_{ij}| + \lambda_2 \sum_{i,j} |b_{ij}| \tag{3.21}$$

via the ECM algorithm where $L_c(\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta}, \nu)$ is given in (3.2).

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

In the E-step, the three expectations $E(W_i|\mathbf{X}, \boldsymbol{Y}, \hat{\Theta}^{(k)})$, $E(\log W_i|\mathbf{X}, \boldsymbol{Y}, \hat{\Theta}^{(k)})$ and $E(Z_i\sqrt{W_i}|\mathbf{X}, \boldsymbol{Y}, \hat{\Theta}^{(k)})$ are needed and can be evaluated using Proposition 1. Note that $E(W_i|\mathbf{X}, \boldsymbol{Y}, \hat{\Theta}^{(k)})$ and $E(Z_i\sqrt{W_i}|\mathbf{X}, \boldsymbol{Y}, \hat{\Theta}^{(k)})$ have closed forms, but $E(\log W_i|\mathbf{X}, \boldsymbol{Y}, \hat{\Theta}^{(k)})$ does not, so we compute the latter numerically using the method in [75]. Denote these conditional expectations by

$$a_i^{(k)} = E(\log W_i|\mathbf{X}, \boldsymbol{Y}, \hat{\Theta}^{(k)}), \quad b_i^{(k)} = E(W_i|\mathbf{X}, \boldsymbol{Y}, \hat{\Theta}^{(k)}),$$
$$c_i^{(k)} = E(Z_i\sqrt{W_i}|\mathbf{X}, \boldsymbol{Y}, \hat{\Theta}^{(k)}). \tag{3.22}$$

Plugging them in $g_c(\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta}, \nu)$, by some algebra, we get the expected log-likelihood function as

$$\begin{aligned} Q_c(\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta}, \nu) &= \frac{1}{n}\sum_{i=1}^{n} b_i^{(k)}(\boldsymbol{y}_i - \mathbf{B}^T\boldsymbol{x}_i)^T\mathbf{\Omega}(\boldsymbol{y}_i - \mathbf{B}^T\boldsymbol{x}_i) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{c_i^{(k)}}{\sqrt{b_i^{(k)}}} - \sqrt{b_i^{(k)}}\boldsymbol{\eta}^T(\boldsymbol{y}_i - \mathbf{B}^T\boldsymbol{x}_i)\right\}^2 \\ &\quad - \log|\mathbf{\Omega}| + \hat{a}(\nu) + \lambda_1\sum_{i\neq j}|\omega_{ij}| + \lambda_2\sum_{i,j}|b_{ij}|. \end{aligned} \tag{3.23}$$

where now

$$\hat{a}(\nu) = 2\log\Gamma\left(\frac{\nu}{2}\right) - \nu\log\left(\frac{\nu}{2}\right) - \frac{1}{n}(\nu + q - 2)\sum_{j=1}^{n} a_j^{(k)} + \frac{\nu}{n}\sum_{j=1}^{n} b_j^{(k)}. \tag{3.24}$$

Let $\tilde{\boldsymbol{y}}_i = \sqrt{b_i^{(k)}}\boldsymbol{y}_i$, $\tilde{\boldsymbol{x}}_i = \sqrt{b_i^{(k)}}\boldsymbol{x}_i$ and $\tilde{z}_i = c_i^{(k)}/\sqrt{b_i^{(k)}}$. If we define $\tilde{\boldsymbol{Y}} = (\tilde{\boldsymbol{y}}_1, \cdots, \tilde{\boldsymbol{y}}_n)^T$, $\tilde{\mathbf{Z}} = (\tilde{z}_1, \cdots, \tilde{z}_n)^T$ and $\tilde{\mathbf{X}} = (\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_n)^T$, then the function in (3.23)

can be written using the matrix notation as

$$\begin{aligned} Q_c(\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta}, \nu) &= \operatorname{tr}\left\{\mathbf{\Omega}\tilde{\boldsymbol{S}}\right\} - \log|\mathbf{\Omega}| + \frac{1}{n}\|\tilde{\mathbf{Z}} - (\tilde{\boldsymbol{Y}} - \tilde{\mathbf{X}}B)\boldsymbol{\eta}\|^2 + \hat{a}(\nu) \\ &\quad + \lambda_1 \sum_{i \neq j} |\omega_{ij}| + \lambda_2 \sum_{i,j} |b_{ij}|. \end{aligned} \tag{3.25}$$

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

Since the degrees of freedom ν is separated from the other three blocks of parameters, the M-step for the skew-t distribution proceed in the following way:

CM1: Given the first three blocks of parameters in $\Theta = (\mathbf{B}, \mathbf{\Omega}, \boldsymbol{\eta}, \nu)$, update ν as $\hat{\nu}^{(k+1)}$ by minimizing the function $\hat{a}(\nu)$ in (3.24).

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

Remark 1: The preceding ECM algorithm for minimizing (3.21) is referred to as the MRST algorithm. We point out the challenge encountered when estimating the degrees of freedom ν. In practice, we have noticed that the sequence $\left\{\hat{\nu}^{(k)}\right\}$ usually converges to a small positive number which is less than 2, whereas $\nu > 2$ is required for the covariance matrix to exist. Thus, the estimate of the degrees of freedom using the MRST algorithm is not satisfactory; the same phenomena occurs when $\boldsymbol{\epsilon}_i's$ have a multivariate t distribution [72]. In most of what follows, we discard the CM1 step in the MRST algorithm and estimate ν separately via the maximum likelihood method [69].

Remark 2: When $\boldsymbol{\alpha} = \mathbf{0}$, the MRST algorithm would reduce to the MRCEI algorithm [72] which is developed to regularize parameters in the general linear model when the errors have a multivariate t distribution. Moreover, if $\boldsymbol{\alpha} = \mathbf{0}$ and ν goes to infinity, the MRST algorithm would reduce to the exact MRCE method in [43].

3.4 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 $S - S_k$ as the training data to estimate the parameters and S_k as the test set to evaluate the prediction error. Then 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,

$$(\hat{\lambda}_1, \hat{\lambda}_2) = \arg \min_{(\lambda_1, \lambda_2)} \frac{1}{Kq} \left\{ \sum_{k=1}^{K} \| \boldsymbol{Y}_{(k)} - \mathbf{X}_{(k)} \hat{\mathbf{B}}_{(-k)}^{\lambda_1, \lambda_2} - \hat{\boldsymbol{\mu}}_{\boldsymbol{E}} \|_{L_2}^2 \right\}, \tag{3.26}$$

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

We measure the performance of our methods in terms of the prediction error (PE) which has the form of

$$PE(\hat{\mathbf{B}}) = \frac{1}{n} \mathrm{tr} \left\{ (\boldsymbol{Y} - \mathbf{X}\hat{\mathbf{B}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{E}})(\boldsymbol{Y} - \mathbf{X}\hat{\mathbf{B}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{E}})^T \right\}. \tag{3.27}$$

The sparsity recognition performance of $\hat{\mathbf{B}}$ is measured by the true positive rate (TPR) as well as the true negative rate (TNR) which are defined in (2.23).

3.5 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 the least square, MRCE [43] for normal and MRCEI [72], for symmetric t distributions, respectively.

3.5.1 Model Design

Throughout this section we will have 50 replications of the multivariate regression with $n = 50$, $p = 22$ 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 we expect each predictor to be relevant for $s_1 q$ of all the response variables. An $n \times p$ predictor matrix $\mathbf{X}$ with $n = 50$ is also generated with rows drawn independently from $N(0, \boldsymbol{\Sigma}_X)$, where $(\boldsymbol{\Sigma}_X)_{ij} = 0.7^{|i-j|}$, as in [25] and [27]. We consider the AR(1) model for the scale matrix of the errors with $(\boldsymbol{\Sigma}_{\boldsymbol{E}})_{ij} = \rho_{\boldsymbol{E}}^{|i-j|}$.

Then each row of the error matrix $\boldsymbol{E}$ is independently drawn from a multivariate skew t distribution $\mathrm{St}_q(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{E}}, \boldsymbol{\alpha}, \nu)$ 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 [43]. This is similar to performing a K-fold cross-validation for the algorithm.

We consider different combinations of $\nu, \boldsymbol{\alpha}, \rho_E, s_1$ and s_2 from the following ranges: (1) $\nu = 10, 20, 40, 100$, (2) $\rho_{\boldsymbol{E}} = 0, 0.5, 0.7, 0.9$, (3) $\boldsymbol{\alpha} = (-1, 1, -1, \cdots, 1)^T$ 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 $\Lambda = \{10^x : x = 0, \pm 1, \cdots, \pm 5\}$ using

5-fold cross-validation. 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.

Table 10. PE for the AR(1) error covariance with $s_1 = 0.1$, $s_2 = 1$ and $\boldsymbol{\alpha} = (1, 1, 1, \cdots, 1)^T$. Average and standard errors in parenthesis are based on 50 replications.

	$\rho_{\boldsymbol{E}}$	OLS	MRCE	MRCEI	MRSN	MRST
$\nu = 10$	0.9	2.27	1.26	1.34	1.18	1.11
		(0.05)	(0.03)	(0.02)	(0.02)	(0.02)
	0.7	2.30	1.34	1.45	1.41	1.40
		(0.04)	(0.02)	(0.01)	(0.06)	(0.01)
	0.5	2.27	1.39	1.47	1.53	1.49
		(0.03)	(0.02)	(0.01)	(0.01)	(0.02)
	0.0	2.23	1.43	1.50	1.59	1.58
		(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
$\nu = 20$	0.9	2.05	1.14	1.18	1.01	1.00
		(0.05)	(0.01)	(0.02)	(0.01)	(0.01)
	0.7	2.00	1.21	1.26	1.25	1.25
		(0.03)	(0.01)	(0.01)	(0.01)	(0.02)
	0.5	1.99	1.28	1.30	1.36	1.34
		(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
	0.0	2.04	1.29	1.28	1.43	1.41
		(0.03)	(0.02)	(0.01)	(0.02)	(0.01)
$\nu = 40$	0.9	1.89	1.07	1.10	0.96	0.92
		(0.04)	(0.02)	(0.02)	(0.02)	(0.02)
	0.7	1.88	1.13	1.17	1.20	1.16
		(0.03)	(0.01)	(0.01)	(0.01)	(0.01)
	0.5	1.93	1.18	1.21	1.31	1.26
		(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
	0.0	1.88	1.22	1.25	1.36	1.34
		(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$\nu = 100$	0.9	1.83	1.07	1.09	0.91	0.90
		(0.04)	(0.02)	(0.02)	(0.02)	(0.02)
	0.7	1.88	1.12	1.14	1.14	1.14
		(0.03)	(0.01)	(0.01)	(0.06)	(0.01)
	0.5	1.91	1.16	1.17	1.25	1.24
		(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
	0.0	1.82	1.19	1.21	1.31	1.32
		(0.02)	(0.01)	(0.01)	(0.01)	(0.01)

Table 11. PE for the AR(1) error covariance with $s_1 = 0.5$, $s_2 = 1$ and $\boldsymbol{\alpha} = (1, 1, 1, \cdots, 1)^T$. Average and standard errors in parenthesis are based on 50 replications.

	$\rho_{\boldsymbol{E}}$	OLS	MRCE	MRCEI	MRSN	MRST
$\nu = 10$	0.9	2.19	1.50	1.45	1.22	1.26
		(0.06)	(0.02)	(0.02)	(0.03)	(0.02)
	0.7	2.23	1.83	1.80	1.58	1.57
		(0.04)	(0.03)	(0.01)	(0.02)	(0.02)
	0.5	2.27	1.98	1.87	1.73	1.69
		(0.04)	(0.03)	(0.02)	(0.02)	(0.02)
	0.0	2.29	2.00	1.99	1.81	1.78
		(0.04)	(0.02)	(0.02)	(0.02)	(0.02)
$\nu = 20$	0.9	2.05	1.33	1.31	1.10	1.16
		(0.05)	(0.02)	(0.02)	(0.02)	(0.03)
	0.7	2.04	1.61	1.60	1.43	1.44
		(0.04)	(0.02)	(0.02)	(0.02)	(0.02)
	0.5	2.01	1.75	1.68	1.56	1.56
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	0.0	2.01	1.77	1.77	1.62	1.66
		(0.03)	(0.02)	(0.02)	(0.01)	(0.02)
$\nu = 40$	0.9	1.88	1.26	1.25	1.02	1.02
		(0.04)	(0.02)	(0.02)	(0.02)	(0.02)
	0.7	1.90	1.51	1.52	1.34	1.32
		(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
	0.5	1.86	1.64	1.61	1.48	1.45
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	0.0	1.89	1.67	1.70	1.54	1.54
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
$\nu = 100$	0.9	1.85	1.24	1.20	1.00	0.99
		(0.04)	(0.02)	(0.02)	(0.02)	(0.02)
	0.7	1.85	1.49	1.47	1.30	1.28
		(0.03)	(0.02)	(0.02)	(0.06)	(0.01)
	0.5	1.86	1.60	1.56	1.42	1.41
		(0.03)	(0.02)	(0.01)	(0.01)	(0.01)
	0.0	1.85	1.63	1.62	1.48	1.50
		(0.02)	(0.02)	(0.02)	(0.01)	(0.01)

3.5.2 Results and Discussion

We report the prediction errors for the AR(1) error covariance in Tables 10 and 11. Note that the OLS always has the largest prediction errors indicating its poor performance relative to the other methods and, in general, the prediction errors tend

to increase as ρ_E decreases. The MRCE method generally outperforms the other methods in terms of prediction errors when $\mathbf{B}$ is more sparse ($s_1 = 0.1$), except for $\rho_E = 0.9$ where the MRSN and MRST have smaller prediction errors. This suggests that the MRSN and MRST perform well for highly correlated data and more sparse $\mathbf{B}$.

Table 12. TPR/TNR for the AR(1) error covariance averaged over 50 replications with $s_1 = 0.1$, $s_2 = 1$ and $\boldsymbol{\alpha} = (1, 1, 1, \cdots, 1)^T$.

	$\rho_{\boldsymbol{E}}$	MRCE	MRCEI	MRSN	MRST
	0.9	0.93/0.59	0.93/0.57	0.94/0.28	0.98/0.13
$\nu = 10$	0.7	0.88/0.63	0.89/0.61	0.93/0.32	0.95/0.27
	0.5	0.85/0.66	0.85/0.64	0.92/0.34	0.92/0.33
	0.0	0.82/0.67	0.83/0.62	0.91/0.36	0.91/0.35
	0.9	0.93/0.59	0.94/0.53	0.94/0.29	0.97/0.18
$\nu = 20$	0.7	0.89/0.62	0.90/0.59	0.94/0.30	0.94/0.31
	0.5	0.85/0.65	0.86/0.63	0.92/0.33	0.93/0.35
	0.0	0.83/0.65	0.85/0.62	0.91/0.35	0.91/0.35
	0.9	0.94/0.59	0.94/0.54	0.95/0.25	0.96/0.22
$\nu = 40$	0.7	0.90/0.61	0.90/0.59	0.95/0.29	0.93/0.32
	0.5	0.87/0.63	0.85/0.61	0.93/0.32	0.92/0.35
	0.0	0.85/0.63	0.84/0.63	0.92/0.35	0.91/0.35
	0.9	0.84/0.63	0.94/0.56	0.95/0.26	0.97/0.22
$\nu = 100$	0.7	0.86/0.64	0.90/0.59	0.95/0.30	0.93/0.33
	0.5	0.88/0.62	0.87/0.61	0.93/0.32	0.91/0.34
	0.0	0.92/0.59	0.85/0.62	0.92/0.35	0.91/0.34

On the other hand, from Table 11 it is evident that when $\mathbf{B}$ is less sparse ($s_1 = 0.5$),

the MRSN and MRST perform quite well in that the prediction errors for the two methods are always smaller than those for the MRCE and MRCEI.

Table 13. TPR/TNR for the AR(1) error covariance averaged over 50 replications with $s_1 = 0.5$, $s_2 = 1$ and $\boldsymbol{\alpha} = (1, 1, 1, \cdots, 1)^T$.

	$\rho_{\boldsymbol{E}}$	MRCE	MRCEI	MRSN	MRST
	0.9	0.91/0.45	0.91/0.43	0.94/0.32	0.98/0.11
$\nu = 10$	0.7	0.87/0.42	0.86/0.48	0.91/0.34	0.94/0.24
	0.5	0.85/0.40	0.84/0.50	0.89/0.36	0.91/0.33
	0.0	0.80/0.52	0.81/0.53	0.87/0.41	0.90/0.34
	0.9	0.92/0.41	0.92/0.44	0.94/0.32	0.97/0.16
$\nu = 20$	0.7	0.86/0.47	0.86/0.47	0.92/0.34	0.93/0.29
	0.5	0.85/0.43	0.84/0.49	0.90/0.35	0.91/0.33
	0.0	0.82/0.51	0.84/0.47	0.89/0.40	0.90/0.34
	0.9	0.93/0.40	0.92/0.44	0.94/0.32	0.97/0.17
$\nu = 40$	0.7	0.88/0.44	0.86/0.49	0.92/0.33	0.93/0.30
	0.5	0.86/0.43	0.84/0.51	0.90/0.34	0.91/0.33
	0.0	0.83/0.52	0.83/0.51	0.89/0.40	0.90/0.34
	0.9	0.93/0.40	0.93/0.43	0.94/0.31	0.96/0.22
$\nu = 100$	0.7	0.87/0.46	0.86/0.49	0.92/0.33	0.92/0.31
	0.5	0.86/0.44	0.87/0.44	0.90/0.35	0.91/0.33
	0.0	0.84/0.47	0.82/0.55	0.88/0.39	0.90/0.35

The corresponding true positive rates (TPR) and true negative rates (TNR) for the AR(1) error covariance are reported in Tables 12 and 13. We note that, with ν fixed, the positive (negative) rates tend to decrease (increase) as ρ_E decreases.

Additionally, the true positive rates for the MRSN and MRST are very large while the true negative rates are very small. Therefore, the regularization methods for the skew distributions give more conservative estimates for $\mathbf{B}$ in the sense that $\hat{\mathbf{B}}$ is less sparse.

3.6 Real Data Analysis

In this section, we re-examine the hourly average electricity spot prices from Australia with the general linear model in (2.26). 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 electricity prices are highly volatile. Because of the apparent skewness in the profile plot, it may be more reasonable to model the error as $\boldsymbol{\epsilon}_i \sim \mathrm{St}_{24}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \nu)$. We apply the MRST algorithm to this model to get sparse estimators for $\mathbf{B}$ and $\boldsymbol{\Omega}$ as well as to improve the prediction accuracy.

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 (2.26). 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 $\Lambda = \left\{2^{-10+20(x-1)/39} : x = 1, 2, \cdots, 40\right\}$. In the case when the Lars-lasso algorithm is used in the CM3 step, we use 80% of the observations in the first 998 days as the training data and the remaining 20% as the validation data. In what follows, for ease of notation, we refer to the results using Lars-lasso and Cod algorithms in the CM3 step as Lars-lasso and Cod, respectively.

The average squared prediction error for each hour in a day for the last 100 days

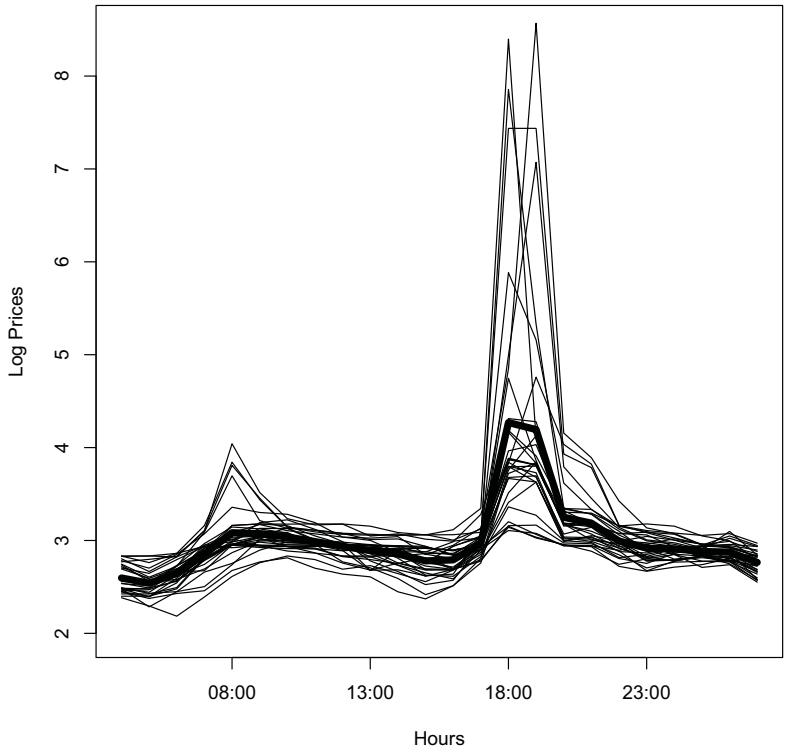

Fig. 2. The profile plot of the hourly electricity wholesale prices. The solid dark curve is the mean profile.

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 Cod turns out to be 0.075 which is the smallest among all the methods considered. (However, the estimate for neither $\mathbf{B}$ nor $\mathbf{\Sigma}$ is sparse.

From Figure 2, since most individual skewness parameters appear to be small, we consider a model with all the elements in the skewness parameter $\boldsymbol{\alpha}$ fixed at zero except for $\alpha_5, \alpha_{15}, \alpha_{16}, \alpha_{17}$. The prediction errors corresponding to this special fixed choice, labeled Cod-fix and Lars-fix, and other models are plotted in Figure 3. Here again Cod-fix has the smallest hourly prediction errors. To assess the meaning and

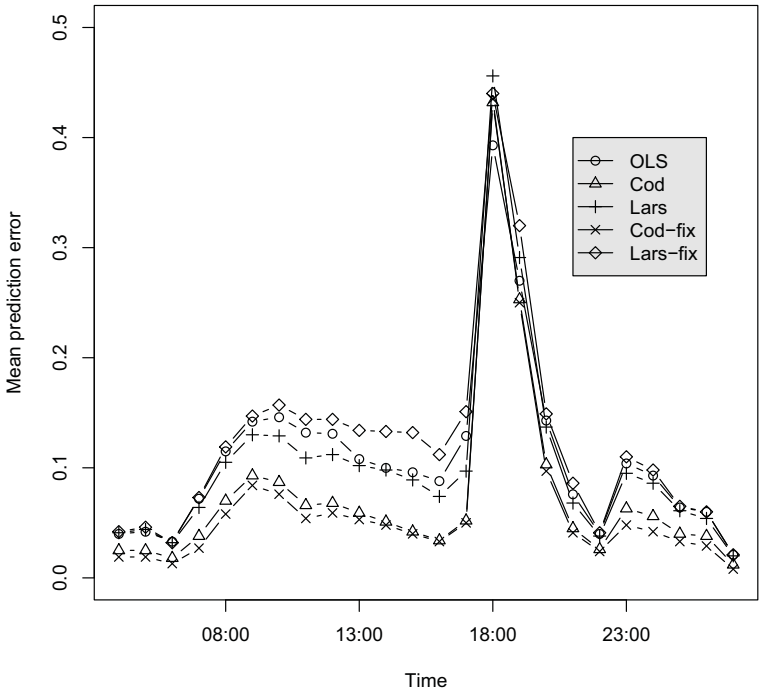

Fig. 3. The average squared prediction error for each hour on a day based on 100 points.

relevance of the skew vector of the fitted model, we compare it with the skewness parameters of the marginal distributions for each component in $\boldsymbol{Y}$ and expect them to agree with the skewness parameter of the estimated density plot for each univariate component. From the marginal skewness parameters (dashed line) in Figure 4 (left panels), it is evident that these are all negative for the estimated models using Cod no matter if $\boldsymbol{\alpha}$ is fixed and very different from those using Lars-lasso which are all positive. In addition, the marginal skewness parameters for the model using Lars-fix present the similar pattern to the estimated $\boldsymbol{\alpha}$. To determine which one of them describe the skewness of the data best, we examine the marginal density plot for each component of $\boldsymbol{Y}$ and find that all the subseries are right-skewed except the first two

components of $\boldsymbol{Y}$. This indicates that the estimated marginal skewness parameters should be positive for all the components of $\boldsymbol{Y}$ expect the first two. From this point of view, it seems that the fitted model using Lars-lasso, whose the marginal skewness parameters are all positive, is more plausible.

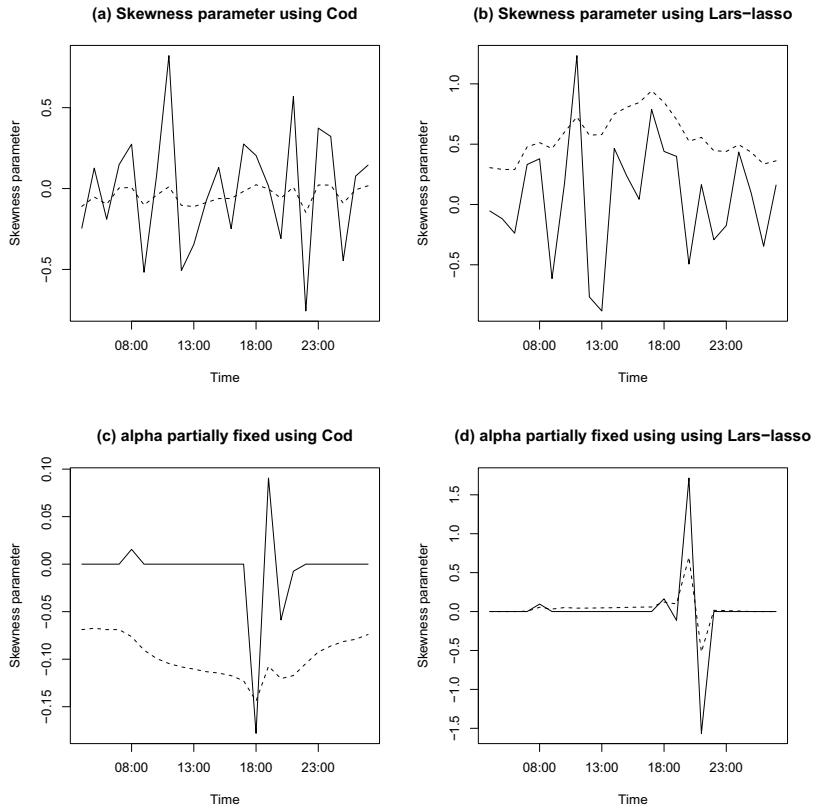

Fig. 4. The estimated skewness parameters using different models and algorithms (The dash line is the marginal skewness parameter and the solid line is the estimate for $\boldsymbol{\alpha}$). (a) Lars-lasso without fixing $\boldsymbol{\alpha}$ (b) Lars-lasso with $\boldsymbol{\alpha}$ fixed (c) Cod without fixing $\boldsymbol{\alpha}$ (d) Cod with $\boldsymbol{\alpha}$ fixed.

3.7 Summary

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 assumption of the skew-normal or skew-t distribution on the errors provides a more flexible probabilistic distribution and enables us to handle the possible skewness in the data. Two algorithms, namely MRSN for skew-normal and MRST for skew-t, which extend the MRCE [43] and MRCEI [72] algorithms are developed to iteratively compute the estimate of the parameters. As pointed out in Section 3.3.4, we are encountered with the same problem as the authors of [72] when estimating the degrees of freedom and suggest estimating it outside the iterations. We have shown the the MRST outperforms the MRCE and MRCEI 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 MRCEI. We have also noticed that the estimated skewness parameter sometimes is far away from the true value. Our further work would focus on improving the estimation of the skewness parameter.

CHAPTER IV

TESTING PROPORTIONALITY OF THE SKEWNESS VECTOR AND EIGENVECTORS OF MULTIVARIATE SKEW-NORMAL DISTRIBUTIONS

In Chapter III, we have introduced the family of multivariate skew-normal distributions with the densities given by (3.4). In this chapter, we focus on the principal component analysis of a skew-normal variable and its connection to the canonical variates. We denote the scale matrix and covariance matrix of a multivariate skew-normal variate by $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ respectively and refer $\boldsymbol{\eta} = \omega^{-1}\boldsymbol{\alpha}$ as the skewness vector. For ease of notation, we simply denote $z \sim \mathrm{SN}_p(0, \boldsymbol{\Omega}, \boldsymbol{\alpha})$ by $z \sim \mathrm{SN}_p(\boldsymbol{\Omega}, \boldsymbol{\alpha})$.

4.1 Introduction

An important property of the family of the skew-normal distributions is that it is closed under the linear transformation. As in [66], there exists a *canonical* transform $y = Wz$ such that $y \sim \mathrm{SN}_p(\xi^\star, I, \boldsymbol{\alpha}^\star)$ where $W = (w_1, w_2, \cdots, w_p)'$ is a $p \times p$ real matrix and at most one element of $\boldsymbol{\alpha}^\star$ is nonzero. This linear transformation converts the multivariate skew-normal vector into the one whose components are independent and defines a so-called *canonical form* of multivariate skew-normal distribution. Without loss of generality, we assume that $\boldsymbol{\alpha}^\star = (\alpha_1^\star, 0, \cdots, 0)'$. Consequently, the canonical variates $w_1'z, \cdots, w_p'z$ are independent and the coefficients $w_i's$ should satisfy

$$w_1 = \frac{\boldsymbol{\eta}}{\sqrt{\boldsymbol{\eta}'\boldsymbol{\Omega}\boldsymbol{\eta}}}, \quad w_i'\boldsymbol{\Omega}w_i = 1, \quad w_i'\boldsymbol{\Omega}w_j = 0, \quad i \neq j.$$

Interestingly, the canonical variates are closely related to the principal components of z when one of the eigenvectors of $\boldsymbol{\Omega}$ is proportional to the skewness vector $\boldsymbol{\eta}$ as shown by the Proposition 2.1 of [76] which is stated below:

Proposition 2.1: Let $z \sim \mathrm{SN}_p(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ and Γ' be a $p \times p$ matrix whose columns $\gamma_1, \cdots, \gamma_p$ are normalized eigenvectors of $\boldsymbol{\Omega}$ corresponding to the eigenvalues $\lambda_1, \cdots, \lambda_p$ with $\gamma_j \propto \boldsymbol{\eta}$ for some j. Then the principal components of z are independent, proportional to the canonical variates, and the variance of z can be represented as

$$\mathrm{Var}(Z) = \Gamma \mathrm{diag}\left\{\lambda_1, \cdots, \frac{\pi\lambda_j + \boldsymbol{\eta}'\boldsymbol{\eta}(\pi-2)\lambda_j^2}{\pi(1+\lambda_j\boldsymbol{\eta}'\boldsymbol{\eta})}, \cdots, \lambda_p\right\}\Gamma'.$$

The proposition shows that if the skewness vector is proportional to an eigenvector of $\boldsymbol{\Omega}$ the eigenvectors of $\boldsymbol{\Sigma}$ would be the same as those of $\boldsymbol{\Omega}$ so that the principal components of z are proportional to the canonical variates and thus independent. In this case, the distributions for the principal components of z are very simple: only one of them is skew-normal and the others are normal.

In this project, we focus on the connection between the canonical variates and principal components of a skew-normal variate as well as the asymptotic properties of the eigenvectors for the maximum likelihood estimate (MLE) of $\boldsymbol{\Omega}$. We firstly investigate the asymptotic distributions for the MLEs of the eigenvectors and eigenvalues and show that these asymptotic distributions would be the same as those when z is normal. Our primary goal is to develop a statistic for testing whether one of the eigenvectors of $\boldsymbol{\Omega}$ is proportional to $\boldsymbol{\eta}$, i.e.,

$$H_0 : \gamma_j \propto \boldsymbol{\eta} \text{ for some } j \text{ vs. } H_a : \gamma_j \not\propto \boldsymbol{\eta} \text{ for any } j. \tag{4.1}$$

To develop a statistic for testing (4.1), instead, we consider p simpler the individual proportionality testing problems between each eigenvector γ_j and $\boldsymbol{\eta}$, i.e.,

$$H_0^{(j)} : \gamma_j \propto \boldsymbol{\eta} \text{ vs. } H_a^{(j)} : \gamma_j \not\propto \boldsymbol{\eta}, \quad j = 1, \cdots, p. \tag{4.2}$$

Then a statistic for testing (4.1) is developed based on the likelihood ratio test (LRT) statistic for testing (4.2).

The chapter is organized as follows. In section 4.2, we briefly review the maximum likelihood estimation of the parameters for multivariate skew-normal distribution and discuss the asymptotic distributions for the estimates of the eigenvectors and eigenvalues of $\mathbf{\Omega}$ and $\mathbf{\Sigma}$. In Section 4.3, we establish the LRT statistic for the hypothesis test in (4.2) and propose a statistic for testing the hypothesis in (4.1). In Section 4.4, we conduct a simulation study to access the performance of the statistic we have proposed. Some comments and conclusions are included in Section 4.5.

4.2 Distributions of the Eigenvalues and Eigenvectors

In this section, we review the procedure for maximizing the skew-normal likelihood and give the asymptotic distributions for the eigenvalues and eigenvectors of the MLEs of $\mathbf{\Omega}$. For simplicity, we assume that all the eigenvalues $\lambda_i's$ are different.

4.2.1 Maximum Likelihood Estimate of $\Theta = (\mathbf{\Omega}, \boldsymbol{\eta})$

To start with, denote its MLE of Θ by $\hat{\Theta} = (\hat{\mathbf{\Omega}}, \hat{\boldsymbol{\eta}})$. Suppose we have n i.i.d. observations $z_1, \cdots, z_n$ from $\mathrm{SN}_p(\mathbf{\Omega}, \boldsymbol{\alpha})$, so the log-likelihood function is

$$\ell(\Theta) = \text{constant} - \frac{n}{2}\log|\mathbf{\Omega}| - \frac{n}{2}\mathrm{tr}\left\{\mathbf{\Omega}^{-1}S_0\right\} + \sum_{i=1}^{n}\log\left[\Phi(\boldsymbol{\eta}'z_i)\right] \tag{4.3}$$

where $S_0 = \frac{1}{n}\sum_{i=1}^{n} z_i z_i'$. Taking derivative of $\ell(\Theta)$ with respective to $\mathbf{\Omega}$ yields

$$\hat{\mathbf{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} z_i z_i'. \tag{4.4}$$

By plugging $\hat{\mathbf{\Omega}}$ into $\ell(\Theta)$, one can obtain the MLE of $\boldsymbol{\eta}$ by numerically maximizing the profile likelihood:

$$\ell(\boldsymbol{\eta}) = \sum_{i=1}^{n}\log\left[\Phi(\boldsymbol{\eta}'z_i)\right] \tag{4.5}$$

Then the matrix of the eigenvectors Γ and eigenvalues Λ of $\mathbf{\Omega}$ can be estimated by $\hat{\Gamma} = (\hat{\gamma}_1, \cdots, \hat{\gamma}_p)$ and $\hat{\Lambda} = \text{diag}\{\hat{\lambda}_1, \cdots, \hat{\Lambda}_p\}$ using the spectral decomposition of $\hat{\mathbf{\Omega}}$:

$$\hat{\Gamma}\hat{\Lambda}\hat{\Gamma}' = \hat{\mathbf{\Omega}}, \qquad \hat{\Gamma}\hat{\Gamma}' = I_p,$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$.

4.2.2 Asymptotic Distributions with $\hat{\Gamma}$ and $\hat{\Lambda}$

When $z \sim \text{SN}_p(\mathbf{\Omega}, \boldsymbol{\alpha})$, it is known that $\hat{\mathbf{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} z_i z_i'$ has a Wishart distribution with the scale matrix $\mathbf{\Omega}$ and degrees of freedom 1 [77]. This implies that the asymptotic distributions of the eigenvectors and eigenvalues would be the same as those for $\mathbf{\Omega}$ when z_i is normal [78]. Therefore, we have

Proposition 2. Suppose $z_1, z_2, \cdots, z_n \sim \text{SN}_p(\mathbf{\Omega}, \boldsymbol{\alpha})$ with ξ known. Let $\hat{\gamma}_i's$ and $\hat{\lambda}_i's$ be the MLEs of the eigenvectors and eigenvalues of $\mathbf{\Omega}$ defined in Proposition 2.1 of [76]. Then $\sqrt{n}(\hat{\gamma}_i - \gamma_i)$ and $\sqrt{n}(\hat{\lambda}_i - \lambda_i)$ for $1 \leq i \leq p$ are mutually and asymptotically independent. Their limiting distributions are

$$\begin{aligned} \sqrt{n}(\hat{\lambda}_i - \lambda_i) &\sim N(0, 2\lambda_i^2) \\ \sqrt{n}(\hat{\gamma}_i - \gamma_i) &\sim N_p\left(0, \mathbf{\Sigma}_{\gamma_i}\right) \end{aligned} \tag{4.6}$$

where $\mathbf{\Sigma}_{\gamma_i} = \lambda_i \sum_{k \neq i} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \gamma_k \gamma_k'$.

Remark 1: It is easy to show that the eigenvalues and the eigenvectors of the common sample covariance matrix estimator $S = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(z_i - \bar{z})'$ for $\mathbf{\Sigma}$ have the same asymptotic distributions as those of $\hat{\mathbf{\Omega}}$ since $\sqrt{n}\hat{\mathbf{\Omega}}$ and $\sqrt{n}S$ have the same limiting distribution.

Remark 2: The authors of [79] have computed the expected Fisher information matrix for Θ and show that it is singular in the absence of asymmetry. To get the asymptotic distribution of the eigenvectors of $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Omega}} - \hat{\mu}\hat{\mu}'$, we assume that $\boldsymbol{\alpha} \neq 0$.

Then $U \triangleq \sqrt{n}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})$ is also asymptotically normal with mean 0 but the covariance structure is not as simple as that for $\hat{\boldsymbol{\Omega}}$. One can determine it using the Delta method and apply Theorem 3.1.7 in [80, pp. 297] to get the asymptotic distributions for the eigenvectors and eigenvalues of $\hat{\boldsymbol{\Sigma}}$. This approach can be exploited to derive the asymptotic distributions of the eigenvectors and eigenvalues for $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\Sigma}}$ in the general cases when the location parameter is included in the skew-normal density.

4.3 The LR Test Statistic

In this section, we propose the likelihood ratio statistic for testing whether the skewness vector $\boldsymbol{\eta}$ is proportional to one of the eigenvectors $\gamma_j's$. The LRT statistic for the hypothesis $H_0^{(j)}$ in (4.2) is

$$T_j = -2[\ell(\hat{\Theta}_0) - \ell(\hat{\Theta})] \sim \chi^2_{p-1}, \tag{4.7}$$

where $\hat{\Theta}_0$ and $\hat{\Theta}$ are the MLEs for the parameters under the null hypothesis $H_0^{(j)}$ and its alternative, respectively. This statistic has an approximate chi-square distribution with degrees of freedom $p-1$ [81]. To test whether one of the eigenvectors of $\boldsymbol{\Sigma}$ is proportional to $\boldsymbol{\eta}$, we use

$$T = \min_{1 \le j \le p} T_j,$$

as the test statistic. At the significant level of α_F, we would reject H_0 in (4.1) if

$$T > \chi^2_{\alpha_F, p-1}, \tag{4.8}$$

where $\chi^2_{\alpha_F,p-1}$ represents the $100(1-\alpha_F)$th percentile of χ^2_{p-1}. In this case, the true type I error for testing H_0 is bounded by α_F since

$$\begin{aligned} P_{H_0}(T > \chi^2_{\alpha_F,p-1}) &= P\left(T > \chi^2_{\alpha_F,p-1}, \bigcup_{j=1}^{p} H_0^{(j)} \,|H_0\right) \\ &= \sum_{j=1}^{p} P\left(T > \chi^2_{\alpha_F,p-1}, H_0^{(j)} \,|H_0\right) \\ &= \sum_{j=1}^{p} P_{H_0^{(j)}}(T > \chi^2_{\alpha_F,p-1}) P(H_0^{(j)}|H_0) \\ &\leq \sum_{j=1}^{p} P_{H_0^{(j)}}(T_j > \chi^2_{\alpha_F,p-1}) P(H_0^{(j)}|H_0) \\ &= \sum_{j=1}^{p} \alpha_F P(H_0^{(j)}|H_0) = \alpha_F. \end{aligned}$$

The second equality above holds because the null hypotheses $H_0^{(j)}$ for $1 \leq j \leq p$ are mutually exclusive.

The log-likelihood function for n observations $z_1, \cdots, z_n$ is given by

$$\ell(\Theta) = \text{constant} - \frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{n}{2}\text{tr}\left\{\boldsymbol{\Omega}^{-1}S_0\right\} + \sum_{i=1}^{n}\log\left\{\Phi(\boldsymbol{\eta}' z_i)\right\}. \tag{4.9}$$

Under the alternative hypothesis $H_a^{(j)}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\eta}$ can be simply estimated by their MLEs as in (4.4) and (4.5). Under the null hypothesis $H_0^{(j)} : \boldsymbol{\eta} \propto \gamma_j$, we have $\boldsymbol{\eta} = b\gamma_j$ for some b and the corresponding likelihood function is

$$\ell(\Theta_1) = \text{constant} - \frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{n}{2}\text{tr}\left\{\boldsymbol{\Omega}^{-1}S_0\right\} + \sum_{i=1}^{n}\log\left\{\Phi(b\gamma_j' z_i)\right\}, \tag{4.10}$$

where $\Theta_1 = (\boldsymbol{\Omega}, b)$. The analytic maximization of $\ell(\Theta_1)$ over Θ_1 is a challenging problem due to a constraint imposed on an eigenvector of $\boldsymbol{\Omega}$. As an alternative, one could accomplish the task numerically with the gradient supplied to an optimization algorithm, but based on our experience such a procedure does not guarantee that

the estimate of $\mathbf{\Omega}$ is positive definite. A possible way to overcome this difficulty is to reparametrize the matrix Γ of eigenvectors is expressed in terms of a product of the Givens rotation matrices [82], [83], [84]. More precisely, let $Q = p(p-1)/2$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_Q)'$, with $\theta_j \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ be the vector of the angles. Then the orthogonal matrix Γ can be rewritten as

$$\Gamma(\boldsymbol{\theta}) = \prod_{m_1=1}^{p-1} \prod_{m_2=m_1+1}^{p} G_k^{(m_1,m_2)}(\theta_k) = G_1^{(1,2)}(\theta_1) \cdots G_{p-1}^{(1,p)}(\theta_{p-1}) G_p^{(2,3)}(\theta_p) \cdots G_Q^{(p-1,p)}(\theta_Q)$$

where $k = m_2 - m_1 + (m_1 - 1)(p - m_1/2)$ and $G_k^{(m_1,m_2)}(\theta_k)$ is a rotation matrix with the elements given by

$$\left[G_k^{(m_1,m_2)}(\theta_k)\right]_{i,j} = \begin{cases} \cos(\theta_k), & \text{if } i = j = m_1 \text{ or } m_2 \\ \sin(\theta_k), & \text{if } i = m_1 \text{ or } j = m_2 \\ -\sin(\theta_k), & \text{if } i = m_2 \text{ or } j = m_1 \\ 1, & \text{if } i = j \neq m_1 \text{ and } i = j \neq m_2 \\ 0, & \text{otherwise.} \end{cases}$$

For example, for $p = 2$, Γ is

$$\Gamma(\boldsymbol{\theta}) = G_1^{(1,2)}(\theta_1) = \begin{bmatrix} \cos(\theta_1) & \sin(\theta_1) \\ -\sin(\theta_1) & \cos(\theta_1) \end{bmatrix}.$$

With this parametrization, the likelihood function for $\Theta_2 = (\Lambda, \boldsymbol{\theta}, b)$ is

$$\ell(\Theta_2) = \text{constant} - \frac{n}{2}\sum_{i=1}^{p} \log \lambda_i - \frac{n}{2}\text{tr}\left\{\Gamma(\boldsymbol{\theta})\Lambda^{-1}\Gamma'(\boldsymbol{\theta})S_0\right\} + \sum_{i=1}^{n} \log\left\{\Phi[b\gamma_j'(\boldsymbol{\theta})z_i]\right\}. \quad (4.11)$$

After some algebra, we have $\hat{\lambda}_i = \gamma_i'(\boldsymbol{\theta})S_0\gamma_i(\boldsymbol{\theta})$ for $i = 1, \cdots, p$ and the profile likelihood for $(\boldsymbol{\theta}, b)$ is

$$\ell(\boldsymbol{\theta}, b) = \text{constant} - \frac{n}{2}\sum_{i=1}^{p} \log\left[\gamma_i'(\boldsymbol{\theta})S_0\gamma_i(\boldsymbol{\theta})\right] + \sum_{i=1}^{n} \log\left\{\Phi\left[b\gamma_j'(\boldsymbol{\theta})z_i\right]\right\}. \quad (4.12)$$

We compute the gradient of $\ell(\boldsymbol{\theta}, b)$ in the Appendix and maximize $\ell(\boldsymbol{\theta}, b)$ numerically to obtain the MLEs for $\boldsymbol{\theta}$ and b.

Another way of parametrization relies on the modified Cholesky decomposition of $\boldsymbol{\Omega}$. It is convenient to reparametrize the problem by writing

$$\boldsymbol{\Omega}^{-1} = A' \text{diag}\{\exp(\rho)\} A = A' D A$$

$$\boldsymbol{\eta} = b\gamma_j$$

where $\rho = (\rho_1, \cdots, \rho_p)'$ and A is an upper triangular $p \times p$ matrix with the diagonal elements equal to 1. The log-likelihood for the parameter $\Theta_3 = (A, \rho, b)$ given by

$$\ell(\Theta_3) = \text{constant} - \frac{n}{2}\sum_{i=1}^{n}\rho_i - \frac{n}{2}\text{tr}\left\{A' D A S_0\right\} + \sum_{i=1}^{n}\log\left\{\Phi(b\gamma_i' z_i)\right\}. \qquad (4.13)$$

can be maximized numerically with its gradient applied to improve the efficiency (See the Appendix).

4.4 A Simulation Study

In this section, we conduct a simulation study to evaluate the performance of the proposed LR test statistic under two scenarios: (1) H_0 is true and (2) H_a is true. The empirical type I error rates and the empirical powers are computed, respectively, at the significant level α_F based on $N = 1000$ replications. Different combinations of n, α_F and p are considered: $n = 50, 100, 250, 1000$; $\alpha_F = 0.05, 0.10$ and $p = 2, 3$. The details of the procedures for simulating data are described below:

(1) Construct a $p \times p$ orthogonal matrix $\Gamma = (\gamma_1, \cdots, \gamma_p)$ using the product of rotation matrices with all the rotation angles equal to $\frac{\pi}{3}$, and use $\Lambda = \text{diag}\{1, 2\}$ for $p = 2$ and $\Lambda = \text{diag}\{1, 1.5, 2\}$ for $p = 3$ as the matrix of the eigenvalues. Then the true scale matrix is determined by $\boldsymbol{\Omega} = \Gamma\Lambda\Gamma'$.

(2) Randomly permutate the eigenvectors $\gamma_k's$ to obtain a new orthogonal matrix $Q = (\mathbf{q}_1, \cdots, \mathbf{q}_p)$ where $\mathbf{q}_k$ is the kth column of Q. Let $\bar{\mathbf{q}} = \frac{1}{\sqrt{p}} \sum_{k=1}^{p} \mathbf{q}_k$ be the mean of the eigenvectors of unit length and denote the acute angle formed by $\boldsymbol{\eta}$ and $\mathbf{q}_k$ as $\pi_k \in \left[0, \frac{\pi}{2}\right]$. Define the distance between $\boldsymbol{\eta}$ and Q by

$$D(\boldsymbol{\eta}, Q) = \min_k \{\pi_k\}.$$

Set $\boldsymbol{\eta} = (1 - a_0)\mathbf{q}_1 + a_0\bar{\mathbf{q}}$ where $0 \leq a_0 \leq 1$ controls the extent to which $\boldsymbol{\eta}$ and $\mathbf{q}_k's$ are far away from the null hypothesis H_0 in terms of the minimum angle between $\boldsymbol{\eta}$ and $\mathbf{q}_k's$. More specifically, we have

$$\begin{aligned} \cos(\pi_1) &= \frac{\boldsymbol{\eta}'\mathbf{q}_1}{\|\boldsymbol{\eta}\|} = \frac{1 - a_0 + \frac{a_0}{\sqrt{p}}}{\sqrt{\left(1 - a_0 + \frac{a_0}{\sqrt{p}}\right)^2 + \frac{p-1}{p}a_0^2}} \\ \cos(\pi_k) &= \frac{\boldsymbol{\eta}'\mathbf{q}_k}{\|\boldsymbol{\eta}\|} = \frac{\frac{a_0}{\sqrt{p}}}{\sqrt{\left(1 - a_0 + \frac{a_0}{\sqrt{p}}\right)^2 + \frac{p-1}{p}a_0^2}}, \; 2 \leq k \leq p. \end{aligned} \tag{4.14}$$

leading to $D(\boldsymbol{\eta}, Q) = \pi_1$, an increasing function of a_0. When $a_0 = 0$, $D(\boldsymbol{\eta}, Q) = 0$ implying that $\boldsymbol{\eta}$ is proportional to one of eigenvectors of $\boldsymbol{\Omega}$; when $a_0 = 1$, $\pi_k's$ are all equal, so $\boldsymbol{\eta}$ is in the equiangular direction of $\gamma_k's$. In the simulation study, we have considered five different values for $a_0 : 0, 0.25, 0.5, 0.75, 1$.

(3) Generate $z_i \sim \mathrm{SN}_p(0, \boldsymbol{\Omega}, \alpha)$ for $1 \leq i \leq n$ and compute the MLEs for the parameters in the full model as in (4.4) and (4.5). Similarly, under $H_0^{(j)}$ compute the MLEs for the parameters by maximizing $\ell(\Theta_0)$ numerically with two different ways of reparametrization of $\boldsymbol{\Omega}$.

(4) For each j, $1 \leq j \leq p$, compute the test statistic $T_j = -2[\ell(\hat{\Theta}_l) - \ell(\hat{\Theta})]$ with the p-value $p_j = \Pr\left(\chi^2_{p-1} > T_j\right)$. If there exist a j_0 such that $p_{j_0} > \alpha_F$, count = count + 1.

(5) Repeat (3) and (4) for N times and compute the ratio = count/N.

We report the estimated ratios corresponding to different values of a_0 in Tables 14 and 15 for $p = 2$ and $p = 3$, respectively. The first column of the tables corresponds to the empirical type I errors, while the others correspond to the empirical powers with varying degrees of departure from the null hypothesis. We see that the powers tend to increase as either a_0 or n increases. The type I errors for Givens Angle agree with the nominal value of the significant level very well, although they are slightly smaller than α_F. However, the type I errors for Modified Cholesky Decomposition (MCD) tends to decrease as n increases and are larger than the nominal α_F especially when p is large and n is small. In addition, we note that all the ratios for the Givens Angle are larger than those for the MCD. This indicates that the test statistic using MCD are more likely to reject H_0 than the statistic using Givens Angle. The performance of the MCD can be explained by the fact that there are p more parameters in (4.13) than (4.12). Therefore, when we maximize them numerically, (4.13) is more likely to obtain a local maximizer than (4.12) leading to a larger value of T_j.

4.5 Data Analysis

The data we analyze is the Australian Institute of Sport (AIS) data examined in [85], which contains various biomedical measurements on $n = 202$ Australian athletes. To illustrate our test, we apply it to subsets of the AIS where the skew-normal distribution is fitted to the variables.

(a) $z = (\text{Ht}, \text{Bmi})'$. Because the MLE for the location parameter is $\hat{\xi} = (180.51, 19.98)'$, the variable z is "centered" by subtracting $\hat{\xi}$ such that the location parameter for the variable $\tilde{z} = z - \hat{\xi}$ is roughly 0. The test statistic for H_0 is 0.019 with $pvalue = 0.890$, so we fail to reject the null hypothesis.

Table 14. Type I error rates and power when nominal $\alpha_F = 0.05$ and $p = 2$

Modified Cholesky Decomposition					
n	$a_0 = 0$	$a_0 = 0.25$	$a_0 = 0.5$	$a_0 = 0.75$	$a_0 = 1$
50	0.065	0.108	0.284	0.488	0.584
100	0.088	0.210	0.576	0.875	0.918
250	0.056	0.389	0.901	0.999	0.999
1000	0.050	0.914	1	1	1
Givens Angle					
n	$a_0 = 0$	$a_0 = 0.25$	$a_0 = 0.5$	$a_0 = 0.75$	$a_0 = 1$
50	0.040	0.092	0.252	0.401	0.481
100	0.053	0.199	0.553	0.823	0.874
250	0.046	0.387	0.900	0.999	0.999
1000	0.047	0.914	1	1	1

(b) $z = (\text{Ht}, \text{Bfat})'$. The MLE for the location parameter is $\hat{\xi} = (182.48, 5.73)'$ and consider $\tilde{z} = z - \hat{\xi}$ as the variable. The test statistic for H_0 is 19.03 with $pvalue < 0.001$, so we reject the null hypothesis.

(c) $z = (\text{Ht}, \text{Bmi}, \text{Bfat})'$. The MLE for the location parameter is $\hat{\xi} = (182.86, 76.07, 5.85)'$ and consider $\tilde{z} = z - \hat{\xi}$ as the variable. The test statistic for H_0 is 47.72 with $pvalue << 0.001$, so we reject the null hypothesis.

4.6 Summary

We investigated the asymptotic distributions for the eigenvalues and eigenvector of the MLEs for $\mathbf{\Omega}$ and $\mathbf{\Sigma}$, and proposed a statistic to test if one of the eigenvectors

Table 15. Type I error rates and power when nominal $\alpha_F = 0.05$ and $p = 3$

Modified Cholesky Decomposition					
n	$a_0 = 0$	$a_0 = 0.25$	$a_0 = 0.5$	$a_0 = 0.75$	$a_0 = 1$
50	0.226	0.259	0.381	0.498	0.558
100	0.208	0.266	0.578	0.866	0.909
250	0.113	0.371	0.866	0.995	0.998
1000	0.051	0.850	1	1	1
Givens Angle					
n	$a_0 = 0$	$a_0 = 0.25$	$a_0 = 0.5$	$a_0 = 0.75$	$a_0 = 1$
50	0.026	0.043	0.114	0.191	0.204
100	0.049	0.114	0.405	0.619	0.562
250	0.044	0.294	0.825	0.973	0.962
1000	0.043	0.849	1	1	1

of $\boldsymbol{\Omega}$ is proportional to the skewness parameter $\boldsymbol{\eta}$. The simulation study shows that the parametrization using the Givens Angle performs better than using the Modified Cholesky Decomposition for small p. However, as p increases, the number of parameters in the profile likelihood would increase quadratically. In this situation, numerical maximization of the profile likelihood is not so reliable since the algorithm would more likely converge to the local maximum resulting in a higher ratio of rejecting the null hypothesis.

CHAPTER V

CONCLUSIONS, EXTENSIONS AND FUTURE WORK

5.1 Regularization of Parameters in Multivariate Linear Regression

In Chapter II and III, instead of imposing the usual normality assumption on the errors, we assume that they have a multivariate t/skew-normal/skew-t distribution and propose an iterative procedure to construct the sparse estimators for both $\mathbf{B}$ and $\mathbf{\Omega}$ in this setup. This extends the MRCE methods [43] and provides a more plausible way to improve the prediction accuracy when the data have outliers or, particularly, exhibit skewness. However, from the simulation study in Chapter III, we have noticed that the skewness parameter estimated through the regularization method sometimes is very different from the true one. In this case, the interpretation of the estimated skewness parameter is difficult since it may not agree with the shape of data. Moreover, the algorithms we have developed are very time consuming especially when p and q are large. Therefore, our future work would focus on how to obtain a better estimate of the skewness parameter in the framework of regularization as well as improve the numerical efficiency.

Our methods can be further extended the multivariate linear mixed model (MLMM) where a random effect is also included in (1.2) with different assumptions on the joint distribution of the random effect and errors [73].

5.2 Principal Component Analysis of a Skew-normal Variable

In Chapter IV, we investigate the asymptotic distributions of the eigenvalues and eigenvectors for the MLE of the scale matrix of the skew-normal distributions. We develop a statistic for testing whether the skewness vector is proportional to an eigen-

vector of the scale matrix based on the LR test statistic. We conclude that the reparametrization via the Givens Angles performs better than that via the Modified Cholesky Decomposition. As discussed in Chapter IV, the maximization of the log-likelihood function under the null hypothesis should be done numerically. As the dimension increases, the MLEs obtained from the numerical maximization are less reliable because they are more likely to be the local maxima. More theoretical work should be done in the future for the MLEs under the null hypothesis when the dimension is moderate or large.

REFERENCES

[1] L. Breiman and J. H. Friedman, "Predicting multivariate responses in multiple linear regression," *Journal of Royal Statistical Society Serie B*, vol. 59, no. 1, pp. 3–54, Jan. 1997.

[2] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. London, UK: Academic Press, 1979.

[3] E. W. Frees, *Longitudinal and Panel Data: Analysis and Application In The Social Sciences*. London: Cambridge University Press, 2004.

[4] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of Royal Statistical Society Serie B*, vol. 68, no. 1, pp. 49–67, Feb. 2006.

[5] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis,* 3rd ed. New York: Wiley, 2003.

[6] A. Zellner, "An efficient method of estimating seemingly unrelated regression equations and tests of aggregation bias," *Journal of American Statistical Association*, vol. 57, no. 298, pp. 348–368, Jun. 1962.

[7] C. Stein, "Inadmissibility of the usual estimator of the mean of a multivariate normal distribution," in *Proceedings of the Third Berkeley Sympossium on Mathematical and Statistical Probability*, vol. 1, UC Berkeley, Dec. 1956, pp. 197–206.

[8] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, Feb. 2004.

[9] T. Anderson, "Estimating linear restrictions on regression coefficients for multivariate normal distributions," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 327–351, Sep. 1951.

[10] A. Izenman, "Reduced-rank regression for the multivariate linear model," *Journal of Multivariate Analysis*, vol. 5, no. 2, pp. 248–264, Jun. 1975.

[11] G. Reinsel, *Elements of Multivariate Time Series Analysis*. New York: Springer, 1997.

[12] E. Bedrick and C. Tsai, "Model selection for multivariate regression in small samples," *Biometrics*, vol. 50, no. 1, pp. 226–231, Mar. 1994.

[13] Y. Fujikoshi and K. Satoh, "Modified aic and cp in multivariate linear regression," *Biometrika*, vol. 84, no. 3, pp. 707–716, Sep. 1997.

[14] R. Lutz and P. Buhlmann, "Boosting for high-multivariate responses in high-dimensional linear regression," *Statistica Sinica*, vol. 16, no. 2, pp. 471–494, Apr. 2006.

[15] P. Brown, M. Vannucci, and T. Fearn, "Bayes model averaging with selection of regressors," *Journal of Royal Statistical Society Serie B*, vol. 64, no. 3, pp. 519–536, Aug. 2002.

[16] F. Li and R. Zhang, "Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics," *Journal of the American Statistical Association*, vol. 105, no. 491, pp. 1202–1214, Sep. 2010.

[17] P. H. Garthwaite, "An interpretation of partial least squares," *Journal of the American Statistical Association*, vol. 89, no. 425, pp. 122–127, Mar. 1994.

[18] V. Vinzi, W. W. Chin, J. Henseler, and H. Wang, *Handbookk of Partial Least Squares: Concepts, Methods and Applications*. New York: Springer, 2010.

[19] M. West, "Bayesian factor regression models in the large p, small n paradigm," in *Bayesian Statistics*. Oxford University Press, 2003, pp. 723–732.

[20] C. Carvalho, J. Chang, J. Lucas, J. R. Nevins, Q. Wang, and M. West, "High-dimensional sparse factor models and latent factor regression," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1438–1456, Dec. 2005.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of Royal Society Series B*, vol. 58, no. 1, pp. 267–288, Jan. 1996.

[22] H. Zou, "The adaptive lasso and its oracle properties," *Journal of American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, Dec. 2006.

[23] I. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–148, May 1993.

[24] H. Zhou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of Royal Statistical Society Serie B*, vol. 67, no. 2, pp. 301–320, Apr. 2005.

[25] J. Peng, J. Zhu, A. Bergamaschi, W. Han, D. Y. Noh, J. R. Pollack, and P. Wang, "Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer," *Annual of Applied Statistics*, vol. 4, no. 1, pp. 53–77, Dec. 2010.

[26] D. Warton, "Penalized normal likelihood and ridge regularization of correlation and covariance matrices," *Journal of the American Statistical Association*, vol. 103, no. 481, pp. 340–349, Mar. 2008.

[27] M. Yuan and Y. Lin, "Model selection and estimation in the gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, Mar. 2007.

[28] A. d'Aspremont, O. Banerjee, and L. El Ghaoui, "First-order methods for sparse covariance selection," *SIAM Journal on Matrix Analysis and its Applications*, vol. 30, no. 1, pp. 55–66, Feb. 2008.

[29] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, Jul. 2008.

[30] A. Rothman, P. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electronic Journal of Staistics*, vol. 2, pp. 494–515, 2008.

[31] Z. Lu, "Smooth optimization approach for sparse covariance selection," *SIAM Journal on Matrix Analysis and its Applications*, vol. 19, no. 4, pp. 1807–1827, Dec. 2008.

[32] P. Bickel and L. Levina, "Regularized estimation of large covariance matrices," *Annals of Statistics*, vol. 36, no. 1, pp. 199–227, Feb. 2008.

[33] A. J. Rothman, E. Levina, and J. Zhu, "Generalized thresholding of large covariance matrices," *Journal of American Statistis Association*, vol. 104, no. 485, pp. 177–186, Mar. 2009.

[34] G. V. Rocha, P. Zhao, and B. Yu, "A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice)," Department of Statistics, UC Berkeley, Tech. Rep., Jul. 2008.

[35] J. Peng, P. Wang, N. Zhou, and J. Zhu, "Partial correlation estimation by joint sparse regression models," *Journal of American Statistics Association*, vol. 104, no. 486, pp. 735–746, Jun. 2009.

[36] M. Pourahmadi, "Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation," *Biometrika*, vol. 86, no. 3, pp. 677–690, Sep. 1999.

[37] W. B. Wu and M. Pourahmadi, "Nonparametric estimation of large covariance matrices of longitudinal data," *Biometrika*, vol. 90, no. 4, pp. 831–844, Dec. 2003.

[38] J. Huang, N. Liu, M. Pourahmadi, and L. Liu, "Covariance matrix selection and estimation via penalized normal likelihood," *Biometrika*, vol. 93, no. 1, pp. 85–98, Mar. 2006.

[39] C. Chang and R. Tsay, "Estimation of covariance matrix via the sparse cholesky factor with lasso," *Journal of Statistical Planning and Inference*, vol. 140, no. 12, pp. 3858–3873, Dec. 2010.

[40] R. Furrer and T. Bengtsson, "Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants," *Journal of Multivariate Analysis*, vol. 98, no. 2, pp. 227–255, Feb. 2007.

[41] P. Bickel and E. Levina, "Covariance regularization by thresholding," *Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, Dec. 2008.

[42] D. M. Witten and R. Tibshirani, "Covariance-regularized regression and classification for high-dimensional problems," *Journal of the Royal Statistical Society, Series B*, vol. 71, no. 3, pp. 615–636, Jun. 2009.

[43] A. Rothman, E. Levina, and J. Zhu, "Sparse multivariate regression with covariance estimation," *Journal of Computational and Graphical Statistics*, vol. 19, no. 4, pp. 947–962, Dec. 2010.

[44] P. J. Huber and E. Ronchetti, *Robust Statistics*. New York: Wiley, 2009.

[45] A. Zellner, "Bayesian and non-bayesian analysis of the regression model with multivariate student t error terms," *Journal of American Statistical Association*, vol. 71, no. 354, pp. 400–405, Jun. 1976.

[46] K. L. Lange, R. J. Little, and J. M. Taylor, "Robust statistical modeling using t distributions," *Journal of American Statistical Association*, vol. 84, no. 408, pp. 881–896, Dec. 1989.

[47] R. J. Adler, R. E. Feldman, and M. S. Taqqu, *A practical guide to heavy tails: statistical techniques and applications*. Birkhauser, Boston: Cambridge University Press, 1998.

[48] S. Resnick, *Heavy-tail Phenomena: Probabilistic and Statistical Modeling*. New York: Springer, 2007.

[49] B. C. Sutradhar and M. M. Ali, "Estimation of the parameters of a regression model with a multivariate t error variable," *Communications in Statistics Theory and Methods*, vol. 15, no. 2, pp. 429–450, 1986.

[50] S. Kotz and S. Nadarajah, *Multivariate t Distributions and Their Applications*. New York: Cambridge University Press, 2004.

[51] J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, Dec. 2007.

[52] C. Liu and D. B. Rubin, "Ml estimation of the t distribution using em and its extensions, ecm and ecme," *Statistica Sinica*, vol. 5, pp. 19–39, 1995.

[53] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society Series B*, vol. 39, no. 1, pp. 1–38, Jan. 1977.

[54] M. Osborne, B. Presnell, and B. Turlach, "A new approach to variable selection in least squares problems," *IMA Journal of Numerical Analysis*, vol. 20, no. 3, pp. 389–403, Jul. 2000.

[55] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, Apr. 2004.

[56] J. C. Pinheiro, C. Liu, and Y. N. Wu, "Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution," *Journal of Computational and Graphical Statistics*, vol. 10, no. 2, pp. 249–276, Jun. 2001.

[57] R. S. Singh, "Estimation of error variance in linear regression models with errors having multivariate student t distribution with unkown degrees of freedom," *Economics Letters*, vol. 27, no. 1, pp. 47–53, Jan. 1988.

[58] J. Luis and G. Cusumano, "A measure of total variability for the multivariate t distribution with application," *Information Sciences*, vol. 92, no. 3, pp. 47–63, Jul. 1996.

[59] W. Hu and A. N. Kercheval, "Portfolio optimization for t and skew-t returns," *Quantitative Finance*, vol. 10, no. 1, pp. 91–105, Jan. 2007.

[60] A. Panagiotelis and M. Smith, "Bayesian density forecasting of intraday electricity prices using multivariate skew t distributions," *International Journal of Forecasting*, vol. 24, no. 4, pp. 710–727, Dec. 2008.

[61] R. Huisman, C. Huurman, and R. Mahieu, "Hourly electricity prices in day-ahead markets," *Energy Economics*, vol. 29, no. 2, pp. 240–248, Mar. 2006.

[62] N. V. Karakatsani and D. W. Bunn, "Forecasting electricity prices: The impact of fundamentals and time-varying coefficients," *International Journal of Forecasting*, vol. 24, no. 4, pp. 764–785, Dec. 2008.

[63] A. K. Diongue, D. Guegan, and B. Vignal, "Forecasting electricity spot market prices with a k-factor gigarch process," *Applied Energy*, vol. 86, no. 4, pp. 505–510, Apr. 2009.

[64] M. A. Hill and W. J. Dixon, "Roubstness in real life: A study of clinical laboratory data," *Biometrics*, vol. 38, no. 2, pp. 377–396, Jun. 1982.

[65] A. Azzalini and A. Dalla-Valle, "The multivariate skew-normal distribution," *Biometrika*, vol. 83, no. 4, pp. 715–726, Dec. 1996.

[66] A. Azzalini and A. Capitanio, "Statistical applications of the multivariate skew normal distribution," *Journal of the Royal Statistical Society Series B*, vol. 61, no. 3, pp. 579–602, Jun. 1999.

[67] A. Azzalini, "The skew-normal distribution and related multivariate families," *Scandinavian Journal of Statistics*, vol. 32, no. 2, pp. 159–188, Jun. 2005.

[68] M. D. Branco and D. K. Dey, "A general class of multivariate skew elliptical distributions," *Journal of Multivariate Analysis*, vol. 79, no. 1, pp. 99–113, Oct. 2001.

[69] A. Azzalini and A. Capitanio, "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution," *Journal of the Royal Statistical Society Series B*, vol. 65, no. 2, pp. 367–389, May 2003.

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

[71] A. Azzalini, "Some recent developments in the theory of distributions and their applications," *Atti della XLIII Riunione Scientifica della Societa Italiana di Statistica*, pp. 51–64, 2006.

[72] L. Chen and M. Pourahmadi, "Robust and sparse multivariate regression with covariance estimation," *Submitted*, 2011.

[73] V. H. Lachos, P. Ghosh, and R. B. Arellano-Valle, "Likelihood based inference for skew-normal independent linear mixed models," *Statistica Sinica*, vol. 20, pp. 303–322, 2010.

[74] X. L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ecm algorithm: A general framework," *Biometrika*, vol. 80, no. 2, pp. 267–278, Jun. 1993.

[75] R. Piessen, E. deDoncker Kapenga, C. W. Uberhueber, and D. K. Kahaner, *QUADPACK, A Subroutine Package for Automatic Integration*. Berlin, Germany: Springer, 1983.

[76] N. Loperfido, "Canonical transformations of skew-normal variates," *Test*, vol. 19, no. 1, pp. 146–165, May 2010.

[77] ——, "Quadratic forms of skew-normal random vectors," *Statistics and Probability Letters*, vol. 54, no. 4, pp. 381–387, Oct. 2001.

[78] T. Anderson, "Asymptotic theory for principal component analysis," *The Annals of Mathematical Statistics*, vol. 34, no. 1, pp. 122–148, Mar. 1963.

[79] R. B. Arellano-Valle and A. Azzalini, "The centred parametrization for the multivariate skew-normal distribution," *Journal of Multivariate Analysis*, vol. 99, no. 7, pp. 1362–1382, Aug. 2008.

[80] T. Kollo and D. Rosen, *Advanced Multivariate Statistics With Matrices*. Dordrecht, The Netherlands: Springer, 2005.

[81] M. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, Mar. 1938.

[82] D. K. Hoffman, R. C. Raffenetti, and K. Reudenberg, "Generalization of euler angles to n-dimensional orthogonal matrices," *Journal of Mathematical Physics*, vol. 13, no. 4, pp. 528–532, Apr. 1972.

[83] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore and London: Johns Hopkins Press, 1996.

[84] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing,* 3rd ed. New York: Cambridge University Press, 2007.

[85] R. D. Cook and S. Weisberg, *An Introduction to Regression Graphics*. New York: Wiley, 1994.

APPENDIX A

PROOF OF PROPOSITION 1

For ease of notation, let $\theta = r = \frac{\nu}{2}$. The computation of the conditional expectations relies on the following result in [69]:

Lemma. If $W \sim \text{Gamma}(r, \theta)$, 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 (3.11) and (3.12), we have

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

and

$$
\begin{aligned}
E\left\{ZW^m|Y\right\} &= \frac{2}{f(\boldsymbol{y})}\int_0^\infty\int_0^\infty zw^m\phi\left\{z-\sqrt{w}\boldsymbol{\eta}^T(\boldsymbol{y}-\boldsymbol{\mu})\right\}\phi_q(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma}/w)\cdot \\
&\quad h(w;\nu/2,\nu/2)dzdw \\
&= \frac{2}{f(\boldsymbol{y})}\int_0^\infty w^m\phi_q(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma}/w)\cdot h(w;\nu/2,\nu/2)\cdot \\
&\quad \int_0^\infty z\phi\left\{z-\sqrt{w}\boldsymbol{\eta}^T(\boldsymbol{y}-\boldsymbol{\mu})\right\}dzdw \\
&= \frac{2}{f(\boldsymbol{y})}\int_0^\infty w^m\phi_q(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma}/w)\cdot h(w;\nu/2,\nu/2)\Bigg[\sqrt{w}\boldsymbol{\eta}^T(\boldsymbol{y}-\boldsymbol{\mu})\cdot \\
&\quad \Phi\{\sqrt{w}\boldsymbol{\eta}^T(\boldsymbol{y}-\boldsymbol{\mu})\}+\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}w(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\eta}\boldsymbol{\eta}^T(\boldsymbol{y}-\boldsymbol{\mu})\right\}\Bigg]dw \\
&= \boldsymbol{\eta}^T(\boldsymbol{y}-\boldsymbol{\mu})\cdot E\left\{W^m|Y\right\}+\frac{1}{\sqrt{2\pi}}C(\theta_2,r_1).
\end{aligned}
$$

APPENDIX B

COMPUTATION OF (3.20)

Expanding $g_c(\mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\eta})$ in (3.15)and ignoring the terms unrelated to $\mathbf{B}$ yields

$$\begin{aligned} g_c(\mathbf{B}) &= \frac{1}{n}\mathrm{tr}\{(\boldsymbol{Y}-\mathbf{XB})^T\boldsymbol{\Omega}(\boldsymbol{Y}-\mathbf{XB})\}+\frac{1}{n}\mathrm{tr}\{(\boldsymbol{Y}-\mathbf{XB})^T\boldsymbol{\eta}\boldsymbol{\eta}^T(\boldsymbol{Y}-\mathbf{XB})\} \\ &\quad -\frac{1}{n}\mathrm{tr}\left\{(\boldsymbol{Y}-\mathbf{XB})\boldsymbol{\eta}\mathbf{Z}^T+\mathbf{Z}\boldsymbol{\eta}^T(\boldsymbol{Y}-\mathbf{XB})^T\right\}. \\ &= \frac{1}{n}\mathrm{tr}\{(\boldsymbol{Y}-\mathbf{XB})^T(\boldsymbol{\Omega}+\boldsymbol{\eta}\boldsymbol{\eta}^T)(\boldsymbol{Y}-\mathbf{XB})-(\boldsymbol{Y}-\mathbf{XB})\boldsymbol{\eta}\mathbf{Z}^T \\ &\quad -\mathbf{Z}\boldsymbol{\eta}^T(\boldsymbol{Y}-\mathbf{XB})^T\} \end{aligned}$$

Then (3.20) can be obtained by completing the square for $(\boldsymbol{Y}-\mathbf{XB})$.

APPENDIX C

THE GRADIENT OF $\ell(\boldsymbol{\theta}, B)$ IN (4.11).

The calculation of the partial derivative of $\ell(\boldsymbol{\theta}, b)$ is straightforward. For ease of notation, let

$$\zeta_0(x) = \log\{2\Phi(x)\} \quad \text{and} \quad \zeta_m(x) = \frac{d^m}{dx^m}\zeta_0(x) \quad (m = 1, 2, \cdots).$$

so we have

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\theta}, b)}{\partial b} &= \sum_{i=1}^{n} \zeta_1\left\{b\gamma_j^T(\boldsymbol{\theta})z_i\right\}\gamma_j^T(\boldsymbol{\theta})z_i \\
\frac{\partial \ell(\boldsymbol{\theta}, b)}{\partial \theta_k} &= -n\sum_{i=1}^{p}\frac{\gamma_i^T(\boldsymbol{\theta})S_0\left[\Gamma(\boldsymbol{\theta}^k)\right]_i}{\gamma_i^T(\boldsymbol{\theta})S_0\gamma_i(\boldsymbol{\theta})} + b\sum_{i=1}^{n}\zeta_1\left\{b\gamma_j^T(\boldsymbol{\theta})z_i\right\}\left[\Gamma(\boldsymbol{\theta}^k)\right]_j^T z_i
\end{aligned}$$

where $\boldsymbol{\theta}^k = \left(\theta_1, \cdots, \theta_k + \frac{\pi}{2}, \cdots, \theta_Q\right)^T$ and $[A]_k$ denotes the kth column of the matrix A.

APPENDIX D

THE GRADIENT OF $\ell(\Theta_3)$ IN (4.12).

Before we calculate the partial derivatives, we introduce the $vec(\cdot)$ operator which stacks the columns of a matrix and the $v(\cdot)$ operator which only stacks the upper triangular of a matrix. We denote the $[p(p+1)/2] \times p^2$ elimination matrix by H such that $v(\Omega) = Hvec(\Omega)$. The computation of the partial derivatives is lengthy but straightforward:

$$\begin{aligned}
\frac{\partial \ell(\Theta_3)}{\partial b} &= \sum_{i=1}^{n} \zeta_1(b\gamma_j^T z_i)\gamma_j^T z_i \\
\frac{\partial \ell(\Theta_3)}{\partial \rho} &= -\frac{n}{2}\mathbf{1}_p + \frac{n}{2}\mathrm{diag}\{D \circ (ASA)^T\} + b\frac{\partial \gamma_j^T}{\partial \rho}\sum_{i=1}^{n} \zeta_1(b\gamma_j^T z_i)z_i \\
\frac{\partial \ell(\Theta_3)}{\partial A_u} &= -n(DAS_0)_u + b\frac{\partial \gamma_j^T}{\partial A_u}\sum_{i=1}^{n} \zeta_1(b\gamma_j^T z_i)z_i
\end{aligned}$$

where $\circ$ is the elementwise product, B_u is defined as $B_u = (b_{12}, b_{13}, b_{23}, \cdots, b_{1p}, \cdots, b_{p-1,p})^T$ for a $p \times p$ matrix B. To get the derivatives of γ_j^T with respective ρ and A_u, we consider the partial derivative of $vec^T(\Gamma)$ with respective ρ and A_u. By some algebra, we have

$$\begin{aligned}
\frac{\partial vec^T(\Gamma)}{\partial \rho} &= \frac{\partial v^T(V)}{\partial \rho}\frac{\partial vec^T(\Gamma)}{\partial v(V)} = D\mathrm{diag}\left\{e_1^T, e_2^T, \cdots, e_p^T\right\}(A \otimes A)H^T\frac{\partial vec^T(\Gamma)}{\partial v(V)} \\
\frac{\partial vec^T(\Gamma)}{\partial A_u} &= \frac{\partial v^T(V)}{\partial A_u}\frac{\partial vec^T(\Gamma)}{\partial v(V)} \\
&= \left[\frac{\partial vec^T(A^T)}{\partial A_u}\{(DA) \otimes I_p\} + \frac{\partial vec^T(A)}{\partial A_u}\{I_p \otimes (DA)\}\right]H^T\frac{\partial vec^T(\Gamma)}{\partial v(V)}
\end{aligned}$$

where $V = \Omega^{-1}$, $\otimes$ is the Kronecker product and $\frac{\partial vec^T(\Gamma)}{\partial v(V)}$ can be determined by the Lemma 3.1.4 [80, pp. 295] since Γ is also the eigenvector matrix of V. Therefore,

$\frac{\partial \gamma_j^T}{\partial \rho}$ and $\frac{\partial v^T(V)}{\partial A_u}$ are the matrices consisting of p columns, from $(j-1)p+1$ to $j \cdot p$, of $\frac{\partial vec^T(\Gamma)}{\partial \rho}$ and $\frac{\partial vec^T(\Gamma)}{\partial A_u}$, respectively.

VITA

Lianfu Chen was born in Fujian, China. He received his Bachelor of Science degree in Statistics in 2006 from University of Science and Technology of China, Hefei. He was admitted to the graduate program in the Department of Statistics at Texas A&M University in 2006. He received his Master's degree in December 2008, and Ph.D. in December 2011. His permanent address is:

Guanfeng South Avenue #85

Shenzhen Village, Neikeng County, Jinjiang, Fujian