JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 516, 1726–1735, Theory and Methods
http://dx.doi.org/10.1080/01621459.2015.1113973

Taylor & Francis
Taylor & Francis Group

Multitask Quantile Regression Under the Transnormal Model

Jianqing Fan[a], Lingzhou Xue[b], and Hui Zou[c]

[a]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, USA; [b]Department of Statistics, Penn State University, University Park, PA, USA; [c]School of Statistics, University of Minnesota, Minneapolis, MN, USA

ABSTRACT

We consider estimating multitask quantile regression under the transnormal model, with focus on high-dimensional setting. We derive a surprisingly simple closed-form solution through rank-based covariance regularization. In particular, we propose the rank-based ℓ_1 penalization with positive-definite constraints for estimating sparse covariance matrices, and the rank-based banded Cholesky decomposition regularization for estimating banded precision matrices. By taking advantage of the alternating direction method of multipliers, nearest correlation matrix projection is introduced that inherits sampling properties of the unprojected one. Our work combines strengths of quantile regression and rank-based covariance regularization to simultaneously deal with nonlinearity and nonnormality for high-dimensional regression. Furthermore, the proposed method strikes a good balance between robustness and efficiency, achieves the "oracle"-like convergence rate, and provides the provable prediction interval under the high-dimensional setting. The finite-sample performance of the proposed method is also examined. The performance of our proposed rank-based method is demonstrated in a real application to analyze the protein mass spectroscopy data. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received December 2013
Revised July 2015

KEYWORDS
Copula model; Optimal transformation; Rank correlation; Cholesky decomposition; Quantile regression; Prediction interval; Alternating direction method of multipliers

1. Introduction

Consider a multitask high-dimensional learning paradigm that independent variables $\boldsymbol{z} = (z_1, \ldots, z_{p_0})'$ are used to simultaneously predict multiple response variables $\boldsymbol{y} = (y_1, \ldots, y_{q_0})'$. Here, dimensions p_0 and q_0 can be of larger order of magnitude than sample size n. This work is motivated by obtaining the optimal prediction $\boldsymbol{t}(\boldsymbol{z}) = (t_1(\boldsymbol{z}), \ldots, t_{q_0}(\boldsymbol{z}))'$ for $\boldsymbol{y}$. To this end, we shall solve optimal transformations $\boldsymbol{t}$ from

$$\min_{\boldsymbol{t}:\, \mathbb{R}^{p_0} \mapsto \mathbb{R}^{q_0}} \sum_{j=1}^{q_0} \mathrm{E}[L(y_j - t_j(\boldsymbol{z}))],$$

where $L(\cdot)$ is some convex loss function. If $\boldsymbol{z}$ and $\boldsymbol{y}$ have a joint normal distribution, nice properties such as linearity and homoscedasticity hold for $\boldsymbol{t}$. Under normality, it is appropriate to specify $L(\cdot)$ as the squared loss, that is, $L(u) = u^2$. Thus, normality makes the neat connection between optimal prediction and ordinary least squares, and optimal transformations can be easily obtained by solving ordinary least squares.

However, observed data are often skewed or heavy-tailed, and rarely normally distributed in real-world applications. Transformations are usually used to achieve normality in regression analysis. Under the traditional low-dimensional setting, estimating transformations for regression has received considerable attention. On one hand, parametric methods focus on the parametric families of transformations, such as the celebrated Box–Cox transformation (Box and Cox 1964). On the other hand, nonparametric estimation of regression transformations is also studied, for instance, projection pursuit regression (Friedman and Stuetzle 1981), alternating conditional expectation (Breiman and Friedman 1985), additivity and variance stabilization (Stone 1985; Tibshirani 1988), among others. However, it is nontrivial to extend them to the high-dimensional setting, since these methods would suffer from the curse of dimensionality. Without requiring the restricted normality, there are significant demands for estimating optimal transformations for high-dimensional regression.

As a nice combination of flexibility and interpretability, the Gaussian copulas have generated a lot of interests in statistics and econometrics, and are deemed as a favorable alternative to the Gaussian model in high-dimensional statistical problems, including linear or quadratic discriminant analysis (Lin and Jeon 2003; Mai and Zou 2015; Fan et al. 2013), graphical modeling (Liu, Lafferty, and Wasserman 2009; Liu et al. 2012; Xue and Zou 2012), covariance matrix estimation (Xue and Zou 2014; Wegkamp and Zhao 2016), and principal component analysis (Han and Liu 2012). The semiparametric Gaussian copula provides a semiparametric generalization of the Gaussian model by assuming the existence of univariate monotone transformations $\boldsymbol{f}$ for $\boldsymbol{x} \in \mathbb{R}^p$ such that $\boldsymbol{f}(\boldsymbol{x}) \sim N_p(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star)$. Throughout this article, we follow Lin and Jeon (2003) to call this copula model as the transnormal model, which is also called the nonparanormal model in Liu, Lafferty, and Wasserman (2009). We will show the power of the transnormal model in estimating optimal transformations for high-dimensional regression. We suppose that $\boldsymbol{x} = (\boldsymbol{z}', \boldsymbol{y}')'$ consists of $\boldsymbol{z} = (x_1, \ldots, x_{p_0})'$ and $\boldsymbol{y} = (x_{p_0+1}, \ldots, x_{p_0+q_0=p})'$. Denote by F_j the marginal cumulative distribution function (CDF) of x_j and $f_j = \Phi^{-1} \cdot F_j$ for $j = 1, \ldots, p_0 + q_0$. Let $\boldsymbol{g} =$

CONTACT Hui Zou zouxx019@umn.edu School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church St. SE, Minneapolis, MN 55455.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

$(f_1, \ldots, f_{p_0})'$ and $\boldsymbol{h} = (f_{p_0+1}, \ldots, f_{p_0+q_0})'$. The transnormal model entails the existence of monotone transformations $\boldsymbol{f} = (\boldsymbol{g}', \boldsymbol{h}')$ such that

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{g}(\boldsymbol{z}) \\ \boldsymbol{h}(\boldsymbol{y}) \end{pmatrix} \sim N_p \left(\boldsymbol{\mu}^\star = \begin{pmatrix} \boldsymbol{\mu}_z^\star \\ \boldsymbol{\mu}_y^\star \end{pmatrix}, \boldsymbol{\Sigma}^\star = \begin{pmatrix} \boldsymbol{\Sigma}_{zz}^\star & \boldsymbol{\Sigma}_{zy}^\star \\ \boldsymbol{\Sigma}_{yz}^\star & \boldsymbol{\Sigma}_{yy}^\star \end{pmatrix} \right), \tag{1}$$

where we may assume that $\boldsymbol{\mu}^\star = \boldsymbol{0}$ and $\boldsymbol{\Sigma}^\star$ is a correlation matrix. The transnormal model essentially assumes the joint normality of marginally normal-transformed variables, and it strikes a balance between model robustness and interpretability. This work aims to estimate transformations of predictors $\boldsymbol{t}(\boldsymbol{z})$ to optimally predict $\boldsymbol{y}$ under the transnormal model (1).

Given transformations $\boldsymbol{g}$ and $\boldsymbol{h}$, we derive the conditional distribution of $\boldsymbol{h}(\boldsymbol{y})$ given $\boldsymbol{g}(\boldsymbol{z})$ as

$$\boldsymbol{h}(\boldsymbol{y})|\boldsymbol{g}(\boldsymbol{z}) \sim N_{q_0}\big(\boldsymbol{\Sigma}_{yz}^\star (\boldsymbol{\Sigma}_{zz}^\star)^{-1} \cdot \boldsymbol{g}(\boldsymbol{z}),\ \boldsymbol{\Sigma}_{yy}^\star - \boldsymbol{\Sigma}_{yz}^\star (\boldsymbol{\Sigma}_{zz}^\star)^{-1} \boldsymbol{\Sigma}_{zy}^\star\big). \tag{2}$$

But, since both $\boldsymbol{g}$ and $\boldsymbol{h}$ are unknown, it is challenging to obtain the explicit conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{z}$. Moreover, linearity and homoscedasticity do not hold for the optimal transformations under the transnormal model. In the presence of nonlinear $\boldsymbol{g}$ and $\boldsymbol{h}$, we should take into account the coordinate-wise heterogeneity among the conditional distributions of $\boldsymbol{y}$ given $\boldsymbol{z}$. To this end, we employ quantile regression to handle such heterogeneity.

Quantile regression is introduced by the seminal article of Koenker and Bassett (1978), and it has received much attention in various topics such as survival analysis (Koenker and Geling 2001), time series analysis (Koenker and Xiao 2006), microarray analysis (Wang and He 2007), and variable selection (Zou and Yuan 2008; Bradic, Fan and Wang 2011). Let $\rho_\tau(u) = u \cdot (\tau - I_{\{u \le 0\}})$ be the check loss function. Then we consider the multitask quantile regression:

$$\min_{\boldsymbol{t}:\, \mathbb{R}^{p_0} \mapsto \mathbb{R}^{q_0}} \sum_{j=1}^{q_0} \mathrm{E}[\rho_\tau(y_j - t_j(\boldsymbol{z}))]. \tag{3}$$

When $\tau = \frac{1}{2}$, $\rho_{\tau=\frac{1}{2}}(u) = |u|$, and (3) reduces to multitask median regression, namely,

$$\min_{\boldsymbol{t}:\, \mathbb{R}^{p_0} \mapsto \mathbb{R}^{q_0}} \sum_{j=1}^{q_0} \mathrm{E}[\rho_{\frac{1}{2}}(y_j - t_j(\boldsymbol{z}))] = \frac{1}{2} \min_{\boldsymbol{t}:\, \mathbb{R}^{p_0} \mapsto \mathbb{R}^{q_0}} \mathrm{E}\|\boldsymbol{y} - \boldsymbol{t}(\boldsymbol{z})\|_{\ell_1}.$$

The use of the ℓ_1 loss in prediction was recommended in Friedman (2001, 2002). Our work shares the similar philosophy and includes multitask median regression as a special case.

In this article, we will show that optimal transformations in multitask quantile regression can be efficiently estimated under the transnormal model. Although estimating optimal transformation is difficult, we derive the surprisingly closed-form optimal prediction without using any smoothing techniques as in Friedman and Stuetzle (1981), Breiman and Friedman (1985), Stone (1985), or Tibshirani (1988). The key ingredient of our proposed method is the positive-definite regularized estimation of large covariance matrices under the transnormal model. We propose two novel rank-based covariance regularization methods: the *rank-based positive definite ℓ_1 penalization* for estimating sparse covariance matrices, and the *rank-based banded Cholesky decomposition regularization* for estimating banded inverse covariance matrices. Our proposed rank-based covariance regularization critically depends on a positive definite correlation matrix that retains the desired sampling properties of the adjusted Spearman's or Kendall's rank correlation matrix (Kendall 1948; Liu et al. 2012; Xue and Zou 2012), which can be solved by an efficient alternating direction method of multipliers. By combining both strengths of quantile regression and rank-based covariance regularization, we simultaneously address issues of nonlinearity, nonnormality and high dimensionality in estimating optimal transformations for high-dimensional regression. Especially, our proposed method achieves the "oracle"-like convergence rate and provides a provable prediction interval in high dimensions, where dimension is on the nearly exponential order of sample size.

The rest of this article is organized as follows. We first present the methodological details of optimal prediction in multitask quantile regression in Section 2. Section 3 establishes the theoretical properties of our proposed method under the transnormal model. Section 4 contains simulation results and a real application to analyze the protein mass spectroscopy data. Technical proofs are presented in the online Appendices.

2. Multitask Quantile Regression

This section presents the optimal prediction in multitask quantile regression under the transnormal model. The transnormal family of distributions retains the nice interpretation of the normal model, and enables us to make a good use of normal model and theory. In particular, it allows us to obtain the closed-form optimal prediction, which is very appealing to deal with nonnormality and high dimensionality. Moreover, the monotone transformation is easy to handle for quantile estimation.

2.1. The Closed-Form Solution

Let $Q_\tau(y_j|\boldsymbol{z})$ be the τth conditional quantile of y_j given $\boldsymbol{z}$. Then, it immediately implies that $Q_\tau(y_j|\boldsymbol{z}) = \min_{t_j} \mathrm{E}[\rho_\tau(y_j - t_j(\boldsymbol{z}))]$. Define $\boldsymbol{Q}_\tau(\boldsymbol{y}|\boldsymbol{z}) = (Q_\tau(y_1|\boldsymbol{z}), \ldots, Q_\tau(y_p|\boldsymbol{z}))'$. The τth equicoordinate conditional quantile $\boldsymbol{Q}_\tau(\boldsymbol{y}|\boldsymbol{z})$ is the analytical solution to (3), that is,

$$\boldsymbol{Q}_\tau(\boldsymbol{y}|\boldsymbol{z}) = \arg \min_{\boldsymbol{t}:\, \mathbb{R}^{p_0} \mapsto \mathbb{R}^{q_0}} \sum_{j=1}^{q_0} \mathrm{E}[\rho_\tau(y_j - t_j(\boldsymbol{z}))].$$

By using the fact that $\boldsymbol{h}$ and $\boldsymbol{g}$ are monotone under the transnormal model, we have

$$\boldsymbol{Q}_\tau(\boldsymbol{y}|\boldsymbol{z}) = \boldsymbol{Q}_\tau(\boldsymbol{y}|\boldsymbol{g}(\boldsymbol{z})) = \boldsymbol{h}^{-1}(\boldsymbol{Q}_\tau(\boldsymbol{h}(\boldsymbol{y})|\boldsymbol{g}(\boldsymbol{z}))).$$

Since $\boldsymbol{h}$ is monotonically nondecreasing, it now follows from (2) that

$$\boldsymbol{Q}_\tau(\boldsymbol{y}|\boldsymbol{z}) = \boldsymbol{h}^{-1}\big(\boldsymbol{\Sigma}_{yz}^\star (\boldsymbol{\Sigma}_{zz}^\star)^{-1} \cdot \boldsymbol{g}(\boldsymbol{z}) + \Phi^{-1}(\tau)\ \mathrm{vdiag}(\boldsymbol{\Sigma}_{yy}^\star - \boldsymbol{\Sigma}_{yz}^\star (\boldsymbol{\Sigma}_{zz}^\star)^{-1} \boldsymbol{\Sigma}_{zy}^\star)\big), \tag{4}$$

where $\Phi(\cdot)$ is the CDF of $N(0, 1)$, and $\mathrm{vdiag}(\boldsymbol{A})$ denotes the vector formed by the diagonal element of $\boldsymbol{A}$. Therefore, the transnormal model enables us to obtain the closed-form solution to multitask quantile regression.

Moreover, the closed-form solution (4) can be used to construct prediction intervals for predicting $\boldsymbol{y}$ given $\boldsymbol{z}$ in high

dimensions. To be more specific, we can obtain the closed-form $100(1-\tau)\%$ prediction interval as

$$[\boldsymbol{Q}_{\frac{\tau}{2}}(\boldsymbol{y}|\boldsymbol{z}), \boldsymbol{Q}_{1-\frac{\tau}{2}}(\boldsymbol{y}|\boldsymbol{z})].$$

For different values of τ, we only need to adjust the value $\Phi(\tau)$. This is also an appealing feature of the transnormal model, especially under the high-dimensional setting.

The solution $\boldsymbol{Q}_\tau(\boldsymbol{y}|\boldsymbol{z})$ uses true covariance matrices and transformations, and it is not a feasible estimator. To use (4), we need to estimate the covariance matrix $\boldsymbol{\Sigma}^\star$ and transformation $\boldsymbol{f}$. It turns out that these two tasks are relatively easy under the transnormal model. Section 2.2 describes how to estimate transformation $\boldsymbol{f}$, and then Section 3 shows how to estimate the large structured covariance matrix $\boldsymbol{\Sigma}^\star$. With estimated transformation $\hat{\boldsymbol{f}} = (\hat{\boldsymbol{g}}, \hat{\boldsymbol{h}})$ and covariance $\hat{\boldsymbol{\Sigma}}$, it is straightforward to obtain the following plug-in estimator

$$\hat{\boldsymbol{Q}}_\tau(\boldsymbol{y}|\boldsymbol{z}) = \hat{\boldsymbol{h}}^{-1}\big(\hat{\boldsymbol{\Sigma}}_{yz}(\hat{\boldsymbol{\Sigma}}_{zz})^{-1}\cdot\hat{\boldsymbol{g}}(\boldsymbol{z}) + \Phi^{-1}(\tau)\ \mathrm{vdiag}(\hat{\boldsymbol{\Sigma}}_{yy} - \hat{\boldsymbol{\Sigma}}_{yz}(\hat{\boldsymbol{\Sigma}}_{zz})^{-1}\hat{\boldsymbol{\Sigma}}_{zy})\big). \quad (5)$$

Given the plug-in estimator (5), we further estimate the $100(1-\tau)\%$ prediction interval as

$$[\hat{\boldsymbol{Q}}_{\frac{\tau}{2}}(\boldsymbol{y}|\boldsymbol{z}), \hat{\boldsymbol{Q}}_{1-\frac{\tau}{2}}(\boldsymbol{y}|\boldsymbol{z})].$$

Remark 1. When $\tau = \frac{1}{2}$, it reduces to optimal prediction in multitask median regression. Since $\Phi^{-1}(\frac{1}{2}) = 0$, the optimal prediction can be further simplified as

$$\boldsymbol{Q}_{\frac{1}{2}}(\boldsymbol{y}|\boldsymbol{z}) = \boldsymbol{h}^{-1}\big(\boldsymbol{\Sigma}^\star_{yz}(\boldsymbol{\Sigma}^\star_{zz})^{-1}\cdot\boldsymbol{g}(\boldsymbol{z})\big),$$

and

$$\hat{\boldsymbol{Q}}_{\frac{1}{2}}(\boldsymbol{y}|\boldsymbol{z}) = \hat{\boldsymbol{h}}^{-1}\big(\hat{\boldsymbol{\Sigma}}^\star_{yz}(\hat{\boldsymbol{\Sigma}}^\star_{zz})^{-1}\cdot\hat{\boldsymbol{g}}(\boldsymbol{z})\big).$$

Remark 2. Compared to multitask mean regression, multitask quantile regression (including median regression) is more robust against outliers in measurements, in addition to delivering the closed-form solution (4). In contrast, by solving ordinary least squares, multitask mean regression is

$$\mathrm{E}(\boldsymbol{y}|\boldsymbol{z}) = \mathrm{E}(\boldsymbol{y}|\boldsymbol{g}(\boldsymbol{z})) = \mathrm{E}(\boldsymbol{h}^{-1}(\boldsymbol{h}(\boldsymbol{y}))|\boldsymbol{g}(\boldsymbol{z})).$$

But unlike the quantile regression, this cannot be simplified further unless $\boldsymbol{h}$ is linear.

Remark 3. Often $\boldsymbol{h}(\cdot)$ in (4) is nonlinear. The nonlinearity of $\boldsymbol{h}(\cdot)$ makes the difference of conditional quantiles at different values of τ depend on $\boldsymbol{z}$ and thus models the effect of heteroscedasticy. In contrast, Wu, Yu and Yu (2010), Zhu, Huang and Li (2012), and Fan et al. (2013) used the single-index or semiparametric quantile regression to model the heteroscedastic effect, by imposing the model $y = \mu(\boldsymbol{z}'\boldsymbol{\beta}) + \sigma(\boldsymbol{z}'\boldsymbol{\beta})\cdot\varepsilon$, where $\sigma(\boldsymbol{z}'\boldsymbol{\beta})\cdot\varepsilon$ is a heteroscedastic error and ε is independent of $\boldsymbol{z}$. Unlike the closed-form expression (4), it would be difficult to employ semiparametric quantile regression to handle nonlinearity and high dimensionality.

2.2. Estimation of Transformation

Note that $f_j(x_j) \sim N(0,1)$ under the transnormal model for any j. Hence, the cumulative distribution function of X_j admits the form

$$F_j(x_j) = \Phi(f_j(x_j)), \quad \text{or} \quad f_j(x_j) = \Phi^{-1}(F_j(x_j)). \quad (6)$$

Equation (6) provides an intuitive way to estimate the transformation function f_j. Define $\tilde{F}_j(u) = \frac{1}{n}\sum_{i=1}^n I_{\{x_{ij}\le u\}}$ as the empirical distribution function. We consider the Winsorized empirical distribution function $\hat{F}_j(\cdot)$ as follows:

$$\hat{F}_j(u) = \delta_n\cdot I_{\{\tilde{F}_j(u)<\delta_n\}} + \tilde{F}_j(u)\cdot I_{\{\delta_n\le\tilde{F}_j(u)\le 1-\delta_n\}} + (1-\delta_n)\cdot I_{\{\tilde{F}_j(u)>1-\delta_n\}}, \quad (7)$$

where δ_n is the Winsorization parameter to avoid infinity values and to achieve better bias-variance tradeoff. Following Liu, Lafferty, and Wasserman (2009) and Mai and Zou (2015), we specify the Winsorization parameter as $\delta_n = \frac{1}{n^2}$, which facilitates both theoretical analysis and practical performance. Next, we estimate the transformation $\boldsymbol{f}$ in the transnormal model as follows:

$$\hat{\boldsymbol{f}} = (\hat{f}_1, \ldots, \hat{f}_p) = (\Phi^{-1}\circ\hat{F}_1, \ldots, \Phi^{-1}\circ\hat{F}_p).$$

Remark that $\hat{f}_1, \ldots, \hat{f}_p$ are nondecreasing, and they can be substituted into (5) to estimate optimal transformations.

2.3. Estimation of Correlation Matrix

This section presents two covariance regularization methods based on the *i.i.d.* transnormal data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$: the *rank-based positive definite* ℓ_1 *penalization* for estimating sparse correlation matrices, and the *rank-based banded Cholesky decomposition regularization* for estimating banded precision matrices. Both estimates achieve the critical positive definiteness, and they can be used in (5) to obtain the optimal prediction.

2.3.1. Positive Definite Sparse Correlation Matrix

Sparse correlation matrices are widely used in many applications, where variables are permutation invariant. By truncating small entries to zero, thresholding (Bickel and Levina 2008b; Fan, Liao, and Mincheva 2013) is a powerful approach to encourage (conditional) sparsity when estimating large correlation matrices. However, the resulting sparse estimator may not be positive definite in practice. Xue, Ma, and Zou (2012) proposed an efficient positive-definite ℓ_1-penalized estimator to address the indefiniteness issue of thresholding. Now, we extend the positive-definite ℓ_1-penalized estimator to the transnormal model.

First, we introduce the "oracle" positive-definite ℓ_1-penalized estimator using the "oracle" transformations $\boldsymbol{f}$, that is,

$$\hat{\boldsymbol{\Sigma}}^o_{\ell_1} = \arg\min_{\boldsymbol{\Sigma}} \frac{1}{2}\|\boldsymbol{\Sigma} - \hat{\boldsymbol{R}}^o\|_F^2 + \lambda\|\boldsymbol{\Sigma}\|_{1,\mathrm{off}} \quad \text{subject to} \quad \mathrm{diag}(\boldsymbol{\Sigma}) = 1;\ \boldsymbol{\Sigma}\succeq\epsilon\boldsymbol{I}$$

where $\hat{\boldsymbol{R}}^o$ is the sample correlation matrix of "oracle" data $\boldsymbol{f}(\boldsymbol{x}_1), \ldots, \boldsymbol{f}(\boldsymbol{x}_n)$, $\|\boldsymbol{\Sigma}\|_{1,\mathrm{off}}$ is the ℓ_1 norm of off-diagonal elements in $\boldsymbol{\Sigma}$, and ϵ is some arbitrarily small constant (say, $\epsilon = 10^{-4}$) satisfying that $\lambda_{\min}(\boldsymbol{\Sigma}^\star) \ge \epsilon$.

Motivated by $\hat{\boldsymbol{\Sigma}}^o_{\ell_1}$, we only need to derive a comparable alternative to $\hat{\boldsymbol{R}}^o$. For space consideration, we focus on

the adjusted Spearman's rank correlation throughout this article, since the same analysis can be easily adapted to the adjusted Kendall's rank correlation. Let $\boldsymbol{s}_j = (s_{1j}, s_{2j}, \ldots, s_{nj})'$ be the ranks for $(x_{1j}, x_{2j}, \ldots, x_{nj})'$. Denote by $\hat{r}_{jl} = \mathrm{corr}(\boldsymbol{s}_j, \boldsymbol{s}_l)$ the Spearman's rank correlation, and by $\tilde{r}^s_{jl} = 2\sin(\frac{\pi}{6}\hat{r}_{jl})$ the adjusted Spearman's rank correlation. It is well-known that $\tilde{r}^s_{jl}$ corrects the bias of $\hat{r}_{jl}$ (Kendall 1948). Now we consider the rank correlation matrix $\tilde{\boldsymbol{R}}^s = (\tilde{r}^s_{jl})_{p\times p}$, which does not require estimating any transformation function. Thus, the rank-based positive-definite ℓ_1 penalization is as follows:

$$\hat{\boldsymbol{\Sigma}}^s_{\ell_1} = \arg\min_{\boldsymbol{\Sigma}} \frac{1}{2}\|\boldsymbol{\Sigma} - \tilde{\boldsymbol{R}}^s\|_F^2 + \lambda\|\boldsymbol{\Sigma}\|_{1,\mathrm{off}} \quad \text{subject to} \quad \mathrm{diag}(\boldsymbol{\Sigma}) = 1;\ \boldsymbol{\Sigma} \succeq \epsilon \boldsymbol{I}. \tag{8}$$

Remark that $\hat{\boldsymbol{\Sigma}}^s_{\ell_1}$ will guarantee the critical positive definiteness, and it can be efficiently solved by the alternating direction method of multiplier in Xue, Ma, and Zou (2012).

2.3.2. Banded Cholesky Decomposition Regularization

When an ordering exists among variables, the bandable structure is used in estimating large covariance matrices. With guaranteed positive definiteness, banded Cholesky decomposition regularization receives much attention; for example, Wu and Pourahmadi (2003); Huang et al. (2006); Bickel and Levina (2008a); and Levina, Rothman, and Zhu (2008).

To motivate our proposal, we first present the "oracle" estimator under the transnormal model. Based on "oracle" data $\boldsymbol{f}(\boldsymbol{x}_1), \ldots, \boldsymbol{f}(\boldsymbol{x}_n)$, the "oracle" banded Cholesky decomposition regularization estimates $\boldsymbol{\Sigma}^\star$ through banding the Cholesky factor of its inverse $\boldsymbol{\Theta}^\star$. Suppose $\boldsymbol{\Theta}^\star$ has the decomposition $\boldsymbol{\Theta}^\star = (\boldsymbol{I} - \boldsymbol{A})'\boldsymbol{D}^{-1}(\boldsymbol{I} - \boldsymbol{A})$, where $\boldsymbol{D}$ is a diagonal matrix and $\boldsymbol{A} = (a_{jl})_{p\times p}$ is a lower triangular matrix with zero diagonal. Let $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p)'$. Since $\boldsymbol{f}(\boldsymbol{x}) \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma}^\star)$, we have $(\boldsymbol{I} - \boldsymbol{A}) \cdot \boldsymbol{f}(\boldsymbol{x}) \sim N_p(\boldsymbol{0}, \boldsymbol{D})$. As in Bickel and Levina (2008a), the "oracle" estimator is derived by regressing $f_j(x_j)$ on its closest $\min\{k, j-1\}$ predecessors, that is, $\hat{\boldsymbol{a}}^o_1 = \boldsymbol{0}$, and for $j = 2, \ldots, p$, $\hat{\boldsymbol{a}}^o_j = \arg\min_{\boldsymbol{a}_j \in \mathcal{A}_j(k)} \frac{1}{n}\sum_{i=1}^n (f_j(x_{ij}) - \boldsymbol{a}_j'\boldsymbol{f}(\boldsymbol{x}_i))^2$, where $\mathcal{A}_j(k) = \{(\alpha_1, \ldots, \alpha_p)' : \alpha_l = 0 \text{ if } l < j-k \text{ or } l \geq j\}$. Then, $\boldsymbol{A}$ is estimated by the k-banded lower triangular matrix $\hat{\boldsymbol{A}}^o = (\hat{\boldsymbol{a}}^o_1, \ldots, \hat{\boldsymbol{a}}^o_p)'$, and $\boldsymbol{D}$ is estimated by $\hat{\boldsymbol{D}}^o = \mathrm{diag}(\hat{d}^o_1, \ldots, \hat{d}^o_p)$ with the residual sum of squares $\hat{d}^o_j = \frac{1}{n}\sum_{i=1}^n (f_j(x_{ij}) - (\hat{\boldsymbol{a}}^o_j)'\boldsymbol{f}(\boldsymbol{x}_i))^2$. Hence, the "oracle" estimator ends up with a positive-definite estimator $\hat{\boldsymbol{\Sigma}}^o_{\mathrm{chol}} = (\boldsymbol{I} - \hat{\boldsymbol{A}}^o)^{-1}(\hat{\boldsymbol{D}}^o)^{-1}[(\boldsymbol{I} - \hat{\boldsymbol{A}}^o)']^{-1}$, which has the k-banded precision matrix $\hat{\boldsymbol{\Theta}}^o_{\mathrm{chol}} = (\boldsymbol{I} - \hat{\boldsymbol{A}}^o)'(\hat{\boldsymbol{D}}^o)^{-1}(\boldsymbol{I} - \hat{\boldsymbol{A}}^o)$.

It is important to observe that $\hat{\boldsymbol{R}}^o = \frac{1}{n}\sum_{i=1}^n \boldsymbol{f}(\boldsymbol{x}_i)(\boldsymbol{f}(\boldsymbol{x}_i))'$ plays the central role there. To see this point, we notice that estimating $\hat{\boldsymbol{a}}^o_j$ and $\hat{d}^o_j$ only depends on the quadratic term

$$\frac{1}{n}\sum_{i=1}^n \big(f_j(x_{ij}) - \boldsymbol{a}_j'\boldsymbol{f}(\boldsymbol{x}_i)\big)^2 = \boldsymbol{a}_j'\hat{\boldsymbol{R}}^o\boldsymbol{a}_j - 2\boldsymbol{a}_j'\hat{\boldsymbol{r}}^o_j + \hat{r}^o_{jj}, \tag{9}$$

where $\hat{\boldsymbol{R}}^o = (\hat{r}^o_{jl})_{p\times p}$ and $\hat{\boldsymbol{r}}^o_j = (\hat{r}^o_{j1}, \ldots, \hat{r}^o_{jp})'$ is its jth row. Thus, we only need a positive-definite comparable alternative to $\hat{\boldsymbol{R}}^o$. The adjusted rank correlation matrix $\tilde{\boldsymbol{R}}^s$ achieves the "oracle"-like rate of convergence, but it is not guaranteed to be positive definite (Devlin, Gnanadesikan and Kettenring 1975). We employ the nearest correlation matrix projection to inherit the desired sampling properties of $\tilde{\boldsymbol{R}}^s$ by solving the following optimization problem:

$$\hat{\boldsymbol{R}}^s = \arg\min_{\boldsymbol{R}} \|\boldsymbol{R} - \tilde{\boldsymbol{R}}^s\|_{\max} \quad \text{subject to} \quad \boldsymbol{R} \succeq \epsilon \boldsymbol{I};\ \mathrm{diag}(\boldsymbol{R}) = 1, \tag{10}$$

where $\|\cdot\|_{\max}$ is the entrywise ℓ_∞ norm and $\epsilon > 0$ is some arbitrarily small constant satisfying that $\lambda_{\min}(\boldsymbol{\Sigma}^\star) \geq \epsilon$. Zhao, Roeder, and Liu (2012) considered a related projection, but used a smooth surrogate function $\|\boldsymbol{R} - \tilde{\boldsymbol{R}}^s\|^\nu_{\max} = \max_{\|\boldsymbol{U}\|_1 \leq 1}\langle \boldsymbol{U}, \boldsymbol{R} - \tilde{\boldsymbol{R}}^s\rangle - \frac{\nu}{2}\|\boldsymbol{U}\|_F^2$, which would introduce approximation error. Qi and Sun (2006) solved a nearest correlation matrix projection under the Frobenius norm. We use the entrywise ℓ_∞-norm for theoretical considerations, as we now demonstrate. Since $\boldsymbol{\Sigma}^\star$ is feasible to (10), we have $\|\hat{\boldsymbol{R}}^s - \tilde{\boldsymbol{R}}^s\|_{\max} \leq \|\boldsymbol{\Sigma}^\star - \tilde{\boldsymbol{R}}^s\|_{\max}$. By using the triangular inequality, $\hat{\boldsymbol{R}}^s$ retains the desired sampling properties as $\tilde{\boldsymbol{R}}^s$, that is,

$$\begin{aligned}\|\hat{\boldsymbol{R}}^s - \boldsymbol{\Sigma}^\star\|_{\max} &\leq \|\hat{\boldsymbol{R}}^s - \tilde{\boldsymbol{R}}^s\|_{\max} + \|\tilde{\boldsymbol{R}}^s - \boldsymbol{\Sigma}^\star\|_{\max} \\ &\leq 2\|\tilde{\boldsymbol{R}}^s - \boldsymbol{\Sigma}^\star\|_{\max}.\end{aligned}$$

The details of nearest correlation matrix projection are presented in the online Appendix.

Now, we propose the rank-based estimator on the basis of $\hat{\boldsymbol{R}}^s$. Let $\hat{\boldsymbol{R}}^s = (\hat{r}^s_{jl})_{p\times p}$ and $\hat{\boldsymbol{r}}^s_j = (\hat{r}^s_{j1}, \ldots, \hat{r}^s_{jp})'$. We substitute $\hat{\boldsymbol{R}}^s$ into the quadratic term (9) as

$$\boldsymbol{a}_j'\hat{\boldsymbol{R}}^s\boldsymbol{a}_j - 2\boldsymbol{a}_j'\hat{\boldsymbol{r}}^s_j + \hat{r}^s_{jj} = \boldsymbol{a}_j'\hat{\boldsymbol{R}}^s\boldsymbol{a}_j - 2\boldsymbol{a}_j'\hat{\boldsymbol{r}}^s_j + 1.$$

Accordingly, we estimate $\boldsymbol{A}$ by

$$\hat{\boldsymbol{A}}^s = \big(\hat{\boldsymbol{a}}^s_1, \ldots, \hat{\boldsymbol{a}}^s_p\big)',$$

with $\hat{\boldsymbol{a}}^s_1 = \boldsymbol{0}$, and for $j = 2, \ldots, p$,

$$\hat{\boldsymbol{a}}^s_j = \arg\min_{\boldsymbol{a}_j \in \mathcal{A}_j(k)} \boldsymbol{a}_j'\hat{\boldsymbol{R}}^s\boldsymbol{a}_j - 2\boldsymbol{a}_j'\hat{\boldsymbol{r}}^s_j + 1,$$

where $\mathcal{A}_j(k) = \{(\alpha_1, \ldots, \alpha_p) : \alpha_i = 0 \text{ if } i < j-k \text{ or } i \geq j\}$. In addition, we estimate $\boldsymbol{D}$ by

$$\hat{\boldsymbol{D}}^s = \mathrm{diag}\big(\hat{d}^s_1, \ldots, \hat{d}^s_p\big),$$

with

$$\hat{d}^s_j = \big(\hat{\boldsymbol{a}}^s_j\big)'\hat{\boldsymbol{R}}^s\hat{\boldsymbol{a}}^s_j - 2\big(\hat{\boldsymbol{a}}^s_j\big)'\hat{\boldsymbol{r}}^s_j + 1.$$

Thus, the rank-based banded Cholesky decomposition regularization yields

$$\hat{\boldsymbol{\Sigma}}^s_{\mathrm{chol}} = (\boldsymbol{I} - \hat{\boldsymbol{A}}^s)^{-1}\hat{\boldsymbol{D}}^s[(\boldsymbol{I} - \hat{\boldsymbol{A}}^s)']^{-1},$$

which has the k-banded precision matrix $\hat{\boldsymbol{\Theta}}^s_{\mathrm{chol}} = (\boldsymbol{I} - \hat{\boldsymbol{A}}^s)'(\hat{\boldsymbol{D}}^s)^{-1}(\boldsymbol{I} - \hat{\boldsymbol{A}}^s)$.

3. Theoretical Properties

This section presents theoretical properties of our proposed methods. We use several matrix norms: the ℓ_1 norm $\|\boldsymbol{U}\|_{\ell_1} = \max_j \sum_i |u_{ij}|$, the spectral norm $\|\boldsymbol{U}\|_{\ell_2} = \lambda_{\max}^{1/2}(\boldsymbol{U}'\boldsymbol{U})$, and the ℓ_∞ norm $\|\boldsymbol{U}\|_{\ell_\infty} = \max_i \sum_j |u_{ij}|$. For a symmetric matrix, its matrix ℓ_1 norm coincides with its matrix ℓ_∞ norm. We use c or C to denote constants that do not depend on n or p.

Throughout this section, we follow Bickel and Levina (2008a, b) to assume that

$$\varepsilon_0 \le \lambda_{\min}(\boldsymbol{\Sigma}^\star) \le \lambda_{\max}(\boldsymbol{\Sigma}^\star) \le \frac{1}{\varepsilon_0}.$$

3.1. A General Theory

First of all, we consider any regularized estimator $\hat{\boldsymbol{\Sigma}}$ satisfying the following condition:

(C1) There exists a regularized estimator $\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{zz} & \hat{\boldsymbol{\Sigma}}_{zy} \\ \hat{\boldsymbol{\Sigma}}_{yz} & \hat{\boldsymbol{\Sigma}}_{yy} \end{pmatrix}$ satisfying that

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\|_{\ell_1} = O_P(\xi_{n,p}), \quad \text{with} \quad \xi_{n,p} = o((\log n)^{-1/2}).$$

In the following theorem, we show that the conditional distribution of $\boldsymbol{h}(\boldsymbol{y})$ given $\boldsymbol{g}(\boldsymbol{z})$ can be well estimated under Condition (C1).

Theorem 1. Assume the data follow the transnormal model. Suppose that there exists $\kappa \in (0, 1)$ such that $n^\kappa \gg \log n \log p_0$. Given Condition (C1), we have the error bounds concerning the estimated conditional distribution of $\boldsymbol{h}(\boldsymbol{y})$ given $\boldsymbol{g}(\boldsymbol{z})$ in (2) as follows:

(i) On the conditional mean vector:

$$\left\| \hat{\boldsymbol{\Sigma}}_{yz}\hat{\boldsymbol{\Sigma}}_{zz}^{-1} \cdot \hat{\boldsymbol{g}}(\boldsymbol{z}) - \boldsymbol{\Sigma}_{yz}^\star(\boldsymbol{\Sigma}_{zz}^\star)^{-1} \cdot \boldsymbol{g}(\boldsymbol{z}) \right\|_{\max} = O_P\left(\sqrt{\frac{\log n \log p_0}{n^\kappa}} + \xi_{n,p}\sqrt{\log n} \right).$$

(ii) On the conditional variance matrix:

$$\left\| \left(\hat{\boldsymbol{\Sigma}}_{yy} - \hat{\boldsymbol{\Sigma}}_{yz}\hat{\boldsymbol{\Sigma}}_{zz}^{-1}\hat{\boldsymbol{\Sigma}}_{zy}\right) - \left(\boldsymbol{\Sigma}_{yy}^\star - \boldsymbol{\Sigma}_{yz}^\star(\boldsymbol{\Sigma}_{zz}^\star)^{-1}\boldsymbol{\Sigma}_{zy}^\star\right) \right\|_{\ell_2} = O_P(\xi_{n,p}).$$

Next, we show that the plug-in estimator $\hat{\boldsymbol{Q}}_\tau(\boldsymbol{y}|\boldsymbol{z})$ is asymptotically as good as $\boldsymbol{Q}_\tau(\boldsymbol{y}|\boldsymbol{z})$ under mild regularity conditions. Let $\psi_j(\cdot)$ be the probability density function of X_j. Define $\boldsymbol{L} = \boldsymbol{\Sigma}_{yz}^\star(\boldsymbol{\Sigma}_{zz}^\star)^{-1} \cdot \boldsymbol{g}(\boldsymbol{z}) + \Phi^{-1}(\tau)\ \mathrm{vdiag}(\boldsymbol{\Sigma}_{yy}^\star - \boldsymbol{\Sigma}_{yz}^\star(\boldsymbol{\Sigma}_{zz}^\star)^{-1}\boldsymbol{\Sigma}_{zy}^\star)$, and let $\boldsymbol{L} = (L_1, \ldots, L_{q_0})'$.

Theorem 2. Under the conditions of Theorem 1, we have the error bound concerning the plug-in estimator in (5) as follows:

$$\|\hat{\boldsymbol{Q}}_\tau(\boldsymbol{y}|\boldsymbol{z}) - \boldsymbol{Q}_\tau(\boldsymbol{y}|\boldsymbol{z})\|_{\max} = O_P\left(\frac{1}{M}\sqrt{\frac{\log q_0}{n}} + \frac{1}{M}\sqrt{\frac{\log n \log p_0}{n^\kappa}} + \frac{\xi_{n,p}}{M}\sqrt{\log n} \right), \tag{11}$$

where M is the minimum of $\psi_{p_0+j}(x)$ over $x \in \mathcal{I}_j = [-|2L_j|, |2L_j|] \cup [-f_{p_0+j}^{-1}(|2L_j|), f_{p_0+j}^{-1}(|2L_j|)]$ for any $j = 1, \ldots, q_0$, that is, $M = \min_{j=1,\ldots,q_0} \min_{x\in\mathcal{I}_j} \psi_{p_0+j}(x)$.

Theorem 2 immediately implies that the plug-in estimator $\hat{\boldsymbol{Q}}_\tau(\boldsymbol{y}|\boldsymbol{z})$ can be used to construct provable prediction intervals in high dimensions. Therefore, we use

$$[\hat{\boldsymbol{Q}}_{\frac{\tau}{2}}(\boldsymbol{y}|\boldsymbol{z}), \hat{\boldsymbol{Q}}_{1-\frac{\tau}{2}}(\boldsymbol{y}|\boldsymbol{z})]$$

to construct the $100(1-\tau)\%$ prediction interval for predicting $\boldsymbol{y}$ given $\boldsymbol{z}$.

In what follows, we show how Condition (C1) is satisfied. Specifically, we consider two parameter spaces for $\boldsymbol{\Sigma}^\star$:

$$\mathcal{G}_q = \left\{ \boldsymbol{\Sigma} : \max_j \sum_{j\neq i} |a_{ij}|^q \le s_0 \right\}$$

$$\mathcal{H}_\alpha = \left\{ \boldsymbol{\Sigma} : \max_j \sum_{j<i-k} |a_{ij}| \le c_0 k^{-\alpha},\ \forall k \right\}.$$

These two parameter spaces were previously studied by Bickel and Levina (2008a,b) and Cai and Zhou (2012). We will show that the proposed estimators achieve the "oracle"-like convergence rate over $\mathcal{G}_q$ and $\mathcal{H}_\alpha$, respectively. As a result, we obtain the optimal prediction result.

3.2. Positive-Definite Sparse Correlation Matrix

First, we derive the estimation bound for the rank-based positive-definite ℓ_1 penalization and show its application to estimate the conditional distribution of $\boldsymbol{h}(\boldsymbol{y})$ given $\boldsymbol{g}(\boldsymbol{z})$.

Theorem 3. Assume the data follow the transnormal model with $\boldsymbol{\Sigma}^\star \in \mathcal{G}_q$, and also assume that $n \gg \log p$. Suppose that $\lambda_{\min}(\boldsymbol{\Sigma}^\star) \ge \varepsilon_0 \gg s_0(\log p/n)^{\frac{1-q}{2}} + \epsilon$. With probability tending to 1, the rank-based positive-definite ℓ_1 penalization with $\lambda = c(\log p/n)^{1/2}$ achieves the following upper bound under the matrix ℓ_1 norm,

$$\sup_{\boldsymbol{\Sigma}^\star\in\mathcal{G}_q} \left\| \hat{\boldsymbol{\Sigma}}_{\ell_1}^s - \boldsymbol{\Sigma}^\star \right\|_{\ell_1} \le C \cdot s_0 \left(\frac{\log p}{n} \right)^{\frac{1-q}{2}}.$$

Remark 4. Both the rank-based estimator $\hat{\boldsymbol{\Sigma}}_{\ell_1}^s$ and its "oracle" counterpart achieve the minimax optimal convergence rate under the matrix ℓ_1 norm (Cai and Zhou 2012).

Corollary 1. Under the conditions of Theorems 1, 2, and 3, with $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_{\ell_1}^s$ we have

$$\sup_{\boldsymbol{\Sigma}^\star\in\mathcal{G}_q} \left\| \hat{\boldsymbol{\Sigma}}_{yz}\hat{\boldsymbol{\Sigma}}_{zz}^{-1} \cdot \hat{\boldsymbol{g}}(\boldsymbol{z}) - \boldsymbol{\Sigma}_{yz}^\star(\boldsymbol{\Sigma}_{zz}^\star)^{-1} \cdot \boldsymbol{g}(\boldsymbol{z}) \right\|_{\max} = O_P\left(\sqrt{\frac{\log n \log p_0}{n^\kappa}} + s_0\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}}\sqrt{\log n} \right),$$

and

$$\sup_{\boldsymbol{\Sigma}^\star\in\mathcal{G}_q} \left\| (\hat{\boldsymbol{\Sigma}}_{yy} - \hat{\boldsymbol{\Sigma}}_{yz}\hat{\boldsymbol{\Sigma}}_{zz}^{-1}\hat{\boldsymbol{\Sigma}}_{zy}) - (\boldsymbol{\Sigma}_{yy}^\star - \boldsymbol{\Sigma}_{yz}^\star(\boldsymbol{\Sigma}_{zz}^\star)^{-1}\boldsymbol{\Sigma}_{zy}^\star) \right\|_{\ell_2}$$

$$= O_P\left(s_0\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}}\right).$$

In light of Theorem 3 and Corollary 1, we can derive the convergence rate for the proposed optimal prediction in the following theorem.

Theorem 4. Under same conditions of Theorems 1, 2, and 3, with $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}^s_{\ell_1}$ we have

$$\sup_{\boldsymbol{\Sigma}^\star \in \mathcal{G}_q} \|\hat{\boldsymbol{Q}}_\tau(\boldsymbol{y}|\boldsymbol{z}) - \boldsymbol{Q}_\tau(\boldsymbol{y}|\boldsymbol{z})\|_{\max} = O_P\left(\frac{1}{M}\sqrt{\frac{\log q_0}{n}} + \frac{1}{M}\sqrt{\frac{\log n \log p_0}{n^\kappa}} + \frac{s_0}{M}\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}}\sqrt{\log n}\right).$$

3.3. Banded Cholesky Decomposition Regularization

Next, we derive the estimation bound for the rank-based banded Cholesky decomposition regularization with applications to estimate the conditional distribution of $\boldsymbol{h}(\boldsymbol{y})$ given $\boldsymbol{g}(\boldsymbol{z})$.

Theorem 5. Assume the data follow the transnormal model with $\boldsymbol{\Sigma}^\star \in \mathcal{H}_\alpha$, and also assume that $n \gg k^2 \log p$. With probability tending to 1, the rank-based banded Cholesky decomposition regularization achieves the following upper bound under the matrix ℓ_1 norm,

$$\sup_{\boldsymbol{\Sigma}^\star \in \mathcal{H}_\alpha} \left\|\hat{\boldsymbol{\Sigma}}^s_{\text{chol}} - \boldsymbol{\Sigma}^\star\right\|_{\ell_1} \le Ck\left(\frac{\log p}{n}\right)^{1/2} + Ck^{-\alpha}.$$

If $k = c \cdot (\frac{\log p}{n})^{\frac{-1}{2(\alpha+1)}}$ in the rank-based banded Cholesky decomposition regularization, we have

$$\sup_{\boldsymbol{\Sigma}^\star \in \mathcal{H}_\alpha} \left\|\hat{\boldsymbol{\Sigma}}^s_{\text{chol}} - \boldsymbol{\Sigma}^\star\right\|_{\ell_1} = O_P\left(\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2(\alpha+1)}}\right).$$

Remark 5. Bickel and Levina (2008a) studied the banded Cholesky decomposition regularization under the normal model. Their analysis directly applies to the "oracle" banded Cholesky decomposition regularization. By Theorem 3 of Bickel and Levina (2008a), we have

$$\sup_{\boldsymbol{\Sigma}^\star \in \mathcal{H}_\alpha} \left\|\hat{\boldsymbol{\Sigma}}^o_{\text{chol}} - \boldsymbol{\Sigma}^\star\right\|_{\ell_1} = O_P\left(\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2(\alpha+1)}}\right).$$

Therefore, the rank-based banded Cholesky decomposition regularization achieves the same convergence rate as the "oracle" counterpart.

Corollary 2. Under the conditions of Theorems 1, 2, and 5, with $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}^s_{\text{chol}}$ we have

$$\sup_{\boldsymbol{\Sigma}^\star \in \mathcal{H}_\alpha} \left\|\hat{\boldsymbol{\Sigma}}_{yz}\hat{\boldsymbol{\Sigma}}_{zz}^{-1} \cdot \hat{\boldsymbol{g}}(\boldsymbol{z}) - \boldsymbol{\Sigma}^\star_{yz}(\boldsymbol{\Sigma}^\star_{zz})^{-1} \cdot \boldsymbol{g}(\boldsymbol{z})\right\|_{\max} = O_P\left(\sqrt{\frac{\log n \log p_0}{n^\kappa}} + \left(\frac{\log p}{n}\right)^{\frac{\alpha}{2(\alpha+1)}}\sqrt{\log n}\right),$$

and

$$\sup_{\boldsymbol{\Sigma}^\star \in \mathcal{H}_\alpha} \left\|\left(\hat{\boldsymbol{\Sigma}}_{yy} - \hat{\boldsymbol{\Sigma}}_{yz}\hat{\boldsymbol{\Sigma}}_{zz}^{-1}\hat{\boldsymbol{\Sigma}}_{zy}\right) - \left(\boldsymbol{\Sigma}^\star_{yy} - \boldsymbol{\Sigma}^\star_{yz}(\boldsymbol{\Sigma}^\star_{zz})^{-1}\boldsymbol{\Sigma}^\star_{zy}\right)\right\|_{\ell_2} = O_P\left(\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2(\alpha+1)}}\right).$$

In light of Theorem 5 and Corollary 2, we can derive the convergence rate for the proposed optimal prediction in the following theorem.

Theorem 6. Under the conditions of Theorems 1, 2 and 5, with $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}^s_{\text{chol}}$ we have

$$\sup_{\boldsymbol{\Sigma}^\star \in \mathcal{G}_q} \left\|\hat{\boldsymbol{Q}}_\tau(\boldsymbol{y}|\boldsymbol{z}) - \boldsymbol{Q}_\tau(\boldsymbol{y}|\boldsymbol{z})\right\|_{\max} = O_P\left(\frac{1}{M}\sqrt{\frac{\log q_0}{n}} + \frac{1}{M}\sqrt{\frac{\log n \log p_0}{n^\kappa}} + \frac{\sqrt{\log n}}{M}\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2(\alpha+1)}}\right).$$

4. Numerical Properties

This section demonstrates the numerical performance of the proposed methods. For space consideration, we only focus on multitask median regression and rank-based banded Cholesky decomposition regularization.

4.1. Simulation Studies

Simulation studies investigate the "oracle" estimator, the proposed rank-based estimator and the "naïve" estimator. The "oracle" estimator serves as a benchmark in the numerical comparison, and the "naïve" estimator directly regularizes on the covariance of the original data. We summarize notation and details of three regularized estimator in Table 1.

In Models 1–3, we consider three different designs for the inverse covariance matrix $\boldsymbol{\Omega}^\star$:

Model 1: $\boldsymbol{\Omega}^\star = (\boldsymbol{I} - \boldsymbol{A})'\boldsymbol{D}^{-1}(\boldsymbol{I} - \boldsymbol{A})$: $d_{ii} = 0.01$, $a_{i+1,i} = 0.8$ and $d_{ij} = a_{ij} = 0$ otherwise;

Model 2: $\boldsymbol{\Omega}^\star$: $\omega^\star_{ii} = 1$, $\omega^\star_{i,i\pm1} = 0.5$, $\omega^\star_{i,i\pm2} = 0.25$ and $\omega^\star_{ij} = 0$ otherwise;

Model 3: $\boldsymbol{\Omega}^\star$: $\omega^\star_{ii} = 1$, $\omega^\star_{i,i\pm1} = 0.4$, $\omega^\star_{i,i\pm2} = \omega^\star_{i,i\pm3} = 0.2$, $\omega^\star_{i,i\pm4} = 0.1$ and $\omega^\star_{ij} = 0$ otherwise.

Models 1–3 are autoregressive models considered by Huang et al. (2006) and Xue and Zou (2012). Based on $\boldsymbol{\Gamma}^\star = (\boldsymbol{\Omega}^\star)^{-1} = (\gamma^\star_{ij})_{p\times p}$, we calculate the true correlation matrix $\boldsymbol{\Sigma}^\star = (\sigma^\star_{ij})_{p\times p} = (\frac{\gamma^\star_{ij}}{\sqrt{\gamma^\star_{ii}\gamma^\star_{jj}}})_{p\times p}$ and also its inverse $\boldsymbol{\Theta}^\star = (\boldsymbol{\Sigma}^\star)^{-1}$.

Table 1. List of three regularized estimators in the simulation study.

Methods	Details
$(\hat{\boldsymbol{\Sigma}}^o_{\text{chol}}, \hat{\boldsymbol{\Theta}}^o_{\text{chol}}, \hat{t}^o_1(z))$	regularizing the "oracle" sample correlation matrix
$(\hat{\boldsymbol{\Sigma}}^s_{\text{chol}}, \hat{\boldsymbol{\Theta}}^s_{\text{chol}}, \hat{t}^s_1(z))$	regularizing the adjusted Spearman's rank correlation matrix
$(\hat{\boldsymbol{\Sigma}}^n_{\text{chol}}, \hat{\boldsymbol{\Theta}}^n_{\text{chol}}, \hat{t}^n_1(z))$	regularizing the usual sample correlation matrix

Table 2. Performance of estimating the correlation matrix with $\hat{\mathbf{\Sigma}}^o_{\text{chol}}$, $\hat{\mathbf{\Sigma}}^s_{\text{chol}}$, and $\hat{\mathbf{\Sigma}}^n_{\text{chol}}$.

	Model 1			Model 2			Model 3		
Method	$p=100$	$p=200$	$p=500$	$p=100$	$p=200$	$p=500$	$p=100$	$p=200$	$p=500$
					matrix ℓ_1 norm				
$\hat{\mathbf{\Sigma}}^o_{\text{chol}}$	1.34 (0.05)	1.60 (0.13)	1.95 (0.27)	0.89 (0.03)	0.91 (0.04)	0.99 (0.08)	1.00 (0.02)	1.04 (0.05)	1.16 (0.12)
$\hat{\mathbf{\Sigma}}^s_{\text{chol}}$	1.49 (0.07)	1.68 (0.11)	2.08 (0.21)	0.92 (0.03)	0.99 (0.05)	1.10 (0.14)	1.03 (0.02)	1.06 (0.04)	1.21 (0.12)
$\hat{\mathbf{\Sigma}}^n_{\text{chol}}$	3.21 (0.06)	3.49 (0.09)	3.78 (0.24)	1.45 (0.02)	1.48 (0.04)	1.63 (0.11)	1.11 (0.03)	1.22 (0.05)	1.28 (0.11)
					matrix ℓ_2 norm				
$\hat{\mathbf{\Sigma}}^o_{\text{chol}}$	0.78 (0.03)	0.95 (0.06)	1.10 (0.13)	0.48 (0.01)	0.49 (0.02)	0.54 (0.04)	0.52 (0.01)	0.54 (0.02)	0.59 (0.05)
$\hat{\mathbf{\Sigma}}^s_{\text{chol}}$	0.84 (0.03)	1.00 (0.06)	1.26 (0.11)	0.49 (0.01)	0.53 (0.02)	0.58 (0.05)	0.53 (0.01)	0.55 (0.02)	0.63 (0.06)
$\hat{\mathbf{\Sigma}}^n_{\text{chol}}$	1.67 (0.05)	1.73 (0.08)	1.89 (0.14)	0.77 (0.01)	0.80 (0.02)	0.84 (0.04)	0.55 (0.01)	0.60 (0.02)	0.68 (0.05)

NOTE: Each metric is averaged over 100 independent replications with standard errors in the bracket.

Then, we generate n independent normal data from $N_p(\mathbf{0}, \mathbf{\Sigma}^\star)$ and transfer the normal data to the desired transnormal data $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ using

$$\left[f_1^{-1}, f_2^{-1}, f_3^{-1}, f_4^{-1}, f_1^{-1}, f_2^{-1}, f_3^{-1}, f_4^{-1}, \ldots\right], \tag{12}$$

where $f_1(x) = x^{\frac{1}{3}}$ (power transformation), $f_2(x) = \log(\frac{x}{1-x})$ (logit transformation), $f_3(x) = \log(x)$ (log transformation) and $f_4(x) = f_1(x)I_{\{x<-1\}} + f_2(x)I_{\{-1\le x\le 1\}} + (f_3(x-1)+1)I_{\{x>1\}}$.

In all cases, we let $n = 200$ and $p = 100$, 200, and 500. For each estimator, tuning parameter is chosen by cross-validation. Estimation accuracy is measured by the matrix ℓ_1-norm and ℓ_2-norm averaged over 100 independent replications.

Tables 2 and 3 summarize the numerical performance of estimating the correlation matrix and its inverse for three estimators. We see that the "naïve" estimator performs the worst in the presence of nonnormality. The rank-based estimator effectively deals with the transnormal data, and performs comparably with the "oracle" estimator. The numeric evidence is consistent with theoretical results presented in Section 3.

Next, we compare the performance of estimating optimal transformations for multitask median regression. In each of the 100 replications, we simulate another $4 \times n$ independent transnormal data $(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{5n})$ as the testing data. Now, we take the first $p_0 = p/2$ variables in $\boldsymbol{x}$ as response $\boldsymbol{y}$ and take the second $q_0 = p/2$ variables in $\boldsymbol{x}$ as predictor $\boldsymbol{z}$. Now, given three estimators $\hat{\mathbf{\Sigma}}^o_{\text{chol}}$, $\hat{\mathbf{\Sigma}}^s_{\text{chol}}$ and $\hat{\mathbf{\Sigma}}^n_{\text{chol}}$, we examine their performance in predicting $\boldsymbol{y}_i$ based on $\boldsymbol{z}_i$ for $i = n+1, \ldots, 5n$. Recall that $\hat{\boldsymbol{t}}_1(\boldsymbol{z}) = \hat{\boldsymbol{h}}^{-1}(\hat{\mathbf{\Sigma}}_{yz}(\hat{\mathbf{\Sigma}}_{zz})^{-1} \cdot \hat{\boldsymbol{g}}(\boldsymbol{z}))$ is the closed-form solution for multitask median regression. In what follows, we consider three optimal prediction methods using $\hat{\mathbf{\Sigma}}^o_{\text{chol}}$, $\hat{\mathbf{\Sigma}}^s_{\text{chol}}$ and $\hat{\mathbf{\Sigma}}^n_{\text{chol}}$ respectively:

- the "oracle" estimator: $\hat{\boldsymbol{t}}^o_1(\boldsymbol{z}) = \boldsymbol{h}^{-1}(\hat{\mathbf{\Sigma}}^o_{yz}(\hat{\mathbf{\Sigma}}^o_{zz})^{-1} \cdot \boldsymbol{g}(\boldsymbol{z}))$
- the rank-based estimator: $\hat{\boldsymbol{t}}^s_1(\boldsymbol{z}) = \hat{\boldsymbol{h}}^{-1}(\hat{\mathbf{\Sigma}}^s_{yz}(\hat{\mathbf{\Sigma}}^s_{zz})^{-1} \cdot \hat{\boldsymbol{g}}(\boldsymbol{z}))$
- the "naïve" estimator: $\hat{\boldsymbol{t}}^n_1(\boldsymbol{z}) = \hat{\boldsymbol{\mu}}_y + \hat{\mathbf{\Sigma}}^n_{yz}(\hat{\mathbf{\Sigma}}^n_{zz})^{-1} \cdot (\boldsymbol{z} - \hat{\boldsymbol{\mu}}_z)$

Table 4 shows their performances. Prediction accuracy is measured by the difference of prediction errors, which is defined

Table 3. Performance of estimating the inverse correlation matrix with $\hat{\mathbf{\Theta}}^o_{\text{chol}}$, $\hat{\mathbf{\Theta}}^s_{\text{chol}}$, and $\hat{\mathbf{\Theta}}^n_{\text{chol}}$.

	Model 1			Model 2			Model 3		
Method	$p=100$	$p=200$	$p=500$	$p=100$	$p=200$	$p=500$	$p=100$	$p=200$	$p=500$
					matrix ℓ_1 norm				
$\hat{\mathbf{\Theta}}^o_{\text{chol}}$	4.07 (0.21)	4.72 (0.37)	5.14 (0.57)	1.83 (0.08)	1.95 (0.12)	1.99 (0.25)	2.21 (0.04)	2.29 (0.09)	2.37 (0.18)
$\hat{\mathbf{\Theta}}^s_{\text{chol}}$	4.21 (0.24)	4.80 (0.32)	5.28 (0.55)	1.91 (0.07)	1.97 (0.14)	2.02 (0.25)	2.27 (0.04)	2.33 (0.07)	2.41 (0.16)
$\hat{\mathbf{\Theta}}^n_{\text{chol}}$	7.18 (0.22)	7.55 (0.35)	8.02 (0.54)	3.29 (0.06)	3.48 (0.14)	3.70 (0.29)	2.81 (0.03)	2.93 (0.10)	3.08 (0.19)
					matrix ℓ_2 norm				
$\hat{\mathbf{\Theta}}^o_{\text{chol}}$	3.18 (0.10)	3.41 (0.27)	4.07 (0.35)	1.26 (0.04)	1.36 (0.07)	1.41 (0.13)	1.68 (0.04)	1.74 (0.08)	1.75 (0.10)
$\hat{\mathbf{\Theta}}^s_{\text{chol}}$	3.29 (0.16)	3.59 (0.25)	4.10 (0.30)	1.35 (0.04)	1.39 (0.07)	1.45 (0.08)	1.69 (0.04)	1.75 (0.08)	1.77 (0.11)
$\hat{\mathbf{\Theta}}^n_{\text{chol}}$	4.98 (0.18)	5.15 (0.25)	5.33 (0.32)	2.64 (0.04)	2.81 (0.11)	2.89 (0.18)	2.29 (0.03)	2.32 (0.07)	2.39 (0.18)

NOTE: Each metric is averaged over 100 independent replications with standard errors in the bracket.

Table 4. Performance of three optimal prediction methods $\hat{t}_1^o(z)$, $\hat{t}_1^s(z)$, and $\hat{t}_1^n(z)$.

	Model 1			Model 2			Model 3		
Method	$p=100$	$p=200$	$p=500$	$p=100$	$p=200$	$p=500$	$p=100$	$p=200$	$p=500$
$\hat{t}_1^s(z)$ vs. $\hat{t}_1^o(z)$	0.02 (0.00)	0.03 (0.00)	0.08 (0.01)	0.16 (0.02)	0.26 (0.03)	0.66 (0.04)	0.12 (0.01)	0.22 (0.01)	0.56 (0.04)
$\hat{t}_1^n(z)$ vs. $\hat{t}_1^o(z)$	0.11 (0.00)	0.19 (0.01)	0.21 (0.03)	13.50 (0.22)	26.07 (0.58)	63.80 (2.16)	7.38 (0.14)	14.70 (0.37)	35.68 (1.28)

NOTE: Prediction accuracy are measured by the relative prediction error and averaged over 100 independent replications, with the standard errors shown in the bracket.

by subtracting the "oracle" prediction error, that is,

$$\mathrm{DPE}(\hat{\boldsymbol{t}}_1(\boldsymbol{z})) = \frac{1}{4n}\sum_{i=n+1}^{5n} \|\hat{\boldsymbol{t}}_1(\boldsymbol{z}_i) - \boldsymbol{y}_i\|_{\ell_1} - \frac{1}{4n}\sum_{i=n+1}^{5n} \|\hat{\boldsymbol{t}}_1^o(\boldsymbol{z}) - \boldsymbol{y}_i\|_{\ell_1}.$$

The relative prediction error $\mathrm{DPE}(\hat{\boldsymbol{t}}_1(\boldsymbol{z}))$ is averaged over 100 independent replications.

As shown in Table 4, we see that the rank-based estimator performs similarly to its oracle counterpart, and outperforms the "naïve" estimator. The numeric evidence is also consistent with theoretical results presented in Section 3.

4.2. Application to the Protein Mass Spectroscopy Data

Recent advances in high-throughput mass spectroscopy technology has enabled biomedical researchers to simultaneously analyze thousands of proteins. This subsection illustrates the power of the proposed rank-based method in an application to study the prostate cancer using the protein mass spectroscopy data (Adam et al. 2002), which was previously analyzed by Levina, Rothman, and Zhu (2008). This dataset consists of the protein mass spectroscopy measurements for the blood serum samples of 157 healthy people and 167 prostate cancer patients. In each blood serum sample, the protein mass spectroscopy measures the intensity for the ordered time-of-flight values, which are related to the mass over charge ratio of proteins. In our analysis, we exclude the measurements with mass over charge ratio below 2000 to avoid chemical artifacts, and perform the same preprocessing as in Levina, Rothman, and Zhu (2008) to smooth the intensity profile. This gives a total of $p = 218$ ordered mass over charge ratio indices for each sample. Then, we have the control data $\boldsymbol{x}_i^{\mathrm{co}} = (x_{i,1}^{\mathrm{co}}, \ldots, x_{i,218}^{\mathrm{co}})$ for $i = 1, \ldots, 157$, and the cancer data $\boldsymbol{x}_j^{\mathrm{ca}} = (x_{j,1}^{\mathrm{ca}}, \ldots, x_{j,218}^{\mathrm{ca}})$ for $j = 1, \ldots, 167$. We refer the readers to Adam et al. (2002) and Levina, Rothman, and Zhu (2008) for more details about this dataset and also the preprocessing procedure. We use the more stable intensity measurements in the latter 168 mass over charge ratio indices to predict the more volatile intensity measurements in the first 50 indices.

The analysis of Levina, Rothman, and Zhu (2008) is based on the multivariate normal assumption. We perform normality tests on both control and cancel data. Table 5 shows the number of rejecting null hypotheses among 218 normality tests. At least 50% of mass spectroscopy measurements appear to be nonnormal. Even after Bonferroni correction, there are at least 30 indices in the control data and 116 indices in the cancer data rejecting all null hypotheses. Figure 1 illustrates nonnormality (e.g., heavy tails, skewness) in two indices (109, 218).

The nonnormality of measurements suggests us to try the transnormal model. Let $\boldsymbol{z}^{\mathrm{co}}$ ($\boldsymbol{z}^{\mathrm{ca}}$) and $\boldsymbol{y}^{\mathrm{co}}$ ($\boldsymbol{y}^{\mathrm{ca}}$) denote the intensity measurements in the latter 168 and first 50 indices of the control (cancer) data. We divide this dataset into training sets $(\boldsymbol{x}_1^{\mathrm{co}}, \ldots, \boldsymbol{x}_{120}^{\mathrm{co}})$, $(\boldsymbol{x}_1^{\mathrm{ca}}, \ldots, \boldsymbol{x}_{120}^{\mathrm{ca}})$ and testing sets $(\boldsymbol{x}_{121}^{\mathrm{co}}, \ldots, \boldsymbol{x}_{157}^{\mathrm{co}})$, $(\boldsymbol{x}_{121}^{\mathrm{ca}}, \ldots, \boldsymbol{x}_{167}^{\mathrm{ca}})$. Note that the sample estimator $\hat{\boldsymbol{t}}_1^{\mathrm{sample}}(\boldsymbol{z}) = \hat{\boldsymbol{\mu}}_y + \hat{\boldsymbol{\Sigma}}_{yz}^{\mathrm{sample}}(\hat{\boldsymbol{\Sigma}}_{zz}^{\mathrm{sample}})^{-1} \cdot (\boldsymbol{z} - \hat{\boldsymbol{\mu}}_z)$ is infeasible since the usual sample correlation matrix $\hat{\boldsymbol{\Sigma}}_{zz}^{\mathrm{sample}}$ is not invertible. We consider two different methods to predict $\boldsymbol{y}$: the proposed rank-based estimator $\hat{\boldsymbol{t}}_1^s(\boldsymbol{z}) = \hat{\boldsymbol{h}}^{-1}(\hat{\boldsymbol{\Sigma}}_{yz}^s(\hat{\boldsymbol{\Sigma}}_{zz}^s)^{-1} \cdot \hat{\boldsymbol{g}}(\boldsymbol{z}))$ by using the proposed rank-based banded Cholesky decomposition regularization and the "naïve" estimator $\hat{\boldsymbol{t}}_1^n(\boldsymbol{z}) = \hat{\boldsymbol{\mu}}_y + \hat{\boldsymbol{\Sigma}}_{yz}^n(\hat{\boldsymbol{\Sigma}}_{zz}^n)^{-1} \cdot (\boldsymbol{z} - \hat{\boldsymbol{\mu}}_z)$ by using the banded Cholesky decomposition regularization (Bickel and Levina 2008a). Following Levina, Rothman, and Zhu (2008), tuning parameters are chosen via cross-validation. The difference of prediction error at jth index is computed as

$$\mathrm{DPE}_j(\hat{\boldsymbol{t}}_1(\boldsymbol{z})) = \mathrm{PE}_j(\hat{\boldsymbol{t}}_1^s(\boldsymbol{z})) - \mathrm{PE}_j(\hat{\boldsymbol{t}}_1^n(\boldsymbol{z})),$$

where $\mathrm{PE}_j(\hat{\boldsymbol{t}}_1(\boldsymbol{z}))$ is prediction error at the jth index computed by averaging the absolute prediction error $|(\hat{\boldsymbol{t}}_1(\boldsymbol{z}_i))_j - y_{ij}|$ over $i = 121, \ldots, 157$ for the control data and over $i = 121, \ldots, 167$ for the cancer data. The differences of prediction errors are shown in Figure 2. As demonstrated in Figure 2, the rank-based method outperforms the "naïve" estimator in 38 out of 50 indices for the control data and 46 out of 50 indices for the cancer data.

Table 5. Testing for normality.

	Significance level	Anderson–Darling	Cramer–von Mises	Kolmogorov–Smirnov
Control data	0.05	158	138	119
	0.05/218	71	52	30
Cancer data	0.05	196	177	156
	0.05/218	136	134	116

NOTE: This table shows the number of rejecting the normality hypothesis at different significance levels among 218 mass over charge ratio indices.

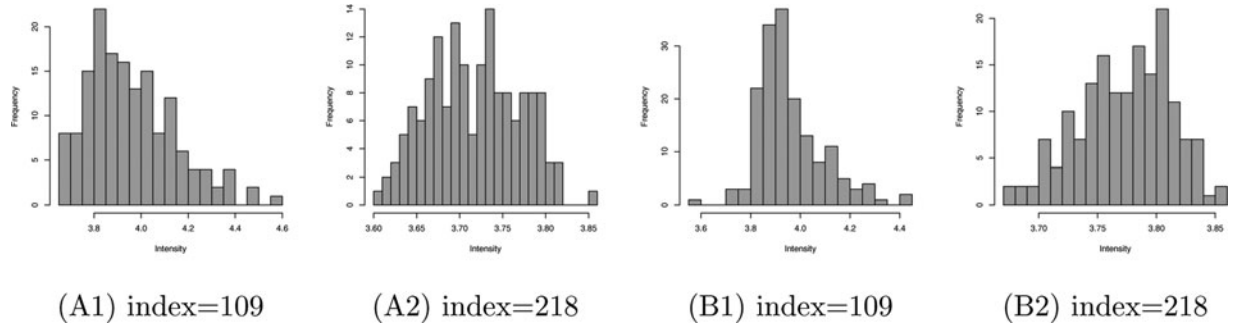

Figure 1. Illustration of the nonnormality issue of the protein mass spectroscopy data: (A1) and (A2) for the control data; (B1) and (B2) for the cancer data.

Moreover, we further use $[\hat{\boldsymbol{Q}}_{\frac{\tau}{2}}(\boldsymbol{y}|\boldsymbol{z}), \hat{\boldsymbol{Q}}_{1-\frac{\tau}{2}}(\boldsymbol{y}|\boldsymbol{z})]$ to construct the $100(1-\tau)\%$ prediction interval. The predict target and prediction intervals are averaged over the testing set. Specifically, we estimate the $100(1-\tau)\%$ prediction interval as follows,

$$100(1-\tau)\%\ \text{PI} = \left[\frac{1}{34}\sum_{i=206}^{239}\hat{\boldsymbol{Q}}_{\frac{\tau}{2}}(\boldsymbol{y}_i|\boldsymbol{z}_i),\ \frac{1}{34}\sum_{i=206}^{239}\hat{\boldsymbol{Q}}_{1-\frac{\tau}{2}}(\boldsymbol{y}_i|\boldsymbol{z}_i)\right].$$

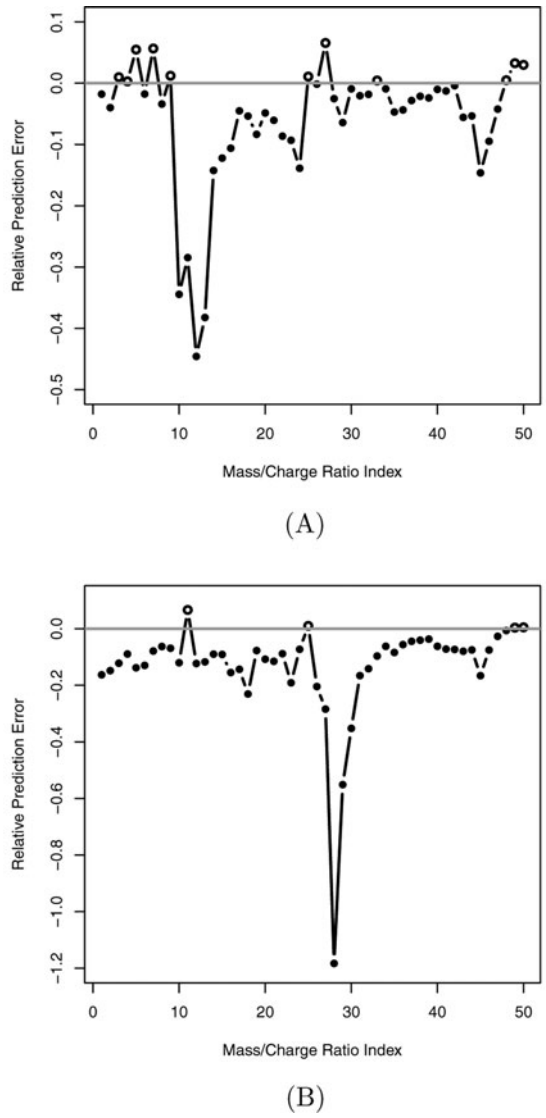

Figure 2. The differences of prediction errors between the existing Cholesky-based method (Bickel and Levina 2008a) and the proposed rank-based estimate with $\tau = 0.5$. The solid dots below the horizontal line represent the mass/charge ratio indices, where our proposed method outperforms the BL method. (A) Relative prediction errors for the control data and (B) relative prediction errors for the cancer data.

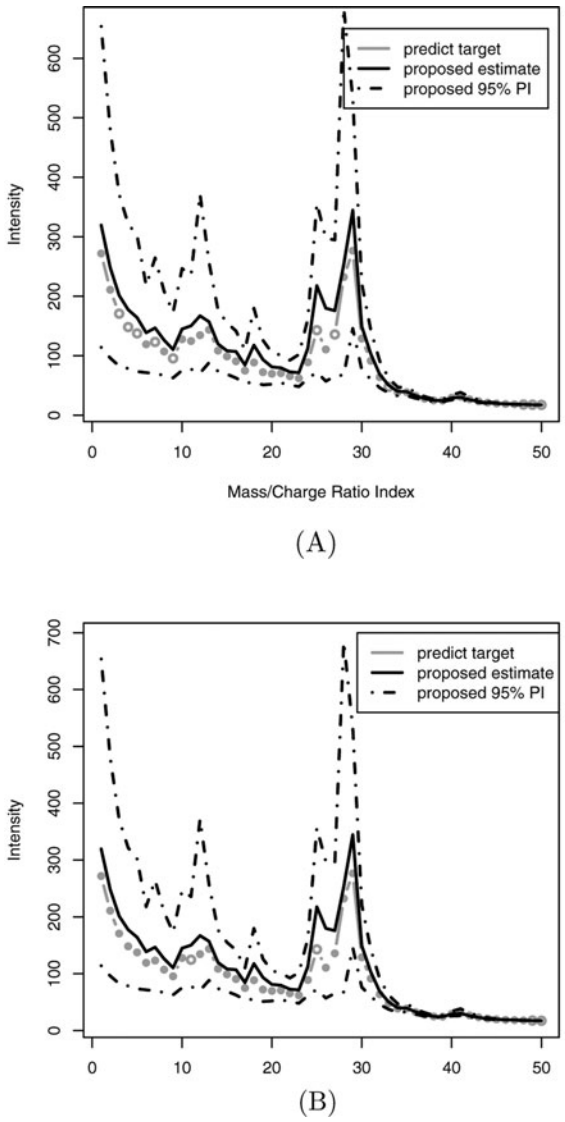

Figure 3. Prediction intervals for predicting protein mass spectroscopy intensities using our proposed method. The predict target, proposed point estimate ($\tau = 0.5$), and 95% prediction intervals are averaged over the testing set. The solid dots are the mass/charge ratio indices in which the proposed method outperforms the BL method. (A) 95% Prediction interval for the control data and (B) 95% prediction interval for the cancer data.

We plot the 95% prediction intervals for both the control data and the cancer data in Figure 3. Figure 3 shows that the estimated prediction intervals cover most of the predict target. This suggests that the transnormal model is a good fit to the data.

Supplementary Materials

The online supplement contains the appendices for the article.

Acknowledgement

The authors sincerely thank the associate editor and referees for their help comments and suggestions.

Funding

Jianqing Fan's research is supported in part by R01GM100474-04 and National Science Foundation grants DMS-1206464 and DMS-1406266. Lingzhou Xue's research is supported by the National Institutes of Health grant R01-GM072611-09 and National Science Foundation grant DMS-1505256. Hui Zou's research is supported by NSF grants DMS-0846068 and DMS-1505111.

References

Adam, B., Qu, Y., Davis, J., Ward, M., Clements, M., Cazares, L., Semmes, O., Schellhammer, P., Yasui, Y., Feng, Z., and Wright, G. (2002), "Serum Protein Fingerprinting Coupled With a Pattern-Matching Algorithm Distinguishes Prostate Cancer From Benign Prostate Hyperplasia And Healthy Men," *Cancer Research*, 62, 3609–3614. [1733]

Bickel, P., and Levina, E. (2008a), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227. [1729,1730,1731,1733]

—— (2008b), "Covariance Regularization by Thresholding," *The Annals of Statistics*, 36, 2577–2604. [1728,1730]

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations" (with discussion), *Journal of the Royal Statistical Society*, Series B, 26, 211–252. [1726]

Bradic, J., Fan, J., and Wang, W. (2011), "Penalized Composite Quasi-likelihood for Ultrahigh Dimensional Variable Selection," *Journal of the Royal Statistical Society*, Series B, 73, 325–349. [1727]

Breiman, L., and Friedman, J. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation" (with discussion), *Journal of the American Statistical Association*, 80, 580–598. [1726,1727]

Cai, T., and Zhou, H. (2012), "Minimax Estimation of Large Covariance Matrices Under ℓ_1 Norm" (with discussion), *Statistica Sinica*, 22, 1319–1378. [1730]

Devlin, S., Gnanadesikan, R., and Kettenring, J. (1975), "Robust Estimation and Outlier Detection With Correlation Coefficients," *Biometrika*, 62, 531–545. [1729]

Fan, Y., Härdle, W., Wang, W., and Zhu, L. (2013), "Composite Quantile Regression for the Single-Index Model," *SFB 649 Discussion Papers*, No.2013-010. [1728]

Fan, J., Ke, T., Liu, H., and Xia, L. (2013), "QUADRO: A Supervised Dimension Reduction Method via Rayleigh Quotient Optimization," *The Annals of Statistics*, to appear. [1726]

Fan, J., Liao, Y., and Mincheva, M. (2013), "Large Covariance Estimation by Thresholding Principal Orthogonal Complements" (with discussion), *Journal of the Royal Statistical Society*, Series B, 75, 603–680. [1728]

Friedman, J. H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29, 1189–1232. [1727]

—— (2002), "Stochastic Gradient Boosting," *Computational Statistics & Data Analysis*, 38, 367–378. [1727]

Friedman, J. H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823. [1726,1727]

Han, F., and Liu, H. (2012), "Semiparametric Principal Component Analysis," *Advances in Neural Information Processing Systems*, 171–179. [1726]

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006), "Covariance Matrix Selection and Estimation via Penalised Normal Likelihood," *Biometrika*, 93, 85–98. [1729,1731]

Kendall, M. (1948), *Rank Correlation Methods*, London: Charles Griffin and Co. Ltd. [1727,1729]

Koenker, R., and Bassett, G. J. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50. [1727]

Koenker, R., and Geling, O. (2001), "Reappraising Medfly Longevity: A Quantile Regression Survival Analysis," *Journal of the American Statistical Association*, 96, 458–468. [1727]

Koenker, R., and Xiao, Z. (2006), "Quantile Autoregression" (with discussion), *Journal of the American Statistical Association*, 101, 980–990. [1727]

Levina, E., Rothman, A., and Zhu, J. (2008), "Sparse Estimation of Large Covariance Matrices via a Nested Lasso Penalty," *The Annals of Applied Statistics*, 2, 245–263. [1729,1733]

Lin, Y., and Jeon, Y. (2003), "Discriminant Analysis Through a Semiparametric Model," *Biometrika*, 90, 379–392. [1726]

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012), "High-Dimensional Semiparametric Gaussian Copula Graphical Models," *The Annals of Statistics*, 40, 2293–2326. [1726,1727]

Liu, H., Lafferty, J., and Wasserman, L. (2009), "The Nonparanormal: Semiparametric Estimation of High-Dimensional Undirected Graphs," *Journal of Machine Learning Research*, 10, 1–37. [1726,1728]

Mai, Q., and Zou, H. (2015), "Semiparametric Sparse Discriminant Analysis in Ultra-high Dimensions," *Journal of Multivariate Analysis*, 135, 175–188. [1726,1728]

Qi, H., and Sun, D. (2006), "A Quadratically Convergent Newton Method for Computing the Nearest Correlation Matrix," *SIAM Journal on Matrix Analysis and Applications*, 28, 360–385. [1729]

Stone, C. J. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 689–705. [1726,1727]

Tibshirani, R. (1988), "Estimating Transformations for Regression via Additivity and Variance Stabilization," *Journal of the American Statistical Association*, 83, 394–405. [1726,1727]

Wang, H., and He, X. (2007), "Detecting Differential Expressions in GeneChip Microarray Studies: A Quantile Approach," *Journal of the American Statistical Association*, 102, 104–112. [1727]

Wegkamp, M., and Zhao, Y. (2016), "Adaptive Estimation of the Copula Correlation Matrix for Semiparametric Elliptical Copulas," *Bernoulli*, 22, 1184–1226. [1726]

Wu, W., and Pourahmadi, M. (2003), "Nonparametric Estimation of Large Covariance Matrices of Longitudinal Data," *Biometrika*, 90, 831–844. [1729]

Wu, T., Yu, K., and Yu, Y. (2010), "Single-Index Quantile Regression," *Journal of Multivariate Analysis*, 101, 1607–1621. [1728]

Xue, L., Ma, S., and Zou, H. (2012), "Positive Definite ℓ_1 Penalized Estimation of Large Covariance Matrices," *Journal of the American Statistical Association*, 107, 1480–1491. [1728,1729]

Xue, L., and Zou, H. (2012), "Regularized Rank Estimation of High-dimensional Nonparanormal Graphical Models," *The Annals of Statistics*, 40, 2541–2571. [1726,1727,1731]

—— (2014), "Rank-Based Tapering Estimation of Bandable Correlation Matrices," *Statistica Sinica*, 24, 83–100. [1726]

Zhao, T., Roeder, K., and Liu, H. (2012), "Smooth-projected Neighborhood Pursuit for High-dimensional Nonparanormal Graph Estimation," *Advances in Neural Information Processing Systems*, 162–170. [1729]

Zhu, L., Huang, M., and Li, R. (2012), "Semiparametric Quantile Regression With High-Dimensional Covariates," *Statistica Sinica*, 22, 1379–1401. [1728]

Zou, H., and Yuan, M. (2008), "Composite Quantileregression and the Oracle Model Selection Theory," *The Annals of Statistics*, 36, 1108–1126. [1727]