Multi-task Quantile Regression under the Transnormal Model

Jianqing Fan, Lingzhou Xue and Hui Zou
Princeton University, Pennsylvania State University and University of Minnesota

Abstract

We consider estimating multi-task 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 $\ell_1$ 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 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.

Key Words: Copula model; Optimal transformation; Rank correlation; Cholesky decomposition; Quantile regression; Prediction interval; Alternating direction method of multipliers.

1 Introduction

Consider a multi-task high-dimensional learning paradigm that independent variables $z = (z_1, \ldots, z_{p_0})'$ are used to simultaneously predict multiple response variables $y = (y_1, \ldots, y_{q_0})'$,
where both dimensions $p_0$ and $q_0$ can be of larger order of magnitude than sample size $n$. In this work, we are interested in estimating optimal transformations $t = (t_1, \ldots, t_{q_0})'$ such that $t(z) = (t_1(z), \ldots, t_{q_0}(z))'$ optimally predict $y$ at the same time. Namely, we shall solve optimal transformations from

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

where $L(\cdot)$ is a convex loss function. If $z$ and $y$ have a joint normal distribution, it is appropriate to specify $L(\cdot)$ as the squared loss, i.e., $L(u) = u^2$. With the squared loss, normality makes the neat connection between optimal transformations and ordinary least squares. Thanks to normality, nice properties such as linearity and homoscedasticity hold for optimal transformations. Thus, optimal transformations can be easily solved by ordinary least squares.

However, observed data are often skewed or heavy-tailed, and rarely normally distributed in real-world applications. Transformations are commonly used in reality to achieve normality in regression analysis, such as the celebrated Box-Cox transformation. Under the classical low-dimensional setting, estimating transformations for regression has received considerable consideration in the statistical literature. On one hand, parametric methods are proposed by focusing on the parametric families of transformations, for example, Box & Tidwell (1962), Box & Cox (1964), and others. On the other hand, nonparametric estimation of regression transformations are also studied, for instance, projection pursuit regression (Friedman & Stuetzle 1981), alternating conditional expectation (Breiman & Friedman 1985), additivity and variance stabilization (Stone 1985, Tibshirani 1988), among others. Although these methods work well under the classical setting, it is nontrivial to extend them to estimate optimal transformations in high dimensions. Under the high-dimensional setting, such parametric and nonparametric methods would suffer from the curse of dimensionality. Therefore, there are significant demands for relaxing normality when estimating optimal transformations for high-dimensional regression.

As a nice combination of flexibility and interpretability, Gaussian copulas have generated a lot of interests in statistics and econometrics. The semiparametric Gaussian copula model is deemed as a favorable alternative to the Gaussian model in several high-dimensional statistical problems, including linear discriminant analysis (Lin & Jeon 2003, Mai & Zou 2012), quadratic discriminant analysis (Fan et al. 2013), graphical modeling (Liu et al. 2009, 2012, Xue & Zou 2012), covariance matrix estimation (Xue & Zou 2014), and prin-
principal component analysis (Han & Liu 2012). The semiparametric Gaussian copula model provides a semiparametric generalization of the Gaussian model by assuming the existence of univariate monotone transformations \( f = (f_1, \cdots, f_p) \) for \( x = (x_1, \ldots, x_p)' \) such that \( f(x) = (f_1(x_1), \ldots, f_p(x_p)) \sim N_p(\mu^*, \Sigma^*) \). Throughout this paper, we follow Lin & Jeon (2003) to call this copula model as the transnormal model, which is also called the nonparamormal model in Liu et al. (2009). In this work, we will show the power of the transnormal model in estimating optimal transformations for high-dimensional regression modeling.

We now suppose that \( x = (z', y')' \) consists of \( z = (x_1, \ldots, x_{p_0})' \) and \( y = (x_{p_0+1}, \cdots, x_{p_0+q_0}=p)' \). The transnormal model entails the existence of monotone transformations \( f = (g, h) = (f_1, \ldots, f_{p_0}, f_{p_0+1}, \ldots, f_{p_0+q_0}) \) such that

\[
\begin{align*}
f(x) &= \begin{pmatrix} g(z) \\ h(y) \end{pmatrix} \sim N_p \left( \begin{pmatrix} \mu^*_z \\ \mu^*_y \end{pmatrix}, \Sigma^* \right), \\
\mu^* &= \begin{pmatrix} \mu^*_z \\ \mu^*_y \end{pmatrix}, \\
\Sigma^* &= \begin{pmatrix} \Sigma^*_{zz} & \Sigma^*_{zy} \\ \Sigma^*_{yz} & \Sigma^*_{yy} \end{pmatrix},
\end{align*}
\]

where we may assume that \( \mu^* = 0 \) and \( \Sigma^* \) is a correlation matrix. Remark that the marginal normality is achieved by transformations, so the transnormal model essentially assumes the joint normality of marginally normal-transformed variables. The transnormal model strikes a better balance between model robustness and interpretability than the normal model. In this work, our aim is to estimate transformations of predictors, \( t(z) : \mathbb{R}^{p_0} \mapsto \mathbb{R}^{q_0} \), to optimally predict multiple response variables \( y \) under the transnormal model (1).

Given the underlying transformations \( g \) and \( h \), we can easily derive the explicit conditional distribution of \( h(y) \) given \( g(z) \), namely

\[
\begin{align*}
h(y) | g(z) &\sim N_{q_0} \left( \Sigma^*_{yz} (\Sigma^*_{zz})^{-1} \cdot g(z), \Sigma^*_{yy} - \Sigma^*_{yz} (\Sigma^*_{zz})^{-1} \Sigma^*_{zy} \right).
\end{align*}
\]

But, since \( g \) and \( h \) are unknown, it is challenging to obtain the explicit conditional distribution of \( y \) given \( z \). Unlike the optimal transformations under the Gaussian model, nice properties such as linearity and homoscedasticity do not hold for the optimal transformations under the transnormal model. In the presence of nonlinear transformations \( g \) and \( h \), it is important to take into account the coordinatewise nonlinearity among the conditional distributions of \( y \) given \( z \). We follow the same spirit of quantile regression (Koenker 2005) to effectively deal with such nonlinearity.

Quantile regression is first introduced by the seminal paper of Koenker & Bassett (1978), and since then, it has received much attention in many topics such as survival analysis (Koenker & Geling 2001), time series analysis (Koenker & Xiao 2006), growth chart analysis (Wei & He 2006, Wei et al. 2006), microarray analysis (Wang & He 2007), variable selection
(Zou & Yuan 2008, Bradic et al. 2011), among others. Denote by $\rho_\tau(u) = u \cdot (\tau - I_{(u \leq 0)})$ the check loss function (Koenker & Bassett 1978). Then we consider multi-task quantile regression:

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

(3)

The use of the $\ell_1$ loss in prediction was recommended in Friedman (2001, 2002). Our work shares the similar philosophy with Friedman (2001, 2002), and includes the $\ell_1$ loss as a special case. In fact, the problem (3) with $\tau = \frac{1}{2}$ reduces to median regression, namely,

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

In this paper, we will show that with the aid of rank-based covariance regularization, the optimal transformations in multi-task quantile regression can be efficiently estimated under the transnormal model. Although estimating optimal transformation is very difficult, we surprisingly show that we can derive the closed-form solution for optimal transformations in (2) without using any smoothing techniques as in Friedman & Stuetzle (1981), Breiman & Friedman (1985) or Stone (1985), Tibshirani (1988). The key ingredient of our proposed method is the positive definite regularized estimation of large covariance matrices under the transnormal model. We introduce two novel rank-based covariance regularization methods to deal with two popular covariance structures respectively: the rank-based positive definite $\ell_1$ 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 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 & Zou 2012).

The aforementioned correlation matrix is not necessarily positive definite (Devlin et al. 1975). We therefore propose a new nearest correlation matrix projection that inherits required sampling properties of the adjusted Spearman’s or Kendall’s rank correlation matrix, which can be solved by an efficient alternating direction method of multipliers. By combining both strengths of quantile regression modeling and rank-based covariance regularization, we can simultaneously address issues of nonlinearity, non-normality 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 under the high-dimensional setting where dimension is on the nearly exponential order of sample size.

The rest of this paper is organized as follows. We first present the methodological details of optimal prediction in multi-task 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 appendices.

2 Multi-task quantile regression: model and method

This section presents the complete methodological details for solving optimal predictions in multi-task quantile regression, i.e., \(\min_{\mathbb{R}^{p_0} \rightarrow \mathbb{R}^{q_0}} \sum_{j=1}^{q_0} E[\rho_{\tau}(y_j - t_j(z))],\) where \(y\) and \(z\) jointly follow a transnormal model. The transnormal family of distributions allows us to obtain the closed-form solution of optimal predictions for multi-task quantile regression, which is very appealing and powerful to simultaneously deal with non-normality and high dimensionality. The transnormal model retains the nice interpretation of the normal model, and enables us to make a good use of normal model and theory. Moreover, the monotone transformation is easy to handle for quantile estimation.

2.1 The closed-form solution

Let \(Q_\tau(y_j|z)\) be the \(\tau\)-th quantile of the conditional distribution of \(y_j\) given \(z\), which is the analytical solution to \(\min_{t_j: \mathbb{R}^{p_0} \rightarrow \mathbb{R}^{q_0}} \sum_{j=1}^{q_0} E[\rho_{\tau}(y_j - t_j(z))].\) Denote by \(Q_\tau(y|z)\) the \(\tau\)-th equicoordinate quantile of the conditional distribution of \(y\) given \(z\), namely, \(Q_\tau(y|z) = (Q_\tau(y_1|z), \ldots, Q_\tau(y_p|z))'\). Thus, the \(\tau\)-th equicoordinate conditional quantile \(Q_\tau(y|z)\) is the exact solution to (3), i.e.

\[
Q_\tau(y|z) = \arg \min_{t: \mathbb{R}^{p_0} \rightarrow \mathbb{R}^{q_0}} \sum_{j=1}^{q_0} E[\rho_{\tau}(y_j - t_j(z))],
\]

By using the fact that \(h\) and \(g\) are monotone under the transnormal model, we have

\[
Q_\tau(y|z) = Q_\tau(y|g(z)) = h^{-1}(Q_\tau(h(y)|g(z))).
\]

Since \(h\) is monotonically nondecreasing, it now follows from (2) that

\[
Q_\tau(y|z) = h^{-1}\left(\Sigma_{yz}^* (\Sigma_{zz})^{-1} \cdot g(z) + \Phi^{-1}(\tau) \text{vdiag}(\Sigma_{yy}^* - \Sigma_{yz}^* (\Sigma_{zz})^{-1} \Sigma_{zy}^*)\right) \tag{4}
\]
where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and $\text{vdiag}(A)$ denotes the vector formed by the diagonal element of $A$. Therefore, the semiparametric Gaussian copula model enables us to obtain the closed-form solution to multi-task quantile regression.

Moreover, the closed-form solution (4) can be used to construct prediction intervals for predicting $y$ given $z$ in high dimensions. To be more specific, we can obtain the closed-form $100(1 - \tau)\%$ prediction interval as

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

For different values of $\tau$, we need only to adjust the value $\Phi(\tau)$. This is also an appealing feature of the transnormal model.

The closed-form solution $Q_\tau(y|z)$ uses true covariance matrices and transformations, and thus it is not a feasible estimator. To utilize (4), we need to estimate the covariance matrix $\Sigma^*$ and transformation functions $f$. It turns out that these two tasks are relatively easy under the transnormal model. Section 2.2 describes how to estimate transformation functions $f$ and then show how to estimate structured covariance matrix under the high-dimensional setting in Section 3. With estimated transformations $\hat{f} = (\hat{g}, \hat{h})$ and structured covariance matrix estimators $\hat{\Sigma}$, we can derive the following plug-in estimator

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

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

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

**Remark 1.** When $\tau = \frac{1}{2}$, it reduces to estimating optimal transformations from multi-task median regression. By using the simple fact that $\Phi^{-1}(\frac{1}{2}) = 0$, it can be further simplified as

$$Q_{\tau=\frac{1}{2}}(y|z) = h^{-1}\left(\Sigma^*_{yz}(\Sigma^*_{zz})^{-1} \cdot g(z)\right).$$

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

$$E(y|z) = E(y|g(z)) = E(h^{-1}(h(y))|g(z)).$$
But unlike the quantile regression, this can not be simplified further unless \( h \) is linear.

**Remark 3.** The nonlinearity of the transformations \( h(\cdot) \) in (4) makes the difference of conditional quantiles at different values of \( \tau \) depend on \( z \) and thus models the effect of heteroscedasticity. In contrast, Wu et al. (2010), Zhu et al. (2012) and Fan et al. (2013) model the heteroscedastic effect using the single-index quantile regression or semiparametric quantile regression by imposing the model \( y = \mu(z'\beta) + \sigma(z'\beta) \cdot \varepsilon \), where \( \sigma(z'\beta) \cdot \varepsilon \) is a heteroscedastic error and \( \varepsilon \) is independent of \( z \). Unlike the closed-form expression (4), it is difficult to employ semiparametric quantile regression to simultaneously deal with 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

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

Equation (6) provides a simple estimation of the transformation function \( f_j \). Let \( \tilde{F}_j(\cdot) \) be the empirical estimator of \( F_j(\cdot) \), i.e. \( \tilde{F}_j(u) = \frac{1}{n} \sum_{i=1}^{n} I\{x_{ij} \leq u\} \). Define the Winsorized empirical distribution function \( \hat{F}_j(\cdot) \) as

\[
\hat{F}_j(u) = \delta_n \cdot I\{\tilde{F}_j(u) < \delta_n\} + \tilde{F}_j(u) \cdot I\{\delta_n \leq \tilde{F}_j(u) \leq 1 - \delta_n\} + (1 - \delta_n) \cdot I\{\tilde{F}_j(u) > 1 - \delta_n\},
\]  

where \( \delta_n \) is the Winsorization parameter to avoid infinity values and to achieve better bias-variance tradeoff. Following Mai & Zou (2012), we specify the Winsorization parameter as \( \delta_n = \frac{1}{n^2} \), which facilitates both theoretical analysis and practical performance. Next, we can estimate the transformation functions \( f \) in the transnormal model as follows:

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

Remark that the estimators are nondecreasing and can be substituted into (4) to estimate the multi-task quantile regression function.

### 2.3 Estimation of correlation matrix

We present two covariance regularization methods based on \( i.i.d. \) transnormal data \( x_1, \ldots, x_n \): the rank-based positive definite \( \ell_1 \) penalization for estimating sparse covariance 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 estimate optimal transformations in multi-task quantile regression.

2.3.1 Positive definite sparse correlation matrix

Sparse covariance matrices are widely used in many applications where variables are permutation invariant. By truncating small entries to zero, thresholding (Bickel & Levina 2008b, Rothman et al. 2009, Fan et al. 2013) is a powerful approach to encourage (conditional) sparsity in estimating large covariance matrices. However, the resulting estimator may not be positive definite in practice. Xue et al. (2012) proposed a computationally efficient positive-definite $\ell_1$-penalized covariance estimator to address the indefiniteness issue of thresholding.

Now we extend Xue et al. (2012) to the transnormal model. First, we introduce the “oracle” positive-definite $\ell_1$-penalized estimator using the “oracle” transformations $f$, i.e.,

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

where $\hat{R}^{o}$ is the “oracle” sample correlation matrix of the “oracle” data $f(x_1), \ldots, f(x_n)$, and $\| \Sigma \|_{1,\text{off}}$ is the $\ell_1$ norm of all off-diagonal elements in $\Sigma$.

Motivated by $\hat{\Sigma}^{o}_{\ell_1}$, we can use the adjusted Spearman’s or Kendall’s rank correlation matrix (Kendall 1948) to derive a correlation matrix that is comparable with $\hat{R}^{o}$. For ease of presentation, we will focus on the adjusted Spearman’s rank correlation matrix throughout this paper, since the same analysis can be adapted to the adjusted Kendall’s rank correlation matrix. Let $r_j = (r_{1j}, r_{2j}, \ldots, r_{nj})'$ be the ranks for $(x_{1j}, x_{2j}, \ldots, x_{nj})'$. Denote by $\hat{r}_{jt} = \text{corr}(r_j, r_l)$ the Spearman’s rank correlation, and by $\tilde{r}_{jt} = 2 \sin(\frac{\pi}{6} \hat{r}_{jt})$ the adjusted Spearman’s rank correlation. It is well-known that $\tilde{r}_{jt}$ corrects the bias of $\hat{r}_{jt}$ (Kendall 1948). Now we consider the rank correlation matrix $\tilde{R}^s = (\tilde{r}_{jt})_{p \times p}$, which does not require estimating any transformation function. Thus, the rank-based positive definite $\ell_1$ penalization is as follows:

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

Remark that $\hat{\Sigma}^{s}_{\ell_1}$ will guarantee the critical positive definiteness, and it can be efficiently solved by the alternating direction method of multiplier in Xue et al. (2012).
2.3.2 Banded Cholesky decomposition regularization

When an ordering exists among variables, the bandable structure is commonly used in estimating large covariance matrices. Banding (Bickel & Levina 2008a) and tapering (Cai et al. 2010) were proposed to estimate bandable covariance matrices. But they have no guarantee of positive definiteness in practice. With the appealing positive definiteness, banded Cholesky decomposition regularization receives much attention, for example, Wu & Pourahmadi (2003), Huang et al. (2006), Bickel & Levina (2008a), Levina et al. (2008) and Rothman et al. (2010). In the sequel we propose the rank-based banded Cholesky decomposition regularization under the transnormal model.

First we introduce the “oracle” estimator to motivate our proposal. By using the “oracle” data \( f(x_1), \ldots, f(x_n) \), the “oracle” banded Cholesky decomposition regularization estimates the covariance matrix \( \Sigma^* \) through banding the Cholesky factor of its inverse \( \Theta^* \). Suppose that \( \Theta^* \) has the Cholesky decomposition \( \Theta^* = (I - A)^\dagger D^{-1}(I - A) \), where \( D \) is a \( p \times p \) diagonal matrix and \( A = (a_{jl})_{p \times p} \) is a lower triangular matrix with \( a_{11} = \cdots = a_{pp} = 0 \). Due to the fact that \( f(x) \sim N_p(0, \Sigma^*) \), it is easy to obtain that \((I - A) \cdot f(x) \sim N_p(0, D)\).

Let \( A = (a_1, \ldots, a_p)' \). As in Bickel & Levina (2008a), the “oracle” estimator is derived by regressing \( f_j(x_j) \) on its closest \( \min \{k, j - 1\} \) predecessors, i.e. \( \hat{\alpha}^o_1 = 0 \), and for \( j = 2, \ldots, p \),

\[
\hat{\alpha}^o_j = \arg \min_{\alpha_j \in A_j(k)} \frac{1}{n} \sum_{i=1}^n (f_j(x_{ij}) - \alpha_j' f(x_i))^2, \tag{8}
\]

where \( A_j(k) = \{(\alpha_1, \ldots, \alpha_p)' : \alpha_l = 0 \text{ if } l < j - k \text{ or } l \geq j\} \). Then \( A \) is estimated by the \( k \)-banded lower triangular matrix \( \hat{A}^o = (\hat{\alpha}^o_1, \ldots, \hat{\alpha}^o_p)' \), and \( D \) is estimated by the diagonal matrix \( \hat{D}^o = \text{diag}(\hat{d}^o_1, \ldots, \hat{d}^o_p) \) with \( \hat{d}^o_j \) being the residual variance, i.e.

\[
\hat{d}^o_j = \frac{1}{n} \sum_{i=1}^n (f_j(x_{ij}) - (\hat{\alpha}^o_j)' f(x_i))^2. \tag{9}
\]

Therefore, the “oracle” estimator ends up with a positive-definite estimator

\[
\hat{\Sigma}^o_{\text{chol}} = (I - \hat{A}^o)^{-1}(\hat{D}^o)^{-1}[(I - \hat{A}^o)^\dagger]^{-1},
\]

which has the \( k \)-banded precision matrix \( \hat{\Theta}^o_{\text{chol}} = (I - \hat{A}^o)'(\hat{D}^o)^{-1}(I - \hat{A}^o) \).

To mimic this “oracle” estimator, it is very important to observe that the “oracle” sample covariance matrix \( \hat{R}^o = \frac{1}{n} \sum_{i=1}^n f(x_i)(f(x_i))' \) plays the central role there. To see this point, we notice that estimating \( \hat{\alpha}^o_j \) and \( \hat{d}^o_j \) only depends on the quadratic term

\[
\frac{1}{n} \sum_{i=1}^n (f_j(x_{ij}) - \alpha_j' f(x_i))^2 = \alpha_j' \hat{R}^o \alpha_j - 2 \alpha_j' \hat{r}^o_j + \hat{r}^o_{jj}, \tag{10}
\]
where $\hat{R}^o = (\hat{r}_{jl})_{p \times p}$ and $\hat{r}_j^o = (\hat{r}_{j1}, \ldots, \hat{r}_{jp})'$ is its $j$-th row. Thus, we only need a positive definite correlation matrix estimator that is comparable with $\hat{R}^o$. The adjusted Spearman’s or Kendall’s rank correlation matrix achieves the “oracle”-like exponential rate of convergence, but it is not guaranteed to be positive definite (Devlin et al. 1975). We employ the nearest correlation matrix projection that inherits the desired sampling properties of the adjusted Spearman’s or Kendall’s rank correlation matrix, namely,

$$\hat{R}^s = \arg\min_{\hat{R}} \|\hat{R} - \hat{R}^s\|_{\text{max}} \text{ subject to } \hat{R} \succeq \epsilon I; \text{ diag}(\hat{R}) = 1, \quad (11)$$

where $\|\cdot\|_{\text{max}}$ is the entrywise $\ell_\infty$ norm and $\epsilon > 0$ is some arbitrarily small constant satisfying $\lambda_{\text{min}}(\Sigma^*) \geq \epsilon$, say $\epsilon = 10^{-4}$. Zhao et al. (2012) considered a related matrix projection, but they used a smooth surrogate function $\|\hat{R} - \hat{R}^s\|_{\nu_{\text{max}}} = \max_{\|U\|_1 \leq 1} \langle U, \hat{R} - \hat{R}^s \rangle - \frac{\nu}{2} \|U\|_F$, which would inevitably introduce unnecessary approximation error. Qi & Sun (2006) solved a related nearest correlation matrix projection under the Frobenius norm. We use the entrywise $\ell_\infty$-norm for theoretical considerations, as we now demonstrate. Notice that $\Sigma^*$ is a feasible solution to the nearest correlation matrix projection (11). Hence, $\|\hat{R}^s - \Sigma^*\|_{\text{max}} \leq \|\hat{R}^s - \hat{R}^s\|_{\text{max}}$ holds by definition. By using the triangular inequality, $\hat{R}^s$ retains almost the same sampling properties as $\hat{R}^s$ in terms of estimation bound, i.e.

$$\|\hat{R}^s - \Sigma^*\|_{\text{max}} \leq \|\hat{R}^s - \hat{R}^s\|_{\text{max}} + \|\hat{R}^s - \Sigma^*\|_{\text{max}} \leq 2\|\hat{R}^s - \Sigma^*\|_{\text{max}}.$$  

The details of nearest correlation matrix projection will be presented in the appendix.

Now we propose a feasible regularized rank estimator on the basis of $\hat{R}^s$. Let $\hat{R}^s = (\hat{r}_{jl})_{p \times p}$ and $\hat{r}_j^s = (\hat{r}_{j1}^s, \ldots, \hat{r}_{jp}^s)'$. We then substitute $\hat{R}^s$ into the quadratic term (10) as

$$a_j' \hat{R}^s a_j - 2a_j' \hat{r}_j^s + \hat{r}_{jj}^s = a_j' \hat{R}^s a_j - 2a_j' \hat{r}_j^s + 1,$$

where the fact that $\hat{r}_{jj}^s = 1$ is used. Accordingly, we estimate $A$ by

$$\hat{A}^s = (\hat{a}_1^s, \ldots, \hat{a}_p^s)'$$

with $\hat{a}_1^s = 0$, and for $j = 2, \ldots, p$,

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

where $A_j(k) = \{(a_1, \ldots, a_p) : a_i = 0 \text{ if } i < j - k \text{ or } i \geq j\}$. In addition, we estimate $D$ by

$$\hat{D}^s = \text{diag}(\hat{d}_1^s, \ldots, \hat{d}_p^s),$$

10
with
\[
\hat{d}_j^s = (\hat{a}_j^s)'\hat{R}^s\hat{a}_j^s - 2(\hat{a}_j^s)'\hat{r}_j^s + 1.
\]

Thus, the rank-based banded Cholesky decomposition regularization yields the estimator
\[
\hat{\Sigma}_{\text{chol}} = (I - \hat{A}^s)^{-1} \hat{D}^s [(I - \hat{A}^s)']^{-1},
\]
which has the \( k \)-banded precision matrix \( \hat{\Theta}_{\text{chol}} = (I - \hat{A}^s)'(\hat{D}^s)^{-1}(I - \hat{A}^s) \).

3 Theoretical properties

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

Throughout this section, we follow Bickel & Levina (2008a,b) to assume that
\[
\varepsilon_0 \leq \lambda_{\text{min}}(\Sigma^*) \leq \lambda_{\text{max}}(\Sigma^*) \leq \frac{1}{\varepsilon_0}.
\]

3.1 A general theory

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

(C1) There exists a regularized estimator \( \hat{\Sigma} \) satisfying the following concentration bound
\[
\| \hat{\Sigma} - \Sigma^* \|_{\ell_1} = O_P(\xi_{n,p}), \quad \text{with} \quad \xi_{n,p} = o((\log n)^{-1/2}).
\]

To simplify notation, we let \( \hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{zz} & \hat{\Sigma}_{zy} \\ \hat{\Sigma}_{yz} & \hat{\Sigma}_{yy} \end{pmatrix} \). In the following theorem, we show that the conditional distribution of \( h(y) \) given \( g(z) \) can be well estimated under Condition (C1).

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

(i) (on the conditional mean vector)
\[
\| \hat{\Sigma}_{yz} \hat{\Sigma}_{zz}^{-1} \cdot \hat{g}(z) - \Sigma^*_{yz}(\Sigma^*_{zz})^{-1} \cdot g(z) \|_{\max} = O_P \left( \sqrt{\frac{\log n \log p_0}{n^\kappa}} + \xi_{n,p} \sqrt{\log n} \right).
\]
(ii) (on the conditional variance matrix)

\[ \| (\hat{\Sigma}_{yy} - \hat{\Sigma}_{yz}\hat{\Sigma}^{-1}_{zz}\hat{\Sigma}_{zy}) - (\Sigma_{yy}^* - \Sigma_{yz}^*(\Sigma_{zz}^*)^{-1}\Sigma_{zy}^*) \|_{\ell_2} = O_P(\xi_{n,p}). \]

Next, we show that the plug-in estimator $\hat{Q}_\tau(y|z)$ is asymptotically as good as the closed-form solution $Q_\tau(y|z)$ under mild regularity conditions. Let $\psi_j(\cdot)$ be the probability density function of $X_j$. Define $L = \Sigma_{yy}^*(\Sigma_{zz}^*)^{-1} \cdot g(z) + \Phi^{-1}(\tau) \mathrm{vdiag}(\Sigma_{yy}^* - \Sigma_{yz}^*(\Sigma_{zz}^*)^{-1}\Sigma_{zy}^*)$ and

$\hat{L} = \hat{\Sigma}_{yz}(\hat{\Sigma}_{zz})^{-1} \cdot \hat{g}(z) + \Phi^{-1}(\tau) \mathrm{vdiag}(\hat{\Sigma}_{yy} - \hat{\Sigma}_{yz}(\hat{\Sigma}_{zz})^{-1}\hat{\Sigma}_{zy})$. Denote $L = (L_1, \ldots, L_{q_0})'$.

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

\[ \| \hat{Q}_\tau(y|z) - Q_\tau(y|z) \|_{\max} = O_P \left( \frac{1}{M} \sqrt{\log \frac{q_0}{n}} \right), \]

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 $j = 1, \ldots, q_0$, i.e., $M = \min_{j=1,\ldots,q_0} \min_{x \in \mathcal{I}_j} \psi_{p_0+j}(x)$.

Theorem 2 immediately implies that the plug-in estimator can be used to construct provable prediction intervals in high dimensions. For instance, we can use

$[\hat{Q}_{\tau_1}(y|z), \hat{Q}_{1-\tau_2}(y|z)]$

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

In what follows, we consider two parameter spaces for $\Sigma^*$ in the transnormal model,

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

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

These two parameter spaces were studied in Bickel & Levina (2008a,b) and Cai et al. (2010). In the sequel we show that the proposed rank-based covariance regularization achieves the “oracle”-like rate of convergence over $\mathcal{G}_q$ and $\mathcal{H}_\alpha$ under the matrix $\ell_1$ norm respectively. Therefore, we can obtain the optimal prediction result on the basis of intermediate theoretical results about rank-based covariance regularization.
3.2 Positive definite sparse correlation matrix

First we derive the estimation bound for the rank-based positive-definite $\ell_1$ penalization and its application to estimate the conditional distribution of $h(y)$ given $g(z)$.

**Theorem 3.** Assume the data follow the transnormal model with $\Sigma^* \in G_q$, and also assume that $n \gg \log p$. Suppose that $\lambda_{\min}(\Sigma^*) \geq \varepsilon_0 \gg s_0(\log p/n)^{1-\frac{q}{2}} + \epsilon$. With probability tending to 1, the rank-based positive-definite $\ell_1$ penalization with $\lambda = c(\log p/n)^{1/2}$ achieves the following upper bound under the matrix $\ell_1$ norm,

$$\sup_{\Sigma^* \in G_q} \| \hat{\Sigma}^*_{\ell_1} - \Sigma^* \|_{\ell_1} \leq C \cdot s_0 \left( \frac{\log p}{n} \right)^{1-\frac{q}{2}}.$$

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

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

and

$$\sup_{\Sigma^* \in G_q} \| \hat{\Sigma}_{yy} - \hat{\Sigma}_{yz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zy} - (\Sigma^*_{yy} - \Sigma^*_{yz} (\Sigma^*_{zz})^{-1} \Sigma^*_{zy}) \|_{\ell_2} = O_P \left( s_0 \left( \frac{\log p}{n} \right)^{1-\frac{q}{2}} \right).$$

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 $\hat{\Sigma} = \hat{\Sigma}^*_{\ell_1}$ we have

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

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 $h(y)$ given $g(z)$.

**Theorem 5.** Assume the data follow the transnormal model with $\Sigma^* \in H_\alpha$, 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 $\ell_1$ norm,

$$\sup_{\Sigma^* \in H_\alpha} \| \hat{\Sigma}^*_{\text{chol}} - \Sigma^* \|_{\ell_1} \leq C k \left( \frac{\log p}{n} \right)^{1/2} + C k^{-\alpha}.$$
If \( k = c \cdot \left( \frac{\log p}{n} \right)^{\frac{1}{\alpha + 1}} \) in the rank-based banded Cholesky decomposition regularization, we have

\[
\sup_{\Sigma^* \in \mathcal{H}_a} \left\| \hat{\Sigma}^{\text{chol}}_{s} - \Sigma^* \right\|_{\ell_1} = O_P \left( \left( \frac{\log p}{n} \right)^{\frac{\alpha}{2(\alpha + 1)}} \right).
\]

**Remark 5.** Bickel & 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 & Levina (2008a), we have

\[
\sup_{\Sigma^* \in \mathcal{H}_a} \left\| \hat{\Sigma}^{\text{chol}}_{o} - \Sigma^* \right\|_{\ell_1} = O_P \left( \left( \frac{\log p}{n} \right)^{\frac{\alpha}{2(\alpha + 1)}} \right).
\]

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 \( \hat{\Sigma} = \hat{\Sigma}^{s}_{\text{chol}} \) we have

\[
\sup_{\Sigma^* \in \mathcal{H}_a} \left\| \hat{\Sigma} - \Sigma^* \right\|_{\ell_1} = O_P \left( \left( \frac{\log p}{n} \right)^{\frac{\alpha}{2(\alpha + 1)}} \right).
\]

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 \( \hat{\Sigma} = \hat{\Sigma}^{s}_{\text{chol}} \) we have

\[
\sup_{\Sigma^* \in \mathcal{H}_a} \left\| \hat{\Omega}_{\tau}(y|z) - Q_{\tau}(y|z) \right\|_{\ell_2} = O_P \left( \left( \frac{\log p}{n} \right)^{\frac{\alpha}{2(\alpha + 1)}} \right).
\]

**4 Numerical properties**

This section examines the finite-sample performance of the proposed methods in Sections 2-4. For space consideration, we focus only on the rank-based banded Cholesky decomposition regularization and its application to multi-task median regression in numerical studies.
4.1 Simulation studies

In this simulation study, we numerically investigate the “oracle” estimator, the proposed rank-based estimator and the “naïve” estimator. We summarize notation and details of three regularized estimator in Table 1. The “oracle” estimator serves a benchmark in the numerical comparison, and the “naïve” estimator directly regularizes on the covariance of the original data.

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(\hat{\Sigma}^o_{\text{chol}}, \hat{\Theta}^o_{\text{chol}}, \hat{t}^o_1(z))$</td>
<td>regularizing the “oracle” sample correlation matrix</td>
</tr>
<tr>
<td>$(\hat{\Sigma}^s_{\text{chol}}, \hat{\Theta}^s_{\text{chol}}, \hat{t}^s_1(z))$</td>
<td>regularizing the adjusted Spearman’s rank correlation matrix</td>
</tr>
<tr>
<td>$(\hat{\Sigma}^n_{\text{chol}}, \hat{\Theta}^n_{\text{chol}}, \hat{t}^n_1(z))$</td>
<td>regularizing the usual sample correlation matrix</td>
</tr>
</tbody>
</table>

In Models 1-3, we consider three different designs for the inverse covariance matrix $\Omega^*$:

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

**Model 2**: $\Omega^*$: $\omega^*_{i,i} = 1$, $\omega^*_{i,i+1} = 0.5$, $\omega^*_{i,i+2} = 0.25$ and $\omega^*_{ij} = 0$ otherwise;

**Model 3**: $\Omega^*$: $\omega^*_{i,i} = 1$, $\omega^*_{i,i+1} = 0.4$, $\omega^*_{i,i+2} = \omega^*_{i,i+3} = 0.2$, $\omega^*_{i,i+4} = 0.1$ and $\omega^*_{ij} = 0$ otherwise.

Models 1-3 are autoregressive (AR) models widely used in time series analysis, and they were considered by Huang et al. (2006) and Xue & Zou (2012). Based upon covariance matrix $\Gamma^* = (\Omega^*)^{-1} = (\gamma^*_{ij})_{p \times p}$, we calculate the true correlation matrix $\Sigma^* = (\sigma^*_{ij})_{p \times p} = \left(\frac{\gamma^*_{ij}}{\sqrt{\gamma^*_{ii} \gamma^*_{jj}}}\right)_{p \times p}$ and also its inverse $\Theta^* = (\Sigma^*)^{-1}$. Then, we generate $n$ independent normal data from $N_p(0, \Sigma^*)$ and then transfer the normal data to the desired transnormal data $(x_1, ..., x_n)$ using the transformation function

$$[f_1^{-1}, f_2^{-1}, f_3^{-1}, f_4^{-1}, f_1^{-1}, f_2^{-1}, f_3^{-1}, f_4^{-1}, ...],$$  \hspace{1cm} (13)

where four monotone transformations are considered: $f_1(x) = x^{\frac{1}{3}}$ (power transformation), $f_2(x) = \log(x)$ (logit transformation), $f_3(x) = \log(x)$ (logarithmic transformation) and $f_4(x) = f_1(x)1_{x<-1} + f_2(x)1_{-1 \leq x \leq 1} + (f_3(x - 1) + 1)1_{x>1}$.

In all cases we let $n = 200$ and $p = 100, 200 & 500$. For each estimator, tuning parameter is chosen by cross-validation. Estimation accuracy is measured by the matrix $\ell_1$-norm and $\ell_2$-norm averaged over 100 independent replications.
Table 2: Performance of estimating the correlation matrix with $\hat{\Sigma}_{chol}^o$, $\hat{\Sigma}_{chol}^s$ and $\hat{\Sigma}_{chol}^n$. Estimation accuracy is measured by the matrix $\ell_1$-norm and $\ell_2$-norm. Each metric is averaged over 100 independent replications with standard errors in the bracket.

<table>
<thead>
<tr>
<th>Method</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p=100$</td>
<td>$p=200$</td>
<td>$p=500$</td>
</tr>
<tr>
<td>$\hat{\Sigma}_{chol}^o$</td>
<td>1.34 1.60 1.95</td>
<td>0.89 0.91 0.99</td>
<td>1.00 1.04 1.16</td>
</tr>
<tr>
<td></td>
<td>(0.05) (0.13) (0.27)</td>
<td>(0.03) (0.04) (0.08)</td>
<td>(0.02) (0.05) (0.12)</td>
</tr>
<tr>
<td>$\hat{\Sigma}_{chol}^s$</td>
<td>1.49 1.68 2.08</td>
<td>0.92 0.99 1.10</td>
<td>1.03 1.06 1.21</td>
</tr>
<tr>
<td></td>
<td>(0.07) (0.11) (0.21)</td>
<td>(0.03) (0.05) (0.14)</td>
<td>(0.02) (0.04) (0.12)</td>
</tr>
<tr>
<td>$\hat{\Sigma}_{chol}^n$</td>
<td>3.21 3.49 3.78</td>
<td>1.45 1.48 1.63</td>
<td>1.11 1.22 1.28</td>
</tr>
<tr>
<td></td>
<td>(0.06) (0.09) (0.24)</td>
<td>(0.02) (0.04) (0.11)</td>
<td>(0.03) (0.05) (0.11)</td>
</tr>
</tbody>
</table>

Tables 2-3 summarize the numerical performance of estimating the correlation matrix and its inverse for three banded Cholesky decomposition regularization methods. We can see that the “naïve” covariance regularization performs the worst in the presence of non-normality. The rank-based covariance regularization effectively deals with the transnormal data, and performs comparably with the “oracle” estimator. This numeric evidence is consistent with theoretical results presented in Section 3.

Next, we compare the performance of estimating optimal transformations for multi-task median regression. In each of the 100 replications, we simulate another $4 \times n$ independent transnormal data ($x_{n+1}, \ldots, x_{5n}$) as testing dataset. Now we take the first $p_0 = p/2$ variables in $x$ as response $y$ and take the second $q_0 = p/2$ variables in $x$ as predictor $z$. Now, given three regularized covariance estimators $\hat{\Sigma}_{chol}^o$, $\hat{\Sigma}_{chol}^s$ and $\hat{\Sigma}_{chol}^n$, we examine their performance in predicting $y_i$ based on $z_i$ for $i = n+1, \ldots, 5n$. Recall that $\hat{t}_1(z) = \hat{h}^{-1}(\hat{\Sigma}_{yz}(\hat{\Sigma}_{zz})^{-1} \hat{g}(z))$
Table 3: Performance of estimating the inverse correlation matrix with $\hat{\Theta}_{chol}^o$, $\hat{\Theta}_{chol}^s$ and $\hat{\Theta}_{chol}^n$. Estimation accuracy is measured by the matrix $\ell_1$-norm and $\ell_2$-norm. Each metric is averaged over 100 independent replications with standard errors in the bracket.

<table>
<thead>
<tr>
<th>Method</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p=100$</td>
<td>$p=200$</td>
<td>$p=500$</td>
</tr>
<tr>
<td></td>
<td>matrix $\ell_1$ norm</td>
<td>matrix $\ell_2$ norm</td>
<td></td>
</tr>
<tr>
<td>$\hat{\Sigma}_{chol}^o$</td>
<td>4.07</td>
<td>4.72</td>
<td>5.14</td>
</tr>
<tr>
<td></td>
<td>(0.21)</td>
<td>(0.37)</td>
<td>(0.57)</td>
</tr>
<tr>
<td>$\hat{\Sigma}_{chol}^s$</td>
<td>4.21</td>
<td>4.80</td>
<td>5.28</td>
</tr>
<tr>
<td></td>
<td>(0.24)</td>
<td>(0.32)</td>
<td>(0.55)</td>
</tr>
<tr>
<td>$\hat{\Sigma}_{chol}^n$</td>
<td>7.18</td>
<td>7.55</td>
<td>8.02</td>
</tr>
<tr>
<td></td>
<td>(0.22)</td>
<td>(0.35)</td>
<td>(0.54)</td>
</tr>
</tbody>
</table>

provides the closed-form solution for multi-task median regression. Then we can derive the closed-form solutions with respect to different regularized covariance estimators as follows:

- the “oracle” estimator: $\hat{t}_1^o(z) = h^{-1}(\hat{\Sigma}_{yz}^o(\hat{\Sigma}_{zz}^o)^{-1} \cdot g(z))$

- the rank-based estimator: $\hat{t}_1^s(z) = \hat{h}^{-1}(\hat{\Sigma}_{yz}^s(\hat{\Sigma}_{zz}^s)^{-1} \cdot \hat{g}(z))$

- the “naïve” estimator: $\hat{t}_1^n(z) = \hat{\mu}_y + \hat{\Sigma}_{yz}^n(\hat{\Sigma}_{zz}^n)^{-1} \cdot (z - \hat{\mu}_z)$

Tables 4 shows their numerical performances. Prediction accuracy is measured by the difference of prediction errors, which is defined by subtracting the “oracle” prediction error, i.e.

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

17
Table 4: Performance of relative prediction errors with $\hat{t}_1^o(z)$, $\hat{t}_1^s(z)$ and $\hat{t}_1^n(z)$. Estimation accuracy are measured by the $\ell_1$-norm and averaged over 100 independent replications, with the standard errors shown in the bracket.

<table>
<thead>
<tr>
<th>Method</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
<th>Model 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=100</td>
<td>p=200</td>
<td>p=500</td>
<td>p=100</td>
<td>p=200</td>
<td>p=500</td>
</tr>
<tr>
<td>$\hat{t}_1^s(z)$ vs. $\hat{t}_1^o(z)$</td>
<td>0.02</td>
<td>0.03</td>
<td>0.08</td>
<td>0.16</td>
<td>0.26</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>$\hat{t}_1^n(z)$ vs. $\hat{t}_1^o(z)$</td>
<td>0.11</td>
<td>0.19</td>
<td>0.21</td>
<td>13.50</td>
<td>26.07</td>
<td>63.80</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.01)</td>
<td>(0.03)</td>
<td>(0.22)</td>
<td>(0.58)</td>
<td>(2.16)</td>
</tr>
</tbody>
</table>

The relative prediction error is averaged over 100 independent replications.

As summarized in Table 4, we can see that the proposed rank-based estimator performs very similarly to its oracle counterpart, and greatly outperforms the “naïve” estimator.

### 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 et al. (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 et al. (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 the control data $\mathbf{x}_i^{co} = (x_{i,1}^{co}, \ldots, x_{i,218}^{co})$ for $i = 1, \ldots, 157$, and the cancer data $\mathbf{x}_j^{ca} = (x_{j,1}^{ca}, \ldots, x_{j,218}^{ca})$ for $j = 1, \ldots, 167$. We refer the readers to Adam et al. (2002) and Levina et al. (2008) for more details about this dataset and also the preprocessing procedure. Our aim now is to use the more stable intensity measurements in the latter 168 mass over charge ratio indices to optimally predict the more volatile intensity
measurements in the first 50 indices.

The analysis of Levina et al. (2008) is on the basis of the multivariate normal covariance matrix. Hence, we perform the normality test for both the control data and the cancel data. Table 5 shows the number of rejecting the null hypotheses among 218 normality tests. There is at least 50% of the mass spectroscopy measurements being non-normal at the significance level of 0.05. Even after the strict Bonferroni correction, there are at least 30 indices in the control data and 116 indices in the cancer data rejecting all the null hypotheses. Figure 1 further illustrates the non-normality issue (e.g. heavy tails, skewness) in two indices (109 & 218).

Table 5: Testing for normality. This table shows the number of rejecting the normality hypothesis at different significance levels among 218 mass over charge ratio indices.

<table>
<thead>
<tr>
<th>significance level</th>
<th>Anderson-Darling</th>
<th>Cramer-von Mises</th>
<th>Kolmogorov-Smirnov</th>
</tr>
</thead>
<tbody>
<tr>
<td>control data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>158</td>
<td>138</td>
<td>119</td>
</tr>
<tr>
<td>0.05/218</td>
<td>71</td>
<td>52</td>
<td>30</td>
</tr>
<tr>
<td>cancer data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>196</td>
<td>177</td>
<td>156</td>
</tr>
<tr>
<td>0.05/218</td>
<td>136</td>
<td>134</td>
<td>116</td>
</tr>
</tbody>
</table>

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

The nonnormality of measurements suggest us to use the transnormal model. Let \( z^{co} \) (\( y^{co} \)) and \( y^{ca} \) (\( y^{ca} \)) denote the intensity measurements in the first 160 and second 58 indices of the control (cancer) data respectively. We divide this dataset into training sets \( (x_{1}^{co}, \ldots, x_{120}^{co}) \)
\& (x_{1}^{co}, \ldots, x_{120}^{co}) \text{ and testing sets } (x_{121}^{co}, \ldots, x_{157}^{co}) \& (x_{121}^{co}, \ldots, x_{167}^{co}). \text{ Note that the sample estimator } \hat{t}_{1}^{\text{sample}}(z) = \hat{\mu}_{y} + \hat{\Sigma}_{yz}^{\text{sample}}(\hat{\Sigma}_{zz}^{\text{sample}})^{-1} \cdot (z - \hat{\mu}_{z}) \text{ is infeasible since the usual sample correlation matrix } \hat{\Sigma}_{zz}^{\text{sample}} \text{ is not invertible. Then we consider two different methods to predict } y: \text{ the proposed rank-based estimator } \hat{t}_{1}^{s}(z) = \hat{h}^{-1}(\hat{\Sigma}_{yz}^{s}(\hat{\Sigma}_{zz}^{s})^{-1} \cdot \hat{g}(z)) \text{ by using the proposed rank-based banded Cholesky decomposition regularization and the “na"ive” estimator } \hat{t}_{1}^{n}(z) = \hat{\mu}_{y} + \hat{\Sigma}_{yz}^{n}(\hat{\Sigma}_{zz}^{n})^{-1} \cdot (z - \hat{\mu}_{z}) \text{ by using the banded Cholesky decomposition regularization (Bickel \& Levina 2008a). Following Levina et al. (2008), tuning parameters are chosen via cross-validation based on the training data. The difference of prediction error at the } j\text{-th index is computed, namely,}

\[ \text{DPE}_{j}(\hat{t}_{1}(z)) = \text{PE}_{j}(\hat{t}_{1}^{s}(z)) - \text{PE}_{j}(\hat{t}_{1}^{n}(z)). \]

where \text{PE}_{j}(\hat{t}_{1}(z)) \text{ is prediction error at the } j\text{-th index computed by averaging the absolute prediction error } |\hat{t}_{1}(z_{i})_{j} - y_{ij}| \text{ over } i = 121, \ldots, 157 \text{ for the control data and over } i = 121, \ldots, 167 \text{ 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 “na"ive” estimator in 38 out of 50 indices for the control data and 46 out of 50 indices for the cancer data. Moreover, we further use } [\hat{Q}_{1\tau}(y|z), \hat{Q}_{1-\tau}(y|z)] \text{ to construct the } 100(1 - \tau)\% \text{ prediction interval. The predict target and prediction intervals are averaged over the testing set. Specifically, we estimate the } 100(1 - \tau)\% \text{ prediction interval as follows,}

\[ 100(1 - \tau)\% \text{ PI} = \left[ \frac{1}{34} \sum_{i=206}^{239} \hat{Q}_{1\tau}(y_{i}|z_{i}), \frac{1}{34} \sum_{i=206}^{239} \hat{Q}_{1-\tau}(y_{i}|z_{i}) \right]. \]

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 our proposal well estimates both end points of prediction intervals in practice, which is consistent with the asymptotic theory in Theorem 2.

**Appendix A: ADMM for nearest matrix projection**

The alternating direction method of multipliers (Boyd et al. 2011) has been widely applied to solving large-scale optimization problems in high-dimensional statistical learning, for example, covariance matrix estimation (Bien \& Tibshirani 2011, Xue et al. 2012) and graphical model selection (Ma et al. 2013, Danaher et al. 2013). Now we will design a new alternating
Figure 2: The differences of prediction errors between the existing Cholesky-based method (Bickel & Levina 2008a) and the proposed rank-based estimate with $\tau = 0.5$. The solid dots are the mass/charge ratio indices in which our proposed method outperforms the BL method. The proposed method outperforms in 38 out of 50 indices for the control data and 46 out of 50 indices 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.
direction method of multipliers to solve the nearest correlation matrix projection (11). To begin with, we introduce a new variable $S = R - \tilde{R}^s$ to split the positive definite constraint and the entrywise $\ell_\infty$ norm. Then we obtain the equivalent convex optimization problem as

$$ (\tilde{R}^s, \tilde{S}) = \arg \min_{(R,S)} \|S\|_{\max} \quad \text{subject to} \quad R \succeq \epsilon I; \quad S = R - \tilde{R}^s; \quad \text{diag}(S) = 0; \quad S = S'. \quad (14) $$

Let $\Lambda$ be the Lagrange multiplier associated with $S = R - \tilde{R}^s$ in (14). Next, we introduce the augmented Lagrangian function

$$ L_{\rho}(R, S; \Lambda) = \|S\|_{\max} - \langle \Lambda, R - S - \tilde{R}^s \rangle + \frac{1}{2\rho} \|R - S - \tilde{R}^s\|_F^2, \quad (15) $$

for any given constant $\rho > 0$. For $i = 0, 1, 2, \ldots$, we iteratively solve $(R^{i+1}, S^{i+1})$ through alternating minimization with respect to $R$ and $S$ respectively, i.e.

$$ (R^{i+1}, S^{i+1}) = \arg \min_{(R,S)} L_{\rho}(R, S; \Lambda^i), $$

and then update the Lagrange multiplier $\Lambda$ by

$$ \Lambda^{i+1} = \Lambda^i - \frac{1}{\rho} (R^{i+1} - S^{i+1} - \tilde{R}^s). \quad (18) $$

To sum up, the entire algorithm proceeds sequentially as follows till convergence,

$$ R \text{ step :} \quad R^{i+1} = \arg \min_{R \succeq \epsilon I} L_{\rho}(R, S^i; \Lambda^i); \quad (16) $$

$$ S \text{ step :} \quad S^{i+1} = \arg \min_{S = S'; \text{diag}(S) = 0} L_{\rho}(R^{i+1}, S; \Lambda^i); \quad (17) $$

$$ \Lambda \text{ step :} \quad \Lambda^{i+1} = \Lambda^i - \frac{1}{\rho} (R^{i+1} - S^{i+1} - \tilde{R}^s). \quad (18) $$

We note that both subproblems (16) and (17) indeed have simple solutions, which we will explain in the sequel, and thus we are able to avoid solving a sequence of complex subproblems. In what follows, we point out the explicit solutions to both (16) and (17).

First we consider the $R$ step. Denote by $(U)_+$ the projection of a matrix $U$ onto the convex set $\{R \succeq \epsilon I\}$. Suppose that $U$ has eigen-decomposition $\sum_{i=1}^p \lambda_i u_i u_i'$. Then, it is easy to see that $(U)_+ = \sum_{i=1}^p \max(\lambda_i, \epsilon) u_i u_i'$. The subproblem (16) is then solved as follows,

$$ R^{i+1} = \arg \min_{R \succeq \epsilon I} -\langle \Lambda^i, R \rangle + \frac{1}{2\rho} \|R - S^i - \tilde{R}^s\|_F^2 = (S^i + \tilde{R}^s + \rho \Lambda^i)_+. $$

We now consider the $S$ step. Define two subspaces $S = \{S : \text{diag}(S) = 0; \quad S = S'\}$ and $T = \{T : \|T\|_1 \leq \rho; \quad \text{diag}(T) = 0; \quad T = T'\}$. We also define

$$ M^{i+1} = R^{i+1} - \tilde{R}^s - \rho \Lambda^i. \quad (19) $$
Note that the subproblem in (17) can be equivalently written as follows,

\[ S^{i+1} = \arg \min_{S \in S} \|S\|_{\text{max}} + \frac{1}{2\rho} \|\hat{R}^{i+1} - S - \hat{R}^s\|_F^2, \]

\[ = \arg \min_{S \in S} \|S\|_{\text{max}} + \frac{1}{2\rho} \|S - M^{i+1}\|_F^2. \]

Lemma 1 will show that (17) can be solved by the projection onto the entrywise \(\ell_1\) ball.

**Lemma 1.** For any given symmetric matrix \(M = (m_{jl})_{p \times p}\), we define

\[ \hat{S} = \arg \min_{S \in S} \|S\|_{\text{max}} + \frac{1}{2\rho} \|S - M\|_F^2 \quad \text{and} \quad \hat{T} = \arg \min_{T \in T} \frac{1}{2\rho} \|T - M\|_F^2. \]

Let \(\hat{S} = (\hat{s}_{jl})_{p \times p}\) and \(\hat{T} = (\hat{t}_{jl})_{p \times p}\). Then for any \((j, l)\), we always have

\[ \hat{s}_{jl} = (m_{jl} - \hat{t}_{jl}) \times I\{j \neq l\}. \]

Now we define

\[ T^{i+1} = \arg \min_{T \in T} \|T - M^{i+1}\|_F^2. \quad (20) \]

We note that \(T^{i+1}\) is essentially a projection of a \(\frac{1}{2}p(p-1)\)-dimensional vector onto the \(\ell_1\) ball in \(\mathbb{R}^{\frac{1}{2}p(p-1)}\), and it can be efficiently solved by applying the exact projection algorithm (Duchi et al. 2008) in the \(O(p^2)\) expected time. Let \(T^{i+1} = (t^{i+1}_{jl})_{p \times p}\) and \(M^{i+1} = (m^{i+1}_{jl})_{p \times p}\). Then by Lemma 1, we can obtain the desired closed-form solution for (17), i.e.

\[ S^{i+1} = ((m^{i+1}_{jl} - t^{i+1}_{jl}) \times I\{j \neq l\})_{p \times p}. \]

---

**Algorithm 1** Proposed alternating direction method of multipliers for obtaining \(\hat{R}^s\).

1. Initialization: \(\rho\), \(S^0\) and \(\Lambda^0\).

2. Iterative alternating direction augmented Lagrangian step: for the \(i\)-th iteration

   2.1 Solve \(R^{i+1} = (S^i + \hat{R}^s + \rho \Lambda^i)_+\);

   2.2 Solve \(S^{i+1} = ((m^{i+1}_{jl} - t^{i+1}_{jl}) \times I\{j \neq l\})_{p \times p}\) with \(M^{i+1}\) and \(T^{i+1}\) as in (19) & (20);

   2.3 Update \(\Lambda^{i+1} = \Lambda^i - \frac{1}{\rho}(R^{i+1} - S^{i+1} - \hat{R}^s)\).

3. Repeat the above cycle till convergence.

The global convergence of Algorithm 1 can be obtained by following Xue et al. (2012). For space consideration, we omit the technical proof in this paper. The global convergence guarantees that Algorithm 1 always converges to the optimal solution \((\hat{R}^s, \hat{S}, \hat{\Lambda})\) of (14) from any initial value \((S^0, \Lambda^0)\) for any specified constant \(\rho > 0\).
Appendix B: Technical Proofs

Proof of Lemma 1

Proof. First we note that the dual norm of the entrywise $\ell_1$ norm is the entrywise $\ell_{\infty}$ norm. Then it is easy to see that $\|S\|_{\max} = \frac{1}{\rho} \max_{T \in \mathcal{T}} \langle T, S \rangle$. Now we have

$$\min_{S \in \mathcal{S}} \|S\|_{\max} + \frac{1}{2\rho} \|S - M\|_F^2$$

$$= \min_{S \in \mathcal{S}} \max_{T \in \mathcal{T}} \langle T, S \rangle + \frac{1}{2\rho} \|S - M\|_F^2$$

$$= \max_{T \in \mathcal{T}} \min_{S \in \mathcal{S}} \langle T, S \rangle + \frac{1}{2\rho} \|S - M\|_F^2$$

$$= \max_{T \in \mathcal{T}} \min_{S \in \mathcal{S}} \frac{1}{2} \|S - M + T\|_F^2 - \frac{1}{2} \|T - M\|_F^2 + \frac{1}{2} \|M\|_F^2.$$  

Notice that the optimal solution for the inner problem with respect to $S$ is given by

$$\hat{S} = (\hat{s}_{jl})_{p \times p} = ((m_{jl} - t_{jl}) \cdot I_{(j \neq l)})_{p \times p}.$$  

After we substitute $\hat{S}$ into the inner problem with respect to $S$, it suffices to solve the outer problem with respect to $T$, namely,

$$\max_{T \in \mathcal{T}} -\frac{1}{2} \|T - M\|_F^2 + \frac{1}{2} \|M\|_F^2 = \min_{T \in \mathcal{T}} \|T - M\|_F^2.$$  

Now by definition, $\hat{T} = (\hat{t}_{jl})_{p \times p}$ gives the optimal solution to the above problem with respect to $T$. Therefore, it is obvious to see that $\hat{s}_{jl} = (m_{jl} - \hat{t}_{jl}) \cdot I_{(j \neq l)}$ always holds for the optimal solutions $\hat{S}$ and $\hat{T}$. This completes the proof of Lemma 1.  

Proof of Theorem 1

Proof. In the first part, we bound $I = \|\hat{\Sigma}_{yz} \hat{\Sigma}_{zz}^{-1} \hat{g}(z) - \Sigma_{yz}^* (\Sigma_{zz}^*)^{-1} \hat{g}(z)\|_{\max}$. To this end, we bound $I_1 = \|\hat{\Sigma}_{yz} \hat{\Sigma}_{zz}^{-1} - \Sigma_{yz}^* (\Sigma_{zz}^*)^{-1}\|_{\ell_\infty}$ and $I_2 = \|\hat{\Sigma}(z) - \Sigma(z)\|_{\max} = O_P(\sqrt{\frac{\log n \log p_0}{n}})$ respectively. Let $\varphi_1 = \|(\Sigma_{zz}^*)^{-1}\|_{\ell_\infty}$ and $\varphi_2 = \|\Sigma_{yz}^*\|_{\ell_\infty}$ for ease of notation. Notice that

$$\hat{\Sigma}_{yz} \hat{\Sigma}_{zz}^{-1} - \Sigma_{yz}^* (\Sigma_{zz}^*)^{-1} = \Sigma_{yz}^* \cdot (\hat{\Sigma}_{zz}^{-1} - (\Sigma_{zz}^*)^{-1}) + (\hat{\Sigma}_{yz} - \Sigma_{yz}^*) \cdot (\Sigma_{zz}^*)^{-1}$$

$$+ (\hat{\Sigma}_{yz} - \Sigma_{yz}^*) \cdot (\hat{\Sigma}_{zz}^{-1} - (\Sigma_{zz}^*)^{-1}).$$  

By using the triangular inequality again, we then have

$$I_1 \leq \|\Sigma_{yz}^*\|_{\ell_\infty} \cdot \|\hat{\Sigma}_{zz}^{-1} - (\Sigma_{zz}^*)^{-1}\|_{\ell_\infty} + \|\Sigma_{zz}^*\|_{\ell_\infty} \cdot \|\hat{\Sigma}_{yz} - \Sigma_{yz}^*\|_{\ell_\infty}$$

$$+ \|\hat{\Sigma}_{yz} - \Sigma_{yz}^*\|_{\ell_\infty} \cdot \|\hat{\Sigma}_{zz}^{-1} - (\Sigma_{zz}^*)^{-1}\|_{\ell_\infty}$$

25
To bound $I_1$, we need to bound $\|\hat{\Sigma}_{yz} - \Sigma^*_y\|_{\ell_\infty}$ and $\|\hat{\Sigma}^{-1}_{zz} - (\Sigma^*_z)^{-1}\|_{\ell_\infty}$. By definition, it is easy to see that $\|\hat{\Sigma}_{yz} - \Sigma^*_y\|_{\ell_\infty} \leq \|\hat{\Sigma} - \Sigma^*\|_{\ell_\infty}$. Moreover, we have

$$\|\hat{\Sigma}^{-1}_{zz} - (\Sigma^*_z)^{-1}\|_{\ell_\infty} \leq C\|\hat{\Sigma}_{zz} - \Sigma^*_zz\|_{\ell_\infty} \cdot \|(\Sigma^*_z)^{-1}\|_{\ell_\infty} \leq C\varphi_1 \cdot \|\hat{\Sigma} - \Sigma^*\|_{\ell_\infty}.$$ 

Thus we can derive the explicit upper bound for $I_1$ as

$$I_1 \leq C(\varphi_1\varphi_2 + \varphi_1) \cdot \|\hat{\Sigma} - \Sigma^*\|_{\ell_\infty} = C(\varphi_1\varphi_2 + \varphi_1) \cdot \|\hat{\Sigma} - \Sigma^*\|_{\ell_1} = O_P(\xi_{n,p}). \quad (21)$$

Next, we prove that

$$I_2 = \|\hat{g}(z) - g(z)\|_{\max} = O_P(\sqrt{\frac{\log n \log p_0}{n^2}}). \quad (22)$$

To this end, note that $g_j(z_j)$ follows the standard normal distribution. Hence we have

$$\Pr(|g_j(z_j)| > \sqrt{(1 - \kappa) \log n}) = \int_{\sqrt{(1 - \kappa) \log n}}^{+\infty} C \cdot \exp(-t^2/2) \cdot dt \leq \frac{C}{\sqrt{\log n}} \cdot \int_{\sqrt{(1 - \kappa) \log n}}^{+\infty} \exp(-t^2/2) \cdot t dt \leq C \cdot \frac{1}{n^{1 - \kappa} \sqrt{\log n}}.$$

Then it immediately implies that

$$\|g(z)\|_{\max} = O_P(\sqrt{\log n}). \quad (23)$$

Let $\eta = \sqrt{(1 - \kappa) \log n}$. Now by using Lemma 1 of Mai & Zou (2012), we have

$$\Pr(\max_{|g_j(z_j)| \leq \eta} |\hat{g}_j(z_j) - g_j(z_j)| \geq \eta) \leq C \exp(-cn^{1/2} \eta^2 \log n) + p_0 \exp(-cn^{1/2} \eta^2 \log n)$$

Thus, we apply the union bound to obtain

$$\Pr(\|\hat{g}(z) - g(z)\|_{\max} \geq \eta) \leq \sum_{j=1}^{p_0} \Pr(|g_j(z_j)| > \eta) + \sum_{j=1}^{p_0} \Pr(\max_{|g_j(z_j)| \leq \eta} |\hat{g}_j(z_j) - g_j(z_j)| \geq \eta) \leq C \cdot \frac{p_0}{n^{1 - \kappa} \sqrt{\log n}} + Cn^{1/2} \exp(-cn^{1/2} \eta^2 \log n) + p_0 \exp(-cn^{1/2} \eta^2 \log n),$$

which immediately implies the upper bound (22). Now since

$$\hat{\Sigma}_{yz} \hat{\Sigma}_{zz}^{-1} \hat{g}(z) - \Sigma^*_y (\Sigma^*_z)^{-1} g(z) = (\hat{\Sigma}_{yz} \hat{\Sigma}_{zz}^{-1} - \Sigma^*_y (\Sigma^*_z)^{-1}) \cdot (\hat{g}(z) - g(z)) + (\Sigma^*_y (\Sigma^*_z)^{-1}) \cdot (\hat{g}(z) - g(z)) + (\hat{\Sigma}_{yz} \hat{\Sigma}_{zz}^{-1} - \Sigma^*_y (\Sigma^*_z)^{-1}) \cdot g(z),$$

26
we can combine (21), (22), (23) and also the triangle inequality to bound $I$ as

$$I \leq \|\hat{\Sigma}_y \hat{\Sigma}_zz^{-1} - \Sigma_{yz}^*(\Sigma_{zz}^*)^{-1}\|_{\ell_\infty} \cdot \|\hat{g}(z) - g(z)\|_{\max} + \|\Sigma_{yz}^*(\Sigma_{zz}^*)^{-1}\|_{\ell_\infty} \cdot \|\hat{g}(z) - g(z)\|_{\max} + \|\hat{\Sigma}_y \hat{\Sigma}_zz^{-1} - \Sigma_{yz}^*(\Sigma_{zz}^*)^{-1}\|_{\ell_\infty} \cdot \|g(z)\|_{\max}$$

$$= O_p(\sqrt{\log n \log p_0 \ell_p} + \xi_{n,p} \sqrt{\log n}).$$

In the second part, we bound $J = \|\hat{\Sigma}_{yy} - \Sigma_{yy}^*\|_{\ell_2} + \|\hat{\Sigma}_{yz} \hat{\Sigma}_zz^{-1} \hat{\Sigma}_{zy} - \Sigma_{yz}^*(\Sigma_{zz}^*)^{-1} \hat{\Sigma}_{zy}\|_{\ell_2} \equiv J_1 + J_2$. By using the triangular inequality, we have

$$J_2 = \hat{\Sigma}_{yz} (\Sigma_{zz}^*)^{-1}(\hat{\Sigma}_{zy} - \Sigma_{zy}^*) + (\hat{\Sigma}_{yz} - \Sigma_{yz}^*)(\Sigma_{zz}^*)^{-1} \hat{\Sigma}_{zy} + \hat{\Sigma}_{yz}(\hat{\Sigma}_zz^{-1} - (\Sigma_{zz}^*)^{-1}) \hat{\Sigma}_{zy}. $$

By using the definition of spectral norm, we have

$$\|\hat{\Sigma} - \Sigma^*\|_{\ell_2} = \max_{u=(u_x',u_y'):\|u\|_{\ell_2} = 1} \|\hat{\Sigma}_{zz} - \Sigma_{zz}^* - \Sigma_{yz}^*(\Sigma_{zz}^*)^{-1} \hat{\Sigma}_{zy} + (\hat{\Sigma}_{yz} - \Sigma_{yz}^*)(\Sigma_{zz}^*)^{-1} \hat{\Sigma}_{zy} + \hat{\Sigma}_{yz}(\hat{\Sigma}_zz^{-1} - (\Sigma_{zz}^*)^{-1}) \hat{\Sigma}_{zy}\|_{\ell_2}$$

$$\geq \max_{u_x:\|u_x\|_{\ell_2} = 1} \|\hat{\Sigma}_{zz} - \Sigma_{zz}^* \hat{\Sigma}_{zy} - \Sigma_{zy}^* (\Sigma_{zz}^*)^{-1} \hat{\Sigma}_{zy}\|_{\ell_2}$$

$$\geq \max_{u_x:\|u_x\|_{\ell_2} = 1} \|\hat{\Sigma}_{yz} - \Sigma_{yz}^* u_x\|_{\ell_2}$$

By similar arguments, we also have $\|\hat{\Sigma} - \Sigma^*\|_{\ell_2} \geq \|\hat{\Sigma}_{zy} - \Sigma_{zy}^*\|_{\ell_2}$. Given that $\varepsilon_0 \leq \lambda_{\min}(\Sigma^*) \leq \lambda_{\max}(\Sigma^*) \leq \frac{1}{\varepsilon_0}$, we then apply the sub-multiplicative property to obtain

$$\|\hat{\Sigma}_{yz} (\Sigma_{zz}^*)^{-1}(\hat{\Sigma}_{zy} - \Sigma_{zy}^*)\|_{\ell_2} \leq C \cdot \|\hat{\Sigma} - \Sigma^*\|_{\ell_2}$$

$$\|(\hat{\Sigma}_{yz} - \Sigma_{yz}^*)(\Sigma_{zz}^*)^{-1} \Sigma_{zy}\|_{\ell_2} \leq C \cdot \|\hat{\Sigma} - \Sigma^*\|_{\ell_2}$$

$$\|\hat{\Sigma}_{yz}(\hat{\Sigma}_zz^{-1} - (\Sigma_{zz}^*)^{-1}) \hat{\Sigma}_{zy}\|_{\ell_2} \leq C \cdot \|\hat{\Sigma} - \Sigma^*\|_{\ell_2}$$

Thus we can combine the above bounds to bound $J_2$ as

$$J_2 = \|\hat{\Sigma}_{yz} \hat{\Sigma}_zz^{-1} \hat{\Sigma}_{zy} - \Sigma_{yz}^*(\Sigma_{zz}^*)^{-1} \hat{\Sigma}_{zy}\|_{\ell_2} \leq C \cdot \|\hat{\Sigma} - \Sigma^*\|_{\ell_2}$$

In addition, due to the fact that $J_1 = \|\hat{\Sigma}_{yy} - \Sigma_{yy}^*\|_{\ell_2} \leq \|\hat{\Sigma} - \Sigma^*\|_{\ell_2}$, we have

$$J \leq J_1 + J_2 \leq C \cdot \|\hat{\Sigma} - \Sigma^*\|_{\ell_2} = C \cdot \|\hat{\Sigma} - \Sigma^*\|_{\ell_1} = O_P(\xi_{n,p})$$

This completes the proof of Theorem 1. □

27
Proof of Theorem 2

Proof. To simplify notation, we define \( \mu_{y|z} = \Sigma_{yz}^*(\Sigma_{zz}^*)^{-1} \cdot g(z) \), \( \Sigma_{y|z} = \Sigma_{y} - \Sigma_{yz}^*(\Sigma_{zz}^*)^{-1} \Sigma_{zy} \), \( \hat{\mu}_{y|z} = \hat{\Sigma}_{yz}(\hat{\Sigma}_{zz})^{-1} \cdot \hat{g}(z) \) and \( \hat{\Sigma}_{y|z} = \hat{\Sigma}_{y} - \hat{\Sigma}_{yz}(\hat{\Sigma}_{zz})^{-1} \hat{\Sigma}_{zy} \). Then \( L = \mu_{y|z} + \Phi^{-1}(\tau) \ vdiag(\Sigma_{y|z}) \) and \( \hat{L} = \hat{\mu}_{y|z} + \Phi^{-1}(\tau) \ vdiag(\hat{\Sigma}_{y|z}) \). In addition, we introduce \( \hat{Q}_\tau(y|z) = h^{-1}(\hat{L}) \).

In order to bound \( \hat{Q}_\tau(y|z) - Q_\tau(y|z) \), we use the triangle inequality to obtain that

\[
\| \hat{Q}_\tau(y|z) - Q_\tau(y|z) \|_{max} \leq \| \hat{Q}_\tau(y|z) - \hat{Q}_\tau(y|z) \|_{max} + \| \hat{Q}_\tau(y|z) - Q_\tau(y|z) \|_{max} \\
\leq \| \hat{h}^{-1}(\hat{L}) - h^{-1}(L) \|_{max} + \| h^{-1}(\hat{L}) - h^{-1}(L) \|_{max} \\
= I + J.
\]

In the sequel, we derive probability bounds for \( I \) and \( J \) respectively.

To bound \( I \), we first use the triangle inequality again to bound

\[
\| \hat{L} - L \|_{max} \leq \| \hat{\mu}_{y|z} - \mu_{y|z} \|_{max} + \Phi^{-1}(\tau) \cdot \| vdiag(\hat{\Sigma}_{y|z}) - vdiag(\Sigma_{y|z}) \|_{max} \\
\leq \| \hat{\mu}_{y|z} - \mu_{y|z} \|_{max} + \Phi^{-1}(\tau) \cdot \| \hat{\Sigma}_{y|z} - \Sigma_{y|z} \|_{\ell_2} \\
\leq O_P \left( \sqrt{\frac{\log n \log p_0}{n^\kappa}} + \xi_{n,p} \sqrt{\log n} \right) + \Phi^{-1}(\tau) \cdot O_P(\xi_{n,p}) \\
= O_P \left( \sqrt{\frac{\log n \log p_0}{n^\kappa}} + \xi_{n,p} \sqrt{\log n} \right)
\]

Notice that \( \sqrt{n^{-\kappa} \log n \log p_0} + \xi_{n,p} \sqrt{\log n} = o(1) \leq |L| \). Thus, with probability 1, we have

\[
|\hat{L}| \leq 2 \cdot |L|, \quad \text{and then,} \quad \Phi(-2 \cdot |L|) \leq \Phi(\hat{L}) \leq \Phi(2 \cdot |L|). \tag{24}
\]

In view of (24), we use the probability inequality from Page 75 of Serfling (1980) to obtain

\[
\Pr(|\hat{F}_{p_0+j}^{-1}(t) - F_{p_0+j}^{-1}(t)| > \epsilon) \leq 2 \exp(-2n\delta^2 \epsilon)
\]

where \( \delta = \min_{t \in [-2L_j, 2L_j]} |F_j(t + \epsilon) - F_j(t)|, \) for any \( j = 1, \ldots, q_0 \) and \( t \in [\Phi(-2|L_j|), \Phi(2|L_j|)] \subset \mathcal{L}_j \). Recall that \( \hat{f} = (\Phi^{-1} \circ \hat{F}_1, \ldots, \Phi^{-1} \circ \hat{F}_p) = (\hat{g}, \hat{h}) \). Note that \( \hat{h}_j^{-1}(\hat{L}_j) - h_j^{-1}(\hat{L}_j) = \hat{F}_j^{-1}(\Phi(\hat{L}_j)) - F_j^{-1}(\Phi(\hat{L}_j)) \) is the \( j \)-th element of \( \hat{h}^{-1}(\hat{L}) - h^{-1}(L) \) for
\(j = 1, \ldots, q_0\). Now we use the union bound to derive

\[
\Pr(\|\hat{\mathbf{Q}}_\tau(y|z) - \check{\mathbf{Q}}_\tau(y|z)\|_{\text{max}} > \epsilon) = \sum_{j=1}^{q_0} \Pr(\|\hat{h}^{-1}_j(\hat{L}_j) - h^{-1}_j(\check{L}_j)\| > \epsilon)
\leq \sum_{j=1}^{q_0} \Pr(\|\hat{F}_{p_{0}+j}(\Phi(\check{L}_j)) - F_{p_{0}+j}^{-1}(\Phi(\check{L}_j))\| > \epsilon)
\leq \sum_{j=1}^{q_0} \max_{t \in [\Phi(-2|L_j|), \Phi(2|L_j|)]} \Pr(\|\hat{F}_{p_{0}+j}(t) - F_{p_{0}+j}^{-1}(t)\| > \epsilon)
\leq q_0 \cdot \exp(-2nM^2\epsilon^2)
\]

where we use the fact that \(\delta_\epsilon \geq \min_{t \in \mathbb{Z}_j} \psi_j(t)\epsilon \geq M\epsilon\) in the last inequality. Thus, we have

\[
\|\hat{\mathbf{Q}}_\tau(y|z) - \check{\mathbf{Q}}_\tau(y|z)\|_{\text{max}} = O_P\left(\frac{1}{M} \sqrt{\frac{\log q_0}{n}}\right)
\]

To bound \(J\), we apply the mean-value theorem to the \(j\)-th element of \(h^{-1}(\hat{L}) - h^{-1}(L)\), that is, \(h^{-1}_j(\hat{L}_j) - h^{-1}_j(L_j) = F^{-1}_j(\Phi(\check{L}_j)) - F^{-1}_j(\Phi(L_j))\). Namely,

\[
h^{-1}_j(\hat{L}_j) - h^{-1}_j(L_j) = (h^{-1}_j)'(L_j) \cdot (\hat{L}_j - L_j) = \frac{\phi(\check{L}_j)}{\psi_{p_{0}+j}(F_{p_{0}+j}^{-1}(\Phi(\check{L}_j)))} \cdot (\hat{L}_j - L_j)
\]

where \(\hat{L}_j\) is no the line segment between \(\hat{L}_j\) and \(L_j\), and in the second equality we use the fact that \((h^{-1}_j)'(t) = (F_{p_{0}+j}^{-1} \circ \Phi)'(t) = \phi(t)/\psi_{p_{0}+j}(F_{p_{0}+j}^{-1}(\Phi(t)))\). With probability 1, we have

\[
\|\hat{L} - L\| \leq \|\hat{L} - L\| + \|L\| \leq \|\hat{L} - L\| + \|L\| \leq 2\|L\|.
\]

Then, \(\phi(\check{L}_j)/\psi_{p_{0}+j}(F_{p_{0}+j}^{-1}(\Phi(\check{L}_j)))) \leq \phi(\check{L}_j)/(\min_{x \in \mathbb{Z}_j} \psi_{p_{0}+j}(x)) \leq (M\sqrt{2\pi})^{-1}.\) Thus we have

\[
\|\check{\mathbf{Q}}_\tau(y|z) - \mathbf{Q}_\tau(y|z)\|_{\text{max}} = \|\hat{\mathbf{h}}^{-1}(\hat{L}) - \mathbf{h}^{-1}(L)\|_{\text{max}}
\leq \max_{j=1, \ldots, q_0} \frac{\phi(\check{L}_j)}{\psi_{p_{0}+j}(F_{p_{0}+j}^{-1}(\Phi(\check{L}_j))))} \cdot \|\hat{L} - L\|_{\text{max}}
\leq O_P\left(\frac{1}{M} \sqrt{\frac{\log n \log p_0}{n^\kappa}} + \frac{\xi_{n,p}}{M} \sqrt{\log n}\right)
\]

Therefore, we obtain the desired bound for \(\hat{\mathbf{Q}}_\tau(y|z) - \mathbf{Q}_\tau(y|z)\) as follows,

\[
\|\hat{\mathbf{Q}}_\tau(y|z) - \mathbf{Q}_\tau(y|z)\|_{\text{max}} \leq I + J = O_P\left(\frac{1}{M} \sqrt{\frac{\log q_0}{n}} + \frac{1}{M} \sqrt{\frac{\log n \log p_0}{n^\kappa}} + \frac{\xi_{n,p}}{M} \sqrt{\log n}\right).
\]

This completes the proof of Theorem 2. \(\Box\)
Proof of Theorem 3

Proof. Define the rank-based soft-thresholding estimator
\[
\hat{\Sigma}_{soft}^s = \left( s_\lambda(\hat{r}_{jl}) \right)_{p \times p}
\]
where \( s_\lambda(\cdot) \) applies the soft-thresholding rule for the off-diagonal elements of \( \hat{R}^s \). Under the event \( \{ \| \hat{R}^s - \Sigma^* \|_{\text{max}} \leq c \cdot (n^{-1} \log p)^{1/2} \} \), we can follow the same line of proof as in Bickel & Levina (2008b) to show that
\[
\sup_{\Sigma^* \in G_q} \| \hat{\Sigma}_{soft}^s - \Sigma^* \|_{\ell_1} \leq C \cdot s_0 \left( \frac{\log p}{n} \right)^{1/2}.
\]
Given that \( \lambda_{\min}(\Sigma^*) \geq \varepsilon_0 \gg s_0 (\log p/n)^{1-q/2} + \epsilon \), it is easy to see that \( \lambda_{\min}(\hat{\Sigma}_{soft}^s) \geq \epsilon \). Thus, \( \hat{\Sigma}_{soft}^s \) is the unique solution to the convex rank-based positive-definite \( \ell_1 \) penalization under the same event, which immediately implies that
\[
\sup_{\Sigma^* \in G_q} \| \hat{\Sigma}_{soft}^s - \Sigma^* \|_{\ell_1} \leq C \cdot s_0 \left( \frac{\log p}{n} \right)^{1/2}.
\]
We cite the entry-wise estimation bound of \( \hat{R}^s \) from Xue & Zou (2012) that
\[
\Pr(\| \hat{R}^s - \Sigma^* \|_{\text{max}} \leq c \cdot (n^{-1} \log p)^{1/2}) \to 1,
\]
as \( n \) tends to \( \infty \). This completes the proof of Theorem 2. \( \square \)

Proof of Theorem 5

Proof. As demonstrated in Bickel & Levina (2008a), it is sufficient to bound \( \| \hat{\Theta}_{chol}^s - \Theta^* \|_{\ell_1} \). Denote by \( B_k(U) = (u_{jl}I_{\{j-l|l| \leq k\}})_{p \times p} \) the \( k \)-banded estimator of \( U = (u_{jl})_{p \times p} \). Let \( A_k = B_k(A) \) be the \( k \)-banded estimator of \( A \), and \( D_k \) be the corresponding diagonal matrix. Define \( \Theta_k = (I - A_k)^t D_k^{-1} (I - A_k) \). Denote by \( B_k(\Theta^*) \) the \( k \)-banded estimator of \( \Theta^* \).

We first apply the triangle inequality to obtain
\[
\| \hat{\Theta}_{chol}^s - \Theta^* \|_{\ell_1} \leq \| \hat{\Theta}_{chol}^s - \Theta_k \|_{\ell_1} + \| \Theta_k - B_k(\Theta^*) \|_{\ell_1} + \| B_k(\Theta^*) - \Theta^* \|_{\ell_1}. \tag{25}
\]
Notice that \( \| \Theta_k - B_k(\Theta^*) \|_{\ell_1} \) in (25) can be bounded as in Bickel & Levina (2008a), i.e.
\[
\| \Theta_k - B_k(\Theta^*) \|_{\ell_1} \leq Ck^{-\alpha}.
\]
Also by the definition of $H_\alpha$, it is easy to see that $\|B_k(\Theta^*) - \Theta^*\|_{\ell_1} \leq C k^{-\alpha}$ holds. Then it is sufficient for us to bound $\|\hat{\Theta}_{\text{chol}}^s - \Theta_k\|_{\ell_2}$ only.

Under the event $\{\|\hat{R}^s - \Sigma^*\|_{\max} \leq \zeta\}$ with $k\zeta = o(1)$, we follow (A14) – (A15) in Bickel & Levina (2008a) to show that $\|\hat{A}^s - A_k\|_{\max} \leq C \zeta$, and meanwhile, follow (A16) – (A17) in Bickel & Levina (2008a) to show that $\|\hat{D}^s - D_k\|_{\max} \leq C k^2 \zeta^2$. Then by using triangular inequality, we obtain the upper bound for $\|\hat{\Theta}_{\text{chol}}^s - \Theta_k\|_{\ell_1}$ under the same event as follows,

$$
\|\hat{\Theta}_{\text{chol}}^s - \Theta_k\|_{\ell_1} = \|(I - \hat{A}^s)'(\hat{D}^s)^{-1}(I - \hat{A}^s) - (I - A_k)'D_k^{-1}(I - A_k)\|_{\ell_1}
\leq \|(A_k - \hat{A}^s)'D_k^{-1}(I - \hat{A}^s)\|_{\ell_1} + \|(I - A_k)'D_k^{-1}(A_k - \hat{A}^s)\|_{\ell_1} +
\|(I - \hat{A}^s)'((\hat{D}^s)^{-1} - D_k^{-1})(I - \hat{A}^s)\|_{\ell_1}
\leq C\|A_k - \hat{A}^s\|_{\ell_1} + C\|\hat{D}^s - D_k\|_{\ell_1},
$$

where the last inequality used the fact that $\|D_k^{-1}\|_{\ell_1} \leq C_1$ and $\|I - A\|_{\ell_1} \leq C_2$ by Lemma A.2 of Bickel & Levina (2008a) and also the fact that $k\zeta = o(1)$. Note that

$$
\|\hat{D}^s - D_k\|_{\ell_1} \leq \|\hat{D}^s - D_k\|_{\max} \leq Ck^2\zeta^2
$$

and

$$
\|A_k - \hat{A}^s\|_{\ell_1} \leq k \cdot \|A_k - \hat{A}^s\|_{\max} \leq Ck\zeta.
$$

By using the fact that $k\zeta = o(1)$ again, we have

$$
\|\hat{\Theta}_{\text{chol}}^s - \Theta_k\|_{\ell_1} \leq Ck\lambda + Ck^2\lambda^2 \leq Ck\zeta.
$$

Next we cite an entry-wise estimation bound from Xue & Zou (2012) that for any positive quantity $\zeta$ satisfying $n\zeta = o(1)$, there exists some fixed constant $c$ such that

$$
\Pr(\|\hat{R}^s - \Sigma^*\|_{\max} > \zeta) \leq p^2 \exp(-cn\zeta^2).
$$

Since $\hat{R}^s$ is the nearest correlation matrix projection of $\tilde{R}^s$, then we have

$$
\|\hat{R}^s - \Sigma^*\|_{\max} \leq 2\|\tilde{R}^s - \Sigma^*\|_{\max} \leq 2\lambda.
$$

Thus, taking $\zeta = c \cdot n^{-1/2} \log^{1/2} p$ yields the upper bound for $\|\hat{\Theta}_{\text{chol}}^s - \Theta^*\|_{\ell_1}$. This completes the proof of Theorem 3. \qed
Acknowledgement

Jianqing Fan’s research is supported in part by R01GM100474-04 and National Science Foundation grants DMS-1206464 and DMS-1406266. Hui Zou’s research is supported in part by grants from the National Science Foundation grant DMS-0846068 and a grant from the Office of Naval Research. Lingzhou Xue was supported by the National Institutes of Health grant R01-GM072611-09.

References


