Statistica Sinica Preprint No: SS-2016-0041R1	
Title	High-Dimensional Gaussian Copula Regression: Adaptive Estimation and Statistical Inference
Manuscript ID	SS-2016-0041R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0041
Complete List of Authors	Linjun Zhang and T. Tony Cai
Corresponding Author	Linjun Zhang
E-mail	linjunz@wharton.upenn.edu, zlj11112222@

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

High-Dimensional Gaussian Copula Regression: Adaptive Estimation and Statistical Inference

T. Tony Cai and Linjun Zhang

University of Pennsylvania

Abstract: We develop adaptive estimation and inference methods for high-dimensional Gaussian copula regression that achieve the same optimality without the knowledge of the marginal transformations as that for high-dimensional linear regression. Using a Kendall's tau based covariance matrix estimator, an ℓ_1 regularized estimator is proposed and a corresponding de-biased estimator is developed for the construction of the confidence intervals and hypothesis tests. Theoretical properties of the procedures are studied and the proposed estimation and inference methods are shown to be adaptive to the unknown monotone marginal transformations. Prediction of the response for a given value of the covariates is also considered. The procedures are easy to implement and perform well numerically. The methods are also applied to analyze the Communities and Crime Unnormalized Data from the UCI Machine Learning Repository.

Key words and phrases: Adaptive estimation, confidence interval, de-biased estimator, Gaussian copula regression, hypothesis testing, Kendall's tau, linear regression, optimal rate of convergence.

1. Introduction

Finding the relationship between a response and a set of covariates is a ubiquitous problem in scientific studies. Linear regression analysis, which occupies a central position in statistics, is arguably the most commonly used method. It has been well studied in both the conventional low-dimensional and contemporary high-dimensional settings. However,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

the assumption of a linear relationship between the predictors and the response is often too restrictive and unrealistic. Data transformations, such as the Box-Cox transformation, Fisher's z transformation, and variance stabilization transformation, have been frequently used to improve the linear fit and to correct violations of model assumptions such as constant error variance. These transformations are often required to be prespecified before applying the linear regression analysis. See, for example, Carroll and Ruppert (1988) for detailed discussions on transformations.

For a response Y and predictors $X_1, ..., X_p$, the following functional form of the relationship has been widely used in a range of applications,

$$f_{\lambda_0}(Y) = \boldsymbol{\beta}_0 + \sum_{j=1}^{p} \boldsymbol{\beta}_j f_{\lambda_j}(X_j) + \epsilon, \tag{1.1}$$

where $f_{\lambda_j}(\cdot)$ are univariate functions and λ_j is the parameter associated with f_{λ_j}. Examples of this model include the additive regression model, single index model, copula regression model, and semiparametric proportional hazards models, see Ravikumar et al. (2009); Meier, Van de Geer, and Buhlmann (2009); Yuan and Zhou (2015); Ni, Cook, and Tsai (2005); Yu et al. (2013); Yi, Wang, and Liu (2015); Foster, Taylor, and Nan (2013); Masarotto and Varin (2012); Pitt, Chan, and Kohn (2006); Luo and Ghosal (2016). For applications in econometrics, computational biology, criminology, and natural language processing, see for example Johnston and DiNardo (1997); McDonald (2009); Osgood (2000); Wang and Hua (2014); Lu, Foster and Ungar (2013). In particular, Yuan and Zhou (2015) and Yi, Wang, and Liu (2015) established the convergence rates for the minimax estimation risk under the high-dimensional additive regression model and the single index model, respectively. For data transformations, it is natural to consider the transformations that are continuous and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

one to one on an interval. Indeed, the functions satisfying these two conditions must be strictly monotonic, see Stein and Shakarchi (2009).

In the present paper, we consider adaptive estimation and statistical inference for high-dimensional sparse Gaussian copula regression. The model can be formulated as follows. Suppose we have an independent and identically distributed random sample $\boldsymbol{Z}_1 = (Y_1, \boldsymbol{X}_1)$, ..., $\boldsymbol{Z}_n = (Y_n, \boldsymbol{X}_n) \in \mathbb{R}^{p+1}$, where $Y_i \in \mathbb{R}$ are the responses and $\boldsymbol{X}_i \in \mathbb{R}^p$ are the covariates. Set $d = p+1$. We say $(Y_i, \boldsymbol{X}_i)$ satisfies a Gaussian copula regression model, if there exists a set of strictly increasing functions $\boldsymbol{f} = \{f_0, f_1, ..., f_p\}$ such that the marginally transformed random vectors $\widetilde{\boldsymbol{Z}}_i = (\widetilde{Y}_i, \widetilde{\boldsymbol{X}}_i) := (f_0(Y_i), f_1(X_{i1}), ..., f_p(X_{ip}))$ satisfy $\widetilde{\boldsymbol{Z}}_i \overset{i.i.d}{\sim} N_d(0, \Sigma)$ for some positive-definite covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ with $\mathrm{diag}(\Sigma) = \mathbf{1}$. The condition $\mathrm{diag}(\Sigma) = \mathbf{1}$ is for identifiability because the scaling and shifting are absorbed in the marginal transformations.

Under the Gaussian copula regression model, one has the following linear relationship for the transformed data:

$$\widetilde{Y}_i = \widetilde{\boldsymbol{X}}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, ..., n, \tag{1.2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ and ϵ_i are i.i.d zero-mean Gaussian variables. Writing in terms of the covariances, one has $\boldsymbol{\beta} = \Sigma_{\widetilde{X}\widetilde{X}}^{-1}\Sigma_{\widetilde{X}\widetilde{Y}}$ and $\epsilon_i \overset{i.i.d}{\sim} N(0, 1 - \Sigma_{\widetilde{Y}\widetilde{X}}\Sigma_{\widetilde{X}\widetilde{X}}^{-1}\Sigma_{\widetilde{X}\widetilde{Y}})$, where $\Sigma_{\widetilde{X}\widetilde{X}} = \mathrm{Cov}(\widetilde{\boldsymbol{X}}_1, \widetilde{\boldsymbol{X}}_1)$ and $\Sigma_{\widetilde{X}\widetilde{Y}} = \mathrm{Cov}(\widetilde{\boldsymbol{X}}_1, \widetilde{Y}_1)$. We focus on the high-dimensional setting where p is comparable to or much larger than n, and $\boldsymbol{\beta}$ is sparse. The fundamental difference between this model and the model at (1.1) is that one observes $\{Y_1, \boldsymbol{X}_1), ..., (Y_n, \boldsymbol{X}_n)\}$, not $\{(\widetilde{Y}_1, \widetilde{\boldsymbol{X}}_1), , , , , (\widetilde{Y}_n, \widetilde{\boldsymbol{X}}_n)\}$ as the transformations f_i are unknown.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

The Gaussian copula regression model has been widely used and well studied in the classical low-dimensional setting, see Sungur (2005); Crane and Hoek (2008); Masarotto and Varin (2012); Noh, Ghouch, and Bouezmarni (2013). For example, Masarotto and Varin (2012) developed a systematic framework to make inference and implement model validation for the Gaussian copula regression model. Noh, Ghouch, and Bouezmarni (2013) proposed a plug-in approach for estimating a regression function based on copulas, and presented the asymptotic normality of the estimator. However, their model and analysis are restricted to the low-dimensional setting and not well adapted to the high-dimensional case. In the high-dimensional setting, Wang and Hua (2014) applied the Gaussian copula regression model to predict financial risks, but the theoretical guarantees are still unclear.

The goal of the present paper is to develop adaptive estimation and inference methods that achieve the optimal performance in terms of the convergence rates without the knowledge of the marginal transformations. The rank-based Kendall's tau is used to extract the covariance information on the transformed data that does not require estimation of the transformations. Based on the covariance matrix estimator, an ℓ_1 regularized estimator is proposed to estimate $\boldsymbol{\beta}$ and a corresponding de-biased estimator is developed for the construction of the confidence intervals and hypothesis tests. In addition, prediction of the response for a given value of the covariates is also considered. One of the main technical challenges is that in the high-dimensional Gaussian copula model, the procedure in Javanmard and Montanari (2014) does not apply and a new method and technical analysis is needed. To achieve the same inferential results as the de-biased Lasso estimator for high-dimensional linear regression, the de-biasing procedure needs to be modified carefully.

Theoretical properties of the procedures for estimation, prediction, and statistical in-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

ference are studied. The proposed estimator is shown to be rate-optimal under regularity conditions. The proposed estimation and inference methods share similar properties as those optimal procedures for the high-dimensional linear regression. They are more flexible in the sense that they are adaptive to unknown monotone marginal transformations. For example, it is of practical interest to test whether a given covariate X_i is related to the response Y. The proposed testing procedure enables one to test this hypothesis without the need of knowing or estimating the marginal transformations. In addition, the procedures are easy to implement and perform well numerically. The methods are applied to analyze the Communities and Crime Unnormalized Data from the UCI Machine Learning Repository.

Compared with other methods such as those for the additive regression model and the single index model, a significant advantage for our proposed estimation and inference procedures is that they do not require estimation of the marginal transformations. For example, one can select the important variables x_i without any knowledge of the transformations f_i. This makes the methods more flexible and adaptive. The estimator achieves the same optimal rate as that for high-dimensional linear regression. We compare our methods and results to the existing literature on the Gaussian copula graphical model, such as Gu et al. (2015) where estimation and inference methods for individual entries of the precision matrix $\Omega = \Sigma^{-1}$ were proposed, based on the observed data $\{(X_{i1}, ..., X_{ip})\}_{i=1}^n$. The inferential result in Gu et al. (2015) requires $(f_1(X_{i1}), ..., f_p(X_{ip})) \sim N(0, \Sigma)$ and Ω to be sparse. Such a matrix sparsity condition is not needed here. In addition, we use a different method to construct the confidence interval. In the present paper, we use the de-biased estimator, while the confidence interval in Gu et al. (2015) was based on the Wald test.

The rest of the paper is organized as follows. After basic notation and definitions are

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

introduced, Section 2 presents the ℓ_1 penalized minimization procedure for estimating $\boldsymbol{\beta}$ that uses a rank-based correlation matrix estimator. Prediction is also considered. Section 3 constructs a de-biased estimator and establishes an asymptotic normality result. Confidence intervals and hypothesis tests are developed based on the limiting distribution. Numerical performance of the proposed estimator and inference procedures are investigated in Section 4. A brief discussion is given in Section 5, and the main results are proved in Section 6. The proofs of some technical lemmas and some additional simulation results are given in the Supplement Cai and Zhang (2016).

2. Adaptive Estimation and Prediction

We consider adaptive estimation and prediction in this section. We first introduce the rank-based correlation matrix estimator to extract covariance information on the transformed data that does not require estimation of the marginal transformations, and then present the estimation and prediction procedures and their theoretical properties.

We begin with the basic notation and definitions. Throughout the paper, we use bold-faced letters for vectors. For a vector $\boldsymbol{u} \in \mathbb{R}^p$ and $1 \leq q \leq \infty$, the ℓ_q norm is defined as $||\boldsymbol{u}||_q = (\sum_{i=1}^p |u_i|^q)^{1/q}$. In particular, $||\boldsymbol{u}||_\infty = \max_i |u_i|$. In addition, $\boldsymbol{u}[i:j]$ denotes the entries of $\boldsymbol{u}$ from i-th to j-th coordinates and $\text{supp}(\boldsymbol{u})$ is the support of $\boldsymbol{u}$. For a matrix $A \in \mathbb{R}^{p\times p}$ and $1 \leq q \leq \infty$, the matrix ℓ_q operator norm is is defined as $||A||_q = \sup_{||\boldsymbol{u}||_q=1} ||A\boldsymbol{u}||_q$. The spectral norm of A is the ℓ_2 operator norm and the ℓ_1 norm is the maximum absolute column sum. For an integer $1 \leq s \leq p$, the s-restricted spectral norm of A is defined as $||A||_{2,s} = \sup_{\boldsymbol{u}\in S^{p-1}, |\boldsymbol{u}|_0 \leq s} ||A\boldsymbol{u}||_2$, where S^{p-1} is the unit ball in $\mathbb{R}^p$. The vector ℓ_∞ norm on matrix A is $|A|_\infty = \max_{i,j} |A_{ij}|$. For a symmetric matrix A, we

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote, respectively, the largest and smallest eigenvalue of A, and $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$ is the condition number. Further, we denote the restricted condition number by $\kappa_s(\Sigma) := \sup\{\lambda_{\max}(\Sigma_{S,S})/\lambda_{\min}(\Sigma_{S,S}) : S \in [n], |S| = s\}$. We write $A \succeq 0$ if A is semidefinite positive. In addition, $\circ$ denotes the matrix element-wise multiplication, and $\otimes$ is the Kronecker product. Moreover, $\mathrm{vec}(\cdot)$ maps an $m \times n$ matrix A to a $\mathbb{R}^{mn}$ vector by laying out the columns of A one by one. For a set of indices I,J, we let $A_{I,J}$ denote the submatrix formed by the rows in I and columns in J. $I_{p\times p}$ is the p by p identity matrix. $e_i^{(n)}$ is the i-th unit vector in $\mathbb{R}^n$ with entries $e_{ij}^{(n)} = I_{\{j=i\}}$, for $j = 1, ..., n$. $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal distribution. $B_r(x)$ denotes the Eucilidean ball centered at x with radius r. For two sequences of nonnegative real numbers, $a_n \lesssim b_n$ implies that there exists a constant C not depending on n, such that $a_n \le Cb_n$. Finally, we use $[d]$ to denote the set $\{1, 2, ..., d\}$.

2.1 Rank-Based Estimator of Correlation Matrix

At (1.2), we use $(\boldsymbol{Y}, X)$ to denote the observed data, with $\boldsymbol{Y} \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ the design matrix with rows $\boldsymbol{X}_1^\top, ..., \boldsymbol{X}_n^\top$, and $(\widetilde{\boldsymbol{Y}}, \widetilde{X})$ to be the original data that possesses the linear relationship. In addition, $\boldsymbol{Z}_i^\top := (Y_i, \boldsymbol{X}_i^\top)$ and $\widetilde{\boldsymbol{Z}}_i^\top := (\widetilde{Y}_i, \widetilde{\boldsymbol{X}}_i^\top)$. An essential quantity in estimation of $\boldsymbol{\beta}$ and inference for the Gaussian copula regression model (1.2) is the covariance matrix (or correlation matrix as the diagonal is 1) Σ. Since the marginal transformations f_i's are unknown and thus $(\widetilde{Y}, \widetilde{\boldsymbol{X}})$ is not directly accessible, the conventional sample covariance matrix is not available as an estimate of Σ. We thus need an alternative method to estimate the covariance/correlation matrix Σ.

Our approach is to use the rank-based Kendall's tau, which can be well estimated from

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

the observed data $(Y_1, \boldsymbol{X}_1^\top)$, ..., $(Y_n, \boldsymbol{X}_n^\top)$. This estimator is based on the following fact (see Section 3 of Kruskal (1958)).

Set $d = p+1$. If $\widetilde{\boldsymbol{Z}}_i \overset{i.i.d.}{\sim} N_d(0, \Sigma)$ with $\Sigma = (\sigma_{jk})_{1\le j,k\le d}$, then

$$\sigma_{jk} = \sin\left(\frac{\pi}{2}\tau_{jk}\right), \tag{2.1}$$

where τ_{jk} is Kendall's tau defined as

$$\tau_{jk} = \mathbb{E}[\mathrm{sgn}(\widetilde{z}_{1j} - \widetilde{z}_{2j})\mathrm{sgn}(\widetilde{z}_{1k} - \widetilde{z}_{2k})], \tag{2.2}$$

with $\widetilde{\boldsymbol{Z}}_i = (\widetilde{z}_{i1}, \widetilde{z}_{i2}, ..., \widetilde{z}_{id})^\top$, $i = 1, 2$, being two independent copies of $N_d(0, \Sigma)$.

Note that τ_{jk} given in (2.2) is invariant under strictly increasing marginal transformations. This leads to an estimate of τ_{ij} based on the observed data $\boldsymbol{Z}_1$, ..., $\boldsymbol{Z}_n$ under the Gaussian copula regression model,

$$\begin{aligned}\widehat{\tau}_{jk} &= \frac{2}{n(n-1)} \sum_{1\le i_1<i_2\le n} \mathrm{sgn}(\widetilde{Z}_{i_1j} - \widetilde{Z}_{i_2j})\mathrm{sgn}(\widetilde{Z}_{i_1k} - \widetilde{Z}_{i_2k}) \\ &= \frac{2}{n(n-1)} \sum_{1\le i_1<i_2\le n} \mathrm{sgn}(Z_{i_1j} - Z_{i_2j})\mathrm{sgn}(Z_{i_1k} - Z_{i_2k}), \quad 1 \le j,k \le d.\end{aligned} \tag{2.3}$$

Denote by $\widehat{T} = (\widehat{\tau}_{jk})_{d\times d}$ the Kendall's tau sample correlation matrix, and its population version $T = (\tau_{jk})_{d\times d}$. If $\boldsymbol{S}_{i,i'} = (\mathrm{sgn}(Z_{i1} - Z_{i'1}), ..., \mathrm{sgn}(Z_{id} - Z_{i'd}))^\top$, then

$$\widehat{T} = (\widehat{\tau}_{jk})_{d\times d} = \frac{1}{n(n-1)} \sum_{i\ne i'}^{n} \boldsymbol{S}_{i,i'}\boldsymbol{S}_{i,i'}^\top. \tag{2.4}$$

Based on Kendall's tau, (2.1) leads to an estimator for the correlation matrix Σ,

$$\widehat{\Sigma} = (\widehat{\sigma}_{jk})_{d\times d} \quad \text{with} \quad \widehat{\sigma}_{jk} = \sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right). \tag{2.5}$$

We divide Σ into four sub-matrices, denoted by $\Sigma_{XX}, \Sigma_{XY}, \Sigma_{YX}, \Sigma_{YY}$, and their corresponding Kendall's tau based estimators are $\widehat{\Sigma}_{YY}, \widehat{\Sigma}_{YX}, \widehat{\Sigma}_{XY}, \widehat{\Sigma}_{XX}$, with $\widehat{\Sigma}_{YX} = \widehat{\Sigma}_{XY}^\top$ and $\Sigma_{YX} = \Sigma_{XY}^\top$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

2.2 Estimation of $\boldsymbol{\beta}$

We introduce the procedure for estimating the sparse coefficient vector $\boldsymbol{\beta}$ in (1.2). If the marginal transformations f_i, $i = 0, 1, ..., p$ are given, then $(\widetilde{Y}_i, \widetilde{\boldsymbol{X}}_i^\top)$ are available and in this case a natural approach to estimating $\boldsymbol{\beta}$ is to use the Lasso estimator

$$\widehat{\boldsymbol{\beta}}_{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\frac{1}{2n}||\widetilde{\boldsymbol{Y}} - \widetilde{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1\}.$$

Rewriting the objective function yields

$$\widehat{\boldsymbol{\beta}}_{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\frac{1}{2n}(\boldsymbol{\beta}^\top\widetilde{X}^\top\widetilde{X}\boldsymbol{\beta} - 2\widetilde{\boldsymbol{Y}}^\top\widetilde{X}) + \lambda||\boldsymbol{\beta}||_1\}. \tag{2.6}$$

Since $(\widetilde{Y}_i, \widetilde{\boldsymbol{X}}_i)$ are not directly accessible as the transformations f_i's are unknown, the estimator given in (2.6) cannot be used. The quantities $\widetilde{X}^\top\widetilde{X}/n$ and $\widetilde{\boldsymbol{Y}}^\top\widetilde{X}/n$ in (2.6) can be viewed as estimators of the covariances Σ_{XX} and Σ_{YX}, respectively. From this perspective, it is natural to replace $\widetilde{X}^\top\widetilde{X}/n$ and $\widetilde{\boldsymbol{Y}}^\top\widetilde{X}/n$ in (2.6) with the alternative covariance estimators $\widehat{\Sigma}_{XX}$ and $\widehat{\Sigma}_{YX}$ based on Kendall's τ, as discussed in Section 2.1. We propose an ℓ_1 penalized minimization procedure for estimating $\boldsymbol{\beta}$.

Algorithm 1 Adaptive estimator of $\boldsymbol{\beta}$

Input: Observed pairs $(Y_1, \boldsymbol{X}_1^\top), ..., (Y_n, \boldsymbol{X}_n^\top)$, parameter $\lambda > 0$.

Output: Regularized estimator $\widehat{\boldsymbol{\beta}}(\lambda)$.

1: Construct Kendall's tau based covariance estimators $\widehat{\Sigma}_{XX}$ and $\widehat{\Sigma}_{XY}$.

2: Set

$$\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\frac{1}{2}(\boldsymbol{\beta}^\top\widehat{\Sigma}_{XX}\boldsymbol{\beta} - 2\widehat{\Sigma}_{YX}\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1\}. \tag{2.7}$$

Remark 1. As $\hat{\Sigma}_{XX}$ may not be positive semidefinite (PSD), the optimization (2.7) may not

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

be convex. Theorem 1 in Loh and Wainwright (2013) developed theory for this nonconvex optimization problem, and showed that the solution obtained by the standard projected gradient descent method lies within statistical error of the true $\boldsymbol{\beta}$. Alternatively, one can project $\hat{\Sigma}_{XX}$ onto the cone of the PSD matrices, $\hat{\Sigma}^{+}_{XX} = \arg\min_{\Sigma \succeq 0} \|\hat{\Sigma}_{XX} - \Sigma\|_{2,s}$. Here we use the $\|\cdot\|_{2,s}$ norm instead of the spectral norm due to theoretical considerations for the results given in Theorem 1. This projection increases the loss by a factor at most two, so in practice $\hat{\Sigma}^{+}_{XX}$ can be used in place of $\hat{\Sigma}_{XX}$.

To consider the properties of the estimator $\widehat{\boldsymbol{\beta}}(\lambda)$ given in Algorithm 1, first define the Restricted Strong Convexity (RSC) condition introduced in Negahban et al. (2009).

Definition 1 (RSC)**.** For a given sparsity level $s \le p$ and constant $\alpha \ge 1$, let $C(s, \alpha) := \{\boldsymbol{\theta} \in \mathbb{R}^p : ||\boldsymbol{\theta}_{S^c}||_1 \le \alpha ||\boldsymbol{\theta}_S||_1, S \subset \{1, ..., p\}, |S| \le s\}$. A matrix $\Sigma \in \mathbb{R}^{p \times p}$ satisfies the restricted strong convexity (RSC) condition, with constants (γ_1, s, α), if

$$\boldsymbol{\theta}^\top \Sigma \boldsymbol{\theta} \ge \gamma_1 ||\boldsymbol{\theta}||_2^2 \quad \text{for all } \boldsymbol{\theta} \in C(s; \alpha).$$

The RSC condition is related to the restricted eigenvalue condition that Bickel, Ritov, and Tsybakov (2009) used in the analysis of high-dimensional linear regression. See Negahban et al. (2009) for more detailed discussion on the RSC.

Theorem 1. *Assume that $\boldsymbol{\beta}$ is s-sparse. Suppose that $\kappa_s(\Sigma) \le M$ for some $M > 0$, and that Σ_{XX} satisfies the RSC with constants $(\gamma_1, s, 3)$. Write $\widehat{\boldsymbol{\beta}}(\lambda)$ defined at (2.7), if $s = o(\frac{n}{\log p})$, and the tuning parameter $\lambda = C_1 \sqrt{\frac{\log p}{n}}$ is chosen with $C_1 > 2M$, then with probability at least $1 - 2p^{-1}$,*

$$||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}||_2 \lesssim \sqrt{\frac{s \log p}{n}} \quad \textit{and} \quad ||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}||_1 \lesssim s \sqrt{\frac{\log p}{n}}. \tag{2.8}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

Furthermore, if $|\Sigma_{X_S, X_{S^c}}|_\infty \le 1-\alpha$ *for some constant* $\alpha > 0$*, where* $S = \mathrm{supp}(\boldsymbol{\beta})$ *and* X_S *is its corresponding index set in* Σ, $\min_{i\in S}|\beta_i| \ge \frac{8M}{\gamma_1}(1+\frac{4(2-\alpha)}{\alpha})\sqrt{\frac{s\log p}{n}}$*, then for* $\lambda = \frac{8M(2-\alpha)}{\alpha}\sqrt{\frac{s\log p}{n}}$*, with probability at least* $1-2p^{-1}$*,*

$$\mathrm{sgn}(\boldsymbol{\beta}) = \mathrm{sgn}(\widehat{\boldsymbol{\beta}}(\lambda)). \tag{2.9}$$

The convergence rates of $\widehat{\boldsymbol{\beta}}(\lambda)$ under the ℓ_1 and ℓ_2 norm losses given in (2.8) match the minimax lower bounds for high-dimensional linear regression, Raskutti, Wainwright and Yu (2011). This implies that $\widehat{\boldsymbol{\beta}}(\lambda)$ is minimax rate optimal under the Gaussian copula regression model and achieves the same optimal rate attained by the regular Lasso for linear regression. Thus the proposed procedure is adaptive to the unknown marginal transformations and gains this added flexibility for free in terms of convergence rate. The result given in (2.9) shows that, under regularity conditions, $\widehat{\boldsymbol{\beta}}(\lambda)$ is sign consistent.

2.3 Prediction

In addition to estimation of $\boldsymbol{\beta}$, a problem of signficant practical interest is that of predicting the response Y^* for a given value of the covariates $\boldsymbol{x}^* = (x_1^*, ..., x_p^*)$ based on the Gaussian copula regression model (1.2). In the oracle setting where the transformations $f_0, ..., f_p$ and the coefficient vector $\boldsymbol{\beta}$ are known, the optimal prediction of the response is

$$\mu^* = f_0^{-1}(\sum_{i=1}^{p} f_i(x_i^*)\beta_i).$$

Our goal is to construct a predictor $\widehat{\mu}^*$, based only on the observed data $(Y_1, \boldsymbol{X}_1), ..., (Y_n, \boldsymbol{X}_n)$, that is close to the oracle predictor μ^*.

Let F_0 be the cumulative distribution function of Y and let F_i be the cumulative distribution function of X_i for $i = 1, ..., p$. As for the sample version, let $\widehat{F}_0$ be the empirical cumulative distribution function of $\{Y_1, ..., Y_n\}$ and let $\widehat{F}_i$ be the empirical cumulative distribution function of $\{X_{i1}, ..., X_{in}\}$ for $i = 1, ..., p$. Set

$$\widehat{f}_0(t) = \Phi^{-1}(\tilde{F}_0(t)), i = 1, 2, ..., n; \tag{2.10}$$

$$\widehat{f}_i(t) = \Phi^{-1}(\widehat{F}_i(t)), i = 1, 2, ..., n, \tag{2.11}$$

where $\tilde{F}_0(t) = \frac{1}{n^2} I(\widehat{F}_0(t) < 1/n^2) + \widehat{F}_0(t) I(\widehat{F}_0(t) \in [1/n^2, 1 - 1/n^2]) + \frac{n^2-1}{n^2} I(\widehat{F}_0(t) > 1 - 1/n^2)$.

For a given value of the covariates $x^* = (x_1^*, ..., x_p^*)$, we define the predictor

$$\widehat{\mu}^* = \widehat{f}_0^{-1}(\sum_{i=1}^{p} \widehat{f}_i(x_i^*) \widehat{\beta}(\lambda)_i), \tag{2.12}$$

where $\widehat{\boldsymbol{\beta}}(\lambda)$ is the estimator given in (2.7) and $\widehat{f}_0^{-1}$ is the generalized inverse of $\widehat{f}_0$:

$$\widehat{f}_0^{-1}(t) = \inf\{x \in \mathbb{R} : \widehat{f}_0(x) \geq t\}.$$

Write $B_r(x)$ as the Eucilidean ball centered at x with radius r, we have the following for the predictor $\widehat{\mu}^*$.

Theorem 2. *Suppose for some constant $c > 0$, $|f_0(v_1) - f_0(v_2)| \geq c|v_1 - v_2|$ for all $v_1, v_2 \in f_0^{-1}(B_r(f_0(\mu^*)))$ with $r \geq Cs\sqrt{\log d/n}$ for a sufficiently large constant C, $f_0(\mu^*) < M$, and $\max_{i=1,...,p} F_i(x_i^*) \in (\delta^*, 1 - \delta^*)$ for some constant $M > 0, \delta^* \in (0, 1)$. If $s = o(\sqrt{\frac{n}{\log p}})$, then under the conditions of Theorem 1 the predictor $\widehat{\mu}^*$ given in (2.12) satisfies, with probability at least $1 - p^{-1} - n^{-1}$,*

$$|\widehat{\mu}^* - \mu^*| \lesssim s\sqrt{\frac{\log p}{n}}.$$

This error bound is tight. $f_0^{-1}(\mu^*) = \sum_{i=1}^{p} f_i(x_i^*)\beta_i$ can be viewed as a linear functional of $\boldsymbol{\beta}$ with unknown weights $f_i(x_i^*)$ (as the marginal transformations f_i's are unknown). For high-dimensional linear regression, inference on the linear functionals of $\boldsymbol{\beta}$ with known weights has

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

been considered in Cai and Guo (2015), where a lower bound of order $s\sqrt{\frac{\log p}{n}}$ was established for estimation error and for the expected length of confidence intervals for linear functionals with "dense" weight vectors.

3. Statistical Inference

We turn to statistical inference for the Gaussian copula regression model. The Lasso estimator is inherently biased as it is essential to trade variance and bias in order to achieve the optimal estimation performance. For statistical inference such as confidence intervals and hypothesis tests, it is desirable to use (nearly) unbiased pivotal estimators. Such an approach has been used in the construction of confidence intervals for high-dimensional linear regression in the recent literature. See, for example, Javanmard and Montanari (2014); Van de Geer et al. (2014); Zhang and Zhang (2014); Cai and Guo (2015). We follow the same principle to de-bias the estimator $\widehat{\boldsymbol{\beta}}(\lambda)$ given in Algorithm 1.

We begin by noting that $\widehat{\boldsymbol{\beta}}(\lambda)$ satisfies the Karush-Kuhn-Tucker (KKT) condition

$$\widehat{\Sigma}_{XX}\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\Sigma}_{XY} + \lambda\partial||\widehat{\boldsymbol{\beta}}(\lambda)||_1 = 0, \tag{3.1}$$

where $\partial||\widehat{\boldsymbol{\beta}}(\lambda)||_1$ is the subgradient of the ℓ_1 norm $||\cdot||_1$. Equation (3.1) can be rewritten as

$$\widehat{\Sigma}_{XX}(\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}) + \lambda\partial||\widehat{\boldsymbol{\beta}}(\lambda)||_1 = \widehat{\Sigma}_{XY} - \widehat{\Sigma}_{XX}\boldsymbol{\beta}.$$

Suppose one has a good approximation of the "inverse" of $\widehat{\Sigma}_{XX}$, say M, and multiplies M on the left:

$$M\widehat{\Sigma}_{XX}(\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}) + \lambda M\partial||\widehat{\boldsymbol{\beta}}(\lambda)||_1 = M(\widehat{\Sigma}_{XY} - \widehat{\Sigma}_{XX}\boldsymbol{\beta}).$$

Then it follows that

$$(\widehat{\boldsymbol{\beta}}(\lambda) + \lambda M\partial||\widehat{\boldsymbol{\beta}}(\lambda)||_1) - \boldsymbol{\beta} = M(\widehat{\Sigma}_{XY} - \widehat{\Sigma}_{XX}\boldsymbol{\beta}) + (I - M\widehat{\Sigma}_{XX})(\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}). \tag{3.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

Let $\widehat{\boldsymbol{\beta}}^u = \widehat{\boldsymbol{\beta}}(\lambda) + \lambda M \partial ||\widehat{\boldsymbol{\beta}}(\lambda)||_1$. By inspection, this leads to

$$\begin{aligned} \sqrt{n}(\widehat{\boldsymbol{\beta}}^u - \boldsymbol{\beta}(\lambda)) =& \sqrt{n}(M\widehat{\Sigma}_{XY} - M\widehat{\Sigma}_{XX}\boldsymbol{\beta}) + \sqrt{n}(I - M\widehat{\Sigma}_{XX})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\lambda)) && (3.3) \\ =& \sqrt{n}(M\widehat{\Sigma}_{XY} - M\widehat{\Sigma}_{XX}\boldsymbol{\beta}) + o(1), && (3.4) \end{aligned}$$

where the second equality use the assumption that M approximates the "inverse" of $\widehat{\Sigma}_{XX}$ well and thus that $(I - M\widehat{\Sigma}_{XX})(\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta})$ is negligible. Then $\sqrt{n}(M\widehat{\Sigma}_{XY} - M\widehat{\Sigma}_{XX}\boldsymbol{\beta})$ plays a major role in the limiting distribution of $\sqrt{n}\widehat{\boldsymbol{\beta}}^u$ and later we will show its asymptotic normality (Theorem 3).

This analysis suggests the de-biasing procedure

$$\widehat{\boldsymbol{\beta}}^u = \widehat{\boldsymbol{\beta}}(\lambda) + \lambda M \partial ||\widehat{\boldsymbol{\beta}}(\lambda)||_1 = \widehat{\boldsymbol{\beta}}(\lambda) + M(\widehat{\Sigma}_{XY} - \widehat{\Sigma}_{XX}\widehat{\boldsymbol{\beta}}(\lambda)),$$

where the second equality is from (3.1).

We then need to construct the matrix M that is a good approximation of the "inverse" of $\widehat{\Sigma}_{XX}$. We proceed with two objectives in mind: to control $|M\widehat{\Sigma}_{XX} - I_{p\times p}|_\infty$ and to control the variance of $\widehat{\beta}_i^u$. The latter is for the precision of the statistical inference procedures. For example, the length of the confidence intervals for β_i is proportional to the standard deviation of $\widehat{\beta}_i^u$.

In the following, we are going to estimate the variance of $\widehat{\beta}_i^u$, and solve for M that minimizes this variance. Assuming that $(I - M\widehat{\Sigma}_{XX})(\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta})$ is negligible, by (3.3), the variance of $\widehat{\beta}_i^u$ is determined by that of $\boldsymbol{m}_i^\top(\widehat{\Sigma}_{XY} - \widehat{\Sigma}_{XX}\boldsymbol{\beta})$, where $\boldsymbol{m}_i$ is the i-th column of M. If $\boldsymbol{u}_i = (0, \boldsymbol{m}_i^\top)^\top$ and $\boldsymbol{v}_0 = (1, -\boldsymbol{\beta}^\top)^\top \in \mathbb{R}^d$, then one has

$$\boldsymbol{m}_i^\top(\widehat{\Sigma}_{XY} - \widehat{\Sigma}_{XX}\boldsymbol{\beta}) = \boldsymbol{u}_i\widehat{\Sigma}\boldsymbol{v}_0^\top.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

It is shown in Lemma 1 in Section 6 that the asymptotic variance of $\sqrt{n}\boldsymbol{u}_i\widehat{\Sigma}\boldsymbol{v}_0^\top$ is

$$\pi^2\sigma^2_{g1(\boldsymbol{u}_i)} := \pi^2\mathrm{Var}(g_1(\boldsymbol{Z};\boldsymbol{u}_i)), \tag{3.5}$$

where $g_1(\boldsymbol{Z};\boldsymbol{u}_i) = \mathbb{E}[g(\boldsymbol{Z},\boldsymbol{Z}';\boldsymbol{u}_i)|\boldsymbol{Z}]$, and $g(\boldsymbol{Z},\boldsymbol{Z}';\boldsymbol{u}_i)$ is defined as

$$g(\boldsymbol{Z},\boldsymbol{Z}';\boldsymbol{u}_i) = \mathrm{sgn}(\boldsymbol{Z}-\boldsymbol{Z}')^\top(\boldsymbol{u}_i\boldsymbol{v}_0^\top \circ \cos(\frac{\pi}{2}T))\mathrm{sgn}(\boldsymbol{Z}-\boldsymbol{Z}')$$

for $\boldsymbol{Z},\boldsymbol{Z}' \overset{i.i.d}{\sim} N_d(0,\Sigma)$ and $\boldsymbol{u}_i \in \mathbb{R}^d$.

Therefore, to estimate the variance of $\widehat{\beta}^u_i$, we need a good estimate of $\sigma^2_{g1(\boldsymbol{u}_i)}$. Note that (3.5) can be further expressed as

$$\sigma^2_{g1(\boldsymbol{u}_i)}=\mathrm{Var}(g_1(\boldsymbol{Z};\boldsymbol{u}_i)) = \mathrm{vec}(\boldsymbol{u}_i\boldsymbol{v}_0^\top \circ \cos(\frac{\pi}{2}T))^\top \cdot \Sigma_{h_Z} \cdot \mathrm{vec}(\boldsymbol{u}_i\boldsymbol{v}_0^\top \circ \cos(\frac{\pi}{2}T)), \tag{3.6}$$

where $\Sigma_{h_Z} = \mathrm{Var}(h_Z(\boldsymbol{Z})) \in \mathbb{R}^{d^2\times d^2}$ is the covariance matrix of $h_Z(\boldsymbol{Z}) = \mathbb{E}[\mathrm{sgn}(\boldsymbol{Z}-\boldsymbol{Z}') \otimes \mathrm{sgn}(\boldsymbol{Z}-\boldsymbol{Z}')|\boldsymbol{Z}] \in \mathbb{R}^{d^2}$. Then we estimate Σ_{h_Z} by

$$\widehat{\Sigma}_{h_Z} = \frac{1}{n}\sum_{i=1}^n(\widehat{h}_Z(\boldsymbol{Z}_i) - \frac{1}{n}\sum_{i'=1}^n\widehat{h}_Z(\boldsymbol{Z}_{i'}))(\widehat{h}_Z(\boldsymbol{Z}_i) - \frac{1}{n}\sum_{i'=1}^n\widehat{h}_Z(\boldsymbol{Z}_{i'}))^\top, \tag{3.7}$$

where $\widehat{h}_Z(\boldsymbol{Z}_i) = \frac{1}{n-1}\sum_{i'\neq i}^n \mathrm{sgn}(\boldsymbol{Z}_i - \boldsymbol{Z}_{i'}) \otimes \mathrm{sgn}(\boldsymbol{Z}_i - \boldsymbol{Z}_{i'})$.

Consequently, a good estimate of $\sigma^2_{g1(\boldsymbol{u}_i)}$ is given by

$$\widehat{\sigma}^2_{g1(\boldsymbol{u}_i)} = \mathrm{vec}(\boldsymbol{u}_i\widehat{\boldsymbol{v}}^\top \circ \cos(\frac{\pi}{2}\widehat{T}))^\top\widehat{\Sigma}_{h_Z}\mathrm{vec}(\boldsymbol{u}_i\widehat{\boldsymbol{v}}^\top \circ \cos(\frac{\pi}{2}\widehat{T})), \tag{3.8}$$

with $\widehat{\boldsymbol{v}} = (1, \widehat{\boldsymbol{\beta}}(\lambda)^\top)^\top$, and this determines the estimate of the variance of $\widehat{\beta}^u_i$.

We are ready to present the de-biasing procedure, which controls $|M\widehat{\Sigma}_{XX} - I_{p\times p}|_\infty$ and minimizes the variance of $\widehat{\beta}^u_i$, where the latter is equivalent to minimizing $\widehat{\sigma}^2_{g1(\boldsymbol{u}_i)}$. To simplify the notation, we define $x(\boldsymbol{u}) : \mathbb{R}^d \to \mathbb{R}^{d^2}$ with $x(\boldsymbol{u}) = \mathrm{vec}(\boldsymbol{u}\boldsymbol{v}_0^\top \circ \cos(\frac{\pi}{2}T))$, and $\widehat{x}(\boldsymbol{u}) : \mathbb{R}^d \to \mathbb{R}^{d^2}$ with $\widehat{x}(\boldsymbol{u}) = \mathrm{vec}(\boldsymbol{u}\widehat{\boldsymbol{v}}^\top \circ \cos(\frac{\pi}{2}\widehat{T}))$. Then (3.5) and (3.8) can be simplified to

$$\sigma^2_{g_1(\boldsymbol{u})} = x(\boldsymbol{u})^\top\Sigma_{h_Z}x(\boldsymbol{u}) \quad \text{and} \quad \widehat{\sigma}^2_{g_1(\boldsymbol{u})} = \widehat{x}(\boldsymbol{u})^\top\widehat{\Sigma}_{h_Z}\widehat{x}(\boldsymbol{u}). \tag{3.9}$$

Let $K = \cos(\frac{\pi}{2}\widehat{T}) = (\boldsymbol{K}_1, ..., \boldsymbol{K}_d)$ and $\check{\boldsymbol{u}} = (\boldsymbol{u}^\top, ..., \boldsymbol{u}^\top)^\top \in \mathbb{R}^{d^2}$. Take $L = (I_{d\times d}, I_{d\times d}, ..., I_{d\times d}) \in \mathbb{R}^{d\times d^2}$ and rewrite $\check{\boldsymbol{u}} = L^\top \boldsymbol{u}$. Define $\check{D} = \mathrm{diag}(v_1 \,\mathrm{diag}(\boldsymbol{K}_1), ..., v_d \,\mathrm{diag}(\boldsymbol{K}_d))$ and set

$$M_\Sigma = L\check{D}\widehat{\Sigma}_{h_Z}\check{D}L^\top. \tag{3.10}$$

Then $\widehat{\sigma}^2_{g_1(\boldsymbol{u})}$ can be rewritten as a convex function of $\boldsymbol{u}$

$$\widehat{\sigma}^2_{g_1(\boldsymbol{u})} = \widehat{x}(\boldsymbol{u})^\top\widehat{\Sigma}_{h_Z}\widehat{x}(\boldsymbol{u}) = \check{\boldsymbol{u}}^\top\check{D}\widehat{\Sigma}_{h_Z}\check{D}\check{\boldsymbol{u}} = \boldsymbol{u}^\top L\check{D}\widehat{\Sigma}_{h_Z}\check{D}L^\top\boldsymbol{u} = \boldsymbol{u}^\top M_\Sigma \boldsymbol{u}. \tag{3.11}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

Algorithm 2 De-biased estimator of $\boldsymbol{\beta}$

Input: Observed pairs $(Y_1, \boldsymbol{X}_1^\top), ..., (Y_n, \boldsymbol{X}_n^\top)$, parameters $a \in (0, \frac{1}{12}), b > 0, \mu > 0, \lambda > 0$.

Output: De-biased estimator $\widehat{\boldsymbol{\beta}}^u$.

1: Construct Kendall's tau based covariance estimators $\widehat{\Sigma}_{XY}$ and $\widehat{\Sigma}_{XX}$, and calculate M_Σ by (3.10).

2: Let

$$\widehat{\boldsymbol{\beta}}(\lambda) = \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \{\frac{1}{2}(\boldsymbol{\beta}^\top \widehat{\Sigma}_{XX} \boldsymbol{\beta} - 2\widehat{\Sigma}_{YX}\boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}||_1\}. \tag{3.12}$$

3: **for** $i = 1, 2, \ldots, p$ **do**

4: Let $\boldsymbol{u}_i$ be a solution of

$$\begin{aligned} \underset{\boldsymbol{u} \in \mathbb{R}^p}{\text{minimize}} \quad & \boldsymbol{u}^\top M_\Sigma \boldsymbol{u} \\ \text{subject to} \quad & ||\widehat{\Sigma}_{XX} \boldsymbol{u}[2:d] - e_i^{(p)}||_\infty \leq \mu \\ & e_1^{(d)\top} \boldsymbol{u} = 0 \\ & b^{-1} n^{-a} \leq ||\boldsymbol{u}||_2 \leq ||\boldsymbol{u}||_1 \leq b n^{a/2} \end{aligned} \tag{3.13}$$

5: Set $M = (\boldsymbol{u}_1[2:d], ..., \boldsymbol{u}_p[2:d])$. If any of the above problems is not feasible, then set $M = I_{p \times p}$.

6: Define $\widehat{\boldsymbol{\beta}}^u$ as

$$\widehat{\boldsymbol{\beta}}^u = \widehat{\boldsymbol{\beta}}(\lambda) + M(\widehat{\Sigma}_{XY} - \widehat{\Sigma}_{XX}\widehat{\boldsymbol{\beta}}(\lambda)). \tag{3.14}$$

Note that (3.13) is a convex program and can be solved efficiently. That $\widehat{\sigma}^2_{g_1(\boldsymbol{u})}$ is convex with respect to $\boldsymbol{u}$ and the constraints of (3.13) construct a convex set of $\boldsymbol{u}$, implies that (3.13) is a convex program. The first constraint in (3.13) is to make sure that M is a good approximation of $\widehat{\Sigma}_{XX}^{-1}$, and the third constraint is for the convenience of theoretical

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

analysis, in practice b can be chosen sufficiently large so that it does not affect the numerical performance of the algorithm.

The following theorem serves as the basis for the construction of statistical inference procedures.

Theorem 3. *Suppose for some constants $M_i > 0$, $i = 1, 2, 3$, that $\frac{1}{M_1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M_1$, $||\Sigma^{-1}||_1 < M_2$, and $\lambda_{\min}(\Sigma_{hZ}) > M_3$. Suppose $s = o(\frac{\sqrt{n}}{\log p})$, $\mu = a\sqrt{\frac{\log p}{n}}$, and $\lambda = c\sqrt{\frac{\log p}{n}}$ in Algorithm 2 are chosen with $a > 4M_2$ and $c > 2M_1^2$. Then for any fixed $1 \leq i \leq p$ and for all $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}, ||\boldsymbol{\beta}||_0 \leq s} \left| P\left(\frac{\sqrt{n}(\widehat{\boldsymbol{\beta}}_i^u - \boldsymbol{\beta}_i)}{\pi\sqrt{\widehat{x}(\boldsymbol{u}_i)^\top \widehat{\Sigma}_{hZ} \widehat{x}(\boldsymbol{u}_i)}} \leq x \right) - \Phi(x) \right| = 0. \tag{3.15}$$

Theorem 3 shows that the estimator $\widehat{\boldsymbol{\beta}}^u$ possesses the distributional property similar to that of the de-biased Lasso estimator in Javanmard and Montanari (2014), although the observed data here have a linear relationship only after unknown transformations.

The asymptotic normality result given in (3.15) can be used to construct confidence intervals and hypothesis tests for any given coordinate β_i. Let $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$.

Corollary 1. *Suppose the conditions of Theorem 3 hold. Then for any given $1 \leq i \leq p$,*

$$CI_i = \left[\beta_i^u - z_{\alpha/2}\pi\sqrt{\frac{\widehat{x}(\boldsymbol{u}_i)^\top \widehat{\Sigma}_{hZ} \widehat{x}(\boldsymbol{u}_i)}{n}}, \quad \beta_i^u + z_{\alpha/2}\pi\sqrt{\frac{\widehat{x}(\boldsymbol{u}_i)^\top \widehat{\Sigma}_{hZ} \widehat{x}(\boldsymbol{u}_i)}{n}} \right] \tag{3.16}$$

is an asymptotically $(1 - \alpha)$ level confidence interval for β_i.

It is of practical interest to test whether a given covariate X_i is related to the response Y. In the context of the Gaussian copula regression model, this can be formulated as testing an individual null hypothesis $H_{0,i} : \beta_i = 0$ versus the alternative $H_{1,i} : \beta_i \neq 0$. To test $H_{0,i}$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

against $H_{1,i}$ at the nominal level α for some $0 < \alpha < 1$, based on Theorem 3, we introduce the test

$$\widehat{\Psi}_i = I\left(\frac{\sqrt{n}|\widehat{\beta}_i^u|}{\pi\sqrt{\widehat{x}(\boldsymbol{u}_i)^\top \widehat{\Sigma}_{hz}\widehat{x}(\boldsymbol{u}_i)}} > z_{\alpha/2}\right). \tag{3.17}$$

Let Ψ_i be any test for testing $H_{0,i} : \beta_i = 0$ versus $H_{1,i} : \beta_i \neq 0$. Define $\alpha_n(\Psi_i)$ be the size of the test over the collection of s-sparse vectors,

$$\alpha_n(\Psi_i) = \sup\{P_{\boldsymbol{\beta}}(\Psi_i = 1) : \boldsymbol{\beta} \in \mathbb{R}^p, ||\boldsymbol{\beta}||_0 \leq s, \boldsymbol{\beta}_i = 0\}.$$

For the power of the test, we consider the collection of s-sparse vectors with $|\beta_i| \geq \gamma$ for some given $\gamma > 0$ and define the power as

$$\zeta_n(\Psi_i; \gamma) = \inf\{P_{\boldsymbol{\beta}}(\Psi_i = 1) : \boldsymbol{\beta} \in \mathbb{R}^p, ||\boldsymbol{\beta}||_0 \leq s, |\beta_i| \geq \gamma\}.$$

Corollary 2. *Suppose the conditions of Theorem 3 hold. The test $\widehat{\Psi}_i$ defined at* (3.17) *satisfies*

$$\lim_{n\to\infty} \alpha_n(\widehat{\Psi}_i) \leq \alpha \quad \text{and} \quad \liminf_{n\to\infty} \frac{\zeta_n(\widehat{\Psi}_i; \gamma)}{\zeta_n^*(\gamma)} \geq 1, \tag{3.18}$$

where $\zeta_n^*(\gamma) := G(\alpha, \frac{\sqrt{n}\gamma}{\pi\sigma_{g_1(\boldsymbol{u})}})$ *and*

$$G(\alpha, u) = 2 - \Phi(z_{\alpha/2} + u) - \Phi(z_{\alpha/2} - u).$$

for $0 < \alpha < 1$ *and* $u \in \mathbb{R}^+$.

Consider the problem of testing an individual null hypothesis $H_{0,i} : \beta_i = 0$ versus the alternative $H_{1,i} : \beta_i \neq 0$ under the linear model

$$\widetilde{Y}_i = \widetilde{\boldsymbol{X}}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, ..., n, \tag{3.19}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

with $\widetilde{\boldsymbol{X}}_i \overset{i.i.d}{\sim} N(0, \Sigma_{XX})$ and $\epsilon_i \sim N(0, \sigma^2)$. As shown in Javanmard and Montanari (2014), for any test Ψ_i, if $\alpha_n(\Psi_i) \leq \alpha$, then

$$\limsup_{n\to\infty} \zeta_n(\Psi_i;\gamma) \leq G(\alpha, \frac{\sqrt{n}\gamma}{\sigma_d}),$$

where

$$\sigma_d = \frac{\sigma}{\sqrt{\sigma_{ii} - \Sigma_{i,S}\Sigma_{S,S}^{-1}\Sigma_{S,i}}}.$$

Hence, our test $\widehat{\Psi}_i$ has nearly optimal power in the following sense: it has power at least as large as the power of any other test Ψ_i based on a sample of size $\frac{n}{C_d}$, where the factor $C_d = \frac{\pi\sigma_{g_1(\boldsymbol{u}_i)}}{\sigma_d}$.

The results show that the proposed confidence intervals and hypothesis tests share the similar properties as those optimal procedures for the high-dimensional linear regression. They are more flexible in the sense that they are adaptive to unknown monotone marginal transformations.

4. Numerical Performance

The proposed estimation and inference procedures are easy to implement. We investigate in this section the numerical performance of the adaptive estimator (2.7), denoted by $\widehat{\boldsymbol{\beta}}_{\text{Copula}}(\boldsymbol{Y}, X)$ in this section, as well as the confidence procedure through simulations. The procedures are also applied to the analysis of the Communities and Crime Unnormalized Data from the UCI Machine Learning Repository.

4.1 Simulation Results for Estimation Accuracy

We first considered the performance of the the proposed estimator $\widehat{\boldsymbol{\beta}}_{\text{Copula}}(\boldsymbol{Y}, X)$ by comparing its estimation ℓ_2 loss and model selection error with those of the oracle Lasso

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

estimator $\widehat{\boldsymbol{\beta}}_{\text{Lasso}}(\widetilde{\boldsymbol{Y}}, \widetilde{X})$ that is performed on the transformed data $(\widetilde{\boldsymbol{Y}}, \widetilde{X})$, in which case we assumed the marginal transformations f_i are known and $\widetilde{\boldsymbol{Y}}$ is linear in $\widetilde{X}$. Then we compare $\widehat{\boldsymbol{\beta}}_{\text{Copula}}(\boldsymbol{Y}, X)$ with the regular Lasso estimator $\widehat{\boldsymbol{\beta}}_{\text{Lasso}}(\boldsymbol{Y}, X)$ and the elastic-net estimator $\widehat{\boldsymbol{\beta}}_{\text{enet}}(\boldsymbol{Y}, X)$, proposed in Zou and Hastie (2005), that are performed on $(\boldsymbol{Y}, X)$ directly.

The detailed simulation settings were as follows. Eight different combinations of the sample size, dimension, and sparsity with (n, p, s)=(100, 500, 10), (100, 500, 20), (100, 1000, 10), (100, 1000, 20), (200, 500, 10), (200, 500, 20), (200, 1000, 10) and (200, 1000, 20), were analyzed. In each case, we considered three different models for the covariance matrix Σ:

Model 1. Random Gaussian matrix: We begin with a random Gaussian matrix $A = (a_{i,j})_{1\le i,j\le d}$ where $d = p + 1$ and $a_{i,j} \overset{i.i.d.}{\sim} N(0, 1)$, and then make the last $p - s$ columns of A orthogonal to the first column of A via the Gram-Schmidt process, to obtain matrix B. The covariance matrix is $\Sigma = D^{-1/2}(B^\top B + I)^{-1}D^{-1/2}$, where $D = \text{diag}((B^\top B + I)^{-1})$.

Model 2. AR(1) matrix: We first generate a random orthogonal matrix $A = (a_{i,j})_{1\le i,j\le d}$ where $d = p + 1$. We then create a new $d \times d$ matrix B with the k-th column $B_k = \sqrt{1-\rho^2}A_k + \rho A_{k-1}$, for $k = 2, 3, ..., d$. The first column of B is the projection of A_1 onto the orthogonal complement of the span of the last $p-s$ columns of B. Define the covariance matrix $\Sigma = D^{-1/2}(B^\top B)^{-1}D^{-1/2}$, where $D = \text{diag}((B^\top B)^{-1})$. From this procedure the resulting covariance matrix Σ_{XX} is the first-order autoregressive (AR(1)) matrix with autocorrelation ρ. In the simulation we set $\rho = 0.5$.

Model 3. Compound symmetric matrix: In this case we start with a random orthogonal matrix $A = (a_{i,j})_{1\le i,j\le d}$ where $d = p + 1$, and create a new $d \times d$ matrix B

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

with k-th column $B_k = \sqrt{1-\rho^2}A_k + \rho A_1$ for $k = 2, 3, ..., d$. We then generate a new random vector $\tilde{A}_1 \sim N_d(0, I_d)$ where the first column of B is the projection of $\tilde{A}_1$ onto the orthogonal complement of the span of the last $p-s$ columns of B. Let the covariance matrix $\Sigma = D^{-1/2}(B^\top B)^{-1}D^{-1/2}$, where $D = \mathrm{diag}((B^\top B)^{-1})$. From this procedure the resulting covariance matrix Σ_{XX} is the compound symmetric matrix with correlation ρ. In the simulation we set $\rho = 0.5$.

After generating Σ from these models, we obtained n samples $(\widetilde{Y}_i, \widetilde{\boldsymbol{X}}_i^\top) \overset{i.i.d.}{\sim} N_d(0, \Sigma)$. For each choice of (n, p, s), we considered two settings. In the first, we set $Y_i = \exp(\widetilde{Y}_i)$, $X_{1j} = \Phi(\widetilde{X}_{ij})^5$, $X_{ij} = 2\widetilde{X}_{ij}^5 + 1$ for $j = 2, .., 10$, $X_{ij} = -\exp(\widetilde{X}_{ij})$ for $j = 11, 12, .., 30$, except for $X_{i,21} = \Phi(\widetilde{X}_{i,21})$, bounded by 0 and 1, while in the second setting we constrained $Y_i \in [0, 1]$ and set $Y_i = \Phi(\widetilde{Y}_i)$ with X_{ij}'s transformed the same way as in the first setting.

In each setting, the simulation was repeated $N_{\mathrm{sim}} = 500$ times and the tuning parameter λ was selected via 5-fold cross validation. The accuracy of the estimators was measured by the average ℓ_2 loss

$$e_{\mathrm{est}} = \frac{1}{N_{\mathrm{sim}}} \sum_{i=1}^{N_{\mathrm{sim}}} ||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2,$$

and the model selection error

$$e_{\mathrm{selection}} = \frac{1}{N_{\mathrm{sim}}} \sum_{i=1}^{N_{\mathrm{sim}}} \left(\frac{1}{p} \sum_{j=1}^{p} I\left(I(\widehat{\beta}_j = 0) \neq I(\beta_j = 0) \right) \right).$$

The simulation results under the first model for the three different estimates $\widehat{\boldsymbol{\beta}}_{\mathrm{Copula}}(\boldsymbol{Y}, X)$, $\widehat{\boldsymbol{\beta}}_{\mathrm{Lasso}}(\widetilde{\boldsymbol{Y}}, \widetilde{X})$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{Lasso}}(\boldsymbol{Y}, X)$ are summarized in Table 1. Results under the second and third models are given in the Supplement Cai and Zhang (2016).

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

Model 1									
	SNR	$\widehat{\boldsymbol{\beta}}_{\text{Copula}}(\boldsymbol{Y}, X)$		$\widehat{\boldsymbol{\beta}}_{\text{Lasso}}(\widetilde{\boldsymbol{Y}}, \widetilde{X})$		$\widehat{\boldsymbol{\beta}}_{\text{Lasso}}(\boldsymbol{Y}, X)$		$\widehat{\boldsymbol{\beta}}_{\text{enet}}(\boldsymbol{Y}, X)$	
(n, p, s)		$e_{\text{selection}}$	e_{est}	$e_{\text{selection}}$	e_{est}	$e_{\text{selection}}$	e_{est}	$e_{\text{selection}}$	e_{est}
$(100, 500, 10)_1$	129.1	0.0033	0.0526	0.0115	0.0351	0.0174	0.8835	0.0157	0.7954
$(100, 500, 10)_2$	129.1	0.0033	0.0526	0.0115	0.0351	0.0152	2.0468	0.0155	0.9489
$(100, 500, 20)_1$	81.6	0.0096	0.0840	0.0138	0.0562	0.0149	0.5452	0.0197	0.6647
$(100, 500, 20)_2$	81.6	0.0096	0.0840	0.0138	0.0562	0.0184	0.4282	0.0142	0.5168
$(100, 1000, 10)_1$	246.7	0.0018	0.0406	0.0090	0.0276	0.0147	1.1532	0.0129	1.0428
$(100, 1000, 10)_2$	246.7	0.0018	0.0406	0.0090	0.0276	0.0126	0.6369	0.0125	0.4932
$(100, 1000, 20)_1$	148.0	0.0052	0.0740	0.0081	0.0379	0.0276	0.8315	0.0142	0.8147
$(100, 1000, 20)_2$	148.0	0.0052	0.0740	0.0081	0.0379	0.0270	2.8695	0.0820	1.6456
$(200, 500, 10)_1$	125.3	0.0030	0.0484	0.0111	0.0251	0.0292	5.1155	0.0162	2.0187
$(200, 500, 10)_2$	125.3	0.0030	0.0484	0.0111	0.0251	0.0308	0.4595	0.0740	0.6657
$(200, 500, 20)_1$	88.8	0.0092	0.0706	0.0132	0.0485	0.0274	3.4115	0.0184	2.7923
$(200, 500, 20)_2$	88.8	0.0092	0.0706	0.0132	0.0485	0.0234	0.4748	0.0842	0.6532
$(200, 1000, 10)_1$	234.5	0.0017	0.0605	0.0092	0.0326	0.0267	4.0319	0.0128	5.6237
$(200, 1000, 10)_2$	234.5	0.0017	0.0605	0.0092	0.0326	0.0260	0.5675	0.0159	0.5145
$(200, 1000, 20)_1$	156.8	0.0044	0.0648	0.0085	0.0258	0.0438	0.6622	0.0141	0.8360
$(200, 1000, 20)_2$	156.8	0.0044	0.0648	0.0085	0.0258	0.0610	0.5130	0.0224	0.4036

Table 1: Simulation results for the synthetic data described under Model 1 in Section 4. The results corresponds to model selection error $e_{\text{selection}}$ and estimation error e_{est} for $\widehat{\boldsymbol{\beta}}_{\text{Copula}}(\boldsymbol{Y}, X)$, $\widehat{\boldsymbol{\beta}}_{\text{Lasso}}(\widetilde{\boldsymbol{Y}}, \widetilde{X})$, $\widehat{\boldsymbol{\beta}}_{\text{Lasso}}(\boldsymbol{Y}, X)$ and $\widehat{\boldsymbol{\beta}}_{\text{enet}}(\boldsymbol{Y}, X)$. The subscript i ($i = 1, 2$) in $(n, p, s)_i$ denotes the i-th setting of transformations

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

Table 1 shows that the performance of the proposed estimator $\widehat{\boldsymbol{\beta}}_{\text{Copula}}(\boldsymbol{Y}, X)$, which does not require the knowledge of the marginal transformations f_i, is as good as the oracle estimator $\widehat{\boldsymbol{\beta}}_{\text{Lasso}}(\widetilde{\boldsymbol{Y}}, \widetilde{X})$, which assumes the full knowledge of the transformations f_i. As expected, applying the Lasso and elastic-net estimator directly to the observed data leads to severely problematic model selection and parameter estimation.

4.2 Simulation Results for Statistical Inference

We considered the performance of the proposed confidence interval CI_i for the i-th coordinate β_i given in (3.16) based on the observed data $(Y_i, \boldsymbol{X}_i^\top)$ in terms of the coverage probability and expected length. In this section we denote the de-biased estimator in (3.14) as $\widehat{\boldsymbol{\beta}}^u_{\text{Copula}}(\boldsymbol{Y}, X)$. The confidence interval was compared with the confidence interval proposed in Javanmard and Montanari (2014) based on the transformed data $(Y_i, \boldsymbol{X}_i^\top)$ with de-biased estimator $\widehat{\boldsymbol{\beta}}^u_{\text{Lasso}}(\boldsymbol{Y}, X)$, and that of $\widehat{\boldsymbol{\beta}}^u_{\text{Lasso}}(\widetilde{\boldsymbol{Y}}, \widetilde{X})$ on the original data $(\widetilde{Y}_i, \widetilde{\boldsymbol{X}}_i^\top)$ while assuming the marginal transformations f_i are known.

In all simulations we set the significance level $\alpha = 0.05$, and considered eight cases: (n, p, s)=((100, 500, 10), (100, 500, 20), (100, 1000, 10), (100, 1000, 20), (200, 500, 10), (200, 500, 20), (200, 1000, 10) and (200, 1000, 20).

In each setting, the simulation was repeated 500 times. The tuning parameter λ was selected via 5-fold cross validation, and μ, a, b in Algorithm 2 were manually set to $\frac{1}{2}\sqrt{\frac{\log p}{n}}$, $\frac{1}{13}$, and 10 respectively. We discovered that the result is robust with respect to the choice of μ, a and b. Recalling that the $\boldsymbol{\beta}$ is constructed with first s elements nonzero, we constructed the 95% confidence intervals for the nonzero (active) coefficient β_1. The simulation results under Model 1 are summarized in Table 2, and the results under Model 2 and 3 are given in

Supplement Cai and Zhang (2016).

Table 2 summarizes the empirical coverage probability of the nominal 95% confidence intervals and the corresponding average lengths of β_1. The results show that the empirical coverage probability of $\widehat{\boldsymbol{\beta}}^u_{\text{Copula}}(\boldsymbol{Y}, X)$ is very close to the desired confidence level, while it is problematic to construct confidence intervals based on $\widehat{\boldsymbol{\beta}}^u_{\text{Lasso}}(\boldsymbol{Y}, X)$. The desired confidence level for the confidence intervals of an active coefficient is always small when we apply the de-biased Lasso estimator directly to the data. The confidence interval constructed by $\widehat{\boldsymbol{\beta}}^u_{\text{Copula}}(\boldsymbol{Y}, X)$ performs as well as that constructed by $\widehat{\boldsymbol{\beta}}^u_{\text{Lasso}}(\widetilde{\boldsymbol{Y}}, \widetilde{X})$, which needs additional information of the transformations. In particular, our method tends to have stable confidence interval lengths, while the length of confidence intervals constructed by $\widehat{\boldsymbol{\beta}}^u_{\text{Lasso}}(\boldsymbol{Y}, X)$ varies a lot according to the scale of data.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

Model 1						
	$\mathrm{CI}(\widehat{\boldsymbol{\beta}}^u_{\mathrm{Copula}}(\boldsymbol{Y}, X))$		$\mathrm{CI}(\widehat{\boldsymbol{\beta}}^u_{\mathrm{Lasso}}(\widetilde{\boldsymbol{Y}}, \widetilde{X}))$		$\mathrm{CI}(\widehat{\boldsymbol{\beta}}^u_{\mathrm{Lasso}}(\boldsymbol{Y}, X))$	
(n, p, s)	$l(\beta_1)$	$C(\beta_1)$	$l(\beta_1)$	$C(\beta_1)$	$l(\beta_1)$	$C(\beta_1)$
$(100, 500, 10)_1$	0.0223	0.956	0.0380	0.958	0.9398	0.332
$(100, 500, 10)_2$	0.0223	0.956	0.0380	0.958	0.1700	0.428
$(100, 500, 20)_1$	0.0241	0.948	0.0562	0.962	1.1152	0.462
$(100, 500, 20)_2$	0.0241	0.948	0.0562	0.962	0.1331	0.574
$(100, 1000, 10)_1$	0.0203	0.958	0.0275	0.956	0.8968	0.296
$(100, 1000, 10)_2$	0.0203	0.958	0.0275	0.956	0.1227	0.092
$(100, 1000, 20)_1$	0.0224	0.962	0.0378	0.962	0.9434	0.782
$(100, 1000, 20)_2$	0.0224	0.962	0.0378	0.962	0.1297	0.294
$(200, 500, 10)_1$	0.0138	0.946	0.0251	0.946	0.7472	0.230
$(200, 500, 10)_2$	0.0138	0.946	0.0251	0.946	0.1301	0.442
$(200, 500, 20)_1$	0.0154	0.952	0.0395	0.958	0.9163	0.068
$(200, 500, 20)_2$	0.0154	0.952	0.0395	0.958	0.1081	0.294
$(200, 1000, 10)_1$	0.0121	0.958	0.0326	0.956	0.7993	0.188
$(200, 1000, 10)_2$	0.0121	0.958	0.0326	0.956	0.1164	0.098
$(200, 1000, 20)_1$	0.0140	0.962	0.0257	0.952	0.8500	0.292
$(200, 1000, 20)_2$	0.0140	0.962	0.0257	0.952	0.1071	0.104

Table 2: Simulation results for the synthetic data under Model 1 described in Section 4. The results corresponds to 95% confidence intervals. $C(\beta_1)$ and $l(\beta_1)$ respectively stand for coverage probability and average lengths of the confidence interval for β_1. The subscript i $(i = 1, 2)$ in $(n, p, s)_i$ denotes the i-th setting of transformations.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

4.3 Analysis of Communities and Crime Unnormalized Data

We applied our estimation and inference procedures to a data example. The Communities and Crime Unnormalized Data from the UCI Machine Learning Repository combines socio-economic data from the 1990 Census, law enforcement data from the 1990 Law Enforcement Management and Administration Stats survey, and crime data from the 1995 FBI UCR. This dataset has been analyzed in Radchenko (2015); Buczak and Gifford (2010). In this example, we focused on explaining the response variable, percentage of women who are divorced, using various community characteristics, such as percentage of population that is African American and percent of people in owner occupied households, as well as law enforcement and crime information, such as percent of officers assigned to drug units. In order to further explore the high-dimensional setting, we used the state-level data of Pennsylvania, whose number of predictors is at least as large as the number of observations.

After removing the variables with NA's and two variables directly related to the response (total and male divorce percentages), the data has 101 observations and 114 predictors. To evaluate the performance of the proposed methods, we randomly split the data into a training set with 90 observations, and a test set with 11 observations. We performed such splits 100 times. Each time the proposed method and the regular Lasso were applied to the training set and the Root Mean Square Errors (RMSE) of the prediction (2.12) were calculated on the test set. The tuning parameters for both methods were selected via 5-fold cross validation over a grid $\lambda \in \{k \cdot \sqrt{\frac{\log p}{n}}\}_{k=1,2,\ldots,20}$. The average number of variables selected and RMSE are summarized in Table 3.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

	RMSE	Number of variable selected
Copula	1.66 (0.66)	4.61 (0.72)
Lasso	2.46 (0.43)	8.01 (0.70)

Table 3: Simulation results for the divorce percentage of women in the Pennsylvania Communities and Crime Data.

In addition, we used the proposed method for model selection. Applying the procedure to the whole Communities and Crime Unnormalized Data leads to four selected variables to explain the percentage of women who are divorced: PctFam2Par (percentage of families that are headed by two parents); PctKidsBornNeverMar (percentage of kids born to never married); PctPersOwnOccup (percent of people in owner occupied households); and PctSameHouse85 (percent of people living in the same house as in 1985). This selection procedure correctly excluded all the law enforcement and crime information and irrelevant features in community characteristics, such as the percentage of population that is African American and percentage of people 16 and over who are employed in manufacturing. In addition, the variables selected were all about family/house, which are directly related to divorce percentage.

5. Discussion

The Gaussian copula regression model is more flexible than the conventional linear model as it allows for unknown marginal monotonic transformations. The present paper proposes procedures for estimation and statistical inference that are adaptive to the unknown transformations. This is a significant advantage over other methods such as those for the additive regression model and single index model. An important observation is that the objective

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

function for the penalized least squares in classical high-dimensional regression only requires the sample covariances among X and $\boldsymbol{Y}$, which can be replaced by a Kendall's tau based estimator under the Gaussian copula regression model.

This idea can also be generalized to high-dimensional sparse multivariate regression. For example, under the linear model, the regularized estimator proposed in Rothman, Levina and Zhu (2010) and the block-structured regularized estimator introduced in Obozinski, Wainwright, and Jordan (2011) only require the knowledge of $X^\top X$ and $X^\top Y$. These can be replaced by the Kendall's tau based estimator $\widehat{\Sigma}_{XX}$ and $\widehat{\Sigma}_{XY}$ under the Gaussian copula model. Analogous analysis can be carried out to establish estimation consistency and inference results.

Similar ideas can be applied to other related models, such as the additive models in a Reproducing Kernel Hilbert Space (RKHS). In RKHS, the fitting procedure only requires the inner products among data points, and the proposed Algorithm 2 can be modified, via dual representation, for the construction of confidence intervals for additive models in RKHS. In addition, it is also possible to extend the model to discrete data and mixed data, by using the similar idea in Fan et al. (2016). These are interesting topics for future work.

Rank-based correlation matrix estimation has been studied in a number of settings, including the nonparanormal graphical model Liu et al. (2012); Xue and Zou (2012); Barber and Kolar (2015), high dimensional structured covariance/precision matrix estimation Xue and Zou (2012); Liu, Lafferty and Wasserman (2009); Liu et al. (2012), and sparse PCA model Han and Liu (2012); Mitra and Zhang (2014).

In the present paper, we only consider Kendall's tau-based estimator. Alternatively, one could use Spearman's rho. The results are similar and the same technique can be applied.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

6. Proofs

We prove the main results in this section. We begin by collecting a few technical lemmas that will be used in the proofs of the main results. These lemmas are proved in the Supplement Cai and Zhang (2016).

6.1 Technical Tools

The first lemma captures the asymptotics of certain U-statistics.

Lemma 1. *For $i = 1, 2, ..., p$, let $H_i = \boldsymbol{u}_i[2:d]^\top(\widehat{\Sigma}_{XY} - \widehat{\Sigma}_{XX}\boldsymbol{\beta}) = \boldsymbol{u}_i^\top \widehat{\Sigma}\boldsymbol{v}_0$, where $\boldsymbol{v}_0 = (1, -\boldsymbol{\beta}^\top)^\top$, then the asymptotic variance of $\sqrt{n}H_i$ is $\pi^2\sigma^2_{g_1(\boldsymbol{u}_i)}$, and moreover,*

$$\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|P(\frac{\sqrt{n}(H_i - \mathbb{E}[H_i])}{\pi\sigma_{g_1(\boldsymbol{u}_i)}} \leq x) - \Phi(x)| = 0,$$

where $\sigma_{g_1(\boldsymbol{u}_i)}$ is defined in (3.6).

Lemmas 2, 3, 4, and 5 control the vanishing terms in the construction of confidence intervals for each coordinate β_i; these lemmas are stated under the conditions of Theorem 3. We use $\boldsymbol{u}$ to denote $\boldsymbol{u}_i$ the solution to (3.13) for any fixed i.

Lemma 2. *If we take $\mu = C\sqrt{\frac{\log p}{n}}$ and $a, b > 0$ in Algorithm 2 for large C, then with probability at least $1 - 2p^{-2}$, the optimization problem* (3.13) *is feasible when n is large,*

$$|\Sigma_{XX}^{-1}\widehat{\Sigma}_{XX} - I|_\infty \leq \mu, \text{ and } b^{-1}n^{-a} \leq ||\boldsymbol{u}||_2 \leq ||\boldsymbol{u}||_1 \leq bn^{a/2}.$$

Lemma 3. *Let $\Sigma_{h_Z} = \mathrm{Var}(h_Z(\boldsymbol{Z})) \in \mathbb{R}^{d^2\times d^2}$ be the covariance matrix of $h_Z(\boldsymbol{Z}) = \mathbb{E}[\mathrm{sgn}(\boldsymbol{Z} - \boldsymbol{Z}') \otimes \mathrm{sgn}(\boldsymbol{Z} - \boldsymbol{Z}')|\boldsymbol{Z}]$, with $\otimes$ being the Kronecker product, and its corresponding estimator $\widehat{\Sigma}_{h_Z} = \frac{1}{n}\sum_i(\widehat{h}_Z(\boldsymbol{Z}_i) - \frac{1}{n}\sum_{i'}\widehat{h}_Z(\boldsymbol{Z}_{i'}))(\widehat{h}_Z(\boldsymbol{Z}_i) - \frac{1}{n}\sum_{i'}\widehat{h}_Z(\boldsymbol{Z}_{i'}))^\top$, with $\widehat{h}_Z(\boldsymbol{Z}_i) = \frac{1}{n-1}\sum_{i'\neq i}\mathrm{sgn}(\boldsymbol{Z}_i - \boldsymbol{Z}_{i'}) \otimes \mathrm{sgn}(\boldsymbol{Z}_i - \boldsymbol{Z}_{i'})$. Then with probability at least $1 - 5p^{-2}$,*

$$|x(\boldsymbol{u})^\top(\widehat{\Sigma}_{h_Z} - \Sigma_{h_Z})x(\boldsymbol{u})| \lesssim \sqrt{\frac{s\log p}{n^{1-2a}}}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

Lemma 4. *If $x(\boldsymbol{u}) = vec(\boldsymbol{u}\boldsymbol{v}_0^\top \circ \cos(\frac{\pi}{2}T))$ and $\widehat{x}(\boldsymbol{u}) = vec(\boldsymbol{u}\widehat{\boldsymbol{v}}^\top \circ \cos(\frac{\pi}{2}\widehat{T}))$, then with probability at least $1 - p^{-2}$,*

$$||x(\boldsymbol{u}) - \widehat{x}(\boldsymbol{u})||_1 \lesssim n^a \sqrt{\frac{s \log p}{n}}.$$

Lemma 5. *If $\sigma_{g_1(\boldsymbol{u})}$ is defined as at (3.6) with $\boldsymbol{u}$ the solution to (3.13) with any fixed i, then*

$$\sigma^2_{g_1(\boldsymbol{u})} \gtrsim n^{-2a}.$$

In addition, we need a few technical results adapted from Barber and Kolar (2015); Han and Liu (2013); Wegkamp and Zhao (2016); Zhao and Wegkamp (2014).

Lemma 6. *(An adapted version from Barber and Kolar (2015)) If $\boldsymbol{Z} \sim N_d(0, \Sigma)$, then $\mathrm{sgn}(\boldsymbol{Z}) = (\mathrm{sgn}(Z_1), ..., \mathrm{sgn}(Z_d))^\top$ is a random vector with subgaussian constant less than $\pi \cdot \kappa(\Sigma)$: for any $\boldsymbol{w} \in S^{d-1}$,*

$$\mathbb{E}[e^{t \cdot \boldsymbol{w}^\top \mathrm{sgn}(\boldsymbol{Z})}] \leq e^{t^2 \pi \cdot \kappa(\Sigma)}.$$

The next lemma characterizes the convergence rates of the Kendall's tau based correlation matrix estimator $\widehat{\Sigma}$ under different norms.

Lemma 7. *(An adapted version from Han and Liu (2013) and Wegkamp and Zhao (2016)) If $\widehat{\Sigma}$ is an estimator of Σ based on Kendall's tau, then*

1. *$P(|\widehat{\Sigma} - \Sigma|_\infty \lesssim \sqrt{\frac{\log p}{n}}) \geq 1 - 2p^{-2}$;*

2. *If $\kappa(\Sigma) \leq M$ for some $M > 0$, then*

$$P(||\widehat{\Sigma} - \Sigma||_2 \lesssim \max\{\sqrt{\frac{p+t}{n}}, \frac{p+t}{n}\}) \geq 1 - e^{-t}.$$

3. *If $\kappa_s(\Sigma) := \sup\{\lambda_{\max}(\Sigma_{S,S})/\lambda_{\min}(\Sigma_{S,S}) : S \subset [n], |S| = s\} \leq M_s$ for some $M_s > 0$, then*

$$P(||\widehat{\Sigma} - \Sigma||_{2,s} \lesssim \sqrt{\frac{s \log p}{n}}) \geq 1 - p^{-s}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

The following lemma provides a tight, pointwise deviation inequality of empirical cumulative distribution function.

Lemma 8. *(Adapted from Zhao and Wegkamp (2014)) If $\widehat{f}_i$ is defined as at* (2.11) *for $i \in \{1, , ..., p\}$, then for any $\epsilon \in (0, \sqrt{2\pi}]$, and $\gamma \in (0, 2)$, and $t \in \mathbb{R}$ such that $|f_i(t)| \leq \sqrt{\gamma \log n}$, we have*

$$P(|\widehat{f}_i(t) - f_i(t)| \geq \epsilon) \leq 2\exp(-\frac{n^{1-\gamma/2}}{12\pi\sqrt{2\pi}\sqrt{\gamma \log n}}\epsilon^2) - 3\log(8\pi n^\gamma \log n)\exp(-\frac{1}{64\sqrt{2\pi}}\frac{n^{1-\gamma/2}}{\sqrt{\log n}}),$$

where $F_i(t) = \Phi(f_i(t))$.

Lemma 9. *(Adapted from Mai and Zou (2013)) If $\widehat{f}_0$ is defined as at* (2.10) *, and for any $\gamma \in (0, 1)$,*

$$I_n := [f_0^{-1}(-\sqrt{2\gamma}\log n), f_0^{-1}(\sqrt{2\gamma}\log n)],$$

then we have

$$P(\sup_{t \in I_n} |\widehat{f}_i(t) - f_i(t)| \geq \epsilon) \leq 2\exp(-\frac{n^{1-\gamma}}{32\pi^2\gamma \log n}\epsilon^2) + \exp(-\frac{1}{16\pi\gamma}\frac{n^{1-\gamma}}{\log n}).$$

6.2 Proof of Theorem 1

This proof relies on Corollary 1 in Negahban et al. (2009) and Theorem 3.4 in Lee, Sun and Taylor (2015):

Lemma 10. *(An adapted version of Corollary 1 in Negahban et al. (2009)) If the loss function*

$$L(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \widehat{\Sigma}_{XX}\boldsymbol{\beta} - 2\widehat{\Sigma}_{YX}\boldsymbol{\beta} + 1$$

satisfies restricted strong convexity (RSC),

$$\delta L(\Delta, \boldsymbol{\beta}) := L(\boldsymbol{\beta} + \Delta) - L(\boldsymbol{\beta}) - \langle \nabla L(\boldsymbol{\beta}), \Delta \rangle \geq \kappa_L ||\Delta||_2^2, \tag{6.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

for some $\kappa_L > 0$ *and* $\Delta \in C(s) := \{\boldsymbol{\theta} \in \mathbb{R}^p : ||\boldsymbol{\theta}_{S^c}||_1 \leq 3||\boldsymbol{\theta}_S||_1, |S| \leq s\}$*, then for* $\lambda \geq 2||\nabla L(\boldsymbol{\beta})||_\infty$*, any optimal solution* $\widehat{\boldsymbol{\beta}}(\lambda)$ *to the convex program* (2.7) *satisfies the bound*

$$||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}||_2 \lesssim \sqrt{s}\lambda, \quad ||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}||_1 \lesssim s\lambda.$$

Lemma 11. *(An adapted version of Theorem 3.4 in Lee, Sun and Taylor (2015)) If we assume* $|\Sigma_{X_S X_{S^c}}|_\infty \leq 1 - \alpha$ *for some* $\alpha > 0$ *and* $S = \mathrm{supp}(\boldsymbol{\beta})$*, and* $\min_{i \in S} |\beta_i| \geq \frac{8}{\gamma_1}(1 + \frac{4(2-\alpha)}{\alpha})M\sqrt{\frac{s \log p}{n}}$*, then for* $\lambda = \frac{8(2-\alpha)}{\alpha}M\sqrt{\frac{s \log p}{n}}$*, with probability at least* $1 - 2p^{-1}$*,*

$$\mathrm{sgn}(\boldsymbol{\beta}) = \mathrm{sgn}(\widehat{\boldsymbol{\beta}}(\lambda)).$$

To prove Theorem 1, it is sufficient to verify (6.1) and calculate $||\nabla L(\boldsymbol{\beta})||_\infty$. We divide this into two steps.

Step 1

By the definition of $\delta L(\Delta, \boldsymbol{\beta})$,

$$\begin{aligned}
\delta L(\Delta, \boldsymbol{\beta}) =& L(\boldsymbol{\beta} + \Delta) - L(\boldsymbol{\beta}) - \langle \nabla L(\boldsymbol{\beta}), \Delta \rangle \\
=& \frac{1}{2}(\boldsymbol{\beta} + \Delta)^\top \widehat{\Sigma}_{XX}(\boldsymbol{\beta} + \Delta) - \widehat{\Sigma}_{YX}(\boldsymbol{\beta} + \Delta) - \frac{1}{2}\boldsymbol{\beta}^\top \widehat{\Sigma}_{XX}\boldsymbol{\beta} \\
& + \widehat{\Sigma}_{YX}\boldsymbol{\beta} - \Delta^\top(\widehat{\Sigma}_{XX}\boldsymbol{\beta} - \widehat{\Sigma}_{XY}) \\
=& \frac{1}{2}\Delta^\top \widehat{\Sigma}_{XX}\Delta.
\end{aligned}$$

Before proving (6.1), we state the adapted version of the reduction principle from Rudelson and Zhou (2013).

Lemma 12. *(The adapted version of Theorem 10 in Rudelson and Zhou (2013)) If* $\delta \in (0, \frac{1}{5})$ *and* $k_0 = 3$*. Then there exists a constant* C_0 *that is not dependent on* n, p, s*, such that* $\widetilde{s} = C_0 s$*. If* $E(\widetilde{s}) = \{\boldsymbol{w} \in \mathbb{R}^p : ||\boldsymbol{w}||_0 = \widetilde{s}\}$ *for* $\widetilde{s} < p$ *and* $E = \mathbb{R}^p$ *otherwise, and* $\widehat{\Sigma}_{XX}$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

satisfies

$$\forall \boldsymbol{w} \in E(\widetilde{s}) \quad (1-\delta)||\boldsymbol{w}||_2^2 \leq \boldsymbol{w}^\top \widehat{\Sigma}_{XX} \boldsymbol{w} \leq (1+\delta)||\boldsymbol{w}||_2^2, \tag{6.2}$$

then for any $\boldsymbol{w} \in C(s)$,

$$(1-5\delta)||\boldsymbol{w}||_2^2 \leq \boldsymbol{w}^\top \widehat{\Sigma}_{XX} \boldsymbol{w} \leq (1+3\delta)||\boldsymbol{w}||_2^2. \tag{6.3}$$

This claim implies that it is sufficient to show, for $\Delta \in E(\widetilde{s}) = \{\boldsymbol{w} \in \mathbb{R}^p : ||\boldsymbol{w}||_0 = \widetilde{s}\}$ and some $\delta \in (0, 1/5)$,

$$|\Delta^\top \widehat{\Sigma}_{XX} \Delta| \geq (1-\delta)||\Delta||^2.$$

Then Lemma 7 with the condition $\kappa_s(\Sigma) \leq M$, together with the fact that the spectral norm of a submatrix is bounded by the spectral norm of the whole matrix, implies that for $\Delta \in \{\boldsymbol{w} \in \mathbb{R}^p : ||\boldsymbol{w}||_0 = \widetilde{s}\}$, with probability at least $1 - p^{-2}$, we have

$$\begin{aligned}
|\Delta^\top \widehat{\Sigma}_{XX} \Delta| =& |\Delta^\top \Sigma_{XX} \Delta + \Delta^\top (\widehat{\Sigma}_{XX} - \Sigma_{XX}) \Delta| \\
\geq & |\Delta^\top \Sigma_{XX} \Delta| - |\Delta^\top (\widehat{\Sigma}_{XX} - \Sigma_{XX}) \Delta| \\
\geq & |\Delta^\top \Sigma_{XX} \Delta| - ||\widehat{\Sigma}_{XX} - \Sigma_{XX}||_{2,\widetilde{s}} \cdot ||\Delta||_2^2 \\
\geq & |\Delta^\top \Sigma_{XX} \Delta| - \sqrt{\frac{C_0 s \log p}{n}} ||\Delta||_2^2 \\
\geq & \gamma_1 ||\Delta||_2^2 - \sqrt{\frac{C_0 s \log p}{n}} ||\Delta||_2^2.
\end{aligned}$$

Therefore (6.1) holds when $s \log p / n \to 0$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

Step 2:

$$\begin{aligned}
||\nabla L(\boldsymbol{\beta})||_\infty &= ||\widehat{\Sigma}_{XX}\boldsymbol{\beta} - \widehat{\Sigma}_{XY}||_\infty = ||\widehat{\Sigma}_{XX}\Sigma_{XX}^{-1}\Sigma_{XY} - \widehat{\Sigma}_{XY}||_\infty \\
=&||(\widehat{\Sigma}_{XX} - \Sigma_{XX})\Sigma_{XX}^{-1}\Sigma_{XY} + \Sigma_{XY} - \widehat{\Sigma}_{XY}||_\infty \\
=&||(\widehat{\Sigma}_{XX} - \Sigma_{XX})\boldsymbol{\beta} + \Sigma_{XY} - \widehat{\Sigma}_{XY}||_\infty \\
\le&||(\widehat{\Sigma} - \Sigma)(1, -\boldsymbol{\beta}^\top)^\top||_\infty \le |\widehat{\Sigma} - \Sigma|_\infty ||(1, -\boldsymbol{\beta}^\top)^\top||_1 \\
\le&\sqrt{\frac{\log p}{n}} \cdot (1 + ||\boldsymbol{\beta}||_1) \le \sqrt{\frac{\log p}{n}} \cdot (1 + \sqrt{s}||\boldsymbol{\beta}||_2) \\
=&\sqrt{\frac{\log p}{n}} \cdot (1 + \sqrt{s}||\Sigma_{XX}^{-1}\Sigma_{XY}||_2) \le \sqrt{\frac{\log p}{n}} \cdot (1 + \sqrt{s}||\Sigma_{XX}^{-1}||_2||\Sigma_{XY}||_2) \\
\le&\sqrt{\frac{s\log p}{n}}M.
\end{aligned}$$

Therefore if we choose λ such that $\lambda > 2M\sqrt{\frac{s\log p}{n}}$, then we have $\lambda_n \ge 2||\nabla L(\boldsymbol{\beta})||_\infty$. Then it follows from Theorem 10 that, when $s\log p/n \to 0$ with probability at least $1 - 2p^{-2}$,

$$\begin{aligned}
||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}||_2 &\lesssim \sqrt{s}\lambda \lesssim \sqrt{\frac{s\log p}{n}}; \\
||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}||_1 &\lesssim s\lambda \lesssim s\sqrt{\frac{\log p}{n}}; \\
\mathrm{sgn}(\boldsymbol{\beta}) &= \mathrm{sgn}(\widehat{\boldsymbol{\beta}}(\lambda)).
\end{aligned}$$

□

6.3 Proof of Theorem 2

According to Lemma 8 and by the union bound

$$\begin{aligned}
P(\max_{i\in\{1,2,\ldots,p\}} |\widehat{f}_i(t) - f_i(t)| \ge \epsilon) \le& 2\exp(\log d - \frac{n^{1-\gamma/2}}{12\pi\sqrt{2\pi}\sqrt{\gamma\log n}}\epsilon^2) \\
&- 3\log(8\pi n^\gamma \log n)\exp(\log d - \frac{1}{64\sqrt{2\pi}}\frac{n}{\sqrt{n^\gamma \log n}}).
\end{aligned}$$

Therefore by taking $\epsilon = \sqrt{\frac{24\pi\sqrt{2\pi}\sqrt{\gamma\log n}\log d}{n^{1-\gamma/2}}}$, then for $t \in \mathbb{R}$ such that $|f_i(t)| \le \sqrt{\gamma\log n}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

with probability at least $1 - d^{-1} - n^{-1}$,

$$\max_{i \in [0,1,2,\ldots,p]} |\widehat{f}_i(t) - f_i(t)| \lesssim \frac{(\gamma \log n)^{1/4}\sqrt{\log d}}{n^{1/2-\gamma/4}}. \tag{6.4}$$

Since $\max_{i=1,\ldots,p} F_i(x_i^*) \in (\delta^*, 1-\delta^*)$, there exists some constant $M_* > 0$, such that

$$\max_{i=1,\ldots,p} f_i(x_i^*) = \max_{i=1,\ldots,p} \Phi^{-1}(F_i(x_i^*)) < M_*.$$

Therefore, if we let $\gamma = \frac{M_*^2}{\log n}$, we have $\max_{i=1,\ldots,p} f_i(x_i^*) \leq \sqrt{\gamma \log n}$. Then by (6.4), with probability at least $1 - d^{-1} - n^{-1}$,

$$\max_{i \in \{1,2,\ldots,p\}} |\widehat{f}_i(x_i^*) - f_i(x_i^*)| \lesssim \sqrt{\frac{\log d}{n}}. \tag{6.5}$$

In addition, use the fact in Theorem 1, with probability at least $1 - d^{-1} - n^{-1}$,

$$\begin{aligned}
&|\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i - \mu^*| = |\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i - \sum_{i=1}^p f_i(x_i^*)\beta(\lambda)_i| \\
\leq& |\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i - \sum_{i=1}^p f_i(x_i^*)\widehat{\beta}(\lambda)_i| + |\sum_{i=1}^p f_i(x_i^*)\widehat{\beta}(\lambda)_i - \sum_{i=1}^p f_i(x_i^*)\beta(\lambda)_i| \\
\lesssim& (||\boldsymbol{\beta}||_1 + s\sqrt{\frac{\log p}{n}}) \cdot \max_{i \in \{1,2,\ldots,p\}} |\widehat{f}_i(t) - f_i(t)| + ||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}||_1 \\
\leq& ||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}||_1 + (s||\boldsymbol{\beta}||_2 + s\sqrt{\frac{\log p}{n}}) \cdot \max_{i \in \{1,2,\ldots,p\}} |\widehat{f}_i(t) - f_i(t)| \\
\lesssim& s\sqrt{\frac{\log d}{n}},
\end{aligned}$$

where the last inequality results from the fact $\boldsymbol{\beta} = \Sigma_{XX}^{-1}\Sigma_{XY}$, and then

$$||\boldsymbol{\beta}||_2 = ||\Sigma_{XX}^{-1}\Sigma_{XY}||_2 \leq \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \leq M.$$

This implies with probability at least $1 - d^{-1} - n^{-1}$,

$$\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i \in f_0^{-1}(B_r(f_0(\mu^*))). \tag{6.6}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

Further, use Lemma 9 and applying the similar derivation as before, we obtain that, with probability at least $1 - d^{-1}$,

$$|\widehat{f}_0(f_0^{-1}(\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i)) - f_0(f_0^{-1}(\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i))| \lesssim \sqrt{\frac{\log d}{n}}. \tag{6.7}$$

Combining (6.5),(6.6) and (6.7), with probability at least $1 - 2/n - 2/d - 1/\log n$,

$$\begin{aligned}
|\mu^* - \widehat{\mu}^*| =& |\widehat{f}_0^{-1}(\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i) - f_0^{-1}(\sum_{i=1}^p f_i(x_i^*)\beta(\lambda)_i)| \\
\le& |\widehat{f}_0^{-1}(\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i) - f_0^{-1}(\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i)| + |f_0^{-1}(\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i) - f_0^{-1}(\sum_{i=1}^p f_i(x_i^*)\beta(\lambda)_i)| \\
\le& |\widehat{f}_0^{-1}(\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i) - f_0^{-1}(\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i)| + \frac{1}{c_2}|\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i - \sum_{i=1}^p f_i(x_i^*)\beta(\lambda)_i| \\
\overset{(i)}{\le}& |\widehat{f}_0(f_0^{-1}(\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i)) - f_0(f_0^{-1}(\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i)) + \frac{1}{c_2}|\sum_{i=1}^p \widehat{f}_i(x_i^*)\widehat{\beta}(\lambda)_i - \sum_{i=1}^p f_i(x_i^*)\beta(\lambda)_i| \\
\lesssim& \sqrt{\frac{\log d}{n}} + s\sqrt{\frac{\log d}{n}} \\
\lesssim& s\sqrt{\frac{\log d}{n}},
\end{aligned}$$

where the inequality (i) is due to the following.

Claim: For two increasing functions f_1, f_2, if $|f_1(f_1^{-1}(t)) - f_2(f_1^{-1}(t))| < c_1$ for some $t \in \mathbb{R}$ and $c_1 > 0$, and if $|f_2(v_1) - f_2(v_2)| \ge c_2|v_1 - v_2|$ for some $c_2 > 0$, then

$$|f_1^{-1}(t) - f_2^{-1}(t)| \le \frac{c_1}{c_2}.$$

In effect, if $|f_1^{-1}(t) - f_2^{-1}(t)| > \frac{c_1}{c_2}$, then

$$\begin{aligned}
|f_1(f_1^{-1}(t)) - f_2(f_1^{-1}(t))| =& |f_1(f_1^{-1}(t)) - f_2(f_2^{-1}(t)) + f_2(f_2^{-1}(t)) - f_2(f_1^{-1}(t))| \\
\ge& |f_2(f_2^{-1}(t)) - f_2(f_1^{-1}(t))| - |f_1(f_1^{-1}(t)) - f_2(f_2^{-1}(t))| \\
>& c_2 \cdot \frac{c_1}{c_2} - 0 = c_1.
\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

This leads to a contradiction.

□

6.4 Proof of Theorem 3

Before proceeding, we should determine μ to make the optimization problem (3.13) feasible. By Lemma 2, it is sufficient to set $\mu = C\sqrt{\frac{\log p}{n}}$ for some sufficient large constant C. According to (3.14) in Algorithm 2,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^u &= \widehat{\boldsymbol{\beta}}(\lambda) + M(\widehat{\Sigma}_{XY} - \widehat{\Sigma}_{XX}\widehat{\boldsymbol{\beta}}(\lambda)) \\
&= \boldsymbol{\beta} - \boldsymbol{\beta} + \widehat{\boldsymbol{\beta}}(\lambda) + M\widehat{\Sigma}_{XY} - M\widehat{\Sigma}_{XX}\widehat{\boldsymbol{\beta}}(\lambda) \\
&= \boldsymbol{\beta} + (M\widehat{\Sigma}_{XY} - M\widehat{\Sigma}_{XX}\boldsymbol{\beta}) + (M\widehat{\Sigma}_{XX} - I)(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\lambda)).
\end{aligned}$$

This implies

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}^u - \boldsymbol{\beta}(\lambda)) = \sqrt{n}(M\widehat{\Sigma}_{XY} - M\widehat{\Sigma}_{XX}\boldsymbol{\beta}) + \sqrt{n}(I - M\widehat{\Sigma}_{XX})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\lambda)). \tag{6.8}$$

We control the two terms on the right hand side separately.

Step 1: $||\sqrt{n}(I - M\widehat{\Sigma}_{XX})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\lambda))||_\infty \to 0$ with high probability.

By Theorem 1 and Lemma 2, with probability at least $1 - 3p^{-2}$,

$$\begin{aligned}
||\sqrt{n}(I - M\widehat{\Sigma}_{XX})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\lambda))||_\infty &\leq \sqrt{n}||I - M\widehat{\Sigma}_{XX}||_\infty||\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\lambda)||_1 \\
&\leq \sqrt{n}\mu \cdot s\sqrt{\frac{\log p}{n}} \lesssim \sqrt{n}\sqrt{\frac{\log p}{n}} \cdot s\sqrt{\frac{\log p}{n}}.
\end{aligned}$$

Therefore, when $\frac{s\log p}{\sqrt{n}} \to 0$, with probability at least $1 - 3p^{-2}$,

$$||\sqrt{n}(I - M\widehat{\Sigma}_{XX})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\lambda))||_\infty \to 0.$$

Step 2: Asymptotics of $\sqrt{n}(\boldsymbol{u}_i'\widehat{\Sigma}_{XY} - \boldsymbol{u}_i'\widehat{\Sigma}_{XX}\boldsymbol{\beta})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

With Lemma 3, Lemma 4, and by $|\Sigma_{h_Z}|_\infty \le 1$, when $\frac{s\log p}{\sqrt{n}} \to 0$, we have with probability at least $1-p^{-2}$,

$$\begin{aligned}
|\sigma^2_{g_1(\boldsymbol{u}_i)} - \widehat{\sigma}^2_{g_1(\boldsymbol{u}_i)}| &= |x(\boldsymbol{u}_i)^\top \Sigma_{h_Z} x(\boldsymbol{u}_i) - \widehat{x}(\boldsymbol{u}_i)^\top \widehat{\Sigma}_{h_Z} \widehat{x}(\boldsymbol{u}_i)| \\
&\le |(x(\boldsymbol{u}_i) - \widehat{x}(\boldsymbol{u}_i))^\top \Sigma_{h_Z}(x(\boldsymbol{u}_i) - \widehat{x}(\boldsymbol{u}_i))| + |x(\boldsymbol{u}_i)^\top(\widehat{\Sigma}_{h_Z} - \Sigma_{h_Z})x(\boldsymbol{u}_i)| \\
&\le ||x(\boldsymbol{u}_i) - \widehat{x}(\boldsymbol{u}_i)||_1^2 + |x(\boldsymbol{u}_i)^\top(\widehat{\Sigma}_{h_Z} - \Sigma_{h_Z})x(\boldsymbol{u}_i)| \\
&\lesssim n^{2a}\frac{s\log p}{n} + \sqrt{\frac{s\log p}{n^{1-2a}}} \lesssim \sqrt{\frac{s\log p}{n^{1-2a}}}.
\end{aligned}$$

Lemma 5 shows that $\sigma^2_{g_1(\boldsymbol{u}_i)} \gtrsim n^{-2a}$. It follows that $|\frac{\widehat{\sigma}^2_{g_1(\boldsymbol{u}_i)}}{\sigma^2_{g_1(\boldsymbol{u}_i)}} - 1| \lesssim \sqrt{\frac{s\log p}{n^{1-6a}}}$. In addition, due to the positiveness of σ_{g1} and $\widehat{\sigma}_{g_1}$, when $\frac{s\log p}{\sqrt{n}} \to 0$ and $a < \frac{1}{12}$, $\widehat{\sigma}_{g_1(\boldsymbol{u}_i)}/\sigma_{g_1(\boldsymbol{u}_i)} \to 1$ in probability. Then according to Lemma 1, for any $\epsilon > 0$,

$$\begin{aligned}
P(\frac{\sqrt{n}(H_i - \mathbb{E}[H_i])}{\pi\widehat{\sigma}_{g_1(\boldsymbol{u}_i)}} \le x) =& P(\frac{\sigma_{g_1(\boldsymbol{u}_i)}}{\widehat{\sigma}_{g_1(\boldsymbol{u}_i)}}\frac{\sqrt{n}(H_i - \mathbb{E}[H_i])}{\pi\sigma_{g_1(\boldsymbol{u}_i)}} \le x) \\
\le& P(\frac{\sqrt{n}(H_i - \mathbb{E}[H_i])}{\pi\sigma_{g_1(\boldsymbol{u}_i)}} \le \frac{x}{1-\epsilon}) + P(\frac{\widehat{\sigma}_{g_1(\boldsymbol{u}_i)}}{\sigma_{g_1(\boldsymbol{u}_i)}} \ge \frac{1}{1-\epsilon}) \\
\to& \Phi(\frac{x}{1-\epsilon}) \qquad \text{as } n \to \infty,
\end{aligned}$$

where the last limit results from Lemma 1.

As $\epsilon \to 0$, we have

$$\limsup_{n\to\infty} P(\frac{\sqrt{n}(H_i - \mathbb{E}[H_i])}{\pi\widehat{\sigma}_{g_1(\boldsymbol{u}_i)}} \le x) \le \Phi(x).$$

Similarly, we have

$$P(\frac{\sqrt{n}(H_i - \mathbb{E}[H_i])}{\pi\widehat{\sigma}_{g_1(\boldsymbol{u}_i)}} \le x) \ge P(\frac{\sqrt{n}(H_i - \mathbb{E}[H_i])}{\pi\sigma_{g_1(\boldsymbol{u}_i)}} \le x(1-\epsilon)) - P(\frac{\widehat{\sigma}_{g_1(\boldsymbol{u}_i)}}{\sigma_{g_1(\boldsymbol{u}_i)}} \le 1-\epsilon)$$

This leads to

$$\liminf_{n\to\infty} P(\frac{\sqrt{n}(H_i - \mathbb{E}[H_i])}{\pi\widehat{\sigma}_{g_1(\boldsymbol{u}_i)}} \le x) \ge \Phi(x).$$

In conclusion, when $\frac{s\log p}{\sqrt{n}} \to 0$, we have

$$\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|P(\frac{\sqrt{n}(H_i - \mathbb{E}[H_i])}{\pi\widehat{\sigma}_{g_1(\boldsymbol{u}_i)}} \le x) - \Phi(x)| = 0.$$

$\square$

Supplemental Materials

In the supplemental materials, we provide the proofs of auxiliary lemmas. Some additional simulation results are given in the supplement.

Acknowledgement

The research of T. Tony Cai was supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334.

References

Barber, R. and Kolar, M. (2015). Rocket: Robust confidence intervals via kendalls tau for transelliptical graphical models. *arXiv preprint*, arXiv:1502.07641.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* **37**, 1705-1732.

Buczak, A. and Gifford, C. (2010). Fuzzy association rule mining for community crime pattern discovery. In *ACM SIGKDD Workshop on Intelligence and Security Informatics*, ACM.

Cai, T. and Guo Z. (2015). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *arXiv preprint*, arXiv:1506.05539v1.

Cai, T. and Zhang, L. (2015). Supplementary material of high-dimensional gaussian copula regression: Adaptive estimation and statistical inference.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

Carroll, R. and Ruppert, D. (1988). *Transformation and Weighting in Regression, volume 30.* CRC Press.

Crane G. and Hoek J. (2008). Conditional expectation formulae for copulas. *Australian & New Zealand Journal of Statistics* **50**, 53-67.

Fan, J., Liu, H., Ning, Y., and Zou, H. (2017). High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society Series B* **79**, 405-421.

Foster, J., Taylor, J., and Nan, B. (2013). Variable selection in monotone single-index models via the adaptive lasso. *Statistics in Medicine* **32**, 3944-3954.

Gu, Q., Cao, Y., Ning, Y., and Liu, H. (2015). Local and global inference for high dimensional nonparanormal graphical models. *arXiv preprint*, arXiv:1502.02347.

Han, F. and Liu, H. (2012). Transelliptical component analysis. *In Advances in Neural Information Processing Systems*, 368-376.

Han, F. and Liu, H. (2013). Optimal rates of convergence for latent generalized correlation matrix estimation in transelliptical distribution. *arXiv preprint*, arXiv:1305.6916.

Javanmard, A. and Montanari, A. (2014). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory* **60**, 6522-6554.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* **15**, 2869-2909.

Johnston, J. and DiNardo, J. (1997). *Econometric Methods.* Cambridge Univ Press.

Kruskal, W. (1958). Ordinal measures of association. *Journal of the American Statistical Association* **53**, 814-861.

Lee, J., Sun, Y., and Taylor, J. (2013). On model selection consistency of M-estimators with geometrically decomposable penalties. *Electronic Journal of Statistics* **9**, 608-642.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric gaussian copula

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

graphical models. *The Annals of Statistics* **40**, 2293- 2326.

Liu, H., Lafferty, L., and Wasserman, L.(2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research* **10**, 2295-2328.

Loh, P. and Wainwright, M.(2013). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *In Advances in Neural Information Processing Systems*, 476-484.

Lu, Y., Dhillon, P., Foster, D. and Ungar, L.(2013) . Faster ridge regression via the subsampled randomized Hadamard transform. *In Advances in Neural Information Processing Systems*, 369-377.

Luo, S., and Ghosal, S. (2016). Forward selection and estimation in high dimensional single index models. *Statistical Methodology* **33**, 172-179.

Mai, Q. and Zou, H., 2013. Semiparametric Sparse Discriminant Analysis. *arXiv preprint*, arXiv:1304.4983.

Masarotto, G., and Varin, C. (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics* **6**, 1517-1549.

McDonald, J. (2009). *Handbook of Biological Statistics*, volume 2. Sparky House Publishing, Baltimore, MD.

Meier, L., Van de Geer, S., and Buhlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics* **37**, 3779-3821.

Mitra, R., and Zhang, C. H. (2014). Multivariate analysis of nonparametric estimates of large correlation matrices. *arXiv preprint*, arXiv:1403.6195.

Negahban, S., Yu, B., Wainwright, M. J., and Ravikumar, P. K. (2009). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *In Advances in Neural Information Processing Systems*, 1348-1356.

Ni, L., Cook, R. D., and Tsai, C. L. (2005). A note on shrinkage sliced inverse regression. *Biometrika* **92**, 242-247.

Noh, H., Ghouch, A. E., and Bouezmarni, T. (2013). Copula-based regression estimation and inference. *Journal of*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

the American Statistical Association **108**, 676-688.

Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics* **39**, 1-47.

Osgood, D. (2000). Poisson-based regression analysis of aggregate crime rates. *Journal of quantitative criminology* **16**, 21-43.

Pitt, M., Chan, D., and Kohn, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika* **93**, 537-554.

Radchenko, P. (2015). High dimensional single index models. *Journal of Multivariate Analysis* **139**, 266-282.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory* **57**, 6976-6994.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman,L.(2009). Sparse additive models. *Journal of the Royal Statistical Society Series B* **71**,1009-1030.

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

Rudelson, M. and Zhou, S.(2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory* **59**, 3434-3447.

Stein,E. and Shakarchi R. (2009). *Real Analysis: Measure Theory, Integration, and Hilbert Spaces.* Princeton University Press.

Sungur, E.(2005). Some observations on copula regression functions. *Communications in Statistics—Theory and Methods* **34**, 1967-1978.

Van de Geer, S., Bhlmann, P., Ritov, Y. A., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42**, 1166-1202.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

Wang, W. and Hua Z. (2014). A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 1155-1165.

Wegkamp, M., and Zhao, Y. (2016). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli* **22**, 1184-1226.

Xue, L., and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics* **40**, 2541-2571.

Yi, X., Wang, Z., Caramanis, C., and Liu, H. (2015). Optimal linear estimation under unknown nonlinear transform. *In Advances in Neural Information Processing Systems*, 1549-1557.

Yu, Z., Zhu, L., Peng, H., and Zhu, L. (2013). Dimension reduction and predictor selection in semiparametric models. *Biometrika* **100**, 641-654.

Yuan, M., and Zhou, D. X. (2015). Minimax Optimal Rates of Estimation in High Dimensional Additive Models: Universal Phase Transition. *arXiv preprint*, arXiv:1503.02817.

Zhang, C. and Zhang, S.(2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B* **76**, 217-242.

Zhao, Y., and Wegkamp, M. (2014). Semiparametric Gaussian copula classification. *arXiv preprint*, arXiv:1411.2944.

Zou, H. and Hastie, T.(2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B* **67**, 301- 320.

T. Tony Cai, University of Pennsylvania

E-mail: tcai@wharton.upenn.edu

Linjun Zhang, University of Pennsylvania

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0041

E-mail: linjunz@wharton.upenn.edu