Coordinatewise Gaussianization: Theories and Applications

Qing Mai, Di He & Hui Zou

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–15: Theory and Methods
https://doi.org/10.1080/01621459.2022.2044825

Taylor & Francis
Taylor & Francis Group

Coordinatewise Gaussianization: Theories and Applications

Qing Mai[a], Di He[b], and Hui Zou[c]

[a]Department of Statistics, Florida State University, Tallahassee, FL; [b]School of Economics, Nanjing University, Nanjing, China; [c]School of Statistics, University of Minnesota, Minneapolis, MN

ABSTRACT

In statistical analysis, researchers often perform coordinatewise Gaussianization such that each variable is marginally normal. The normal score transformation is a method for coordinatewise Gaussianization and is widely used in statistics, econometrics, genetics and other areas. However, few studies exist on the theoretical properties of the normal score transformation, especially in high-dimensional problems where the dimension p diverges with the sample size n. In this article, we show that the normal score transformation uniformly converges to its population counterpart even when $\log p = o\,(n/\log n)$. Our result can justify the normal score transformation prior to any downstream statistical method to which the theoretical normal transformation is beneficial. The same results are established for the Winsorized normal transformation, another popular choice for coordinatewise Gaussianization. We demonstrate the benefits of coordinatewise Gaussianization by studying its applications to the Gaussian copula model, the nearest shrunken centroids classifier and distance correlation. The benefits are clearly shown in theory and supported by numerical studies. Moreover, we also point out scenarios where coordinatewise Gaussinization does not help and even causes damages. We offer a general recommendation on how to use coordinatewise Gaussianization in applications. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received February 2021
Accepted February 2022

KEYWORDS
Gaussianization; Gaussianized distance correlation; Heavy tails; Normal score transformation; Nearest shrunken centroids classifier.

1. Introduction

In statistical analysis, researchers often perform coordinatewise Gaussianization such that each variable is marginally normal. The Gaussianization benefits subsequent analysis in two ways. On the one hand, there is a rich literature on statistical models developed under normality assumptions. Guassianization allows us to borrow the strengths of these works. On the other hand, normal variables have sub-Gaussian tails. A large number of high-dimensional methods require variables to be sub-Gaussian in order to succeed in ultra-high dimensions, while heavy tails often negatively impact the performance of these methods. With coordinatewise Gaussianization, these methods can be readily applied.

We are interested in two closely related and popular methods for coordinatewise Gaussianization; namely, the normal score (NS) estimator and the Winsorized estimator. Consider $\mathbf{X} = (X_1, \ldots, X_p)$, where $X_j \in \mathbb{R}$ for $j = 1, \ldots, p$. Recall that, for any continuous X_j, we have

$$T_j(X_j) = \Phi^{-1} \circ F_j(X_j) \sim N(0, 1), \tag{1}$$

where Φ is the cumulative distribution function (CDF) for the standard normal random variable, and F_j is the CDF for X_j. Hence, if we knew F_j and hence, $\Phi^{-1} \circ F_j$, we could transform our data to be marginally normal according to (1). However, in practice F_j is generally not available. Consider n independent copies of $\mathbf{X}$, $\mathbf{X}^i, i = 1, \ldots, n$ and let $\widehat{F}_j$ be the empirical CDF for X_j. The NS estimator and the Winsorized estimator are defined as follows:

- The NS estimator:

$$\widehat{T}_j^{(ns)} = \Phi^{-1} \circ (\frac{n}{n+1}\widehat{F}_j); \tag{2}$$

- The Winsorized estimator:

$$\widehat{T}_j^{(w)} = \Phi^{-1} \circ \widehat{F}_j^{(w)}, \tag{3}$$

where, with $\delta_n > 0$ being a small number chosen by the user,

$$\widehat{F}_j^{(w)}(x) = \begin{cases} \delta_n, & \text{if } \widehat{F}_j^{(w)}(x) \le \delta_n; \\ \widehat{F}_j^{(w)}(x), & \text{if } \delta_n < \widehat{F}_j^{(w)}(x) < 1 - \delta_n; \\ 1 - \delta_n, & \text{if } \widehat{F}_j^{(w)}(x) \ge 1 - \delta_n. \end{cases} \tag{4}$$

Note that both estimators shrink $\widehat{F}_j$ to prevent it from achieving 0 or 1, because $\Phi^{-1}(1) = \infty$ and $\Phi^{-1}(0) = -\infty$. These two intuitive estimators have a long history, and have become standard tools in statistics, biostatistics, education and behavior sciences, among other research fields. For example, the widely used statistical software SAS and SPSS provide built-in functions to perform the normal score transformation. In education and psychology, Glass and Hopkins (1996) discussed in their clas-

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

sical book the application of the normal score transformation in removing skewness and kurtosis. In econometrics, Berkowitz (2001) demonstrated how to construct more powerful tests through normal score transformation in forecast evaluation.

Moreover, these estimators are among the rare classical methods that continue to be applicable in high dimensions without any modification, at least empirically. Many researchers have applied the normal score transformation to high-dimensional data and observed empirical successes. For example, Cai, Li, and Liu (2016) demonstrated the application of the normal score transformation in comparing multiple clinical trial endpoints. In genetics study, Peng et al. (2007) proposed to apply the normal score transformation before using variance-components and regression-based methods to map quantitative trait loci. Other applications in genetics research include Wu et al. (2002), Anokhin, Heath, and Ralano (2003), Dixon et al. (2007), Lambregts-Rommelse et al. (2008), Scuteri et al. (2007), Fan et al. (2013), Wang et al. (2015), and Nansel et al. (2015), among others. The Winsorized estimator plays an important role in various high-dimensional statistics methods, such as the Gaussian copula model (Liu, Lafferty, and Wasserman 2009; Xue and Zou 2012), semiparametric discriminant analysis (Mai and Zou 2015c), principal component analysis (Han and Liu 2014) and sufficient dimension reduction (Mai and Zou 2015b).

However, theoretical supports for these estimators in high dimensions are much weaker. The existing theoretical studies for the NS estimator in the literature typically focus on the fixed $p, n \to \infty$ paradigm. See Klaassen and Wellner (1997), Serfling (2009), and Hoff, Niu, and Wellner (2014) for example. The theoretical properties of the NS estimator in high dimensions are generally unknown. On the other hand, the Winsorized estimator is shown to be consistent when p is larger than n, but p can only grow at a relatively slow rate of n. For example, Liu, Lafferty, and Wasserman (2009) established the consistency of the Winsorized estimator when p grows as a polynomial function of n, while Mai and Zou (2015c) showed the consistency when $\log p = o(n^{1/3-\gamma})$ for any $0 < \gamma < 1/3$. Note that we often hope a method to handle dimensions at the rate as close to $\log p = o(n)$ as possible. The relatively lower dimension that can be handled by the Winsorized estimator has led to the belief that the estimated transformation fundamentally hurt the data analysis. Consequently, many statisticians have spent significant amount of efforts in avoiding the transformation of data in the Gaussian copula model, see the rank-based approach for estimating the graphical model in Liu et al. (2012) and Xue and Zou (2012). Despite their success for Gaussian graphical models, the rank-based approach cannot be easily used for other applications, such as the nearest shrunken centroid classifier and the distance correlation.

Contrary to current beliefs, we show in this article that coordinatewise Gaussianization has a minimal effect on statistical analysis. We conduct a systematic investigation on the theoretical properties and applications of coordinatewise Gaussianization achieved by the NS estimator and the Winsorized estimator. With careful calculation, we show that both estimators are consistent under nearly optimal dimensionality, $\log p = o(n/\log n)$. We further study the implications of our results in several important applications. In many applications, coordinatewise Gaussianization removes moment conditions to deliver strong theoretical results. Our major contributions are listed below:

- We present theoretical results that the two coordinatewise Gaussianization estimators $\widehat{T}_j^{(ns)}$ and $\widehat{T}_j^{(w)}$ uniformly converge to T_j over $j = 1, \dots, p$ when $\log p = o\left(\dfrac{n}{\log n}\right)$. Our result is very general, without any regard to the downstream method. Also, the theoretical studies are far from trivial from the technical aspect. Note that our estimators are composites of the (shrunken) empirical CDF $\widehat{F}_j$ and Φ^{-1}. Although it is known that $\widehat{F}_j$ converges to F_j uniformly at a fast rate, Φ^{-1} amplifies the estimation error in $\widehat{F}_j$. Our proof involves intensive study on the variability and bias of the estimated transformations, which may be useful to other theoretical studies as well.
- We study the statistical properties of several important statistical methods when they are combined with coordinatewise Gaussianization, including the Gaussian copula model, the nearest shrunken centroids classifier and distance correlation screening. The major findings are listed below.

  - For Gaussian copula model, we show that graphical model estimators after coordinatewise Gaussianization enjoy similar theoretical properties of the rank-based estimators, which clarifies a misbelief in the literature. Previous theories only support the use of coordinate Gaussianization when the dimension is much lower than that handled by rank-based estimators.
  - For the nearest shrunken centroid classifier, our theory reveals the fundamental impact of tail behavior on the performance of the classifier. Heavy tails in the input variables negatively impacts the nearest shrunken centroid classifier, while light tails can be helpful. In this case, coordinatewise Gaussianization prior to fitting the nearest shrunken centroid classifier is shown to eliminate such negative impact and hence, improves the classification performance.
  - For the distance correlation application, we propose the Gaussianized distance correlation and its empirical version. It is viewed as the distance correlation after coordinatewise Gaussianization. The Gaussianized distance correlation is invariant under any monotone transformation, a property not shared by the distance correlation. Furthermore, when used for variable screening, Gaussianized distance correlation screening does not require the sub-Gaussian tail assumptions that are necessary for distance correlation screening in order to have the sure screening property. In this sense, coordinatewise Gaussianization improves the robustness of distance correlation screening.

  All the above theoretical findings are supported with empirical experiments as well.
- We clarify the applicability of the normal score transformation in high dimensions with several cautionary examples. Such explanation sheds light on the different influences of coordinatewise Gaussianization on low-dimensional and high-dimensional data analysis. We give

a general recommendation on how to use coordinatewise Gaussianization in statistical learning.

The rest of this article is organized as follows. In Section 2 we present the uniform convergence result for the coordinatewise Gaussianization. Section 3 gives a general guideline for determining whether coordinatewise Gaussianization is appropriate. Section 4 contains the applications of our results to the Gaussian copula model, nearest shrunken centroid classifier and distance correlation screening. Numerical studies are presented in Section 5. We conclude the article with a discussion. For the sake of space, additional simulations and all the technical proofs are relegated to the supplementary materials.

2. Uniform Convergence Rates of Coordinatewise Gaussianization

2.1. The Normal Score Transformation

Throughout the rest of the article, we assume that $\mathbf{X}$ is continuous, because we rarely directly transform discrete variable to a continuous one. We make no further distributional assumption on $\mathbf{X}$. The collection of the population coordinatewise Gaussianization transformations is denoted as $\mathbf{T} = (T_1, \dots, T_p)$, where T_j is defined in (1). Similarly, we let the collection of the normal score estimator be denoted as $\widehat{\mathbf{T}}^{(ns)} = (\widehat{T}_1^{(ns)}, \dots, \widehat{T}_p^{(ns)})$, and the Winsorized estimator be denoted as $\widehat{\mathbf{T}}^{(w)} = (\widehat{T}_1^{(w)}, \dots, \widehat{T}_p^{(w)})$. We use the capital letter C to denote a generic constant that could vary from line to line. Recall that the normal score transformation $\widehat{T}_j^{(ns)}$ is defined in (2). We have the following results for $\widehat{T}_j^{(ns)}$.

Theorem 1. There exists a generic positive constant M that does not depend on n or p such that, for any $\epsilon > 0$, when $M\dfrac{\log n}{\sqrt{n}} < \epsilon$, for each $j = 1, \dots, p$, we have

$$\Pr\Big(\frac{1}{n}\sum_{i=1}^{n} |\widehat{T}_j^{(ns)}(X_j^i) - T_j(X_j^i)| \geq \epsilon\Big) \leq C\exp\Big(-C\frac{n\epsilon^2}{\log n}\Big). \quad (5)$$

Consequently,

$$\Pr\left(\max_j \left\{\frac{1}{n}\sum_{i=1}^{n} |\widehat{T}_j^{(ns)}(X_j^i) - T_j(X_j^i)|\right\} \geq \epsilon\right) \leq Cp\exp\Big(-C\frac{n\epsilon^2}{\log n}\Big). \quad (6)$$

Sketch of proof. Theorem 1 can be shown in the following steps. Denote $a_j = \frac{1}{n}\sum_{i=1}^{n} |\widehat{T}_j^{(ns)}(X_j^i) - T_j(X_j^i)|$ and $T_j^*(x) = T_j(x)1(|T_j(x)| \leq \sqrt{2\log n}) + \mathrm{sign}(T_j(x))\sqrt{2\log n}$. We have $a_j \leq b_j + e_j$, where

$$b_j = \frac{1}{n}\sum_{i=1}^{n} |\widehat{T}_j^{(ns)}(X_j^i) - T_j^*(X_j^i)|, \quad e_j = \frac{1}{n}\sum_{i=1}^{n} |T_j^*(X_j^i) - T_j(X_j^i)|. \quad (7)$$

Therefore, it suffices to provide bounds for b_j and e_j. It can be shown that b_j is a function with bounded differences, so we use McDiarmid's inequality (McDiarmid 1989) to provide a bound on b_j. Meanwhile, e_j is an average of independent variables closely related to the normal distribution. Hence, we use properties of the normal distribution to provide a bound for e_j. Combine the bounds for b_j and e_j and we have (5). Then we use the union bound argument to show (6). □

Theorem 1 is virtually free of assumptions and is thus, widely applicable. Moreover, Theorem 1 provides bounds on the average estimation error of $\widehat{T}_j^{(ns)}$ over all the X_j^i's. Such bounds guarantee that $\widehat{\mathbf{T}}^{(ns)}(\mathbf{X})$ is overall an accurate approximation of $\mathbf{T}(\mathbf{X})$, which can be used to show the consistency of the follow-up analysis. See Section 4 for details. Purely for interpretation purposes, we translate Theorem 1 to an asymptotic result in the following corollary.

Corollary 1. If $\log p = o\left(\dfrac{n}{\log n}\right)$ and $n \to \infty$, we have $\max_{j=1,\dots,p}\left\{\frac{1}{n}\sum_{i=1}^{n} |\widehat{T}_j^{(ns)}(X_j^i) - T_j(X_j^i)|\right\} = o_P(1)$.

Corollary 1 confirms the consistency of the normal score transformation when $\log p = o(\frac{n}{\log n})$. When the normal score transformation is used, the practitioner wishes to treat the computed $\widehat{\mathbf{T}}^{(ns)}(\mathbf{X})$ as the theoretically normal variables $\mathbf{T}(\mathbf{X})$. Theorem 1 and Corollary 1 show that there is a very small difference between the actual data to be used for the downstream statistical method and the theoretically desired data, as long as the dimension does not grow faster than an exponential rate relative to the sample size.

Remark 1. In addition to the transformation in (2), a family of its variants are widely applied as well. For a constant $c \geq 0$, one could also consider the transformation $\widehat{T}_j^c(x) = \Phi^{-1}(\frac{n}{n-2c+1}(\widehat{F}_j(x) - \dfrac{c}{n}))$.

Popular choices of c include $0, 1/3, 3/8, 1/2$ (Van der Waerden 1952; Blom 1958; Tukey 1962; Bliss 1967). When $c = 0$, we recover the transformation in (2). It has been observed in practice that the choice of c does not have a noticeable impact on the analysis (Beasley, Erickson, and Allison 2009). Indeed, we can rigorously prove that all these choices of c have theoretical results similar to those in Theorem 1 and Corollary 1. For simplicity, the readers may focus on $c = 0$ case to understand our results.

2.2. The Winsorized Estimator

Now we turn to the Winsorized estimator. First we need to choose the Winsorization parameter δ_n in (4). This choice can be viewed as a parameter for variance-bias tradeoff. Larger δ_n introduces larger bias, while smaller δ_n inflates the variance. It is important to use a proper δ_n in theory. We consider the choice of $\delta_n = 1/n$, the reason for which will be discussed after we present the theoretical results.

Theorem 2. There exists a generic positive constant M that does not depend on n or p such that, for any $\epsilon > 0$, when $M\dfrac{\log n}{\sqrt{n}} < \epsilon$, for each $j = 1, \dots, p$, we have

$$\Pr\Big(\frac{1}{n}\sum_{i=1}^{n} |\widehat{T}_j^{(w)}(X_j^i) - T_j(X_j^i)| \ge \epsilon\Big) \le C\exp\Big(-C\frac{n\epsilon^2}{\log n}\Big). \tag{8}$$

Consequently,

$$\Pr\left(\max_j \left\{\frac{1}{n}\sum_{i=1}^{n} |\widehat{T}_j^{(w)}(X_j^i) - T_j(X_j^i)|\right\} \ge \epsilon\right) \le Cp\exp\Big(-C\frac{n\epsilon^2}{\log n}\Big). \tag{9}$$

We again rewrite Theorem 2 into asymptotic results purely for interpretation purposes.

Corollary 2. If $\log p = o\left(\dfrac{n}{\log n}\right)$ and $n \to \infty$, we have $\max_{j=1,\dots,p}\left\{\dfrac{1}{n}\sum_{i=1}^{n} |\widehat{T}_j^{(w)}(X_j^i) - T_j(X_j^i)|\right\} = o_P(1)$.

Remark 2. It can be seen that that the Winsorized estimator has the same theoretical properties as the normal score estimator. The Winsorized estimator can also handle dimensionality of $\log p = o\left(\dfrac{n}{\log n}\right)$. We further note that the fast convergence of the Winsorized estimator is closely related to our choice of δ_n. Liu, Lafferty, and Wasserman (2009) considered a smaller δ_n and showed the polynomial rate, while Mai and Zou (2015c) used a larger δ_n and the rate is shown to be $\log p = o(n^{1/3-\gamma})$. Our choice of $\delta_n = 1/n$ has a higher convergence rate than both of them, because it strikes a good balance between the variance and bias. Also, our proof is fundamentally different from Liu, Lafferty, and Wasserman (2009) and Mai and Zou (2015c). These two papers partition the real line into several nonoverlapping line segments $\mathcal{A}_1, \dots, \mathcal{A}_R$. Because $\dfrac{1}{n}\sum_{i=1}^{n} |\widehat{T}_j^{(w)}(X_j^i) - T_j(X_j^i)| \le \frac{1}{n}\sum_{r=1}^{R} \#\{X_j^i \in \mathcal{A}_r\} \sup_{x\in\mathcal{A}_r} |\widehat{T}_j(x) - T_j(x)|$, their proofs reduce to finding bounds for $\#\{X_j^i \in \mathcal{A}_r\}$ and $\sup_{x\in\mathcal{A}_r} |\widehat{T}_j(x) - T_j(x)|$. However, within $\mathcal{A}_r$, likely many $|\widehat{T}_j^{(w)}(X_j^i) - T_j(X_j^i)|$ are much smaller than $\sup_{x\in\mathcal{A}_r} |\widehat{T}_j(x) - T_j(x)|$, and the resulting upper bound may be loose. In contrast, our Corollary 2 is proved by showing that the Winsorized estimator is close to the NS estimator, the properties of which are obtained in Theorem 1. As can be seen in the sketch of proof for Theorem 1, we never consider the supreme of the estimation error over line segments. We instead leverage the stability of the NS estimator (i.e., b_j having bounded difference) to obtain a sharper rate.

Remark 3. Because the normal score estimator and the Winsorized estimator have the same theoretical properties, in what follows we only discuss the application for the normal score estimator for ease of presentation. But all the results hold for the Winsorized estimator as well. We also suppress the superscripts (ns) or (w) to avoid proliferation of notation.

3. Applicability of Coordinatewise Gaussianization

Coordinatewise Gaussianization is only the first step of the data analysis. The end results also depend on the downstream statistical method. Since $\widehat{\mathbf{T}}(\mathbf{X})$ is close to $\mathbf{T}(\mathbf{X})$, one could determine the applicability of the normal score transformation by investigating if the analysis is appropriate on $\mathbf{T}(\mathbf{X})$. If it is easier to analyze $\mathbf{T}(\mathbf{X})$ than $\mathbf{X}$ or the statistical analysis becomes easier with $\mathbf{T}(\mathbf{X})$ than $\mathbf{X}$, then the normal score transformation is helpful and can be applied. In many applications, it is indeed beneficial to perform analysis on $\mathbf{T}(\mathbf{X})$. For example, in the Gaussian copula model, the conditional independence structure of $\mathbf{T}(\mathbf{X})$ can be fully characterized by the precision matrix. It is easier to work on $\mathbf{T}(\mathbf{X})$ than $\mathbf{X}$ without changing the problem. In nearest shrunken centroid classifier, it is much easier to estimate the centroids of $\mathbf{T}(\mathbf{X})$. In distance correlation screening, the problem becomes easier on $\mathbf{T}(\mathbf{X})$. In these cases, the normal score transformation often improves the accuracy, because $\widehat{\mathbf{T}}(\mathbf{X})$ is a very good approximation of $\mathbf{T}(\mathbf{X})$ as justified in Theorem 1. See Section 4 for rigorous establishment of these statements.

However, there are also scenarios where coordinatewise Gaussinization does not help and even causes damages. We discuss two important cases here. First, some methods are invariant under monotone transformations, and yield exactly the same results on $\mathbf{X}$ and $\mathbf{T}(\mathbf{X})$. The large family of tree-based methods are typical examples of this kind (Hastie, Tibshirani, and Friedman 2008). When we build trees, we recursively find points x_j and split X_j into two regions $X_j \le x_j$ and $X_j > x_j$. Apparently, this is equivalent to splitting $T_j(X_j)$ into $T_j(X_j) \le T_j(x_j)$ and $T_j(X_j) > T_j(x_j)$. Thus, we do not gain anything by combining the normal score transformation with tree-based methods. Second, coordinatewise Gaussinization forces all the variables to have the same marginal distribution. Consequently, if a method exploits the difference among marginal characteristics of variables, it should not be combined with coordinatewise Gaussinization. For example, Johnstone and Lu (2009) rank variables by their marginal variance, and only keep the top ranked ones for principal component analysis. Apparently, this approach cannot be combined with coordinatewise Gaussinization, as all variables have variance of 1 afterwards. Another example is the proposal by Jin and Wang (2016) for high-dimensional clustering. Their method assumes that important variables have the mixture normal distribution, while the noise variables are normal. The Kolmogorov–Smirnov test is used to identify the important variables by checking for deviation from normality. This method does not work after coordinatewise Gaussinization because all variables will be normal after transformation and hence, discarded as noise features.

The main point here is that one should not use normal score transformation blindly without thinking about the whole procedure of statistical analysis from the beginning to the end. As mentioned before, many statistical methods do benefit from coordinatewise Gaussianization. We discuss some important examples of them in the next section.

4. Statistical Learning after the Normal Score Transformation

4.1. Unsupervised Learning: The Gaussian Copula Model

4.1.1. Model

Copula models are popular statistical tools for understanding the dependence among variables. By Sklar (1959), for any distribution F on $\mathbb{R}^p$ with marginal distribution functions $F_1, \ldots, F_p$, there exists a unique copula, Co : $\mathbb{R}^p \mapsto \mathbb{R}^p$, such that $F(x_1, \ldots, x_p) = \text{Co}(F_1(x_1), \ldots, F_p(x_p))$. The copula Co is often taken as a summary of the dependence among $\mathbf{X}$. A particularly interesting copula is the Gaussian copula. Define $\Phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ as the multivariate normal CDF with parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. The Gaussian copula model assumes that, for any $(u_1, \ldots, u_p) \in [0,1]^p$, we have $\text{Co}(u_1, \ldots, u_p) = \Phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p))$.

Because copula models focus on the dependence structure, the location and the scale of the distribution are generally irrelevant. Hence, conventionally it is assumed that $\boldsymbol{\mu} = 0$ and the diagonal elements of $\boldsymbol{\Sigma}$ are all equal to 1. The Gaussian copula model can also be viewed as a transformation model. If $\mathbf{X}$ follows the Gaussian copula model, then there exist marginal transformations $\mathbf{G} = (g_1, \ldots, g_p)$ such that

$$\mathbf{G}(\mathbf{X}) = (g_1(X_1), \ldots, g(X_p)) \sim N(0, \boldsymbol{\Sigma}). \tag{10}$$

The nonparametric transformation $\mathbf{G}$ makes the Gaussian copula model more flexible than the normal model, while the parametric distribution of $\mathbf{G}(\mathbf{X})$ often leads to easy estimation and interpretation. Many methods have been proposed under the Gaussian copula model (Klaassen and Wellner 1997; Lin and Jeon 2003; Chen and Fan 2006; Hoff 2007; Liu, Lafferty, and Wasserman 2009; Xue and Zou 2012; Liu et al. 2012; Hoff, Niu, and Wellner 2014; Fan, Xue, and Zou 2015; Mai and Zou 2015b, 2015c, Cai and Zhang 2018).

The Gaussian copula model is closely related to the normal score transformation. It can be shown that, the transformation $\mathbf{G}$ in (10) has to coincide with $\mathbf{T} = (T_1, \ldots, T_p)$, where $T_j = \Phi^{-1} \circ F_j$. Therefore, it is straightforward to estimate the Gaussian copula model in two steps: (i) estimate $\mathbf{G}$ with the normal score transformation and (ii) perform normality-based analysis on the transformed data. Indeed, this was the approach in Klaassen and Wellner (1997), Serfling (2009) and Hoff, Niu, and Wellner (2014) in low-dimensional problems. We emphasize though that the normal score transformation does not require joint normality on its own. Rather, the joint normality is introduced by the Gaussian copula model.

4.1.2. Methods

Suppose that $\mathbf{X}$ follows the Gaussian copula model in (10). In graphical learning, our goal is to identify pairs of (X_j, X_k) that are conditionally independent given all the other variables. Denote $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. It can be shown that $\theta_{jk} = 0$ if and only if X_j, X_k are conditionally independent given all the other variables. Therefore, to recover the conditional independence structure among $\mathbf{X}$, it suffices to construct a sparse estimator for the precision matrix $\boldsymbol{\Theta}$.

In the special case where $\mathbf{T}$ is known, (10) reduces to the Gaussian graphical model on $\mathbf{T}(\mathbf{X})$. Many methods have been proposed for the Gaussian graphical model, including the neighborhood lasso regression (Meinshausen and Bühlmann 2006), the graphical lasso (Friedman, Hastie, and Tibshirani 2008), SPACE (Peng et al. 2009), the neighborhood Dantzig selector (Yuan 2010), constrained ℓ_1-minimization for inverse matrix estimation (CLIME; Cai, Liu, and Luo (2011)), the penalized D-Trace estimator (Zhang and Zou 2014), among others. When $\mathbf{T}$ is unknown, we can first obtain its estimate $\widehat{\mathbf{T}}$, and then apply normality-based methods on $\widehat{\mathbf{T}}(\mathbf{X})$. We demonstrate this approach with the graphical lasso, which is the most popular method for estimating Gaussian graphical model in practice.

We first describe the method with the oracle information about $\mathbf{T}$. For any matrix $\mathbf{V} \in \mathbb{R}^{q_1 \times q_2}$, define $\|\mathbf{V}\|_{\max} = \max_{i,j} |V_{ij}|$, $\|\mathbf{V}\|_\infty = \max_i \sum_{j=1}^{q_2} |V_{ij}|$, and $\|\mathbf{V}\|_1 = \max_j \sum_{i=1}^{q_1} |V_{ij}|$. If $q_1 = q_2 = p$, $\mathbf{V}_k \in \mathbb{R}^{p \times 1}$ is the kth column of $\mathbf{V}$, while $\mathbf{V}_{(k)} \in \mathbb{R}^{(p-1)\times(p-1)}$ is $\mathbf{V}$ excluding the kth row and the kth column. For any $\mathbf{v} \in \mathbb{R}^p$, we denote $\mathbf{v}_{(k)} \in \mathbb{R}^{p-1}$ as $\mathbf{v}$ excluding v_k. Define the oracle covariance estimator that uses the information of $\mathbf{T}$ as $\widehat{\boldsymbol{\Sigma}}^{(\text{o})} = \frac{1}{n} \sum_{i=1}^n \mathbf{T}(\mathbf{X}^i)(\mathbf{T}(\mathbf{X}^i))^{\mathrm{T}}$. We use λ to denote a positive tuning parameter. The oracle graphical lasso is defined as follows:

$$\widehat{\boldsymbol{\Theta}}^{gl.o} = \arg\min_{\boldsymbol{\Theta} \succ 0} \{-\log\det(\boldsymbol{\Theta}) + \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}^{(o)} \boldsymbol{\Theta}) + \lambda \sum_{i \neq j} |\theta_{ij}|\}.$$

Let $\widehat{\boldsymbol{\Sigma}}$ be the normal score estimator of $\boldsymbol{\Sigma}$:

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{T}}(\mathbf{X}^i)(\widehat{\mathbf{T}}(\mathbf{X}^i))^{\mathrm{T}}. \tag{11}$$

Then the normal score estimators replace $\widehat{\boldsymbol{\Sigma}}^{(\text{o})}$ in the oracle estimators with $\widehat{\boldsymbol{\Sigma}}$. The normal score graphical lasso is defined as follows:

$$\widehat{\boldsymbol{\Theta}}^{gl} = \arg\min_{\boldsymbol{\Theta} \succ 0} \{-\log\det(\boldsymbol{\Theta}) + \mathrm{tr}(\widehat{\boldsymbol{\Sigma}} \boldsymbol{\Theta}) + \lambda \sum_{i \neq j} |\theta_{ij}|\}.$$

4.1.3. Theories

Our theories for the Gaussian copula model contain two parts. First, we show that the normal score estimator $\widehat{\boldsymbol{\Sigma}}$ converges to $\boldsymbol{\Sigma}$ in an elementwise manner when p grows at an exponential rate of n. Second, when we combine $\widehat{\boldsymbol{\Sigma}}$ with sparse methods, we obtain consistent estimators of $\boldsymbol{\Theta}$ in ultra-high dimensions.

Theorem 3. Under the Gaussian copula model, there exists generic constants M, ϵ_0 that do not depend on n or p such that, for any $0 < \epsilon < \epsilon_0$, when $M \frac{\log n}{\sqrt{n}} < \epsilon$, we have, for any $j, k = 1, \ldots, p$,

$$\Pr(|\widehat{\sigma}_{jk} - \sigma_{jk}| \geq \epsilon) \leq C \exp(-\frac{Cn\epsilon^2}{\log^2 n}). \tag{12}$$

Consequently, $\Pr(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \geq \epsilon) \leq Cp^2 \exp\left(-\frac{Cn\epsilon^2}{\log^2 n}\right)$.

Theorem 3 indicates that $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} = o_P(1)$ as long as $\log p = o(\frac{n}{\log^2 n})$. Theorem 3 is a key step in showing the consistency of these methods in estimating $\boldsymbol{\Theta}$ later. Now we present the theoretical properties of the normal score estimators for $\boldsymbol{\Theta}$. Define $K_{\boldsymbol{\Sigma}} = \|\boldsymbol{\Sigma}\|_\infty$, $\mathbf{H} = \boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}$ and $K_{\mathbf{H}} = \|\mathbf{H}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty$. Define d as the number of nonzero off-diagonal elements in $\boldsymbol{\Theta}$. We have the following results.

Theorem 4. Assume $\frac{d^2 \log^2 n \log p}{n} \to 0$. If $\|\mathbf{H}_{\mathcal{A}\mathcal{A}^c}(\mathbf{H}_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty < 1 - \kappa$ for $\kappa \in (0,1)$ and $\frac{\log n \sqrt{\log p}}{\sqrt{n}} \ll \lambda <$

$\frac{1}{6(1+\kappa/4)K_{\mathbf{\Sigma}}K_{\mathbf{H}}\max\{1,(1+4/\kappa)K_{\mathbf{\Sigma}}^2K_{\mathbf{H}}\}}\cdot\frac{1}{d}$, then $\|\widehat{\mathbf{\Theta}}^{gl}-\mathbf{\Theta}\|_{\max}=o_P(1)$. The support of $\widehat{\mathbf{\Theta}}^{gl}$ exactly recovers that of $\mathbf{\Theta}$ with probability going to 1.

Remark 4. As discussed in Section 1, previously theoretical studies on the Winsorized estimator for the Gaussian graphical model can only handle dimensions up to polynomial order (Liu, Lafferty, and Wasserman 2009) or $\log p = o(n^{1/3-\gamma})$ for some constant $\gamma \in (0, 1/3)$ (Mai and Zou 2015c). Theorem 4 pushes the dimension limit to $\log p = O(n^{\gamma})$ for some constant $\gamma \in (0, 1)$.

Remark 5. The normal score transformation can be combined with other estimators such as the neighborhood Dantzig selector and the CLIME estimator. The condition $\|\mathbf{H}_{\mathcal{A}\mathcal{A}^c}(\mathbf{H}_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty < 1-\kappa$ is the irrepresentable condition that Ravikumar et al. (2011) used to study the theoretical properties of the oracle graphical lasso. This irrepresentable condition is not needed in the theory if we use the neighborhood Dantzig selector and the CLIME estimator. With Theorem 3, the convergence rates under other matrix norms such as Frobenius norm or matrix ℓ_1 norm can be established similarly by using the same arguments in Xue and Zou (2012).

4.2. Supervised Learning: The Nearest Shrunken Centroids Classifier

4.2.1. Method and Cautionary Remarks

Supervised learning covers all applications in which we need to predict an outcome variable (response). Numerous supervised learning methods have been developed, such as logistic regression, nearest neighborhood, neural networks, boosting, random forest, support vector machines, just to name a few (Hastie, Tibshirani, and Friedman 2008). The normal score transformation may be desirable in supervised learning if one wishes to remove heavy tails in the features. For example, if we want to use a distance-based classifier, such as the nearest shrunken centroids classifier (NSC) to be discussed shortly, we need data to be reasonably light-tailed in order to well estimate their centroids.

Recall that the application of the normal score transformation is not universally beneficial in supervised learning. For example, we discussed in Section 3 that tree-based methods cannot be improved by the normal score transformation because they are invariant under monotone marginal transformations of $\mathbf{X}$. Hence, the application of the normal score transformation in supervised learning should be closely tied to the classifier of interest. An exhaustive study of the normal score transformation in supervised learning is apparently impossible within the scope of this manuscript. Instead, we focus on the nearest shrunken centroids classifier (NSC) (Tibshirani et al. 2002, 2003) as a demonstration for the potential benefit of the normal score transformation in supervised learning.

Consider $\{Y, \mathbf{X}\}$, where $Y \in \{+1, -1\}$ and $\mathbf{X} \in \mathbb{R}^p$. Our goal is to predict Y based on $\mathbf{X}$. Define $\bar{\mu}_j = \frac{1}{n}\sum_{i=1}^n X_j^i$ as the overall centroid for X_j, s_j as the pooled within-class standard deviation, and $\tilde{\mu}_{yj} = \frac{1}{n_y}\sum_{Y^i=y} X_j^i$ as the within-class centroid, where n_y is the sample size within Class y. Then $d_{yj}^* = \frac{\tilde{\mu}_{yj}-\bar{\mu}_j}{m_y s_j + s_0}$ estimates the standardized difference between the within-class centroid and the overall centroid on the j'th predictor, where $m_y = \sqrt{1/n_y - 1/n}$ and $s_0 \geq 0$ is a constant to improve numeric stability in practice. NSC soft-thresholds d_{yj}^* by some user-chosen $\Delta > 0$ to obtain $d_{yj} = \mathrm{sign}(d_{yj}^*)(|d_{yj}^*| - \Delta)_+$. Then the centroids are estimated by the shrunken estimates$\widehat{\mu}_{yj} = \bar{\mu}_j + m_y s_j d_{yj}$. A new observation $\mathbf{X}^{\text{new}}$ is classified to Class +1 if and only if

$$-2\log\frac{n_+}{n_-} + \sum_{j=1}^{p}\frac{(X_j^{\text{new}} - \widehat{\mu}_{+j})^2}{s_j^2} < \sum_{j=1}^{p}\frac{(X_j^{\text{new}} - \widehat{\mu}_{-j})^2}{s_j^2}, \tag{13}$$

where n_y are sample sizes of Class y. Note that if $d_{yj} = 0$, then $\widehat{\mu}_{+j} = \widehat{\mu}_{-j}$, indicating that X_j is excluded from NSC. In order to justify this selection scheme we need a statistical model, to be introduced in Section 4.2.2. Otherwise, NSC selection is not always consistent and hence, can lead to bias selection and classification (Mai, Zou, and Yuan 2012).

Apparently, the accuracy of $\widehat{\boldsymbol{\mu}}_+, \widehat{\boldsymbol{\mu}}_-$ is critical to the variable selection and prediction in NSC. We will see later that the behavior of these estimates greatly depends on the tail conditions on $\mathbf{X}$, and the normal score transformation can be beneficial. To apply the normal score transformation in NSC, we first obtain the transformed data $\widehat{\mathbf{T}}(\mathbf{X})$, and then apply NSC on $(Y, \widehat{\mathbf{T}}(\mathbf{X}))$. We refer to this method as NS-NSC. Based on the pseudo dataset, we obtain $\widehat{\boldsymbol{\eta}}_y = (\widehat{\eta}_{y1}, \ldots, \widehat{\eta}_{yp})$ as the shrunken centroid of class y. In the transformed space, a feature X_j is important if and only if $\widehat{\eta}_{+j} \neq \widehat{\eta}_{-j}$.

4.2.2. Theories

To study the statistical properties of NSC and its combination with the normal score transformation, we consider the following *invariant contrast in mean* or *invariant contrast in median* model. We assume that $\Pr(Y = y) = \pi_y \in (0, 1)$, and within Class $Y = y$,

$$X_j = \mu_{yj} + \epsilon_j, \tag{14}$$

where either ϵ_j has mean 0 when $E\epsilon_j$ exists or the distribution of ϵ_j is symmetric about 0 when $E\epsilon_j$ does not exist. So μ_{yj} is interpreted as either the conditional mean or the conditional median. We use the acronym *ICIM* to name the model. The model in (14) underlines the application of NSC in that the distribution of X_j only differs in the mean across classes (or median when the mean does not exist). We further assume that ϵ_j are independent, because NSC may produce inconsistent variable selection results when ϵ_j are dependent (Mai, Zou, and Yuan 2012). In order to show that the normal score transformation can help NSC, it only makes sense to have a theoretical setup where NSC can be a good classifier in principle. Otherwise, the comparison is meaningless. In the theoretical study of NSC and NS-NSC we set $s_0 = 0$ because there is no numerical instability issue in the theoretical analysis.

Under the ICIM model, observations can be classified based on their distances to the centroids. It is easy to show that only the variables in $\mathcal{D}$ are important for classification, where $\mathcal{D} = \{j : \mu_{+j} \neq \mu_{-j}\}$. Denote $\eta_{yj} = E(T_j(X_j) \mid Y = y)$. Note that η_{yj}

is always finite because $T_j(X_j)$ is marginally sub-Gaussian. We have the following invariance result.

Lemma 1. Define $\mathcal{D}' = \{j : \eta_{+j} \neq \eta_{-j}\}$. Then we must have that $\mathcal{D}' = \mathcal{D}$.

Lemma 1 indicates that $\mathcal{D}$ is invariant under the transformation of $\mathbf{T}$. This is why we refer to it as the invariant model. The invariance guarantees that the target set $\mathcal{D}$ remains identical after transformation and hence, on the population level we can apply the normal score transformation. In other words, coordinatewise Gaussianization does not change the problem. Nevertheless, a successful recovery of $\mathcal{D}$ in practice depends on the accurate estimation of $\boldsymbol{\mu}_y$ or $\boldsymbol{\eta}_y$. The consistent estimation of μ_{yj} typically requires tail conditions on ϵ_j, while when data are heavy-tailed, it is much easier to estimate $\boldsymbol{\eta}_y$ with the normal score transformation. We discuss this point in detail in the next section. First, we present an example to show that NSC can completely fail while NS-NSC succeeds, which highlights the importance of the tail behavior of ϵ_j critically to the performance of the original NSC (without any data transformation). Recall that, NSC selects the set $\widehat{\mathcal{D}} = \{j : \widehat{\mu}_{+j} \neq \widehat{\mu}_{-j}\}$.

Lemma 2. If ϵ_j are standard Cauchy random variables in the ICIM model, then for any threshold Δ, we have $\Pr(\widehat{\mathcal{D}} = \mathcal{D}) \to 0$ as long as $d \to \infty, p - d \to \infty$, where $d = |\mathcal{D}|$.

Lemma 2 shows that, if ϵ_j follows the Cauchy distribution, it is impossible to recover $\mathcal{D}$ by applying NSC on $\mathbf{X}$ directly. In the next theorem, we further consider the variable selection consistency of NSC under two different tail conditions on ϵ_j. For simplicity, we assume that the data have been standardized such that $\bar{\mu}_j = 0, s_j = 1$ for all j. Define the minimum signal strength $\delta > 0$ such that $\min_{j\in\mathcal{D}, y=\pm1}\{|\mu_{yj}|, |\eta_{yj}|\} \geq \delta$. Also recall that $\pi_y = \Pr(Y = y)$.

Theorem 5. Assume that there exists a constant $C_\pi > 0$ such that $\pi_y \geq C_\pi$. For any $0 < \rho \leq 1/2$, if $\rho\sqrt{n}\delta \leq \Delta \leq (1 - \rho)\sqrt{n}\delta$, we have that

1. if there exists a positive integer k and a constant M such that $E|\epsilon_j|^l \leq M^l$ for all j and $l \leq 2k$,
$$\Pr(\widehat{\mathcal{D}} = \mathcal{D}) \geq 1 - \frac{Cpk^{2k}M^{2k}}{(\sqrt{n}\rho\delta)^{2k}} - C\exp(-Cn). \qquad (15)$$
Consequently, $\Pr(\widehat{\mathcal{D}} = \mathcal{D}) \to 1$ if $\delta \gg \dfrac{kp^{1/(2k)}}{n^{1/2}}$.
2. if there exists $\sigma^2 > 0$ such that $E\exp(t\epsilon_j) \leq \exp(\sigma^2 t^2)$ for all $t > 0$ and $j = 1, \ldots, p$, then
$$\Pr(\widehat{\mathcal{D}} = \mathcal{D}) \geq 1 - Cp\exp(-Cn\rho^2\delta^2) - C\exp(-Cn). \qquad (16)$$
Consequently, $\Pr(\widehat{\mathcal{D}} = \mathcal{D}) \to 1$ if $\delta \gg \dfrac{\sqrt{\log p}}{\sqrt{n}}$.

Theorem 5 reveals the effect of the tail behaviors on NSC. When the predictors are sub-Gaussian, NSC can consistently select all the important predictors even when $\log p = o(n)$ under mild regularity conditions. However, when the predictors only have finite kth moments, we are only guaranteed to achieve variable selection consistency when p grows at a polynomial rate of the sample size n. Hence, when data are not sub-Gaussian, the applicability of NSC in high dimensions is limited. This is where the normal score transformation can provide a great lift. We can show that, without imposing any tail condition on ϵ_js, NS-NSC can consistently recover $\mathcal{D}$ with an overwhelming probability. Specifically, write $\widehat{\eta}^{(o)}_{yj}$ as the estimated centroid of $T_j(X_j)$ within Class y given by NSC, and $\widehat{\eta}_{yj}$ as that of $\widehat{T}_j(X_j)$, where $\widehat{T}_j$ is the normal score transformation defined in (2). The estimate $\widehat{\eta}^{(o)}_{yj}$ uses oracle information about T_j, and is hence, only a baseline for theoretical studies. Define the selected sets by the oracle and the normal score estimator as $\widehat{\mathcal{D}}^{(o)}$ and $\widehat{\mathcal{D}}$, respectively, where

$$\widehat{\mathcal{D}}^{(o)} = \{j : \widehat{\eta}^{(o)}_{+j} \neq \widehat{\eta}^{(o)}_{-j}\}, \quad \widehat{\mathcal{D}} = \{j : \widehat{\eta}_{+j} \neq \widehat{\eta}_{-j}\}. \qquad (17)$$

Theorem 6. Assume that there exists a constant $C_\pi > 0$ such that $\pi_y \geq C_\pi$. For any $0 < \rho \leq 1/2$, if $\rho\sqrt{n}\delta \leq \Delta \leq (1 - \rho)\sqrt{n}\delta$, we have that

1. for the oracle estimate $\widehat{\mathcal{D}}^{(o)}$, $\Pr(\widehat{\mathcal{D}}^{(o)} = \mathcal{D}) \geq 1 - Cp\exp(-Cn\rho^2\delta^2)$. Consequently, $\Pr(\widehat{\mathcal{D}}^{(o)} = \mathcal{D}) \to 1$ if $\delta \gg \dfrac{\sqrt{\log p}}{\sqrt{n}}$.
2. for the normal score estimate $\widehat{\mathcal{D}}$, $\Pr(\widehat{\mathcal{D}} = \mathcal{D}) \geq 1 - Cp\exp(-Cn\rho^2\delta^2/\log n)$. Consequently, $\Pr(\widehat{\mathcal{D}}^{(o)} = \mathcal{D}) \to 1$ if $\delta \gg \dfrac{\sqrt{\log p \log n}}{\sqrt{n}}$.

To see how the influence of the transformations, again consider the case where δ does not change with (n, p). Theorem 6 shows that, when we know the transformation, NS-NSC is consistent if $\log p = o(n)$, while if we estimate the transformation, NS-NSC is consistent if $\log p = o\left(\dfrac{n}{\log n}\right)$. Hence, NS-NSC is almost optimal up to a factor of $\log n$. Moreover, Theorem 6 requires no tail condition on ϵ_j, indicating that NS-NSC is potentially better than NSC on heavy-tailed data. Recall that Lemma 1 shows that NSC fails when the error is Cauchy, but NS-NSC can still perform well according to Theorem 6.

Theorems 5 and 6 are derived under independence assumption on ϵ_j. This independence assumption is imposed because NSC is indifferent to the correlation structure. It sums up the squared Euclidean distance at each coordinate. The NS transformation modifies the way we evaluate the coordinatewise Euclidean distance, but still utilizes the total Euclidean distance for classification, which, after all, is the core of NSC. Empirically, NS-NSC still works well when a reasonable amount of correlation exists; see Models N5 and N6 in Section 5.1.2 and the real data analysis in Section 5.2. For theoretical considerations, if strong correlation is present, NSC and NS-NSC will still consistently select the set $\mathcal{D}$ under respective conditions, but $\mathcal{D}$ may not be the best set for classification. For example, Cai and Liu (2011), Fan, Feng, and Tong (2012) and Mai, Zou, and Yuan (2012) showed that when $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$, that is, under the linear discriminant analysis (LDA) model, $\mathcal{D}$ could lead to inferior classification depending on the interplay between $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_k$. One way to resolve the issue of correlation is to combine the NS transformation with a method, such as LDA, that models the correlation. Similar ideas have been explored by Lin and Jeon (2003) and Mai and Zou (2015c), but our results on the NS transformation could potentially lead to a classifier with better statistical properties.

4.3. Variable Screening: Distance Correlation Screening

4.3.1. Model and Cautionary Remarks

Variable screening contains a large family of computationally efficient methods for variable selection in high dimensions. Consider the response $\mathbf{Y} \in \mathbb{R}^q$ and the predictors $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$, where $\mathbf{X}_k \in \mathbb{R}^{p_k}$ for $p_k \geq 1$ could either be univariate or multivariate. In high-dimensional data where p is very large, we are interested in identifying

$$\mathcal{D} = \{k : F(\mathbf{y} \mid \mathbf{X}_k) \text{ functionally depends on } \mathbf{X}_k \text{ for some } \mathbf{y}\}, \tag{18}$$

where $F(\mathbf{y} \mid \mathbf{X}_k)$ is the conditional CDF of $\mathbf{Y}$ given $\mathbf{X}_k$. It is often assumed that $|\mathcal{D}|$ is much smaller than p. Screening methods aim to detect a set $\mathcal{S}$ such that $\mathcal{D} \subset \mathcal{S}$. Once screening methods find an estimate of the set $\mathcal{S}$, denoted as $\hat{\mathcal{S}}$, refined analysis will be applied to the much smaller data matrix, $\mathbf{X}_{\hat{\mathbf{S}}}$. This two-stage approach is shown to produce good results.

Since the second-stage analysis is only applied to $\mathbf{X}_{\hat{\mathcal{S}}}$, it is crucial that the first-stage screening should give an $\hat{\mathcal{S}}$ that includes all the predictors in $\mathcal{D}$. A screening method that succeeds in this is said to enjoy the SURE screening property. Fan and Lv (2008) first proposed the marginal Pearson correlation screening for the linear model, and later researchers have proposed numerous methods that can handle more complicated models (Fan and Song 2010; Fan, Feng, and Song 2011; Zhu et al. 2011; Li, Zhong, and Zhu 2012; Chang, Tang, and Wu 2013; Mai and Zou 2013, 2015a; Cui, Li, and Zhong 2015; Chang, Tang, and Wu 2016, among others).

Similar to supervised learning, the applicability of coordinatewise Gaussianization depends on the specific screening method we employ. Many screening methods require variables to be sub-Gaussian; see Fan and Lv (2008) and Li, Zhong, and Zhu (2012) for example. Such methods can be combined with coordinatewise Gaussianization, as variables become sub-Gaussian afterwards. However, there are also several screening methods that are invariant under marginal monotone transformations. For example, Li et al. (2012) proposed using Kendall's τ correlation to perform screening under a semiparametric single-index model with a monotone link function. Since Kendall's τ remains the same whether we transform the variables or not, coordinatewise Gaussianization gains nothing when combined with this method. Similarly, the (fused) Kolmogorov filter (Mai and Zou 2013, 2015a) is invariant under variable transformation, and there is no need to perform coordinatewise Gaussianization.

In what follows, we focus on the impact of the coordinatewise Gaussianization when it is applicable. We choose to consider its combination with distance correlation screening (DCS, Li, Zhong, and Zhu (2012)). DCS is a well-known and successful screening method with the SURE screening property in the model-free context. Importantly, its nice theoretical properties hinge on moment conditions. The empirical results in Mai and Zou (2015a) suggest that the performance of DCS can be poor without sub-Gaussian assumptions. A direct remedy is to apply coordinatewise Gaussianization prior to computing distance correlation. The resulting correlation is named Gaussianized distance correlation. In what follows, we first briefly review the distance correlation for the sake of completeness.

4.3.2. Method

We start with the definition of distance correlation (Székely, Rizzo, and Bakirov 2007). For a complex-valued function f, define $|f|^2 = f \cdot \bar{f}$, where $\bar{f}$ is the complex conjugate of f. For a vector $\mathbf{t}$, $\|\mathbf{t}\|$ is its Euclidean norm. The distance covariance between $\mathbf{X}_k$ and $\mathbf{Y}$ is written as $\mathrm{dcov}(\mathbf{X}_k, \mathbf{Y})$ with

$$\mathrm{dcov}^2(\mathbf{X}_k, \mathbf{Y}) = \int_{R^{p_k+q}} |\psi_{\mathbf{X}_k,\mathbf{Y}}(\mathbf{t}, \mathbf{s}) - \psi_{\mathbf{X}_k}(\mathbf{t})\psi_{\mathbf{Y}}(\mathbf{s})|^2 w(\mathbf{t}, \mathbf{s}) d\mathbf{t} d\mathbf{s}, \tag{19}$$

where $\psi_{\mathbf{X}_k,\mathbf{Y}}$ is the joint characteristic function of $(\mathbf{X}_k, \mathbf{Y})$, $\psi_{\mathbf{X}_k}$ is the characteristic function of $\mathbf{X}_k$, $\psi_{\mathbf{Y}}$ is the characteristic function of $\mathbf{Y}$, $\mathbf{t} \in \mathbb{R}^{p_k}$, $\mathbf{s} \in \mathbb{R}^q$ and $w(\mathbf{t}, \mathbf{s})$ is a weight function. With a random sample $\{\mathbf{X}^i, \mathbf{Y}^i\}_{i=1}^n$, we can find the empirical distance covariance between $\mathbf{X}_k$ and $\mathbf{Y}$, denoted as $\widehat{\mathrm{dcov}^2}(\mathbf{X}_k, \mathbf{Y})$, as in Székely, Rizzo, and Bakirov (2007). Then the empirical distance correlation between $\mathbf{X}_k$ and $\mathbf{Y}$ is $\widehat{\mathrm{dcor}}(\mathbf{X}_k, \mathbf{Y}) = \frac{\widehat{\mathrm{dcov}}(\mathbf{X}_k, \mathbf{Y})}{\sqrt{\widehat{\mathrm{dcov}}(\mathbf{X}_k, \mathbf{X}_k)\widehat{\mathrm{dcov}}(\mathbf{Y}, \mathbf{Y})}}$. The distance correlation screening (DCS) is applied as follows. For each predictor $\mathbf{X}_k$, we compute $\hat{\omega}_k^{\mathrm{DC}} = \widehat{\mathrm{dcor}}(\mathbf{X}_k, \mathbf{Y})$. Then we keep the predictors with large $\hat{\omega}_k^{\mathrm{DC}}$. Note that DCS is a model-free screening method. However, the success of DCS apparently depends on whether $\hat{\omega}_k^{\mathrm{DC}}$ accurately approximates ω_k^{DC}. Li, Zhong, and Zhu (2012) assumed that $\mathbf{X}_k$ and $\mathbf{Y}$ are sub-Gaussian to establish estimation consistency when p grows at an exponential rate of n. However, when data are heavy-tailed, there is no guarantee that DCS enjoys the SURE screening property. See the simulation results in Section 5.1.3. To resolve this issue, one could apply coordinatewise Gaussianization to remove the heavy tails. We first find either the normal score estimator or the Winsorized estimator $\widehat{\mathbf{T}}$. Then we compute

$$\hat{\omega}_k = \widehat{\mathrm{dcor}}(\widehat{T}_{X_k}(\mathbf{X}_k), \widehat{T}_Y(\mathbf{Y})). \tag{20}$$

Clearly, $\hat{\omega}_k$ is the empirical version of the Gaussianized distance correlation (GDC) defined as

$$\omega_k = \mathrm{dcor}(T_{X_k}(\mathbf{X}_k), T_Y(\mathbf{Y})). \tag{21}$$

Like DC, zero GDC implies independence. Unlike DC, GDC is invariant under monotone transformations. Moreover, we can easily derive the following lemma:

Lemma 3. For ω_k defined in (21), we have (i) $0 \leq \omega_k \leq 1$; and(ii) when $(\mathbf{X}_k, \mathbf{Y})$ is bivariate normal with Pearson correlation ρ_k, ω_k is a strictly monotone function of $|\rho_k|$.

The predictors with large $\hat{\omega}_k$'s are regarded as important, while those with small $\hat{\omega}_k$'s are regarded as unimportant. More specifically, for a threshold γ_n, the kept subset is $\hat{\mathcal{S}}(\gamma_n) = \{k : \hat{\omega}_k > \gamma_n\}$. Alternatively, we can pick the d_nth largest value of $\hat{\omega}_k$, where d_n is a predefined positive integer. Hence, for a predefined d_n, the procedure reserves the subset

$$\hat{\mathcal{S}}(d_n) = \{k : \hat{\omega}_k \text{ is among the } d_n\text{th largest of } \hat{\omega}_j, j = 1, \ldots, p\}.$$

We refer to the above procedure as the *Gaussianized distance correlation screening* (GDCS).

4.3.3. Theories

We assume that the dimension of $\mathbf{Y}$, q, is fixed, and the dimension for each predictor $\mathbf{X}_k$, p_k, are uniformly bounded above, but the number of predictors, p, is allowed to diverge with n.

Theorem 7. For some constants $M > 0$, $0 \le \kappa < 1/2$ and any $0 < \gamma < \frac{1}{2} - \kappa$, we have

$$\Pr(|\widehat{\mathrm{dcor}}(\hat{T}_{X_k}(\mathbf{X}_k), \hat{T}_Y(\mathbf{Y})) - \mathrm{dcor}(T_{X_k}(\mathbf{X}_k), T_Y(\mathbf{Y}))| \ge Mn^{-\kappa}) \le C[\exp(-Cn^{1-2(\kappa+\gamma)}) + n\exp(-Cn^{\gamma})]$$

for any $k = 1, \ldots, p$. Consequently,

$$\Pr(\sup_{k=1,\ldots,p} |\hat{\omega}_k - \omega_k| \ge Mn^{-\kappa}) \le Cp[\exp(-Cn^{1-2(\kappa+\gamma)}) + n\exp(-Cn^{\gamma})].$$

Theorem 7 implies that $\hat{\omega}_k$'s uniformly converge to their population counterparts in ultra-high dimensions without any condition on the distribution of $\mathbf{X}$ or $\mathbf{Y}$. We further compare the results in Theorem 7 with the convergence rate of DCS without Gaussianization. Li, Zhong, and Zhu (2012) considered the following condition:

(C1) There exists a positive constant s_0 such that for any $0 < s < s_0$, we have

$$\sup_p \max_{1\le k\le p} E(\exp(s\|\mathbf{X}_k\|^2)) < \infty, \text{ and } E(\exp(s\|\mathbf{Y}\|^2)) < \infty. \tag{22}$$

Proposition 1. Under Condition (C1), for some constants $M > 0$, $0 \le \kappa < 1/2$ and any $0 < \gamma < 1/2 - \kappa$, we have

$$\Pr(\sup_{k=1,\ldots,p} |\hat{\omega}_k^{\mathrm{DC}} - \omega_k^{\mathrm{DC}}| \ge Mn^{-\kappa}) \le O(p[\exp(-Cn^{1-2(\kappa+\gamma)}) + n\exp(-Cn^{\gamma})]). \tag{23}$$

It is easy to see that GDCS and DCS has the same theoretical properties. However, DCS requires the additional moment condition in Condition (C1) to achieve such results. When data are heavy-tailed and (C1) does not hold, there is no longer any theoretical guarantee for DCS. In contrast, with coordinatewise Gaussianization, GDCS does not rely on any moment conditions.

Now we show the SURE screening properties of GDCS when we fix d_n. For a generic set $\mathcal{A}$, denote $\Delta_{\mathcal{A}} = \min_{k\in\mathcal{A}} \omega_k - \max_{k\in\mathcal{A}^C} \omega_k$. We consider the following condition:

(C2) There exists $\mathcal{S}$ and $c > 0$, $0 \le \kappa < 1/2$, such that $\mathcal{D} \subset \mathcal{S}$ and $\Delta_{\mathcal{S}} > cn^{-\kappa}$.

Theorem 8. Define $\hat{\mathcal{D}}(d_n) = \{k : \hat{\omega}_k \text{ is among the } d_n\text{'th largest}\}$. Under Condition (C2), for any $d_n > |\mathcal{S}|$, $0 < \gamma < 1/2 - \kappa$, we have

$$\Pr(\mathcal{D} \subset \hat{\mathcal{D}}(d_n)) \ge 1 - Cp[\exp(-Cn^{1-2(\kappa+\gamma)}) + n\exp(-Cn^{\gamma})].$$

Condition (C2) is very similar to the common assumption in screening literature that assumes a gap between the marginal signals of true predictors and those of noise predictors (Mai and Zou 2015a). Note that GDCS does not require Condition (C1) to enjoy the sure screening property. In the literature, popular choices of d_n are n or $c[n/\log n]$, $c = 1$ or 2.

As suggested by referees, we compare the theoretical properties for GDCS with the Uniform-transformed distance correlation screening (UDCS; Székely and Rizzo 2009; Zhong et al. 2016; Chen, Chen, and Wang 2018). UDCS first transforms the variables by the empirical CDFs so that they are roughly uniformly distributed, and then computes the distance correlation on the transformed data. The uniform transformation serves a similar purpose to the Gaussianization. It removes moment conditions and improves robustness against heavy tails. As a result, UDCS has the same SURE screening property as GDCS.[1]

However, there are some noticeable differences between UDCS and GDCS. For example, in Section 5.1.3 we test the performance of UDCS and NS-DCS under a wide range of models, and GDCS outperforms UDCS empirically. On the other hand, as shown in Lemma 3, the Gaussianized distance correlation is a monotone function of the Pearson correlation when data is jointly normal. This is a desirable property as the Pearson correlation is a well accepted dependence measurement for normal data. However, it is unclear if UDCS can be interpreted in the same way for normal data.

5. Numerical Studies

5.1. Simulations

5.1.1. The Gaussian Copula Model

We illustrate our results for semiparametric graphical lasso by numerical studies. Following Xue and Zou (2012), in all the simulated studies we let $p = 100$, $n = 300$. Consider the model $\mathbf{T}(\mathbf{X}) \sim N(0, \mathbf{\Sigma})$. We are interested in $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$. We first generated $\mathbf{V} \sim N(0, \mathbf{\Sigma})$ and then transformed $\mathbf{X} = \mathbf{T}^{-1}(\mathbf{V})$, where $\mathbf{T}^{-1} = (g_1, \ldots, g_5, g_1, \ldots, g_5, \ldots)$. The selection of g_i's are as follows:

$$g_1(x) = x, g_2(x) = \mathrm{sign}(x)\sqrt{|x|}, g_3(x) = \Phi(x), g_4(x) = x^3, g_5(x) = \exp(x) \tag{24}$$

Model G1: $\theta_{ii} = 1$ for $i = 1, \ldots, p$, $\theta_{i,i+1} = \theta_{i+1,i} = 0.5$ for $i = 1, \ldots, 3$ and $\theta_{ij} = 0$ otherwise.

Model G2: $\theta_{ii} = 1$ for $i = 1, \ldots, p$, $\theta_{2i,2i-1} = \theta_{2i-1,2i} = 0.5$ for $i = 1, \ldots, 4$ and $\theta_{ij} = 0$ otherwise.

We estimate these two models with the following methods for comparison: (i) Oracle: The oracle graphical lasso with oracle information on $\mathbf{T}$; (ii) Raw: Graphical lasso on the raw data; (iii) Normal score (NS): The normal score estimator; (iv) Winsorized: The Winsorized estimator with $\delta_n = \frac{1}{n}$. Our results are based on 500 replicates. The average receiver operating characteristic (ROC) curves are plotted in Figure 1. We only plot the results for the NS estimator and the Raw estimator, because the Oracle estimator and the Winsorized estimator are almost

[1] To be rigorous, Zhong et al. (2016) proposed to only transform the response Y, and thus, moment conditions are still imposed on $\mathbf{X}$. But such moment conditions should be easy to avoid if $\mathbf{X}$ is transformed as well.

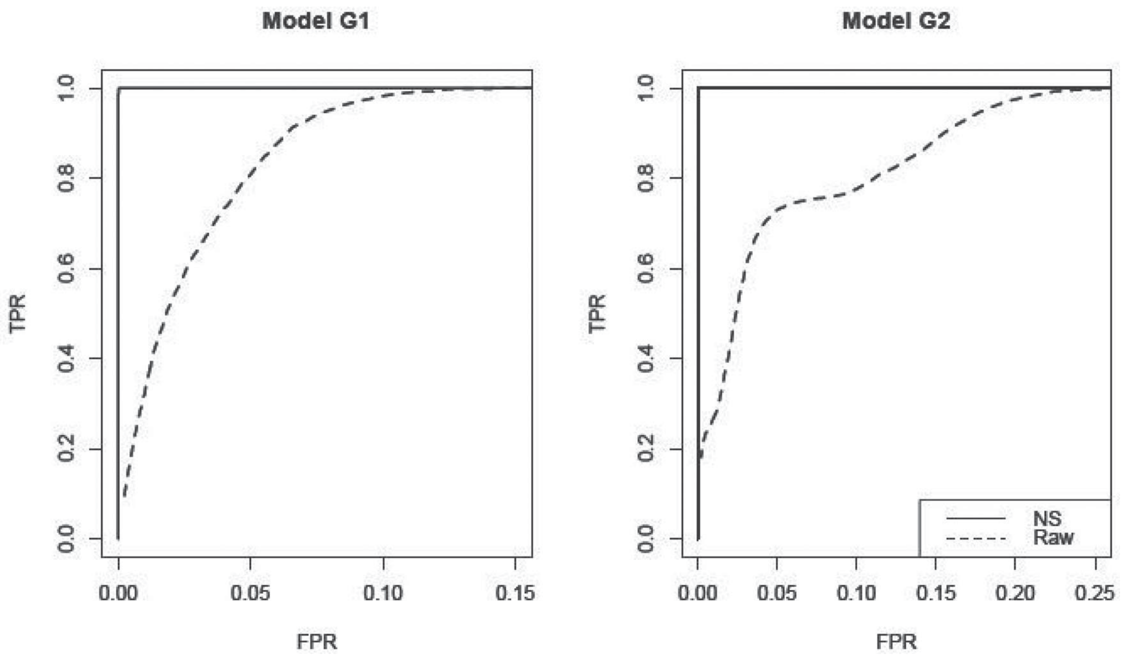

Figure 1. Receiver operating characteristic (ROC) curves for Models G1 & G2; x-axis: false positive rate (FPR), y-axis: true positive rate (TPR). The Oracle estimator and the Winsorized estimator are omitted because they are almost identical to the NS estimator.

identical to the NS estimator. The Raw method, on the other hand, is notably different from the other three estimators. Such results are expected based on our theoretical studies.

5.1.2. Nearest Shrunken Centroids Classifier

We present simulation results for normal score transformation in nearest shrunken centroids classifier. In all the simulations, we set $n = 200, p = 2000, d = 10$, where d is the number of important predictors. We considered the six models, the first five of which take the form that $\mathbf{X} = \boldsymbol{\mu}_y + \boldsymbol{\epsilon}$ if $Y = y$. Without loss of generality, we fix $\boldsymbol{\mu}_1 = 0$ in all models.

Model N1: $\boldsymbol{\epsilon} \sim N(0, \mathbf{I})$ and $\boldsymbol{\mu}_2 \propto (\mathbf{1}_d, \mathbf{0}_{p-d})$. The scale of $\boldsymbol{\mu}_2$ is chosen such that the Bayes error is 10%.

Model N2: Each entry ϵ_j is an independent t_3 random variable and $\boldsymbol{\mu}_2 = (\mathbf{1}_d, \mathbf{0}_{p-d})$.

Model N3: Each entry ϵ_j is an independent standard Cauchy random variable and $\boldsymbol{\mu}_2 = 2 \cdot (\mathbf{1}_d, \mathbf{0}_{p-d})$.

Model N4: Each entry ϵ_j is an independent Weibull random variable with both the shape and scale parameters equal to 1, and $\boldsymbol{\mu}_2 = 0.5 \cdot (\mathbf{1}_d, \mathbf{0}_{p-d})$.

Model N5: $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = AR(0.5)$, and $\boldsymbol{\mu}_2 \propto (\mathbf{1}_d, \mathbf{0}_{p-d})$. The scale of $\boldsymbol{\mu}_2$ is chosen such that the Bayes error is 10%.

Model N6: For $\mathcal{D} = \{1, \dots, d\}$, $\mathbf{V}_\mathcal{D} \mid Y = y \sim N(\boldsymbol{\mu}_{\mathcal{D},y}, \boldsymbol{\Sigma}_\mathcal{D})$, where $\boldsymbol{\Sigma}_\mathcal{D} = AR(0.5)$, $\boldsymbol{\mu}_{\mathcal{D},1} = 0, \boldsymbol{\mu}_{\mathcal{D},2} \propto \mathbf{1}_d$. The scale of $\boldsymbol{\mu}_2$ is chosen such that the Bayes error is 10%. Then $V_{\mathcal{D}^C} \perp V_\mathcal{D}$. For all the odd numbers $j \in \mathcal{D}^C$, $V_j \sim N(0, 1)$ independently; for all the even numbers $j \in \mathcal{D}^C$, $V_j = \delta_j V_{j-1}$, where δ_j is an independent Bernoulli random variable with the parameter 0.5. $\mathbf{X} = \exp(2\mathbf{V})$.

We considered nearest shrunken centroids classifier on $\mathbf{T}(\mathbf{X}), \mathbf{X}, \widehat{\mathbf{T}}(\mathbf{X})$ and the Winsorized data $\widehat{\mathbf{T}}^{(w)}(\mathbf{X})$, yielding the oracle-NSC, raw-NSC, NS-NSC and Wins-NSC, respectively. Tuning parameters are chosen to minimize the 10-fold cross-validation classification errors. The testing classification errors, the false negatives and the false positives are presented in Table 1. Across all models, NS-NSC have classification errors close to the oracle method, which supports Theorems 1 and 6. In Models N1 & N5, all the predictors are sub-Gaussian, and all methods perform similarly. However, in Models N2, N3 & N6, the predictors are heavy-tailed. Raw-NSC has high classification errors because the centroids cannot be estimated accurately. In Model N4, the predictors are heavily skewed, and NS-NSC and Wins-NSC outperform raw-NSC in prediction and variable selection. Also note that Models N5 & N6 are not ICIM models, but coordinatewise Gaussianization is still beneficial when data are heavy-tailed in Model N6.

5.1.3. Distance Correlation Screening

We denote GDCS via normal score estimator as NS-DCS, and GDCS via Winsorized estimator as Wins-DCS. We compare them with the original DCS (Li, Zhong, and Zhu 2012) and UDCS that combines the uniform transformation with DCS. We also compare our methods with many additional screening methods, but these results are relegated to the supplementary materials for the sake of space.

We repeat each experiment 200 times and assess the performance under the following criteria adopted by Li, Zhong, and Zhu (2012). (1). $\mathcal{M}$: the minimum model size to include all the true variables. We report the 5%, 25%, 50%, 75%, and 95% quantiles of $\mathcal{M}$ out of 200 replications. (2). $\mathcal{P}_a$: the proportion that all true variables are selected for a given model size d in the 200 replications.

We present the simulation results of $\mathcal{P}_a$ with $d = 2[n/\log n]$. We also tried $d = [n/\log n]$ with quite similar outcomes and hence, omit such results here for the sake of space. The simulation results for Model D1–D6 are summarized in Table 2. The response in this section is univariate unless otherwise specified.

Table 1. Simulation results for nearest shrunken centroids classifier.

	Oracle		Raw		NS-NSC		Wins-NSC	
				Model N1				
Error(%)	12.9	(0.13)	12.6	(0.11)	13.0	(0.10)	13.2	(0.13)
FN	0	(0)	0	(0)	0	(0)	0	(0)
FP	19	(1.3)	19	(1.5)	21	(1.7)	19	(1.5)
				Model N2				
Error(%)	17.1	(0.11)	21.9	(0.17)	16.8	(0.10)	17.0	(0.11)
FN	0	(0.0)	0	(0)	0	(0)	0	(0)
FP	22	(1.5)	23	(1.2)	22.5	(1.4)	23	(1.2)
				Model N3				
Error(%)	13.6	(0.08)	49.6	(0.06)	13.6	(0.11)	13.8	(0.11)
FN	0	(0)	7	(0.5)	0	(0)	0	(0)
FP	23.5	(2.1)	384.5	(42.1)	25	(2.1)	24	(2.1)
				Model N4				
Error(%)	15.9	(0.14)	31.6	(0.22)	15.4	(0.12)	15.7	(0.14)
FN	0	(0)	1	(0)	0	(0)	0	(0)
FP	24	(1.8)	48.5	(4.1)	24	(1.2)	24	(1.2)
				Model N5				
Error(%)	12.3	(0.03)	12.2	(0.03)	12.3	(0.03)	12.3	(0.03)
FN	0	(0)	0	(0)	0	(0)	0	(0)
FP	18	(1.9)	21	(3.8)	13	(2.2)	17	(2.7)
				Model N6				
Error(%)	11.1	(0.02)	25.7	(0.23)	11.3	(0.04)	11.3	(0.03)
FN	0	(0)	0	(0.1)	0	(0)	0	(0)
FP	13	(2.8)	7.5	(2.5)	14	(2.3)	15	(1.9)

NOTE: All the numbers are medians based on 500 replicates. The standard errors are in parentheses.

Note that during the computation of GDCS in Model D5, the categorical response is not transformed.

Model D1: This example is adopted from Li, Zhong, and Zhu (2012). We generate $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ from multivariate normal distribution with zero mean and covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$, and the error term ϵ from standard normal distribution. Here, we consider $\sigma_{ij} = 0.5^{|i-j|}$. The sample size n is set to be 200, the dimension p is 2000. The response is generated from the following four submodels:

$$Y = c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 1(X_{12} < 0) + c_4\beta_4 X_{22} + \epsilon, \tag{25}$$

$$Y = c_1\beta_1 X_1 X_2 + c_3\beta_2 1(X_{12} < 0) + c_4\beta_3 X_{22} + \epsilon, \tag{26}$$

$$Y = c_1\beta_1 X_1 X_2 + c_3\beta_2 1(X_{12} < 0) X_{22} + \epsilon, \tag{27}$$

$$Y = c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 1(X_{12} < 0) + \exp\{c_4|X_{22}|\}\epsilon, \tag{28}$$

where $1(\cdot)$ is an indicator function. The regression functions $E(Y|\mathbf{X})$ in (25)–(28) are all nonlinear in X_{12}. Moreover, (26) and (27) contain interaction terms, and (28) is heteroscedastic. Following Fan and Lv (2008), we set $(c_1, c_2, c_3, c_4) = (2, 0.5, 3, 2)$ and choose $\beta_j = (-1)^U(a + |Z|)$ for $j = 1, \ldots, 4$, where $a = 4\log n/\sqrt{n}$, $U \sim$ Bernoulli(0.4) and $Z \sim N(0, 1)$. Especially, the parameters $(\beta_1, \beta_2, \beta_3, \beta_4)$ we generated is $(-3.9, 1.8, -2.4, -2.3)$.

Model D2: (Heavy-tailed single index regression model). As in Mai and Zou (2015a), we consider

$$Y = (X_1 + X_2 + 1)^3 + \epsilon, \tag{29}$$

where X_k's independently follow the Cauchy distribution and ϵ following $N(0, 1)$ is independent of covariates. We let $n = 200, p = 5000$.

Model D3: (Additive model). Case 1: Following Meier, Van de Geer, and Bühlmann (2009) and Cui, Li, and Zhong (2015), we define the following four functions: $f_1(x) = -\sin(2x)$, $f_2(x) = x^2 - 25/12$, $f_3(x) = x$, $f_4(x) = e^{-x} - 2/5 \cdot \sinh(5/2)$. Then we consider the following additive model

$$Y = 3f_1(X_1) + f_2(X_2) - 1.5f_3(X_3) + f_4(X_4) + \epsilon, \tag{30}$$

where the predictors are generated independently from Unif$(-2.5, 2.5)$. To examine the robustness of each screening approach, we consider two cases for the error term ϵ: (1) $\epsilon \sim N(0, 1)$; (2) $\epsilon \sim t(1)$.
Case 2: This nonlinear additive model has been analyzed in Meier, Van de Geer, and Bühlmann (2009) and Fan, Feng, and Song (2011). Let $g_1(x) = x$, $g_2(x) = (2x - 1)^2$, $g_3(x) = \sin(2\pi x)/(2 - \sin(2\pi x))$, and $g_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x)$. The following model is studied:

$$Y = 5g_1(X_1) + 3g_2(X_2) + 4g_3(X_3) + 6g_4(X_4) + \sqrt{1.74}\epsilon, \tag{31}$$

where the covariates are independently simulated according to Unif$(0, 1)$, and ϵ is independent of the covariates and follows the standard normal distribution. We let $(n, p) = (200, 2000)$ for Case 1 and $(400, 1000)$ for Case 2.

Model D4: (Heteroscedastic regression model; Zhu et al. (2011)). The predictor vector $(X_1, X_2, \cdots, X_p)$ is generated in the same way as that in Model D1, the error term $\epsilon \sim N(0, 1)$, and $(n, p) = (200, 2000)$. The response is generated from the following model:

$$Y = X_1 + 0.8X_2 + 0.6X_3 + 0.4X_4 + 0.2X_5 + \exp(X_{20} + X_{21} + X_{22}) \cdot \epsilon. \tag{32}$$

Model D5: (Discriminant analysis model; Cui, Li, and Zhong (2015)): We generate $Y_i \in \{1, 2, \ldots, R\}$ from two different distributions: (i) balanced, a discrete uniform distribution with $P(Y_i = r) = 1/R$ for every $1 \leq r \leq R$; (ii) unbalanced, the sequence of probabilities is an arithmetic progression with $\max_{1\leq r\leq R} p_r = 2\min_{1\leq r\leq R} p_r$. For example, Y is binary when $R = 2$ and $p_1 = 1/3, p_2 = 2/3$. Given $Y_i = r$, the ith predictor X_i is generated by letting $X_i = \mu_r + \epsilon_i$, where $\mu_r = (0, \ldots, 0, 3, 0, \ldots, 0)$ is a p-dimensional vector with rth component being 3 but others being all zero, and $\epsilon = (\epsilon_{i1}, \ldots, \epsilon_{ip})$ is a p-dimensional error term. We consider three cases of the error term: (1) $\epsilon_{ij} \sim N(0, 1)$; (2) $\epsilon_{ij} \sim t(2)$; (3) $\epsilon_{ij} \sim t(1)$ independently for every $j = 1, \ldots, p$. We consider $(R, n, p) = (10, 200, 2000)$, corresponding to a 10-categorical response case. Because a value of the response Y is a nominal number in this case, to apply DCS and GDCS for this problem, we transfer the 10-categorical response to nine dummy binary variables according to Cui, Li, and Zhong (2015), which are together considered as a new multiple response. The active predictors are $X_1, X_2, \ldots, X_{10}$.

Model D6 (The Box-Cox transformation model; (Li et al. 2012)):

$$H(Y) = \mathbf{X}^T\boldsymbol{\beta} + \epsilon, \tag{33}$$

In the simulations, we consider the Box-Cox transformation:

$$H(Y) = \frac{|Y|^\lambda \mathrm{sgn}(Y) - 1}{\lambda}, \text{when } \lambda = 0.25, 0.5, 0.75, 1;$$
$$H(Y) = \log Y, \text{when } \lambda = 0.$$

Table 2. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $\mathcal{M}$ out of 200 replications for Model D1–D6, and the proportion of $\mathcal{P}_a$ with model size $d = 2[n/\log n]$.

			$\mathcal{M}$					$\mathcal{P}_a$				$\mathcal{M}$					$\mathcal{P}_a$
Model		Method	5%	25%	50%	75%	95%		Model		Method	5%	25%	50%	75%	95%	
D1	(25)	DCS	4.0	4.0	4.0	6.0	24.2	0.98	D2		DCS	18.0	69.8	212.0	740.5	2285.2	0.26
[4]		UDCS	4.0	4.0	5.0	6.0	44.4	0.97	[2]		UDCS	2.0	2.0	2.0	2.0	2.0	1.00
		NS-DCS	4.0	4.0	4.0	6.0	28.3	0.98			NS-DCS	2.0	2.0	2.0	2.0	2.0	1.00
		Wins-DCS	4.0	4.0	4.0	6.0	28.3	0.98			Wins-DCS	2.0	2.0	2.0	2.0	2.0	1.00
	(26)	DCS	4.0	5.0	7.0	11.0	38.6	0.98	D3	Case 1:	DCS	10.0	28.0	53.0	92.2	198.6	0.65
		UDCS	6.0	12.0	24.0	50.2	152.6	0.85	[4]	$\epsilon \sim N(0, 1)$	UDCS	11.0	29.8	54.0	94.0	211.0	0.66
		NS-DCS	4.0	6.0	9.0	15.2	52.3	0.97			NS-DCS	5.0	13.0	23.5	46.2	108.3	0.90
		Wins-DCS	4.0	6.0	9.0	15.2	52.3	0.97			Wins-DCS	5.0	13.0	24.0	47.0	106.2	0.90
	(27)	DCS	18.9	79.0	211.5	473.2	1274.5	0.25		Case 1:	DCS	28.0	95.5	234.0	427.0	1071.8	0.19
		UDCS	22.9	88.2	217.5	426.8	976.0	0.18		$\epsilon \sim t(1)$	UDCS	21.0	79.5	137.0	208.2	403.4	0.23
		NS-DCS	22.9	73.8	167.5	365.8	1099.4	0.25			NS-DCS	11.0	43.8	79.0	143.2	361.5	0.48
		Wins-DCS	22.0	74.5	166.0	363.2	1085.1	0.25			Wins-DCS	10.0	43.0	77.5	142.2	367.7	0.48
	(28)	DCS	4.0	7.0	19.5	79.5	400.3	0.73		Case 2	DCS	6.0	15.0	36.0	95.5	262.0	0.80
		UDCS	34.9	110.5	211.5	385.2	748.6	0.15			UDCS	6.0	17.8	39.0	100.8	294.1	0.80
		NS-DCS	7.0	20.0	39.5	88.2	256.1	0.71			NS-DCS	5.0	8.8	20.0	56.0	215.3	0.89
		Wins-DCS	7.0	19.0	37.0	81.0	248.1	0.73			Wins-DCS	5.0	8.8	20.0	56.0	210.5	0.90
D4		DCS	26.0	193.8	413.0	758.8	1393.9	0.13	D5	$\epsilon \sim N(0, 1)$:	DCS	10.0	10.0	10.0	10.0	13.1	1.00
[8]		UDCS	14.0	23.0	40.5	108.0	571.9	0.66	[10]	Balanced	UDCS	10.0	10.0	11.0	20.2	96.1	0.93
		NS-DCS	9.0	13.0	23.5	95.8	601.2	0.72			NS-DCS	10.0	10.0	10.0	11.0	31.1	0.98
		Wins-DCS	9.0	13.0	23.5	96.0	600.0	0.71			Wins-DCS	10.0	10.0	10.0	11.0	28.0	0.98
D6	λ=0	DCS	192.6	468.5	836.0	1250.5	1748.7	0.00		$\epsilon \sim N(0, 1)$:	DCS	10.0	10.0	12.0	28.5	322.1	0.85
[3]		UDCS	5.0	19.8	88.0	267.8	713.7	0.33		Unbalanced	UDCS	11.0	20.0	60.5	203.8	828.5	0.55
		NS-DCS	4.0	14.0	55.5	179.8	580.8	0.38			NS-DCS	10.0	12.0	25.0	86.2	566.9	0.72
		Wins-DCS	4.0	14.0	55.5	174.0	580.9	0.40			Wins-DCS	10.0	12.0	24.0	82.8	545.2	0.73
	λ=0.25	DCS	11.0	69.0	173.0	434.0	1001.1	0.16		$\epsilon \sim t(2)$:	DCS	10.0	11.0	16.0	43.8	251.1	0.79
		UDCS	6.0	25.0	92.0	304.2	899.3	0.30		Balanced	UDCS	11.0	16.0	33.0	84.2	333.4	0.72
		NS-DCS	4.0	18.8	66.5	243.2	703.4	0.37			NS-DCS	10.0	12.0	22.0	74.2	266.4	0.75
		Wins-DCS	4.0	17.0	65.0	242.0	691.8	0.37			Wins-DCS	10.0	12.0	21.0	70.8	262.1	0.76
	λ=0.5	DCS	4.0	14.8	56.0	183.0	759.3	0.41		$\epsilon \sim t(2)$:	DCS	15.0	37.8	154.5	331.2	1033.4	0.36
		UDCS	5.0	24.0	101.5	278.2	787.4	0.30		Unbalanced	UDCS	22.9	65.5	201.0	429.2	901.2	0.29
		NS-DCS	4.0	14.0	71.5	204.0	598.7	0.34			NS-DCS	19.9	60.8	181.5	404.0	981.9	0.30
		Wins-DCS	4.0	14.0	70.5	196.8	561.5	0.34			Wins-DCS	19.9	59.8	177.0	400.5	973.5	0.30
	λ=0.75	DCS	3.0	10.0	42.0	178.2	514.0	0.43		$\epsilon \sim t(1)$:	DCS	373.9	1194.2	1704.0	1905.2	1987.3	0.01
		UDCS	4.0	19.8	86.5	320.2	699.2	0.30		Balanced	UDCS	29.0	85.8	201.0	382.5	1045.0	0.21
		NS-DCS	4.0	15.0	62.5	234.8	596.5	0.38			NS-DCS	24.0	76.0	198.0	399.5	1024.0	0.25
		Wins-DCS	4.0	14.8	61.5	236.2	551.9	0.39			Wins-DCS	23.0	74.8	198.5	398.2	1018.9	0.25
	λ=1	DCS	3.0	14.8	51.5	164.0	495.9	0.39		$\epsilon \sim t(1)$:	DCS	540.8	1300.2	1700.5	1911.2	1981.1	0.00
		UDCS	4.0	24.0	84.5	305.2	745.0	0.29		Unbalanced	UDCS	60.0	200.8	397.5	842.2	1423.9	0.10
		NS-DCS	4.0	17.8	64.0	200.2	522.6	0.36			NS-DCS	47.9	189.8	383.5	837.0	1472.2	0.11
		Wins-DCS	4.0	16.0	63.0	202.0	507.2	0.36			Wins-DCS	46.9	189.5	381.5	828.5	1470.2	0.11

NOTE: The numbers in the brackets are the true numbers of variables.

The predictor $(X_1, X_2, \ldots, X_p)$ is generated from a multivariate normal distribution $N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$ has entries $\sigma_{ii} = 1, i = 1, \ldots, p$ and $\sigma_{ij} = 0.5, i \neq j$. The noise ϵ follows the standard normal distribution, $\boldsymbol{\beta} = (3, 1.5, 2, 0, \ldots, 0)^T, n = 70, p = 2000$.

In all the models, GDCS is either comparable to or significantly better than UDCS, although UDCS can outperform DCS when heavy tails are present. For Model D1, in all cases, GDCS behaves comparably with DCS and better than UDCS. In the presence of nonlinearity and heavy-tailed data in Model D2, DCS has much more false discoveries than GDCS and UDCS, where the latter two are comparable. The variable transformation approach with DC can handle the issue of heavy-tailed data well. For Models D3–D6, GDCS is consistently promising and robust with the best results. In Model D6, it can also be inferred from the results that GDCS has invariance property under monotonic transformation. The little difference across different λ is due to different random errors generated for models. When the model deviates from a linear model and Y from normal (λ decreases from one), the performance of DCS quickly deteriorates due to the existence of the nonlinearity and heavy-tailed response. See Figure S1 in the supplementary materials for a visual demonstration.

5.2. A Real Dataset Example

We demonstrate the application of NS-NSC with the malaria dataset (Ockenhouse et al. 2006). This dataset contains measurements of 22,283 gene expressions of 71 human subjects. Twenty-two of the human subjects are healthy, and the rest have malaria. Prior knowledge is available on some of the genes collected in this dataset. For example, the gene IRF1 was known to be related to the immune response of human. Fan and Fan (2008) proposed to rank the importance of the genes by the absolute values of their t-statistics. On the original dataset, the gene IRF1 is ranked as the 125th most important gene by t-statistics ranking. It is also ranked as the 497th most important gene by DCS ranking. With the normal score transformation, IRF1 is recognized as the second most important gene by both methods. This suggests that the normal score transformation gives a more meaningful ranking.

To further investigate the effect of normal score transformation, we randomly split the dataset in a balanced manner with a

Table 3. The average error rates (%) using RF or LDA combined with different screening methods in 100 randomly split malaria data, with their standard errors shown in parentheses.

Methods	DCS	NS-DCS	Win-DCS	Methods	DCS	NS-DCS	Win-DCS
RF	5.3 (0.40)	3.8 (0.32)	4.3 (0.36)	LDA	15.1 (0.76)	10.2 (0.57)	11.6 (0.69)

1:1 ratio into training and testing datasets. We fit classifiers on the training set and evaluate the testing error on the testing set. On the raw dataset, NSC has an error rate of 8.6%, with a standard error of 1.37%. If we apply the normal score transformation or Winsorized transformation, both NS-NSC and Wins-NSC lower the error rate to 5.7% with standard errors of 0.70% and 0.20%, respectively. Paired t-test indicate that the improvement is significant, with p-values less than 10^{-4}.

In addition, we include the DCS, NS-DCS, Wins-DCS to carry out the classification for comparison. Again, we randomly split the dataset in the same way aforementioned, and apply each screening method to the training set to select top $d = 2[n_{\text{train}}/\log n_{\text{train}}]$ genes, where n_{train} is the training sample size. Then, we fit a random forest (RF) or a linear discriminant analysis (LDA) model using selected features to do classification and make prediction in the testing set. The above procedure is repeated 100 times. The average error rates are reported in Table 3. Paired t-tests indicate that the improvement of error rate obtained by fitting a RF or LDA model after coordinatewise Gaussianization in the screening stage is significant, with all p-values less than 10^{-3}. It is also not surprising that DCS cannot identify the gene IRF1 in those 100 trials while NS-DCS and Wins-DCS select IRF1 for 66 and 60 times, respectively.

6. Discussion

In this article, we establish the uniform convergence of coordinatewise Gaussianization as long as $\log p = o(\frac{n}{\log n})$. This result is independent of any downstream statistical method to be used after the variable transformation. We have also provided three concrete statistical methods to show that when the theoretical normal transformation is helpful, coordinatewise Gaussianization achieves similar performance.

We have considered two methods for coordinatewise Gaussianization: the NS estimator and the Winsorized estimator. The two methods have identical theoretical properties in ultra-high dimensions. Throughout our numerical studies, their performance also exhibits minimal difference. Hence, if one wishes to perform coordinatewise Gaussianization in practice, either of them is expected to achieve the goal. However, we note that the NS estimator has a much longer history and is more widely applied in many areas such as statistics, biostatistics, education and econometrics. In comparison, the Winsorized estimator was more recently proposed mainly for theoretical studies. Therefore, if there is no strong reason to prefer the Winsorized estimator in the problem at hand, the NS estimator may be more coherent with studies in the past.

We emphasize again that coordinatewise Gaussianization should be avoided when the theoretical normal transformation does not help or even harm the downstream statistical method. For example, popular tree-building algorithms are invariant under monotone transformations, and coordinatewise Gaussianization does not have any effect in these tree-based methods and ensemble-trees learning. On the other hand, methods that rely on marginal distribution of the variables usually should not be combined with the coordinatewise Gaussianization, because all of them have the same distribution after the transformation.

Based on the above counter-examples, we recommend the following procedure for using coordinatewise Gaussianization in applications. We should always do a careful analysis of the downstream method on the theoretically transformed data to see if the transformation provides any benefit. Only after getting an affirmative conclusion, we then carry out coordinatewise Gaussianization and proceed with the intended statistical method.

Supplementary Materials

For the sake of space, all the technical proofs and additional simulation results are relegated to the supplementary file.

Acknowledgments

The authors would like to thank the associate editor and three referees for helpful comments.

Funding

Qing Mai is supported in part by NSF CCF 1908969. Di He is joint first author of this article. His research is supported by the Youth Program of National Natural Science Foundation of China (11901289). Hui Zou is supported in part by NSF DMS 1915842 and 2015120.

References

Anokhin, A. P., Heath, A. C., and Ralano, A. (2003), "Genetic Influences on Frontal Brain Function: WCST Performance in Twins," *Neuroreport*, 14, 1975–1978. [2]

Beasley, T. M., Erickson, S., and Allison, E. B. (2009), "Rank-Based Inverse Normal Transformations are Increasingly Used, but are they Merited?" *Behavior Genetics*, 35, 580–595. [3]

Berkowitz, J. (2001), "Testing Density Forecasts, with Applications to Risk Management," *Journal of Business & Economic Statistics*, 19, 465–474. [2]

Bliss, C. I. (1967), *Statistics in Biology*, New York: McGraw-Hill. [3]

Blom, G. (1958), *Statistical Estimates and Transformed Beta-Variables*, New York: Wiley. [3]

Cai, T., and Liu, W. (2011), "A Direct Estimation Approach to Sparse Linear Discriminant Analysis," *Journal of the American Statistical Association*, 106, 1566–1577. [7]

Cai, T., Liu, W., and Luo, X. (2011), "A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation," *Journal of American Statistical Association*, 106, 594–607. [5]

Cai, T. T., and Zhang, L. (2018), "High-Dimensional Gaussian Copula Regression: Adaptive Estimation and Statistical Inference," *Statistica Sinica*, 28, 963–993. [5]

Cai, X., Li, H., and Liu, A. (2016), "A Marginal Rank-Based Inverse Normal Transformation Approach to Comparing Multiple Clinical Trial Endpoints," *Statistics in Medicine*, 35, 3259–3271. [2]

Chang, J., Tang, C. Y., and Wu, Y. (2013), "Marginal Empirical Likelihood and Sure Independence Feature Screening," *The Annals of Statistics*, 41, 2123–2148. [8]

Chang, J., Tang, C. Y., and Wu, Y. (2016), "Local Independence Feature Screening for Nonparametric and Semiparametric Models by Marginal Empirical Likelihood," *Annals of Statistics*, 44, 515–539. [8]

Chen, X., Chen, X., and Wang, H. (2018), "Robust Feature Screening for Ultra-High Dimensional Right Censored Data via Distance Correlation," *Computational Statistics & Data Analysis*, 119, 118–138. [9]

Chen, X., and Fan, Y. (2006), "Estimation of Copula-Based Semiparametric Time Series Models," *Journal of Econometrics*, 130, 307–335. [5]

Cui, H., Li, R., and Zhong, W. (2015), "Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis," *Journal of American Statistical Association*, 110, 630–641. [8,11]

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C., Taylor, J., Burnett, E., Gut, I., Farrall, M., Mark Lathrop, G., Abecasis, G. R., and Cookson, W. O. C. (2007), "A Genome-Wide Association Study of Global Gene Expression," *Nature Genetics*, 39, 1202–1207. [2]

Fan, J., and Fan, Y. (2008), "High Dimensional Classification Using Features Annealed Independence Rules," *The Annals of Statistics*, 36, 2605–2637. [12]

Fan, J., Feng, Y., and Song, R. (2011), "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models," *Journal of the American Statistical Association*, 106, 544–557. [8,11]

Fan, J., Feng, Y., and Tong, X. (2012), "A ROAD to Classification in High Dimensional Space," *Journal of the Royal Statistical Society,* Series B, 74, 745–771. [7]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultra-High Dimensional Feature Space," *Journal of the Royal Statistical Society,* Series B, 20, 101–148. [8,11]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models with NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [8]

Fan, J., Xue, L., and Zou, H. (2015), "Multi-Task Quantile Regression Under the Transnormal Model," *Journal of American Statistical Association*, 111, 1726–1735. [5]

Fan, R., Wang, Y., Mills, J. L., Wilson, A. F., Bailey-Wilson, J. E. and Xiong, M. (2013), "Functional Linear Models for Association Analysis of Quantitative Traits," *Genetic Epidemiology*, 37, 726–742. [2]

Friedman, J. H., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation with the Graphical Lasso," *Biostatistics*, 9, 432–441. [5]

Glass, G. V., and Hopkins, K. D. (1996), *Statistical Methods in Education and Psychology* (3rd ed.), Boston: Allyn & Bacon. [1]

Han, F., and Liu, H. (2014), "High Dimensional Semiparametric Scale-Invariant Principal Component Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36, 2016–2032. [2]

Hastie, T., Tibshirani, R., and Friedman, J. H. (2008), *Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), New York: Springer Verlag. [4,6]

Hoff, P. D. (2007), "Extending the Rank Likelihood for Semiparametric Copula Estimation," *Annals of Applied Statistics*, 1, 265–283. [5]

Hoff, P. D., Niu, X., and Wellner, J. A. (2014), "Information Bounds for Gaussian Copulas," *Bernoulli*, 20, 604–622. [2,5]

Jin, J., and Wang, W. (2016), "Influential Features PCA for High Dimensional Clustering," *The Annals of Statistics*, 44, 2323–2359. [4]

Johnstone, I. M., and Lu, A. Y. (2009), "On Consistency and Sparsity for Principal Component Analysis in High Dimensions," *Journal of the American Statistical Association*, 104, 682–693. [4]

Klaassen, C., and Wellner, J. (1997), "Efficient Estimation in the Bivariate Normal Copula Model: Normal Margins are Least Favourable," *Bernoulli*, 3, 55–77. [2,5]

Lambregts-Rommelse, N., Arias-Vasquez, A., Altink, M., Buschgens, C., Fliers, E., Asherson, P., Faraone, S., Buitelaar, J., Sergeant, J., Oosterlaan, J., Franke, B. (2008), "Neuropsychological Endophenotype Approach to Genome-Wide Linkage Analysis Identifies Susceptibility Loci for adhd on 2q21. 1 and 13q12. 11," *American Journal of Human Genetics*, 83, 99–105. [2]

Li, G., Peng, H., Zhang, J., and Zhu, L.-X. (2012), "Robust Rank Correlation Based Screening," *The Annals of Statistics*, 40, 1846–1877. [8,11]

Li, R., Zhong, W., and Zhu, L.-P. (2012), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [8,9,10,11]

Lin, Y., and Jeon, Y. (2003), "Discriminant Analysis Through a Semiparametric Model," *Biometrika*, 90, 379–392. [5,8]

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012), "High-dimensional Semiparametric Gaussian Copula Graphical Models," *Annals of Statistics*, 40, 2293–2326. [2,5]

Liu, H., Lafferty, J., and Wasserman, L. (2009), "The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs," *Journal of Machine Learning Research*, 10, 2295–2328. [2,4,5,6]

Mai, Q., and Zou, H. (2013), "The Kolmogorov Filter for Variable Screening in High-Dimensional Binary Classification," *Biometrika*, 100, 229–234. [8]

——— (2015a), "The Fused Kolmogorov Filter: A Nonparametric Model-Free Screening Method," *Annals of Statistics*, 43, 1471–1497. [8,9,11]

——— (2015b), "Nonparametric Variable Transformation in Sufficient Dimension Reduction," *Technometrics*, 57, 1–10. [2,5]

——— (2015c), "Sparse Semiparametric Discriminant Analysis," *Journal of Multivariate Analysis*, 135, 175–188. [2,4,5,6,8]

Mai, Q., Zou, H., and Yuan, M. (2012), "A Direct Approach to Sparse Discriminant Analysis in Ultra-High Dimensions," *Biometrika*, 99, 29–42. [6,7]

McDiarmid, C. (1989), "On the Method of Bounded Differences," in *Surveys in Combinatorics, (Norwich, 1989)* Volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188, Cambridge: Cambridge University Press. [3]

Meier, L., Van de Geer, S., and Bühlmann, P. (2009), "High-dimensional Additive Modeling," *The Annals of Statistics*, 37, 3779–3821. [11]

Meinshausen, N., and Bühlmann, P. (2006), "High Dimensional Graphs and Variable Selection with the Lasso," *Annals of Statistics*, 34, 1436–1462. [5]

Nansel, T. R., Laffel, L. M., Haynie, D. L., Mehta, S. N., Lipsky, L. M., Volkening, L. K., Butler, D. A., Higgins, L. A., and Liu, A. (2015), "Improving Dietary Quality in Youth with Type 1 Diabetes: Randomized Clinical Trial of a Family-Based Behavioral Intervention," *International Journal of Behavioral Nutrition and Physical Activity*, 12, 1–11. [2]

Ockenhouse, C. F., Hu, W. C., Kester, K. E., Cummings, J. F., Stewart, A., Heppner, D. G., Jedlicka, A. E., Scott, A. L., Wolfe, N. D., Vahey, M., and Burke, D. S. (2006), "Common and Divergent Immune Response Signaling Pathways Discovered in Peripheral Blood Mononuclear Cell Gene Expression Patterns in Presymptomatic and Clinically Apparent Malaria," *Infection and Immunity*, 74, 5561–5573. [12]

Peng, B., Robert, K. Y., DeHoff, K. L., and Amos, C. I. (2007), "Normalizing a Large Number of Quantitative Traits Using Empirical Normal Quantile Transformation," *BMC Proceedings*, 1, S156. doi:10.1186/1753-6561-1-S1-S156. [2]

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial Correlation Estimation by Joint Sparse Regression Models," *Journal of American Statistical Association*, 104, 735–746. [5]

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011), "High-dimensional Covariance Estimation by Minimizing l1 Log-Determinant Divergence," *Electronic Journal of Statistics*, 5, 935–980. [6]

Scuteri, A., Sanna, S., Chen, W.-M., Uda, M., Albai, G., Strait, J., Najjar, S., Nagaraja, R., Orrú, M., Usala, G., Dei, M., Lai, S., Maschio, A., Busonero, F., Mulas, A., Ehret, G. B., Fink, A. A., Weder, A. B., Cooper, R. S., Galan, P., Chakravarti, A., Schlessinger, D., Cao, A., Lakatta, E., and Abecasis, G. R. (2007), "Genome-wide Association Scan Shows Genetic Variants in the FTO Gene are Associated with Obesity-Related Traits," *PLoS Genetics*, 3, e115. [2]

Serfling, R. J. (2009), *Approximation Theorems of Mathematical Statistics* (Vol. 162), New York: Wiley. [2,5]

Sklar, M. (1959), "Fonctions de Repartition an Dimensions et Leurs Marges," *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231. [4]

Székely, G. J., and Rizzo, M. L. (2009), "Brownian Distance Covariance," *Annals of Applied Statistics*, 3, 1236–1265. [9]

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and Testing Dependence by Correlation of Distances," *The Annals of Statistics*, 35, 2769–2794. [8]

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002), "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proceedings of the National Academy of Sciences*, 99, 6567–6572. [6]

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003), "Class Prediction by Nearest Shrunken Centroids, with Applications to DNA," *Statistical Science*, 18, 104–117. [6]

Tukey, J. W. (1962), "The Future of Data Analysis," *Annals of Mathematical Statistics*, 33, 1–67. [3]

Van der Waerden, B. (1952), "Order Tests for the Two-Sample Problem and their Power," in *Indagationes Mathematicae (Proceedings)* (Vol. 55), pp. 453–458, Elsevier. [3]

Wang, Y., Liu, A., Mills, J. L., Boehnke, M., Wilson, A. F., Bailey-Wilson, J. E., Xiong, M., Wu, C. O., and Fan, R. (2015), "Pleiotropy Analysis of Quantitative Traits at Gene Level by Multivariate Functional Linear Models," *Genetic Epidemiology*, 39, 259–275. [2]

Wu, X., Cooper, R. S., Borecki, I., Hanis, C., Bray, M., Lewis, C. E., Zhu, X., Kan, D., Luke, A., and Curb, D. (2002), "A Combined Analysis of Genomewide Linkage Scans for Body Mass Index, from the National Heart, Lung, and Blood Institute Family Blood Pressure Program," *The American Journal of Human Genetics*, 70, 1247–1256. [2]

Xue, L., and Zou, H. (2012), "Regularized Rank-Based Estimation of High-Dimensional Nonparanormal Graphical Models," *The Annals of Statistics*, 40, 2541–2571. [2,5,6,9]

Yuan, M. (2010), "High Dimensional Inverse Covariance Matrix Estimation via Linear Programming," *Journal of Machine Learning Research*, 11, 2261–2286. [5]

Zhang, T., and Zou, H. (2014), "Sparse Precision Matrix Estimation via Lasso Penalized d-Trace Loss," *Biometrika*, 101, 103–120. [5]

Zhong, W., Zhu, L., Li, R., and Cui, H. (2016), "Regularized Quantile Regression and Robust Feature Screening for Single Index Models," *Statistica Sinica*, 26, 69–95. [9]

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011), "Model-Free Feature Screening for Ultrahigh Dimensional Data," *Journal of the American Statistical Association*, 106, 1464–1475. [8,11]