Nonparanormal Information Estimation

Shashank Singh [1]   **Barnabás Póczos** [1]

Abstract

We study the problem of using i.i.d. samples from an unknown multivariate probability distribution p to estimate the mutual information of p. This problem has recently received attention in two settings: (1) where p is assumed to be Gaussian and (2) where p is assumed only to lie in a large nonparametric smoothness class. Estimators proposed for the Gaussian case converge in high dimensions when the Gaussian assumption holds, but are brittle, failing dramatically when p is not Gaussian, while estimators proposed for the nonparametric case fail to converge with realistic sample sizes except in very low dimension. Hence, there is a lack of robust mutual information estimators for many realistic data. To address this, we propose estimators for mutual information when p is assumed to be a nonparanormal (or Gaussian copula) model, a semiparametric compromise between Gaussian and nonparametric extremes. Using theoretical bounds and experiments, we show these estimators strike a practical balance between robustness and scalability.

1. Introduction

This paper is concerned with the problem of estimating entropy or mutual information of an unknown probability density p over $\mathbb{R}^D$, given n i.i.d. samples from p. Entropy and mutual information are fundamental information theoretic quantities, and consistent estimators for these quantities have a host of applications within machine learning, statistics, and signal processing. For example, entropy estimators have been used for goodness-of-fit testing (Goria et al., 2005), parameter estimation in semi-parametric models (Wolsztynski et al., 2005), texture classification and image registration (Hero et al., 2001; 2002), change point de-

tection (Bercher & Vignat, 2000), and anomaly detection in networks (Noble & Cook, 2003; Nychis et al., 2008; Bereziński et al., 2015). Mutual information is a popular nonparametric measure of dependence, whose estimators have been used in feature selection (Peng et al., 2005; Shishkin et al., 2016), clustering (Aghagolzadeh et al., 2007), learning graphical models (Chow & Liu, 1968), fMRI data processing (Chai et al., 2009), prediction of protein structures (Adami, 2004), boosting and facial expression recognition (Shan et al., 2005), and fitting deep nonlinear models (Hunter & Hodas, 2016). Estimators for both entropy and mutual information have been used in independent component and subspace analysis (Learned-Miller & Fisher, 2003; Szabó et al., 2007a).

Motivated by these and other applications, several very recent lines of work (discussed in Section 3) have studied information estimation,[1] focusing largely on two settings:

1. **Gaussian Setting:** If p is known to be Gaussian, there exist information estimators with mean squared error (MSE) at most $-2\log\left(1 - \frac{D}{n}\right)$ and an (almost matching) minimax lower bound of $2D/n$ (Cai et al., 2015).

2. **Nonparametric Setting:** If p is assumed to lie in a nonparametric smoothness class, such an s-order[2] Hölder or Sobolev class, then the minimax MSE is of asymptotic order $\asymp \max\left\{n^{-1}, n^{-\frac{8s}{4s+D}}\right\}$ (Birgé & Massart, 1995).

In the Gaussian setting, consistent estimation is tractable even in the high-dimensional case where D increases fairly quickly with n, as long as $D/n \to 0$. However, optimal estimators for the Gaussian setting rely heavily on the assumption of joint Gaussianity, and their performance can degrade quickly when the data deviate from Gaussian. Especially in high dimensions, it is unlikely that data are jointly Gaussian, making these estimators brittle in practice. In the nonparametric setting, the theoretical convergence rate decays exponentially with D, and, it has been found empirically that information estimators for this setting fail to converge at realistic sample sizes in all but very low dimensions. Also, most nonparametric estimators are sensitive to tuning of bandwidth parameters, which is chal-

---

[1]Carnegie Mellon University, Pittsburgh, USA. Correspondence to: Shashank Singh <sss1@andrew.cmu.edu>.

---

[1]We will collectively call the closely related problems of entropy and mutual information estimation *information estimation*.

[2]Here, s encodes the degree of smoothness, roughly corresponding to the number of continuous derivatives of p.

lenging for information estimation, since no empirical error estimate is available for cross-validation.

Given these factors, though the Gaussian and nonparametric cases are fairly well understood in theory, there remains a lack of practical information estimators for the common case where data are neither exactly Gaussian nor very low dimensional. The **main goal of this paper** is to fill the gap between these two extreme settings by studying information estimation in a semiparametric compromise between the two, known as the "nonparanormal" (a.k.a. "Gaussian copula") model (see Definition 2). The nonparanormal model, analogous to the additive model popular in regression (Friedman & Stuetzle, 1981), limits complexity of interactions among variables but makes minimal assumptions on the marginal distribution of each variable. The result scales better with dimension than nonparametric models, while being more robust than Gaussian models.

Proofs of our main results, as well as additional experiments and details are given in the extended version of this paper [3]. All code can be found on GitHub[4].

2. Problem statement and notation

There are a number of distinct generalizations of mutual information to more than two variables. The definition we consider is simply the difference between the sum of marginal entropies and the joint entropy:

Definition 1. (Multivariate mutual information) *Let* $X_1, \ldots, X_D$ *be* $\mathbb{R}$*-valued random variables with a joint probability density* $p : \mathbb{R}^D \to [0, \infty)$ *and marginal densities* $p_1, ..., p_D : \mathbb{R} \to [0, \infty)$. *The* multivariate mutual information $I(X)$ *of* $X = (X_1, \ldots, X_D)$ *is defined by*

$$I(X) := \mathop{\mathbb{E}}_{X \sim p} \left[\log \left(\frac{p(X)}{\prod_{j=1}^{D} p_j(X_j)} \right) \right]$$
$$= \sum_{j=1}^{D} H(X_j) - H(X), \quad (1)$$

where $H(X) = -\mathbb{E}_{X \sim p}[\log p(X)]$ *denotes entropy of* X.

This notion of multivariate mutual information, originally due to Watanabe (1960) (who called it "total correlation") measures total dependency, or redundancy, within a set of D random variables. It has also been called the "multivariate constraint" (Garner, 1962) and "multi-information" (Studený & Vejnarová, 1998). Many related information theoretic quantities can be expressed in

terms of $I(X)$, and can thus be estimated using estimators of $I(X)$. Examples include pairwise mutual information $I(X, Y) = I((X, Y)) - I(X) - I(Y)$, which measures dependence between (potentially multivariate) random variables X and Y, conditional mutual information

$$I(X|Z) = I((X, Z)) - \sum_{j=1}^{D} I((X_j, Z)),$$

which is useful for characterizing how much dependence within X can be explained by a latent variable Z (Studený & Vejnarová, 1998), and transfer entropy (a.k.a. "directed information") $T_{X \to Y}$, which measures predictive power of one time series X on the future of another time series Y. $I(X)$ is also related to entropy via Eq. (1), but, unlike the above quantities, this relationship depends on the marginal distributions of X, and hence involves some additional considerations, as discussed in Section 8.

We now define the class of nonparanormal distributions, from which we assume our data are drawn.

Definition 2. (Nonparanormal distribution, a.k.a. Gaussian copula model) *A random vector* $X = (X_1, \ldots, X_D)^T$ *is said to have a* nonparanormal distribution *(denoted* $X \sim \mathcal{NPN}(\Sigma; f)$*) if there exist functions* $\{f_j\}_{j=1}^{D}$ *such that each* $f_j : \mathbb{R} \to \mathbb{R}$ *is a diffeomorphism* [5] *and* $f(X) \sim \mathcal{N}(0, \Sigma)$*, for some (strictly) positive definite* $\Sigma \in \mathbb{R}^{D \times D}$ *with 1's on the diagonal (i.e., each* $\sigma_j = \Sigma_{j,j} = 1$*).* [6] Σ *is called the* latent covariance *of* X *and* f *is called the* marginal transformation *of* X.

The nonparanormal family relaxes many constraints of the Gaussian family. Nonparanormal distributions can be multi-modal, skewed, or heavy-tailed, can encode noisy nonlinear dependencies, and need not be supported on $\mathbb{R}^D$. Minimal assumptions are made on the marginal distributions; any desired continuously differentiable marginal cumulative distribution function (CDF) F_i of variable X_i corresponds to marginal transformation $f_i(x) = \Phi^{-1}(F_i(x))$ (where Φ is the standard normal CDF). As examples, for a Gaussian variable Z, the 2-dimensional case, $X_1 \sim \mathcal{N}(0, 1)$, and $X_2 = T(X_1 + Z)$ is nonparanormal when $T(x) = x^3$, $T = \tanh$, $T = \Phi$, or any other diffeomorphism. On the other hand, the limits of the Gaussian copula appear, for example, when $T(x) = x^2$, which is not bijective; then, if $\mathbb{E}[Z] = 0$, the Gaussian copula approximation of (X_1, X_2) models X_1 and X_2 as independent.

---

[5]A diffeomorphism is a continuously differentiable bijection $g : \mathbb{R} \to R \subseteq \mathbb{R}$ such that g^{-1} is continuously differentiable.

[6]Setting $\mathbb{E}[f(X)] = 0$ and each $\sigma_j = 1$ ensures model identifiability, but does not reduce the model space, since these parameters can be absorbed into the marginal transformation f.

We are now ready to formally state our problem:

Formal Problem Statement: *Given n i.i.d. samples $X_1, ..., X_n \sim \mathcal{NPN}(\Sigma; f)$, where Σ and f are both unknown, we would like to estimate $I(X)$.*

Other notation: D denotes the dimension of the data (i.e., $\Sigma \in \mathbb{R}^{D \times D}$ and $f : \mathbb{R}^D \rightarrow \mathbb{R}^D$). For a positive integer k, $[k] = \{1, ..., k\}$ denotes the set of positive integers less than k (inclusive). For consistency, where possible, we use $i \in [n]$ to index samples and $j \in [D]$ to index dimensions (so that, e.g., $X_{i,j}$ denotes the jth dimension of the ith sample). Given a data matrix $X \in \mathbb{R}^{n \times D}$, our estimators depend on the empirical rank matrix

$$R \in [n]^{n \times D} \quad \text{with} \quad R_{i,j} := \sum_{k=1}^{n} 1_{\{X_{i,j} \geq X_{k,j}\}}. \quad (2)$$

For a square matrix $A \in \mathbb{R}^{k \times k}$, $|A|$ denotes the determinant of A, A^T denotes the transpose of A, and

$$\|A\|_2 := \max_{\substack{x \in \mathbb{R}^k \\ \|x\|_2 = 1}} \|Ax\|_2 \quad \text{and} \quad \|A\|_F := \sqrt{\sum_{i,j \in [k]} A_{i,j}^2}$$

denote the spectral and Frobenius norms of A, respectively. When A is symmetric, $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_D(A)$ are its eigenvalues.

3. Related Work and Our Contributions

3.1. The Nonparanormal

Nonparanormal models have been used for modeling dependencies among high-dimensional data in a number of fields, such as graphical modeling of gene expression data (Liu et al., 2012), of neural data (Berkes et al., 2009), and of financial time series (Malevergne & Sornette, 2003; Wilson & Ghahramani, 2010; Hernández-Lobato et al., 2013), extreme value analysis in hydrology (Renard & Lang, 2007; Aghakouchak, 2014), and informative data compression (Rey & Roth, 2012).

With one recent exception (Ince et al., 2017), previous information estimators for the nonparanormal case (Calsaverini & Vicente, 2009; Ma & Sun, 2011; Elidan, 2013), rely on fully nonparametric information estimators as subroutines, and hence suffer strongly from the curse of dimensionality. Very recently, Ince et al. (2017) proposed what we believe is the first mutual information estimator tailored specifically to the nonparanormal case; their estimator is equivalent to one of the estimators (I_G, described in Section 4.1) we study. However, they focused on its applications to neuroimaging data analysis, and did not study its performance theoretically or empirically.

3.2. Information Estimation

Our motivation for studying the nonparanormal family comes from trying to bridge two recent approaches to information estimation. The first has studied fully nonparametric entropy estimation, assuming only that data are drawn from a smooth probability density p, where smoothness is typically quantified by a Hölder or Sobolev exponent $s \in (0, \infty)$, roughly corresponding to the continuous differentiability of s. In this setting, the minimax optimal MSE rate has been shown by Birgé & Massart (1995) to be $O\left(\max\left\{n^{-1}, n^{-\frac{8s}{4s+D}}\right\}\right)$. This rate slows exponentially with the dimension D, and, while many estimators have been proposed (Pál et al., 2010; Sricharan et al., 2011; 2013; Singh & Póczos, 2014a;b; Krishnamurthy et al., 2014; Moon & Hero, 2014b;a; Singh & Póczos, 2016a; Moon et al., 2017) for this setting, their practical use is limited to a few dimensions[7].

The second area is in the setting where data are assumed to be drawn from a truly Gaussian distribution. Here the high-dimensional case is far more optimistic. While this case had been studied previously (Ahmed & Gokhale, 1989; Misra et al., 2005; Srivastava & Gupta, 2008), Cai et al. (2015) recently provided a precise finite-sample analysis based on deriving the exact probability law of the log-determinant $\log|\widehat{\Sigma}|$ of the scatter matrix $\widehat{\Sigma}$. From this, they derived a deterministic bias correction, giving an estimator for which they prove an MSE upper bound of $-2\log\left(1 - \frac{D}{n}\right)$ and a high-dimensional central limit theorem for the case $D \rightarrow \infty$ as $n \rightarrow \infty$ (but $D < n$).

Cai et al. (2015) also prove a minimax lower bound of $2D/n$ on MSE, with several interesting consequences. First, consistent information estimation is possible only if $D/n \rightarrow 0$. Second, since, for small x, $-\log(1 - x) \approx x$, this lower bound essentially matches the above upper bound when D/n is small. Third, they show this lower bound holds even when restricted to diagonal covariance matrices. Since the upper bound for the general case and the lower bound for the diagonal case essentially match, it follows that Gaussian information estimation is not made easier by structural assumptions such as Σ being bandable, sparse, or Toeplitz, as is common in, for example, stationary Gaussian process models (Cai & Yuan, 2012).

This $2D/n$ lower bound extends to our more general nonparanormal setting. However, we provide a minimax lower bound suggesting that the nonparanormal setting is strictly harder, in that optimal rates depend on Σ. Our results imply

---

[7]"Few" depends on s and n, but Kandasamy et al. (2015) suggest nonparametric estimators should only be used with D at most 4-6. Rey & Roth (2012) tried using several nonparametric information estimators on the *Communities and Crime* UCI data set ($n = 2195$, $D = 10$), but found all too unstable to be useful.

nonparanormal information estimation *does* become easier if Σ is assumed to be bandable or Toeplitz.

A closely related point is that known convergence rates for the fully nonparametric case require the density p to be bounded away from 0 or have particular tail behavior, due to singularity of the logarithm near 0 and resulting sensitivity of Shannon information-theoretic functionals to regions of low but non-zero probability. In contrast, Cai et al. (2015) need no lower-bound-type assumptions in the Gaussian case. In the nonparanormal case, we show *some* such condition is needed to prove a uniform rate, but a weaker condition, a positive lower bound on $\lambda_D(\Sigma)$, suffices.

The **main contributions** of this paper are the following:

1. We propose three estimators, $\widehat{I}_G$, $\widehat{I}_\rho$, and $\widehat{I}_\tau$,[8] for the mutual information of a nonparanormal distribution.
2. We prove upper bounds, of order $O(D^2/(\lambda_D^2(\Sigma)n))$ on the mean squared error of $\widehat{I}_\rho$, providing the first upper bounds for a nonparanormal information estimator. This bound suggests nonparanormal estimators scale far better with D than nonparametric estimators.
3. We prove a minimax lower bound suggesting that, unlike the Gaussian case, difficulty of nonparanormal information estimation depends on the true Σ.
4. We give simulations comparing our proposed estimators to Gaussian and nonparametric estimators. Besides confirming and augmenting our theoretical predictions, these help characterize the settings in which each nonparanormal estimator works best.
5. We present entropy estimators based on $\widehat{I}_G$, $\widehat{I}_\rho$, and $\widehat{I}_\tau$. Though nonparanormal entropy estimation requires somewhat different assumptions from mutual information estimation, we show that entropy can also be estimated at the rate $O(D^2/(\lambda_D^2(\Sigma)n))$.

4. Nonparanormal Information Estimators

In this section, we present three different estimators, I_G, I_ρ, and I_τ, for the mutual information of a nonparanormal distribution. We begin with a lemma providing common motivation for all three estimators.

Since mutual information is invariant to diffeomorphisms of individual variables, it is easy to see that the mutual information of a nonparanormal random variable is the same as that of the latent Gaussian random variable. Specifically:

Lemma 3. (Nonparanormal mutual information): *Suppose $X \sim \mathcal{NPN}(\Sigma; f)$. Then,*

$$I(X) = -\frac{1}{2}\log|\Sigma|. \qquad (3)$$

---

[8]Ince et al. (2017) proposed $\widehat{I}_G$ for use in neuroimaging data analysis. To the best of our knowledge, $\widehat{I}_\rho$ and $\widehat{I}_\tau$ are novel.

Lemma 3 shows that mutual information of a nonparanormal random variable depends only the latent covariance Σ; the marginal transformations are nuisance parameters, allowing us to avoid difficult nonparametric estimation; the estimators we propose all plug different estimates of Σ into Eq. (3), after a regularization step described in Section 4.3.

4.1. Estimating Σ by Gaussianization

The first estimator $\widehat{\Sigma}_G$ of Σ proceeds in two steps. First, the data are transformed to have approximately standard normal marginal distributions, a process Szabó et al. (2007b) referred to as "Gaussianization". By the nonparanormal assumption, the Gaussianized data are approximately jointly Gaussian. Then, the latent covariance matrix is estimated by the empirical covariance of the Gaussianized data.

More specifically, letting Φ^{-1} denote the quantile function of the standard normal distribution and recalling the rank matrix R defined in (2), the Gaussianized data

$$\widetilde{X}_{i,j} := \Phi^{-1}\left(\frac{R_{i,j}}{n+1}\right) \quad (\text{for } i \in [n], j \in [D])$$

are obtained by transforming the empirical CDF of the each dimension to approximate Φ. Then, we estimate Σ by the empirical covariance $\widehat{\Sigma}_G := \frac{1}{n}\sum_{i=1}^n \widetilde{X}_i \widetilde{X}_i^T$.

4.2. Estimating Σ by rank correlation

The second estimator actually has two variants, I_ρ and I_τ, respectively based on relating the latent covariance to two classic rank-based dependence measures, Spearman's ρ and Kendall's τ. For two random variables X and Y with CDFs $F_X, F_Y : \mathbb{R} \to [0,1]$, ρ and τ are defined by

$$\rho(X,Y) := \text{Corr}(F_X(X), F_Y(Y))$$

$$\text{and} \quad \tau(X,Y) := \text{Corr}(\text{sign}(X - X'), \text{sign}(Y - Y')),$$

respectively, where

$$\text{Corr}(X,Y) = \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}}$$

denotes the standard Pearson correlation operator and (X', Y') is an IID copy of (X, Y). ρ and τ generalize to the D-dimensional setting in the form of rank correlation matrices $\rho, \tau \in [-1,1]^{D \times D}$ with $\rho_{j,k} = \rho(X_j, X_k)$ and $\tau_{j,k} = \tau(X_j, X_k)$ for each $j, k \in [D]$.

I_ρ and I_τ are based on a classical result relating the correlation and rank-correlation of a bivariate Gaussian:

Theorem 4. (Kruskal, 1958): *Suppose (X,Y) has a Gaussian joint distribution with covariance Σ. Then,*

$$\text{Corr}(X,Y) = 2\sin\left(\frac{\pi}{6}\rho(X,Y)\right) = \sin\left(\frac{\pi}{2}\tau(X,Y)\right).$$

ρ and τ are often preferred over Pearson correlation for their relative robustness to outliers and applicability to non-numerical ordinal data. While these are strengths here as well, the main reason for their relevance is that they are invariant to marginal transformations (i.e., for diffeomorphisms $f, g : \mathbb{R} \to \mathbb{R}$, $\rho(f(X), g(Y)) = \pm\rho(X, Y)$ and $\tau(f(X), g(Y)) = \pm\tau(X, Y)$). As a consequence, the identity provided in Theorem 4 extends unchanged to the case $(X, Y) \sim \mathcal{NPN}(\Sigma; f)$. This suggests an estimate for Σ based on estimating ρ or τ and plugging this element-wise into the transform $x \mapsto 2\sin\left(\frac{\pi}{6}x\right)$ or $x \mapsto \sin\left(\frac{\pi}{2}x\right)$, respectively. Specifically, $\widehat{\Sigma}_\rho$ is defined by

$$\widehat{\Sigma}_\rho := 2\sin\left(\frac{\pi}{6}\widehat{\rho}\right), \quad \text{where} \quad \widehat{\rho} = \widehat{\mathrm{Corr}}(R)$$

is the empirical correlation of the rank matrix R, and sine is applied element-wise. Similarly, $\widehat{\Sigma}_\tau := \sin\left(\frac{\pi}{2}\widehat{\tau}\right)$, where

$$\widehat{\tau}_{j,k} := \frac{1}{\binom{n}{2}} \sum_{i \neq \ell \in [n]} \mathrm{sign}(X_{i,j} - X_{\ell,j})\,\mathrm{sign}(X_{i,k} - X_{\ell,k}).$$

4.3. Regularization and estimating I

Unfortunately, unlike usual empirical correlation matrices, none of $\widehat{\Sigma}_G$, $\widehat{\Sigma}_\rho$, or $\widehat{\Sigma}_\tau$ is almost surely strictly positive definite. As a result, directly plugging into the mutual information functional (3) may give ∞ or be undefined. To correct for this, we propose a regularization step, in which we project each estimated latent covariance matrix onto the (closed) cone $\mathcal{S}(z)$ of symmetric matrices with minimum eigenvalue $z > 0$. Specifically, for any $z > 0$, let

$$\mathcal{S}(z) := \left\{ A \in \mathbb{R}^{D \times D} : A = A^T, \lambda_D(A) \geq z \right\}.$$

For any symmetric matrix $A \in \mathbb{R}^{D \times D}$ with eigendecomposition $\widehat{\Sigma} = Q\Lambda Q^{-1}$ (i.e., $QQ^T = Q^TQ = I_D$ and Λ is diagonal), the projection A_z of A onto $\mathcal{S}(z)$ is defined as $A_z := Q\Lambda_z Q^{-1}$, where Λ_z is the diagonal matrix with j^{th} nonzero entry $(\Lambda_z)_{j,j} = \max\{z, \Lambda_{j,j}\}$. We call this a "projection" because $A_z = \mathrm{argmin}_{B \in \mathcal{S}(z)} \|A - B\|_F$ (see, e.g., Henrion & Malick (2012)).

Applying this regularization to $\widehat{\Sigma}_G$, $\widehat{\Sigma}_\rho$, or $\widehat{\Sigma}_\tau$ gives a strictly positive definite estimate $\widehat{\Sigma}_{G,z}$, $\widehat{\Sigma}_{\rho,z}$, or $\widehat{\Sigma}_{\tau,z}$, respectively, of Σ. We can then estimate I by plugging this into Equation (3), giving our three estimators:

$$\widehat{I}_{G,z} := -\frac{1}{2}\log\left|\widehat{\Sigma}_{G,z}\right|, \qquad \widehat{I}_{\rho,z} := -\frac{1}{2}\log\left|\widehat{\Sigma}_{\rho,z}\right|$$

$$\text{and} \quad \widehat{I}_{\tau,z} := -\frac{1}{2}\log\left|\widehat{\Sigma}_{\tau,z}\right|.$$

5. Upper Bounds on the Error of $\widehat{I}_{\rho,z}$

Here, we provide finite-sample upper bounds on the error of the estimator $\widehat{I}_\rho$ based on Spearman's ρ. Proofs are given in the Appendix. We first bound the bias of $\widehat{I}_\rho$:

Proposition 5. *Suppose* $X_1, ..., X_n \overset{i.i.d.}{\sim} \mathcal{NPN}(\Sigma; f)$. *Then, there exists a constant* $C > 0$ *such that, for any* $z > 0$, *the bias of* $\widehat{I}_{\rho,z}$ *is at most*

$$\left|\mathbb{E}\left[\widehat{I}_{\rho,z}\right] - I\right| \leq C\left(\frac{D}{z\sqrt{n}} + \log\frac{|\Sigma_z|}{|\Sigma|}\right),$$

where Σ_z *is the projection of* Σ *onto* $\mathcal{S}(z)$.

The first term of the bias stems from nonlinearity of the log-determinant function in Equation 3, which we analyze via Taylor expansion. The second term,

$$\log\frac{|\Sigma_z|}{|\Sigma|} = \sum_{\lambda_j(\Sigma) < z} \log\left(\frac{z}{\lambda_j(\Sigma)}\right),$$

is due to the regularization step and is actually exact, but is difficult to simplify or bound without more assumptions on the spectrum of Σ and choice of z, which we discuss later. We now turn to bounding the variance of $\widehat{I}_{\rho,z}$. We first provide an exponential concentration inequality for $\widehat{I}_{\rho,z}$ around its expectation, based on McDiarmid's inequality:

Proposition 6. *Suppose* $X_1, ..., X_n \overset{i.i.d.}{\sim} \mathcal{NPN}(\Sigma; f)$. *Then, for any* $z, \varepsilon > 0$,

$$\mathbb{P}\left[\left|\widehat{I}_{\rho,z} - \mathbb{E}\left[\widehat{I}_{\rho,z}\right]\right| > \varepsilon\right] \leq 2\exp\left(-\frac{nz^2\varepsilon^2}{18\pi^2 D^2}\right).$$

Such exponential concentration bounds are useful when one wants to simultaneously bound the error of multiple uses of an estimator, and hence we present it separately as it may be independently useful. However, for the purpose of understanding convergence rates, we are more interested in the variance bound that follows as an easy corollary:

Corollary 7. *Suppose* $X_1, ..., X_n \overset{i.i.d.}{\sim} \mathcal{NPN}(\Sigma; f)$. *Then, for any* $z > 0$, *the variance of* $\widehat{I}_{\rho,z}$ *is at most*

$$\mathbb{V}\left[\widehat{I}_{\rho,z}\right] \leq \frac{36\pi^2 D^2}{z^2 n}.$$

Given these bias and variance bounds, a bound on the MSE of $\widehat{I}_{\rho,z}$ follows via the usual bias-variance decomposition:

Theorem 8. *Suppose* $X \sim \mathcal{NPN}(\Sigma; f)$. *Then, there exists a constant* C *such that*

$$\mathbb{E}\left[\left(\widehat{I}_{\rho,z} - I\right)^2\right] \leq C\left(\frac{D^2}{z^2 n} + \log^2\frac{|\Sigma_z|}{|\Sigma|}\right). \quad (4)$$

A natural question is now how to optimally select the regularization parameter z. While the bound (4) is clearly convex in z, it depends crucially on the unknown spectrum of Σ, and, in particular, on the smallest eigenvalues of Σ. As a result, it is difficult to choose z optimally in general, but we can do so for certain common subclasses of covariance matrices. For example, if Σ is Toeplitz or bandable (i.e., for some $c \in (0,1)$, all $|\Sigma_{i,j}| \leq c^{|i-j|}$), then the smallest eigenvalue of Σ can be bounded below (Cai & Yuan, 2012). When Σ is bandable, as we show in the Appendix, this bound can be independent of D. In these cases, the following somewhat simpler MSE bound can be used:

Corollary 9. *Suppose* $X \sim \mathcal{NPN}(\Sigma; f)$, *and suppose* $z \leq \lambda_D(\Sigma)$. *Then, there exists a constant* $C > 0$ *such that*

$$\mathbb{E}\left[\left(\widehat{I}_{\rho,z} - I\right)^2\right] \leq \frac{CD^2}{z^2 n}.$$

6. Lower Bounds in terms of Σ

If $X_1, ..., X_n \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma)$ are Gaussian, for the plug-in estimator

$$\widehat{I} = -\tfrac{1}{2}\log\left|\widehat{\Sigma}\right| \quad \text{(where} \quad \widehat{\Sigma} = \tfrac{1}{n}\sum_{i=1}^n X_i X_i^T$$

is the empirical covariance matrix), Cai et al. (2015) showed that the distribution of $\widehat{I} - I$ is independent of the true correlation matrix Σ. This follows from the "stability" of Gaussians (i.e., that nonsingular linear transformations of Gaussian random variables are Gaussian). In particular,

$$\widehat{I} - I = \log|\widehat{\Sigma}| - \log|\Sigma| = \log|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}|,$$

and $\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}$ has the same distribution as $\log\widehat{\Sigma}$ does in the special case that $\Sigma = I_D$ is the identity. This property is both somewhat surprising, given that $I \to \infty$ as $|\Sigma| \to 0$, and useful, leading to a tight analysis of the error of $\widehat{I}$ and confidence intervals that do not depend on Σ.

It would be convenient if any nonparanormal information estimators satisfied this property. Unfortunately, the main result of this section is a negative one, showing that this property is unlikely to hold without additional assumptions:

Proposition 10. *Consider the 2-dimensional case*

$$X_1, ..., X_n \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma), \quad with \quad \Sigma = \begin{bmatrix} 1 & \sigma \\ \sigma & 1 \end{bmatrix}, \quad (5)$$

and let $\sigma_* \in (0,1)$. *Suppose an estimator* $\widehat{I} = \widehat{I}(R)$ *of* $I_\sigma = -\tfrac{1}{2}\log(1 - \sigma^2)$ *is a function of the empirical rank matrix* $R \in \mathbb{N}^{n \times 2}$ *of* X. *Then, there exists a constant* $C > 0$, *depending only* n, *such that the worst-case MSE of* $\widehat{I}$ *over* $\sigma \in (0, \sigma_*)$ *satisfies*

$$\sup_{\sigma \in (0,\sigma^*)} \mathbb{E}\left[\left(\widehat{I}(R) - I_\sigma\right)^2\right] \geq \frac{1}{64}\left(C - \log(1 - \sigma_*^2)\right)^2$$

Clearly, this lower bound tends to ∞ as $\sigma \to 1$. As written, this result lower bounds the error of *rank-based estimators* in the Gaussian case when $\sigma \approx 1$. However, to the best of our knowledge, all methods for estimating Σ in the nonparanormal case are functions of R, and prior work (Hoff, 2007) has shown that the rank matrix R is a generalized sufficient statistic for Σ (and hence for I) in the nonparanormal model. Thus, it is reasonable to think of lower bounds for rank-based estimators in the Gaussian case as lower bounds for any estimator in the nonparanormal case.

The proof of this result is based on the simple observation that the rank matrix can take only finitely many values. Hence, as $\sigma \to 1$, R tends to be perfectly correlated, providing little information about σ, whereas the dependence of the estimand I_σ on σ increases sharply. This is intuition is formalized in the Appendix using Le Cam's lemma for lower bounds in two-point parameter estimation problems.

7. Empirical Results

We compare 5 mutual information estimators:

- $\widehat{I}$: Gaussian plug-in estimator with bias-correction (see Cai et al. (2015)).

- $\widehat{I}_G$: Nonparanormal estimator using Gaussianization.

- $\widehat{I}_\rho$: Nonparanormal estimator using Spearman's ρ.

- $\widehat{I}_\tau$: Nonparanormal estimator using Kendall's τ.

- $\widehat{I}_{kNN}$: Nonparametric estimator using k-nearest neighbor (kNN) statistics.

For I_ρ and I_τ, we used a regularization constant $z = 10^{-3}$. We did not regularize for I_G. Although this implies $\mathbb{P}[I_G = \infty] > 0$, this is extremely unlikely for even moderate values of n and never occurred during our experiments, which all use $n \geq 32$. We thus omit denoting dependence on z. For I_{kNN}, except as noted in Experiment 3, $k = 2$, based on recent analysis (Singh & Póczos, 2016b) suggesting that small values of k are best for estimation.

Sufficient details to reproduce experiments are given in the Appendix, and MATLAB source code is available at [Omitted for anonymity]. We report MSE based on 1000 i.i.d. trials of each condition. 95% confidence intervals were consistently smaller than plot markers and hence omitted to avoid cluttering plots. Except as specified otherwise, each experiment had the following basic structure: In each trial, a correlation matrix Σ was drawn by normalizing a random covariance matrix from a Wishart distribution, and data $X_1, ..., X_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$ drawn. All 5 estimators were computed from $X_1, ..., X_n$ and squared error from true mutual information (computed from Σ) was recorded. Unless specified otherwise, $n = 100$ and $D = 25$.

Since our nonparanormal information estimators are func-

tions of ranks of the data, neither the true mutual information nor our non-paranormal estimators depend on the marginal transformations. Thus, except in Experiment 2, where we show the effects of transforming marginals, and Experiment 3, where we add outliers to the data, we perform all experiments on truly Gaussian data, with the understanding that this setting favors the Gaussian estimator.

All experimental results are displayed in Figure 1.

Experiment 1 (Dependence on n): We first show non-paranormal estimators have "parametric" $O(n^{-1})$ dependence on n, unlike $\widehat{I}_{k\mathrm{NN}}$, which converges far more slowly. For large n, MSEs of $\widehat{I}_G$, $\widehat{I}_\rho$, and $\widehat{I}_\tau$ are close to that of $\widehat{I}$.

Experiment 2 (Non-Gaussian Marginals): Next, we show nonparanormal estimators are robust to non-Gaussianity of the marginals, unlike $\widehat{I}$. We applied a nonlinear transformation f to a fraction $\alpha \in [0,1]$ of dimensions of Gaussian data. That is, we drew $Z_1, ..., Z_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$ and then used data $X_1, ..., X_n$, where

$$X_{i,j} = \begin{cases} T(Z_{i,j}) & \text{if } j < \alpha D \\ Z_{i,j} & \text{if } j \geq \alpha D \end{cases}, \quad \forall i \in [n], j \in [D],$$

for a diffeomorphism T. Here, we use $T(z) = e^z$. The Appendix shows similar results for several other T. $\widehat{I}$ performs poorly even when α is quite small. Poor performance of $\widehat{I}_{k\mathrm{NN}}$ may be due to discontinuity of the density at $x = 0$.

Experiment 3 (Outliers): We now show that nonparanormal estimators are far more robust to the presence of outliers than $\widehat{I}$ or $\widehat{I}_{k\mathrm{NN}}$. To do this, we added outliers to the data according to the method of Liu et al. (2012). After drawing Gaussian data, we independently select $\lfloor \beta n \rfloor$ samples in each dimension, and replace each i.i.d. uniformly at random from $\{-5, +5\}$. Performance of $\widehat{I}$ degrades rapidly even for small β. $\widehat{I}_{k\mathrm{NN}}$ can fail for atomic distributions, $\widehat{I}_{k\mathrm{NN}} = \infty$ whenever at least k samples are identical. This mitigate this, we increased k to 20 and ignored trials where $\widehat{I}_{k\mathrm{NN}} = \infty$, but $\widehat{I}_{k\mathrm{NN}}$ ceased to give any finite estimates when β was sufficiently large.

For small values of β, nonparanormal estimators surprisingly improve. We hypothesize this is due to convexity of the mutual information functional Eq. (3) in Σ. By Jensen's inequality, estimators which plug-in an approximately unbiased estimate $\widehat{\Sigma}$ of Σ are biased towards overestimating I. Adding random (uncorrelated) noise reduces estimated dependence, moving the estimate closer to the true value. If this nonlinearity is indeed a major source of bias, it may be possible to derive a von Mises-type bias correction (see Kandasamy et al. (2015)) accounting for higher-order terms in the Taylor expansion of the log-determinant.

Experiment 4 (Dependence on Σ): Here, we verify our results in Section 6 showing that MSE of rank-based estima-

tors approaches ∞ as $|\Sigma| \to 0$, while MSE of $\widehat{I}$ is independent of Σ. Here, we set $D = 2$ and Σ as in Eq. (5), varying $\sigma \in [0, 1]$. Indeed, the MSE of $\widehat{I}$ does not change, while the MSEs of $\widehat{I}_G$, $\widehat{I}_\rho$, and $\widehat{I}_\tau$ all increase as $\sigma \to 1$. This increase seems mild in practice, with performance worse than of $\widehat{I}$ only when $\sigma > 0.99$. $\widehat{I}_\tau$ appears to perform far better than $\widehat{I}_G$ and $\widehat{I}_\rho$ in this regime. Performance of $I_{k\mathrm{NN}}$ degrades far more quickly as $\sigma \to 1$. This phenomenon is explored by Gao et al. (2015), who lower bound error of $I_{k\mathrm{NN}}$ in the presence of strong dependencies, and proposed a correction to improve performance in this case.

It is also interesting that errors of $\widehat{I}_\rho$ and $\widehat{I}_\tau$ drop as $\sigma \to 0$. This is likely because, for small σ, the main source of error is the variance of $\widehat{\rho}$ and $\widehat{\tau}$ (as $-\log(1 - \sigma^2) \approx \sigma^2$ when $\sigma \approx 0$). When $n \to \infty$ and D is fixed, both $2 \sin(\pi \widehat{\rho}/6)$ and $\sin(\pi \widehat{\tau}/2)$ are asymptotically normal estimates of σ, with asymptotic variances proportional to $(1 - \sigma^2)^2$ (Klaassen & Wellner, 1997). By the delta method, since $\frac{dI}{d\sigma} = \frac{\sigma}{1 - \sigma^2}$, $\widehat{I}_\rho$ and $\widehat{I}_\tau$ are asymptotically normal estimates of I, with asymptotic variances proportional to σ^2 and hence vanishing as $\sigma \to 0$.

8. Estimating Entropy

Thus far, we have discussed estimation of mutual information $I(X)$. Mutual information is convenient because it is invariant under marginal transformation, and hence $I(X) = I(f(X))$ depends only on Σ. While the entropy $H(X)$ does depend on the marginal transform f, fortunately, by Eq. (1), $H(X)$ differs from $I(X)$ only by a sum of univariate entropies. Univariate nonparametric estimation of entropy in has been studied extensively, and there exist several estimators (e.g., based on sample spacings (Beirlant et al., 1997), kernel density estimates (Moon et al., 2016) or k-nearest neighbor methods (Singh & Póczos, 2016b)) that can estimate $H(X_j)$ at the rate $\asymp n^{-1}$ in MSE under relatively mild conditions on the marginal density p_j. While the precise assumptions vary with the choice of estimator, they are mainly (a) that p_j be lower bounded on its support or have particular (e.g., exponential) tail behavior, and (b) that p_j be smooth, typically quantified by a Hölder or Sobolev condition. Details of these assumptions are in the Appendix.

Under these conditions, since there exist estimators $\widehat{H}_1, ..., \widehat{H}_D$ and a constant $C > 0$ such that

$$\mathbb{E}[(\widehat{H}_j - H(X_j))^2] \leq C/n, \quad \forall j \in [D]. \qquad (6)$$

Combining these estimators with an estimator, say $\widehat{I}_{\rho, z}$, of mutual information gives an estimator of entropy:

$$\widehat{H}_{\rho, z} := \sum_{j=1}^D \widehat{H}_j - \widehat{I}_{\rho, z}.$$

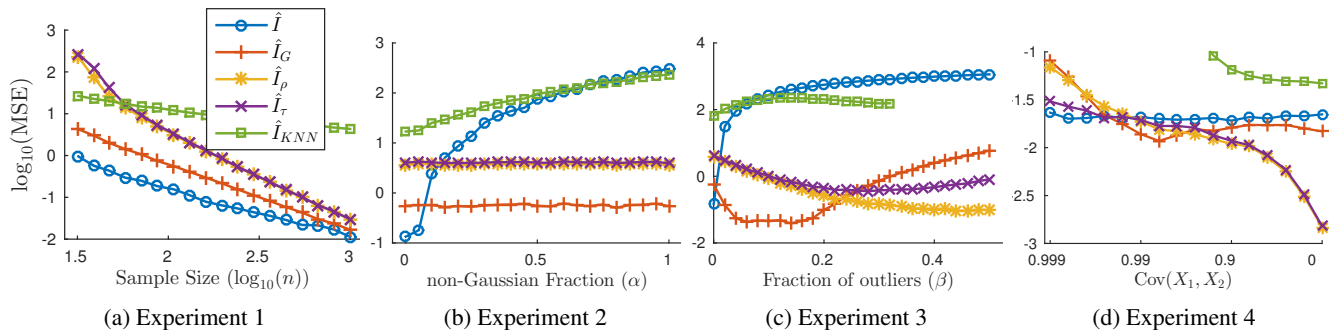If we assume $z = \lambda_D^{-1}(\Sigma)$ is bounded below by a positive

Figure 1. Plots of $\log_{10}(\text{MSE})$ plotted over (a) log-sample-size $\log_{10}(n)$, (b) fraction α of dimensions with non-Gaussian marginals, (c) fraction β of outlier samples in each dimension, and (d) covariance $\Sigma_{1,2} = \text{Cov}(X_1, X_2)$. Note that the x-axis in (d) is decreasing.

constant, combining inequality (6) with Corollary 9 gives

$$\mathbb{E}\left[\left(\widehat{H}_{\rho,z} - H(X)\right)^2\right] \leq \frac{CD^2}{n},$$

where C differs from in (6) but is independent of n and D.

9. Conclusions and Future Work

This paper suggests nonparanormal information estimation as a practical compromise between the intractable nonparametric case and the limited Gaussian case. We proposed three estimators for this problem and provided the first upper bounds for nonparanormal information estimation. We also gave lower bounds showing how dependence on Σ differs from the Gaussian case and demonstrated empirically that nonparanormal estimators are more robust than Gaussian estimators, even in dimensions too high for nonparametric estimators.

Collectively, these results suggest that, by scaling to moderate or high dimensionality without relying on Gaussianity, nonparanormal information estimators may be effective tools with a number of machine learning applications. While the best choice of information estimator inevitably depends on context, as an off-the-shelf guide for practitioners, the estimators we suggest, in order of preference, are:

- fully nonparametric if $D < 6, n > \max\{100, 10^D\}$.
- $\widehat{I}_\rho$ if D^2/n is small and data may have outliers.
- $\widehat{I}_\tau$ if D^2/n is small and dependencies may be strong.
- $\widehat{I}_G$ otherwise.
- $\widehat{I}$ only given strong belief that data are nearly Gaussian.

There are many natural open questions in this line of work. First, in the nonparanormal model, we focused on estimating mutual information $I(X)$, which does not depend on marginal transforms f, and entropy, which decomposes into $I(X)$ and 1-dimensional entropies. In both cases, additional structure imposed by the nonparanormal model allows estimation in higher dimensions than fully nonparametric models. Can nonparanormal assumptions lead to

higher dimensional estimators for the many other useful nonlinear functionals of densities (e.g., L_p norms/distances and more general (e.g., Rényi or Tsallis) entropies, mutual informations, and divergences) that do not decompose?

Second, there is a gap between our upper bound rate of $\|\Sigma^{-1}\|_2^2 D^2/n$ and the only known lower bound of $2D/n$ (from the Gaussian case), though we also showed that bounds for rank-based estimators depend on Σ. Is quadratic dependence on D optimal? How much do rates improve under structural assumptions on Σ? Upper bounds should be derived for other estimators, such as $\widehat{I}_G$ and $\widehat{I}_\tau$. The $2D/n$ lower bound proof of Cai et al. (2015) for the Gaussian case, based on the Cramer-Rao inequality (Van den Bos, 2007), is unlikely to tighten in the nonparanormal case, since Fisher information is invariant to diffeomorphisms of the data. Hence, a new approach is needed if the lower bound in the nonparanormal case is to be raised.

Finally, our work applies to estimating the log-determinant $\log|\Sigma|$ of the latent correlation in a nonparanormal model. Besides information estimation, the work of Cai et al. (2015) on estimating $\log|\Sigma|$ in the Gaussian model was motivated by the role of $\log|\Sigma|$ in other multivariate statistical tools, such as quadratic discriminant analysis (QDA) and MANOVA (Anderson, 1984). Can our estimators lead to more robust nonparanormal versions of these tools?

References

Adami, Christoph. Information theory in molecular biology. *Physics of Life Reviews*, 1(1):3–22, 2004.

Aghagolzadeh, Mehdi, Soltanian-Zadeh, Hamid, Araabi, B, and Aghagolzadeh, Ali. A hierarchical clustering based on mutual information maximization. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, volume 1, pp. I–277. IEEE, 2007.

Aghakouchak, Amir. Entropy–copula in hydrology and climatology. *J. Hydrometeorology*, 15(6):2176–2189, 2014.

Ahmed, Nabil Ali and Gokhale, DV. Entropy expressions and their estimators for multivariate distributions. *IEEE Trans. on Information Theory*, 35(3):688–692, 1989.

Anderson, TW. Multivariate statistical analysis. *Wiley and Sons, New York, NY*, 1984.

Beirlant, Jan, Dudewicz, Edward J, Györfi, László, and Van der Meulen, Edward C. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

Bercher, J-F and Vignat, Christophe. Estimating the entropy of a signal with applications. *IEEE Trans. on Signal Processing*, 48 (6):1687–1694, 2000.

Bereziński, Przemysław, Jasiul, Bartosz, and Szpyrka, Marcin. An entropy-based network anomaly detection method. *Entropy*, 17(4):2367–2408, 2015.

Berkes, Pietro, Wood, Frank, and Pillow, Jonathan W. Characterizing neural dependencies with copula models. In *Advances in Neural Information Processing Systems*, pp. 129–136, 2009.

Bickel, Peter J and Levina, Elizaveta. Regularized estimation of large covariance matrices. *Annals of Stat.*, pp. 199–227, 2008.

Birgé, Lucien and Massart, Pascal. Estimation of integral functionals of a density. *Annals of Stat.*, pp. 11–29, 1995.

Cai, T Tony and Yuan, Ming. Adaptive covariance matrix estimation through block thresholding. *Annals of Stat.*, 40(4): 2014–2042, 2012.

Cai, T Tony, Liang, Tengyuan, and Zhou, Harrison H. Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions. *J. of Multivariate Analysis*, 137:161–172, 2015.

Calsaverini, Rafael S and Vicente, Renato. An information-theoretic approach to statistical dependence: Copula information. *EPL (Europhysics Letters)*, 88(6):68003, 2009.

Chai, Barry, Walther, Dirk, Beck, Diane, and Fei-Fei, Li. Exploring functional connectivities of the human brain using multivariate information analysis. In *Advances in neural information processing systems*, pp. 270–278, 2009.

Chow, C and Liu, Cong. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

Efron, Bradley and Stein, Charles. The jackknife estimate of variance. *The Annals of Statistics*, pp. 586–596, 1981.

Elidan, Gal. Copulas in machine learning. In *Copulae in mathematical and quantitative finance*, pp. 39–60. Springer, 2013.

Friedman, Jerome H and Stuetzle, Werner. Projection pursuit regression. *JASA*, 76(376):817–823, 1981.

Gao, Shuyang, Ver Steeg, Greg, and Galstyan, Aram. Efficient estimation of mutual information for strongly dependent variables. In *AISTATS*, 2015.

Garner, Wendell R. Uncertainty and structure as psychological concepts. 1962.

Gershgorin, Semyon Aranovich. Uber die abgrenzung der eigenwerte einer matrix. (6):749–754, 1931.

Goria, Mohammed Nawaz, Leonenko, Nikolai N, Mergel, Victor V, and Novi Inverardi, Pier Luigi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Nonparametric Statistics*, 17(3):277–297, 2005.

Han, Insu, Malioutov, Dmitry, and Shin, Jinwoo. Large-scale log-determinant computation through stochastic chebyshev expansions. In *ICML*, pp. 908–917, 2015.

Henrion, Didier and Malick, Jérôme. Projection methods in conic optimization. In *Handbook on Semidefinite, Conic and Polynomial Optimization*, pp. 565–600. Springer, 2012.

Hernández-Lobato, José Miguel, Lloyd, James R, and Hernández-Lobato, Daniel. Gaussian process conditional copulas with applications to financial time series. In *Advances in Neural Information Processing Systems*, pp. 1736–1744, 2013.

Hero, Alfred O, Ma, Bing, Michel, Olivier, and Gorman, John. Alpha-divergence for classification, indexing and retrieval (revised). 2001.

Hero, Alfred O, Ma, Bing, Michel, Olivier JJ, and Gorman, John. Applications of entropic spanning graphs. *IEEE Signal Processing Magazine*, 19(5):85–95, 2002.

Hoff, Peter D. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, pp. 265–283, 2007.

Hunter, Jacob S and Hodas, Nathan O. Mutual information for fitting deep nonlinear models. *arXiv preprint arXiv:1612.05708*, 2016.

Ince, Robin AA, Giordano, Bruno L, Kayser, Christoph, Rousselet, Guillaume A, Gross, Joachim, and Schyns, Philippe G. A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Human brain mapping*, 38(3):1541–1573, 2017.

Kandasamy, Kirthevasan, Krishnamurthy, Akshay, Poczos, Barnabas, Wasserman, Larry, and Robins, James M. Nonparametric von mises estimators for entropies, divergences and mutual informations. In *Advances in Neural Information Processing Systems*, pp. 397–405, 2015.

Klaassen, Chris AJ and Wellner, Jon A. Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, 3(1):55–77, 1997.

Krishnamurthy, Akshay, Kandasamy, Kirthevasan, Poczos, Barnabas, and Wasserman, Larry. Nonparametric estimation of renyi divergence and friends. In *International Conference on Machine Learning*, pp. 919–927, 2014.

Kruskal, William H. Ordinal measures of association. *JASA*, 53 (284):814–861, 1958.

Learned-Miller, E. G. and Fisher, J. W. ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, 4: 1271–1295, 2003.

Liu, Han, Han, Fang, Yuan, Ming, Lafferty, John, and Wasserman, Larry. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

Ma, Jian and Sun, Zengqi. Mutual information is copula entropy. *Tsinghua Science & Tech.*, 16(1):51–54, 2011.

Malevergne, Yannick and Sornette, Didier. Testing the gaussian copula hypothesis for financial assets dependences. *Quantitative Finance*, 3(4):231–250, 2003.

Misra, Neeraj, Singh, Harshinder, and Demchuk, Eugene. Estimation of the entropy of a multivariate normal distribution. *J. Multivariate Analysis*, 92(2):324–342, 2005.

Mitra, Ritwik and Zhang, Cun-Hui. Multivariate analysis of nonparametric estimates of large correlation matrices. *arXiv preprint arXiv:1403.6195*, 2014.

Moon, Kevin and Hero, Alfred. Multivariate f-divergence estimation with confidence. In *Advances in Neural Information Processing Systems*, pp. 2420–2428, 2014a.

Moon, Kevin R and Hero, Alfred O. Ensemble estimation of multivariate f-divergence. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 356–360. IEEE, 2014b.

Moon, Kevin R, Sricharan, Kumar, Greenewald, Kristjan, and Hero, Alfred O. Improving convergence of divergence functional ensemble estimators. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 1133–1137. IEEE, 2016.

Moon, Kevin R, Sricharan, Kumar, and Hero III, Alfred O. Ensemble estimation of mutual information. *arXiv preprint arXiv:1701.08083*, 2017.

Noble, Caleb C and Cook, Diane J. Graph-based anomaly detection. In *KDD*, pp. 631–636. ACM, 2003.

Nychis, George, Sekar, Vyas, Andersen, David G, Kim, Hyong, and Zhang, Hui. An empirical evaluation of entropy-based traffic anomaly detection. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, pp. 151–156. ACM, 2008.

Pál, Dávid, Póczos, Barnabás, and Szepesvári, Csaba. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Advances in Neural Information Processing Systems*, pp. 1849–1857, 2010.

Peng, Hanchuan, Long, Fuhui, and Ding, Chris. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

Renard, Benjamin and Lang, Michel. Use of a Gaussian copula for multivariate extreme value analysis: some case studies in hydrology. *Advances in Water Resources*, 30(4):897–912, 2007.

Rey, Mélanie and Roth, Volker. Meta-Gaussian information bottleneck. In *Advances in Neural Information Processing Systems*, pp. 1916–1924, 2012.

Shan, Caifeng, Gong, Shaogang, and McOwan, Peter W. Conditional mutual infomation based boosting for facial expression recognition. In *BMVC*, 2005.

Shishkin, Alexander, Bezzubtseva, Anastasia, Drutsa, Alexey, Shishkov, Ilia, Gladkikh, Ekaterina, Gusev, Gleb, and Serdyukov, Pavel. Efficient high-order interaction-aware feature selection based on conditional mutual information. In *Advances in Neural Information Processing Systems*, pp. 4637–4645, 2016.

Singh, Shashank and Póczos, Barnabás. Exponential concentration of a density functional estimator. In *Advances in Neural Information Processing Systems*, pp. 3032–3040, 2014a.

Singh, Shashank and Póczos, Barnabás. Generalized exponential concentration inequality for Rényi divergence estimation. In *ICML*, pp. 333–341, 2014b.

Singh, Shashank and Póczos, Barnabás. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. In *Advances in Neural Information Processing Systems*, pp. 1217–1225, 2016a.

Singh, Shashank and Póczos, Barnabás. Analysis of k-nearest neighbor distances with application to entropy estimation. *arXiv preprint arXiv:1603.08578*, 2016b.

Sricharan, Kumar, Raich, Raviv, and Hero, Alfred O. k-nearest neighbor estimation of entropies with confidence. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 1205–1209. IEEE, 2011.

Sricharan, Kumar, Wei, Dennis, and Hero, Alfred O. Ensemble estimators for multivariate entropy estimation. *IEEE Transactions on Information Theory*, 59(7):4374–4388, 2013.

Srivastava, Santosh and Gupta, Maya R. Bayesian estimation of the entropy of the multivariate Gaussian. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 1103–1107. IEEE, 2008.

Studenỳ, Milan and Vejnarová, Jirina. The multiinformation function as a tool for measuring stochastic dependence. In *Learning in graphical models*, pp. 261–297. Springer, 1998.

Szabó, Zoltán, Póczos, Barnabás, and Lőrincz, András. Undercomplete blind subspace deconvolution. *Journal of Machine Learning Research*, 8(May):1063–1095, 2007a.

Szabó, Zoltán, Póczos, Barnabás, Szirtes, Gábor, and Lőrincz, András. Post nonlinear independent subspace analysis. In *International Conference on Artificial Neural Networks*, pp. 677–686. Springer, 2007b.

Tsybakov, A.B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, 1st edition, 2008.

Van den Bos, Adriaan. *Parameter estimation for scientists and engineers*. John Wiley & Sons, 2007.

Varga, Richard S. *Matrix Iterative Analysis*, volume 27. Springer Science & Business Media, 2009.

Watanabe, Satosi. Information theoretical analysis of multivariate correlation. *IBM J. of research and development*, 4(1):66–82, 1960.

Wilson, Andrew and Ghahramani, Zoubin. Copula processes. In *Advances in Neural Information Processing Systems*, pp. 2460–2468, 2010.

Wolsztynski, E., Thierry, E., and Pronzato, L. Minimum-entropy estimation in semi-parametric models. *Signal Process.*, 85(5): 937–949, 2005. ISSN 0165-1684.

A. Lemmas

Our proofs rely on the following lemmas.

Lemma 11. (Convexity of the inverse operator norm): *The function $A \mapsto \|A^{-1}\|_2$ is convex over $A \succ 0$.*

Proof: For $A, B \succ 0$, let $C := \tau A + (1 - \tau)B$. Then,

$$
\begin{aligned}
\|\widehat{C}^{-1}\|_2 &= \frac{1}{\inf_{x \in \mathbb{R}^D} x^T C x} \\
&= \frac{1}{\inf_{x \in \mathbb{R}^D} \tau x^T A x + (1 - \tau)x^T B x} \\
&\leq \frac{1}{\tau \inf_{x \in \mathbb{R}^D} x^T A x + (1 - \tau)\inf_{x \in \mathbb{R}^D} x^T B x} \\
&\leq \tau \frac{1}{\inf_{x \in \mathbb{R}^D} x^T A x} + (1 - \tau)\frac{1}{\inf_{x \in \mathbb{R}^D} x^T B x} \\
&= \tau \left\|A^{-1}\right\|_2 + (1 - \tau)\left\|B^{-1}\right\|_2
\end{aligned}
$$

via convexity of the function $x \mapsto 1/x$ on $(0, \infty)$. ∎

Lemma 12. (Mean-Value Bound on the Log-Determinant): *Matrix derivative of log-determinant. Suppose $A, B \succ 0$. Then, for $\lambda := \min\{\lambda_D(A), \lambda_D(B)\}$,*

$$
|\log |A| - \log |B|| \leq \frac{1}{\lambda}\|A - B\|_F.
$$

Proof: *Proof:* First recall that the log-determinant is continuously differentiable over the strict positive definite cone, with $\nabla_X \log |X| = X^{-1}$ for any $X \succ 0$. Hence, by the matrix-valued version of the mean value theorem,

$$
\log |A| - \log |B| = tr(C^{-1}(A - B)),
$$

where $C = \tau A + (1 - \tau)B$ for some $\tau \in (0, 1)$. Since for positive definite matrices, the inner product can be bounded by the product of the operator and Frobenius norms, and clearly $C \succ 0$, we have

$$
|\log |A| - \log |B|| = \|C^{-1}\|_2 \|A - B\|_F.
$$

Finally, it follows by Lemma 11 that

$$
|\log |A| - \log |B|| \leq \frac{1}{\lambda}\|A - B\|_F.
$$

∎

B. Proofs of Main Results

Here, we give proofs of our main theoretical results, beginning with upper bounds on the MSE of $\widehat{I}_\rho$ and proceeding to minimax lower bounds in terms of Σ.

C. Upper bounds on the MSE of $\widehat{I}_\rho$

Proposition 13.

$$
\begin{aligned}
&\left|\mathbb{E}\left[\log |\widehat{\Sigma}_z|\right] - \log |\Sigma|\right| \\
&\leq C\left(\|\Sigma\|_2^2 \frac{D}{z^2 n} + \left(\sum_{\lambda_j(\Sigma) < z} \log\left(\frac{z}{\lambda_j(\Sigma)}\right)\right)^2\right).
\end{aligned}
$$

Proof: By the triangle inequality,

$$
\begin{aligned}
\left|\mathbb{E}\left[\log |\widehat{\Sigma}_z|\right] - \log |\Sigma|\right| &\leq \left|\mathbb{E}\left[\log |\widehat{\Sigma}_z|\right] - \log |\Sigma_z|\right| \\
&+ |\log |\Sigma_z| - \log |\Sigma||
\end{aligned}
$$

For the first term, applying the matrix mean value theorem (Lemma 12) and the inequality $\|A\|_F \leq \sqrt{D}\|A\|_2$

$$
\begin{aligned}
\left|\mathbb{E}\left[\log\left|\widehat{\Sigma}_z\right|\right] - \log |\Sigma_z|\right| &\leq \mathbb{E}\left[\left|\log\left|\widehat{\Sigma}_z\right| - \log |\Sigma_z|\right|\right] \\
&\leq \frac{1}{z}\mathbb{E}\left[\left\|\widehat{\Sigma}_z - \Sigma_z\right\|_F\right] \\
&\leq \frac{\sqrt{D}}{z}\mathbb{E}\left[\left\|\widehat{\Sigma}_z - \Sigma_z\right\|_2\right] \\
&\leq \frac{C_{MZ}\|\Sigma\|_2 D}{z\sqrt{n}},
\end{aligned}
$$

where we used Theorem 1 of Mitra & Zhang (2014), which gives a constant C_{MZ} such that

$$
\mathbb{E}\left[\left\|\widehat{\Sigma}_z - \Sigma_z\right\|_2\right] \leq C_{MZ}\|\Sigma\|_2 \sqrt{\frac{D}{n}}.
$$

Via the bound $\|\Sigma\|_2 \leq \sqrt{D}\|\Sigma\|_\infty$, this reduces to

$$
\mathbb{E}\left[\left\|\widehat{\Sigma}_z - \Sigma_z\right\|_2\right] \leq C_{MZ}\frac{D}{\sqrt{n}}.
$$

∎

Proposition 14.

$$
\mathbb{V}\left[\widehat{I}\right] \leq \frac{36\pi^2 D^2}{z^2 n}.
$$

Proof: By the Efron-Stein inequality (Efron & Stein, 1981), since $X_1, \ldots, X_n$ are independent and identically

distributed,

$$\mathbb{V}\left[\widehat{I}\right] \le \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\left(\log|\widehat{\Sigma}_z| - \log|\widehat{\Sigma}_z^{(i)}|\right)^2\right]$$

$$= \frac{n}{2}\mathbb{E}\left[\left(\log|\widehat{\Sigma}_z| - \log|\widehat{\Sigma}_z^{(1)}|\right)^2\right],$$

where $\widehat{\Sigma}_z^{(1)}$ is our estimator after independently re-sampling the first sample X_1. Applying the multivariate mean-value theorem (Lemma 12), we have

$$\left|\log|\widehat{\Sigma}_z| - \log|\widehat{\Sigma}_z^{(1)}|\right| \le \frac{1}{z}\|\widehat{\Sigma}_z - \widehat{\Sigma}_z^{(1)}\|_F.$$

$\|\widehat{\Sigma}_\tau^{-1}\|_2 \le \frac{1}{z}$. Since $\mathcal{S}(z)$ is convex and the Frobenius norm is supported by an inner product, the operation of projecting onto $\mathcal{S}(z)$ is a contraction. In particular, $\left\|\left(\widehat{\Sigma}_z - \widehat{\Sigma}_z^{(1)}\right)\right\|_F \le \left\|\left(\widehat{\Sigma} - \widehat{\Sigma}^{(1)}\right)\right\|_F$ Applying the mean value theorem to the function $x \mapsto 2\sin\left(\frac{\pi}{6}x\right)$,

$$\left\|\left(\widehat{\Sigma} - \widehat{\Sigma}^{(1)}\right)\right\|_F^2 = \sum_{j,k=1}^{D}\left(\widehat{\Sigma} - \widehat{\Sigma}^{(1)}\right)_{j,k}^2 \quad (7)$$

$$\le \frac{\pi^2}{9}\sum_{j,k=1}^{D}\left(\widehat{\rho}_{j,k} - \widehat{\rho}_{j,k}^{(1)}\right)^2 \quad (8)$$

$$= \frac{\pi^2}{9}\left\|\widehat{\rho} - \widehat{\rho}^{(1)}\right\|_F^2. \quad (9)$$

From the formula

$$\widehat{\rho}_{j,k} = 1 - \frac{6\sum_{i=1}^{n}d_{i,j,k}^2}{n(n^2-1)},$$

(where $d_{i,j,k}$ denotes the difference in ranks of $X_{i,j}$ and $X_{i,k}$ in $X_{1,j},...,X_{n,j}$ and $X_{1,k},...,Y_{n,k}$, respectively), one can see, since $|d_{1,j,k} - d'_{1,j,k}| \le n$ and, for $i \ne 1$, $|d_{i,j,k} - d'_{i,j,k}| \le 1$, that

$$\left|\widehat{\rho}_{j,k} - \widehat{\rho}_{j,k}^{(1)}\right| \le \frac{18}{n},$$

and hence that

$$\|\widehat{\rho} - \widehat{\rho}^{(1)}\|_F \le \frac{18D}{n}. \quad (10)$$

It follows from inequality (9) that

$$\|\widehat{\Sigma}_z - \widehat{\Sigma}_z^{(1)}\|_F \le \frac{6\pi D}{n}.$$

Altogether, this gives

$$\left|\log|\widehat{\Sigma}_z| - \log|\widehat{\Sigma}_z^{(1)}|\right| \le \frac{6\pi D}{zn}.$$

Then, McDiarmid's Inequality gives, for all $\varepsilon > 0$,

$$\mathbb{P}\left[\left|\widehat{I} - \mathbb{E}\left[\widehat{I}\right]\right| > \varepsilon\right] = 2\exp\left(-\frac{nz^2\varepsilon^2}{18\pi^2D^2}\right).$$

This translates to a variance bound of

$$\mathbb{V}\left[\widehat{I}\right] \le \frac{36\pi^2D^2}{z^2n}.$$

∎

C.1. Lower bound for rank-based estimators in terms of Σ

One (perhaps surprising) result of Cai et al. (2015) is that, as long as $D/n \to 0$, the convergence rate of the estimator is independent of the true correlation structure Σ. Here, we show that this desirable property does not hold in the nonparanormal case.

Proposition 15. *Consider the 2-dimensional case*

$$X_1,...,X_n \overset{i.i.d}{\sim} \mathcal{N}(0,\Sigma), \quad with \quad \Sigma = \begin{bmatrix} 1 & \sigma \\ \sigma & 1 \end{bmatrix}, \quad (11)$$

and let $\sigma_* \in (0,1)$. Suppose an estimator $\widehat{I} = \widehat{I}(R)$ of $I_\sigma = -\frac{1}{2}\log(1-\sigma^2)$ is a function of the empirical rank matrix $R \in \mathbb{N}^{n\times 2}$ of X (as defined in (2)). Then, there exists a constant $C > 0$, depending only n, such that the worst-case MSE of $\widehat{I}$ over $\sigma \in (0,\sigma_*)$ satisfies

$$\sup_{\sigma\in(0,\sigma^*]}\mathbb{E}\left[\left(\widehat{I}(R) - I_\sigma\right)^2\right] \ge \frac{1}{64}\left(C - \log(1-\sigma_*^2)\right)^2$$

$$\to \infty \quad as \quad \sigma_* \to 1.$$

Proof: Note that the rank matrix R can take only finitely many values. Let $\mathcal{R}$ be the set of all $(n!)^D$ possible rank matrices and let $\mathcal{R}_1 \subseteq \mathcal{R}$ be the set of $n!$ rank matrices that are perfectly correlated. Then, as $\sigma \to 1$, $\mathbb{P}[R \in \mathcal{R}_1] \to 1$, so, in particular, we can pick σ_0 (depending only on n) such that, for all $\sigma \ge \sigma_0$, $\mathbb{P}[R \in \mathcal{R}_1] \ge \frac{1}{2}$. Since the data are i.i.d., all rank matrices in $\mathcal{R}_1$ have equal probability. It follows that

$$D_{TV}(\mathbb{P}_0||\mathbb{P}_1) = \frac{1}{2}\|\mathbb{P}_0 - \mathbb{P}_1\|_1 \le \frac{1}{2},$$

where D_{TV} denotes total variation distance. Finally, by Le Cam's Lemma (see, e.g., Section 2.3 of Tsybakov (2008)),

$$\inf_{\widehat{I}}\sup_{\sigma\in\{\sigma_0,\sigma_1\}}\mathbb{E}\left[\left(\widehat{I} - I_\sigma\right)^2\right]$$

$$\ge \frac{(I_{\sigma_*} - I_{\sigma_0})^2}{8}\left(1 - D_{TV}(P_{\sigma_0},P_{\sigma_1})\right)$$

$$\ge \frac{(\log(1-\sigma_0^2) - \log(1-\sigma_*^2))^2}{64}$$

∎

D. Details of Experimental Methods

Here, we present details needed to reproduce our numerical simulations. Note that MATLAB source code for these experiments is available at [Omitted for anonymity.], including a single runnable script that performs all experiments and generates all figures presented in this paper. Specific details needed to reproduce experiments are given in the Appendix,

In short, experiments report empirical mean squared errors based on 100 i.i.d. trials of each condition. We initially computed 95% confidence intervals, but these intervals were consistently smaller than marker sizes, so we omitted them to avoid cluttering plots. Except as specified otherwise, each experiment followed the same basic structure, as follows: In each trial, a random correlation matrix $\Sigma \in [-1, 1]^{D \times D}$ was drawn by normalizing a covariance matrix from a Wishart distribution $W(I_D, D)$ with identity scale matrix and D degrees of freedom. Data $X_1, ..., X_n$ were then drawn i.i.d. from $\mathcal{N}(0, \Sigma)$. All estimators were applied to the same data. Unless specified otherwise, $n = 100$ and $D = 25$.

D.1. Computational Considerations

In general, the running time of all the nonparanormal estimators considered is $O(Dn \log n + D^2 n + D^3)$ (i.e., $O(Dn \log n)$ to rank or Gaussianize the variables in each dimension, $D^2 n$ to compute the covariance matrix, and $O(D^3)$ to compute the log-determinant). All log-determinants $\log |\Sigma|$ were computed by summing the logarithms of the diagonal of the Cholesky decomposition of Σ, as this is widely considered to be a fast and numerically stable approach. Note however that faster ($O(D)$-time) randomized algorithms (Han et al., 2015) have been proposed to approximate the log-determinant).

E. Additional Experimental Results

Here, we present variants on the experiments presented in the main paper, which support but are not necessary for illustrating our conclusions.

E.1. Effects of Other Marginal Transformations

In Section 7, we showed that the Gaussian estimator $\widehat{I}$ is highly sensitive to failure of the Gaussian assumption for even a small fraction of marginals. Figure 1(b), illustrates this for the transformation $x \mapsto \exp(x)$, but we show here that this is not specific to the exponential transformation. As shown in Figures 2 nearly identical results hold when the marginal transformation f is the hyperbolic tangent function $x \mapsto \tanh(x)$, the cubic function $x \mapsto x^3$, sigmoid function $x \mapsto \frac{1}{1+e^{-x}}$, or standard normal CDF.

F. Specific Assumptions for Estimating $H(X)$

As shown in the main paper, to estimate the entropy of a nonparanormal distribution at the rate $O(D^2/n)$, it suffices to the univariate entropy of each variable X_j at the rate $O(1/n)$. To do this, additional assumptions are required on the marginal densities p_j. Here, we give detailed sufficient conditions for this.

Letting $S_j \subseteq \mathbb{R}$ denote the support of p_j, the two key assumptions can be roughly classified as follows:

(a) $\frac{1}{2}$-order smoothness[9]; e.g., a Hölder condition:

$$\sup_{x \neq y \in S_j} \frac{|p_j(x) - p_j(y)|}{|x - y|^{1/2}} < L,$$

or a (slightly weaker) Sobolev condition:

$$\int_{S_j} p_j^2(x)\, dx < \infty \ \text{ and } \ \int_{S_j} \left(|\xi|^{1/2} |\mathcal{F}\left[p_j\right](\xi)| \right)^2 d\xi < L,$$

(where $\mathcal{F}\left[p_j\right](\xi)$ denotes the Fourier transform of p_j evaluated at ξ) for some constant $L > 0$.

(b) absolute bounds $p_j(x) \in [\kappa_1, \kappa_2]$ for all $x \in S_j$ or (a_j, b_j)-exponential tail bounds

$$\frac{f(x)}{\exp(-a_j x^{b_j})} \in [\kappa_1, \kappa_2] \quad \text{for all } x \in S_j$$

for some $\kappa_1, \kappa_2 \in (0, \infty)$.

Under these assumptions, there are a variety of nonparametric univariate entropy estimators that have been shown to converge at the rate $O(1/n)$ (Beirlant et al., 1997; Kandasamy et al., 2015; Singh & Póczos, 2016b; Moon et al., 2016).

G. Lower bounding the eigenvalues of a bandable matrix

Recall that, for $c \in (0, 1)$, a matrix $\Sigma \in \mathbb{R}^{D \times D}$ is called *c-bandable* if there exists a constant $c \in (0, 1)$ such that, for all $i, j \in D$, $|\Sigma_{i,j}| \leq c^{|i-j|}$.

Here, we show simple bounds on the eigenvalues of a bandable correlation matrix Σ. While this result is fairly straightforward, a brief search the literature turned up no comparable results. Bickel & Levina (2008), who originally introduced the class of bandable covariance matrices, separately assumed the existence of lower and upper

---

[9] This is stronger than the $\frac{1}{4}$-order smoothness mandated by the minimax rate for entropy estimation (Birgé & Massart, 1995), but appears necessary for most practical entropy estimators. See Section 4 of Kandasamy et al. (2015) for further details.

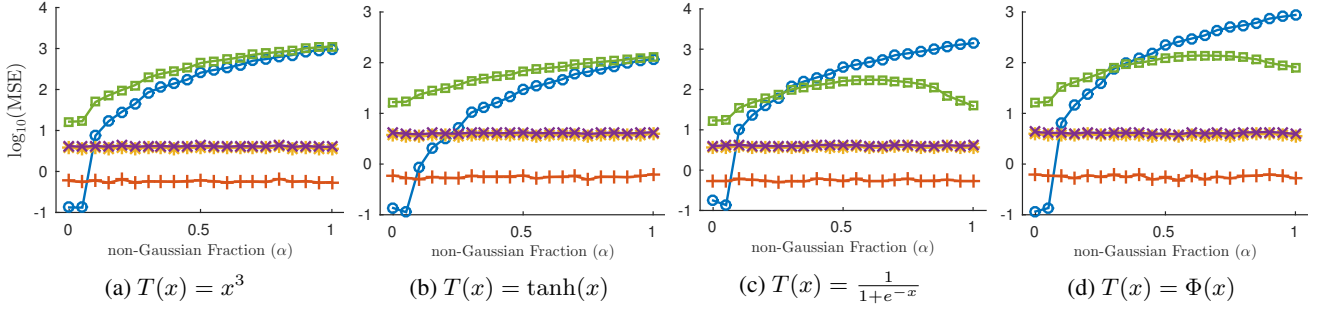(a) $T(x) = x^3$ (b) $T(x) = \tanh(x)$ (c) $T(x) = \frac{1}{1+e^{-x}}$ (d) $T(x) = \Phi(x)$

Figure 2. Semi-log plot of mean squared error of various estimators over the fraction of non-Gaussian marginals $\alpha \in [0, 1]$, for various marginal transforms T.

bounds on the eigenvalues to prove their results. In the context of information estimation, this results of particular interest because, when $c < 1/3$ it implies a dimension-free positive lower bound on the minimum eigenvalue of Σ, hence complementing our upper bound in Theorem 8.

Proposition 16. *Suppose a symmetric matrix* $\Sigma \in \mathbb{R}^{D \times D}$ *is c-bandable and has identical diagonal entries* $\Sigma_{j,j} = 1$. *Then, the eigenvalues* $\lambda_1(\Sigma), ..., \lambda_D(\Sigma)$ *of* Σ *can be bounded as*

$$\frac{1 - 3c}{1 - c} \le \lambda_1(\Sigma), ..., \lambda_D(\Sigma) \le \frac{1 + c}{1 - c}.$$

In particular, when $c < 1/3$, *we have*

$$0 < \frac{1 - 3c}{1 - c} \le \lambda_D(\Sigma).$$

Proof: The proof is based on the Gershgorin circle theorem (Gershgorin, 1931; Varga, 2009). In the case of a real symmetric matrix Σ, this states that the eigenvalues of Σ lie within a union of intervals

$$\{\lambda_1(\Sigma), ..., \lambda_D(\Sigma)\} \subseteq \bigcup_{j=1}^{D} [\Sigma_{j,j} - R_j, \Sigma_{j,j} + R_j], \quad (12)$$

where $R_j := \sum_{k \neq j} |\Sigma_{j,k}|$ is the sum of the absolute values of the non-diagonal entries of the jth row of Σ. In our case, since the diagonal entries of Σ are all $\Sigma_{j,j} = 1$, we simply have to bound

$$\max_{j \in [D]} R_j \le \sum_{k \neq j} c^{|k-j|}.$$

This geometric sum is maximized when $j = \lceil D/2 \rceil$, giving

$$R_j \le 2 \sum_{\delta=1}^{\lfloor D/2 \rfloor} c^{\delta} = 2c \frac{1 - c^{\lfloor D/2 \rfloor}}{1 - c} \le \frac{2c}{1 - c}.$$

Finally, the inclusion (12) gives

$$\lambda_D(\Sigma) \ge 1 - \frac{2c}{1 - c} = \frac{1 - 3c}{1 - c} > 0$$

when $c < 1/3$. $1 + \frac{2c}{1-c} = \frac{1+c}{1-c}$. ∎