Nonparanormal Information Estimation by Singh and Poczos

Daniel J. McDonald

7 Sept. 2017

BASIC STRUCTURE

- Introduction — Brief sketch of applications, overview of previous work, no definitions or notation

  *"The **main goal of this paper** is to fill the gap between these two extreme settings by studying informa- tion estimation in a semiparametric compromise between the two…"*

- Problem statement and notation — careful definitions of important concepts with discussion, "formal problem statement", collection of other notational choices
- Related work and contributions — more detailed overview of previous work (what does each do), itemized list of the important contributions of the paper
- The remainder proposes 3 estimators, the main result is an upper bound on the quality of the estimator, discuss lower bounds, compare their estimators in experiments, suggest techniques for a related quantity, and conclude.

Definitions and concepts

- Let $X_1, \ldots, X_D$ be $\mathbb{R}$-valued random variables with a joint probability density p and marginal densities $p_1, \ldots, p_D$.

The mutual information $I(X)$ of $X = (X_1, \ldots, X_D)$ is

$$I(X) = \mathbb{E}\left[\log\left(\frac{p(X)}{\prod_i p_i(X_i)}\right)\right].$$

- A random vector X has **nonparanormal distribution** $\mathcal{NPN}(\Sigma; f)$ if there exist some functions g_j such that $g_j(X_j) \sim N(0, 1)$ for all j and and the joint distribution of $f(X) = (g_1(X_1), \ldots, g_D(X_D)) \sim N(0, \Sigma)$.

This is a generalization of Gaussian distributions which allow for very odd marginals. This is also called a Gaussian copula.

Goal: Estimate $I(X)$ using a sample $X_1, \ldots, X_n \sim \mathcal{NPN}(\Sigma; f)$.

Context

- If $X_1, \ldots, X_n$ are multivariate normal, $I(X_1) = -\frac{1}{2}\log|\Sigma|$.
- In this case, there exists an estimator which has $MSE = -2\log(1 - D/n)$.
- It is also known that there is a matrix Σ such that **any** estimator will have $MSE \geq 2\frac{D}{n}$. This has 2 consequences: (1) $D/n \to 0$ is necessary for consistent estimation, and (2) if $D/n \approx 0$, then the bound is tight—$-2\log(1 - D/n) \approx 2\frac{D}{n}$.
- So this problem is nearly solved.
- If f is allowed to be any density in a Hölder class with smoothness parameter s, there exist nonparametric (minimax) estimators with

$$MSE = O\left(n^{-\frac{8s}{4s+D}}\right)$$

- $\mathcal{NPN}$ is not quite Gaussian, but it's not nearly as general as nonparametric.
- Turns out we still have $I(X_1) = -\frac{1}{2}\log|\Sigma|$.

THE ESTIMATORS

$\widehat{\Sigma}_G$:

1. Define $R_{ij} = \sum_{k=1}^n \mathbf{1}(X_{ij} \geq X_{kj})$.
2. "Gaussianize" the data: $\tilde{X}_{ij} = \Phi^{-1}\left(\frac{R_{ij}}{n+1}\right)$.
3. Estimate Σ with $\widehat{\Sigma_G} = \frac{1}{n}\sum_{i=1}^n \tilde{X}_i \tilde{X}_i^\top$.

$\widehat{\Sigma}_\rho$ and $\widehat{\Sigma}_\tau$:

1. Define $\rho(X,Y) = \mathrm{Corr}(F_X(X),\ F_Y(Y))$ and $\tau(X,Y) = \mathrm{Corr}(sgn(X - X'),\ sgn(Y - Y'))$. Turns out these are invariant to the marginal transformation f.
2. Set $\widehat{\Sigma}_\rho = 2\sin\left(\frac{\pi}{6}\widehat{\rho}\right)$ with $\widehat{\rho} = \widehat{\mathrm{Corr}}(R)$ and $\widehat{\Sigma}_\tau = \sin\left(\frac{\pi}{2}\widehat{\tau}\right)$ with

$$\widehat{\tau} = \frac{1}{\binom{n}{2}} \sum_{i \neq \ell} sgn(X_{ij} - X_{\ell j}) sgn(X_{ik} - X_{\ell k}).$$

- These may not be positive definite.
- Project onto the cone

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

- By hard thresholding the eigenvalues of a matrix at z, you project onto this cone.

That is

$$A_z = \arg \min_{B \in S(z)} \|A - B\|_F .$$

- So, estimate, threshold, then plug in to $I(X) = -\frac{1}{2} \log |\Sigma|$.

THEORY

They prove upper bounds for the bias and variance of their estimator based on ρ:

$$\text{bias}^2 \leq C\left(\frac{D}{z\sqrt{n}} + \log\frac{|\Sigma_z|}{|\Sigma|}\right)$$

This is done with a Taylor expansion, a typical trick for these.

$$\text{Var} \leq \frac{36\pi^2 D^2}{z^2 n}$$

The second one comes from a concentration equality based on Hoeffding's inequality (more on that later).

These are standard techniques. The next step would be to minimize their sum in z, but this is not possible because of Σ_z. They give a simpler result in a special case.

- They argue that the lower bound technique in the Gaussian case won't work.
- They give an example illustrating why.
- They argue in the paper that the Gaussian lower bound should apply because R is sufficient for Σ (and hence for I).
- Essentially there is a gap: the upper bound is $\frac{\lambda_{\min}(\Sigma)^2 D^2}{n}$ while the lower bound is $\frac{2D}{n}$.

Conclusions

- I find this paper very nicely done.
- It illustrates what the requirements are for publishing in a top conference, and illustrates how to deal with difficulties.
- The structure and formatting make it easy for reviewers to grasp.
- There is some nice future work here which would make a good project.
- It is worth examining the simulations to see how to incorporate those.
- Something that might be beneficial would be to include a real data example. They give a special case that shows