Supplemental material: High-dimensional estimation of mutual information via classification error

Charles Zheng and Yuval Benjamini

April 24, 2016

1 Proofs

Lemma 1. *Suppose $(Z_1, Z_2, \ldots, Z_k)$ are jointly multivariate normal, with $\boldsymbol{E}[Z_1 - Z_i] = \alpha$, $Var(Z_1) = \beta$, $Cov(Z_1, Z_i) = \gamma$, $Var(Z_i) = \delta$, and $Cov(Z_i, Z_j) = \epsilon$ for all $i, j = 2, \ldots, k$, such that $\beta + \epsilon - 2\gamma > 0$. Then, letting*

$$\mu = \frac{\boldsymbol{E}[Z_1 - Z_i]}{\sqrt{\frac{1}{2} Var(Z_i - Z_j)}} = \frac{\alpha}{\sqrt{\delta - \epsilon}},$$

$$\nu^2 = \frac{Cov(Z_1 - Z_i, Z_1 - Z_j)}{\frac{1}{2} Var(Z_i - Z_j)} = \frac{\beta + \epsilon - 2\gamma}{\delta - \epsilon},$$

we have

$$\begin{aligned} \Pr[Z_1 < \max_{i=2}^{k} Z_i] &= \Pr[W < M_{k-1}] \\ &= 1 - \int \frac{1}{\sqrt{2\pi\nu^2}} e^{-\frac{(w-\mu)^2}{2\nu^2}} \Phi(w)^{k-1} dw, \end{aligned}$$

where $W \sim N(\mu, \nu^2)$ and M_{k-1} is the maximum of $k-1$ independent standard normal variates, which are independent of W.

Proof. We can construct independent normal variates $G_1, G_2, \ldots, G_k$ such that

$$G_1 \sim N(0, \beta + \epsilon - 2\gamma)$$

$$G_i \sim N(0, \delta - \epsilon) \text{ for } i > 1$$

such that

$$Z_1 - Z_i = \alpha + G_1 + G_i \text{ for } i > 1.$$

Hence

$$\begin{aligned}
\Pr[Z_1 < \max_{i=2}^{k} Z_i] &= \Pr[\min_{i>1} Z_1 - Z_i < 0]. \\
&= \Pr[\min_{i=2}^{k} G_1 + G_i + \alpha < 0] \\
&= \Pr[\min_{i=2}^{k} G_i < -\alpha - G_1] \\
&= \Pr[\min_{i=2}^{k} \frac{G_i}{\sqrt{\delta - \epsilon}} < -\frac{\alpha - G_1}{\sqrt{\delta - \epsilon}}].
\end{aligned}$$

Since $\frac{G_i}{\sqrt{\delta-\epsilon}}$ are iid standard normal variates, and since $-\frac{\alpha-G_1}{\sqrt{\delta-\epsilon}} \sim N(\mu, \nu^2)$ for μ and ν^2 given in the statement of the Lemma, the proof is completed via a straightforward computation. □

Theorem 1. Let $p^{[d]}(x, y)$ be a sequence of joint densities for $d = 1, 2, \ldots$ as given above. Further assume that

A1. $\lim_{d\to\infty} I(X^{[d]}; Y^{[d]}) = \iota < \infty$.

A2. There exists a sequence of scaling constants $a_{ij}^{[d]}$ and $b_{ij}^{[d]}$ such that the random vector $(a_{ij}\ell_{ij}^{[d]} + b_{ij}^{[d]})_{i,j=1,\ldots,k}$ converges in distribution to a multivariate normal distribution.

A3. There exists a sequence of scaling constants $a^{[d]}$, $b^{[d]}$ such that

$$a^{[d]}u(X^{(1)}, Y^{(2)}) + b^{[d]}$$

converges in distribution to a univariate normal distribution.

A4. For all $i \neq k$,

$$\lim_{d\to\infty} \mathrm{Cov}[u(X^{(i)}, Y^{(j)}), u(X^{(k)}, Y^{(j)})] = 0.$$

Then for $e_{ABE,k}$ as defined above, we have

$$\lim_{d\to\infty} e_{ABE,k} = \pi_k(\sqrt{2\iota})$$

where

$$\pi_k(c) = 1 - \int_{\mathbb{R}} \phi(z-c)\Phi(z)^{k-1} dz$$

where ϕ and Φ are the standard normal density function and cumulative distribution function, respectively.

Proof.

For $i = 2, \ldots, k$, define

$$Z_i = \log p(Y^{(1)}|X^{(i)}) - \log p(Y^{(1)}|X^{(1)}).$$

Then, we claim that $\vec{Z} = (Z_2, \ldots, Z_k)$ converges in distribution to

$$\vec{Z} \sim N\left(-2\iota, \begin{bmatrix} 4\iota & 2\iota & \cdots & 2\iota \\ 2\iota & 4\iota & \cdots & 2\iota \\ \vdots & \vdots & \ddots & \vdots \\ 2\iota & 2\iota & \cdots & 4\iota \end{bmatrix}\right).$$

Combining the claim with the lemma (stated below this proof) yields the desired result.

To prove the claim, it suffices to derive the limiting moments

$$\mathbf{E}[Z_i] \to -2\iota,$$

$$\text{Var}[Z_i] \to 4\iota,$$

$$\text{Cov}[Z_i, Z_j] \to 2\iota,$$

for $i \neq j$, since then assumption A2 implies the existence of a multivariate normal limiting distribution with the given moments.

Before deriving the limiting moments, note the following identities. Let $X' = X^{(2)}$ and $Y = Y^{(1)}$.

$$\mathbf{E}[e^{u(X',Y)}] = \int p(x)p(y)e^{u(x,y)} dxdy = \int p(x,y) dxdy = 1.$$

Therefore, from assumption A3 and the formula for gaussian exponential moments, we have

$$\lim_{d\to\infty} \mathbf{E}[u(X',Y)] - \frac{1}{2}\text{Var}[u(X',Y)] = 0.$$

Let $\sigma^2 = \lim_{d\to\infty} \mathrm{Var}[u(X', Y)]$. Meanwhile, by applying assumption A2,

$$\begin{aligned}
\lim_{d\to\infty} I(X;Y) &= \lim_{d\to\infty} \int p(x,y)u(x,y)dxdy = \lim_{d\to\infty} \int p(x)p(y)e^{u(x,y)}u(x,y)dxdy \\
&= \lim_{d\to\infty} \mathbf{E}[e^{u(X,Y')}u(X,Y')] \\
&= \int_{\mathbb{R}} e^z z \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z+\sigma^2/2)^2}{2\sigma^2}} \text{ (applying A2)} \\
&= \int_{\mathbb{R}} z \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\sigma^2/2)^2}{2\sigma^2}} \\
&= \frac{1}{2}\sigma^2.
\end{aligned}$$

Therefore,

$$\sigma^2 = 2\iota,$$

and

$$\lim_{d\to\infty} \mathbf{E}[u(X', Y)] = -\iota.$$

Once again by applying A2, we get

$$\begin{aligned}
\lim_{d\to\infty} \mathrm{Var}[u(X,Y)] &= \lim_{d\to\infty} \int (u(x,y)-\iota)^2 p(x,y)dxdy \\
&= \lim_{d\to\infty} \int (u(x,y)-\iota)^2 e^{u(x,y)} p(x)p(y)dxdy \\
&= \lim_{d\to\infty} \mathbf{E}[(u(X',Y)-\iota)^2 e^{u(X',Y)}] \\
&= \int (z-\iota)^2 e^z \frac{1}{\sqrt{4\pi\iota}} e^{-\frac{(z+\iota)^2}{4\iota}} dz \text{ (applying A2)} \\
&= \int (z-\iota)^2 \frac{1}{\sqrt{4\pi\iota}} e^{-\frac{(z-\iota)^2}{4\iota}} dz \\
&= 2\iota.
\end{aligned}$$

We now proceed to derive the limiting moments. We have

$$\begin{aligned}
\lim_{d\to\infty} \mathbf{E}[Z] &= \lim_{d\to\infty} \mathbf{E}[\log p(Y|X') - \log p(Y|X)] \\
&= \lim_{d\to\infty} \mathbf{E}[u(X',Y) - u(X,Y)] = -2\iota.
\end{aligned}$$

Also,

$$\begin{aligned}
\lim_{d\to\infty} \mathrm{Var}[Z] &= \lim_{d\to\infty} \mathrm{Var}[u(X', Y) - u(X, Y)] \\
&= \lim_{d\to\infty} \mathrm{Var}[u(X', Y)] + \mathrm{Var}[u(X, Y)] \text{ (using assumption A4)} \\
&= 4\iota,
\end{aligned}$$

and similarly

$$\begin{aligned}
\lim_{d\to\infty} \mathrm{Cov}[Z_i, Z_j] &= \lim_{d\to\infty} \mathrm{Var}[u(X, Y)] \text{ (using assumption A4)} \\
&= 2\iota.
\end{aligned}$$

This concludes the proof. $\square$.

1.1 Assumptions of theorem 1

Assumptions A1-A4 are satisfied in a variety of natural models. One example is a multivariate Gaussian model where

$$X \sim N(0, \Sigma_d)$$

$$E \sim N(0, \Sigma_e)$$

$$Y = X + E$$

where Σ_d and Σ_e are $d \times d$ covariance matrices, and where X and E are independent. Then, if $d\Sigma_d$ and Σ_e have limiting spectra H and G respectively, the joint densities $p(x, y)$ for $d = 1, \ldots,$ satisfy assumptions A1 - A4.

We can also construct a family of densities satisfying A1 - A4, which we call an *exponential family sequence model* since each joint distribution in the sequence is a member of an exponential family. A given exponential family sequence model is specified by choice of a base carrier function $b(x, y)$ and base sufficient statistic $t(x, y)$, with the property that carrier function factorizes as

$$b(x, y) = b_x(x) b_y(y)$$

for marginal densities b_x and b_y. Note that the dimensions of x and y in the base carrier function are arbitrary; let p denote the dimension of x and q the dimension of y for the base carrier function. Next, one specifies a sequence of scalar parameters $\kappa_1, \kappa_2, \ldots$ such that

$$\lim_{d\to\infty} d\kappa_d = c < \infty.$$

for some constant c. For the dth element of the sequence, $X^{[d]}$ is a pd-dimensional vector, which can be partitioned into blocks

$$X^{[d]} = (X_1^{[d]}, \ldots, X_d^{[d]})$$

where each $X_i^{[d]}$ is p-dimensional. Similarly, $Y^{[d]}$ is partitioned into $Y_i^{[d]}$ for $i = 1, \ldots, d$. The density of $(X^{[d]}, Y^{[d]})$ is given by

$$p^{[d]}(x^{[d]}, y^{[d]}) = Z_d^{-1} \left(\prod_{i=1}^{d} b(x_i^{[d]}, y_i^{[d]}) \right) \exp \left[\kappa_d \sum_{i=1}^{d} t(x_i^{[d]}, y_i^{[d]}) \right],$$

where Z_d is a normalizing constant. Hence $p^{[d]}$ can be recognized as the member of an exponential family with carrier measure

$$\left(\prod_{i=1}^{d} b(x_i^{[d]}, y_i^{[d]}) \right)$$

and sufficient statistic

$$\sum_{i=1}^{d} t(x_i^{[d]}, y_i^{[d]}).$$

One example of such an exponential family sequence model is a multivariate Gaussian model with limiting spectra $H = \delta_1$ and $G = \delta_1$, but scaled so that the marginal variance of the components of X and Y are equal to one. This corresponds to a exponential family sequence model with

$$b_x(x) = b_y(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

and

$$t(x, y) = xy.$$

Another example is a multivariate logistic regression model, given by

$$X \sim N(0, I)$$

$$Y_i \sim \text{Bernoulli}(e^{\beta X_i}/(1 + e^{\beta X_i}))$$

This corresponds to an exponential family sequence model with

$$b_x(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$b_y(y) = \frac{1}{2} \text{ for } y = \{0, 1\},$$

and

$$t(x, y) = x\delta_1(y) - x\delta_0(y).$$

The multivariate logistic regression model (and multivariate Poisson regression model) are especially suitable for modeling neural spike count data; we simulate data from such a multivariate logistic regression model in section X.

2 Additional simulation results

Multiple-response logistic regression model

$$X \sim N(0, I_p)$$

$$Y \in \{0, 1\}^q$$

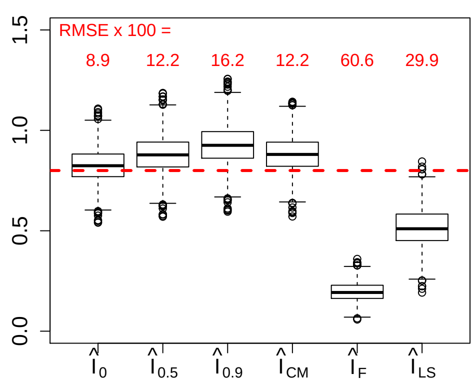
$$Y_i | X = x \sim \text{Bernoulli}(x^T B_i)$$

where B is a $p \times q$ matrix.

Methods.

- Nonparametric: $\hat{I}_0$ naive estimator, $\hat{I}_\alpha$ anthropic correction.
- ML-based: $\hat{I}_{CM}$ confusion matrix, $\hat{I}_F$ Fano, $\hat{I}_{HD}$ high-dimensional method.

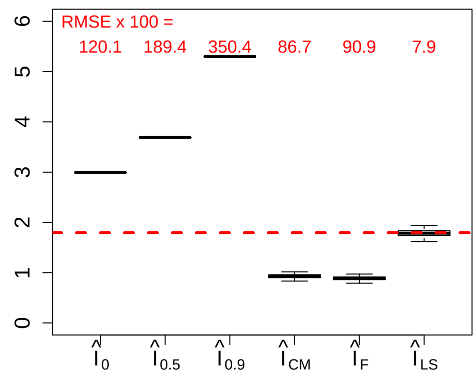
Sampling distribution of $\hat{I}$ for $\{p = 3,\ B = \frac{4}{\sqrt{3}} I_3,\ K = 20,\ r = 40\}$.
True parameter $I(X; Y) = 0.800$ *(dotted line.)*



Naïve estimator performs best! $\hat{I}_{HD}$ not effective.

Sampling distribution of $\hat{I}$ for $\{p = 50,\ B = \frac{4}{\sqrt{50}} I_{50},\ K = 20,\ r = 8000\}$.

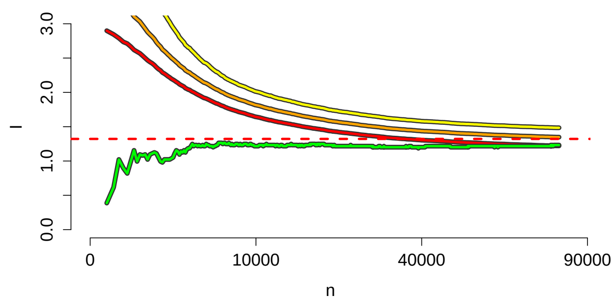
True parameter $I(X;Y) = 1.794$ *(dashed line.)*



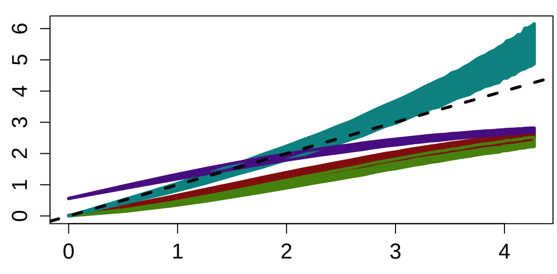
Non-parametric methods extremely biased.

Estimation path of $\hat{I}_{HD}$ and $\hat{I}_\alpha$ as n ranges from 10 to 8000.

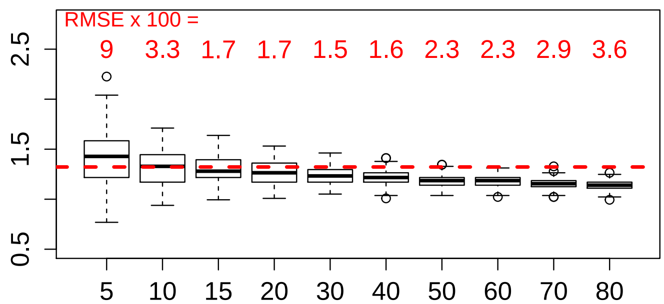
$\{p = 10,\ B = \frac{4}{\sqrt{10}} I_{10},\ K = 20\}$. True parameter $I(X;Y) = 1.322$ *(dashed line.)*



Estimated $\hat{I}$ vs true I.

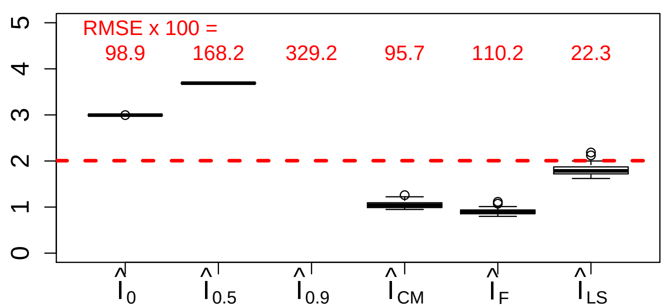


Sampling distribution of $\hat{I}_{HD}$ for $\{p = 10,\ B = \frac{4}{\sqrt{10}} I_{10},\ N = 80000\}$, and $K = \{5, 10, 15, 20, \ldots, 80\}$, $r = N/k$.
True parameter $I(X;Y) = 1.322$ *(dashed line.)*



Decreasing variance as K increases. Bias at large and small K.
$p = 20$ and $q = 40$, entries of B are iid $N(0, 0.025)$.
$K = 20$, $r = 8000$, true $I(X;Y) = 1.86$ *(dashed line.)*

Sampling distribution of $\hat{I}$.



3 References

- Cover and Thomas. Elements of information theory.
- Muirhead. Aspects of multivariate statistical theory.
- van der Vaart. Asymptotic statistics.