Supplementary Material: A Robust-Equitable Copula Dependence Measure for Feature Selection

Yale Chang
Department of ECE
Northeastern University
Yi Li
Department of Mathematics
Northeastern University
Adam(Aidong Ding)
Department of Mathematics
Northeastern University
Jennifer G. Dy
Department of ECE
Northeastern University

1 Proof of Theorem 3

For simplicity, we focus on the bivariate case (X and Y are each one-dimensional variables). The extension to proof in multivariate case is straight forward. We first work on the mutual information, then show the similar arguments on the copula distances. To prove the theorem, we use Le Cam [1973]'s method to find the lower bound on the minimax risk of the estimating mutual information MI. To do this, we will use a more convenient form of Le Cam's method developed by Donoho and Liu [1991]. Define the module of continuity of a functional T over the class $\mathbf{F}$ with respect to Hellinger distance as in equation (1.1) of Donoho and Liu [1991]:

$$w(\varepsilon) = \sup\{|T(F_1) - T(F_2)| : H(F_1, F_2) \leq \varepsilon, F_i \in \mathbf{F}\}. \tag{1}$$

Here $H(F_1, F_2)$ denotes the Hellinger distance between F_1 and F_2. Then the minimax rate of convergence for estimating $T(F)$ over the class $\mathbf{F}$ is bounded below by $w(n^{-1/2})$.

We now look for a pair of density functions $c_1(u, v)$ and $c_2(u, v)$ on the unit square for distributions that are close in Hellinger distance but far away in their mutual information. This provides a lower bound on the module of continuity for mutual information MI over the class $\mathfrak{C}$, and hence leads to a lower bound on the minimax risk. We outline the proof here.

We first divide the unit square into three disjoint regions R_1, R_2 and R_3 with $R_1 \cup R_2 \cup R_3 = [0,1] \times [0,1]$. The first density function $c_1(u,v)$ puts probability masses δ, a and $1-a-\delta$ respectively on the regions R_1, R_2 and R_3 each uniformly. The a is an arbitrary small fixed value, for example, $a = 0.01$. For now, we take δ to be another small fixed value. The area of the region is chosen so that $c_1(u,v) = M$ on region R_2 and $c_1(u,v) = M^*$ on region R_1 for a very big M^*. The second density function $c_2(u,v)$, compared to $c_1(u,v)$, moves a small probability mass ε from R_1 to R_2. We will see that the Hellinger distance between c_1 and c_2 is of the same order as ε, but the change in MI is unbounded for big M^*. Hence module of continuity $w(\varepsilon)$ is unbounded for mutual information MI. Therefore the MI can not be consistently estimated over the class $\mathfrak{C}$.

Specifically, the region R_1 is chosen to be a narrow strip immediately above the diagonal, $R_1 = \{(u,v) : -\delta_1 < u - v < 0\}$; and R_2 is chosen to be a narrow strip immediately below the diagonal, $R_2 = \{(u,v) : 0 \le u - v < \delta_2\}$. The remaining region is $R_3 = [0,1] \times [0,1] \setminus (R_1 \cup R_2)$. The values of δ_1 and δ_2 are chosen so that the areas of regions R_1 and R_2 are δ/M^* and a/M respectively. Then clearly $c_1(u,v) = M^*$ on R_1; $c_1(u,v) = M$ on R_2; $c_1(u,v) = (1-a-\delta)/(1-a/M-\delta/M^*)$ on R_3. And $c_2(u,v) = M^* - \varepsilon(M^*/\delta)$ on R_1; $c_2(u,v) = M + \varepsilon(M/a)$ on R_2; $c_2(u,v) = c_1(u,v)$ on R_3. See the Figure 1.

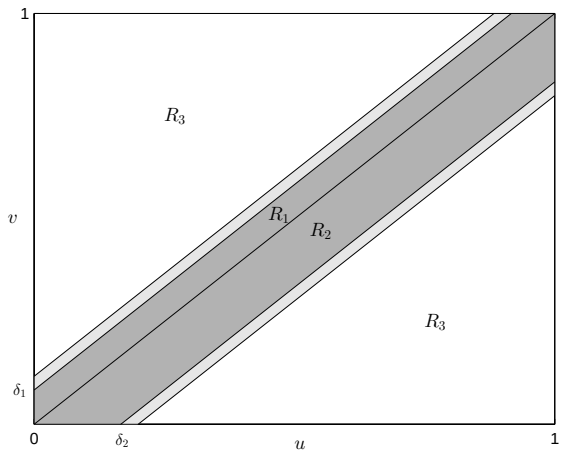

Figure 1: The plot shows the regions R_1, R_2 and R_3. The other two narrow strips neighboring R_1 and R_2 are for the continuity correction mentioned at the end of the proof.

Then we have

$$\begin{aligned} 2H^2(c_1,c_2) &= \iint(\sqrt{c_2(u,v)}-\sqrt{c_1(u,v)})^2 dudv \\ &= (\sqrt{M^*-\varepsilon(M^*/\delta)}-\sqrt{M^*})^2\delta/M^* + (\sqrt{M+\varepsilon(M/a)}-\sqrt{M})^2 a/M \\ &= \delta(\sqrt{1-\varepsilon/\delta}-1)^2 + a(\sqrt{1+\varepsilon/a}-1)^2 \\ &= \delta(\varepsilon/2\delta)^2 + a(\varepsilon/2a)^2 + o(\varepsilon^2) \\ &= \varepsilon^2(\tfrac{1}{4\delta}+\tfrac{1}{4a}) + o(\varepsilon^2). \end{aligned}$$

Hence the Hellinger distance is of the same order as ε:

$$H(c_1,c_2) = \varepsilon\sqrt{\frac{1}{8\delta}+\frac{1}{8a}} + o(\varepsilon).$$

On the other hand, the difference in the mutual information is

$$\begin{aligned} & \mathrm{MI}(c_1) - \mathrm{MI}(c_2) \\ = \;& \delta\log(M^*) + a\log(M) - (\delta-\varepsilon)\log[M^*-\varepsilon(M^*/\delta)] - (a+\varepsilon)\log[M+\varepsilon(M/a)] \\ = \;& \varepsilon\log(M^*) - \varepsilon\log(M) - (\delta-\varepsilon)\log(1-\varepsilon/\delta) - (a+\varepsilon)\log(1+\varepsilon/a). \end{aligned} \tag{2}$$

Here M, δ and a are fixed constants. Hence when $M^* \to \infty$, this difference in MI also goes to ∞. For example, if we let $M^* = e^{1/(\varepsilon)^2}$, then the module of continuity $w(\varepsilon) \geq O(1/\varepsilon)$. That means, the rate of convergence is at least $O(w(n^{-1/2})) = O(n^{1/2}) \to \infty$. In other words, MI can not be consistently estimated.

Now, let us consider the $\mathrm{CD}_\alpha = \iint_{I^2} |c(u,v)-1|^\alpha dudv$, for $\alpha > 1$, where I^2 is the unit square.

$$\begin{aligned} & \mathrm{CD}_\alpha(c_1) - \mathrm{CD}_\alpha(c_2) \\ = \;& |M^*-1|^\alpha\delta/M^* + |M^-1|^\alpha a/M - |M^*-1-\varepsilon(M^*/\delta)|^\alpha\delta/M^* + |M-1+\varepsilon(M/a)|^\alpha a/M \\ = \;& [|M^*-1|^\alpha - |M^*-1-\varepsilon(M^*/\delta)|^\alpha]\delta/M^* + [|M^-1|^\alpha - |M-1+\varepsilon(M/a)|^\alpha]a/M \\ = \;& \alpha[(M^*-1)^{\alpha-1}M^*/\delta - (M-1)^{\alpha-1}M/\alpha]\varepsilon + o(\varepsilon^2). \end{aligned} \tag{3}$$

Again, M, δ and a are fixed constants. Hence when $M^* \to \infty$, this difference in $\mathrm{CD}_\alpha, \alpha > 1$ also goes to ∞. For example, if we let $M^* = (\varepsilon^{-2} + M^\alpha)^{\frac{1}{\alpha-1}} + 1$, then the module of continuity $w(\varepsilon) \geq O(1/\varepsilon)$. Note that $\alpha > 1$ is essential here. That means, the rate of convergence is at least $O(w(n^{-1/2})) = O(n^{1/2}) \to \infty$. In other words, $\mathrm{CD}_\alpha, \alpha > 1$ can not be consistently estimated.

The above outlines the main idea of the proof, ignoring some mathematical subtleties. One is that the example densities c_1 and c_2 are only piecewise continuous on the three regions, but not truly continuous as required for the class $\mathfrak{C}$. This can be easily remedied by connecting the three pieces linearly. Specifically we set the densities $c_i(u,v) = M$, $i = 1, 2$, on the boundary between R_1 and R_3, $\{(u,v) : u - v = -\delta_1\}$, and on the boundary between R_2 and R_3, $\{(u,v) : u - v = \delta_2\}$. Then we use two narrow strips within R_3, $\{(u,v) : -\delta_3 \leq u-v \leq -\delta_1\}$ and $\{(u,v) : \delta_2 \leq u-v \leq \delta_4\}$ to connect the constant $c_i(u,v)$ values on the rest of region R_3 with the boundary value $c_i(u,v) = M$ continuously through linear (in $u - v$) $c_i(u,v)$'s on the two strips that satisfies the Hölder condition (17) of the main text. By the Hölder conditionp, the

connection can be made with strips of width at most $(M-1+a+\delta)/M_1$. This continuity modification does not affect the calculation of the difference $\text{MI}(c_1) - \text{MI}(c_2)$ or $\text{CD}_\alpha(c_1) - \text{CD}_\alpha(c_2)$ above as c_1 and c_2 only differ on regions R_1 and R_2. Within regions R_1 and R_2, the densities c_1 and c_2 can be further similarly connected continuously linearly in $u - v$. As there is no Hölder condition on A_M^c, the connection within R_1 and R_2 can be as steep as we want. Clearly the order obtained through above calculations will not change if we make these connections very steep so that their effect is negligible.

Another technical subtlety is that the c_1 and c_2 defined above are only densities on the unit square but not copula densities which require uniform marginal distributions. However, it is clear that the marginal densities for c_is are uniform over the interval $(\delta_3, 1-\delta_4)$ and linear in the rest of interval near the two end points 0 and 1. The copulas densities c_i^*'s corresponding to c_i's can be calculated directly through Sklar's decomposition (1) in the main text. It is easy to see that the order for the module of continuity $w(\varepsilon)$ remains the same for using the corresponding copula densities c_i^*'s.

2 Proof of Theorem 4

Proof. The first two terms in (9) corresponds to bias and standard deviation of kernel density estimation when the copula density is bounded. When the copula density is unbounded, the kernel density estimation $\hat{c}(\mathbf{Z})$ is not consistent. However, a smaller order $O(\frac{1}{nh^d})$ term bounds the overall error contribution to $\widehat{RCD}$ resulting from $\hat{c}(\mathbf{Z})$ in the unbounded copula density region.

Let $M_2 = \frac{M+1}{2}$, $A_{M_2} = \{\mathbf{Z}|c(\mathbf{Z}) \le M_2\}$, $T_1(c) = \int_{A_{M_2}} (1-c(\mathbf{Z}))_+ d\mathbf{Z}$, $T_2(c) = \int_{A_{M_2}^c} (1-c(\mathbf{Z}))_+ d\mathbf{Z}$, $RCD = T_1(c) + T_2(c)$, and $\widehat{RCD} = T_1(\hat{c}) + T_2(\hat{c})$

Firstly, we consider the region A_{M_2} with bounded copula density. Here we calculate the bias and variance of the kernel density estimator using standard methods first.

$$\bar{c}_n(\mathbf{Z}) = E[\hat{c}_{kde}(\mathbf{Z})] = \frac{1}{h^d}\int K(\frac{\mathbf{z}-\mathbf{Z}}{h})c(\mathbf{z})d\mathbf{z} = \int K(s)c(\mathbf{Z}+sh)ds.$$

Hence

$$\begin{aligned} |Bias(\mathbf{Z})| = |\int K(s)c(\mathbf{Z}+sh)ds - c(\mathbf{Z})| &\le \int_{\mathbb{B}_0} K(s)|c(\mathbf{Z}+sh)-c(\mathbf{Z})|ds \\ &\le \int_{\mathbb{B}_0} K(s)M_1 h ds \\ &= M_1 h. \end{aligned} \tag{4}$$

$$\begin{aligned} |Var(\mathbf{Z})| &= \tfrac{1}{n}Var[\tfrac{1}{h^d}K(\tfrac{\mathbf{Z}_1-\mathbf{Z}}{h})] \le \tfrac{1}{n}E[\tfrac{1}{h^{2d}}K^2(\tfrac{\mathbf{Z}_1-\mathbf{Z}}{h})] \\ &= \tfrac{1}{nh^d}\int_{\mathbb{B}_0} K^2(s)c(\mathbf{Z}+sh)ds \\ &\le \tfrac{1}{nh^d}\int_{\mathbb{B}_0} K^2(s)[c(\mathbf{Z})+M_1 h]ds \\ &= \tfrac{\mu_2^2}{nh^d}[c(\mathbf{Z})+M_1 h]. \end{aligned} \tag{5}$$

Hence the integrated mean square error of the density estimator $\hat{c}_n(\mathbf{Z})$ over regions A_{M_2} is

$$\begin{aligned} IMSE(\mathbf{Z}) &= \textstyle\int_{A_{M_2}} [Bias^2(\mathbf{Z}) + Var(\mathbf{Z})] d\mathbf{Z} \\ &\leq \textstyle\int_{A_{M_2}} [M_1^2 h^2 + \frac{\mu_2^2}{nh^d}[c(\mathbf{Z}) + M_1 h] d\mathbf{Z} \ \leq M_1^2 h^2 + \frac{\mu_2^2}{nh^d}[1 + M_1 h] \\ &\leq M_1^2 h^2 + \frac{2\mu_2^2}{nh^d} \end{aligned} \tag{6}$$

Hence the error of $\widehat{RCD}$ on A_{M_2} is bounded by

$$\begin{aligned} E|T_1(\hat{c}) - T_1(c)| &\leq E \textstyle\int_{A_{M_2}} |(1 - \hat{c}_n(\mathbf{Z}))_+ - (1 - c(\mathbf{Z}))_+| d\mathbf{Z} \\ &\leq E \textstyle\int_{A_{M_2}} |\hat{c}_n(\mathbf{Z}) - c(\mathbf{Z})| d\mathbf{Z} \\ &\leq \sqrt{E \textstyle\int_{A_{M_2}} (\hat{c}_n(\mathbf{Z}) - c(\mathbf{Z}))^2 d\mathbf{Z}} \\ &\leq \sqrt{d^2 M_1^2 h^2 + \frac{2\mu_2^2}{nh^d}} \\ &\leq dM_1 h + \sqrt{2}\mu_2 (\tfrac{1}{nh^d})^{1/2}. \end{aligned}$$

Now we consider the region $A_{M_2}^c$ with unbounded copula density. For $\mathbf{Z} \in A_{M_2}^c$, $\hat{c}(\mathbf{Z})$ does not have a finite variance bound in (5). But we can bound the variance by the expectation $\bar{c}_n(\mathbf{Z}) = E[\hat{c}_n(\mathbf{Z})]$. Let $M_3 = \frac{M_2+1}{2}$, when h small, $\mathbf{Z} \in A_{M_2}^c$ implies $\mathbf{Z} + sh \in A_{M_3}^c$. Hence

$$|Var(\mathbf{Z})| \leq \frac{1}{nh^d} \int_{\mathbb{B}_0} K^2(s) c(\mathbf{Z} + sh) ds \leq \frac{M_K}{nh^d} \int_{\mathbb{B}_0} K(s) c(\mathbf{Z} + sh) ds = \frac{M_K}{nh^d} \bar{c}_n(\mathbf{Z})$$

Using Chebyshev's inequality,

$$\begin{aligned} E[1_{\{\hat{c}_n(\mathbf{Z}) < 1\}}] = P(\hat{c}_n(\mathbf{Z}) < 1) &\leq P(|\bar{c}_n(\mathbf{Z}) - \hat{c}_n(\mathbf{Z})| > \bar{c}_n(\mathbf{Z}) - 1) \\ &\leq \frac{Var[\hat{c}_n(\mathbf{Z})]}{[\hat{c}_n(\mathbf{Z}) - 1]^2} \\ &\leq \frac{M_K}{nh^d} \frac{\bar{c}_n(\mathbf{Z})}{[\bar{c}_n(\mathbf{Z}) - 1]^2} \ \leq \frac{M_K M_4}{nh^d} \end{aligned}$$

where $M_4 = \frac{M_3}{(M_3 - 1)^2}$.

Hence the error of $\widehat{RCD}$ on $A_{M_2}^c$ is bounded by

$$E|T_2(\hat{c}) - T_2(c)| = E[T_2(\hat{c})] \leq \int_{A_{M_2}^c} E[1_{\{\hat{c}_n(\mathbf{Z}) < 1\}}] d\mathbf{Z} \leq \frac{M_K M_4}{nh^d}$$

Combining the above results:

$$E[|\widehat{RCD} - RCD|] \leq M_1 h + \frac{\sqrt{2}\mu_2}{\sqrt{n} h^{d/2}} + \frac{M_K M_4}{nh^d}. \tag{7}$$

□

Note that we can use any L_p norm ($1 \leq p \leq \infty$) in the Hölder condition: equation (3) in maintext. The kernel K is then assumed to have support in the unit ball $\mathbb{B}_0$ corresponding to that L_p norm. The proof remains exactly the same. We in fact will use L_∞ norm in our estimator for computational simplicity. In that case, the unit ball $\mathbb{B}_0 = \{\mathbf{Z} : \|\mathbf{Z}\|_{l_\infty} \leq 1\}$ is in fact the d-dimensional cube.

3 Proof of Theorem 5

We can estimate RCD by plugging in the k-NN estimator [Loftsgaarden and Quesenberry, 1965] of the copula density

$$\hat{c}_{knn}(\mathbf{Z}) = \frac{\frac{k(n)}{n}}{A_{r(k(n),n),\mathbf{Z}}}, \tag{8}$$

where $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n$ are the copula based observations, $r(k(n), n)$ is the distance from $\mathbf{Z}$ to the k^{th} closest of $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n$ and $A_{r(k(n),n),\mathbf{Z}}$ is the volume of the d-dimensional hyper-ball with radius $r(k(n), n)$.

In the following, without ambiguity, we denote $r(k(n), n)$ by r, and $k(n)$ by k. Hence the volume $A_{r,\mathbf{Z}}$ is $v_d \cdot r^d$, where v_d is the volume of the d-dimensional unit ball $\mathbb{B}_0$. And $\hat{c}_{knn}(\mathbf{Z}) = k/(v_d n r^d)$. For l_2 norm, $v_d = \pi^{d/2}/\Gamma[(d+2)/2]$ where $\Gamma(\cdot)$ denotes the Gamma function.

Moore and Yackel [1977a] showed that, for bounded densities, there is equivalence between the consistency of the KDE density estimator and the consistency of the k-NN estimator.

To cite the results of [Moore and Yackel, 1977a], we assume a slightly stronger version of the Hölder condition: equation (3) in maintext. That is, we assume that c also has bounded continuous *second order* derivative in A_M. Let $Q(\mathbf{Z}) = tr[\frac{\partial^2 c(\mathbf{Z})}{\partial \mathbf{Z}^2}]$ denote the trace of the Hessian matrix of copula density $c(\mathbf{Z})$. For the d-dimensional vector $\mathbf{Z} = (z_1, ..., z_d)$, the Hessian matrix $\partial^2 c(\mathbf{Z})/\partial \mathbf{Z}^2$ has entries

$$[\frac{\partial^2 c(\mathbf{Z})}{\partial \mathbf{Z}^2}]_{ij} = \frac{\partial^2 c(\mathbf{Z})}{\partial z_i \partial z_j}.$$

We derive convergence properties for the k-NN estimator $\widehat{RCD}_{knn} = RCD(\hat{c}_{knn})$.

Theorem 5. *We assume c in $\mathfrak{C}$ has bounded and continuous second order derivative in A_M, $k \to \infty$ and $\frac{k}{n} \to 0$ when $n \to \infty$. Then the plug-in estimator $\widehat{RCD} = RCD(\hat{c})$ has a risk bound*

$$\sup_{C \in \mathfrak{C}} E[|\widehat{RCD} - RCD|] \leq 2\bar{Q}(\frac{k}{n\epsilon})^{\frac{2}{d}} + \frac{2M}{\sqrt{k}} + 2\epsilon, \tag{9}$$

where $\bar{Q} = \frac{1}{2(d+2)\pi}\Gamma^{2/d}(\frac{d+2}{2}) \sup_{\mathbf{Z} \in A_M} Q(\mathbf{Z})$, and $\epsilon = \epsilon(n)$ is any sequence converging to 0 slower than k/n. We suppress the n from the notation in ϵ without ambiguity as in k and r above.

Proof. We shall use the following asymptotic results on k-NN density estimator in Mack and Rosenblatt [1979]. Denote $\tilde{Q}(\mathbf{Z}) = \frac{1}{2(d+2)\pi}\Gamma^{2/d}(\frac{d+2}{2})Q(\mathbf{Z})$. Then

$$\begin{aligned} Bias[\hat{c}_{knn}(\mathbf{Z})] &= \frac{\tilde{Q}(\mathbf{Z})}{c(\mathbf{Z})^{2/d}}(\frac{k}{n})^{2/d} + O(\frac{c(\mathbf{Z})}{k}) + o((\frac{k}{n})^{2/d}), \\ Var[\hat{c}_{knn}(\mathbf{Z})] &= \frac{c^2(\mathbf{Z})}{k} + o(\frac{1}{k}). \end{aligned} \tag{10}$$

These expressions provide control on the error contribution of $\hat{c}(\mathbf{Z})$ to $\widehat{RCD}$ when $c(\mathbf{Z})$ is bounded both from above and from below. Similar to the proof of KDE-based $\widehat{RCD}$, we prove that the error contribution to $\widehat{RCD}$ from the big copula density region is of a smaller order $O(1/k)$. Different from the KDE, the k-NN density estimator also does not have finite bias bound in (10) when the copula density $c(\mathbf{Z})$ is not bounded below. Therefore, we also need to control the error contribution to $\widehat{RCD}$ from the small ($< \epsilon$) copula density region separately.

As before, let M_2 be a constant between 1 and M, say, $M_2 = \frac{M+1}{2}$. We now separate the three regions by copula density: $A^c_{M_2} = \{\mathbf{Z} : c(\mathbf{Z}) > M_2\}$ (big), $A_{M_2,\epsilon} = \{\mathbf{Z} : \epsilon \leq c(\mathbf{Z}) \leq M_2\}$ (middle) and $A_\epsilon = \{\mathbf{Z} : c(\mathbf{Z}) < \epsilon\}$ (small). Then we can separate RCD into three components $RCD = T_1(c) + T_2(c) + T_3(c)$: $T_1(c) = \int_{A_{M_2^c}}[1 - c(\mathbf{Z})]_+ d\mathbf{Z}$, $T_2(c) = \int_{A_{M_2,\epsilon}}[1 - c(\mathbf{Z})]_+ d\mathbf{Z}$ and $T_3(c) = \int_{A_\epsilon}[1 - c(\mathbf{Z})]_+ d\mathbf{Z}$.

Firstly, we look at the error bound on $A^c_{M_2}$, the region of big copula density. Similar to the KDE, the error in $\hat{c}_{knn}(\mathbf{Z})$ can be arbitrarily large for $\mathbf{Z} \in A^c_{M_2}$. However, the error only leads to the error in $\widehat{RCD}$ if $\hat{c}_{knn}(\mathbf{Z}) < 1$. From equation (8), $\hat{c}_{knn}(\mathbf{Z}) < 1$ if and only if

$$r > (\frac{k}{nv_d})^{1/d} \stackrel{\text{def}}{=} \bar{r}.$$

This occurs when at most $k-1$ of observations $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n$ fall into the ball $\mathbb{B}(\mathbf{Z}; \bar{r})$ which is centered at $\mathbf{Z}$ with radius $\bar{r}$.

Let $\bar{N}(\mathbf{Z})$ denotes the number of observations falling into $\mathbb{B}(\mathbf{Z}; \bar{r})$. Then $\bar{N}(\mathbf{Z})$ follows a binomial distribution with mean $n\bar{p}$, where $\bar{p} = \int_{\mathbb{B}(\mathbf{Z};\bar{r})} c(\mathbf{z})d\mathbf{z}$. Since $k/n \to 0$, $\bar{r} \to 0$. Hence $M_1\bar{r} < (M_2 - 1)/2$ when n is large enough. Then the whole ball $\mathbb{B}(\mathbf{Z}; \bar{r})$ is contained in $A^c_{M_3}$ with $M_3 = (M_2 + 1)/2$ as before. Hence $\bar{p} = \int_{\mathbb{B}(\mathbf{Z};\bar{r})} c(\mathbf{z})d\mathbf{z} \geq M_3 v_d \bar{r}^d = M_3 k/n$. Using Chebyshev's inequality,

$$\begin{aligned} Pr[\hat{c}_{knn}(\mathbf{Z}) < 1] &= E[\mathbb{1}\{\bar{N}(\mathbf{Z}) < k\}] \leq \frac{Var[\bar{N}(\mathbf{Z})]}{\{E[\bar{N}(\mathbf{Z})] - k\}^2} = \frac{n\bar{p}(1-\bar{p})}{(n\bar{p}-k)^2} \\ &\leq \frac{1}{n\bar{p}[1-k/(n\bar{p})]^2} \leq \frac{1}{M_3 k[1-1/M_3]^2} \\ &= O(\frac{1}{k}). \end{aligned}$$

Hence

$$E|T_1(c) - T_1(\hat{c}_{knn})| = \int_{A^c_{M_2}} E[1_{\{\hat{c}_{knn}(\mathbf{Z})<1\}}]d\mathbf{Z} \leq \frac{1}{M_3 k[1 - 1/M_3]^2} = O(\frac{1}{k}).$$

Secondly, we look at the error bound on $A_{M_2,\epsilon}$, the region of middle copula density.

Using (10), for $\mathbf{Z} \in A_{M_2,\epsilon}$, the mean squared error of $\hat{c}_{knn}(\mathbf{Z})$ is

$$\begin{aligned} E[(\hat{c}_n(\mathbf{Z}) - c(\mathbf{Z}))^2] &= bias^2(\mathbf{Z}) + Var(\mathbf{Z}) \\ &= [\frac{\tilde{Q}(\mathbf{Z})}{c(\mathbf{Z})^{2/d}}(\frac{k}{n})^{2/d}]^2 + \frac{c^2(\mathbf{Z})}{k} + o((\frac{k}{n})^{4/d} + \frac{1}{k}) \\ &\leq (\frac{\bar{Q}^2}{\epsilon^{4/d}})(\frac{k}{n})^{4/d} + \frac{M_2^2}{k} + o((\frac{k}{n})^{4/d} + \frac{1}{k}). \end{aligned}$$

Hence

$$E|\hat{c}_{knn}(\mathbf{Z}) - c(\mathbf{Z})| \leq \sqrt{E[(\hat{c}_n(\mathbf{Z}) - c(\mathbf{Z}))^2]} \leq \sqrt{2}[\bar{Q}(\frac{k}{n\epsilon})^{2/d} + \frac{M_2}{\sqrt{k}}][1 + o(1)].$$

We get

$$E[T_2(\hat{c}_{knn}) - T_2(c)] \leq E[\int_{A_{M_2,\epsilon}} |\hat{c}_{knn}(\mathbf{Z}) - c(\mathbf{Z})| d\mathbf{Z}] \leq \sqrt{2}[\bar{Q}(\frac{k}{n\epsilon})^{2/d} + \frac{M_2}{\sqrt{k}}][1 + o(1)]. \tag{11}$$

Thirdly, we look at the error bound on A_ϵ, the region of small copula density. From equation (8), $\hat{c}_{knn}(\mathbf{Z}) \geq 2\epsilon$ if and only if

$$r \leq (\frac{k}{n2\epsilon v_d})^{1/d} \stackrel{\text{def}}{=} r^*.$$

This occurs when at least k of observations $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n$ fall into the ball $\mathbb{B}(\mathbf{Z}; r^*)$. Since $k/(n\epsilon) \to 0$, $r^* \to 0$.

Let $N^*(\mathbf{Z})$ denotes the number of observations falling into $\mathbb{B}(\mathbf{Z}; r^*)$. Then $\bar{N}(\mathbf{Z})$ follows a binomial distribution with mean np^*, where $p^* = \int_{\mathbb{B}(\mathbf{Z};r^*)} c(\mathbf{z}) d\mathbf{z}$.

Using Taylor expansion, we have (from last line page 228 in Biau et al. [2011])

$$\int_{\mathbb{B}(\mathbf{Z};r)} c(\mathbf{z}) d\mathbf{z} = c(\mathbf{Z}) v_d r^d + \tilde{Q}(\mathbf{Z}) v_d r^{d+2} + o(r^{d+2}).$$

Therefore, using $r^* \to 0$, we have $p^* = c(\mathbf{Z}) v_d (r^*)^d [1 + o(1)] \leq \epsilon v_d (r^*)^d [1 + o(1)]$ which converges to $k/(2n)$. Hence for n big, $p^* < 0.6k/n$. Using Chebyshev's inequality,

$$\begin{aligned} Pr[\hat{c}_{knn}(\mathbf{Z}) > 2\epsilon] &= E[\mathbb{1}\{N^*(\mathbf{Z}) < k\}] \leq \tfrac{Var[N^*(\mathbf{Z})]}{[k - E(N^*(\mathbf{Z}))]^2} = \tfrac{np^*(1-p^*)}{(k-np^*)^2} \\ &\leq \tfrac{0.6k}{(0.4k)^2} < \tfrac{3}{k} \\ &= O(\tfrac{1}{k}). \end{aligned}$$

Since $c(\mathbf{Z}) \leq \epsilon$, if $\hat{c}_{knn}(\mathbf{Z}) \leq 2\epsilon$, then $|\hat{c}_{knn}(\mathbf{Z}) - c(\mathbf{Z})| \leq 2\epsilon$. Hence

$$\begin{aligned} E|T_3(c) - T_3(\hat{c}_{knn})| &\leq \textstyle\int_{A_\epsilon} E|\hat{c}_{knn}(\mathbf{Z}) - c(\mathbf{Z})| d\mathbf{Z} \leq \int_{A_\epsilon} \{2\epsilon + Pr[\hat{c}_{knn}(\mathbf{Z}) > 2\epsilon]\} d\mathbf{Z} \\ &< 2\epsilon + \tfrac{3}{k} = O(\epsilon + \tfrac{1}{k}). \end{aligned}$$

Finally, when combining the three parts, the terms $O(1/k) = o(1/\sqrt{k}) < (2 - \sqrt{2})/\sqrt{k}$. We have

$$\sup_{C \in \mathfrak{C}} E[|\widehat{RCD} - RCD|] \leq 2[\bar{Q}(\frac{k}{n\epsilon})^{2/d} + \frac{M_2}{\sqrt{k}} + \epsilon]. \tag{12}$$

□

Note that we can use other l_p norms, which changes the v_d in the proof to the volume of the unit ball under the corresponding norm. The rate does not change.

We can also prove the consistency under Hölder condition (equation (3) in main-text) without assuming continuous second derivatives. However, that involve tedious derivation of bias and variance bounds similar to (10) for k-NN density estimators. We provide the simple proof here by citing (10) from Mack and Rosenblatt [1979].

To minimize the error bound in (9), we get $\epsilon = (k/n)^{2/(d+2)}$ and $k = n^{4/(d+6)}$. So in bivariate ($d = 2$) case, we take $k = O(n^{1/2})$. Taking k below the $n^{4/(d+6)}$ rate will make the $O(1/\sqrt{k})$ term dominant in the error bound. In that case, the asymptotic results on the k-NN density estimation states that $\sqrt{k}[\hat{c}(\mathbf{Z}) - c(\mathbf{Z})]/c(\mathbf{Z})$ converge to a standard Gaussian distribution. Then $\sqrt{k}[\widehat{RCD} - RCD]$ converges to an integral of a Gaussian process.

4 Selection of Tuning Parameter in the Practical Estimator

For a practical estimator for $\widehat{RCD}$, we need to decide the bandwidth in KDE-based estimator or the number of neighbors k in KNN-based estimator. Theorem 4 and Theorem 5 provides the rates. For bivariate case, $h = O(n^{-1/4})$ and $k = O(\sqrt{n})$. To decide the constant coefficient, we used empirical simulations.

First, for KDE estimators, we tested $\widehat{RCD}$ on nine functions (listed in Table 1) with various levels of additive noises. Four sample sizes of $n = 10^2$, 10^3, 10^4 and 10^5 are used.Figure 2 plots the simulation results using $h = 0.25n^{-1/4}$. We can see that the performance of $\widehat{RCD}$ improves as sample size increases, and gives very accurate estimates for RCD under big sample sizes. For illustration, we showed the plots with bandwidth $h = 0.1n^{-1/4}$ and $h = 0.5n^{-1/4}$ in Figure 3 and Figure 4 respectively. Those bandwidth choices are clearly either too small ($h = 0.1n^{-1/4}$ estimator overshoot in several cases when RCD is small) or too big ($h = 0.5n^{-1/4}$ estimator converges slowly when RCD is large). Hence the bandwidth $h = 0.25n^{-1/4}$ is a good choice.

A	Linear	$y = x$
B	Quadratic	$y = x^2$
C	Square Root	$y = \sqrt{x}$
D	Cubic	$y = x^3$
E	Centered Cubic	$y = 4(x - 1/2)^3$
F	Centered Quadratic	$y = 4x(1 - x)$
G	Cosine (Period 1)	$y = [\cos(2\pi x) + 1]/2$
H	Circle	$(x - 1/2)^2 + y^2 = 1/4$
I	Cross	$y = \pm(x - 1/2)$

Table 1: The function relationships used in Figures 2 - 7.

According to the equivalence results between the KDE and the KNN estimator by

[Moore and Yackel, 1977b], the k in the KNN density estimation corresponds to the bandwidth in KDE estimator as $c(\mathbf{z})(2h)^2 = k/n$. As the mean of copula density $c(\alpha)$ is one, $h = 0.25n^{-1/4}$ corresponds to $k = n(2h)^2 = 0.25\sqrt{n}$. The simulation results for KNN-based $\widehat{RCD}$ with $k = 0.25\sqrt{n}$, $k = 0.1\sqrt{n}$ and $k = 0.5\sqrt{n}$ are plotted in Figures 5 - 7. Similar pattern as in KDE-based estimator are observed. Hence we propose the practical KNN-based $\widehat{RCD}$ to use $k = 0.25\sqrt{n}$.

Furthermore, we also checked the KNN-based $\widehat{RCD}$ on the mixture noise setting used in definition 2: a proportion (p) of deterministic function is hidden in independent continuous noise. Six types of deterministic function are used, as listed in Table 2. When $n = 5000$, the $\widehat{RCD}$ is close to the true value p in the simulations. And compared to the two choices of $k = 0.1\sqrt{n}$ and $k = 0.5\sqrt{n}$, $k = 0.25\sqrt{n}$ provide a good balance of approximating the true values when RCD is small or large.

A	Linear	$y = x$
B	Centered Quadratic	$y = 4(x - 1/2)^2$
C	Cosine	$y = \cos(4\pi x)$
D	Cross	$y = \pm x 1_{\{0 \leq x \leq 1\}}$
E	Circle	$(2x - 1)^2 + y^2 = 1$
F	Cross 2	$y = \pm(x - 1/2)1_{\{0 \leq x \leq 1\}}$

Table 2: The function relationships used in Figures 8.

5 The Performance of the HSNIC Estimator

In this section, we examine the performance of the convergence of the HSNIC estimator to its theoretical value CD_2 with a simulation. Let us consider the regression model $Y = X + \epsilon$ with support $[0, 1]$. For simplicity, ϵ follows uniform distribution with various bandwidth. Two estimators, the HSNIC and the KDE, are compared with increasing sample sizes. As we can see from figure 9, the x-axis is the noise level for the uniform noise ϵ, while the y-axis is the value for the estimated CD_2. The sample sizes ranges from $N = 1000$ to $N = 20000$. Results show that the HSNIC converges relatively slow J Reddi and Póczos [2013] to the true value, especially when the signal is strong. On the other hand, the KDE estimator for CD_2 becomes better when the sample size increases.

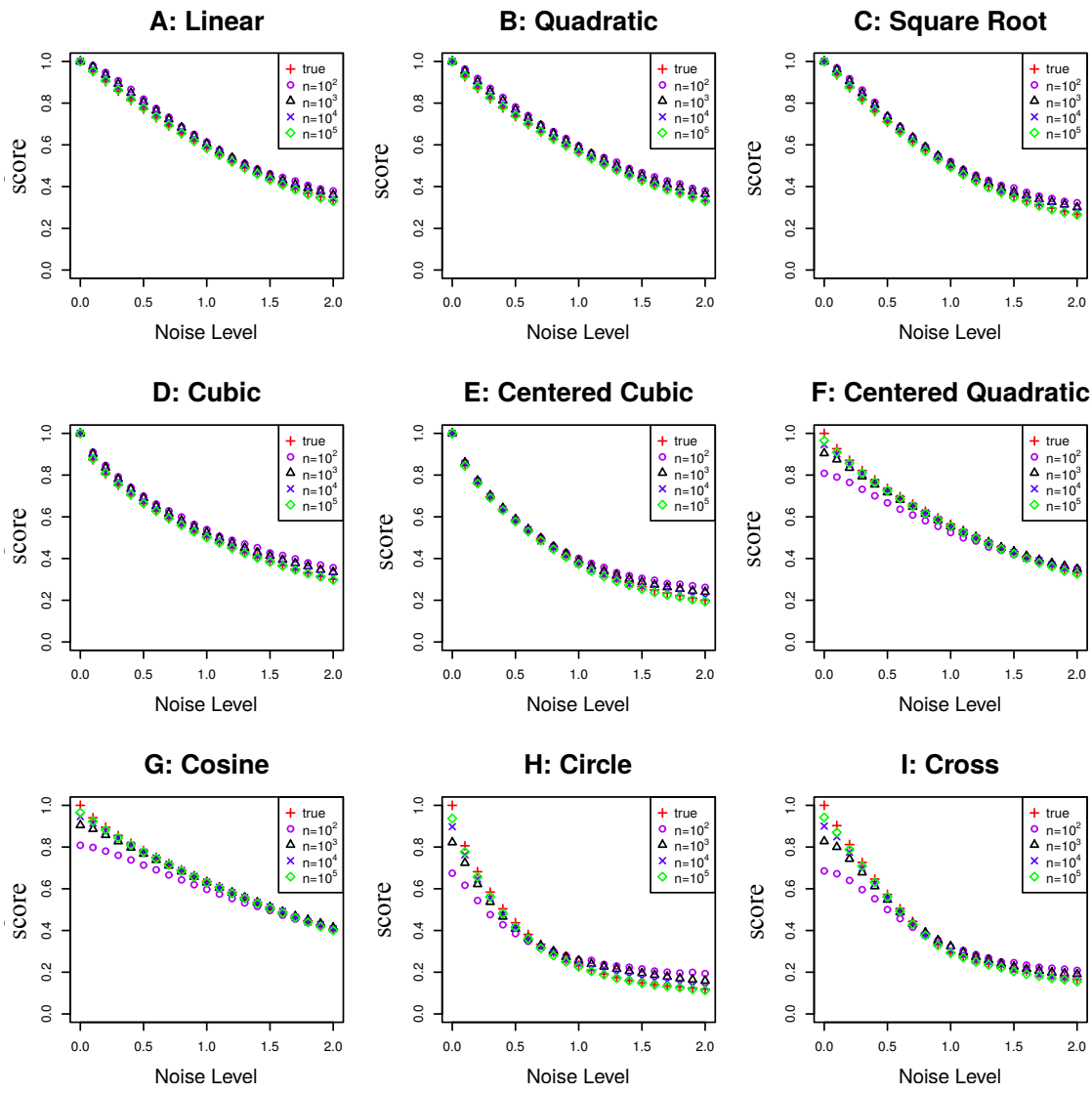

Figure 2: The comparison of RCD with its estimated values under different sample sizes. This estimator uses the square kernel density estimator with bandwidth $h = 0.25n^{-1/4}$.

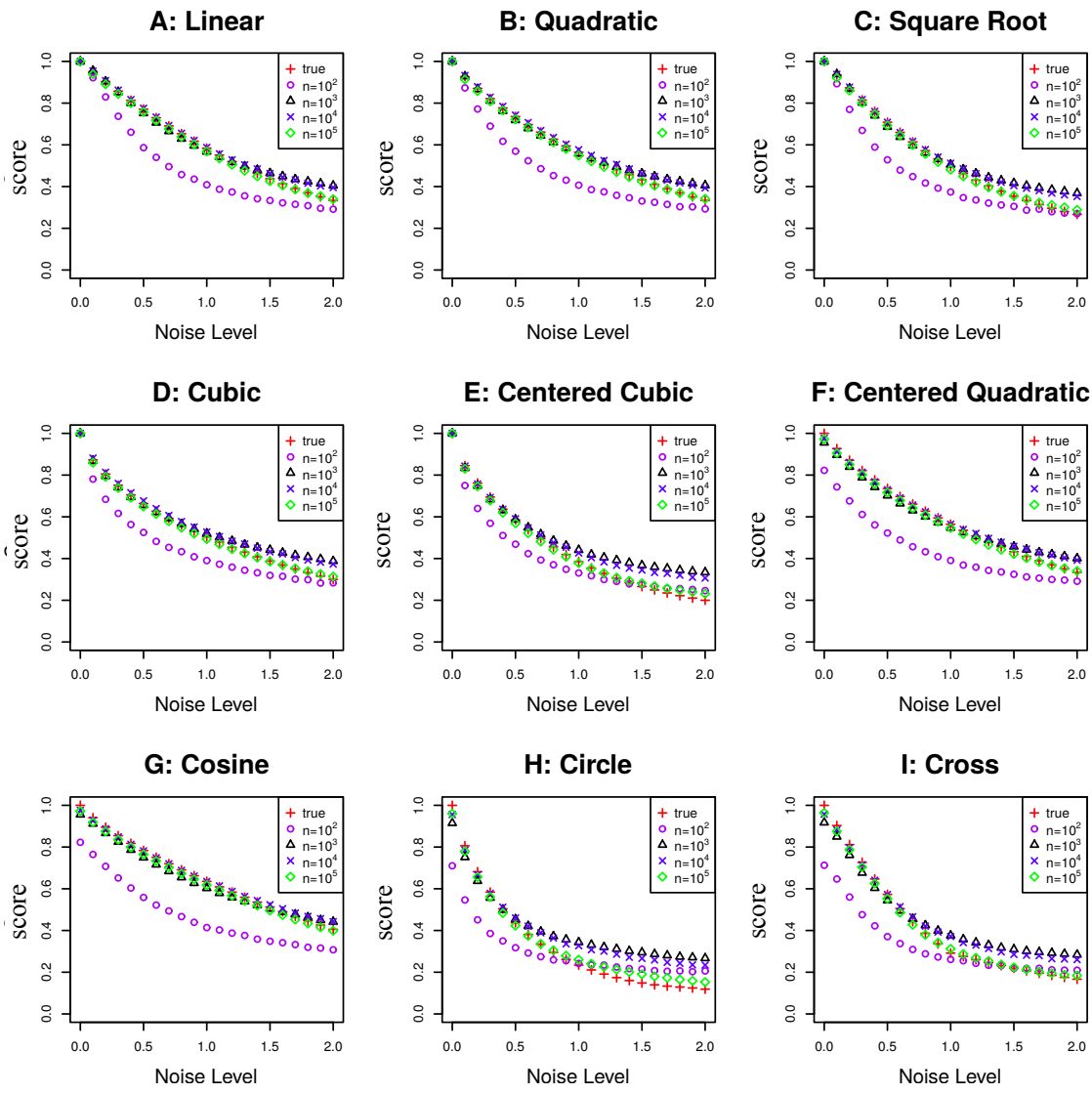

Figure 3: The comparison of RCD with its estimated values under different sample sizes. This estimator uses the square kernel density estimator with bandwidth $h = 0.1n^{-1/4}$.

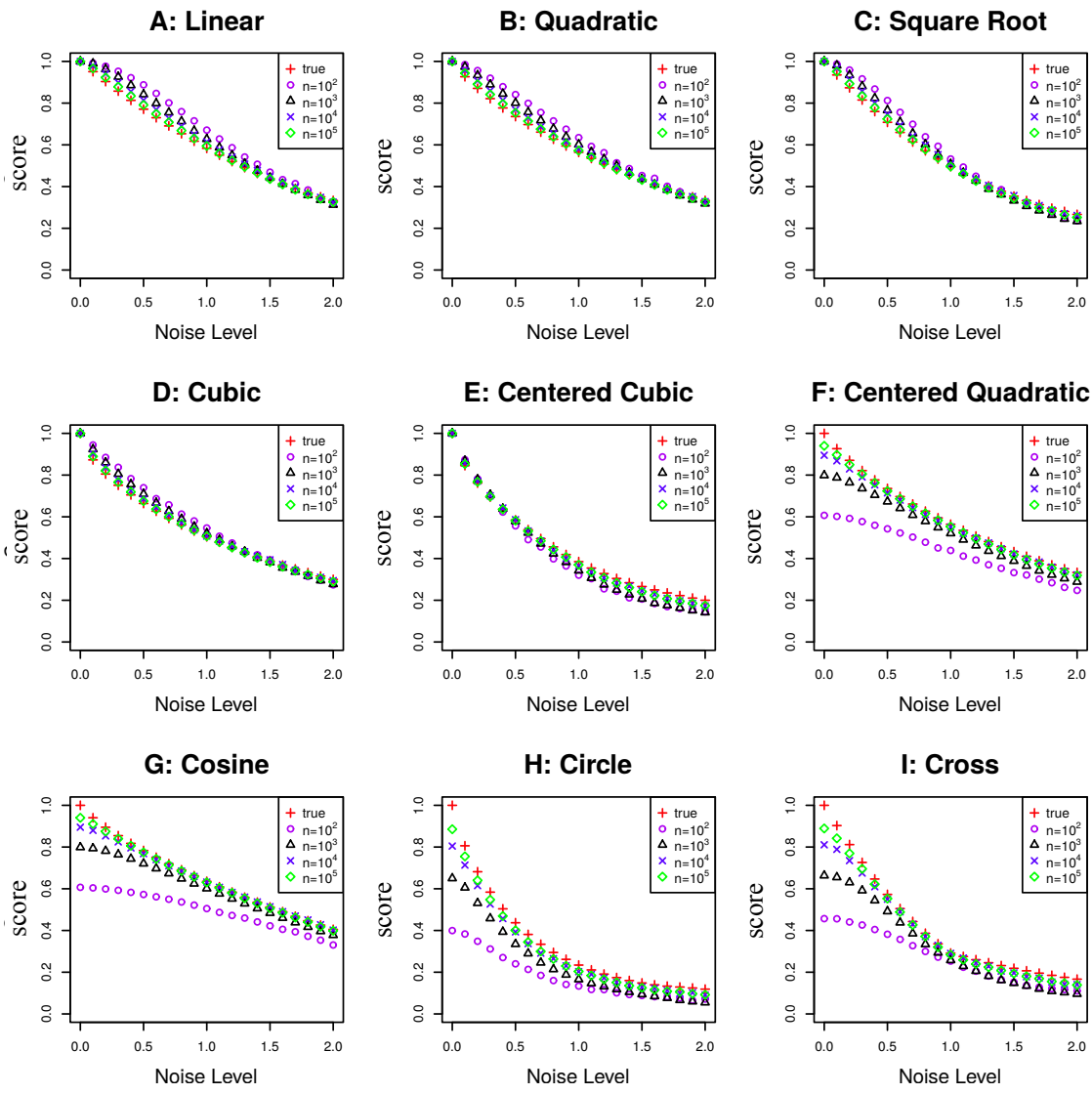

Figure 4: The comparison of RCD with its estimated values under different sample sizes. This estimator uses the square kernel density estimator with bandwidth $h = 0.5n^{-1/4}$.

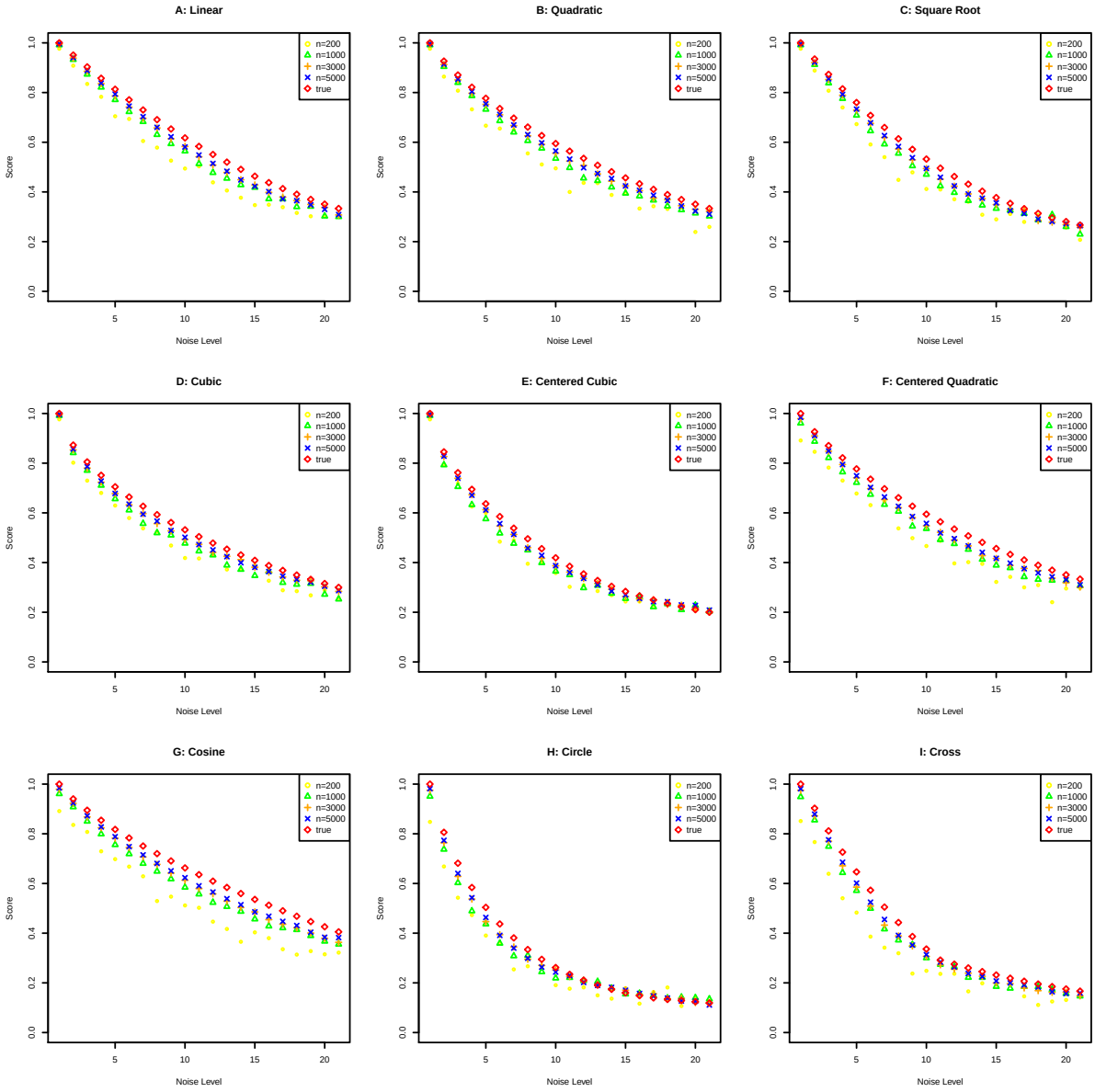

Figure 5: Additive noise with $k = 0.25\sqrt{n}$.

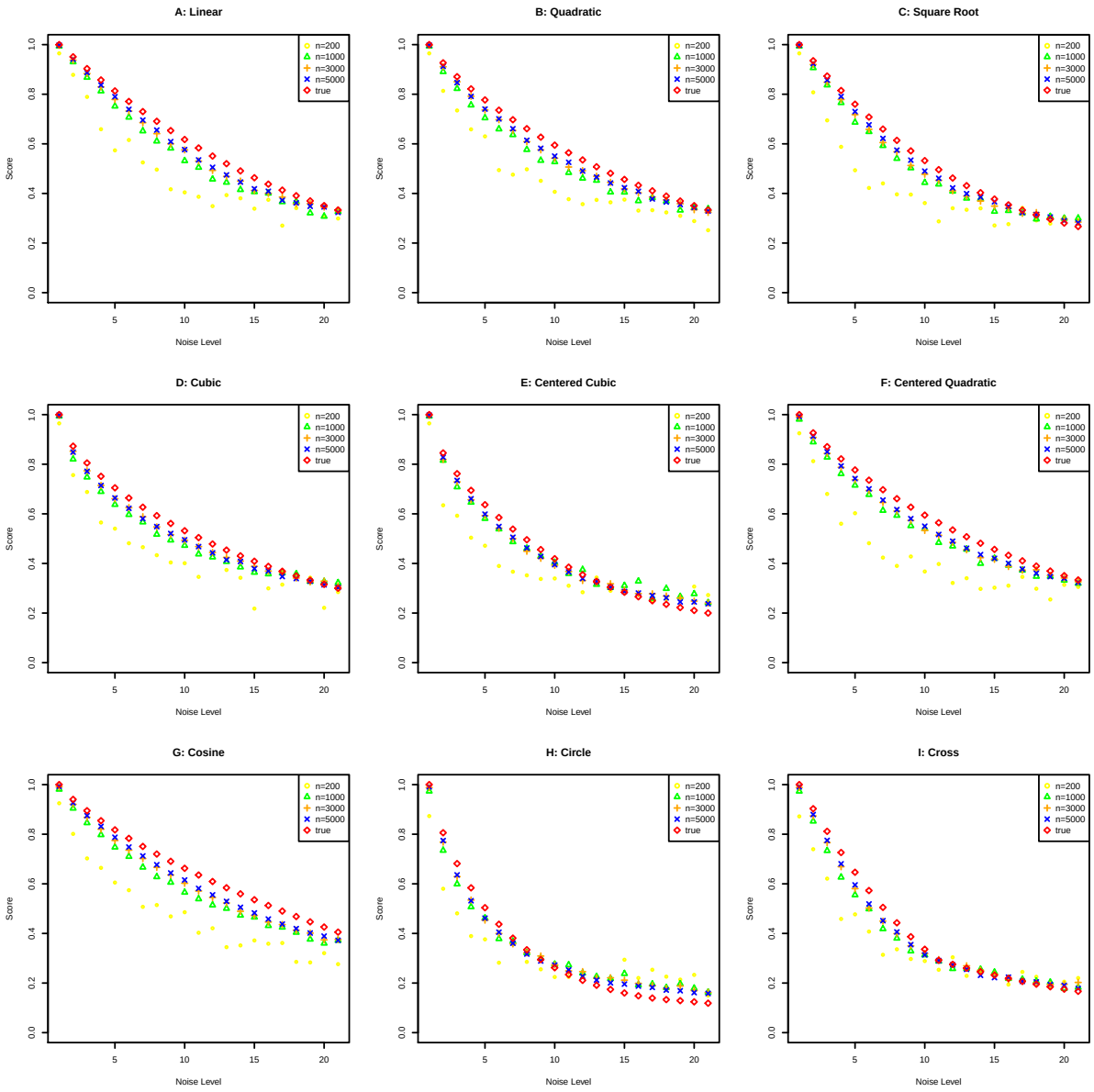

Figure 6: Additive noise with $k = 0.1\sqrt{n}$.

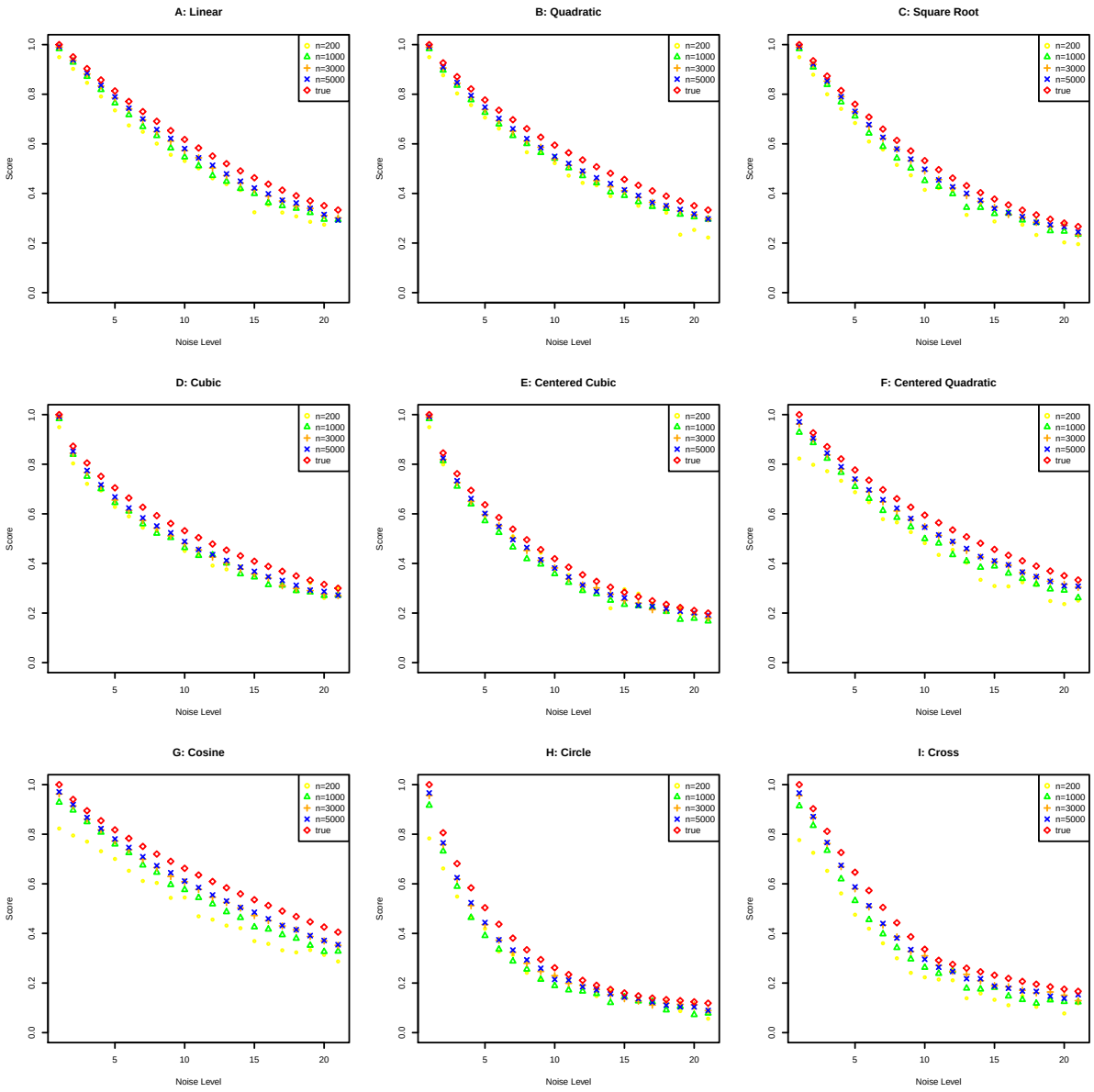

Figure 7: Additive noise with $k = 0.5\sqrt{n}$.

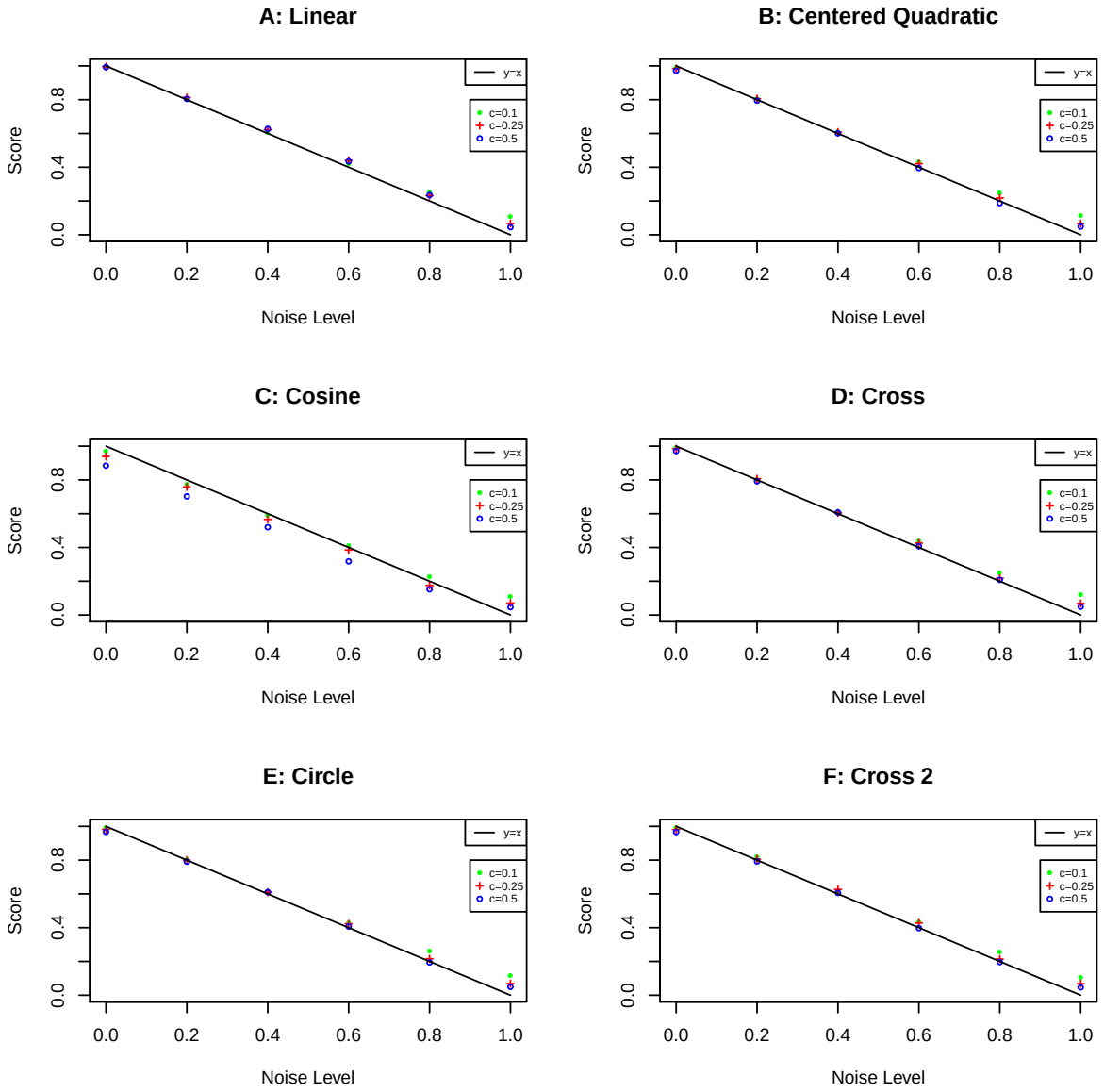

Figure 8: Mixture noise with $k = c\sqrt{n}$, where $c = 0.1, 0.25, 0.5$.

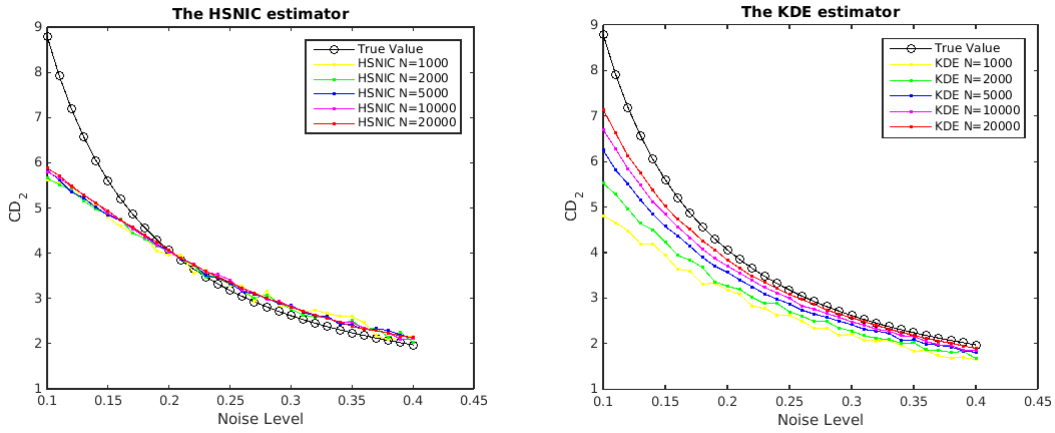

Figure 9: The comparison of the estimator of HSNIC and the KDE estimator.

References

Gérard Biau, Frédéric Chazal, David Cohen-Steiner, Luc Devroye, and Carlos Rodríguez. A weighted k-nearest neighbor density estimate for geometric inference. *Electron. J. Statist.*, 5:204–237, 2011. doi: 10.1214/11-EJS606.

David L. Donoho and Richard C. Liu. Geometrizing rates of convergence, ii. *The Annals of Statistics*, 19(2):pp. 633–667, 1991. ISSN 00905364. URL `http://www.jstor.org/stable/2242077`.

Sashank J Reddi and Barnabás Póczos. Scale invariant conditional dependence measures. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1355–1363, 2013.

Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

D. O. Loftsgaarden and C. P. Quesenberry. A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.*, 36(3):1049–1051, 06 1965. doi: 10.1214/aoms/1177700079.

Y. P. Mack and M. Rosenblatt. Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9(1):1–15, 1979.

David S. Moore and James W. Yackel. Consistency properties of nearest neighbor density function estimators. *Ann. Statist.*, 5(1):143–154, 01 1977a. doi: 10.1214/aos/1176343747.

David S. Moore and James W. Yackel. Large sample properties of nearest neighbor density function estimators. *Statistical Decision Theory and Related Topics*, II:269–279, 1977b.