Supplementary Material for "Nonparametric Variable Screening with Optimal Decision Stumps"

Jason M. Klusowski **Peter M. Tian**

In this supplement, we include the proofs which were omitted from the main text due to space constraints. In Supplement A, we first describe two technical lemmas from other papers which will be crucial for our proofs. Next, we prove Theorem 1 on linear models with Gaussian variates in Supplement B and we prove Theorem 2 on model selection consistency for general additive models in Supplement C. We prove Example 1 in Supplement D and use Theorem 1 to determine a sufficient sample size for model selection consistency (mentioned in Section 5) in Supplement E.

A PRELIMINARY LEMMAS

Our first lemma, developed by the first author in recent work, reveals the crucial role that optimization (of a nonlinear model) plays in assessing whether a particular variable is relevant or irrelevant—by relating the impurity reduction for a particular variable X_j to the sample correlation between the response variable Y and *any* function of X_j. This lemma also highlights a key departure from other approaches in past decision tree literature that do not consider splits that depend on *both* input and output data (see, for example, DSTUMP (Kazemitabar et al., 2017)).

Lemma A.1 (Lemma A.4, Supplementary Material in (Klusowski, 2020))**.** *Almost surely, uniformly over all functions $h(\cdot)$ of X_j, we have*

$$\widehat{\Delta}(X_j, Y) \geq \frac{4\widehat{\mathrm{Var}}(h(X_j))}{TV^2(h)} \times \widehat{\mathrm{Cov}}^2\left(\frac{h(X_j)}{\sqrt{\widehat{\mathrm{Var}}(h(X_j))}}, Y\right), \tag{A.1}$$

where $TV(h)$ is the total variation of $h(\cdot)$. Furthermore, almost surely, uniformly over all monotone functions $h(\cdot)$ of X_j, we have

$$\widehat{\Delta}(X_j, Y) \geq \frac{1}{1+\log(2n)} \times \widehat{\mathrm{Cov}}^2\left(\frac{h(X_j)}{\sqrt{\widehat{\mathrm{Var}}(h(X_j))}}, Y\right). \tag{A.2}$$

Remark A.1. *The bound* (A.2) *is tight (up to universal constant factors), since $\widehat{\Delta}(X_j, Y) = 1$ and $\widehat{\mathrm{Var}}(Y) \asymp \log(n)$ when, for example, $X_{1j} \leq \cdots \leq X_{nj}$, $h(X_{ij}) = Y_i$, and*

$$Y_i = \sqrt{(i-1)(n-i+1)} - \sqrt{i(n-i)}.$$

Proof sketch of Lemma A.1. For self-containment, we sketch the proof when $h(\cdot)$ is differentiable. The essential idea is to construct an empirical prior Π on the split points z and lower bound $\widehat{\Delta}(X_j, Y)$ by

$$\int \widehat{\Delta}(z; X_j, Y) d\Pi(z).$$

Recall from Section 2.4 that $N_L = N_L(z)$ and $N_R = N_R(z)$ are the number of samples in the left and right daughter nodes, respectively, if the j^{th} variable is split at z. The special prior we choose has density

$$\frac{d\Pi(z)}{dz} = \frac{|h'(z)|\sqrt{N_L(z)N_R(z)}}{\int |h'(z')|\sqrt{N_L(z')N_R(z')}dz'},$$

with support between the minimum and maximum values of the data $\{X_{ij}\}$.[1] This then yields

$$\widehat{\Delta}(X_j, Y) \geq C(h) \times \widehat{\mathrm{Cov}}^2\left(\frac{h(X_j)}{\sqrt{\widehat{\mathrm{Var}}(h(X_j))}}, Y\right), \qquad \text{where} \quad C(h) = \frac{\widehat{\mathrm{Var}}(h(X_j))}{\left(\int |h'(z')|\sqrt{N_L(z')N_R(z')}dz'\right)^2}.$$

To prove (A.1), we simply note that the denominator in $C(h)$ is at most $\mathrm{TV}^2(h)/4$. To prove (A.2), the factor $C(h)$ can be minimized (by solving a simple quadratic program) over all monotone functions $h(\cdot)$, yielding the desired result.

We direct the reader to (Klusowski, 2020, Lemma A.4, Supplementary Material) for the full proof. □

Ignoring the factor $4\widehat{\mathrm{Var}}(h(X_j))/\mathrm{TV}^2(h)$ in (A.1) and focusing only on the squared sample covariance, note that choosing $h(\cdot)$ to be the marginal projection $f_j(\cdot)$, we have

$$\widehat{\mathrm{Cov}}^2\left(\frac{f_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(f_j(X_j))}}, Y\right) \approx \mathrm{Cov}^2\left(\frac{f_j(X_j)}{\sqrt{\mathrm{Var}(f_j(X_j))}}, Y\right) = \mathrm{Var}(f_j(X_j)),$$

[1] In case $h(\cdot)$ is not differentiable, once can replace $h'(\cdot)$ above with the divided difference of $h(\cdot)$ at two adjacent X_{ij}.

where the last equality can be deduced from the fact that the marginal projection $f_j(X_j)$ is orthogonal to the residual $Y - f_j(X_j)$. Thus, in an ideal setting, Lemma A.1 enables us to asymptotically lower bound $\widehat{\Delta}(X_j, Y)$ by a multiple of the variance of the marginal projections—which can then be used to screen for important variables and control the number of false negatives.

To summarize, the previous lemma shows that $\widehat{\Delta}(X_j, Y)$ is large for variables X_j such that $g_j(X_j)$ or $f_j(X_j)$ is strongly correlated with Y. Conversely, our next lemma will be used to show that $\widehat{\Delta}(X_j, Y)$ is small when Y does not depend on X_j. A special instance of this lemma, namely, when Y is independent of X_j, was stated in (Li et al., 2019, Lemma 1) and serves as the inspiration for our proof.

Lemma A.2. *Suppose that $Z_j = Y - f_j(X_j)$ is conditionally sub-Gaussian given X_j, with variance parameter $\sigma^2_{Z_j}$, i.e., $\mathbb{E}[\exp(\lambda Z_j)|X_j] \le \exp(\lambda^2\sigma^2_{Z_j}/2)$ for all $\lambda \in \mathbb{R}$. With probability at least $1 - 4n\exp(-n\xi^2/(12\sigma^2_{Z_j}))$,*

$$\widehat{\Delta}(X_j, Y) \le 3\widehat{\mathrm{Var}}(f_j(X_j)) + \xi^2.$$

Proof. Let π be a permutation of the data such that $X_{\pi(1)j} \le X_{\pi(2)j} \le \cdots \le X_{\pi(n)j}$. Recall from the representation (6) that we have

$$\widehat{\Delta}(X_j, Y) = \max_{1\le k\le n} \frac{k}{n}\Big(1 - \frac{k}{n}\Big)\Bigg(\underbrace{\frac{1}{k}\sum_{X_{ij}\le X_{\pi(k)j}} Y_i - \frac{1}{n-k}\sum_{X_{ij} > X_{\pi(k)j}} Y_i}_{\text{(III)}}\Bigg)^2.$$

Now, since

$$\sum_{i=1}^n \left(\frac{\mathbf{1}(X_{ij} \le X_{\pi(k)j})}{k} - \frac{\mathbf{1}(X_{ij} > X_{\pi(k)j})}{n-k}\right) = 0,$$

we can rewrite (III) as

$$\underbrace{\frac{1}{k}\sum_{X_{ij}\le X_{\pi(k)j}} (Y_i - f_j(X_{ij}))}_{\text{(a)}} - \underbrace{\frac{1}{n-k}\sum_{X_{ij} > X_{\pi(k)j}} (Y_i - f_j(X_{ij}))}_{\text{(b)}}$$
$$+ \underbrace{\sum_{i=1}^n \left(f_j(X_{ij}) - \frac{1}{n}\sum_{i=1}^n f_j(X_{ij})\right)\left(\frac{\mathbf{1}(X_{ij} \le X_{\pi(k)j})}{k} - \frac{\mathbf{1}(X_{ij} > X_{\pi(k)j})}{n-k}\right)}_{\text{(c)}}.$$

Therefore, we have that

$$\begin{aligned}\widehat{\Delta}(X_j, Y) &= \max_{1\le k\le n} \frac{k}{n}\Big(1 - \frac{k}{n}\Big)(\text{(a)} - \text{(b)} + \text{(c)})^2 \\ &\le 3\max_{1\le k\le n} \frac{k}{n}\Big(1 - \frac{k}{n}\Big)\text{(a)}^2 + 3\max_{1\le k\le n} \frac{k}{n}\Big(1 - \frac{k}{n}\Big)\text{(b)}^2 + 3\max_{1\le k\le n} \frac{k}{n}\Big(1 - \frac{k}{n}\Big)\text{(c)}^2,\end{aligned} \tag{A.3}$$

where we use, in succession, the inequality $(x - y + z)^2 \le 3(x^2 + y^2 + z^2)$ for any real numbers x, y, and z, and the fact that the maximum of a sum is at most the sum of the maxima. To finish the proof, we will bound the terms involving (a)^2, (b)^2, and (c)^2 separately.

For the last term in (A.3), notice that by the Cauchy-Schwartz inequality we have

$$\begin{aligned}\frac{k}{n}\Big(1 - \frac{k}{n}\Big)\text{(c)}^2 &= \frac{k}{n}\Big(1 - \frac{k}{n}\Big)\left[\sum_{i=1}^n \Big(f_j(X_{ij}) - \frac{1}{n}\sum_{i=1}^n f_j(X_{ij})\Big) \times \left(\frac{\mathbf{1}(X_{ij} \le X_{\pi(k)j})}{k} - \frac{\mathbf{1}(X_{ij} > X_{\pi(k)j})}{n-k}\right)\right]^2 \\ &\le \frac{k}{n}\Big(1 - \frac{k}{n}\Big)\sum_{i=1}^n \Big(f_j(X_{ij}) - \frac{1}{n}\sum_{i=1}^n f_j(X_{ij})\Big)^2 \sum_{i=1}^n \left(\frac{\mathbf{1}(X_{ij} \le X_{\pi(k)j})}{k} - \frac{\mathbf{1}(X_{ij} > X_{\pi(k)j})}{n-k}\right)^2,\end{aligned}$$

which is exactly equal to

$$\frac{k}{n}\Big(1-\frac{k}{n}\Big)\Big[n\widehat{\mathrm{Var}}(f_j(X_j))\Big(k\cdot\frac{1}{k^2}+(n-k)\cdot\frac{1}{(n-k)^2}\Big)\Big]=\widehat{\mathrm{Var}}(f_j(X_j)).$$

Therefore we have shown that

$$\max_{1\le k\le n}\frac{k}{n}\Big(1-\frac{k}{n}\Big)(\mathrm{c})^2\le\widehat{\mathrm{Var}}(f_j(X_j)). \tag{A.4}$$

To bound the first term in (A.3), by a union bound we have that

$$\begin{aligned}\mathbb{P}\Big(\max_{1\le k\le n}\frac{k}{n}\Big(1-\frac{k}{n}\Big)(\mathrm{a})^2>\frac{\xi^2}{6}\Big)&\le\sum_{k=1}^n\mathbb{P}\Big(\frac{k}{n}\Big(1-\frac{k}{n}\Big)(\mathrm{a})^2>\frac{\xi^2}{6}\Big)\\&=\sum_{k=1}^n\mathbb{P}\Big(\frac{k}{n}\Big(1-\frac{k}{n}\Big)\Big(\frac{1}{k}\sum_{X_{ij}\le X_{\pi(k)j}}(Y_i-f_j(X_{ij}))\Big)^2>\frac{\xi^2}{6}\Big).\end{aligned} \tag{A.5}$$

Next, notice that, conditional on $X_{1j},\dots,X_{nj}$, $\sum_{X_{ij}\le X_{\pi(k)j}}(Y_i-f_j(X_{ij}))$ is a sum of k independent, sub-Gaussian, mean zero random variables. Thus, by the law of total probability, we have that (A.5) is equal to

$$\sum_{k=1}^n\mathbb{E}\left[\mathbb{P}\left(\left|\frac{1}{k}\sum_{X_{ij}\le X_{\pi(k)j}}(Y_i-f_j(X_{ij}))\right|>\xi\sqrt{\frac{n^2}{6k(n-k)}}\;\;\middle|\;\;X_{1j},\dots,X_{nj}\right)\right]$$

and, by Hoeffding's inequality for sub-Gaussian random variables, is bounded by

$$\sum_{k=1}^n 2\exp\Big(-k\frac{\xi^2n^2}{12k(n-k)\sigma_{Z_j}^2}\Big)\le 2n\exp\Big(-\frac{\xi^2n}{12\sigma_{Z_j}^2}\Big).$$

Note that here we have implicitly used the fact that $\mathbb{E}[\exp(\lambda Z_j)|X_j]\le\exp(\lambda^2\sigma_{Z_j}^2/2)$. It thus follows that with probability at least $1-2n\exp\big(-\frac{\xi^2n}{12\sigma_{Z_j}^2}\big)$ that

$$\max_{1\le k\le n}\frac{k}{n}\Big(1-\frac{k}{n}\Big)(\mathrm{a})^2\le\frac{\xi^2}{6}. \tag{A.6}$$

A similar argument shows that with probability at least $1-2n\exp\Big(-\frac{\xi^2n}{12\sigma_{Z_j}^2}\Big)$, the second terms in (A.3) obeys

$$\max_{1\le k\le n}\frac{k}{n}\Big(1-\frac{k}{n}\Big)(\mathrm{b})^2\le\frac{\xi^2}{6}. \tag{A.7}$$

Therefore, substituting (A.4), (A.6), and (A.7) into (A.3) and using a union bound, it follows that with probability at least $1-4n\exp\big(-\frac{\xi^2n}{12\sigma_{Z_j}^2}\big)$,

$$\widehat{\Delta}(X_j,Y)\le 3\widehat{\mathrm{Var}}(f_j(X_j))+\xi^2.$$

□

B PROOF OF THEOREM 1

The goal of this section is to prove Theorem 1. First, we include a proof sketch below to highlight the main ideas. Then in the first subsection, we rigorously prove the lower bound (9) and in the second subsection, we rigorously prove the upper bound (10).

Throughout this section, for brevity, we let $\rho_j=\mathrm{Cor}(X_j,Y)\neq 0$.

Proof sketch of Theorem 1. The first step in proving the lower bound (9) is to apply (A.2) from Lemma A.1 with $h(X_j)=X_j$ (a monotone function) to see that

$$\widehat{\Delta}(X_j,Y)\ge\frac{\widehat{\mathrm{Var}}(Y)}{\log(2n)+1}\times\widehat{\rho}^{\,2}(X_j,Y). \tag{B.1}$$

Next, we can apply asymptotic tail bounds for Pearson's sample correlation coefficient $\widehat{\rho}\,(X_j, Y)$ between two correlated Gaussian distributions (Hotelling, 1953) to show that with high probability, $|\widehat{\rho}\,(X_j, Y)| \geq (1-\delta)|\rho(X_j, Y)|$. Finally, we divide (B.1) by $\widehat{\mathrm{Var}}(Y)$, use (7), and take square roots to complete the proof of the high probability lower bound (9).

To prove the upper bound (10), notice that since X_j and Y are jointly Gaussian with mean zero, we have $f_j(X_j) = \rho_j \frac{\sigma_Y}{\sigma_{X_j}} X_j$, where $\rho_j = \rho(X_j, Y)$. Thus, by Lemma A.2 with $\sigma^2_{Z_j} = (1-\rho_j^2)\sigma_Y^2$ and $\xi^2 = (1-\rho_j^2)\sigma_Y^2\delta^2$, with probability at least $1 - 4n\exp(-n\delta^2/12)$,

$$\widehat{\Delta}(X_j, Y) \leq 3\widehat{\mathrm{Var}}(f_j(X_j)) + \delta^2(1-\rho_j^2)\sigma_Y^2 3\rho_j^2 \frac{\sigma_Y^2}{\sigma_{X_j}^2}\widehat{\mathrm{Var}}(X_j) + \delta^2(1-\rho_j^2)\sigma_Y^2. \tag{B.2}$$

We further upper bound (B.2) by obtaining high probability upper and lower bounds, respectively, for $\widehat{\mathrm{Var}}(X_j)$ and $\widehat{\mathrm{Var}}(Y)$ in terms of $\sigma^2_{X_j}$ and σ^2_Y, with a standard chi-squared concentration bound, per the Gaussian assumption. This yields that with high probability,

$$\widehat{\Delta}(X_j, Y) \lesssim \rho_j^2\widehat{\mathrm{Var}}(Y) + \delta^2\widehat{\mathrm{Var}}(Y). \tag{B.3}$$

Finally, dividing both sides of (B.3) by $\widehat{\mathrm{Var}}(Y)$, using (7), and taking square roots proves (10). □

B.1 Proof of the Lower Bound (9)

Choosing $h(X_j) = X_j$ (which is monotone) in Lemma A.1 to get that

$$\widehat{\Delta}(X_j, Y) \geq \frac{1}{\log(2n)+1} \times \widehat{\mathrm{Cov}}^2\left(\frac{X_j}{\sqrt{\widehat{\mathrm{Var}}(X_j)}}, Y\right) = \frac{\widehat{\mathrm{Var}}(Y)}{\log(2n)+1} \times \widehat{\rho}^{\,2}(X_j, Y).$$

Now observe that $\widehat{\rho}\,(X_j, Y)$ is the empirical Pearson sample correlation between two correlated normal distributions. If $\rho_j > 0$, by (Hotelling, 1953, Equation (44)), we have that

$$\begin{aligned}
\mathbb{P}(\widehat{\rho}\,(X_j, Y) > (1-\delta)\rho_j) &= 1 - \mathbb{P}(\widehat{\rho}\,(-X_j, Y) > -(1-\delta)\rho_j) \\
&\sim 1 - (2\pi)^{-1/2}\frac{\Gamma(n)}{\Gamma(n+1/2)}(1-\rho_j^2)^{n/2}(1-[(1-\delta)\rho_j]^2)^{(n-1)/2} \\
&\quad \times (-(1-\delta)\rho_j - (-\rho_j))^{-1}(1-(-\rho_j)(-(1-\delta)\rho_j))^{-n+3/2}(1+\mathcal{O}(n^{-1})).
\end{aligned} \tag{B.4}$$

If $\rho_j < 0$ we can show the same bound on $\mathbb{P}(\widehat{\rho}\,(X_j, Y) < (1-\delta)\rho_j)$. Again by (Hotelling, 1953, Equation (44)), we have the similar bound

$$\begin{aligned}
\mathbb{P}(\widehat{\rho}\,(X_j, Y) < (1-\delta)\rho_j) &= 1 - \mathbb{P}(\widehat{\rho}\,(X_j, Y) > (1-\delta)\rho_j) \\
&\sim 1 - (2\pi)^{-1/2}\frac{\Gamma(n)}{\Gamma(n+1/2)}(1-\rho_j^2)^{n/2}(1-[(1-\delta)\rho_j]^2)^{(n-1)/2} \\
&\quad \times ((1-\delta)\rho_j - \rho_j)^{-1}(1-\rho_j \times (1-\delta)\rho_j))^{-n+3/2}(1+\mathcal{O}(n^{-1})).
\end{aligned} \tag{B.5}$$

Therefore because of (B.4) and (B.5), regardless of the sign of ρ_j, it follows that there exists a universal constant C_0 for which

$$\begin{aligned}
\mathbb{P}(|\widehat{\rho}\,(X_j, Y)| > (1-\delta)|\rho_j|) &\geq 1 - \frac{C_0}{\sqrt{2\pi}\delta|\rho_j|}\frac{\Gamma(n)}{\Gamma(n+1/2)}(1-\rho_j^2)^{\frac{n}{2}}(1-(1-\delta)^2\rho_j^2)^{\frac{n-1}{2}}(1-(1-\delta)\rho_j^2)^{-n+\frac{3}{2}} \\
&\geq 1 - \frac{C_0}{\sqrt{n}\delta|\rho_j|}\exp\big(-\rho_j^2 n/2 - (1-\delta)^2\rho_j^2(n-1)/2 + (1-\delta)\rho_j^2(n-3/2)\big) \\
&= 1 - \frac{C_0}{\sqrt{n\delta^2\rho_j^2}}\exp\big(-\rho_j^2 n\delta^2/2 + \rho_j^2(1-\delta)^2/2 - 3(1-\delta)\rho_j^2/2\big) \\
&\geq 1 - \frac{C_0}{\sqrt{n\delta^2\rho_j^2}}\exp(-\rho_j^2 n\delta^2/2),
\end{aligned} \tag{B.6}$$

where we used $\exp(x) \geq 1 + x$ and Wendel's inequality (Wendel, 1948) $\frac{\Gamma(n)}{\Gamma(n+1/2)} \leq \sqrt{\frac{n+1/2}{n}} \frac{1}{\sqrt{n}} \leq \sqrt{\frac{2\pi}{n}}$ in the second inequality (B.6).

Thus, we have that with probability at least $1 - \frac{C_0}{\sqrt{n\delta^2\rho_j^2}} \exp(-\rho_j^2 n\delta^2/2)$ that

$$\widehat{\Delta}(X_j, Y) \geq \frac{(1-\delta)^2 \widehat{\mathrm{Var}}(Y)\rho_j^2}{\log(2n)+1} \iff \widehat{\rho}^2(\widehat{Y}(X_j), Y) \geq \frac{(1-\delta)^2\rho_j^2}{\log(2n)+1}.$$

This completes the first half of the proof of Theorem 1.

B.2 Proof of the Upper Bound (10)

We first state the following sample variance concentration inequality, which will be helpful.

Lemma B.1. *Let $Z_1, \ldots, Z_n$ be i.i.d. $\mathcal{N}(0, \sigma_Z^2)$. For any $0 < \delta < 1$, we have*

$$\mathbb{P}(\widehat{\mathrm{Var}}(Z) \geq (1-\delta)\frac{n-1}{n}\sigma_Z^2) \geq 1 - \exp(-\delta^2(n-1)/4) \tag{B.7}$$

and

$$\mathbb{P}(\widehat{\mathrm{Var}}(Z) \leq (1+\delta)\frac{n-1}{n}\sigma_Z^2) \geq 1 - \exp(-(n-1)(1+\delta-\sqrt{1+2\delta})/2). \tag{B.8}$$

Proof of Lemma B.1. Since Z_i are independent and normally distributed, by Cochran's theorem we have $\widehat{\mathrm{Var}}(Z) \sim \frac{\sigma_Z^2}{n}\chi_{n-1}^2$. In the notation of (Laurent and Massart, 2000), choosing $D = n-1$ and $x = \delta^2(n-1)/4$ for the chi-squared concentration inequality (4.4) in (Laurent and Massart, 2000), we have that

$$\mathbb{P}\Big(\widehat{\mathrm{Var}}(Z) \geq (1-\delta)\frac{n-1}{n}\sigma_Z^2\Big) = 1 - \mathbb{P}(\chi_{n-1}^2 < (1-\delta)(n-1))1 - \exp(-\delta^2(n-1)/4),$$

proving (B.7). For (B.8), choosing $D = n-1$ and $x = (n-1)(1+\delta-\sqrt{1+2\delta})/2$ in (Laurent and Massart, 2000, Equation (4.3)) we see that

$$\mathbb{P}\Big(\widehat{\mathrm{Var}}(Z) \leq (1+\delta)\frac{n-1}{n}\sigma_Z^2\Big) = 1 - \mathbb{P}(\chi_{n-1}^2 > (1+\delta)(n-1)) \geq 1 - \exp(-(n-1)(1+\delta-\sqrt{1+2\delta})/2).$$

□

Now we are ready to prove the upper bound (10). We begin with the inequality (B.2), as shown in the proof sketch of Theorem 1. We aim to upper bound the right hand side of (B.2) using Lemma B.1. Since the samples $X_{1j}, \ldots, X_{nj}$ are i.i.d., using (B.8) and choosing $\delta = 1$, we find that with probability at least $1 - \exp\big(-(n-1)\frac{2-\sqrt{3}}{2}\big) \geq 1 - \exp(-(n-1)/16)$, we have that $\widehat{\mathrm{Var}}(X_j) \leq 2\sigma_{X_j}^2$. Similarly, choosing $\delta = 1/2$ in (B.7), we also have that with probability at least $1 - \exp(-(n-1)/16)$ that $\widehat{\mathrm{Var}}(Y) \geq \sigma_Y^2/4$. Substituting these concentration inequalities into the right hand side of (B.2), it follows by a union bound that with probability at least $1 - 4n\exp(-n\delta^2/12) - 2\exp(-(n-1)/16)$,

$$\widehat{\Delta}(X_j, Y) \leq 24\widehat{\mathrm{Var}}(Y)\rho_j^2 + 4\delta^2\widehat{\mathrm{Var}}(Y) \iff \widehat{\rho}^2(\widehat{Y}(X_j), Y) \leq 24\rho_j^2 + 4\delta^2.$$

Finally, noticing that $\sqrt{24\rho_j^2 + 4\delta^2} < 5|\rho_j| + 2\delta$ completes the proof.

C PROOF OF THEOREM 2

The high level idea of the proof will be to show that the impurity reductions for relevant variables dominate those for irrelevant variables with high probability, meaning that relevant and irrelevant variables are correctly ranked. The following two propositions, which we prove separately in C.1 and C.2, provide high probability lower bounds on the impurity reduction for relevant variables.

C.1 Impurity Reduction Lower Bound for Relevant Variables

Our first result deals with general, smooth marginal projections. Remarkably, it shows that, with high probability, $\widehat{\Delta}(X_j, Y)$ captures a portion of the variance in the marginal projection.

Proposition C.1. *Under Assumptions 1, 2, 4, and 5, with probability at least* $1 - 3\exp(-nC_1\mathrm{Var}(f_j(X_j)))$*, we have*

$$\widehat{\Delta}(X_j, Y) \geq C_2(\mathrm{Var}(f_j(X_j)))^{1+1/d},$$

where $C_1 = c_1(B^2+\sigma^2)^{-1}$ *and* $C_2 = c_2 L^{-2/d}$ *and* c_1 *and* c_2 *are universal positive constants.*

Proof sketch. The main idea is to apply Lemma A.1 with $h(\cdot)$ equal to a trigonometric polynomial approximation $T_M(\cdot)$ to $f_j(\cdot)$. This is done to temper the effect of the factor $4\widehat{\mathrm{Var}}(h(X_j))/(\int_0^1 |h'(x)|dx)^2$ from Lemma A.1.

To construct such a function, we employ a Jackson-type estimate in conjunction with Assumption 2 to show the existence of a good trigonometric polynomial approximation $T_M(x) = a_0 + \sum_{k=1}^M a_k\sqrt{2}\cos(2\pi kx) + \sum_{k=1}^M b_k\sqrt{2}\sin(2\pi kx)$ (of degree M) to $f_j(\cdot)$. Because $T_M(\cdot)$ is a sum of orthogonal functions,

$$\mathrm{Var}(T_M(X_j)) = \sum_{k=1}^M (a_k^2 + b_k^2), \tag{C.1}$$

and

$$\Big(\int_0^1 |T_M'(x)|dx\Big)^2 \leq \int_0^1 |T_M'(x)|^2 dx = (2\pi)^2 \sum_{k=1}^M k^2(a_k^2+b_k^2). \tag{C.2}$$

Combining (C.1) and (C.2), we find that, with high probability,

$$\frac{\widehat{\mathrm{Var}}(T_M(X_j))}{(\int_0^1 |T_M'(x)|dx)^2} \approx \frac{\mathrm{Var}(T_M(X_j))}{(\int_0^1 |T_M'(x)|dx)^2} \geq \frac{\sum_{k=1}^M (a_k^2+b_k^2)}{(2\pi)^2\sum_{k=1}^M k^2(a_k^2+b_k^2)} \geq \frac{1}{(2\pi)^2 M^2}.$$

Plugging these values into the lower bound (A.1) in Lemma A.1 and choosing $M \asymp (\mathrm{Var}(f_j(X_j)))^{-1/(2d)}$, we find that with high probability,

$$\widehat{\Delta}(X_j, Y) \gtrsim (\mathrm{Var}(f_j(X_j)))^{1/d} \times \widehat{\mathrm{Cov}}^2\left(\frac{T_M(X_j)}{\sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}}, Y\right). \tag{C.3}$$

Thus, the lower bound in Proposition C.1 will follow if we can show that the squared sample covariance factor in (C.3) exceeds $\mathrm{Var}(f_j(X_j))$ with high probability. To this end, note that

$$\widehat{\mathrm{Cov}}\left(\frac{T_M(X_j)}{\sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}}, Y\right) = \underbrace{\widehat{\mathrm{Cov}}\left(\frac{T_M(X_j)}{\sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}}, f_j(X_j)\right)}_{\text{(I)}} + \underbrace{\widehat{\mathrm{Cov}}\left(\frac{T_M(X_j)}{\sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}}, Y - f_j(X_j)\right)}_{\text{(II)}}. \tag{C.4}$$

With high probability, (I) can be lower bounded by $C\sqrt{\mathrm{Var}(f_j(X_j))}$, per the choice M from above, using the approximation properties of $T_M(\cdot)$ for $f_j(\cdot)$ and a concentration inequality for the sample variance of $f_j(X_j)$. Here, C is some positive constant. Furthermore, since $Y - f_j(X_j)$ is orthogonal to any function of X_j, a Hoeffding type concentration inequality shows that (II) is larger than any (strictly) negative constant, including $-(C/2)\sqrt{\mathrm{Var}(f_j(X_j))}$, with high probability. Combining this analysis from (C.3) and (C.4), we obtain the high probability lower bound on $\widehat{\Delta}(X_j, Y)$ given in Proposition C.1. □

Now we present the proof of Proposition C.1 more rigorously. First we will present several key ideas and lemmas which are used in the proof. At the end of the section, we will complete the proof of Proposition C.1.

First, we state and prove a lemma that will be used in later proofs. Though stated in terms of general probability measures, we will be specifically interested in the case where $\mathbb{P}$ is the empirical probability measure $\mathbb{P}_n$ and $\mathbb{E}$ is the empirical expectation $\mathbb{E}_n$, both with respect to a sample of size n.

Lemma C.1. *For any random variables U and V with finite second moments with respect to a probability measure $\mathbb{P}$,*

$$\mathrm{Cov}_{\mathbb{P}}\Big(\frac{U}{\sqrt{\mathrm{Var}_{\mathbb{P}}(U)}}, V\Big) \geq \sqrt{\mathrm{Var}_{\mathbb{P}}(V)} - 2\sqrt{\mathbb{E}_{\mathbb{P}}[(U-V)^2]}.$$

Proof of Lemma C.1. First note that by the triangle inequality,

$$\big|\sqrt{\mathrm{Var}_{\mathbb{P}}(U)} - \sqrt{\mathrm{Var}_{\mathbb{P}}(V)}\big| \leq \sqrt{\mathrm{Var}_{\mathbb{P}}(U-V)} \leq \sqrt{\mathbb{E}_{\mathbb{P}}[(U-V)^2]}. \tag{C.5}$$

To complete the proof, we write

$$\mathrm{Cov}_{\mathbb{P}}(U,V) = \sqrt{\mathrm{Var}_{\mathbb{P}}(U)}\sqrt{\mathrm{Var}_{\mathbb{P}}(V)} + \big(\mathrm{Cov}_{\mathbb{P}}(U,V) - \sqrt{\mathrm{Var}_{\mathbb{P}}(U)}\sqrt{\mathrm{Var}_{\mathbb{P}}(V)}\big) \tag{C.6}$$

and apply (C.5) to arrive at

$$\begin{aligned}
&\big|\mathrm{Cov}_{\mathbb{P}}(U,V) - \sqrt{\mathrm{Var}_{\mathbb{P}}(U)}\sqrt{\mathrm{Var}_{\mathbb{P}}(V)}\big| && \text{(C.7)}\\
&= \big|\mathrm{Cov}_{\mathbb{P}}(U,V) - \mathrm{Var}_{\mathbb{P}}(U) + \sqrt{\mathrm{Var}_{\mathbb{P}}(U)}\big(\sqrt{\mathrm{Var}_{\mathbb{P}}(U)} - \sqrt{\mathrm{Var}_{\mathbb{P}}(V)}\big)\big| \\
&\leq \big|\mathrm{Cov}_{\mathbb{P}}(U,V) - \mathrm{Var}_{\mathbb{P}}(U)\big| + \sqrt{\mathrm{Var}_{\mathbb{P}}(U)} \times \big|\sqrt{\mathrm{Var}_{\mathbb{P}}(U)} - \sqrt{\mathrm{Var}_{\mathbb{P}}(V)}\big| \\
&\leq 2\sqrt{\mathrm{Var}_{\mathbb{P}}(U)} \times \big|\sqrt{\mathrm{Var}_{\mathbb{P}}(U)} - \sqrt{\mathrm{Var}_{\mathbb{P}}(V)}\big| && \text{(C.8)}\\
&\leq 2\sqrt{\mathrm{Var}_{\mathbb{P}}(U)}\sqrt{\mathbb{E}_{\mathbb{P}}[(U-V)^2]},
\end{aligned}$$

where the penultimate line (C.8) follows from the Cauchy-Schwarz inequality. Substituting (C.7) into (C.6), we get

$$\mathrm{Cov}_{\mathbb{P}}(U,V) \geq \sqrt{\mathrm{Var}_{\mathbb{P}}(U)}\big(\sqrt{\mathrm{Var}_{\mathbb{P}}(V)} - 2\sqrt{\mathbb{E}_{\mathbb{P}}[(U-V)^2]}\big),$$

which proves the claim. □

The following sample variance concentration inequality will also come in handy.

Lemma C.2 (Equation 5, (Maurer and Pontil, 2009))**.** *Let U be a random variable bounded by B. Then for all $\gamma > 0$,*

$$\mathbb{P}\Big(\frac{n}{n-1}\widehat{\mathrm{Var}}(U) \geq \mathrm{Var}(U) - \gamma\Big) \geq 1 - \exp\Big(-\frac{(n-1)\gamma^2}{8B^2\mathrm{Var}(U)}\Big).$$

As explained in the main text, the key step in the proof of Proposition C.1 is to apply Lemma A.1 with a good trigonometric polynomial approximation $T_M(\cdot)$ to the marginal projection $f_j(\cdot)$.

Notice that we can extend $f_j(\cdot)$ to $[-1,2]$ so that $f_j(-1) = f_j(2)$ while also preserving Assumption 2. Then by a Jackson-type estimate for trigonometric polynomials (Korneichuk, 1991, Section 6.2.4), there exists a trigonometric polynomial $T_M(x) = a_0 + \sum_{k=1}^M a_k\sqrt{2}\cos(2\pi kx) + \sum_{k=1}^M b_k\sqrt{2}\sin(2\pi kx)$ such that $\sup_{x\in[0,1]}|T_M(x) - f_j(x)| \leq KL(M+1)^{-d}$, where K is a universal positive constant.

Next, we set $M = \lfloor (4^{-1}\tau^2(KL)^{-2}\mathrm{Var}(f_j(X_j)))^{-1/(2d)}\rfloor$, where τ is a constant less than $1/4$, so that $\sup_{x\in[0,1]}|T_M(x) - f_j(x)| \leq KL(M+1)^{-d} \leq \frac{\tau\sqrt{\mathrm{Var}(f_j(X_j))}}{2}$. Since $f_j(\cdot)$ is bounded in magnitude by B, the approximation properties of $T_M(\cdot)$ imply that $T_M(\cdot)$ is bounded in magnitude by $B_0 = (1+\tau/2)B$. By Lemma C.2 with $\gamma = (1/2)\mathrm{Var}(T_M(X_j))$, we find that

$$\widehat{\mathrm{Var}}(T_M(X_j)) \geq \frac{n-1}{2n}\mathrm{Var}(T_M(X_j)) \geq \frac{1}{4}\mathrm{Var}(T_M(X_j)), \tag{C.9}$$

with probability at least $1 - \exp\Big(-\frac{(n-1)\mathrm{Var}(T_M(X_j))}{32B_0^2}\Big)$. The same computations as the proof sketch of Proposition C.1 in the main text yield $\frac{\mathrm{Var}(T_M(X_j))}{\mathrm{TV}^2(T_M)} \geq \frac{\sum_{k=1}^M(a_k^2+b_k^2)}{\sum_{k=1}^M k^2(a_k^2+b_k^2)} \geq \frac{1}{(2\pi)^2M^2}$. Thus, by (C.9), with probability at least $1 - \exp\Big(-\frac{(n-1)\mathrm{Var}(T_M(X_j))}{32B_0^2}\Big)$, we have

$$\widehat{\Delta}(X_j,Y) \geq \frac{4\widehat{\mathrm{Var}}(T_M(X_j))}{\mathrm{TV}^2(T_M)} \times \widehat{\mathrm{Cov}}^2\left(\frac{T_M(X_j)}{\sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}}, Y\right) \geq \frac{1}{(2\pi)^2M^2} \times \widehat{\mathrm{Cov}}^2\left(\frac{T_M(X_j)}{\sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}}, Y\right). \tag{C.10}$$

The choice of M and the uniform approximation properties of $T_M(\cdot)$ above gives

$$\sqrt{\mathbb{E}_n[(f_j(X_j) - T_M(X_j))^2]} \le KL(M+1)^{-d} \le \frac{\tau\sqrt{\mathrm{Var}(f_j(X_j))}}{2}. \tag{C.11}$$

It follows by Lemma C.1 along with (C.11) that

$$\widehat{\mathrm{Cov}}(T_M(X_j), f_j(X_j)) \ge \sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}\Big(\sqrt{\widehat{\mathrm{Var}}(f_j(X_j))} - \tau\sqrt{\mathrm{Var}(f_j(X_j))}\Big). \tag{C.12}$$

In the next lemma, we use (C.12) along with Lemma C.2 to obtain a lower bound on the squared covariance factor in (C.10).

Lemma C.3. *With probability at least* $1 - \exp\Big(-\frac{(n-1)(1-8\tau^2)^2\mathrm{Var}(f_j(X_j))}{8B^2}\Big) - \exp\Big(-\frac{n\tau^2\mathrm{Var}(f_j(X_j))}{8(B^2+\sigma^2)}\Big)$*, we have that*

$$\widehat{\mathrm{Cov}}\left(\frac{T_M(X_j)}{\sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}}, Y\right) \ge (\tau/2)\sqrt{\mathrm{Var}(f_j(X_j))}.$$

Proof of Lemma C.3. Recalling (C.4), we will prove Lemma C.3 by first getting a concentration bound on (I) and then getting a concentration bound on (II).

To get a concentration bound on (I), we need Lemma C.2 to lower bound the sample variance on the right hand side of inequality (C.12). Choosing $U = f_j(X_j) \in [-B, B]$ and $\gamma = \mathrm{Var}(f_j(X_j))(1-8\tau^2)$ (which is greater than zero by assumption that $\tau < 1/4$), notice that Lemma C.2 gives us

$$\mathbb{P}\Big(\widehat{\mathrm{Var}}(f_j(X_j)) \ge \frac{8\tau^2(n-1)}{n}\mathrm{Var}(f_j(X_j))\Big) \ge 1 - \exp\Big(-\frac{(n-1)(1-8\tau^2)^2\mathrm{Var}(f_j(X_j))}{8B^2}\Big),$$

so that by (C.12),

$$\begin{aligned}\widehat{\mathrm{Cov}}\left(\frac{T_M(X_j)}{\sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}}, f_j(X_j)\right) &\ge \sqrt{\widehat{\mathrm{Var}}(f_j(X_j))} - \tau\sqrt{\mathrm{Var}(f_j(X_j))} \\ &\ge \tau(\sqrt{8}\sqrt{1-1/n} - 1)\sqrt{\mathrm{Var}(f_j(X_j))} \\ &\ge \tau\sqrt{\mathrm{Var}(f_j(X_j))},\end{aligned}$$

with probability at least $1 - \exp\Big(-\frac{(n-1)(1-8\tau^2)^2\mathrm{Var}(f_j(X_j))}{8B^2}\Big)$.

Now we need to get a concentration bound for (II). Let $s_i = \frac{T_M(X_{ij}) - \frac{1}{n}\sum_{k=1}^n T_M(X_{kj})}{\sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}}$. We need to bound

$$\widehat{\mathrm{Cov}}\left(\frac{T_M(X_j)}{\sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}}, Y - f_j(X_j)\right) = \frac{1}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i) - f_j(X_{ij}) + \varepsilon_i).$$

For notational simplicity, we let $\underline{\mathbf{X}} = (X_{ij})$ be the $n \times p$ data matrix with $\mathbf{X}_i$ as rows. First notice that

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i) - f_j(X_{ij}) + \varepsilon_i)\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i) - f_j(X_{ij}) + \varepsilon_i)\right)\middle|\underline{\mathbf{X}}\right]\right] \\ &= \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i) - f_j(X_{ij}))\right)\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n s_i\varepsilon_i\right)\middle|\underline{\mathbf{X}}\right]\right].\end{aligned}$$

Now, by the sample independence of the errors ϵ_i, we can write the above as

$$\begin{aligned}
&\mathbb{E}\bigg[\exp\bigg(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i)-f_j(X_{ij}))\bigg)\prod_{i=1}^n \mathbb{E}\Big[\exp\Big(\frac{\lambda}{n}s_i\varepsilon_i\Big)\Big|\underline{\mathbf{X}}\Big]\bigg]\\
&\leq \mathbb{E}\bigg[\exp\bigg(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i)-f_j(X_{ij}))\bigg)\prod_{i=1}^n \exp\Big(\frac{\lambda^2 s_i^2\sigma^2}{2n^2}\Big)\bigg]\\
&= \mathbb{E}\bigg[\exp\bigg(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i)-f_j(X_{ij}))\bigg)\bigg]\exp\Big(\frac{\lambda^2\sigma^2}{2n}\Big),
\end{aligned}$$

where we used Assumption 5 and the fact that $\frac{1}{n}\sum_{i=1}^n s_i^2 = 1$. Recalling that s_i depends on $(X_{1j}, X_{2j}, \ldots, X_{nj})^\top$, we have

$$\begin{aligned}
\mathbb{E}\bigg[\exp\bigg(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i)-f_j(X_{ij}))\bigg)\bigg] &= \mathbb{E}\bigg[\mathbb{E}\bigg[\exp\bigg(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i)-f_j(X_{ij}))\bigg)\bigg|X_{1j}, X_{2j}, \ldots, X_{nj}\bigg]\bigg]\\
&= \mathbb{E}\bigg[\prod_{i=1}^n \mathbb{E}\Big[\exp\Big(\frac{\lambda}{n}s_i(g(\mathbf{X}_i)-f_j(X_{ij}))\Big)\Big|X_{1j}, X_{2j}, \ldots, X_{nj}\Big]\bigg],
\end{aligned} \tag{C.13}$$

where we used sample independence in the second equality. Finally, applying Hoeffding's Lemma along with the fact that $\|g\|_\infty \leq B$, we have that (C.13) is bounded above by

$$\mathbb{E}\bigg[\prod_{i=1}^n \exp\Big(\frac{\lambda^2 s_i^2 B^2}{2n^2}\Big)\bigg] \leq \exp\Big(\frac{\lambda^2 B^2}{2n}\Big).$$

Having bounded the moment generating function, we can now apply Markov's inequality to see that

$$\begin{aligned}
\mathbb{P}\bigg(\frac{1}{n}\sum_{i=1}^n s_i(g(X)-f_j(X_j)+\varepsilon_i) \leq -\gamma\bigg) &= \mathbb{P}\bigg(\exp\bigg(-\frac{\lambda}{n}\sum_{i=1}^n s_i(g(X)-f_j(X_j)+\varepsilon_i)\bigg) \geq \exp(\lambda\gamma)\bigg)\\
&\leq \exp\bigg(\frac{\lambda^2(B^2+\sigma^2)}{2n} - \gamma\lambda\bigg)\\
&\leq \exp\bigg(-\frac{n\gamma^2}{2(B^2+\sigma^2)}\bigg),
\end{aligned}$$

where the last inequality follows by maximizing over λ. Choosing $\gamma = (\tau/2)\sqrt{\mathrm{Var}(f_j(X_j))}$, we have by a union bound that,

$$\begin{aligned}
\widehat{\mathrm{Cov}}\bigg(\frac{T_M(X_j)}{\sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}}, Y\bigg) &= \widehat{\mathrm{Cov}}\bigg(\frac{T_M(X_j)}{\sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}}, f_j(X_j)\bigg) + \widehat{\mathrm{Cov}}\bigg(\frac{T_M(X_j)}{\sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}}, Y - f_j(X_j)\bigg)\\
&\geq \tau\sqrt{\mathrm{Var}(f_j(X_j))} - (\tau/2)\sqrt{\mathrm{Var}(f_j(X_j))}\\
&= (\tau/2)\sqrt{\mathrm{Var}(f_j(X_j))},
\end{aligned} \tag{C.14}$$

with probability at least $1 - \exp\Big(-\frac{(n-1)(1-8\tau^2)^2\mathrm{Var}(f_j(X_j))}{8B^2}\Big) - \exp\Big(-\frac{n\tau^2\mathrm{Var}(f_j(X_j))}{8(B^2+\sigma^2)}\Big)$. □

With this setup, we are now ready to finish the proof of Proposition C.1.

Proof of Proposition C.1. By the triangle inequality, the approximation properties of $T_M(\cdot)$, and the choice of M,

$$\sqrt{\mathrm{Var}(T_M(X_j))} \geq \sqrt{\mathrm{Var}(f_j(X_j))} - \sqrt{\mathbb{E}[(f_j(X_j)-T_M(X_j))^2]} \geq (1-\tau/2)\sqrt{\mathrm{Var}(f_j(X_j))}.$$

Therefore, from (C.10), with probability at least $1-\exp\Big(-\frac{(n-1)\mathrm{Var}(T_M(X_j))}{32B_0^2}\Big) \geq 1-\exp\Big(-\frac{(n-1)(1-\tau/2)^2\mathrm{Var}(f_j(X_j))}{32B_0^2}\Big)$,

$$\widehat{\Delta}(X_j,Y) \geq \frac{1}{(2\pi)^2M^2} \times \widehat{\mathrm{Cov}}^2\left(\frac{T_M(X_j)}{\sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}},Y\right). \tag{C.15}$$

Combining (C.14) and (C.15) with a union bound, it follows that with probability at least $1-\exp\Big(-\frac{(n-1)(1-8\tau^2)^2\mathrm{Var}(f_j(X_j))}{8B^2}\Big)-\exp\Big(-\frac{n\tau^2\mathrm{Var}(f_j(X_j))}{8(B^2+\sigma^2)}\Big)-\exp\Big(-\frac{(n-1)(1-\tau/2)^2\mathrm{Var}(f_j(X_j))}{32B_0^2}\Big) \geq 1-3\exp(-C_1n\mathrm{Var}(f_j(X_j)))$ that

$$\begin{aligned}\widehat{\Delta}(X_j,Y) &\geq \frac{1}{(2\pi)^2M^2} \times \widehat{\mathrm{Cov}}^2\left(\frac{T_M(X_j)}{\sqrt{\widehat{\mathrm{Var}}(T_M(X_j))}},Y\right)\\ &\geq \frac{4^{-1}\tau^2\mathrm{Var}(f_j(X_j))}{(2\pi)^2\lfloor(4^{-1}\tau^2(KL)^{-2}\mathrm{Var}(f_j(X_j)))^{-1/(2d)}\rfloor^2}\\ &\geq C_2(\mathrm{Var}(f_j(X_j)))^{1+1/d},\end{aligned}$$

where $C_1 = c_1(B^2+\sigma^2)^{-1}$ and $C_2 = c_2L^{-2/d}$ and c_1 and c_2 are universal positive constants. □

C.2 Impurity Reduction Upper Bound for Irrelevant Variables

Next, we need to ensure that there is a sufficient separation in the impurity reductions between relevant and irrelevant variables. To do so, we use Lemma A.2 along with the partial orthogonality assumption in Section 4.1 to show that the impurity reductions for irrelevant variables will be small with high probability.

Lemma C.4. *Under Assumptions 1, 3 and 5, for each $j \in \mathcal{S}^c$, with probability at least $1-4n\exp(-n\xi^2/(12(B^2+\sigma^2)))$,*

$$\widehat{\Delta}(X_j,Y) \leq \xi^2.$$

In other words, if $j \in \mathcal{S}^c$, then $\widehat{\Delta}(X_j,Y) = \mathcal{O}(n^{-1}\log(n))$ with probability at least $1-n^{-\Omega(1)}$.

Proof of Lemma C.4. Observe that

$$\begin{aligned}\mathbb{E}[\exp(\lambda(Y-f_j(X_j)))|X_j] &= \mathbb{E}\big[\exp(\lambda(g(\mathbf{X})-f_j(X_j)))\mathbb{E}\big[\exp(\lambda\varepsilon)\big|\mathbf{X}\big]\big|X_j\big]\\ &\leq \mathbb{E}\big[\exp(\lambda(g(\mathbf{X})-f_j(X_j)))\big|X_j\big]\exp(\lambda^2\sigma^2/2) &\text{(C.16)}\\ &\leq \mathbb{E}\big[\exp(\lambda^2(B^2+\sigma^2)/2)\big], &\text{(C.17)}\end{aligned}$$

where we used Assumption 5 in the penultimate inequality (C.16) and Hoeffding's Lemma together with Assumption 1 in the last inequality (C.17). Using Assumption 3 along with Lemma A.2 with $\sigma_{Z_j}^2 = B^2+\sigma^2$ proves Lemma C.4. □

C.3 Finishing the Proof of Theorem 2

In this section, we use Proposition C.1 along with Lemma C.4 to complete the proof of Theorem 2.

Proof of Theorem 2. The high-level idea is to show that the upper and lower bounds on the impurity reductions for irrelevant and relevant variables from Lemma C.4 and Proposition C.1, respectively, are well-separated.

By Proposition C.1 for all variables $j \in \mathcal{S}$ and a union bound, we see that with probability at least $1-3s\exp(-C_1nv)$, we have

$$\widehat{\Delta}(X_j,Y) \geq C_2v^{1+1/d} \qquad \forall\, j \in \mathcal{S}. \tag{C.18}$$

By Lemma C.4 and a union bound over all $p-s$ variables in $\mathcal{S}^c$, we have that with probability at least $1-4n(p-s)\exp(-n\xi^2/(12(B^2+\sigma^2)))$,

$$\widehat{\Delta}(X_j,Y) \leq \xi^2 \qquad \forall\, j \in \mathcal{S}^c. \tag{C.19}$$

Recall that if we know the size s of the support $\mathcal{S}$, then $\widehat{\mathcal{S}}$ consists of the top s impurity reductions. Note that choosing $\xi^2 = \frac{C_2 v^{1+1/d}}{2}$ in (C.19) will give us a high probability upper bound on $\widehat{\Delta}(X_j, Y)$ for irrelevant variables which is dominated by the lower bound on $\widehat{\Delta}(X_j, Y)$ for relevant variables in (C.18). Thus, by a union bound, it follows that with probability at least $1 - 3s\exp(-C_1 nv) - 4n(p-s)\exp\Big(-\frac{nC_2 v^{1+1/d}}{24(B^2+\sigma^2)}\Big)$, we have $\widehat{\mathcal{S}} = \mathcal{S}$. $\square$

D PROOF OF STATEMENT IN EXAMPLE 1

In this section, we prove the statement in Example 1.

Proof. Let σ_1^2 and σ_2^2 be the respective variances of X_1 and X_2. By the Gaussian assumption, write $X_2 = \rho(\sigma_2/\sigma_1)X_1 + \sigma_2\sqrt{1-\rho^2}Z$, where $\rho \neq \pm 1$ is the correlation between X_1 and X_2 and $Z \sim \mathcal{N}(0,1)$ is independent of X_1. Let $p(z) = 1 - q(z) = \mathbb{P}(X_2 \leq z)$ and $p(z|z') = 1 - q(z|z') = \mathbb{P}(X_2 \leq z | Z = z')$. Then,

$$\Delta(z, X_2, Y) = \frac{\mathrm{Cov}^2(\mathbf{1}(X_2 \leq z), Y)}{p(z)q(z)} = \frac{(\mathbb{E}_Z[\mathrm{Cov}(\mathbf{1}(X_2 \leq z), Y|Z)])^2}{p(z)q(z)}.$$

By the Cauchy-Schwarz inequality,

$$\frac{(\mathbb{E}_Z[\mathrm{Cov}(\mathbf{1}(X_2 \leq z), Y|Z)])^2}{p(z)q(z)} \leq \frac{\mathbb{E}_Z[p(z|Z)q(z|Z)]}{p(z)q(z)} \times \mathbb{E}_Z\Big[\frac{\mathrm{Cov}^2(\mathbf{1}(X_2 \leq z), Y|Z)}{p(z|Z)q(z|Z)}\Big].$$

Next, observe that $\mathbb{E}_Z[p(z|Z)] = p(z)$ and by Jensen's inequality, $\mathbb{E}_Z[p^2(z|Z)] > p^2(z)$ (the inequality is strict since $\rho \neq \pm 1$). Thus,

$$\Delta(z, X_2, Y) < \mathbb{E}_Z\Big[\frac{\mathrm{Cov}^2(\mathbf{1}(X_2 \leq z), Y|Z)}{p(z|Z)q(z|Z)}\Big].$$

Finally,

$$\Delta(X_2, Y) < \max_z \mathbb{E}_Z\Big[\frac{\mathrm{Cov}^2(\mathbf{1}(X_2 \leq z), Y|Z)}{p(z|Z)q(z|Z)}\Big] \leq \mathbb{E}_Z\Big[\max_z \frac{\mathrm{Cov}^2(\mathbf{1}(X_2 \leq z), Y|Z)}{p(z|Z)q(z|Z)}\Big] = \mathbb{E}_Z[\Delta(X_1, Y)] = \Delta(X_1, Y),$$

where the penultimate equality comes from $\max_z \frac{\mathrm{Cov}^2(\mathbf{1}(X_2 \leq z), Y|Z)}{p(z|Z)q(z|Z)} = \max_u \frac{\mathrm{Cov}^2(\mathbf{1}(X_1 \leq u), Y)}{\mathbb{P}(X_1 \leq u)\mathbb{P}(X_1 > u)} = \Delta(X_1, Y)$ with $u = (z - \sigma_2\sqrt{1-\rho^2}Z)/(\rho\sigma_2/\sigma_1)$. $\square$

E PROOF OF MODEL SELECTION CONSISTENCY FOR LINEAR MODELS

Recall the setting mentioned in the heading "***Minimum sample size for consistency***" in Section 5, which considers the same linear model with Gaussian variates from Theorem 1. To reiterate, we assume that $\mathbf{\Sigma} = \mathbf{I}_{p\times p}$ is the $p \times p$ identity matrix, $\sum_{k=1}^p \beta_k^2 = \mathcal{O}(1)$, and $\min_{j\in\mathcal{S}} |\beta_j|^2 \asymp 1/s$, all of which are special cases of the more general setting considered in (Wainwright, 2009, Corollary 1). Under these assumptions, we then have $\rho^2(X_j, Y) = \beta_j^2/(\sigma^2 + \sum_{k=1}^p \beta_k^2) \gtrsim 1/s$ for any $j \in \mathcal{S}$ and $\rho(X_j, Y) = 0$ for $j \in \mathcal{S}^c$. Our goal is to show that $n \asymp s\log(n)\log(n(p-s))$ samples suffice for high probability model selection consistency.

Choosing $\delta = 1/2$ in (9) applied to $j \in \mathcal{S}$ and using (7), there exists a universal positive constant C_0 such that with probability at least $1 - \frac{2C_0}{\sqrt{n\rho^2(X_j,Y)}}\exp(-n\rho^2(X_j, Y)/8)$, we have

$$\widehat{\Delta}(X_j, Y) \geq \widehat{\mathrm{Var}}(Y) \times \frac{\rho^2(X_j, Y)}{4(\log(2n)+1)} = \frac{\widehat{\mathrm{Var}}(Y)}{4(\log(2n)+1)} \frac{\beta_j^2}{\sigma^2 + \sum_{k=1}^p \beta_k^2} \gtrsim \frac{\widehat{\mathrm{Var}}(Y)}{s\log(n)} \tag{E.1}$$

Therefore by a union bound over all s relevant variables, we have that with probability at least $1 - s\max_{j\in\mathcal{S}}\{\frac{2C_0}{\sqrt{n\rho^2(X_j,Y)}}\exp(-n\rho^2(X_j, Y)/8)\}$,

$$\widehat{\Delta}(X_j, Y) \gtrsim \frac{\widehat{\mathrm{Var}}(Y)}{s\log(n)} \qquad \forall j \in \mathcal{S}. \tag{E.2}$$

Furthermore, by applying (10) for $j \in \mathcal{S}^c$ (and noting that $\rho(X_j, Y) = 0$) and using (7), with probability at least $1 - 4n\exp(-\delta^2 n/64) - 3\exp(-(n-1)/16)$, we have

$$\widehat{\Delta}(X_j, Y) \le \widehat{\mathrm{Var}}(Y)\delta^2.$$

Therefore by a union bound over all $p - s$ irrelevant variables we have that with probability at least $1 - 4n(p-s)\exp(-\delta^2 n/64) - 3(p-s)\exp(-(n-1)/16)$,

$$\widehat{\Delta}(X_j, Y) \le \widehat{\mathrm{Var}}(Y)\delta^2 \qquad \forall j \in \mathcal{S}^c.$$

Now, choosing $\delta^2 = \frac{C_3}{s\log(n)}$ for some appropriate constant $C_3 > 0$ which only depends on σ^2 to match (E.1), we see by a union bound that

$$\mathbb{P}(\widehat{\mathcal{S}} = \mathcal{S}) \ge$$
$$1 - \max_{j\in\mathcal{S}}\Big\{\frac{2C_0 s}{\sqrt{n\rho^2(X_j,Y)}}\exp\Big(-\frac{n\rho^2(X_j,Y)}{8}\Big)\Big\} - 4n(p-s)\exp\Big(-\frac{C_3 n}{64 s\log(n)}\Big) - 3(p-s)\exp\Big(-\frac{(n-1)}{16}\Big).$$

Since $\rho^2(X_j, Y) \gtrsim 1/s$ for all $j \in \mathcal{S}$, the above implies that if $n(p-s)\exp\Big(-\frac{C_3 n}{64 s\log(n)}\Big) \to 0$, then $\mathbb{P}(\widehat{\mathcal{S}} = \mathcal{S}) \to 1$. Hence, a sufficient sample size for consistent support recovery is

$$n \asymp s\log(n)\log(n(p-s)),$$

as desired.

References

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B*, 15(2):193–232.

Kazemitabar, J., Amini, A., Bloniarz, A., and Talwalkar, A. S. (2017). Variable importance using decision trees. In *Advances in Neural Information Processing Systems*, pages 426–435.

Klusowski, J. M. (2020). Sparse learning with CART. In *Advances in Neural Information Processing Systems*.

Korneichuk, N. (1991). *Exact Constants in Approximation Theory*. Cambridge University Press. Cambridge, England.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338.

Li, X., Wang, Y., Basu, S., Kumbier, K., and Yu, B. (2019). A debiased MDI feature importance measure for random forests. In *Advances in Neural Information Processing Systems 32*, pages 8049–8059. Curran Associates, Inc.

Maurer, A. and Pontil, M. (2009). Empirical Bernstein bounds and sample-variance penalization. In *COLT*.

Wainwright, M. J. (2009). Information-theretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741.

Wendel, J. (1948). Note on the gamma function. *The American Mathematical Monthly*, 55(9):563–564.