The Annals of Statistics
2021, Vol. 49, No. 6, 3070–3102
https://doi.org/10.1214/21-AOS2073

A SIMPLE MEASURE OF CONDITIONAL DEPENDENCE

BY MONA AZADKIA* AND SOURAV CHATTERJEE†

Department of Statistics, Stanford University, *mazadkia@stanford.edu; †souravc@stanford.edu*

We propose a coefficient of conditional dependence between two random variables Y and Z given a set of other variables $X_1, \ldots, X_p$, based on an i.i.d. sample. The coefficient has a long list of desirable properties, the most important of which is that under absolutely no distributional assumptions, it converges to a limit in $[0, 1]$, where the limit is 0 if and only if Y and Z are conditionally independent given $X_1, \ldots, X_p$, and is 1 if and only if Y is equal to a measurable function of Z given $X_1, \ldots, X_p$. Moreover, it has a natural interpretation as a nonlinear generalization of the familiar partial R^2 statistic for measuring conditional dependence by regression. Using this statistic, we devise a new variable selection algorithm, called Feature Ordering by Conditional Independence (FOCI), which is model-free, has no tuning parameters, and is provably consistent under sparsity assumptions. A number of applications to synthetic and real data sets are worked out.

1. Introduction. The problem of measuring the amount of dependence between two random variables is an old problem in statistics. Numerous methods have been proposed over the years. For recent surveys, see [13, 34]. The literature on measures of *conditional dependence*, on the other hand, is not so large, especially in the nonparametric setting.

The nonparametric conditional independence testing problem can be relatively easily solved for discrete data using the classical Cochran–Mantel–Haenszel test [15, 38]. This test can be adapted for continuous random variables by binning the data [32] or using kernels [18, 28, 48, 52, 63].

Besides these, there are methods based on estimating conditional cumulative distribution functions [37, 42], conditional characteristic functions [53], conditional probability density functions [54], empirical likelihood [55], mutual information and entropy [33, 44, 47], copulas [5, 51, 58], distance correlation [23, 56, 60] and other approaches [49]. A number of interesting ideas based on resampling and permutation tests have been proposed in recent years [6, 11, 48].

The first contribution of this paper is a new coefficient of conditional dependence between two random variables Y and Z given a set of other variables $X_1, \ldots, X_p$, based on i.i.d. data. The coefficient is inspired by a similar measure of univariate dependence recently proposed in [13]. The main features of our coefficient are the following:

1. it has a simple expression,
2. it is fully nonparametric,
3. it has no tuning parameters,
4. there is no need for estimating conditional densities, conditional characteristic functions or mutual information,
5. it can be estimated from data very quickly, in time $O(n \log n)$ where n is the sample size,
6. asymptotically, it converges to a limit in $[0, 1]$, where the limit is 0 if and only if Y and Z are conditionally independent given $X_1, \ldots, X_p$, and is 1 if and only if Y is equal to a measurable function of Z given $X_1, \ldots, X_p$,

Received July 2020; revised January 2021.
MSC2020 subject classifications. 62G05, 62H20.
Key words and phrases. Conditional dependence, nonparametric measures of association, variable selection.

7. the limit has a natural interpretation as a nonlinear generalization of the familiar partial R^2 statistic for measuring the conditional dependence of Y and Z given $X_1, \ldots, X_p$, and
8. all of the above hold under absolutely no assumptions on the laws of the random variables.

The second contribution of this paper is a new variable selection algorithm based on the above measure of conditional dependence, called *Feature Ordering by Conditional Independence (FOCI)*, which is model-free, has no tuning parameters, and is provably consistent under sparsity assumptions. More importantly, it appears to perform very well in simulated and real data sets. The development of FOCI and the proof of its consistency are the major new contributions of this paper over [13]. It is not possible to devise such an algorithm using the univariate coefficient from [13].

The paper is organized as follows. The definition and properties of our coefficient are presented in Section 2. Section 3 discusses how to interpret the coefficient as a nonlinear generalization of partial R^2. A theorem about its rate of convergence is presented in Section 4. Our variable selection method is introduced in Section 5 and a theorem about its consistency is stated in Section 6. The special case of linear regression with Gaussian predictors is illustrated in Section 7. Applications to simulated and real data sets are presented in Section 8. The remaining sections are devoted to proofs.

2. The coefficient. Let Y be a random variable and $\mathbf{X} = (X_1, \ldots, X_p)$ and $\mathbf{Z} = (Z_1, \ldots, Z_q)$ be random vectors, all defined on the same probability space. Here, $q \geq 1$ and $p \geq 0$. The value $p = 0$ means that $\mathbf{X}$ has no components at all. Let μ be the law of Y. We propose the following quantity as a measure of the degree of conditional dependence of Y and $\mathbf{Z}$ given $\mathbf{X}$:

$$T = T(Y, \mathbf{Z}|\mathbf{X}) := \frac{\int \mathbb{E}(\operatorname{Var}(\mathbb{P}(Y \geq t|\mathbf{Z}, \mathbf{X})|\mathbf{X}))\, d\mu(t)}{\int \mathbb{E}(\operatorname{Var}(1_{\{Y \geq t\}}|\mathbf{X}))\, d\mu(t)}. \tag{2.1}$$

In the denominator, $1_{\{Y \geq t\}}$ is the indicator of the event $\{Y \geq t\}$. If the denominator equals zero, T is undefined. (We will see below that this happens if and only if Y is almost surely equal to a measurable function of $\mathbf{X}$, which is a degenerate case that we will ignore.) If $p = 0$, then $\mathbf{X}$ has no components, and the conditional expectations and variances given $\mathbf{X}$ should be interpreted as unconditional expectations and variances. In this case, we will write $T(Y, \mathbf{Z})$ instead of $T(Y, \mathbf{Z}|\mathbf{X})$.

Although the statistic T has a somewhat complicated looking expression, it has a natural interpretation as a nonlinear generalization of the partial R^2 statistic for measuring the proportion of variation in Y that is explained by $(\mathbf{Z}, \mathbf{X})$ but cannot be explained solely by $\mathbf{X}$. This is discussed in the next section. Specifically, see equation (3.1).

Note that T is a nonrandom quantity that depends only the joint law of $(Y, \mathbf{X}, \mathbf{Z})$. Before stating our theorem about T, let us first see why T is a reasonable measure of conditional dependence. Since taking conditional expectation decreases variance, we have that for any t,

$$\operatorname{Var}(1_{\{Y \geq t\}}|\mathbf{X}) \geq \operatorname{Var}\big(\mathbb{P}(Y \geq t|\mathbf{Z}, \mathbf{X})|\mathbf{X}\big).$$

This shows that the numerator in (2.1) is less than or equal to the denominator, and so T is always between 0 and 1. Now, if Y and $\mathbf{Z}$ are conditionally independent given $\mathbf{X}$, then $\mathbb{P}(Y \geq t|\mathbf{Z}, \mathbf{X})$ is a function of $\mathbf{X}$ only, and hence $\operatorname{Var}(\mathbb{P}(Y \geq t|\mathbf{Z}, \mathbf{X})|\mathbf{X}) = 0$. Therefore, in this situation, $T = 0$. We will show later that the converse is also true. On the other hand, if Y is almost surely equal to a measurable function of $\mathbf{Z}$ given $\mathbf{X}$, then $\mathbb{P}(Y \geq t|\mathbf{Z}, \mathbf{X}) = 1_{\{Y \geq t\}}$ for any t. Therefore, in this case, $T = 1$. Again, we will prove later that the converse is true. The following theorem summarizes these properties of T.

THEOREM 2.1. *Suppose that Y is not almost surely equal to a measurable function of $\mathbf{X}$ (when $p = 0$, this means that Y is not almost surely a constant). Then T is well defined and $0 \le T \le 1$. Moreover, $T = 0$ if and only if Y and $\mathbf{Z}$ are conditionally independent given $\mathbf{X}$, and $T = 1$ if and only if Y is almost surely equal to a measurable function of $\mathbf{Z}$ given $\mathbf{X}$. When $p = 0$, conditional independence given $\mathbf{X}$ simply means unconditional independence.*

The statistic T is a generalization of a similar univariate measure defined in [13, 17]. Having defined T, the main question is whether T can be efficiently estimated from data. We will now present a consistent estimator of T, which is our conditional dependence coefficient. This generalizes a similar univariate estimator defined in [13]. Our data consists of n i.i.d. copies $(Y_1, \mathbf{X}_1, \mathbf{Z}_1), \ldots, (Y_n, \mathbf{X}_n, \mathbf{Z}_n)$ of the triple $(Y, \mathbf{X}, \mathbf{Z})$, where $n \ge 2$. For each i, let $N(i)$ be the index j such that $\mathbf{X}_j$ is the nearest neighbor of $\mathbf{X}_i$ with respect to the Euclidean metric on $\mathbb{R}^p$, where ties are broken uniformly at random. Let $M(i)$ be the index j such that $(\mathbf{X}_j, \mathbf{Z}_j)$ is the nearest neighbor of $(\mathbf{X}_i, \mathbf{Z}_i)$ in $\mathbb{R}^{p+q}$, again with ties broken uniformly at random. Let R_i be the rank of Y_i, that is, the number of j such that $Y_j \le Y_i$. If $p \ge 1$, our estimate of T is

$$T_n = T_n(Y, \mathbf{Z}|\mathbf{X}) := \frac{\sum_{i=1}^n (\min\{R_i, R_{M(i)}\} - \min\{R_i, R_{N(i)}\})}{\sum_{i=1}^n (R_i - \min\{R_i, R_{N(i)}\})}.$$

If $p = 0$, let L_i be the number of j such that $Y_j \ge Y_i$, let $M(i)$ denote the j such that $\mathbf{Z}_j$ is the nearest neighbor of $\mathbf{Z}_i$ (ties broken uniformly at random), and let

$$T_n = T_n(Y, \mathbf{Z}) := \frac{\sum_{i=1}^n (n \min\{R_i, R_{M(i)}\} - L_i^2)}{\sum_{i=1}^n L_i(n - L_i)}.$$

In both cases, T_n is undefined if the denominator is zero. The following theorem proves that T_n is indeed a consistent estimator of T.

THEOREM 2.2. *Suppose that Y is not almost surely equal to a measurable function of $\mathbf{X}$. Then as $n \to \infty$, $T_n \to T$ almost surely.*

REMARKS. (1) If p and q are fixed, the statistic T_n can be computed in $O(n \log n)$ time because nearest neighbors can be determined in $O(n \log n)$ time [27] and ranks can also be calculated in $O(n \log n)$ time [35].

(2) No assumptions on the joint law of $(Y, \mathbf{X}, \mathbf{Z})$ are needed other than the nondegeneracy condition that Y is not almost surely equal to a measurable function of $\mathbf{X}$. This condition is inevitable, because if this does not hold, then given $\mathbf{X}$, Y is a constant; in this circumstance, Y is both a function of $\mathbf{Z}$ given $\mathbf{X}$ *and* independent of $\mathbf{Z}$ given $\mathbf{X}$, and so there can be no reasonable measure of the degree of conditional dependence of Y and $\mathbf{Z}$ given $\mathbf{X}$.

(3) Although the limit of T_n is guaranteed to be in $[0, 1]$, the actual value of T_n for finite n may lie outside this interval.

(4) It is not easy to explain why T_n is a consistent estimator of T without going into the details of the proof, so we will not make that attempt here.

(5) We have not given a name to T_n, but if an acronym is desired for easy reference, one may call it CODEC, which is an acronym for Conditional Dependence Coefficient. In fact, this is the acronym that we use in the R code for computing T_n.

(6) We have prepared an R package, called FOCI, that has a function for computing T_n and a function for executing the variable selection algorithm FOCI presented in Section 5 below. The package is available for download on CRAN [3].

(7) Besides variable selection, another natural area of applications of our coefficient is graphical models. This is currently under investigation.

(8) The consistency of T_n raises the possibility of constructing a consistent test for conditional independence based on T_n. However, it is known that this is an impossible task, even for a single alternative hypothesis, if we demand that the level of the test be asymptotically uniformly bounded by some given α over the whole null hypothesis space [50]. This is why the problem of nonparametric conditional independence testing for continuous random variables is essentially unsolvable unless one is willing to impose unverifiable assumptions. This contrasts starkly with the problem of nonparametric testing of *unconditional* independence, for which there are many useful and popular methods (see [13] for a survey).

(9) In view of Theorem 2.2, it is natural to be curious about the rate of convergence of T_n to T. This is investigated in Section 4.

3. Interpreting the coefficient. To interpret $T(Y, \mathbf{Z}|\mathbf{X})$, it is instructive to first consider the case of binary Y. Suppose that Y is $\{0, 1\}$-valued. Then μ is supported on $\{0, 1\}$. Since $Y \geq 0$ always, we have $\mathbb{P}(Y \geq 0|\mathbf{Z}, \mathbf{X}) = 1_{\{Y \geq 0\}} = 1$ always. Thus,

$$\operatorname{Var}\big(\mathbb{P}(Y \geq 0|\mathbf{Z}, \mathbf{X})|\mathbf{X}\big) = \operatorname{Var}(1_{\{Y \geq 0\}}|\mathbf{X}) = 0.$$

Moreover, $Y = 1_{\{Y \geq 1\}}$. Combining all of this, we get that for binary Y,

$$T(Y, \mathbf{Z}|\mathbf{X}) = \frac{\mathbb{E}(\operatorname{Var}(\mathbb{E}(Y|\mathbf{Z}, \mathbf{X})|\mathbf{X}))}{\mathbb{E}(\operatorname{Var}(Y|\mathbf{X}))}.$$

But, by the law of total variance,

$$\operatorname{Var}(Y|\mathbf{X}) = \mathbb{E}\big(\operatorname{Var}(Y|\mathbf{Z}, \mathbf{X})|\mathbf{X}\big) + \operatorname{Var}\big(\mathbb{E}(Y|\mathbf{Z}, \mathbf{X})|\mathbf{X}\big).$$

Thus, for binary Y,

$$T(Y, \mathbf{Z}|\mathbf{X}) = 1 - \frac{\mathbb{E}(\operatorname{Var}(Y|\mathbf{Z}, \mathbf{X}))}{\mathbb{E}(\operatorname{Var}(Y|\mathbf{X}))}.$$

But this is just the *partial* R^2 that measures the proportion of variation in Y that is explained by $(\mathbf{Z}, \mathbf{X})$ but cannot be explained solely by $\mathbf{X}$. Therefore, when Y is binary, we have the identity

$$T(Y, \mathbf{Z}|\mathbf{X}) = R^2_{Y, \mathbf{Z}|\mathbf{X}}.$$

For a general Y, let $Y_t := 1_{\{Y \geq t\}}$ for each t. Then by the same calculation as above, we get

$$T(Y, \mathbf{Z}|\mathbf{X}) = 1 - \frac{\int \mathbb{E}(\operatorname{Var}(Y_t|\mathbf{Z}, \mathbf{X}))\, d\mu(t)}{\int \mathbb{E}(\operatorname{Var}(Y_t|\mathbf{X}))\, d\mu(t)}.$$

Let us now define a probability measure ν on $\mathbb{R}$, which has density proportional to $\mathbb{E}(\operatorname{Var}(Y_t|\mathbf{X}))$ with respect to μ. Then the above formula can be rewritten as

$$T(Y, \mathbf{Z}|\mathbf{X}) = \int R^2_{Y_t, \mathbf{Z}|\mathbf{X}}\, d\nu(t). \tag{3.1}$$

Thus, $T(Y, \mathbf{Z}|\mathbf{X})$ is a weighted average of $R^2_{Y_t, \mathbf{Z}|\mathbf{X}}$ over all $t \in \mathbb{R}$. But the random variable Y is a linear combination of the binary variables $\{Y_t\}_{t \in \mathbb{R}}$. Therefore, $T(Y, \mathbf{Z}|\mathbf{X})$ is *also a measure of the proportion of variation in Y that is explained by $(\mathbf{Z}, \mathbf{X})$ but cannot be explained solely by $\mathbf{X}$*. It generalizes the usual partial R^2 statistic $R^2_{Y, \mathbf{Z}|\mathbf{X}}$ in a nonlinear way — by breaking up Y as a linear combination of binary variables, computing the partial R^2 for each binary variable, and combining these partial R^2 statistics by taking a weighted average.

4. Rate of convergence. Suppose that $p \geq 1$, so that $\mathbf{X}$ has at least one component. (Recall that q is always at least 1.) To obtain a rate of convergence of T_n to T, we need to make some assumptions about the distribution of $(Y, \mathbf{X}, \mathbf{Z})$, because otherwise, we believe that the convergence may be arbitrarily slow. The main issue is that we need some kind of control on the sensitivity of the conditional distribution of Y given $\mathbf{X}$ and $\mathbf{Z}$ on the values of $\mathbf{X}$ and $\mathbf{Z}$. This is handled by the first assumption below. The second assumption is a matter of technical convenience.

(A1) There are nonnegative real numbers β and C such that for any $t \in \mathbb{R}$, $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^p$ and $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^q$,

$$\begin{aligned}&\bigl|P(Y \geq t|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) - P\bigl(Y \geq t|\mathbf{X} = \mathbf{x}', \mathbf{Z} = \mathbf{z}'\bigr)\bigr| \\ &\quad \leq C\bigl(1 + \|\mathbf{x}\|^\beta + \|\mathbf{x}'\|^\beta + \|\mathbf{z}\|^\beta + \|\mathbf{z}'\|^\beta\bigr)\bigl(\|\mathbf{x} - \mathbf{x}'\| + \|\mathbf{z} - \mathbf{z}'\|\bigr),\end{aligned}$$

and

$$\begin{aligned}&\bigl|P(Y \geq t|\mathbf{X} = \mathbf{x}) - P\bigl(Y \geq t|\mathbf{X} = \mathbf{x}'\bigr)\bigr| \\ &\quad \leq C\bigl(1 + \|\mathbf{x}\|^\beta + \|\mathbf{x}'\|^\beta\bigr)\|\mathbf{x} - \mathbf{x}'\|.\end{aligned}$$

(A2) There are positive numbers C_1 and C_2 such that for any $t > 0$, $\mathbb{P}(\|\mathbf{X}\| \geq t)$ and $\mathbb{P}(\|\mathbf{Z}\| \geq t)$ are bounded by $C_1 e^{-C_2 t}$.

Note that assumption (A1) means that the conditional distribution of Y given $(\mathbf{X}, \mathbf{Z}) = (\mathbf{x}, \mathbf{z})$ is a locally Lipschitz function of $(\mathbf{x}, \mathbf{z})$, where the Lipschitz constant is allowed to grow at most polynomially in $\|\mathbf{x}\|$ and $\|\mathbf{z}\|$. Local Lipschitzness is a fairly relaxed assumption. It only excludes esoteric cases where the conditional distribution of Y given $(\mathbf{X}, \mathbf{Z}) = (\mathbf{x}, \mathbf{z})$ is a very rough function of $(\mathbf{x}, \mathbf{z})$ (e.g., like a Brownian path), which do not arise in any model used in practice.

Under the above assumptions, the following theorem shows that T_n converges to T essentially at the rate $n^{-1/(p+q)}$, up to an extra logarithmic term.

THEOREM 4.1. *Suppose that $p \geq 1$ and $q \geq 1$, and that the assumptions* (A1) *and* (A2) *hold with some β and C. Then, as $n \to \infty$,*

$$T_n - T = O_P\biggl(\frac{(\log n)^{p+q+\beta+1}}{n^{1/(p+q)}}\biggr).$$

We believe that the rate $n^{-1/(p+q)}$ in Theorem 4.1 is the true rate of convergence of T_n to T when the variables are continuous. It is not clear if there is some other statistic with the same properties as T_n but with a better rate of convergence.

Note that the case $p = 0$ is not covered by Theorem 4.1. This is the case where T_n is a measure of unconditional, rather than conditional, dependence. When $p = 0$ and $q = 1$, we conjecture that $T_n - T = O_P(n^{-1/2})$. Moreover, under independence, we conjecture that $\sqrt{n}T_n$ obeys a central limit theorem when $p = 0$ and $q = 1$. At this moment, we do not know how to prove these conjectures.

Conditions (A1) and (A2) are trivially satisfied if the support of $(Y, \mathbf{X}, \mathbf{Z})$ is a finite set, by choosing $\beta = 0$ and a suitably large C. Another situation where it is easy to see that (A1) and (A2) hold is when $(Y, \mathbf{X}, \mathbf{Z})$ is normal, because then the conditional distribution of Y given $(\mathbf{X}, \mathbf{Z})$ is again normal with a mean that is a linear function of $\mathbf{X}$ and $\mathbf{Z}$, and a variance that does not depend on $\mathbf{X}$ and $\mathbf{Z}$.

More generally, the following result shows that (A1) is satisfied for a large class of densities with certain regularity and decay properties (and (A2) holds widely anyway).

PROPOSITION 4.2. *Let $f(y|\mathbf{x})$ be the conditional probability density function of Y given $\mathbf{X} = \mathbf{x}$, assuming it exists. Suppose that f is nonzero everywhere and differentiable with respect to $\mathbf{x}$, and for each i, the function*

$$\left| \frac{\partial}{\partial x_i} \log f(y|\mathbf{x}) \right|$$

is bounded above by a polynomial in $|y|$ and $\|\mathbf{x}\|$. Next, suppose that for any compact set $K \subseteq \mathbb{R}^p$, the function $g(y) := \max_{\mathbf{x} \in K} f(y|\mathbf{x})$ is bounded and decays faster than any negative power of $|y|$ as $|y| \to \infty$. Lastly, assume that for any $k \geq 1$, $\mathbb{E}(Y^{2k}|\mathbf{X} = \mathbf{x})$ is bounded above by a polynomial in $\|\mathbf{x}\|$. Then the second inequality in assumption (A1) *holds for some C and β. A similar set of conditions on the conditional density of Y given $\mathbf{X} = \mathbf{x}$ and $\mathbf{Z} = \mathbf{z}$ ensures that the first inequality in* (A1) *holds.*

5. Feature Ordering by Conditional Independence (FOCI). In this section, we propose a new variable selection algorithm for multivariate regression using a forward stepwise algorithm based on our measure of conditional dependence. The commonly used variable selection methods in the statistics literature use linear or additive models. This includes classical methods [7, 14, 22, 26, 29, 30, 41, 57] as well as modern ones [12, 24, 45, 61, 64, 65]. These methods are powerful and widely used in practice. However, they sometimes run into problems when significant interaction effects or nonlinearities are present. We will later show an example where methods based on linear and additive models fail to select any of the relevant predictors, *even in the complete absence of noise*.

Such problems can sometimes be overcome by model-free methods [2, 4, 8–11, 25, 30, 31, 59]. These, too, are powerful and widely used techniques, and they perform better than model-based methods if interactions are present. On the flip side, their theoretical foundations are usually weaker than those of model-based methods.

The method that we are going to propose below, called *Feature Ordering by Conditional Independence (FOCI)*, attempts to combine the best of both worlds by being fully model-free, as well as having a proof of consistency under a set of assumptions.

The method is as follows. Let Y be the response variable and let $\mathbf{X} = (X_j)_{1 \leq j \leq p}$ be the set of predictors. The data consists of n i.i.d. copies of $(Y, \mathbf{X})$. First, choose j_1 to be the index j that maximizes $T_n(Y, X_j)$. Having obtained $j_1, \ldots, j_k$, choose j_{k+1} to be the index $j \notin \{j_1, \ldots, j_k\}$ that maximizes $T_n(Y, X_j | X_{j_1}, \ldots, X_{j_k})$. Continue like this until arriving at the first k such that $T_n(Y, X_{j_{k+1}} | X_{j_1}, \ldots, X_{j_k}) \leq 0$, and then declare the chosen subset to be $\hat{S} := \{j_1, \ldots, j_k\}$. If there is no such k, define $\hat{S}$ to be the whole set of variables. It may also happen that $T_n(Y, X_{j_1}) \leq 0$. In that case, declare $\hat{S}$ to be empty.

Although it is not required theoretically, we recommend that the predictor variables be standardized before running the algorithm. We will see later that FOCI performs well in examples, even if the true dependence of Y on $\mathbf{X}$ is nonlinear in a complicated way. In the next section, we prove the consistency of FOCI under a set of assumptions on the law of $(Y, \mathbf{X})$.

If computational time is not an issue, one can try to add $m \geq 2$ variables at each step instead of just one. Although we do not explore this idea in this paper, it is possible that this gives improved results in certain situations. Similarly, one can try a forward–backward version of FOCI, analogous to the forward–backward version of ordinary stepwise selection.

One can also consider implementing a forward stepwise algorithm like FOCI with other measures of conditional dependence. To the best of our knowledge, that has not yet been done. The closest cousin in the literature is an algorithm based on mutual information [4], but unlike FOCI, it does not have a well-defined stopping rule.

One deficiency of FOCI is that it only selects a subset of predictors, without actually fitting a predictive model. Following a suggestion from Rob Tibshirani, we recommend doing the

following: First, select a subset using FOCI, and then use random forests [9] to fit a predictive model with the selected variables. This has two advantages over simply fitting random forests with the full set of predictors: (1) It picks out a small set of "important" variables, which may be useful for various reasons, and (2) it is computationally much less expensive. We saw that in real data sets, the prediction error of FOCI followed by random forests is only slightly worse than fitting random forests with the full set of predictors. On the other hand, the number of variables selected by FOCI is usually very small compared to the total number of variables. Some examples are given in Section 8.

The stopping rule for FOCI may not be the best one. One can think of various other ways of using our measure of conditional (or any other) for variable selection. Our stopping rule seems to work well in practice, and we are able to prove consistency of variable selection for this rule. It is possible that there are other, better rules, which are also provably consistent. This merits further investigation.

6. Consistency of FOCI. Let $(Y, \mathbf{X})$ be as in the previous section. For any subset of indices $S \subseteq \{1, \ldots, p\}$, let $\mathbf{X}_S := (X_j)_{j \in S}$, and let $S^c := \{1, \ldots, p\} \setminus S$. In the machine learning literature, a subset S is sometimes called *sufficient* [59] if Y and $\mathbf{X}_{S^c}$ are conditionally independent given $\mathbf{X}_S$. This includes the possibility that S is the empty set, when it simply means that Y and $\mathbf{X}$ are independent. Sufficient subsets are known as *Markov blankets* in the literature on graphical models [43], Section 3.2.1, and are closely related to the concept of *sufficient dimension reduction* in classical statistics [1, 16, 36]. If we can find a small subset of predictors that is sufficient, then our job is done, because these predictors contain all the relevant predictive information about Y among the given set of predictors, and the statistician can then fit a predictive model based on this small subset of predictors.

Define $Q(\varnothing) := 0$, and for any nonempty set $S \subseteq \{1, \ldots, p\}$, let

$$Q(S) := \int \operatorname{Var}\big(\mathbb{P}(Y \geq t | \mathbf{X}_S)\big)\, d\mu(t), \tag{6.1}$$

where μ is the law of Y. We will prove later (Lemma 11.2) that $Q(S') \geq Q(S)$ whenever $S' \supseteq S$, with equality if and only Y and $\mathbf{X}_{S' \setminus S}$ are conditionally independent given $\mathbf{X}_S$. Thus if $S' \supseteq S$, the difference $Q(S') - Q(S)$ is a measure of how much extra predictive power is added by appending $\mathbf{X}_{S' \setminus S}$ to the set of predictors $\mathbf{X}_S$.

Let δ be the largest number such that for any *insufficient* subset S, there is some $j \notin S$ such that $Q(S \cup \{j\}) \geq Q(S) + \delta$. In other words, if S is insufficient, there exists some index $j \notin S$ such that appending X_j to $\mathbf{X}_S$ increases the predictive power by at least δ. The main result of this section, stated below, says that if δ is not too close to zero, then under some regularity assumptions on the law of $(Y, \mathbf{X})$, the subset selected by FOCI is sufficient with high probability. Note that a sparsity assumption is hidden in the condition that δ is not very small, because the definition of δ ensures that there is at least one sufficient subset of size $\leq 1/\delta$. An interpretation of δ in the familiar setting of linear regression with Gaussian predictors is discussed in the next section.

To prove our result, we need the following two technical assumptions on the joint distribution of $(Y, \mathbf{X})$. They are generalizations of the assumptions (A1) and (A2) from Section 4.

(A1′) There are nonnegative real numbers β and C such that for any set $S \subseteq \{1, \ldots, p\}$ of size $\leq 1/\delta + 2$, any $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^S$ and any $t \in \mathbb{R}$,

$$\begin{aligned} &\big|P(Y \geq t | \mathbf{X}_S = \mathbf{x}) - P\big(Y \geq t | \mathbf{X}_S = \mathbf{x}'\big)\big| \\ &\qquad \leq C\big(1 + \|\mathbf{x}\|^\beta + \|\mathbf{x}'\|^\beta\big)\|\mathbf{x} - \mathbf{x}'\|. \end{aligned}$$

(A2′) There are positive numbers C_1 and C_2 such that for any S of size $\leq 1/\delta + 2$ and any $t > 0$, $\mathbb{P}(\|\mathbf{X}_S\| \geq t) \leq C_1 e^{-C_2 t}$.

Proposition 4.2 shows that the above assumptions are satisfied in a wide variety of situations. The following theorem shows that under the above assumptions, the subset chosen by FOCI is sufficient with high probability.

THEOREM 6.1. *Suppose that $\delta > 0$, and that the assumptions* (A1′) *and* (A2′) *hold. Let $\hat{S}$ be the subset selected by FOCI with a sample of size n. There are positive real numbers L_1, L_2 and L_3 depending only on C, β, C_1, C_2 and δ such that $\mathbb{P}(\hat{S}$ is sufficient$) \geq 1 - L_1 p^{L_2} e^{-L_3 n}$.*

The main implication of Theorem 6.1 is that if δ is not too close to zero, and $n \gg \log p$, then with high probability, FOCI chooses a sufficient set of predictors. In particular, this theorem allows p to be quite large compared to n, as long as δ is not too small.

Although Theorem 6.1 works under the assumption that δ is fixed (because the constants L_1, L_2 and L_3 depend on δ in an unspecified manner), it is possible that a deeper analysis can allow us to take $\delta \to 0$ as $n \to \infty$. For that, the dependences of L_1, L_2 and L_3 on δ will have to be made explicit. It is possible that such an improvement can be made by carefully reworking the steps in the proof, to at least get a consistency result when $\delta \to 0$ slower than $(\log n)^{-1}$. To get anything better than that will probably require entirely new ideas.

Theorem 6.1 gives conditions under which the subset selected by FOCI is sufficient with high probability. This is intended to be useful for practitioners, by giving them confidence that the selected subset is indeed sufficient. In practice, as we will see in Section 8, the subsets selected by FOCI are quite small. However, Theorem 6.1 does not guarantee the smallness of the subset size. It would be desirable to have an improved version of Theorem 6.1, which not only guarantees that $\hat{S}$ is sufficient with high probability, but also that $|\hat{S}|$ is small with high probability. As of now, we do not know how to prove such a result.

7. Interpreting δ. In this section, we will try to understand the meaning of the quantity δ defined in the previous section, in the familiar context of linear regression with normally distributed predictor variables. Suppose that $\mathbf{X}$ is a normal random vector with zero mean and arbitrary covariance structure, and that

$$Y = \beta \cdot \mathbf{X} + \varepsilon,$$

where $\beta \in \mathbb{R}^p$ is a vector of coefficients and $\varepsilon \sim N(0, \sigma^2)$ is independent of $\mathbf{X}$, with nonzero σ. Then Y is also a normal random variable with mean zero. Let $\tau^2 := \mathrm{Var}(Y)$. Let δ be the quantity defined in the previous section, for this Y and $\mathbf{X}$.

For any nonempty $S \subsetneq \{1, \ldots, p\}$ and any $j \in \{1, \ldots, p\} \setminus S$, let $\rho(S, j)$ be the partial R^2 of Y and X_j given $\mathbf{X}_S$. Let $\rho(\varnothing, j)$ be the usual R^2 (that is, squared correlation) between Y and X_j.

Note that if S is a sufficient set of predictors, then $\rho(S, j) = 0$ for any $j \notin S$. Conversely, if $\rho(S, j) = 0$ for all $j \notin S$, then by normality, Y and $\mathbf{X}_{S^c}$ are conditionally independent given $\mathbf{X}_S$, and hence S is sufficient. Thus, S is sufficient if and only if $\rho(S, j) = 0$ for all $j \notin S$. So if S is insufficient, then there is at least one $j \notin S$ such that $\rho(S, j) > 0$.

Let δ' be the largest number such that for any insufficient set S, there is some $j \notin S$ such that $\rho(S, j) \geq \delta'$. The following result shows that δ' is comparable to δ, up to constant multiples depending only on σ and τ.

THEOREM 7.1. *Let all notation be as above. There are positive universal constants C_1 and C_2 such that*

$$\frac{C_1 \sigma^2}{\tau^2} \delta' \leq \delta \leq \frac{C_2 \tau^2}{\sigma^2} \delta'.$$

Thus, in the Gaussian setup, δ is equivalent to the analogous quantity computed using the usual partial R^2 instead of our measure of conditional dependence. The above result is proved in Section 17.

8. Examples. In this section, we present some applications of our methods to simulated examples and real data sets. In all examples, the covariates were standardized prior to the analysis.

EXAMPLE 8.1. Let X_1 and X_2 be independent Uniform[0, 1] random variables, and define

$$Y := X_1 + X_2 \pmod 1.$$

The relationship between Y and (X_1, X_2) has three main features:

1. Y is a function of (X_1, X_2),
2. unconditionally, Y is independent of X_2, and
3. conditional on X_1, Y is a function of X_2.

Let $n = 1000$. In about 95% of our simulations, $T_n(Y, (X_1, X_2))$ took values between 0.88 and 0.94, $T_n(Y, X_2|X_1)$ was between 0.88 and 0.94, and $T_n(Y, X_2)$ was between -0.07 and 0.07, in agreement with the above properties. Other measures of conditional dependence, such as conditional distance correlation [60], were unable to gauge the strength of the conditional dependency between Y and X_2 given X_1.

EXAMPLE 8.2. Let X_1 and X_2 be independent $N(0, 1)$ random variables, and define

$$Y := X_1^2 + X_2^2, \qquad Z := \arctan(X_1/X_2).$$

Then unconditionally, Y is independent of Z, and conditional on X_1, Y is a function of Z. Let $n = 1000$. In about 95% of our simulations, $T_n(Y, Z)$ took values between -0.06 and 0.05, and $T_n(Y, Z|X_1)$ was between 0.79 and 0.84, in agreement with the above properties. Again, other measures of conditional dependence were unable to capture the strength of the conditional dependence between Y and Z given X_1.

EXAMPLE 8.3. Let $X_1, \ldots, X_{1000}$ be independent $N(0, 1)$ random variables and let

$$Y = X_1X_2 + \sin(X_1X_3).$$

With a sample of size 2000 from the above model, FOCI was able to select the correct subset $\{X_1, X_2, X_3\}$ more than $90^i\%$ of the time. On the other hand, popular variable selection algorithms based on linear models, such as ordinary forward stepwise, Lasso [57], the Dantzig selector [12], and SCAD [24] were essentially never able to pick out the correct subset. (The tuning parameters for Lasso, Dantzig selector and SCAD were chosen using 10-fold cross-validation, and the AIC criterion was used for stopping in forward stepwise.) Even methods based on nonlinear additive models, such as SPAM [45], were generally unable to find the correct subset. The only other methods that successfully detected the importance of X_1, X_2 and X_3 were random forests [9] and mutual information [4], but the computational times for these methods were many times greater than that of FOCI.

EXAMPLE 8.4. Again, let $X_1, \ldots, X_{1000}$ be independent $N(0, 1)$ random variables and let

$$Y = X_1X_2 + X_1 - X_3 + \varepsilon,$$

where $\varepsilon \sim N(0, 1)$ is a noise term that is independent of X_i's. With a sample of size 2000 from this model, FOCI was able to select the correct subset $\{X_1, X_2, X_3\}$ in 99.5% of simulations. Methods based on linear models were generally able to pick out X_1 and X_3 but almost never detected the role of X_2. SPAM was able to pick out all three variables in about a quarter of the simulations. Again, the only other methods that successfully detected the importance of X_1, X_2 and X_3 were random forests and mutual information, but at a far greater computational cost than FOCI.

TABLE 1
Applications of FOCI to real data

	Spambase data		Polish companies data		Million song data	
Method	Subset size	MSPE	Subset size	MSPE	Subset size	MSPE
FOCI	14	0.045	5	0.022	17	88.085
Forward stepwise	56	0.039	62	0.020	90	87.226
Lasso	55	0.041	48	0.021	86	87.260
Dantzig selector	53	0.041	7	0.023	90	87.226
SCAD	38	0.041	4	0.025	85	87.319

EXAMPLE 8.5. We tried out FOCI on the following three benchmark real data examples, all from the UCI Machine Learning Repository [19]:

1. *Spambase data*. Consists of 4601 observations, each corresponding to one email, and 57 features for each observation. The response variable is binary, indicating whether the email is a spam email or not.
2. *Polish companies bankruptcy data*. Consists of 19,967 observations with 64 features. Each sample corresponds to a company in Poland. The response variable is binary, indicating whether or not the company was bankrupted after a period of time.
3. *Million song data*. Consists of 515,345 observations with 90 features. Each sample corresponds to the audio features of a song published sometime ranging from 1922 to 2011. The response variable is the year that the song was published.

FOCI was compared with forward stepwise, Lasso, Dantzig selector and SCAD. For each method, after selecting the variables, a predictive model was fitted to a training set using random forests. As before, the tuning parameters for Lasso, Dantzig selector and SCAD were chosen using 10-fold cross-validation, and the AIC criterion was used for stopping in forward stepwise. Mean squared prediction errors (MSPE) were estimated using a test set. The sizes of the selected subsets and the MSPEs are reported in Table 1. In all three examples, FOCI attained similar prediction errors as the other methods, but with a significantly fewer number of variables.

EXAMPLE 8.6. In Section 5, we recommended fitting a predictive model using random forests with the set of variables selected by FOCI. To test the validity of this approach, we computed the prediction errors for random forests with the full set of predictors versus FOCI followed by random forests, in the three real data sets considered above. The results are displayed in Table 2. We see that FOCI followed by random forests attains almost the same MSPE as random forests with the full set of variables; but in each case, the number of variables selected by FOCI is small compared to the total number of variables.

9. Restatement of Theorems 2.1 and 2.2. Beginning with this section, the rest of the paper is devoted to proofs. Throughout the rest of the manuscript, whenever we say that a random variable Y is a function of another variable X, we will mean that $Y = f(X)$ almost surely for some measurable function f.

First, we focus on Theorems 2.1 and 2.2. To prove these theorems, it is convenient to break up the estimators into pieces. This gives certain "elaborate" versions of Theorems 2.1 and 2.2, which are interesting in their own right. First, suppose that $p \geq 1$. Define

$$Q_n(Y, \mathbf{Z}|\mathbf{X}) := \frac{1}{n^2} \sum_{i=1}^{n} \big(\min\{R_i, R_{M(i)}\} - \min\{R_i, R_{N(i)}\}\big) \tag{9.1}$$

and

$$S_n(Y, \mathbf{X}) := \frac{1}{n^2} \sum_{i=1}^{n} \bigl(R_i - \min\{R_i, R_{N(i)}\}\bigr). \tag{9.2}$$

Let μ denote the law of Y. We will see later that the following theorem implies both Theorem 2.1 and Theorem 2.2 in the case $p \geq 1$.

THEOREM 9.1. *Suppose that $p \geq 1$. As $n \to \infty$, the statistics $Q_n(Y, \mathbf{Z}|\mathbf{X})$ and $S_n(Y, \mathbf{X})$ converge almost surely to deterministic limits. Call these limit a and b, respectively. Then*:

(i) $0 \leq a \leq b$.
(ii) *Y is conditionally independent of $\mathbf{Z}$ given $\mathbf{X}$ if and only if $a = 0$.*
(iii) *Y is conditionally a function of $\mathbf{Z}$ given $\mathbf{X}$ if and only if $a = b$.*
(iv) *Y is not a function of $\mathbf{X}$ if and only if $b > 0$.*

Explicitly, the values of a and b are given by

$$a = \int \mathbb{E}\bigl(\mathrm{Var}\bigl(\mathbb{P}(Y \geq t|\mathbf{Z}, \mathbf{X})|\mathbf{X}\bigr)\bigr)\, d\mu(t)$$

and

$$\begin{aligned} b &= \int \mathbb{E}\bigl(\mathrm{Var}(1_{\{Y \geq t\}}|\mathbf{X})\bigr)\, d\mu(t) \\ &= \int \mathbb{E}\bigl(\mathbb{P}(Y \geq t|\mathbf{X})\bigl(1 - \mathbb{P}(Y \geq t|\mathbf{X})\bigr)\bigr)\, d\mu(t). \end{aligned}$$

Next, suppose that $p = 0$. Define

$$Q_n(Y, \mathbf{Z}) := \frac{1}{n^2} \sum_{i=1}^{n} \left(\min\{R_i, R_{M(i)}\} - \frac{L_i^2}{n}\right) \tag{9.3}$$

and

$$S_n(Y) := \frac{1}{n^3} \sum_{i=1}^{n} L_i(n - L_i). \tag{9.4}$$

We will prove later that the following theorem implies Theorems 2.1 and 2.2 when $p = 0$.

THEOREM 9.2. *As $n \to \infty$, $Q_n(Y, \mathbf{Z})$ and $S_n(Y)$ converge almost surely to deterministic limits c and d, satisfying the following properties*:

(i) $0 \leq c \leq d$.
(ii) *Y is independent of $\mathbf{Z}$ if and only if $c = 0$.*
(iii) *Y is a function of $\mathbf{Z}$ if and only if $c = d$.*
(iv) *$d > 0$ if and only if Y not a constant.*

TABLE 2
Comparison with random forests

Data set	FOCI subset size/Total set size	MSPE FOCI	MSPE random forest
Spambase	14/57	0.045	0.040
Polish companies	5/64	0.022	0.020
Million song	17/90	88.085	87.260

Explicitly,

$$c = \int \operatorname{Var}\big(\mathbb{P}(Y \geq t|\mathbf{Z})\big)\, d\mu(t),$$

and

$$\begin{aligned} d &= \int \operatorname{Var}(1_{\{Y \geq t\}})\, d\mu(t) \\ &= \int \mathbb{P}(Y \geq t)\big(1 - \mathbb{P}(Y \geq t)\big)\, d\mu(t). \end{aligned}$$

It is not difficult to see that whenever Y has a continuous distribution, $d = 1/6$. In this case, there is no need for estimating d using $S_n(Y, \mathbf{Z})$. On the other hand, the value of d may be dependent on the distribution of Y when the distribution is not continuous. In such cases, d needs to be estimated from the data using $S_n(Y, \mathbf{Z})$.

10. Proofs of Theorems 2.1 and 2.2 using Theorems 9.1 and 9.2. Suppose that $p \geq 1$. Recall the quantities a and b from the statement of Theorem 9.1, and notice that $T = a/b$. Suppose that Y is not a function of $\mathbf{X}$. Then by conclusion (iv) of Theorem 9.1, $b > 0$, and hence T is well defined. Moreover, conclusion (i) implies that $0 \leq T \leq 1$, conclusion (ii) implies that $T = 0$ if and only if Y and $\mathbf{Z}$ are conditionally independent given $\mathbf{X}$, and conclusion (iii) implies that Y is a function of $\mathbf{Z}$ given $\mathbf{X}$ if and only if $T = 1$. This proves Theorem 2.1 when $p \geq 1$. Next, note that $T_n = Q_n/S_n$, where $Q_n = Q_n(Y, \mathbf{Z}|\mathbf{X})$ and $S_n = S_n(Y, \mathbf{X})$, as defined in (9.1) and (9.2). By Theorem 9.1, $Q_n \to a$ and $S_n \to b$ in probability. Thus, $T_n \to a/b = T$ in probability. This proves Theorem 2.2 when $p \geq 1$.

Next, suppose that $p = 0$. The proof proceeds exactly as before, but using Theorem 9.2. Here, $T = c/d$, where c and d are the quantities from Theorem 9.2. Suppose that Y is not a function of $\mathbf{X}$, which in this case just means that Y is not a constant. Then by conclusion (iv) of Theorem 9.2, $d > 0$, and hence T is well defined. Moreover, conclusion (i) implies that $0 \leq T \leq 1$, conclusion (ii) implies that $T = 0$ if and only if Y and $\mathbf{Z}$ are independent and conclusion (iii) implies that Y is a function of $\mathbf{Z}$ if and only if $T = 1$. This proves Theorem 2.1 when $p = 0$. Next, note that $T_n = Q_n/S_n$, where $Q_n = Q_n(Y, \mathbf{Z})$ and $S_n = S_n(Y)$, as defined in (9.3) and (9.4). By Theorem 9.2, $Q_n \to c$ and $S_n \to d$ in probability. Thus, $T_n \to c/d = T$ in probability. This proves Theorem 2.2 when $p = 0$.

11. Preparation for the proofs of Theorems 9.1 and 9.2. In this section, we prove some lemmas that are needed for the proofs of Theorems 9.1 and 9.2. Let Y be a random variable and $\mathbf{X}$ be an $\mathbb{R}^p$-valued random vector, defined on the same probability space. Define

$$F(t) := \mathbb{P}(Y \leq t), \qquad G(t) := \mathbb{P}(Y \geq t).$$

By the existence of regular conditional probabilities on regular Borel spaces (see, e.g., [20], Theorem 2.1.15 and Exercise 5.1.16), for each Borel set $A \subseteq \mathbb{R}$ there is a measurable map $\mathbf{x} \mapsto \mu_{\mathbf{x}}(A)$ from $\mathbb{R}^p$ into $[0, 1]$, such that:

(i) for any A, $\mu_{\mathbf{X}}(A)$ is a version of $\mathbb{P}(Y \in A|\mathbf{X})$, and
(ii) with probability one, $\mu_{\mathbf{X}}$ is a probability measure on $\mathbb{R}$.

In the above sense, $\mu_{\mathbf{x}}$ is the conditional law of Y given $\mathbf{X} = \mathbf{x}$. For each t, let

$$F_{\mathbf{X}}(t) := \mu_{\mathbf{X}}\big((-\infty, t]\big), \qquad G_{\mathbf{X}}(t) := \mu_{\mathbf{X}}\big([t, \infty)\big).$$

Define

$$Q(Y, \mathbf{X}) := \int \operatorname{Var}\big(G_{\mathbf{X}}(t)\big)\, d\mu(t). \tag{11.1}$$

LEMMA 11.1. *Let $Q(Y, \mathbf{X})$ be as above. Then $Q(Y, \mathbf{X}) = 0$ if and only if Y and $\mathbf{X}$ are independent.*

PROOF. If Y and $\mathbf{X}$ are independent, then for any t, $\mathbb{P}(Y \geq t|\mathbf{X}) = \mathbb{P}(Y \geq t)$ almost surely. Thus, $G_{\mathbf{X}}(t) = G(t)$ almost surely, and so $\operatorname{Var}(G_{\mathbf{X}}(t)) = 0$. Consequently, $Q(Y, \mathbf{X}) = 0$.

Conversely, suppose that $Q(Y, \mathbf{X}) = 0$. Then there is a set $A \subseteq \mathbb{R}$ such that $\mu(A) = 1$ and $\operatorname{Var}(G_{\mathbf{X}}(t)) = 0$ for every $t \in A$. Since $\mathbb{E}(G_{\mathbf{X}}(t)) = G(t)$, $G_{\mathbf{X}}(t) = G(t)$ almost surely for each $t \in A$. We claim that $A = \mathbb{R}$.

To show this, take any $t \in \mathbb{R}$. If $\mu(\{t\}) > 0$, then clearly t must be a member of A and there is nothing more to prove. So assume that $\mu(\{t\}) = 0$. This implies that G is right continuous at t.

There are two possibilities. First, suppose that $G(s) < G(t)$ for all $s > t$. Then for each $s > t$, $\mu([t, s)) > 0$, and hence A must intersect $[t, s)$. This shows that there is a sequence r_n in A such that r_n decreases to t. Since $G_{\mathbf{X}}(r_n) = G(r_n)$ almost surely for each n, this implies that with probability one,

$$G_{\mathbf{X}}(t) \geq \lim_{n\to\infty} G_{\mathbf{X}}(r_n) = \lim_{n\to\infty} G(r_n) = G(t).$$

But $\mathbb{E}(G_{\mathbf{X}}(t)) = G(t)$. Thus, $G_{\mathbf{X}}(t) = G(t)$ almost surely.

The second possibility is that there is some $s > t$ such that $G(s) = G(t)$. Take the largest such s, which exists because G is left continuous. If $s = \infty$, then $G(t) = G(s) = 0$, and hence $G_{\mathbf{X}}(t) = 0$ almost surely because $\mathbb{E}(G_{\mathbf{X}}(t)) = G(t)$. Suppose that $s < \infty$. Then either $\mu(\{s\}) > 0$, which implies that $G_{\mathbf{X}}(s) = G(s)$ almost surely, or $\mu(\{s\}) = 0$ and $G(r) < G(s)$ for all $r > s$, which again implies that $G_{\mathbf{X}}(s) = G(s)$ almost surely, by the previous paragraph. Therefore, in either case, with probability one,

$$G_{\mathbf{X}}(t) \geq G_{\mathbf{X}}(s) = G(s) = G(t).$$

Since $\mathbb{E}(G_{\mathbf{X}}(t)) = G(t)$, this implies that $G_{\mathbf{X}}(t) = G(t)$ almost surely.

This completes the proof of our claim that $\operatorname{Var}(G_{\mathbf{X}}(t)) = 0$ for every $t \in \mathbb{R}$. In particular, for each $t \in \mathbb{R}$, $G_{\mathbf{X}}(t) = G(t)$ almost surely. Therefore, for any $t \in \mathbb{R}$ and any Borel set $B \subseteq \mathbb{R}^p$,

$$\begin{aligned}\mathbb{P}\big(\{Y \geq t\} \cap \{\mathbf{X} \in B\}\big) &= \mathbb{E}\big(\mathbb{P}(Y \geq t|\mathbf{X})1_{\{\mathbf{X}\in B\}}\big)\\ &= G(t)\mathbb{P}(\mathbf{X} \in B) = \mathbb{P}(Y \geq t)\mathbb{P}(\mathbf{X} \in B).\end{aligned}$$

This proves that Y and $\mathbf{X}$ are independent. □

Let $\mathbf{Z}$ be an $\mathbb{R}^q$-valued random vector defined on the same probability space as Y and $\mathbf{X}$, and let $\mathbf{W} = (\mathbf{X}, \mathbf{Z})$ be the concatenation of $\mathbf{X}$ and $\mathbf{Z}$.

LEMMA 11.2. *Let $\mathbf{W}$ be as above. Then $Q(Y, \mathbf{W}) \geq Q(Y, \mathbf{X})$, and equality holds if and only if Y and $\mathbf{Z}$ are conditionally independent given $\mathbf{X}$.*

PROOF. Since $G_{\mathbf{X}}(t) = \mathbb{E}(G_{\mathbf{W}}(t)|\mathbf{X})$, it follows that for each t,

$$\operatorname{Var}\big(G_{\mathbf{X}}(t)\big) \leq \operatorname{Var}\big(G_{\mathbf{W}}(t)\big).$$

Consequently, $Q(Y, \mathbf{W}) \geq Q(Y, \mathbf{X})$. If Y and $\mathbf{Z}$ are conditionally independent given $\mathbf{X}$, then for any t,

$$G_{\mathbf{W}}(t) = \mathbb{P}(Y \geq t|\mathbf{X}, \mathbf{Z}) = \mathbb{P}(Y \geq t|\mathbf{X}) = G_{\mathbf{X}}(t).$$

Thus, $Q(Y, \mathbf{W}) = Q(Y, \mathbf{X})$. Conversely, suppose that $Q(Y, \mathbf{W}) = Q(Y, \mathbf{X})$. Notice that

$$\begin{aligned}\operatorname{Var}\big(G_{\mathbf{W}}(t)\big) - \operatorname{Var}\big(G_{\mathbf{X}}(t)\big) &= \operatorname{Var}\big(G_{\mathbf{W}}(t)\big) - \operatorname{Var}\big(\mathbb{E}\big(G_{\mathbf{W}}(t)|\mathbf{X}\big)\big) \\ &= \mathbb{E}\big(\operatorname{Var}\big(G_{\mathbf{W}}(t)|\mathbf{X}\big)\big) \\ &= \mathbb{E}\big(G_{\mathbf{W}}(t) - G_{\mathbf{X}}(t)\big)^2.\end{aligned}$$

Thus,

$$Q(Y, \mathbf{W}) - Q(Y, \mathbf{X}) = \int \mathbb{E}\big(G_{\mathbf{W}}(t) - G_{\mathbf{X}}(t)\big)^2 \, d\mu(t).$$

So, if $Q(Y, \mathbf{W}) = Q(Y, \mathbf{X})$, then there is a Borel set $A \subseteq \mathbb{R}$ such that $\mu(A) = 1$ and $G_{\mathbf{W}}(t) = G_{\mathbf{X}}(t)$ almost surely for every $t \in A$. We claim that $A = \mathbb{R}$. Let us now prove this claim. The proof is similar to the proof of the analogous claim in Lemma 11.1, with a few additional complications.

Take any $t \in \mathbb{R}$. If $\mu(\{t\}) > 0$, then clearly t must be a member of A. So assume that $\mu(\{t\}) = 0$. As before, this implies that G is right continuous at t. Take any sequence t_n decreasing to t. Then $G(t) - G(t_n) \to 0$. But

$$G(t) - G(t_n) = \mathbb{E}\big(G_{\mathbf{X}}(t) - G_{\mathbf{X}}(t_n)\big),$$

and $G_{\mathbf{X}}(t) - G_{\mathbf{X}}(t_n)$ is a nonnegative random variable. Thus, $G_{\mathbf{X}}(t) - G_{\mathbf{X}}(t_n) \to 0$ in probability and, therefore, there is a subsequence n_k such that $G_{\mathbf{X}}(t_{n_k})$ converges to $G_{\mathbf{X}}(t)$ almost surely. But from the properties of the regular conditional probability $\mu_{\mathbf{x}}$ we know that $G_{\mathbf{X}}$ is a nonincreasing function almost surely. Thus, it follows that $G_{\mathbf{X}}$ is right continuous at t almost surely.

Now, as before, there are two possibilities. First, suppose that $G(s) < G(t)$ for all $s > t$. Then for each $s > t$, $\mu([t, s)) > 0$, and hence A must intersect $[t, s)$. This shows that there is a sequence r_n in A such that r_n decreases to t. Since $G_{\mathbf{W}}(r_n) = G_{\mathbf{X}}(r_n)$ almost surely for each n and $G_{\mathbf{X}}$ is right continuous at t with probability one, this implies that with probability one,

$$G_{\mathbf{W}}(t) \geq \lim_{n\to\infty} G_{\mathbf{W}}(r_n) = \lim_{n\to\infty} G_{\mathbf{X}}(r_n) = G_{\mathbf{X}}(t).$$

But $\mathbb{E}(G_{\mathbf{W}}(t)|\mathbf{X}) = G_{\mathbf{X}}(t)$. Thus, $G_{\mathbf{W}}(t) = G_{\mathbf{X}}(t)$ almost surely.

The second possibility is that there is some $s > t$ such that $G(s) = G(t)$. Take the largest such s, which exists because G is left continuous. If $s = \infty$, then $G(t) = G(s) = 0$, and hence $G_{\mathbf{W}}(t) = G_{\mathbf{X}}(t) = 0$ almost surely because $\mathbb{E}(G_{\mathbf{W}}(t)) = \mathbb{E}(G_{\mathbf{X}}(t)) = G(t)$. Suppose that $s < \infty$. Then either $\mu(\{s\}) > 0$, which implies that $G_{\mathbf{W}}(s) = G_{\mathbf{X}}(s)$ almost surely (by the previous step), or $\mu(\{s\}) = 0$ and $G(r) < G(s)$ for all $r > s$, which again implies that $G_{\mathbf{W}}(s) = G_{\mathbf{X}}(s)$ almost surely (also by the previous step). Therefore in either case, with probability one,

$$G_{\mathbf{W}}(t) \geq G_{\mathbf{W}}(s) = G_{\mathbf{X}}(s).$$

Now, $\mathbb{P}(Y \in [t, s)) = 0$, and hence $\mathbb{P}(Y \in [t, s)|\mathbf{X}) = 0$ almost surely. In other words, $G_{\mathbf{X}}(t) = G_{\mathbf{X}}(s)$ almost surely. Thus, $G_{\mathbf{W}}(t) \geq G_{\mathbf{X}}(t)$ almost surely. Since $\mathbb{E}(G_{\mathbf{W}}(t)|\mathbf{X}) = G_{\mathbf{X}}(t)$, this implies that $G_{\mathbf{W}}(t) = G_{\mathbf{X}}(t)$ almost surely. This completes the proof of our claim that $A = \mathbb{R}$.

Therefore, for any $t \in \mathbb{R}$ and any Borel set $B \subseteq \mathbb{R}^{p'}$,

$$\begin{aligned}\mathbb{P}\big(\{Y \geq t\} \cap \{\mathbf{Z} \in B\}|\mathbf{X}\big) &= \mathbb{E}\big(\mathbb{P}\big(\{Y \geq t\} \cap \{\mathbf{Z} \in B\}|\mathbf{W}\big)|\mathbf{X}\big) \\ &= \mathbb{E}\big(\mathbb{P}(Y \geq t|\mathbf{W}) 1_{\{\mathbf{Z}\in B\}}|\mathbf{X}\big) \\ &= \mathbb{E}\big(G_{\mathbf{X}}(t) 1_{\{\mathbf{Z}\in B\}}|\mathbf{X}\big) \\ &= \mathbb{P}(Y \geq t|\mathbf{X})\mathbb{P}(\mathbf{Z} \in B|\mathbf{X}).\end{aligned}$$

This proves that Y and $\mathbf{Z}$ are conditionally independent given $\mathbf{X}$. □

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ be an infinite sequence of i.i.d. copies of $\mathbf{X}$. For each $n \geq 2$ and each $1 \leq i \leq n$, let $\mathbf{X}_{n,i}$ be the Euclidean nearest-neighbor of $\mathbf{X}_i$ among $\{\mathbf{X}_j : 1 \leq j \leq n, j \neq i\}$. Ties are broken at random.

LEMMA 11.3. *With probability one,* $\mathbf{X}_{n,1} \to \mathbf{X}_1$ *as* $n \to \infty$.

PROOF. Let ν be the law of $\mathbf{X}$. Let A be the support of ν. Recall that A is the set of all $\mathbf{x} \in \mathbb{R}^p$ such that any open ball containing $\mathbf{x}$ has strictly positive ν-measure. From this definition, it follows easily that the complement of A is a countable union of open balls of ν-measure zero. Consequently, $\mathbf{X} \in A$ with probability one.

Take any $\varepsilon > 0$. Let B be the ball of radius ε centered at $\mathbf{X}_1$. Then

$$\mathbb{P}\big(\|\mathbf{X}_1 - \mathbf{X}_{n,1}\| \geq \varepsilon | \mathbf{X}_1\big) \leq \big(1 - \nu(B)\big)^{n-1}.$$

Since $\mathbf{X}_1 \in A$ almost surely, it follows that $\nu(B) > 0$ almost surely. Thus,

$$\lim_{n\to\infty} \mathbb{P}\big(\|\mathbf{X}_1 - \mathbf{X}_{n,1}\| \geq \varepsilon | \mathbf{X}_1\big) = 0$$

almost surely, and hence

$$\lim_{n\to\infty} \mathbb{P}\big(\|\mathbf{X}_1 - \mathbf{X}_{n,1}\| \geq \varepsilon\big) = 0.$$

This proves that $\|\mathbf{X}_1 - \mathbf{X}_{n,1}\| \to 0$ in probability. But $\|\mathbf{X}_1 - \mathbf{X}_{n,1}\|$ is decreasing in n. Therefore, $\|\mathbf{X}_1 - \mathbf{X}_{n,1}\| \to 0$ almost surely. □

Take any particular realization of $\mathbf{X}_1, \ldots, \mathbf{X}_n$. In this realization, for each $1 \leq i \leq n$, let $K_{n,i}$ be the number of j such that $\mathbf{X}_i$ is a nearest neighbor of $\mathbf{X}_j$ (not necessarily the randomly chosen one) and $\mathbf{X}_j \neq \mathbf{X}_i$. The following is a well-known geometric fact (see, e.g., [62], p. 102).

LEMMA 11.4. *There is a deterministic constant* $C(p)$, *depending only on the dimension* p, *such that* $K_{n,1} \leq C(p)$ *always.*

PROOF. Consider a triangle with vertices $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ in $\mathbb{R}^p$, where $\mathbf{y} \neq \mathbf{x}$ and $\mathbf{z} \neq \mathbf{x}$. Suppose that the angle at $\mathbf{x}$ is strictly less than 60° and $\|\mathbf{x} - \mathbf{y}\| \leq \|\mathbf{x} - \mathbf{z}\|$. Then

$$\frac{(\mathbf{y} - \mathbf{x}) \cdot (\mathbf{z} - \mathbf{x})}{\|\mathbf{y} - \mathbf{x}\| \|\mathbf{z} - \mathbf{x}\|} > \cos 60^\circ = \frac{1}{2}.$$

Consequently,

$$\begin{aligned}
\|\mathbf{z} - \mathbf{y}\|^2 &= \|\mathbf{z} - \mathbf{x}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 + 2(\mathbf{z} - \mathbf{x}) \cdot (\mathbf{x} - \mathbf{y}) \\
&< \|\mathbf{z} - \mathbf{x}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 - \|\mathbf{y} - \mathbf{x}\| \|\mathbf{z} - \mathbf{x}\| \\
&\leq \|\mathbf{z} - \mathbf{x}\|^2,
\end{aligned}$$

where the last inequality holds because $\|\mathbf{x} - \mathbf{y}\| \leq \|\mathbf{x} - \mathbf{z}\|$. Thus, if K is a cone at $\mathbf{x}$ of aperture less than 60°, and $\mathbf{x}_1, \ldots, \mathbf{x}_m$ is a finite list of points in $K \setminus \{\mathbf{x}\}$ (not necessarily distinct), then there can be at most one i such that the nearest neighbor of $\mathbf{x}_i$ in $\{\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_m\}$ is $\mathbf{x}$.

Now it is not difficult to see that there is a deterministic constant $C(p)$ depending only on p such that the whole of $\mathbb{R}^p$ can be covered by at most $C(p)$ cones of apertures less than 60° based at any given point. Take this point to be $\mathbf{X}_1$. Then within each cone, there can be at most one $\mathbf{X}_j$, which is not equal to $\mathbf{X}_1$, and whose nearest neighbor is $\mathbf{X}_1$. This shows that there can be at most $C(p)$ points distinct from $\mathbf{X}_1$ whose nearest neighbor is $\mathbf{X}_1$, completing the proof of the lemma. □

LEMMA 11.5. *There is a constant $C(p)$ depending only on p, such that for any measurable $f:\mathbb{R}^p\to[0,\infty)$ and any n, $\mathbb{E}(f(\mathbf{X}_{n,1}))\le C(p)\mathbb{E}(f(\mathbf{X}_1))$.*

PROOF. Since f is nonnegative,

$$\begin{aligned}\mathbb{E}\big(f(\mathbf{X}_{n,i})\big)&\le\mathbb{E}\big(f(\mathbf{X}_i)\big)+\mathbb{E}\big(f(\mathbf{X}_{n,i})1_{\{\mathbf{X}_{n,i}\neq X_i\}}\big)\\&\le\mathbb{E}\big(f(\mathbf{X}_i)\big)+\sum_{j=1}^n\mathbb{E}\big(f(\mathbf{X}_j)1_{\{\mathbf{X}_j=\mathbf{X}_{n,i},\,\mathbf{X}_j\neq\mathbf{X}_i\}}\big).\end{aligned}$$

Therefore by symmetry,

$$\begin{aligned}\mathbb{E}\big(f(\mathbf{X}_{n,1})\big)&=\frac{1}{n}\sum_{i=1}^n\mathbb{E}\big(f(\mathbf{X}_{n,i})\big)\\&\le\frac{1}{n}\sum_{i=1}^n\mathbb{E}\big(f(\mathbf{X}_i)\big)+\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\mathbb{E}\big(f(\mathbf{X}_j)1_{\{\mathbf{X}_j=\mathbf{X}_{n,i},\,\mathbf{X}_j\neq X_i\}}\big)\\&=\mathbb{E}\big(f(\mathbf{X}_1)\big)+\frac{1}{n}\sum_{j=1}^n\mathbb{E}\Bigg(f(\mathbf{X}_j)\sum_{i=1}^n1_{\{\mathbf{X}_j=\mathbf{X}_{n,i},\,\mathbf{X}_j\neq\mathbf{X}_i\}}\Bigg)\\&\le\mathbb{E}\big(f(\mathbf{X}_1)\big)+\frac{1}{n}\sum_{j=1}^n\mathbb{E}\big(f(\mathbf{X}_j)K_{n,j}\big)=\mathbb{E}\big(f(\mathbf{X}_1)(1+K_{n,1})\big).\end{aligned}$$

By Lemma 11.4, this completes the proof. □

For the next result, we will need the following version of Lusin's theorem (proved, e.g., by combining [46], Theorem 2.18 and Theorem 2.24).

LEMMA 11.6 (Special case of Lusin's theorem). *Let $f:\mathbb{R}^p\to\mathbb{R}$ be a measurable function and γ be a probability measure on $\mathbb{R}^p$. Then, given any $\varepsilon>0$, there is a compactly supported continuous function $g:\mathbb{R}^p\to\mathbb{R}$ such that $\gamma(\{\mathbf{x}:f(\mathbf{x})\neq g(\mathbf{x})\})<\varepsilon$.*

LEMMA 11.7. *For any measurable $f:\mathbb{R}^p\to\mathbb{R}$, $f(\mathbf{X}_1)-f(\mathbf{X}_{n,1})$ tends to 0 in probability as $n\to\infty$.*

PROOF. Fix some $\varepsilon>0$. Let g be a function as in Lemma 11.6, for the given f and ε, and $\gamma=$ the law of $\mathbf{X}_1$. Then note that for any $\delta>0$,

$$\begin{aligned}&\mathbb{P}\big(|f(\mathbf{X}_1)-f(\mathbf{X}_{n,1})|>\delta\big)\\&\quad\le\mathbb{P}\big(|g(\mathbf{X}_1)-g(\mathbf{X}_{n,1})|>\delta\big)+\mathbb{P}\big(f(\mathbf{X}_1)\neq g(\mathbf{X}_1)\big)\\&\qquad+\mathbb{P}\big(f(\mathbf{X}_{n,1})\neq g(\mathbf{X}_{n,1})\big).\end{aligned}$$

By Lemma 11.3 and the continuity of g,

$$\lim_{n\to\infty}\mathbb{P}\big(|g(\mathbf{X}_1)-g(\mathbf{X}_{n,1})|>\delta\big)=0.$$

By the construction of g,

$$\mathbb{P}\big(f(\mathbf{X}_1)\neq g(\mathbf{X}_1)\big)<\varepsilon.$$

Finally, by Lemma 11.5,

$$\mathbb{P}\big(f(\mathbf{X}_{n,1})\neq g(\mathbf{X}_{n,1})\big)\le C(p)\mathbb{P}\big(f(\mathbf{X}_1)\neq g(\mathbf{X}_1)\big)\le C(p)\varepsilon.$$

Putting it all together, we get

$$\limsup_{n\to\infty} \mathbb{P}\big(|f(\mathbf{X}_1) - f(\mathbf{X}_{n,1})| > \delta\big) \le \varepsilon + C(p)\varepsilon.$$

Since ε and δ are arbitrary, this completes the proof of the lemma. $\square$

Let $(Y_1, \mathbf{X}_1), \dots, (Y_n, \mathbf{X}_n)$ be i.i.d. copies of $(Y, \mathbf{X})$. Let F_n be the empirical distribution function of $Y_1, \dots, Y_n$, that is,

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i \le t\}}.$$

Also let

$$G_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i \ge t\}}.$$

For each i, let $N(i)$ be the index j such that $\mathbf{X}_j = \mathbf{X}_{n,i}$ (ties broken at random). Define

$$Q_n = Q_n(Y, \mathbf{X}) := \frac{1}{n}\sum_{i=1}^{n}\big(\min\{F_n(Y_i), F_n(Y_{N(i)})\} - G_n(Y_i)^2\big). \tag{11.2}$$

Note that this is exactly the statistic $Q_n(Y, \mathbf{X})$ defined in equation (9.3) of Section 9.

LEMMA 11.8. *Let Q_n be defined as above. Then*

$$\lim_{n\to\infty} \mathbb{E}\big(Q_n(Y, \mathbf{X})\big) = Q(Y, \mathbf{X}).$$

PROOF. Let

$$Q_n' := \frac{1}{n}\sum_{i=1}^{n}\big(\min\{F(Y_i), F(Y_{N(i)})\} - G(Y_i)^2\big) \tag{11.3}$$

and let

$$\Delta_n := \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| + \sup_{t\in\mathbb{R}} |G_n(t) - G(t)|.$$

Then by the triangle inequality,

$$|Q_n' - Q_n| \le 3\Delta_n. \tag{11.4}$$

On the other hand, by the Glivenko–Cantelli theorem, $\Delta_n \to 0$ almost surely as $n \to \infty$. Since Δ_n is bounded by 2, this implies that

$$\lim_{n\to\infty} \mathbb{E}|Q_n' - Q_n| = 0.$$

Thus, it suffices to show that $\mathbb{E}(Q_n')$ converges to $Q(Y, \mathbf{X})$. First, notice that

$$\min\{F(Y_1), F(Y_{N(1)})\} = \int 1_{\{Y_1 \ge t\}} 1_{\{Y_{N(1)} \ge t\}}\, d\mu(t).$$

Let $\mathcal{F}$ be the σ-algebra generated by $\mathbf{X}_1, \dots, \mathbf{X}_n$ and the random variables used for breaking ties in the selection of nearest neighbors. Then for any t,

$$\mathbb{E}(1_{\{Y_1 \ge t\}} 1_{\{Y_{N(1)} \ge t\}} | \mathcal{F}) = G_{\mathbf{X}_1}(t) G_{\mathbf{X}_{N(1)}}(t).$$

Note that $\mathbf{X}_{N(1)} = \mathbf{X}_{n,1}$. Also, recall that by the properties of the regular conditional probability $\mu_{\mathbf{x}}$, the map $\mathbf{x} \mapsto G_{\mathbf{x}}(t)$ is measurable. Therefore, by the above identity and Lemma 11.7, we have

$$\lim_{n\to\infty} \mathbb{E}(1_{\{Y_1 \geq t\}} 1_{\{Y_{N(1)} \geq t\}}) = \mathbb{E}\big(G_{\mathbf{X}}(t)^2\big).$$

Thus,

$$\lim_{n\to\infty} \mathbb{E}(Q_n') = \int \big(\mathbb{E}\big(G_{\mathbf{X}}(t)^2\big) - G(t)^2\big)\, d\mu(t).$$

Since $\mathbb{E}(G_{\mathbf{X}}(t)) = G(t)$, this completes the proof of the lemma. □

LEMMA 11.9. *There are positive constants C_1 and C_2 depending only on the dimension p such that for any n and any $t \geq 0$,*

$$\mathbb{P}\big(\big|Q_n - \mathbb{E}(Q_n)\big| \geq t\big) \leq C_1 e^{-C_2 n t^2}.$$

PROOF. Throughout this proof, $C(p)$ will denote any constant that depends only on p. The value of $C(p)$ may change from line to line.

In addition to the variables $\mathbf{X}_i$ and Y_i, in this proof we will make use of i.i.d. Uniform[0, 1] random variables $U_1, \ldots, U_n$, where U_i is used for breaking ties if $\mathbf{X}_i$ has multiple nearest neighbors.

Our plan is to use the bounded difference concentration inequality [40]. For that, we have to get a bound on the maximum possible change in Q_n if one $(Y_i, \mathbf{X}_i, U_i)$ is replaced by some alternative value $(Y_i', \mathbf{X}_i', U_i')$. We first write $Q_n = A_n + B_n$, where

$$A_n := \frac{1}{n}\sum_{i=1}^n \min\big\{F_n(Y_i), F_n(Y_{N(i)})\big\}, \qquad B_n := \frac{1}{n}\sum_{i=1}^n G_n(Y_i)^2.$$

It is not hard to see that after the above replacement, each $G_n(Y_j)$ can change by at most $1/n$, and since these quantities are in $[0, 1]$, B_n can change by at most $2/n$. Therefore, the bounded difference inequality gives

$$\mathbb{P}\big(\big|B_n - \mathbb{E}(B_n)\big| \geq t\big) \leq 2e^{-nt^2/8}. \tag{11.5}$$

Unfortunately, A_n is not well behaved with respect to this kind of perturbation, so we have to first replace A_n by some more manageable quantity. Take a realization of $(Y_1, \mathbf{X}_1, U_1), \ldots, (Y_n, \mathbf{X}_n, U_n)$. Define an equivalence relation on $\{1, \ldots, n\}$ by declaring that i and j are equivalent if $\mathbf{X}_i = \mathbf{X}_j$. Call an equivalence class a "cluster" if its size is greater than one, and a "singleton" otherwise. Note that if i belongs to a cluster $\mathcal{C}$, then $N(i)$ must necessarily be also a member of the same cluster. In fact, $N(i)$ would be chosen uniformly at random (using U_i) from $\mathcal{C} \setminus \{i\}$.

Let $\mathfrak{C}$ denote the set of all clusters and $\mathfrak{S}$ denote the set of all singletons. For convenience, let us define

$$a_{i,j} := \min\big\{F_n(Y_i), F_n(Y_j)\big\},$$

so that

$$A_n = \frac{1}{n}\sum_{\mathcal{C}\in\mathfrak{C}}\sum_{i\in\mathcal{C}} a_{i,N(i)} + \frac{1}{n}\sum_{i\in\mathfrak{S}} a_{i,N(i)}.$$

Let $\mathcal{G}$ denote the σ-algebra generated by $(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)$ and $(U_i)_{i\in\mathfrak{S}}$. Define $A_n' := \mathbb{E}(A_n|\mathcal{G})$. Then it is clear that

$$A_n' = \frac{1}{n}\sum_{\mathcal{C}\in\mathfrak{C}} b(\mathcal{C}) + \frac{1}{n}\sum_{i\in\mathfrak{S}} a_{i,N(i)}, \tag{11.6}$$

where

$$b(\mathcal{C}) := \frac{1}{|\mathcal{C}| - 1} \sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{C} \setminus \{i\}} a_{i,j}.$$

We will now use the bounded difference inequality to get a tail bound for the difference $A_n - A_n'$. Conditional on $\mathcal{G}$, A_n is a function of $(U_i)_{i \notin \mathfrak{S}}$. If one such U_i is replaced by some other value U_i', then only $N(i)$ may be affected. Thus, A_n changes by at most $1/n$. Therefore, the bounded difference inequality gives

$$\mathbb{P}\big(|A_n - A_n'| \geq t | \mathcal{G}\big) \leq 2e^{-nt^2/2}.$$

Since the right side is deterministic, we can remove the conditioning on the left. But then the tail bound gives $\mathbb{E}|A_n - A_n'| < 3n^{-1/2}$. Therefore,

$$\begin{aligned} &\mathbb{P}\big(|A_n - \mathbb{E}(A_n)| \geq 3n^{-1/2} + t\big) \\ &\quad \leq \mathbb{P}\big(|A_n - A_n'| \geq t/2\big) + \mathbb{P}\big(|A_n' - \mathbb{E}(A_n')| \geq t/2\big) \\ &\quad \leq 2e^{-nt^2/8} + \mathbb{P}\big(|A_n' - \mathbb{E}(A_n')| \geq t/2\big). \end{aligned} \tag{11.7}$$

So we now need to get a tail bound for $A_n' - \mathbb{E}(A_n')$. Fortunately, A_n' is well behaved with respect to perturbing one coordinate. Let us now try to figure out the maximum possible change in A_n' if some $(Y_i, \mathbf{X}_i, U_i)$ is replaced by an alternative value $(Y_i', \mathbf{X}_i', U_i')$. We will do this in stages. First, let us replace $\mathbf{X}_i$ by $\mathbf{X}_i'$, keeping Y_i and U_i fixed. We know by Lemma 11.4 that in any configuration, for any i there can be at most $C(p)$ singletons j such that i is a nearest neighbor of j (not necessarily the chosen one). This fact will be used many times in the following argument. There are several cases to consider:

1. Suppose that i is in some cluster $\mathcal{C}$ of size ≥ 3 in the original configuration, and lands up in some other cluster $\mathcal{C}'$ in the new configuration. Then the set of singletons is the same in the two configurations. If j is a singleton, then $N(j)$ can change only if i is a nearest neighbor of j in either the original configuration or the final configuration. As noted above, there can be at most $C(p)$ such j. Therefore, due to these changes, A_n' can change by at most $C(p)/n$. On the other hand, $b(\mathcal{C})$ changes by at most 2, as seen from the following computation:

$$\begin{aligned} &\big|b(\mathcal{C}) - b\big(\mathcal{C} \setminus \{i\}\big)\big| \\ &= \Big| \frac{1}{|\mathcal{C}| - 1} \sum_{j \in \mathcal{C}} \sum_{k \in \mathcal{C} \setminus \{j\}} a_{j,k} - \frac{1}{|\mathcal{C}| - 2} \sum_{j \in \mathcal{C} \setminus \{i\}} \sum_{k \in \mathcal{C} \setminus \{i,j\}} a_{j,k} \Big| \\ &= \Big| \frac{1}{|\mathcal{C}| - 1} \sum_{k \in \mathcal{C} \setminus \{i\}} a_{i,k} + \frac{1}{|\mathcal{C}| - 1} \sum_{j \in \mathcal{C} \setminus \{i\}} a_{j,i} \\ &\quad - \frac{1}{(|\mathcal{C}| - 1)(|\mathcal{C}| - 2)} \sum_{j \in \mathcal{C} \setminus \{i\}} \sum_{k \in \mathcal{C} \setminus \{i,j\}} a_{j,k} \Big| \leq 2, \end{aligned}$$

where the last inequality holds because the $a_{i,j}$'s are in $[0, 1]$. A similar calculation shows that $b(\mathcal{C}')$ also changes by at most 1. Thus, overall, A_n' changes by at most $C(p)/n$.

2. Suppose that i is in some cluster $\mathcal{C}$ of size ≥ 3 in the original configuration, and pairs up with a singleton to form a new cluster in the new configuration. Again, $b(\mathcal{C})$ changes by at most 2, and the contributions from the singletons in (11.6) changes by at most $C(p)/n$, by the same logic as in case (1). The formation of the new cluster causes a change of at most $2/n$. Therefore, again, the change in A_n' is at most $C(p)/n$.

3. Suppose that i is in some cluster $\mathcal{C}$ of size ≥ 3 in the original configuration, and becomes a singleton in the new configuration. Then just as before, $b(\mathcal{C})$ changes by at most 2, and the contributions from singletons changes by at most $C(p)/n$.

4. Suppose that i is in some cluster $\mathcal{C}$ of size 2 in the original configuration, and pairs up with a singleton to form a new cluster in the new configuration. Again, the number of singletons j for which $N(j)$ changes due to this operation is bounded by $C(p)$, and the contributions from the clusters terms in (11.6) also changes by at most a bounded amount. Thus, the change in A'_n is at most $C(p)/n$.

5. Suppose that i is in some cluster $\mathcal{C}$ of size 2 in the original configuration, and becomes a singleton in the new configuration. Proceeding as before, we see that A'_n changes by at most $C(p)/n$.

6. Suppose that i is a singleton in the original configuration and remains so in the new configuration. Again, it is clear that the change in A'_n is at most $C(p)/n$.

7. All other cases are just reverses of the situations considered above. For example, if i is a singleton in the original configuration and becomes part of a cluster of size ≥ 3 in the new configuration, that is just the reverse of case (3).

Thus, we conclude that changing $\mathbf{X}_i$ to $\mathbf{X}'_i$ changes A'_n by at most $C(p)/n$. Next, let us change Y_i to Y'_i. Then $F_n(Y_j)$ changes by at most $1/n$ for each $j \neq i$, and $F_n(Y_i)$ changes by at most 1. Therefore, each $a_{j,k}$ changes by at most $1/n$ if $j \neq i$ and $k \neq i$, and by at most 1 if either index equals i. From this, it is easy to see that A'_n can change by at most $1/n$. Finally, let us replace U_i by U'_i. Then only $N(i)$ can change, and hence A'_n can change by at most $1/n$. Combing all three steps, we get

$$\mathbb{P}\big(\big|A'_n - \mathbb{E}\big(A'_n\big)\big| \geq t\big) \leq 2e^{-C(p)nt^2}.$$

Therefore, by (11.5) and (11.7), we get

$$\mathbb{P}\big(\big|A_n - \mathbb{E}(A_n)\big| \geq 3n^{-1/2} + t\big) \leq 6e^{-C(p)nt^2}.$$

If $t \geq 3n^{-1/2}$, this bound holds for $\mathbb{P}(|A_n - \mathbb{E}(A_n)| \geq 2t)$. If $t < 3n^{-1/2}$, we can choose $C_1 \geq 6$ so that $C_1 e^{-C(p)nt^2} \geq 1$, so that it is trivially a bound for $\mathbb{P}(|A_n - \mathbb{E}(A_n)| \geq 2t)$. This completes the proof. □

Combining Lemmas 11.8 and 11.9, we get the following corollary.

COROLLARY 11.10. *As $n \to \infty$, $Q_n(Y, \mathbf{X}) \to Q(Y, \mathbf{X})$ almost surely.*

12. Proof of Theorem 9.2. Note that convergence of $Q_n(Y, \mathbf{Z})$ to the deterministic limit c is the result of Corollary 11.10 (applied to the pair $(Y, \mathbf{Z})$ instead of $(Y, \mathbf{X})$). Showing that $S_n(Y)$ converges to d is easier. Let

$$S'_n(Y) = \frac{1}{n}\sum_{i=1}^{n} G(Y_i)\big(1 - G(Y_i)\big),$$

and

$$\Delta_n := \sup_{t \in \mathbb{R}} \big|G_n(t) - G(t)\big|.$$

Then by triangle inequality $|S_n(Y) - S'_n(Y)| \leq 4\Delta_n$, and by the Glivenko–Cantelli theorem $\Delta_n \to 0$ almost surely. So it is enough to show that $S'_n(Y)$ converges almost surely to d. But that is a consequence of the strong law of large numbers, since the Y_i's are i.i.d. and

$$\mathbb{E}\big(G(Y_i)\big(1 - G(Y_i)\big)\big) = \int G(t)\big(1 - G(t)\big)\, d\mu(t) = d.$$

This completes the proof of the convergence claims in the theorem. Next, by combining Corollary 11.10 and Lemma 11.1, we see that if Y and $\mathbf{X}$ are independent, then $c = 0$. This proves claim (i) in the theorem. On the other hand, if Y is a function of $\mathbf{Z}$, say $Y = f(\mathbf{Z})$ almost surely, then

$$\begin{aligned} c &= \int \operatorname{Var}\bigl(\mathbb{P}(Y \geq t|\mathbf{Z})\bigr)\, d\mu(t) \\ &= \int \operatorname{Var}\bigl(\mathbb{E}(1_{\{Y \geq t\}}|\mathbf{Z})\bigr)\, d\mu(t) \\ &= \int \operatorname{Var}(1_{\{f(\mathbf{Z}) \geq t\}})\, d\mu(t) \\ &= \int \mathbb{E}(1_{\{f(\mathbf{Z}) \geq t\}})\bigl(1 - \mathbb{E}(1_{\{f(\mathbf{Z}) \geq t\}})\bigr)\, d\mu(t) = d, \end{aligned}$$

which proves claim (ii) in the theorem. Finally, by the law of total variance we have

$$\operatorname{Var}(1_{\{Y \geq t\}}) = \mathbb{E}\bigl(\operatorname{Var}(1_{\{Y \geq t\}}|\mathbf{Z})\bigr) + \operatorname{Var}\bigl(\mathbb{P}(Y \geq t|\mathbf{Z})\bigr),$$

therefore, $0 \leq c \leq d$. Note that by Lemma 11.1, $c = 0$ if and only if Y is independent of $\mathbf{Z}$. To complete the proof of claim (iii), we have to show that if $c = d$ then Y is almost surely a function of $\mathbf{Z}$. If $c = d$, then

$$\int \mathbb{E}\bigl(G_{\mathbf{Z}}(t) - G_{\mathbf{Z}}(t)^2\bigr)\, d\mu(t) = 0,$$

which implies that $\mathbb{P}(E) = 1$, where E is the event

$$\int G_{\mathbf{Z}}(t)\bigl(1 - G_{\mathbf{Z}}(t)\bigr)\, d\mu(t) = 0. \tag{12.1}$$

Let A be the support of μ. Define

$$a_{\mathbf{Z}} := \sup\bigl\{t : G_{\mathbf{Z}}(t) = 1\bigr\}, \qquad b_{\mathbf{Z}} := \inf\bigl\{t : G_{\mathbf{Z}}(t) = 0\bigr\},$$

so that $a_{\mathbf{Z}} \leq b_{\mathbf{Z}}$. Now suppose that the event $\{a_{\mathbf{Z}} < b_{\mathbf{Z}}\} \cap E$ takes place. Since $G_{\mathbf{Z}}(t) \in (0, 1)$ for all $t \in (a_{\mathbf{Z}}, b_{\mathbf{Z}})$, the condition (12.1) implies that $\mu((a_{\mathbf{Z}}, b_{\mathbf{Z}})) = 0$. Since $(a_{\mathbf{Z}}, b_{\mathbf{Z}})$ is an open interval, this shows that $(a_{\mathbf{Z}}, b_{\mathbf{Z}}) \subseteq A^c$. On the other hand, under the given circumstance, we also have $\mathbb{P}(Y \in (a_{\mathbf{Z}}, b_{\mathbf{Z}})|\mathbf{Z}) > 0$. Thus, $\mathbb{P}(Y \in A^c|\mathbf{Z}) > 0$.

The above argument implies that if $\mathbb{P}(\{a_{\mathbf{Z}} < b_{\mathbf{Z}}\} \cap E) > 0$, then $\mathbb{P}(Y \in A^c) > 0$. But this is impossible, since A is the support of μ. Therefore, $\mathbb{P}(\{a_{\mathbf{Z}} < b_{\mathbf{Z}}\} \cap E) = 0$. But $\mathbb{P}(E) = 1$. Therefore $\mathbb{P}(a_{\mathbf{Z}} = b_{\mathbf{Z}}) = 1$. This implies that Y is almost surely a function of $\mathbf{Z}$.

13. Proof of Theorem 9.1. For the proof of Theorem 9.1, we need some additional lemmas.

LEMMA 13.1. *Let $Q_n(Y, \mathbf{Z}|\mathbf{X})$ be defined as in (9.1). Then $Q_n(Y, \mathbf{Z}|\mathbf{X})$ converges to $Q(Y, \mathbf{Z}|\mathbf{X})$ almost surely as $n \to \infty$, where*

$$Q(Y, \mathbf{Z}|\mathbf{X}) := \int \mathbb{E}\bigl(\operatorname{Var}\bigl(G_{\mathbf{W}}(t)|\mathbf{X}\bigr)\bigr)\, d\mu(t),$$

where, as before, $\mathbf{W} = (\mathbf{X}, \mathbf{Z})$.

PROOF. Note that $Q_n(Y, \mathbf{Z}|\mathbf{X}) = Q_n(Y, \mathbf{W}) - Q_n(Y, \mathbf{X})$. Also,

$$\mathbb{E}\bigl(G_{\mathbf{W}}(t)|\mathbf{X}\bigr) = G_{\mathbf{X}}(t),$$

which, by the law of total variance, gives

$$\operatorname{Var}\bigl(G_{\mathbf{W}}(t)\bigr) - \operatorname{Var}\bigl(G_{\mathbf{X}}(t)\bigr) = \mathbb{E}\bigl(\operatorname{Var}\bigl(G_{\mathbf{W}}(t)|\mathbf{X}\bigr)\bigr).$$

Thus,

$$Q(Y, \mathbf{Z}|\mathbf{X}) = Q(Y, \mathbf{W}) - Q(Y, \mathbf{X}).$$

The result now follows by Corollary 11.10. □

LEMMA 13.2. *For $S_n(Y, \mathbf{X})$ defined in (9.2),*

$$\lim_{n\to\infty} \mathbb{E}\bigl(S_n(Y, \mathbf{X})\bigr) = S(Y, \mathbf{X})$$

where $S(Y, \mathbf{X}) := \int \mathbb{E}(\operatorname{Var}(1_{\{Y\geq t\}}|\mathbf{X}))\, d\mu(t)$.

PROOF. The proof uses techniques developed in the proof of Lemma 11.8. Let

$$S'_n(Y, \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\bigl(F(Y_i) - \min\bigl\{F(Y_i), F(Y_{N(i)})\bigr\}\bigr),$$

and

$$\Delta_n := \sup_{t\in\mathbb{R}}\bigl|F_n(t) - F(t)\bigr|.$$

By the triangle inequality,

$$\bigl|S'_n(Y, \mathbf{X}) - S_n(Y, \mathbf{X})\bigr| \leq 4\Delta_n.$$

By the Glivenko–Cantelli theorem, $\Delta_n \to 0$ almost surely and since Δ_n is bounded by 1, we can conclude that

$$\lim_{n\to\infty} \mathbb{E}\bigl|S'_n(Y, \mathbf{X}) - S_n(Y, \mathbf{X})\bigr| = 0.$$

Then it is enough to show that $\mathbb{E}(S'_n(Y, \mathbf{X}))$ converges to $S(Y, \mathbf{X})$. Proceeding as in the proof of Lemma 11.8, we get

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\bigl(S'_n(Y, \mathbf{X})\bigr) &= \int \bigl(G(t) - \mathbb{E}\bigl(G_{\mathbf{X}}(t)^2\bigr)\bigr)\, d\mu(t)\\ &= \int \mathbb{E}\bigl(G_{\mathbf{X}}(t) - G_{\mathbf{X}}(t)^2\bigr)\, d\mu(t) = S(Y, \mathbf{X}),\end{aligned}$$

which completes the proof. □

LEMMA 13.3. *There are positive constants C_1 and C_2 depending only on p such that for any n and any $t \geq 0$,*

$$\mathbb{P}\bigl(\bigl|S_n(Y, \mathbf{X}) - \mathbb{E}\bigl(S_n(Y, \mathbf{X})\bigr)\bigr| \geq t\bigr) \leq C_1 e^{-C_2 n t^2}$$

PROOF. The concentration for the second term in the definition (9.2) was already argued in the proof of Lemma 11.9. For the first term, a simple application of the bounded difference inequality suffices. □

Finally, we are ready to prove Theorem 9.1.

PROOF OF THEOREM 9.1. Convergence of $Q_n(Y, \mathbf{Z}|\mathbf{X})$ almost surely to $a = Q(Y, \mathbf{Z}|\mathbf{X})$ is the content of Lemma 13.1, and convergence of $S_n(Y, \mathbf{X})$ to $b = S(Y, \mathbf{X})$ follows by Lemmas 13.2 and 13.3.

Let us now prove the claims (i), (ii) and (iii) of the theorem. First, let us prove (i). It is not hard to see that $a = Q(Y, \mathbf{W}) - Q(Y, \mathbf{X})$. Thus, if Y and $\mathbf{Z}$ are conditionally independent given $\mathbf{X}$, then by Lemma 11.2, $a = 0$. This proves (i). Next, note that

$$\begin{aligned} b - a &= \int \mathbb{E}\big(\mathrm{Var}(1_{\{Y \geq t\}}|\mathbf{X}) - \mathrm{Var}\big(\mathbb{E}(1_{\{Y \geq t\}}|\mathbf{Z}, \mathbf{X})|\mathbf{X}\big)\big)\, d\mu(t) \\ &= \int \mathbb{E}\big(\mathbb{E}\big(\mathrm{Var}(1_{\{Y \geq t\}}|\mathbf{Z}, \mathbf{X})|\mathbf{X}\big)\big)\, d\mu(t) \\ &= \int \mathbb{E}\big(\mathrm{Var}(1_{\{Y \geq t\}}|\mathbf{Z}, \mathbf{X})\big)\, d\mu(t). \end{aligned}$$

Now, if with probability one Y is a function of $\mathbf{Z}$ conditional on $\mathbf{X}$, then $\mathrm{Var}(1_{\{Y \geq t\}}|\mathbf{Z}, \mathbf{X}) = 0$ almost surely. Thus, the above expression shows that $a = b$ in this situation.

Finally, let us prove claim (iii). Note that the above expression for $b - a$ also shows that $0 \leq a \leq b$, since $\mathrm{Var}(1_{\{Y \geq t\}}|\mathbf{Z}, \mathbf{X}) \geq 0$. Thus, it suffices to prove the opposite implications for (i) and (ii).

If $a = 0$, then again by Lemma 11.2, we get that Y and $\mathbf{Z}$ are conditionally independent given $\mathbf{X}$. If $a = b$, then there exists a set $A \subseteq \mathbb{R}$ such that $\mu(A) = 1$ and for any $t \in A$ we have

$$\mathrm{Var}(1_{\{Y \geq t\}}|\mathbf{Z}, \mathbf{X}) = 0$$

almost surely. Proceeding as the last part of the proof of Theorem 9.2, we can now conclude that Y is almost surely equal to a function of $\mathbf{W}$. This implies that Y is almost surely a function of $\mathbf{Z}$ conditional on $\mathbf{X}$. □

14. Proof of Theorem 4.1. Throughout this section, we will assume that the assumptions (A1) and (A2) from Section 4 hold. In the following lemma, $\mathbf{X}_{n,1}$ is the nearest neighbor of $\mathbf{X}_1$ among $\mathbf{X}_2, \ldots, \mathbf{X}_n$ (with ties broken at random), as in previous sections.

LEMMA 14.1. *Let C_1 and C_2 be as in assumption* (A2). *Then there is some C_3 depending only on C_1, C_2 and p such that*

$$\mathbb{E}\big(\min\{\|\mathbf{X}_1 - \mathbf{X}_{n,1}\|, 1\}\big) \leq \begin{cases} C_3 n^{-1} (\log n)^3 & \text{if } p = 1, \\ C_3 n^{-1/p} (\log n)^{p+1} & \text{if } p \geq 2. \end{cases}$$

PROOF. Throughout this proof, C will denote any constant that depends only on C_1, C_2 and p. Take any $t > 0$ and $\varepsilon \in (n^{-1/p}, 1)$. Let B be the ball of radius t in $\mathbb{R}^p$ centered at the origin. Partition B into at most $Ct^p\varepsilon^{-p}$ small sets of diameter $\leq \varepsilon$. Let S be the small set containing $\mathbf{X}_1$. Then

$$\mathbb{P}\big(\|\mathbf{X}_1 - \mathbf{X}_{n,1}\| \geq \varepsilon\big) \leq \mathbb{P}(\mathbf{X}_1 \notin B) + \mathbb{P}(\mathbf{X}_2 \notin S, \ldots, \mathbf{X}_n \notin S).$$

Now note that

$$\mathbb{P}(\mathbf{X}_2 \notin S, \ldots, \mathbf{X}_n \notin S|\mathbf{X}_1) = \big(1 - \mathbb{P}(\mathbf{X}_2 \in S|\mathbf{X}_1)\big)^{n-1} = \big(1 - \nu(S)\big)^{n-1},$$

where ν is the law of $\mathbf{X}$. Let A be the collection of all small sets with ν-mass less than δ. Since there are at most $Ct^p\varepsilon^{-p}$ small sets, we get

$$\begin{aligned} \mathbb{E}\big[\big(1 - \nu(S)\big)^{n-1}\big] &\leq (1 - \delta)^{n-1} + \mathbb{P}(\mathbf{X}_1 \in A) \\ &\leq (1 - \delta)^{n-1} + Ct^p\varepsilon^{-p}\delta. \end{aligned}$$

Since $\mathbb{P}(\mathbf{X}_1 \notin B) \leq C_1 e^{-C_2 t}$, this gives

$$\mathbb{P}\big(\|\mathbf{X}_1 - \mathbf{X}_{n,1}\| \geq \varepsilon\big) \leq C_1 e^{-C_2 t} + (1 - \delta)^{n-1} + Ct^p\varepsilon^{-p}\delta.$$

Now choosing $\delta = Kn^{-1}\log n$ and $t = K\log(n\varepsilon^p)$ for some large enough K, we get

$$\mathbb{P}\big(\|\mathbf{X}_1 - \mathbf{X}_{n,1}\| \geq \varepsilon\big) \leq \frac{C(\log n)^{p+1}}{n\varepsilon^p}.$$

Thus,

$$\begin{aligned}\mathbb{E}\big(\min\{\|\mathbf{X}_1 - \mathbf{X}_{n,1}\|, 1\}\big) &= n^{-1/p} + \int_{n^{-1/p}}^1 \mathbb{P}\big(\|\mathbf{X}_1 - \mathbf{X}_{n,1}\| \geq \varepsilon\big)d\varepsilon \\ &\leq n^{-1/p} + \frac{C(\log n)^{p+1}}{n}\int_{n^{-1/p}}^1 \varepsilon^{-p}d\varepsilon.\end{aligned}$$

This is bounded by $Cn^{-1}(\log n)^3$ if $p = 1$, and $Cn^{-1/p}(\log n)^{p+1}$ if $p \geq 2$. □

In the next lemma, let $Q = Q(Y, \mathbf{X})$ be defined as in equation (11.1) and $Q_n = Q_n(Y, \mathbf{X})$ be defined as in equation (9.3).

LEMMA 14.2. *Let C and β be as in assumption* (A1) *and C_1 and C_2 be as in assumption* (A2). *Then there are K_1, K_2 and K_3 depending only on C, β, C_1, C_2 and p such that for any $t \geq 0$,*

$$\mathbb{P}\big(|Q_n - Q| \geq K_1 n^{-\min\{1/p,1/2\}}(\log n)^{p+\beta+1} + t\big) \leq K_2 e^{-K_3 nt^2}.$$

PROOF. Let Q_n' and Δ_n be as in the proof of Lemma 11.8. By the Dvoretzky–Kiefer–Wolfowitz inequality [21, 39], we know that for any $x \geq 0$,

$$\mathbb{P}(\sqrt{n}\Delta_n \geq x) \leq 2e^{-2x^2}.$$

From this, it follows that $\mathbb{E}(\Delta_n) \leq n^{-1/2}$ and, therefore, by (11.4),

$$\mathbb{E}\big|Q_n' - Q_n\big| \leq 3n^{-1/2}. \tag{14.1}$$

Arguing as in the proof of Lemma 11.8, we get

$$\mathbb{E}\big(Q_n'\big) = \int \big(\mathbb{E}\big(G_{\mathbf{X}_1}(t)G_{\mathbf{X}_{n,1}}(t)\big) - G(t)^2\big)\,d\mu(t).$$

On the other hand,

$$Q = \int \big(\mathbb{E}\big(G_{\mathbf{X}}(t)^2\big) - G(t)^2\big)\,d\mu(t).$$

Since $G_{\mathbf{x}}(t) \in [0, 1]$ for all $\mathbf{x}$ and t, this gives

$$\big|\mathbb{E}\big(Q_n'\big) - Q\big| \leq \int \mathbb{E}\big|G_{\mathbf{X}_1}(t) - G_{\mathbf{X}_{n,1}}(t)\big|\,d\mu(t).$$

Now note that by assumption (A1),

$$\big|G_{\mathbf{X}_{n,1}}(t) - G_{\mathbf{X}_1}(t)\big| \leq C\big(1 + \|\mathbf{X}_{n,1}\|^\beta + \|\mathbf{X}_1\|^\beta\big)\|\mathbf{X}_1 - \mathbf{X}_{n,1}\|.$$

Next, note that by assumption (A2), $\mathbb{P}(\|\mathbf{X}_1\| \geq t) \leq C_1 e^{-C_2 t}$. Therefore, by Lemma 11.5, the law of $\|\mathbf{X}_{n,1}\|$ also has an exponentially decaying tail. Lastly, note that $|G_{\mathbf{X}_{n,1}}(t) - G_{\mathbf{X}_1}(t)| \leq 1$. So, letting E be the event that the maximum of $\|\mathbf{X}_1\|$ and $\|\mathbf{X}_{n,1}\|$ is bigger than $K\log n$ for some suitably large K, we get

$$\begin{aligned}\mathbb{E}\big|G_{\mathbf{X}_{n,1}}(t) - G_{\mathbf{X}_1}(t)\big| &\leq \mathbb{P}(E) + \mathbb{E}\big(\big|G_{\mathbf{X}_{n,1}}(t) - G_{\mathbf{X}_1}(t)\big|1_{E^c}\big) \\ &\leq n^{-1} + L(\log n)^\beta \mathbb{E}\big(\min\{\|\mathbf{X}_1 - \mathbf{X}_{n,1}\|, 1\}\big)\end{aligned}$$

for some large constant L. It is now easy to complete the proof using Lemma 14.1, inequality (14.1), and Lemma 11.9. □

We are now ready to prove Theorem 4.1.

PROOF OF THEOREM 4.1. Recall from Section 10 that

$$T_n(Y, \mathbf{Z}|\mathbf{X}) = \frac{Q_n(Y, \mathbf{Z}|\mathbf{X})}{S_n(Y, \mathbf{X})},$$

and

$$T(Y, \mathbf{Z}|\mathbf{X}) = \frac{Q(Y, \mathbf{Z}|\mathbf{X})}{S(Y, \mathbf{X})},$$

where the quantity $Q(Y, \mathbf{Z}|\mathbf{X})$ is defined in Lemma 13.1 and $S(Y, \mathbf{X})$ is defined in Lemma 13.2. Now, as we observed in the proof of Lemma 13.1,

$$Q_n(Y, \mathbf{Z}|\mathbf{X}) = Q_n(Y, \mathbf{W}) - Q_n(Y, \mathbf{X}),$$

where $\mathbf{W} = (\mathbf{X}, \mathbf{Z})$. Therefore, by Lemma 14.2,

$$Q_n(Y, \mathbf{Z}|\mathbf{X}) - Q(Y, \mathbf{Z}|\mathbf{X}) = O_P\left(\frac{(\log n)^{p+q+\beta+1}}{n^{1/(p+q)}}\right).$$

By an exactly similar argument,

$$S_n(Y, \mathbf{X}) - S(Y, \mathbf{X}) = O_P\left(\frac{(\log n)^{p+\beta+1}}{n^{\min\{1/p, 1/2\}}}\right).$$

Finally, by part (iv) of Theorem 9.1, $S(Y, \mathbf{X}) \neq 0$. The proof is completed by combining these observations. □

15. Proof of Proposition 4.2. Take a bounded open ball B in $\mathbb{R}^p$ and let K be the closure of B. Let $g(y) := \max_{\mathbf{x}\in K} f(y|\mathbf{x})$. By assumption, $g(y)$ is bounded and decays faster than any negative power of $|y|$ as $|y| \to \infty$. Also, since K is bounded, the assumption on the derivatives of $\log f(y|\mathbf{x})$ implies that

$$h(y) := \max_{\mathbf{x}\in K}\left|\frac{\partial}{\partial x_i}\log f(y|\mathbf{x})\right|$$

is bounded above by a polynomial in $|y|$. Thus, for $\mathbf{x} \in K$ and $y \in \mathbb{R}$,

$$\begin{aligned}\left|\frac{\partial}{\partial x_i} f(y|\mathbf{x})\right| &= \left|f(y|\mathbf{x})\frac{\partial}{\partial x_i}\log f(y|\mathbf{x})\right| \\ &\le g(y)h(y),\end{aligned}$$

and $g(y)h(y)$ is an integrable function of y. This allows us to apply the dominated convergence theorem and conclude that for any $\mathbf{x} \in B$ (and hence any $\mathbf{x} \in \mathbb{R}^p$),

$$\begin{aligned}\left|\frac{\partial}{\partial x_i}\mathbb{P}(Y \ge t|\mathbf{X} = \mathbf{x})\right| &= \left|\frac{\partial}{\partial x_i}\int_t^\infty f(y|\mathbf{x})\,dy\right| \\ &= \left|\int_t^\infty \frac{\partial}{\partial x_i} f(y|\mathbf{x})\,dy\right| \\ &\le \int_{-\infty}^\infty \left|\frac{\partial}{\partial x_i}\log f(y|\mathbf{x})\right| f(y|\mathbf{x})\,dy.\end{aligned}$$

Now applying the assumption about the derivatives of $\log f(y|\mathbf{x})$, and the condition that $\mathbb{E}(Y^{2k}|\mathbf{X}=\mathbf{x})$ is bounded by a polynomial in $\|\mathbf{x}\|$ for any k, it follows easily that

$$\left|\frac{\partial}{\partial x_i}\mathbb{P}(Y \geq t|\mathbf{X}=\mathbf{x})\right|$$

is bounded above by a polynomial in $\|\mathbf{x}\|$. The second inequality in (A1) follows directly from this.

16. Proof of Theorem 6.1. Let $j_1, j_2, \ldots, j_p$ be the complete ordering of all variables produced by the stepwise algorithm in FOCI. Let $S_0 := \varnothing$, and for each $1 \leq k \leq p$, let $S_k := \{j_1, \ldots, j_k\}$. For $k > p$, let $S_k := S_p$. For any subset S, let $Q(Y, \mathbf{X}_S)$ be defined as in (11.1) and let $Q_n(Y, \mathbf{X}_S)$ be defined as in (9.3). Notice that $Q(Y, \mathbf{X}_S)$ is the same as the quantity $Q(S)$ defined in (6.1). Define these quantities to be zero if $S = \varnothing$. Let K be the integer part of $1/\delta + 2$. Let E' be the event that $|Q_n(Y, \mathbf{X}_{S_k}) - Q(Y, \mathbf{X}_{S_k})| \leq \delta/8$ for all $1 \leq k \leq K$, and let E be the event that S_K is sufficient.

LEMMA 16.1. *Suppose that E' has happened, and also that*

$$Q_n(Y, \mathbf{X}_{S_k}) - Q_n(Y, \mathbf{X}_{S_{k-1}}) \leq \frac{\delta}{2} \tag{16.1}$$

for some $1 \leq k \leq K$. Then S_{k-1} is sufficient.

PROOF. Take any $k \leq K$ such that (16.1) holds. If $k > p$ there is nothing to prove. So let us assume that $k \leq p$. An examination of the formula for T_n shows that for each k, j_k is the index j that maximizes $Q_n(Y, \mathbf{X}_{S_{k-1}\cup\{j\}})$ among all $j \notin S_{k-1}$. Since E' has happened, this implies that for any $j \notin S_{k-1}$,

$$\begin{aligned} Q(Y, \mathbf{X}_{S_{k-1}\cup\{j\}}) - Q(Y, \mathbf{X}_{S_{k-1}}) &\leq Q_n(Y, \mathbf{X}_{S_{k-1}\cup\{j\}}) - Q_n(Y, \mathbf{X}_{S_{k-1}}) + \frac{\delta}{4} \\ &\leq Q_n(Y, \mathbf{X}_{S_k}) - Q_n(Y, \mathbf{X}_{S_{k-1}}) + \frac{\delta}{4} \\ &\leq \frac{3\delta}{4}. \end{aligned}$$

Therefore, since $\delta > 0$, the definition of δ implies that S_{k-1} must be a sufficient subset of predictors. □

LEMMA 16.2. *The event E' implies E.*

PROOF. Suppose that E' has happened. Suppose also that (16.1) is violated for every $1 \leq k \leq K$. Since E' has happened, this implies that for each $k \leq K$,

$$\begin{aligned} Q(Y, \mathbf{X}_{S_k}) - Q(Y, \mathbf{X}_{S_{k-1}}) &\geq Q_n(Y, \mathbf{X}_{S_k}) - Q_n(Y, \mathbf{X}_{S_{k-1}}) - \frac{\delta}{4} \\ &\geq \frac{\delta}{4}. \end{aligned}$$

This gives

$$\begin{aligned} Q(Y, \mathbf{X}_{S_K}) &= \sum_{k=1}^{K} \big(Q(Y, \mathbf{X}_{S_k}) - Q(Y, \mathbf{X}_{S_{k-1}})\big) \\ &\geq \frac{K\delta}{4} \geq \left(\frac{1}{\delta} + 1\right)\frac{\delta}{4} > \frac{1}{4}. \end{aligned}$$

But the variance of any [0, 1]-valued random variable is bounded by $1/4$, which implies that $1/4$ is the maximum possible value of the statistic Q. This yields a contradiction, proving that (16.1) must hold for some $k \le K$. Therefore, by Lemma 16.1, S_K is sufficient. □

LEMMA 16.3. *There are positive constants L_1, L_2 and L_3 depending only on C, β, C_1, C_2 and K, such that*

$$\mathbb{P}(E') \ge 1 - L_1 p^{L_2} e^{-L_3 n}.$$

PROOF. Throughout this proof, $L_1, L_2, \ldots$ will denote constants that depend only on C, β, C_1, C_2 and K. By assumptions (A1′) and (A2′), and Lemma 14.2, there exist L_1, L_2 and L_3 such that for any S of size $\le K$ and any $t \ge 0$,

$$\begin{aligned}&\mathbb{P}\bigl(|Q_n(Y, \mathbf{X}_S) - Q(Y, \mathbf{X}_S)| \ge L_1 n^{-\min\{1/K, 1/2\}} (\log n)^{K+\beta+1} + t\bigr)\\ &\quad \le L_2 e^{-L_3 n t^2}.\end{aligned}$$

Call the event on the left $A_{S,t}$. Let

$$A_t := \bigcup_{|S| \le K} A_{S,t}.$$

Then by a simple union bound,

$$\mathbb{P}(A_t) \le L_2 p^K e^{-L_3 n t^2}.$$

Now choose $t = \delta/16$. If n is so large that

$$L_1 n^{-\min\{1/K, 1/2\}} (\log n)^{K+\beta+1} \le \frac{\delta}{16}, \tag{16.2}$$

then the above bound implies that

$$\mathbb{P}(E') \ge 1 - L_2 p^K e^{-L_4 n}. \tag{16.3}$$

Now the condition (16.2) can be written as $n \ge L_5$. Choose a constant $L_6 \ge L_2$ so large that for any $n < L_5$,

$$L_6 p^K e^{-L_3 n} \ge 1.$$

Then if $n < L_5$, we have $\mathbb{P}(E') \ge 1 - L_6 p^K e^{-L_3 n}$. Combining with (16.3), we see that this inequality holds without any constraint on n. □

LEMMA 16.4. *The event E' implies that $\hat{S}$ is sufficient.*

PROOF. Suppose that E' has happened. Consider two cases. First, suppose that FOCI has stopped at step K or later. Then $S_K \subseteq \hat{S}$. By Lemma 16.2, E has also happened, and hence S_K is sufficient. Therefore, in this case, $\hat{S}$ is sufficient. Next, suppose that FOCI has stopped at step $k - 1 < K$. Then by the definition of the stopping rule, we see that

$$Q_n(Y, \mathbf{X}_{S_k}) \le Q_n(Y, \mathbf{X}_{S_{k-1}}).$$

In particular, (16.1) holds. Since E' has happened, Lemma 16.1 now implies that $\hat{S} = S_{k-1}$ is sufficient. □

It is clear that Lemmas 16.3 and 16.4 together imply Theorem 6.1.

17. Proof of Theorem 7.1. We start with the following lemma about the variance of a certain kind of function of normal random variables.

LEMMA 17.1. *Let Φ be the standard normal c.d.f. and let $Z \sim N(0,1)$. There are positive constants C_1 and C_2 such that for any $a, b \in \mathbb{R}$,*

$$C_1 b^2 e^{-(a^2+b^2)} \le \operatorname{Var}\big(\Phi(a+bZ)\big) \le C_2 b^2.$$

PROOF. Since Z has the same law as $-Z$, and $\Phi(-x) = 1 - \Phi(x)$ for all x, it is easy to see that there is no loss of generality in assuming that a and b are nonnegative. Moreover, since the result is trivial when $b = 0$, let us also assume that $b > 0$. Let Z, Z' be i.i.d. $N(0,1)$ random variables, so that

$$\operatorname{Var}\big(\Phi(a+bZ)\big) = \frac{1}{2}\mathbb{E}\big[\big(\Phi(a+bZ) - \Phi\big(a+bZ'\big)\big)^2\big].$$

Let $\varphi = \Phi'$ be the standard normal p.d.f. Now,

$$\Phi(a+bZ) - \Phi\big(a+bZ'\big) = \varphi(Y) b\big(Z - Z'\big)$$

for some Y lying between $a+bZ$ and $a+bZ'$. Suppose that in a particular realization, Z and Z' both turn out to be in $[-1,1]$. Then Y lies between $a-b$ and $a+b$, which implies that $\varphi(Y)$ is at least as large as the minimum of $\varphi(a-b)$ and $\varphi(a+b)$. Thus, in this situation,

$$\begin{aligned}\varphi(Y) &\ge \frac{1}{\sqrt{2\pi}} \exp\biggl(-\frac{1}{2}\max\bigl\{(a-b)^2, (a+b)^2\bigr\}\biggr) \\ &\ge \frac{1}{\sqrt{2\pi}} \exp\bigl(-\bigl(a^2+b^2\bigr)\bigr).\end{aligned}$$

This shows that

$$\begin{aligned}\operatorname{Var}\big(\Phi(a+bZ)\big) &\ge \frac{1}{2}\mathbb{E}\big[\big(\Phi(a+bZ) - \Phi\big(a+bZ'\big)\big)^2; |Z| \le 1, \big|Z'\big| \le 1\big] \\ &\ge C_1 b^2 e^{-(a^2+b^2)} \mathbb{E}\big[\big(Z - Z'\big)^2; |Z| \le 1, \big|Z'\big| \le 1\big] \\ &= C_2 b^2 e^{-(a^2+b^2)},\end{aligned}$$

where C_1 and C_2 are positive universal constants. This proves the lower bound. For the upper bound, simply observe that since φ is uniformly bounded by $1/\sqrt{2\pi}$, we have

$$\operatorname{Var}\big(\Phi(a+bZ)\big) \le \frac{b^2}{4\pi}\mathbb{E}\big[\big(Z - Z'\big)^2\big] = \frac{b^2}{2\pi}.$$

This completes the proof of the lemma. □

The next lemma compares one of our measures of conditional dependence with partial R^2 in the case of normal random variables.

LEMMA 17.2. *Let $(Y, \mathbf{X}, \mathbf{Z})$ be jointly normal, with $Y \sim N(0, \tau^2)$ for some $\tau > 0$. Let $Q(Y, \mathbf{Z}|\mathbf{X})$ be defined as in the statement of Lemma 13.1. Let $\alpha^2 := \operatorname{Var}(Y|\mathbf{X}, \mathbf{Z})$ and $\beta^2 := \operatorname{Var}(Y|\mathbf{X})$, and assume that these numbers are nonzero. Let $R^2_{Y,\mathbf{Z}|\mathbf{X}}$ be the partial R^2 of Y and $\mathbf{Z}$ given $\mathbf{X}$. There are positive universal constants C_1 and C_2 such that*

$$\frac{C_1 \beta^2 e^{-\beta^2/\alpha^2}}{\alpha\tau} R^2_{Y,\mathbf{Z}|\mathbf{X}} \le Q(Y, \mathbf{Z}|\mathbf{X}) \le \frac{C_2 \beta^2}{\alpha^2} R^2_{Y,\mathbf{Z}|\mathbf{X}}.$$

The same bounds hold if we replace $Q(Y, \mathbf{Z}|\mathbf{X})$ by $Q(Y, \mathbf{Z})$ (defined in equation (11.1)) and $R^2_{Y,\mathbf{Z}|\mathbf{X}}$ by $R^2_{Y,\mathbf{Z}}$ (the usual R^2 between Y and Z) on both sides, and define β^2 as $\operatorname{Var}(Y)$.

PROOF. Let $\mathbf{W} := (\mathbf{X}, \mathbf{Z})$, and for each $t \in \mathbb{R}$, let

$$Y_t := \mathbb{P}(Y \geq t|\mathbf{W}).$$

Let $U := \mathbb{E}(Y|\mathbf{W})$ and $V := \mathbb{E}(Y|\mathbf{X})$. Given $\mathbf{W}$, Y is normal with mean U and variance α^2. Thus,

$$\begin{aligned} Y_t &= \mathbb{P}\big((Y-U)/\alpha \geq (t-U)/\alpha|\mathbf{W}\big) \\ &= 1 - \Phi\big((t-U)/\alpha\big) = \Phi\big((U-t)/\alpha\big). \end{aligned} \tag{17.1}$$

Recall that

$$Q(Y, \mathbf{Z}|\mathbf{X}) = \int \mathbb{E}\big(\mathrm{Var}(Y_t|\mathbf{X})\big)\, d\mu(t),$$

where μ is the $N(0, \tau^2)$ probability measure. So, by (17.1),

$$Q(Y, \mathbf{Z}|\mathbf{X}) = \int \mathbb{E}\big(\mathrm{Var}\big(\Phi\big((U-t)/\alpha\big)|\mathbf{X}\big)\big)\, d\mu(t). \tag{17.2}$$

Now note that $\mathbb{E}(U|\mathbf{X}) = V$. Next, note that

$$\mathbb{E}\big[(Y-U)(U-V)|\mathbf{X}\big] = \mathbb{E}\big[\mathbb{E}\big((Y-U)(U-V)|\mathbf{W}\big)|\mathbf{X}\big] = 0,$$

since $\mathbb{E}(Y|\mathbf{W}) = U$, and U and V are functions of $\mathbf{W}$. Thus,

$$\begin{aligned} \mathbb{E}\big[(Y-V)^2|\mathbf{X}\big] &= \mathbb{E}\big[(Y-U)^2|\mathbf{X}\big] + \mathbb{E}\big[(U-V)^2|\mathbf{X}\big] \\ &= \mathbb{E}\big[\mathbb{E}\big((Y-U)^2|\mathbf{W}\big)|\mathbf{X}\big] + \mathrm{Var}(U|\mathbf{X}) \\ &= \mathbb{E}\big[\mathrm{Var}(Y|\mathbf{W})|\mathbf{X}\big] + \mathrm{Var}(U|\mathbf{X}) \\ &= \alpha^2 + \mathrm{Var}(U|\mathbf{X}). \end{aligned}$$

But $\mathbb{E}[(Y-V)^2|\mathbf{X}] = \beta^2$. Thus, $\mathrm{Var}(U|\mathbf{X}) = \beta^2 - \alpha^2$. Therefore, given $\mathbf{X}$, U is normal with mean V and variance $\beta^2 - \alpha^2$. Let

$$b := \frac{\sqrt{\beta^2 - \alpha^2}}{\alpha}.$$

Then, by (17.2), we get

$$Q(Y, \mathbf{Z}|\mathbf{X}) = \int \mathbb{E}\big(\mathrm{Var}\big(\Phi\big(bZ + (V-t)/\alpha\big)|\mathbf{X}\big)\big)\, d\mu(t),$$

where Z is a standard normal random variable, independent of all else. By Lemma 17.1,

$$C_1 b^2 e^{-(V-t)^2/\alpha^2 - b^2} \leq \mathrm{Var}\big(\Phi\big(bZ + (V-t)/\alpha\big)|\mathbf{X}\big) \leq C_2 b^2.$$

Plugging these bounds into the previous display, we get

$$C_1 b^2 e^{-b^2} \int \mathbb{E}\big(e^{-(V-t)^2/\alpha^2}\big)\, d\mu(t) \leq Q(Y, \mathbf{Z}|\mathbf{X}) \leq C_2 b^2. \tag{17.3}$$

Now note that

$$\int \mathbb{E}\big(e^{-(V-t)^2/\alpha^2}\big)\, d\mu(t) = \mathbb{E}\big(e^{-(V-\xi)^2/\alpha^2}\big),$$

where $\xi \sim N(0, \tau^2)$ and is independent of V. But $V - \xi \sim N(0, 2\tau^2 - \beta^2)$, since

$$\mathrm{Var}(V) = \mathrm{Var}(Y) - \mathbb{E}\big(\mathrm{Var}(Y|\mathbf{X})\big) = \tau^2 - \beta^2.$$

Thus, a simple computation gives

$$\mathbb{E}\big(e^{-(V-\xi)^2/\alpha^2}\big) = \left(1 + \frac{2(2\tau^2 - \beta^2)}{\alpha^2}\right)^{-1/2} = \frac{\alpha}{\sqrt{\alpha^2 + 2(2\tau^2 - \beta^2)}} \geq \frac{\alpha}{2\tau},$$

where the last inequality holds because $\alpha \leq \beta$. Plugging this lower bound into (17.3) and using $b^2 \leq \beta^2/\alpha^2$, we get

$$\frac{C_1 \alpha e^{-\beta^2/\alpha^2} b^2}{2\tau} \leq Q(Y, \mathbf{Z}|\mathbf{X}) \leq C_2 b^2.$$

Finally, observe that

$$b^2 = \frac{\beta^2}{\alpha^2}\frac{\beta^2 - \alpha^2}{\beta^2} = \frac{\beta^2}{\alpha^2} R^2_{Y,\mathbf{Z}|\mathbf{X}}$$

and substitute this in the previous display. This completes the proof of the first assertion of the lemma. The second assertion follows similarly, by retracing the steps in the proof and making suitable changes at the appropriate places. □

We are now ready to prove Theorem 7.1.

PROOF OF THEOREM 7.1. Let S be an insufficient subset of predictors. Then there is some $j \notin S$ such that $Q(Y, X_j|\mathbf{X}_S) \geq \delta$, because if $Q(S)$ is defined as in (6.1), then it is not hard to see that

$$Q(Y, X_j|\mathbf{X}_S) = Q\big(S \cup \{j\}\big) - Q(S).$$

But then, by Lemma 17.2,

$$\rho(S, j) = R^2_{Y,X_j|\mathbf{X}_S} \geq \frac{\alpha^2}{C_2\beta^2} Q(Y, X_j|\mathbf{X}_S) \geq \frac{\alpha^2\delta}{C_2\beta^2},$$

where $\alpha^2 = \mathrm{Var}(Y|\mathbf{X}_{S\cup\{j\}})$ and $\beta^2 = \mathrm{Var}(Y|\mathbf{X}_S)$. Since $\alpha^2 \geq \mathrm{Var}(Y|\mathbf{X}) = \sigma^2$ and $\beta^2 \leq \mathrm{Var}(Y) = \tau^2$, this shows that

$$\rho(S, j) \geq \frac{\sigma^2\delta}{C_2\tau^2}.$$

This proves that $\delta' \geq \sigma^2\delta/C_2\tau^2$. Conversely, for any insufficient set S, there is some $j \notin S$ such that $\rho(S, j) \geq \delta'$. So by Lemma 17.2,

$$Q(Y, X_j|\mathbf{X}_S) \geq \frac{C_1\beta^2 e^{-\beta^2/\alpha^2}\delta'}{\alpha\tau}.$$

Now, $\beta^2 \geq \mathrm{Var}(Y|\mathbf{X}) = \sigma^2$, $\alpha^2 \leq \mathrm{Var}(Y) = \tau^2$, and $\alpha^2 \leq \beta^2$. Thus,

$$Q(Y, X_j|\mathbf{X}_S) \geq \frac{C_1\sigma^2 e^{-1}\delta'}{\tau^2}.$$

Thus, $\delta \geq C_1 e^{-1}\sigma^2\delta'/\tau^2$. This completes the proof of the theorem. □

Acknowledgments. We are grateful to Mohsen Bayati, Persi Diaconis, Adityanand Guntuboyina, Susan Holmes, Bodhisattva Sen and Rob Tibshirani for helpful comments, and to Nima Hamidi, Norm Matloff and Balasubramanian Narasimhan for help with preparing the R package FOCI. We also thank the anonymous referees and the Associate Editor for various useful suggestions that helped improve the paper.

Funding. The second author was supported in part by NSF Grants DMS-1608249 and DMS-1855484.

REFERENCES

[1] ADRAGNI, K. P. and COOK, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4385–4405. MR2546393 https://doi.org/10.1098/rsta.2009.0110

[2] AMIT, Y. and GEMAN, D. (1997). Shape quantization and recognition with randomized trees. *Neural Comput.* **9** 1545–1588.

[3] AZADKIA, M., CHATTERJEE, S. and MATLOFF, N. (2020). FOCI: Feature Ordering by Conditional Independence. R Package. https://CRAN.R-project.org/package=FOCI.

[4] BATTITI, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* **5** 537–550.

[5] BERGSMA, W. P. (2004). Testing conditional independence for continuous random variables.

[6] BERRETT, T. B., WANG, Y., BARBER, R. F. and SAMWORTH, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 175–197. MR4060981

[7] BREIMAN, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37** 373–384. MR1365720 https://doi.org/10.2307/1269730

[8] BREIMAN, L. (1996). Bagging predictors. *Mach. Learn.* **26** 123–140.

[9] BREIMAN, L. (2018). Random forests. *Mach. Learn.* **45** 5–32.

[10] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees. Wadsworth Statistics/Probability Series*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

[11] CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 551–577. MR3798878 https://doi.org/10.1111/rssb.12265

[12] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644 https://doi.org/10.1214/009053606000001523

[13] CHATTERJEE, S. (2019). A new coefficient of correlation. *J. Amer. Statist. Assoc.* Available at https://arxiv.org/abs/1909.10140.

[14] CHEN, S. and DONOHO, D. (1994). Basis pursuit. In *Proc. of 28th Asilomar Conf. on Signals, Systems and Computers* **1** 41–44. IEEE, Los Alamitos.

[15] COCHRAN, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10** 417–451. MR0067428 https://doi.org/10.2307/3001616

[16] COOK, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.* **22** 1–26. MR2408655 https://doi.org/10.1214/088342306000000682

[17] DETTE, H., SIBURG, K. F. and STOIMENOV, P. A. (2013). A copula-based non-parametric measure of regression dependence. *Scand. J. Stat.* **40** 21–41. MR3024030 https://doi.org/10.1111/j.1467-9469.2011.00767.x

[18] DORAN, G., MUANDET, K., ZHANG, K. and SCHÖLKOPF, B. (2014). A permutation-based kernel conditional independence test. *Uncertainty Artif. Intell.* **30** 132–141.

[19] DUA, D. and GRAFF, C. (2019). *UCI Machine Learning Repository* [*http://archive.ics.uci.edu/ml*]. Univ. California, School of Information and Computer Science, Irvine, CA.

[20] DURRETT, R. (2010). *Probability: Theory and Examples*, 4th ed. *Cambridge Series in Statistical and Probabilistic Mathematics* **31**. Cambridge Univ. Press, Cambridge. MR2722836 https://doi.org/10.1017/CBO9780511779398

[21] DVORETZKY, A., KIEFER, J. and WOLFOWITZ, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Stat.* **27** 642–669. MR0083864 https://doi.org/10.1214/aoms/1177728174

[22] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166 https://doi.org/10.1214/009053604000000067

[23] FAN, J., FENG, Y. and XIA, L. (2020). A projection-based conditional dependence measure with applications to high-dimensional undirected graphical models. *J. Econometrics* **218** 119–139. MR4111748 https://doi.org/10.1016/j.jeconom.2019.12.016

[24] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

[25] FREUND, Y. and SCHAPIRE, R. (1996). Experiments with a new boosting algorithm. In *Machine Learning*: *Proceedings of the Thirteenth International Conference*. pp. 148–156.

[26] FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–141. MR1091842 https://doi.org/10.1214/aos/1176347963

[27] FRIEDMAN, J. H., BENTLEY, J. L. and FINKEL, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Software* **3** 209–226.

[28] FUKUMIZU, K., GRETTON, A., SUN, X. and SCHÖLKOPF, B. (2008). Kernel measures of conditional dependence. *Adv. Neural Inf. Process. Syst.* **20** 489–496.

[29] GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

[30] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*. *Springer Series in Statistics*. Springer, New York. MR1851606 https://doi.org/10.1007/978-0-387-21606-5

[31] HO, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **20** 832–844.

[32] HUANG, T.-M. (2010). Testing conditional independence using maximal nonlinear conditional correlation. *Ann. Statist.* **38** 2047–2091. MR2676883 https://doi.org/10.1214/09-AOS770

[33] JOE, H. (1989). Relative entropy measures of multivariate dependence. *J. Amer. Statist. Assoc.* **84** 157–164. MR0999674

[34] JOSSE, J. and HOLMES, S. (2016). Measuring multivariate association and beyond. *Stat. Surv.* **10** 132–167. MR3573303 https://doi.org/10.1214/16-SS116

[35] KNUTH, D. E. (1998). *The Art of Computer Programming*. *Vol.* 3: *Sorting and Searching*. Addison-Wesley, Reading, MA. MR3077154

[36] LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117

[37] LINTON, O. and GOZALO, P. (1996). Conditional independence restrictions: Testing and estimation. Cowles Foundation Discussion Paper, no. 1140.

[38] MANTEL, N. and HAENSZEL, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22** 719–748.

[39] MASSART, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* **18** 1269–1283. MR1062069

[40] MCDIARMID, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, 1989 (*Norwich*, 1989). *London Mathematical Society Lecture Note Series* **141** 148–188. Cambridge Univ. Press, Cambridge. MR1036755

[41] MILLER, A. J. (1990). *Subset Selection in Regression*. *Monographs on Statistics and Applied Probability* **40**. CRC Press, London. MR1072361 https://doi.org/10.1007/978-1-4899-2939-6

[42] PATRA, R. K., SEN, B. and SZÉKELY, G. J. (2016). On a nonparametric notion of residual and its applications. *Statist. Probab. Lett.* **109** 208–213. MR3434980 https://doi.org/10.1016/j.spl.2015.10.011

[43] PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*: *Networks of Plausible Inference*. *The Morgan Kaufmann Series in Representation and Reasoning*. Morgan Kaufmann, San Mateo, CA. MR0965765

[44] PÓCZOS, B. and SCHNEIDER, J. (2012). Nonparametric estimation of conditional information and divergences. *Proc. Mach. Learn. Res.* **22** 914–923.

[45] RAVIKUMAR, P., LAFFERTY, J., LIU, H. and WASSERMAN, L. (2009). Sparse additive models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 1009–1030. MR2750255 https://doi.org/10.1111/j.1467-9868.2009.00718.x

[46] RUDIN, W. (1987). *Real and Complex Analysis*, 3rd ed. McGraw-Hill, New York. MR0924157

[47] RUNGE, J. (2018). Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. *Proc. 21st Internat. Conf. Artif. Intell. Statist.* **84** 938–947.

[48] SEN, R., SURESH, A. T., SHANMUGAM, K., DIMAKIS, A. G. and SHAKKOTTAI, S. (2017). Model-powered conditional independence test. *Adv. Neural Inf. Process. Syst.* **31** 2955–2965.

[49] SETH, S. and PRÍNCIPE, J. C. (2012). Conditional association. *Neural Comput.* **24** 1882–1905.

[50] SHAH, R. D. and PETERS, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.* **48** 1514–1538. MR4124333 https://doi.org/10.1214/19-AOS1857
[51] SONG, K. (2009). Testing conditional independence via Rosenblatt transforms. *Ann. Statist.* **37** 4011–4045. MR2572451 https://doi.org/10.1214/09-AOS704
[52] STROBL, E. V., ZHANG, K. and VISWESWARAN, S. (2019). Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *J. Causal Inference* **7**.
[53] SU, L. and WHITE, H. (2007). A consistent characteristic function-based test for conditional independence. *J. Econometrics* **141** 807–834. MR2413488 https://doi.org/10.1016/j.jeconom.2006.11.006
[54] SU, L. and WHITE, H. (2008). A nonparametric Hellinger metric test for conditional independence. *Econometric Theory* **24** 829–864. MR2428851 https://doi.org/10.1017/S0266466608080341
[55] SU, L. and WHITE, H. (2014). Testing conditional independence via empirical likelihood. *J. Econometrics* **182** 27–44. MR3212759 https://doi.org/10.1016/j.jeconom.2014.04.006
[56] SZÉKELY, G. J. and RIZZO, M. L. (2014). Partial distance correlation with methods for dissimilarities. *Ann. Statist.* **42** 2382–2412. MR3269983 https://doi.org/10.1214/14-AOS1255
[57] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
[58] VERAVERBEKE, N., OMELKA, M. and GIJBELS, I. (2011). Estimation of a conditional copula and association measures. *Scand. J. Stat.* **38** 766–780. MR2859749 https://doi.org/10.1111/j.1467-9469.2011.00744.x
[59] VERGARA, J. R. and ESTÉVEZ, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **24** 175–186.
[60] WANG, X., PAN, W., HU, W., TIAN, Y. and ZHANG, H. (2015). Conditional distance correlation. *J. Amer. Statist. Assoc.* **110** 1726–1734. MR3449068 https://doi.org/10.1080/01621459.2014.993081
[61] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574 https://doi.org/10.1111/j.1467-9868.2005.00532.x
[62] YUKICH, J. E. (1998). *Probability Theory of Classical Euclidean Optimization Problems. Lecture Notes in Math.* **1675**. Springer, Berlin. MR1632875 https://doi.org/10.1007/BFb0093472
[63] ZHANG, K., PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2011). Kernel-based conditional independence test and application in causal discovery. *Uncertainty Artif. Intell.* **27** 804–813.
[64] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469 https://doi.org/10.1198/016214506000000735
[65] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x