Machine Learning
https://doi.org/10.1007/s10994-021-06030-6

Testing conditional independence in supervised learning algorithms

David S. Watson[1] · **Marvin N. Wright**[2,3]

Received: 30 January 2021 / Revised: 11 May 2021 / Accepted: 16 June 2021

Abstract
We propose the conditional predictive impact (CPI), a consistent and unbiased estimator of the association between one or several features and a given outcome, conditional on a reduced feature set. Building on the knockoff framework of Candès et al. (J R Stat Soc Ser B 80:551–577, 2018), we develop a novel testing procedure that works in conjunction with any valid knockoff sampler, supervised learning algorithm, and loss function. The CPI can be efficiently computed for high-dimensional data without any sparsity constraints. We demonstrate convergence criteria for the CPI and develop statistical inference procedures for evaluating its magnitude, significance, and precision. These tests aid in feature and model selection, extending traditional frequentist and Bayesian techniques to general supervised learning tasks. The CPI may also be applied in causal discovery to identify underlying multivariate graph structures. We test our method using various algorithms, including linear regression, neural networks, random forests, and support vector machines. Empirical results show that the CPI compares favorably to alternative variable importance measures and other nonparametric tests of conditional independence on a diverse array of real and synthetic datasets. Simulations confirm that our inference procedures successfully control Type I error with competitive power in a range of settings. Our method has been implemented in an `R` package, `cpi`, which can be downloaded from https://github.com/dswatson/cpi.

Keywords Knockoffs · Machine learning · Conditional independence · Markov blanket · Variable importance

Editors: Annalisa Appice, Sergio Escalera, Jose A. Gamez, Heike Trautmann.

✉ David S. Watson
david.watson@ucl.ac.uk

[1] Department of Statistical Science, University College London, London, UK

[2] Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany

[3] Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany

1 Introduction

Variable importance (VI) is a major topic in statistics and machine learning. It is the basis of most if not all feature selection methods, which analysts use to identify key drivers of variation in an outcome of interest and/or create more parsimonious models (Guyon & Elisseeff, 2003; Kuhn & Johnson, 2019; Meinshausen & Bühlmann, 2010). Many importance measures have been proposed in recent years, either for specific algorithms or more general applications. Several different notions of VI—some overlapping, some inconsistent—have emerged from this literature. We examine these in greater detail in Sect. 2.1.

One fundamental difference between various importance measures is whether they test the marginal or conditional independence of features. To evaluate response variable Y's marginal dependence on predictor X_j, we test against the following hypothesis:

$$H_0^m \ : \ X_j \perp Y, \boldsymbol{X}_{-j},$$

where $\boldsymbol{X}_{-j}$ represents a set of covariates.[1] A measure of conditional dependence, on the other hand, tests against a different null hypothesis:

$$H_0^c \ : \ X_j \perp Y | \boldsymbol{X}_{-j}.$$

Note that X_j's marginal VI may be high due to its association with either Y or $\boldsymbol{X}_{-j}$. This is why measures of marginal importance tend to favor correlated predictors. Often, however, our goal is to determine whether X_j adds any *new* information—in other words, whether Y is dependent on X_j even after conditioning on $\boldsymbol{X}_{-j}$. This becomes especially important when the assumption of feature independence is violated.

Tests of conditional independence (CI) are common in the causal modelling literature. For instance, the popular PC algorithm (Spirtes et al., 2000), which converges on a set of directed acyclic graphs (DAGs) consistent with some observational data, relies on the results of CI tests to recursively remove the edges between nodes. Common parametric examples include the partial correlation test for continuous variables or the χ^2 test for categorical data. A growing body of literature in recent years has examined nonparametric alternatives to these options. We provide an overview of several such proposals in Sect. 2.2.

In this paper, we introduce a new CI test to measure VI. The conditional predictive impact (CPI) quantifies the contribution of one or several features to a given algorithm's predictive performance, conditional on a complementary feature subset. Our work relies on so-called "knockoff'" variables (formally defined in Sect. 2.3) to provide negative controls for feature testing. Because knockoffs are, by construction, exchangeable with their observed counterparts and conditionally independent of the response, they enable a paired testing approach without any model refitting. Unlike the original knockoff filter, however, our methods are not limited to certain types of datasets or algorithms.

The CPI is extremely modular and general. It can be used with any combination of knockoff sampler, supervised learner, and loss function. It can be efficiently computed in high dimensions without sparsity constraints. We demonstrate that the CPI is an unbiased estimator, provably consistent under minimal assumptions. We develop statistical inference procedures for evaluating its magnitude, precision, and significance. Finally, we demonstrate the measure's utility on a variety of real and simulated datasets.

[1] We denote variables using uppercase italicized letters, e.g. X; matrices using uppercase bold letters, e.g. $\boldsymbol{X}$; scalars using lowercase italicized letters, e.g. x; and row-vectors using lowercase bold letters, e.g. $\boldsymbol{x}$.

The remainder of this paper is structured as follows. We review related work in Sect. 2. We present theoretical results in Sect. 3, where we also outline an efficient algorithm for estimating the CPI, along with corresponding p-values and confidence intervals. We test our procedure on real and simulated data in Sect. 4, comparing its performance with popular alternatives under a variety of regression and classification settings. Following a discussion in Sect. 5, we conclude in Sect. 6.

2 Related work

In this section, we survey the relevant literature on VI estimation, CI tests, and the knockoff filter.

2.1 Variable importance measures

The notion of VI may feel fairly intuitive at first, but closer inspection reveals a number of underlying ambiguities. One important dichotomy is that between global and local measures, which respectively quantify the impact of features on all or particular predictions. This distinction has become especially important with the recent emergence of interpretable machine learning techniques designed to explain individual outputs of black box models (e.g., Lundberg & Lee, 2017; Ribeiro et al., 2016; Wachter et al., 2018). In what follows, we restrict our focus to global importance measures.

Another important dichotomy is that between model-specific and model-agnostic approaches. For instance, a number of methods have been proposed for estimating importance in linear regression (Barber & Candès, 2015; Grömping, 2007; Lindeman et al., 1980), random forests (Breiman, 2001; Kursa & Rudnicki, 2010; Strobl et al., 2008), and neural networks (Bach et al., 2015; Gevrey et al., 2003; Shrikumar et al., 2017). These measures have the luxury of leveraging an algorithm's underlying assumptions and internal architecture for more precise and efficient VI estimation.

Other, more general techniques have also been developed. Van der Laan (2006) derives efficient influence curves and inference procedures for a variety of VI measures. Hubbard et al. (2018) build on this work, proposing a data-adaptive method for estimating the causal influence of variables within the targeted maximum likelihood framework (van der Laan & Rose, 2018). Williamson et al. (2021) describe an ANOVA-style decomposition of a regressor's R^2 into feature-wise contributions. Feng et al. (2018) design a neural network to efficiently compute this decomposition using multi-task learning. Fisher et al. (2019) propose a number of "reliance" statistics, calculated by integrating a loss function over the empirical distribution of covariates while holding a given feature vector constant.

Perhaps the most important distinction between various competing notions of VI is the aforementioned split between marginal and conditional measures. The topic has received considerable attention in the random forest literature, where Breiman's popular permutation technique (2001) has been criticized for failing to properly account for correlations between features (Gregorutti et al., 2015; Nicodemus et al., 2010). Conditional alternatives have been developed (Mentch & Hooker, 2016; Strobl et al., 2008), but we do not consider them here, as they are specific to tree ensembles.

Our proposed measure resembles what Fisher et al. (2019) call "algorithm reliance" (AR). The authors do not have much to say about AR in their paper, the majority of which is instead devoted to two related statistics they term "model reliance" (MR) and "model

class reliance" (MCR). These measure the marginal importance of a feature subset in particular models or groups of models, respectively. Only AR measures the importance of the subset conditional on remaining covariates for a given supervised learner, which is our focus here. Fisher et al. (2019) derive probabilistic bounds for MR and MCR, but not AR. They do not develop hypothesis testing procedures for any of their reliance statistics.

2.2 Conditional independence tests

CI tests are the cornerstone of constraint-based and hybrid methods for causal graph inference and Bayesian network learning (Koller & Friedman, 2009; Korb & Nicholson, 2009; Scutari & Denis, 2014). Assuming the causal Markov condition and faithfulness—which together state (roughly) that statistical independence implies graphical independence and vice versa—a number of algorithms have been developed that use CI tests to discover an equivalence class of DAGs consistent with a set of observational data (Maathuis et al., 2009; Spirtes et al., 2000; Verma & Pearl, 1991).

Shah and Peters (2020) have shown that there exists no uniformly valid CI test. Parametric assumptions are typically deployed to restrict the range of alternative hypotheses, which is default behavior for most causal discovery software (e.g., Kalisch et al., 2012; Scutari, 2010). However, more flexible methods have been introduced. Much of this literature relies on techniques that embed the data in a reproducing kernel Hilbert space (RKHS). For instance, Fukumizu et al. (2008) use a normalized cross-covariance operator to test the association between features in the RKHS. A null distribution is approximated via permutation. Doran et al. (2014) build on Fukumizu et al.'s work with a modified permutation scheme intended to capture the effects of CI. Zhang et al. (2011) derive a test statistic from the traces of kernel matrices, using a gamma null distribution to compute statistical significance.

Another general family of methods for CI testing is rooted in information theory. Fleuret (2004) proposes a fast binary variable selection procedure using conditional mutual information. Similar techniques have been used to infer directionality in networks (Vejmelka & Paluš, 2008), cluster features together for dimensionality reduction (Martínez Sotoca & Pla, 2010), and reason about the generalization properties of supervised learners (Steinke & Zakynthinou, 2020). These approaches typically rely either on discretization procedures to convert all inputs to categorical data, or strong parametric assumptions to handle continuous spaces.

Several authors have proposed alternative tests to avoid the inefficiencies of kernel methods and the binning often required by information theoretic algorithms. For instance, Strobl et al. (2018) employ a fast Fourier transform to reduce the complexity of matrix operations. Methods have been developed for estimating regularized, nonlinear partial correlations (Ramsey, 2014; Shah & Peters, 2020). Lei et al. (2018) and Rinaldo et al. (2019) study the leave-one-covariate-out (LOCO) procedure, in which an algorithm is trained on data with and without the variable of interest. The predictive performance of nested models is compared to evaluate the conditional importance of the dropped feature.

Our proposal is conceptually similar to LOCO, which can in principle be extended to feature subsets of arbitrary dimension. The method enjoys some desirable statistical properties when used in conjunction with sample splitting. For instance, Rinaldo et al. (2019) derive convergence rates for LOCO parameters, while Lei et al. (2018) prove finite sample error control using conformal inference. However, retraining an algorithm for each CI test is potentially infeasible, especially for complex learners and/or large

datasets. With knockoffs, we can directly import LOCO's statistical guarantees without any model refitting.

2.3 The Knockoff framework

Our work builds on the knockoff procedure originally conceived by Barber and Candès (2015) and later refined by Candès et al. (2018). Central to this approach is the notion of a knockoff variable. Given an $n \times p$ input matrix $\boldsymbol{X}$, we define a knockoff matrix of equal dimensionality $\tilde{\boldsymbol{X}}$ as any matrix that meets the following two criteria:

(a) *Pairwise exchangeability*. For any proper subset $S \subset [p] = (1,\ldots,p)$

$$(\boldsymbol{X}, \tilde{\boldsymbol{X}})_{swap(S)} =^{d} (\boldsymbol{X}, \tilde{\boldsymbol{X}}),$$

where $=^{d}$ represents equality in distribution and the swapping operation is defined below.

(b) *Conditional independence*. $\tilde{\boldsymbol{X}} \perp Y | \boldsymbol{X}$.

A swap is obtained by switching the entries X_j and $\tilde{X}_j$ for each $j \in S$. For example, with $p = 3$ and $S = \{1,\ 3\}$:

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)_{swap(S)} =^{d} (\tilde{X}_1,\ X_2, \tilde{X}_3, X_1, \tilde{X}_2, X_3).$$

Knockoffs provide negative controls for conditional independence testing. The intuition behind the method is that if X_j does not significantly outperform $\tilde{\boldsymbol{X}}_j$ by some relevant importance measure, then the original feature may be safely removed from the final model.

Practical implementation requires both a method for generating knockoffs and a decision procedure for variable selection. The subject has quickly become a busy one in statistics and machine learning, with most authors focusing on the former task. In this paper, we instead tackle the latter, developing a general framework for testing conditional variable importance.

Constructing nontrivial knockoffs is a considerable challenge. Numerous methods have been proposed, including but not limited to: second-order Gaussian knockoffs (Candès et al., 2018); conditional permutation sampling (Berrett et al., 2020); hidden Markov models (Sesia et al., 2019); generative deep neural networks (Romano et al., 2020); Metropolis–Hastings sampling (Bates et al., 2020); conditional density estimation (Tansey et al., 2021); and normalizing flows (Hansen et al., 2021). A complete review of these proposals is beyond the scope of this chapter. Bates et al. (2020) demonstrate that no efficient knockoff sampler exists for arbitrary probability distributions, suggesting that algorithms will have to make some assumptions about the data generating process to strike a reasonable balance between sensitivity and specificity.

The original knockoff papers introduce a novel algorithm for controlling the false discovery rate (FDR) in variable selection problems. The goal is to find the minimal subset $\mathcal{S} \subset [p]$ such that, conditional on $\{X_j\}_{j \in \mathcal{S}}$, Y is independent of all other variables. Call this the *Markov blanket* of Y (Pearl, 1988). Null features form a complementary set $\mathcal{R} = [p] \backslash \mathcal{S}$ such that $k \in \mathcal{R}$ if and only if $X_k \perp Y | \{X_j\}_{j \in \mathcal{S}}$. The FDR is given by the expected proportion of false positives among all declared positives:

$$\text{FDR} = \mathbb{E}\left[\frac{\left|\widehat{\mathcal{S}} \cap \mathcal{R}\right|}{\left|\widehat{\mathcal{S}} \vee 1\right|}\right],$$

where $\widehat{\mathcal{S}}$ is the output of the decision procedure and the "$\vee 1$" in the denominator enforces the convention that FDR $= 0$ when $\left|\widehat{\mathcal{S}}\right| = 0$.

Barber and Candès (2015) demonstrate a method for guaranteed finite sample FDR control when (i) statistics for null variables are symmetric about zero and (ii) large positive statistics indicate strong evidence against the null. We will henceforth refer to this method as the adaptive thresholding test (ATT). Unlike other common techniques for controlling the FDR (e.g., Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; Storey, 2002), the ATT does not require p-values as an intermediate step. Candès et al. (2018) argue that this is a benefit in high-dimensional settings, where p-value calculations can be unreliable.

Acknowledging that p-values may still be desired in some applications, however, Candès et al. also propose the conditional randomization test (CRT), which provides one-sided Monte Carlo p-values by repeatedly sampling from the knockoff distribution. Experiments indicate that the CRT is slightly more powerful than the ATT, but the authors caution that the former is computationally intensive and do not recommend it for large datasets. That has not stopped other groups from advancing formally similar proposals (e.g., Berrett et al., 2020; Tansey et al., 2021).

We highlight several important shortcomings of the ATT: (1) Not all algorithms provide feature scoring statistics. (2) The ATT requires a large number of variables to reliably detect true positives. (3) Because the ATT does not perform individual hypothesis tests, it cannot provide confidence or credible intervals for particular variables. In what follows, we present alternative inference procedures for conditional independence testing designed to address all three issues.

3 Conditional predictive impact

The basic intuition behind our approach is that important features should be informative—that is, their inclusion should improve the predictive performance of an appropriate algorithm as measured by some preselected loss function. Moreover, the significance of improvement should be quantifiable so that error rates can be controlled at user-specified levels.

Consider an $n \times p$ feature matrix $\boldsymbol{X} \in \mathcal{X}$ and corresponding $n \times 1$ response variable $Y \in \mathcal{Y}$, which combine to form the dataset $\boldsymbol{Z} = (\boldsymbol{X}, Y) \in \mathcal{Z}$.[2] Each observation $z_i = (\boldsymbol{x}_i, y_i)$ is an i.i.d. sample from a fixed but unknown joint probability distribution, $\mathbb{P}(\boldsymbol{Z}) = \mathbb{P}(\boldsymbol{X}, Y)$. Let $\boldsymbol{X}^S \subseteq (X_1, \ldots, X_p)$ denote some subset of features whose predictive impact we intend to quantify, conditional on the (possibly empty) set of remaining covariates $\boldsymbol{X}^R = \boldsymbol{X} \backslash \boldsymbol{X}^S$. Data can now be expressed as a triple, $\boldsymbol{Z} = (\boldsymbol{X}^S, \boldsymbol{X}^R, Y)$. We remove the predictive information in $\boldsymbol{X}^S$ while preserving the covariance structure of the predictors by replacing the submatrix with the corresponding knockoff variables, $\tilde{\boldsymbol{X}}^S$, rendering a new dataset, $\tilde{\boldsymbol{Z}} = (\tilde{\boldsymbol{X}}^S, \boldsymbol{X}^R, Y)$.

Define a function $f \in \mathcal{F}, \mathcal{F} : \mathcal{X} \to \mathcal{Y}$ as a mapping from features to outcomes. We evaluate a model's performance using some real-valued, non-negative loss function

[2] We leave the more general case of multidimensional response variables to future work.

$L : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$. Define the risk of f with respect to $\boldsymbol{Z}$ as its expected loss over the joint probability distribution $\mathbb{P}(\boldsymbol{Z})$, $R(f, \boldsymbol{Z}) = \mathbb{E}\big[L(f, \boldsymbol{Z})\big]$. Our strategy is to replace the conditional null hypothesis defined in Sect. 1 with the following:

$$H_0 : R(f, \boldsymbol{Z}) \geq R\big(f, \tilde{\boldsymbol{Z}}\big).$$

In other words, we test whether the model performs better using the original or the knockoff data. We note that this is potentially weaker than H_0^C, as many popular loss functions restrict attention to just the first moment. However, we argue that if such a loss function is appropriate, then a conditional predictive test is a better choice than a conditional independence test. For instance, if the goal is simply to minimize out-of-sample L_2 loss, then we have no need for features that do not improve mean square error, even if they encode information about higher moments (e.g., predictive variance or skewness). Alternatively, if confidence intervals are important for a given task, then the loss function should reflect that, as likelihood-based measures do. In general, recovering the full Markov blanket of Y may be unnecessary.

3.1 Consistency and convergence

The CPI of submatrix $\boldsymbol{X}^S$ measures the extent to which the feature subset improves predictions made using model f. Assume that the loss function L can be evaluated for each sample i.[3] We define the following random variable:

$$\Delta_i = L\big(f, \tilde{z}_i\big) - L\big(f, z_i\big). \tag{1}$$

This vector represents the difference in sample-wise loss between predictions made using knockoff data and original data. We define the CPI by taking its expectation:

$$\mathrm{CPI}\big(\boldsymbol{X}^S\big) = \mathbb{E}[\Delta]. \tag{2}$$

Note that the CPI is always a function of some feature subset $\boldsymbol{X}^S$. We suppress the dependency for notational convenience moving forward.

To consistently estimate this statistic, it is necessary and sufficient to show that we can consistently estimate the risk of model f. The population parameter $R(f, \boldsymbol{Z})$ is estimated using the empirical risk formula:

$$R_{\mathrm{emp}}(f, \boldsymbol{Z}) = \frac{1}{m} \sum_{i=1}^{m} L\big(f, z_i\big). \tag{3}$$

The goal in estimating risk is to evaluate how well the model generalizes beyond its training data, so the m samples in Eq. 3 constitute a test set drawn independently from $\boldsymbol{Z}$, distinct from the n samples used to fit f. In practice, this is typically achieved by some resampling procedure like cross-validation or bootstrapping. In what follows, we presume that unit-level loss $L(f, z_i)$ is always an out-of-sample evaluation, such that f was trained on data excluding z_i.

[3] For loss functions that do not have this property, such as the area under the receiver operating characteristic curve, the following arguments can easily be modified to apply to each fold in a cross-validation.

The empirical risk minimization (ERM) principle is a simple decision procedure in which we select the function f that minimizes empirical risk in some function space $\mathcal{F}$. A celebrated result of Vapnik and Chervonenkis (1971), independently derived by Sauer (1972) and Shelah (1972), is that the ERM principle is consistent with respect to $\mathcal{F}$ if and only if the function space is of finite VC dimension. Thus, for any algorithm that meets this minimal criterion, the empirical risk $R_{\text{emp}}(f, \boldsymbol{Z})$ converges uniformly in probability to $R(f, \boldsymbol{Z})$ as $n \to \infty$, which means the estimate:

$$\begin{aligned} \widehat{\text{CPI}} &= \frac{1}{n} \sum_{i}^{n} L\big(f, \tilde{z}_i\big) - L\big(f, z_i\big) \\ &= R_{\text{emp}}\big(f, \tilde{\boldsymbol{Z}}\big) - R_{\text{emp}}(f, \boldsymbol{Z}) \\ &= \frac{1}{n} \sum_{i}^{n} \Delta_i \end{aligned} \tag{4}$$

is likewise guaranteed to converge. Because the CPI inherits the convergence properties of the learner f, it imposes no additional smoothness, sparsity, parametric, or dimensionality constraints upon the data.

Though finite complexity thresholds have been derived for many algorithms—e.g., projective planes, decision trees, boosting machines, and neural networks (Shalev-Shwartz & Ben-David, 2014)—it is worth noting that some popular supervised learners do in fact have infinite VC dimension. This is the case, for instance, with methods that rely on the radial basis function kernel, widely used in support vector machines and Gaussian process regression. The learning theoretic properties of these algorithms are better described with other measures such as the Rademacher complexity and PAC-Bayes bounds (Guedj, 2019). However, as we show in Sect. 4, the CPI shows good convergence properties even when used with learners of infinite VC dimension.

Inference procedures for the CPI can be designed using any paired difference test. Familiar frequentist examples include the t-test and the Fisher exact test, which we use for large- and small-sample settings, respectively. Bayesian analogues can easily be implemented as well. Rouder et al. (2009) advocate an analytic strategy for calculating Bayes factors for t-tests. Wetzels et al. (2009) and Kruschke (2013) propose more general methods based on Markov chain Monte Carlo sampling, although they differ in their proposed priors and decision procedures. Care should be taken when selecting a prior distribution in the Bayesian setting, especially with small sample sizes. Tools for Bayesian inference are implemented in the accompanying `cpi` package; however, for brevity's sake, we restrict the following sections to frequentist methods.

3.2 Large sample inference: paired *t*-tests

By the central limit theorem, empirical risk estimates will tend to be normally distributed around the true population parameter value. Thus we use paired, one-sided t-tests to evaluate statistical significance when samples are sufficiently large ($n \geq 30$ or thereabouts).

The variable Δ has mean $\widehat{\text{CPI}}$ and standard error $\text{SE} = s/\sqrt{n}$, where s denotes the sample standard deviation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\Delta_i - \widehat{\text{CPI}})^2}.$$

The t-score for $\widehat{\text{CPI}}$ is given by $t = \widehat{\text{CPI}}/\text{SE}$, and we may compute p-values by comparing this statistic to the most tolerant distribution consistent with $H_0 : R(f, \boldsymbol{Z}) \geq R(f, \boldsymbol{Z})$, namely t_{n-1}. To control Type I error at level α, we reject H_0 for all t greater than or equal to the $(1-\alpha)$ quantile of t_{n-1}. This procedure can easily be modified to adjust for multiple testing.

We can relax the assumption of homoskedasticity if reliable estimates of predictive precision are available. Construct a $2n \times (n+1)$ feature matrix $\boldsymbol{X}$ with columns for each unit $i = \{1,\ldots,n\}$, as well as an indicator variable for data type D (original vs. knockoff). Let $\boldsymbol{W}$ be a $2n \times 2n$ diagonal matrix such that $\boldsymbol{W}_{ii}$ denotes the weight assigned to the i^{th} prediction. For instance, in a regression setting, this could be the inverse of the expected residual variance for i. Then solve a weighted least squares regression, with the response variable $\boldsymbol{y}$ equal to the observed loss for each unit-data type combination:

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}$$

The t-statistic and p-value associated with coefficient $\hat{\gamma}_D$ can then be used to test the CPI of the substituted variable(s) under a heteroskedastic error model.

Confidence intervals around $\widehat{\text{CPI}}$ may be constructed in the typical manner. The lower bound is set by subtracting from our point estimate the product of SE and $F_{n-1}^{-1}(1-\alpha)$, where $F_{n-1}(\bullet)$ denotes the CDF of t_{n-1}. Using this formula, we obtain a 95% confidence interval for $\widehat{\text{CPI}}$ by calculating $\left[\widehat{\text{CPI}} - \text{SE} \times F_{n-1}^{-1}(0.95),\ \infty\right)$. As n grows large, this interval converges to the Wald-type interval, $\left[\widehat{\text{CPI}} - \text{SE} \times \Phi^{-1}(0.95),\ \infty\right)$, where Φ represents the standard normal CDF.

The t-testing framework also allows for analytic power calculations. Let t^* denote the critical value $t^* = F_{n-1}^{-1}(1-\alpha)$. Then Type II error is given by the formula $\beta = F_{n-1}(t^* - \delta)$, where δ represents the postulated effect size. Statistical power is just the complement $1-\beta$, and rearranging this equation with simple algebra allows us to determine the sample size required to detect a given effect at some fixed Type I error α.

3.3 Small sample inference: Fisher exact tests

The applicability of the central limit theorem is dubious when sample sizes are small. In such cases, exact p-values may be computed for a slightly modified null hypothesis using Fisher's method (1935). Rather than focusing on overall risk, this null hypothesis states that replacing $\boldsymbol{X}^S$ with the knockoff submatrix $\tilde{\boldsymbol{X}}^S$ has no impact on unit-level loss. More formally, we test against the following:

$$H_0^{\text{FEP}} : L\big(f, z_i\big) \geq L\big(f, \tilde{z}_i\big), \quad i = 1, \ldots, n.$$

Under this null hypothesis, which is sufficient but not necessary for the conditional predictive H_0, we may implement a permutation scheme in which the CPI is calculated for all possible assignments of data type D. Consider a $2n \times 3$ matrix with columns for unit index $U = \{1,1,\ldots,n,n\}$, data type $D \in \{0,1\}$, and loss L. We permute the rows of D subject to the constraint that every sample's loss is recorded under both original and knockoff predictions. For each possible vector D, compute the resulting CPI and compare the value of our observed statistics, $\widehat{\text{CPI}}$, to the complete distribution. Note that this paired setup

dramatically diminishes the possible assignment space from an unmanageable $\binom{2n}{n}$, corresponding to a Bernoulli trial design, to a more reasonable 2^n. The one-tailed Fisher exact p-value (FEP) is given by the formula:

$$\mathrm{FEP}\left(\widehat{\mathrm{CPI}}\right) = \frac{1}{2^n}\sum_{b=1}^{2^n} \mathbb{I}\left(\widetilde{\mathrm{CPI}}_b \geq \widehat{\mathrm{CPI}}\right),$$

where $\mathbb{I}(\bullet)$ represents the indicator function and $\widetilde{\mathrm{CPI}}_b$ is the CPI resulting from the bth permutation of D.

To construct a confidence interval for $\widehat{\mathrm{CPI}}$ at level $1-\alpha$, we use our empirical null distribution. Find the critical value CPI* such that FEP(CPI*)$=\alpha$. Then a $(1-\alpha)\times 100\%$ confidence interval for $\widehat{\mathrm{CPI}}$ is given by $[\widehat{\mathrm{CPI}} - \mathrm{CPI}^*, \infty)$. For n large, approximate calculations can be made by sampling from the set of 2^n permissible permutations. In this case, however, we add 1 to both the numerator and denominator to ensure unbiased inference (Phipson & Smyth, 2010).

3.4 Computational complexity

To summarize, we outline the proposed algorithm for testing the conditional importance of feature subsets for supervised learners in pseudocode below. This algorithm executes in $\mathcal{O}(ak+g+h)$ time and $\mathcal{O}(a+k+g+h)$ space. We take the complexity of the learner a and knockoff sampler g to be given. The empirical risk estimator k can be made more or less complex depending on the resampling procedure. The most efficient option for evaluating generalization error is the holdout method, in which a model is trained on a random subset of the available data and tested on the remainder. Unfortunately, this procedure can be unreliable with small sample sizes. Popular alternatives include the bootstrap and cross-validation. Both require considerable model refitting, which can be costly when a is complex.

The inference procedure h is quite efficient in the parametric case—on the order of $\mathcal{O}(n)$ for the t-test—but scales exponentially with the sample size when using the permutation-based approach. As noted above, the complexity of the Fisher test can be bounded by setting an upper limit on the number of permutations B used to approximate the empirical null distribution. The standard error of a p-value estimate made using such an approximation is $\sqrt{p^*(1-p^*)/B}$, where p^* represents the true p-value. This expression is maximized at $p^* = 0.5$, corresponding to a standard error of $1/(2\sqrt{B})$. Thus, to guarantee a standard error of at most 0.001, it would suffice to use $B=250{,}000$ permutations, an eminently feasible computation on a modern laptop. Space complexity is dominated by the learner a and knockoff sampler g because most resampling procedures refit the same learner a fixed number of times and the inference procedures described above execute in $\mathcal{O}(1)$ space.

Algorithm 1: CPI

Inputs: Dataset $\boldsymbol{Z}$, submatrix $\boldsymbol{X}^S$, ERM learner a, risk functional R, knockoff sampler g, risk estimator k, inference procedure h

\\ Learn f via a by minimizing empirical risk on $\boldsymbol{Z}$

1. $f \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} R_{\text{emp}}(f, \boldsymbol{Z})$

\\ Sample knockoffs via g

2. $\widetilde{\mathbf{X}}^S \leftarrow g(\boldsymbol{X}^S)$

\\ Use risk estimator k to compute unit-level loss

3. $\widehat{\text{CPI}} \leftarrow n^{-1} \sum_{i=1}^{n} L(f, \tilde{\boldsymbol{z}}_i) - L(f, \boldsymbol{z}_i)$

\\ Apply inference procedure h to compute p-value (p) and confidence interval (ci)

4. $\{p, \text{ci}\} \leftarrow h(\widehat{\text{CPI}})$

Output: $\widehat{\text{CPI}}$, p, ci

4 Experiments

All experiments were conducted in the R statistical computing environment, version 3.6.2. Code for reproducing all results and figures can be found in the dedicated GitHub repository: https://github.com/dswatson/cpi_paper

4.1 Simulated data

We report results from a number of simulation studies. First, we analyze the statistical properties of the proposed tests under null and alternative hypotheses. We proceed to compare the sensitivity and specificity of the CPI to those of several alternative measures.

Data were simulated under four scenarios, corresponding to all combinations of independent vs. correlated predictors and linear vs. nonlinear outcomes. Because conditional importance is most relevant in the case of correlated predictors, results for the two scenarios with independent features are left to the Supplement. In the linear setting, ten variables were drawn from a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, with covariance matrix$\Sigma_{ij} = 0.5^{|i-j|}$. A continuous outcome Y was calculated as$Y = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$, where $\boldsymbol{\beta} = (0.0, 0.1, \ldots, 0.9)'$ and $\varepsilon \sim \mathcal{N}(0, 1)$. In the nonlinear scenario, we keep the same predictors but generate the response from a transformed matrix,$Y = \boldsymbol{X}^* \boldsymbol{\beta} + \varepsilon$, where

$$x_{ij}^* = \begin{cases} +1, & \text{if } \Phi^{-1}(0.25) \leq x_{ij} \leq \Phi^{-1}(0.75) \\ -1, & \text{else} \end{cases}$$

with the same $\boldsymbol{\beta}$ and ε as in the linear case. A similar simulation was performed for a classification outcome, where the response Y was drawn from a binomial distribution with probability $[1 + \exp(-\boldsymbol{X}\boldsymbol{\beta})]^{-1}$ and $[1 + \exp(-\boldsymbol{X}^*\boldsymbol{\beta})]^{-1}$ for the linear and nonlinear scenarios, respectively.

Knockoffs for all simulated data were generated using the second-order Gaussian technique described in (Candès et al., 2018) and implemented in the knockoff package, version 0.3.2 (Patterson & Sesia, 2018).

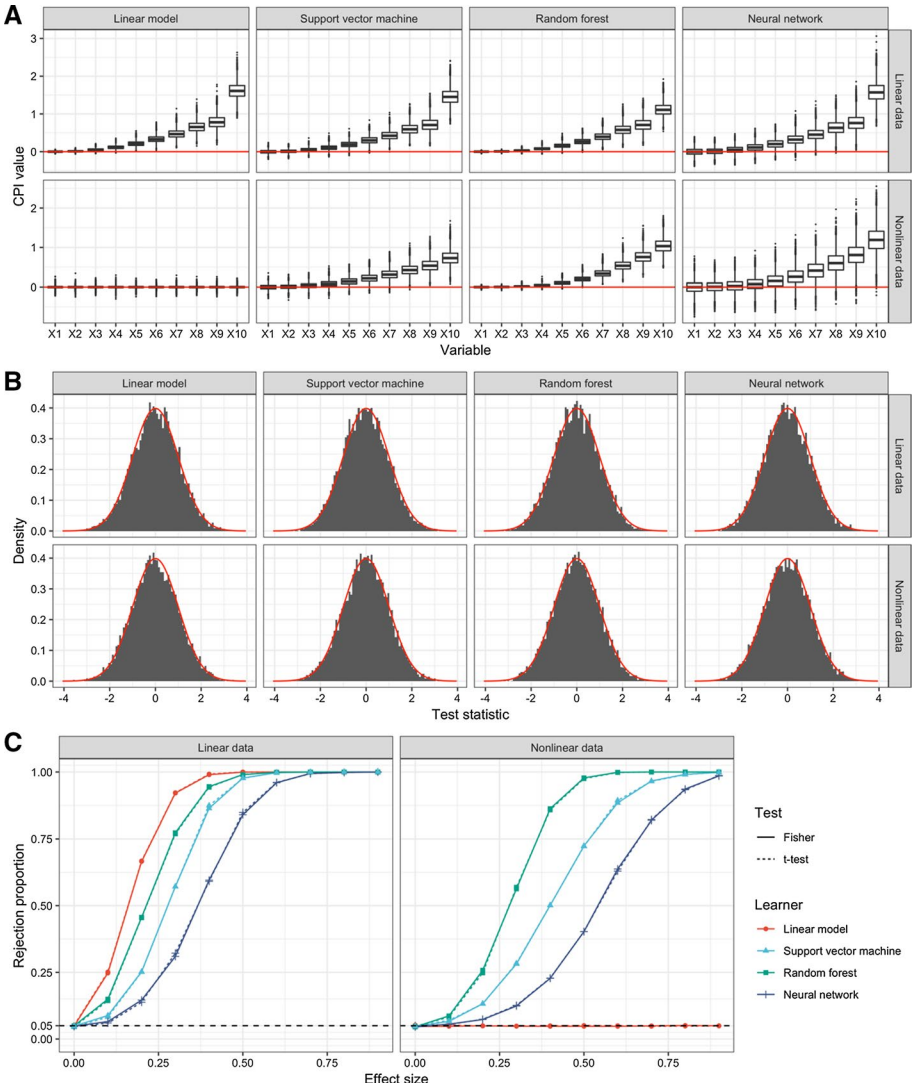

Fig. 1 Simulation results for continuous outcome with MSE loss and correlated predictors. **A**: Boxplots of simulated CPI values of variables $X_1, \ldots, X_{10}$ with increasing effect size. The red line indicates a CPI value of 0, corresponding to a completely uninformative predictor. **B**: Histograms of simulation replications of t-statistics of variables with effect size 0. The distribution of the expected t-statistic under the null hypothesis is shown in red. **C**: Proportion of rejected hypotheses at $\alpha = 0.05$ as a function of effect size. Results at effect size 0 correspond to the Type I error, at effect sizes > 0 to statistical power. The dashed line indicates the nominal level, $\alpha = 0.05$

4.2 Type I and Type II errors

We simulate 10^4 datasets with $n = 1000$ observations and compute the CPI using four different learning algorithms: linear/logistic regression (LM), random forest (RF), artificial neural network (ANN), and support vector machine (SVM). Risk was estimated using

holdout sampling with a train/test ratio of 2:1. For regression models, we used mean square error (MSE) and mean absolute error (MAE) loss functions; for classification, we used cross entropy (CE) and mean misclassification error (MMCE). We computed p-values via the inference procedures described in Sect. 3, i.e. paired t-tests and Fisher exact tests. For Fisher tests we used 10^4 permutations.

Linear and logistic regressions were built using the functions `lm` and `glm`, respectively, from the `R` package stats (R Core Team, 2019). RFs were built using the `ranger` package (Wright & Ziegler, 2017), with 500 trees. ANNs were built with the `nnet` package (Venables & Ripley, 2002), with 20 hidden units and a weight decay of 0.1. SVMs were built with the `e1071` package (Meyer et al., 2018), using a Gaussian radial basis function (RBF) kernel and $\sigma = 1$. Unless stated otherwise, all parameters were left to their default values. Resampling was performed with the `mlr` package (Bischl et al., 2016).

Significance levels for all tests were fixed at $\alpha = 0.05$. For each simulation, we calculated the CPI values, Type I errors, Type II errors, empirical coverage, and t-statistics, where applicable. Results for MSE loss with $p = 10$ are shown in Fig. 1. Similar plots for MAE, CE, and MMCE loss functions are presented in Figs. S1–S10 of the Supplement. Coverage probabilities are shown in Tables S1–S8 of the Supplement.

For continuous outcomes, CPI controlled Type I error with all four learners and reached 100% power under all settings, with the exception of the LM on nonlinear data. No observable differences are detected between the MSE and MAE loss functions. We found similar results for categorical outcomes. The CPI controlled Type I error for the MMCE and CE loss functions with all four learners. The LM once again performed poorly on nonlinear data, as expected. The Fisher test had slightly increased power compared to the t-test. Statistical power was generally greater with CE loss than with MMCE loss.

4.3 Comparative performance

We use the same simulation setup to compare the CPI's performance to that of three other global, nonparametric, model-agnostic measures of CI, each of which relies on identical or stronger testing assumptions:

- **ANOVA**: Williamson et al. (2021)'s nonparametric ANOVA-inspired VI.
- **LOCO**: Lei et al. (2018)'s leave-one-covariate-out procedure.
- **GCM**: Shah and Peters (2020)'s generalized covariance measure.

Unfortunately, software for Hubbard et al. (2018)'s targeted maximum likelihood VI statistic was still under development at the time of testing, and beta versions generated errors. Candès et al. (2018)'s probabilistic knockoff procedure can be extended to nonparametric models, but requires an algorithm-specific VI measure, which not all learners provide. We consider this method separately in Sect. 4.1.3. Kernel methods do not work with arbitrary algorithms and were therefore excluded. Information theoretic measures typically require a discretization step that does not extend well to high dimensions, and were therefore also excluded. We restrict this section to the regression setting, as none of the other methods considered here are designed for classification problems.

Training and test sets are of equal size, with $n \in \{100, 500, 1000\}$. In each case, we fit LM, RF, ANN, and SVM regressions, as described previously. We estimate the VI

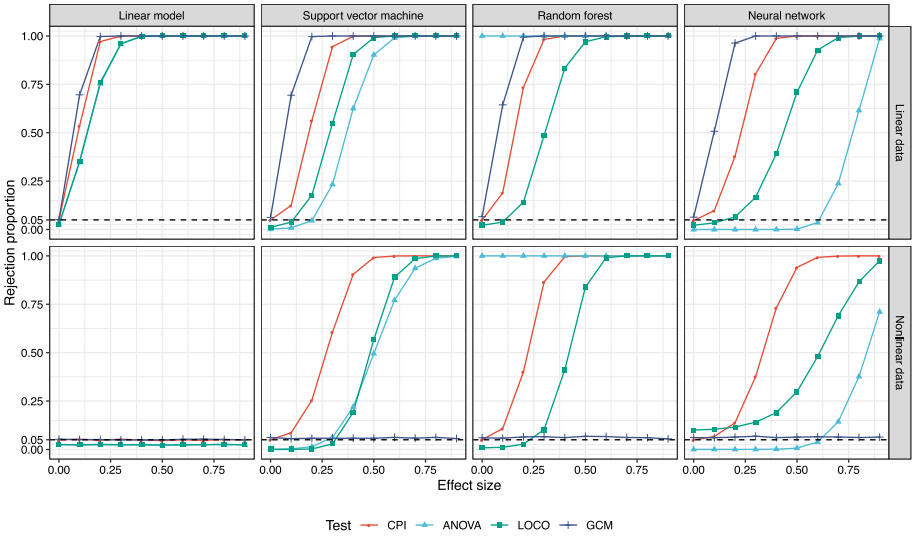

Fig. 2 Comparative performance of VI measures across different simulations and algorithms. Plots depict the proportion of rejected hypotheses at $\alpha = 0.05$ as a function of effect size. Results at effect size 0 correspond to Type I error, at effect sizes > 0 to statistical power. The dashed line indicates the nominal level, $\alpha = 0.05$. These results were computed using training and test samples of $n = 1000$ and $p = 10$. Similar results were obtained for sample sizes of $n = \{100, 500\}$ and $p = \{20, 50, 100\}$ (see Supplement, Sect. 2)

of all features on the test set for every model. This procedure was repeated 10^4 times. Results for $n = 1000$ and $p = 10$ are plotted in Fig. 2. Similar results for $n = \{100, 500\}$ and $p = \{20, 50, 100\}$ are included in the Supplement.

All methods have high Type II error rates when fitting an LM to nonlinear data, highlighting the dangers of model misspecification. GCM appears to dominate in the linear setting but struggles to detect VI in nonlinear simulations. LOCO is somewhat conservative, often falling short of the nominal Type I error rate under the null hypothesis. However, the method fails to control Type I error in the case of an ANN trained on nonlinear data. The nonparametric ANOVA generally performs poorly, especially with RF regressions, where we may observe Type I error rates of up to 100%.

The CPI outperforms all competitors with nonlinear data, and achieves greater power than ANOVA or LOCO in the linear case. GCM is the only other method to control Type I error under all simulation settings, but it has nearly zero power with nonlinear data.

4.4 Knockoff filter

To compare the performance of the CPI with that of the original knockoff filter, we followed the simulation procedure described in Sect. 4 of (Candès et al., 2018). A $n = 300 \times p = 1000$ matrix was sampled from a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, with covariance matrix$\Sigma_{ij} = \rho^{|i-j|}$. A continuous outcome Y was calculated as$Y = \mathbf{X}\boldsymbol{\beta} + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$ and the coefficient vector $\boldsymbol{\beta}$ contains just 60 nonzero entries, with random signs and variable effect sizes. We vary ρ with fixed nonzero $|\beta| = 1$, and vary effect size with fixed $\rho = 0$.

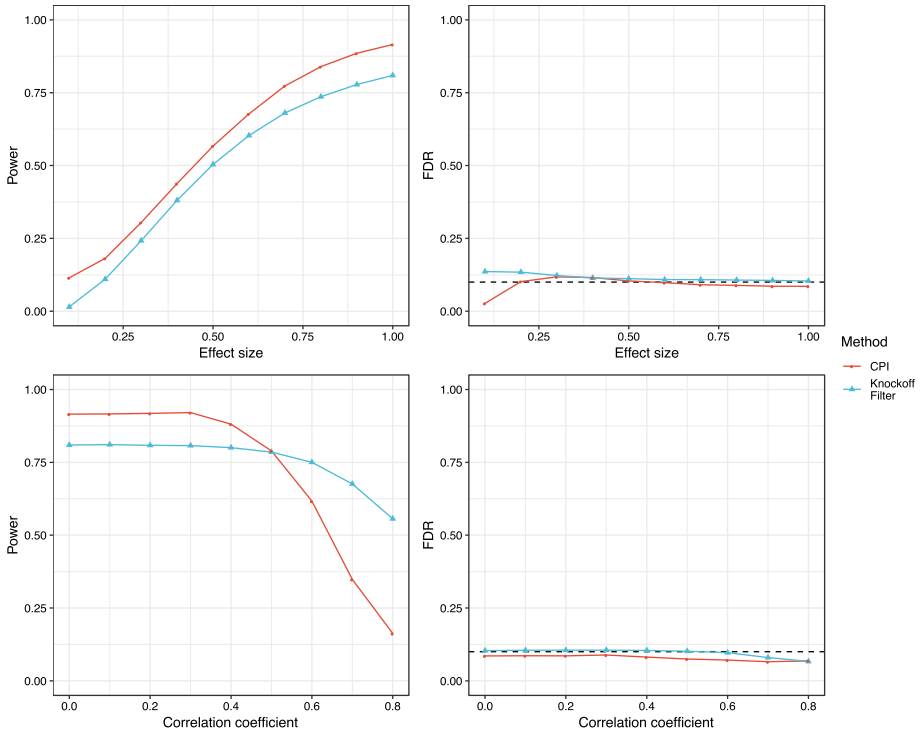

Fig. 3 Power and FDR as a function of effect size and autocorrelation for CPI and knockoff filter. Target FDR is 10%. Results are from a lasso regression with $n = 300$ and $p = 1000$. Each point represents 10^4 replications. Similar results were obtained for $p = 2000$ (see Supplement, Sect. 3)

We train a series of lasso regressions (Tibshirani, 1996) on the data using the original design matrix and tenfold cross-validation to calculate the CPI, and the expanded $n \times 2p$ design matrix for the knockoff filter. VI for the latter was estimated using the difference statistic originally proposed by Barber and Candès (2015):

$$W_j = \left|\hat{\beta}_j\right| - \left|\hat{\beta}_{j+p}\right|,$$

where $\widehat{\beta}_j$ and $\widehat{\beta}_{j+p}$ represent coefficients associated with a feature and its knockoff, respectively, at some fixed value of the Lagrange multiplier λ. Variables are selected based on the ATT method described in Sect. 2.3. The hyperparameter λ is tuned via tenfold cross-validation, per the default settings of the `glmnet` package (Friedman et al., 2010). Power and FDR are averaged over 10^4 iterations for each combination of effect size and autocorrelation coefficient. FDR is estimated via the ATT procedure for the knockoff filter and via the Benjamini–Hochberg algorithm (1995) for the CPI.

The CPI is more powerful than the original knockoff filter for all effect sizes at $\rho = 0$, but less powerful for high autocorrelation of $\rho = 0.5$ (see Fig. 3). Both methods generally control the FDR at the target rate of 10%. The only exceptions are under small effect sizes, where the knockoff filter shows slightly inflated errors. Similar results for $p = 2000$ are included in the Supplement.

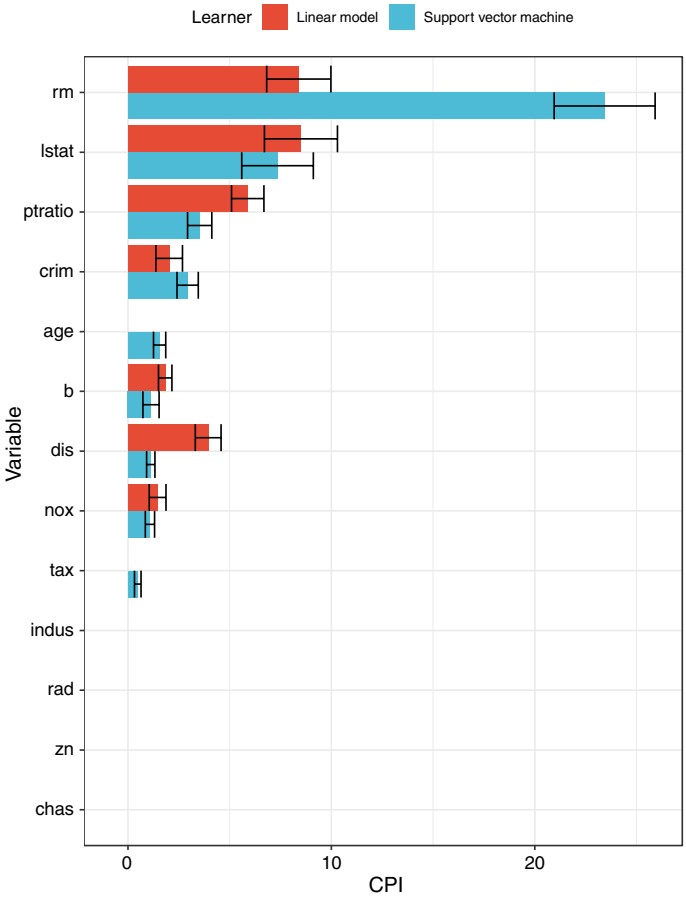

Fig. 4 Results of the Boston housing experiment. For each variable in the data set, the CPI value is shown, computed with a linear model and a support vector machine. Whiskers represent standard errors. Non-significant variables at $\alpha = 0.05$ after adjustment for multiple testing are shaded

Note that in addition to being a more powerful test under most conditions, the CPI has other important advantages over the ATT. Whereas the latter can only be applied to algorithms with inbuilt feature scoring statistics, the former requires nothing more than a valid loss function. Whereas the ATT struggles to select important variables in low-dimensional settings, the CPI imposes no dimensionality restraints. Finally, the CPI is more informative, insomuch as it provides feature-level p-values and confidence (or credible) intervals.

4.5 Real data

In this section, we apply the CPI to real datasets of low- and high-dimensionality.

4.5.1 Boston housing

We analyzed the Boston housing data (Harrison & Rubinfeld, 1978), which consists of 506 observations and 14 variables. This benchmark dataset is available in the UCI Machine Learning Repository (Dua & Graff, 2017). The dependent variable is the median price of owner-occupied houses in census tracts in the Boston metropolitan area in 1970. The independent variables include the average number of rooms, crime rates, and air pollution.

Using LM and SVM regressions, we computed CPI, standard errors, and t-test p-values for each feature, adjusting for multiple testing using Holm's (1979) procedure. We used an RBF kernel for the SVM, measured performance via MSE, and used 5 subsampling iterations to evaluate empirical risk. The results are shown in Fig. 4. We found significant effects at $\alpha = 0.05$ for the average number of rooms (`rm`), percentage of lower status of the population (`lstat`), pupil-teacher ratio (`ptratio`), and several other variables with both LM and SVM, which is in line with previous analyses (Friedman & Popescu, 2008; Williamson et al., 2021). Interestingly, the SVM assigned a higher CPI value to several variables compared to the LM. For example, the proportion of owner-occupied units built prior to 1940 (`age`) significantly increased the predictive performance of the SVM but had approximately zero impact on the LM. The reason for this difference might be a nonlinear interaction between `rm` and `age`, which was also observed by Friedman and Popescu (2008).

4.5.2 Breast cancer

We examined gene expression profiles of human breast cancer samples downloaded from GEO series GSE3165. Only the 94 arrays of platform GPL887 (Agilent Human 1A Microarray V2) were included. These data were originally analysed by Herschkowitz et al. (2007) and later studied by Lim et al. (2009). We follow their pre-processing pipeline, leaving 13,064 genes. All samples were taken from tumor tissue and classified into one of six molecular subtypes: basal-like, luminal A, luminal B, Her2, normal-like, and claudin-low.

Basal-like breast cancer (BLBC) is an especially aggressive form of the disease, and BLBC patients generally have a poor prognosis. Following Wu and Smyth (2012), we defined a binary response vector to indicate whether samples are BLBC. Gene sets were downloaded from the curated C2 collection of the MSigDB and tested for their association with this dichotomous outcome.

We trained an RF classifier with 10^4 trees to predict BLBC based on microarray data. Second-order knockoffs were sampled using an approximate semidefinite program with block-diagonal covariance matrices of maximum dimension 4000×4000. We test the CPI for each of the 2609 gene sets in the C2 collection for which at least 25 genes were present in the expression matrix. Models were evaluated using the CE loss function on out-of-bag samples.

We calculate p-values for each CPI via the t-test and corresponding q-values using the Benjamini–Hochberg procedure. We identify 660 significantly enriched gene sets at $q \leq 0.05$, including 24 of 73 explicitly breast cancer derived gene sets and 6 of 13 gene sets indicative of basal signatures. Almost all top results are from cancer studies or other biologically relevant research (see Fig. 5). These include 4 sets of BRCA1 targets, genetic mutations known to be associated with BLBC (Turner & Reis-Filho, 2006), and 4 sets of ESR1 targets, which are markers for the luminal A subtype (Sørlie et al., 2003).

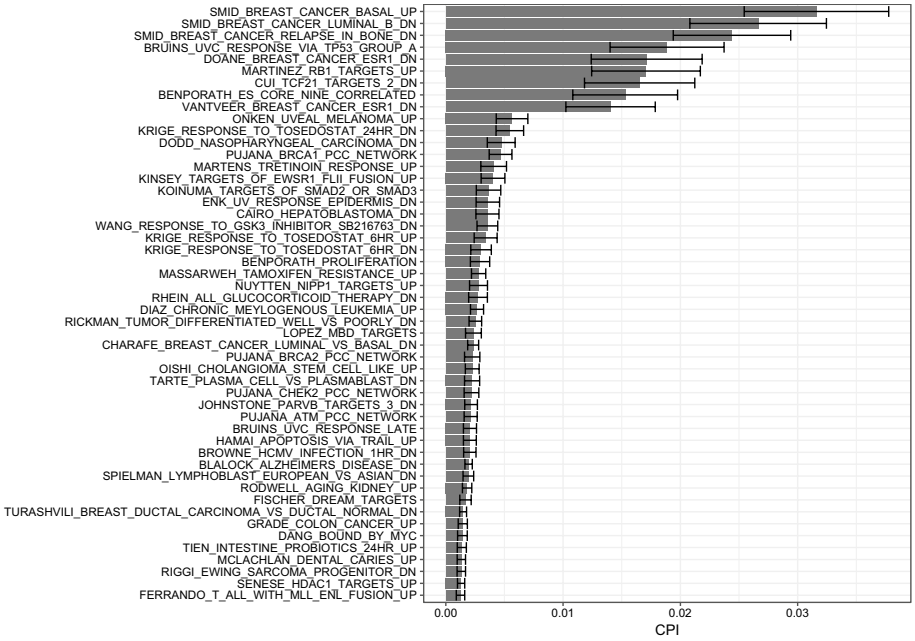

Fig. 5 Results for the top 50 gene sets. For each gene set, the CPI value is shown, computed with a random forest. Whiskers represent standard errors

By comparison, popular pathway enrichment tests like GSEA (Subramanian et al., 2005) and CAMERA (Wu & Smyth, 2012) respectively identify 137 and 74 differentially enriched pathways in this dataset at 5% FDR. These results are especially notable given that those methods rely on marginal associations between gene expression and clinical outcomes, whereas the CPI is a conditional test with a more restrictive null hypothesis, and should theoretically have less power to detect enrichment when features within a gene set are correlated with others outside it. Despite collinearity between genes, the CPI still identifies a large number of biologically meaningful gene sets differentiating BLBC tumors from other breast cancer subtypes.

5 Discussion

Shah and Peters (2020) have demonstrated that no CI test can be uniformly valid against arbitrary alternatives, a sort of no-free-lunch (NFL) theorem for CI. Bates et al. (2020) prove a similar NFL theorem for constructing knockoff variables, showing that no algorithm can efficiently compute nontrivial knockoffs for arbitrary input distributions. The original NFL theorem for optimization is well-known (Wolpert & Macready, 1997). Together, these results delimit the scope of the CPI. The method is highly modular, in the sense that it works with any well-chosen combination of supervised learner, loss function, and knockoff sampler. However, it is simultaneously constrained by these choices. The CPI cannot be guaranteed to control Type I error or have any power against the null when

knockoffs are poorly constructed or models are misspecified. If the selected loss function ignores distributional information from higher moments, then the CPI will as well.

In our experiments, we employed a variety of risk estimators, including cross-validation, subsampling, out-of-bag estimates, and the holdout method. Results did not depend on these choices, suggesting that analysts may use whichever is most efficient for the problem at hand. We also used a number of different learning algorithms and found that all showed good convergence properties in finite samples—even the SVM, which is known to have infinite VC dimension, and therefore violates the consistency criterion cited in Sect. 3.

Computational bottlenecks can complicate the use of this procedure for high-dimensional datasets. It took approximately 49 hours to generate second-order knockoffs for the gene expression matrix described in Sect. 4.2.2. However, as noted in Sect. 2.3, knockoff sampling is an active area of research, and it is reasonable to expect future advances to speed up the procedure considerably.

6 Conclusion

We propose the conditional predictive impact (CPI), a global measure of variable importance and general test of conditional independence. It works for regression and classification problems using any combination of knockoff sampler, supervised learning algorithm, and loss function. It imposes no parametric or sparsity constraints, and can be efficiently computed on data with many observations and features. Our inference procedures are fast and effective, able to simultaneously control Type I error and achieve high power in finite samples. We have shown that our approach is consistent and unbiased under minimal assumptions. Empirical results demonstrate that the method performs favorably against a number of alternatives for a range of supervised learners and data generating processes.

We envision several avenues for future research in this area. Localized versions of the CPI algorithm could be used to detect the conditional importance of features on particular predictions. Model-specific methods could be implemented to speed up the procedure. We are currently working on applications for causal discovery and inference, an especially promising direction for this approach.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10994-021-06030-6.

Acknowledgements We thank our anonymous reviewers for their helpful feedback. DSW received funding for this project from ONR Grant N62909-19-1-2096. MNW received funding for this project from the German Research Foundation (DFG), Emmy Noether Grant 437611051.

References

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE, 10*(7), 1–46.

Barber, R. F., & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Annals of Statistics, 43*(5), 2055–2085.

Bates, S., Candès, E., Janson, L., & Wang, W. (2020). Metropolized knockoff sampling. *Journal of the American Statistical Association*, 1–15.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (statistical Methodology), 57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29*(4), 1165–1188.

Berrett, T. B., Wang, Y., Barber, R. F., & Samworth, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (statistical Methodology)*, *82*(1), 175–197. https://doi.org/10.1111/rssb.12340

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., et al. (2016). mlr: Machine learning in R. *Journal of Machine Learning Research, 17*(170), 1–5.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 1–33.

Candès, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (statistical Methodology), 80*(3), 551–577.

Doran, G., Muandet, K., Zhang, K., & Schölkopf, B. (2014). A permutation-based kernel conditional independence test. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence* (pp. 132–141).

Dua, D., & Graff, C. (2017). *UCI machine learning repository*. University of California, School of Information and Computer Science.

Feng, J., Williamson, B., Simon, N., & Carone, M. (2018). Nonparametric variable importance using an augmented neural network with multi-task learning. In *Proceedings of the International Conference on Machine Learning* (pp. 1496–1505).

Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*(177), 1–81.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research, 5*, 1531–1555.

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics, 2*(3), 916–954.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1–41.

Fukumizu, K., Gretton, A., Sun, X., & Schölkopf, B. (2008). Kernel measures of conditional dependence. *Advances in Neural Information Processing Systems*, *20*, 489–496.

Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling, 160*(3), 249–264.

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2015). Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis, 90*, 15–35.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician, 61*(2), 139–147.

Guedj, B. (2019). A primer on PAC-Bayesian learning. *arXiv* preprint, 1901.05353.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*(7/8), 1157–1182.

Hansen, D., Manzo, B., & Regier, J. (2021). Normalizing flows for knockoff-free controlled feature selection. *arXiv* preprint, 2106.01528.

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management, 5*(1), 81–102.

Herschkowitz, J. I., Simin, K., Weigman, V. J., Mikaelian, I., Usary, J., Hu, Z., et al. (2007). Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biology, 8*(5), R76.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.
Hubbard, A. E., Kennedy, C. J., & van der Laan, M. J. (2018). Data-adaptive target parameters. In M. J. van der Laan & S. Rose (Eds.), *Targeted learning in data science* (pp. 125–142). Springer.
Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., & Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software, 47*(11), 1–26.
Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press.
Korb, K. B., & Nicholson, A. E. (2009). *Bayesian artificial Intelligence* (2nd ed.). Chapman and Hall/CRC.
Kruschke J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146
Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC.
Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software, 36*(11), 1–13.
Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association, 113*(523), 1094–1111.
Lim, E., Vaillant, F., Wu, D., Forrest, N. C., Pal, B., Hart, A. H., et al. (2009). Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nature Medicine, 15*, 907.
Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Longman.
Lundberg, S. M., & Lee, S. -I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
Maathuis, M. H., Kalisch, M., & Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *Annals of Statistics, 37*(6A), 3133–3164.
Martínez Sotoca, J., & Pla, F. (2010). Supervised feature selection by clustering using conditional mutual information-based distances. *Pattern Recognition, 43*(6), 2068–2081.
Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (statistical Methodology), 72*(4), 417–473.
Mentch, L., & Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research, 17*(1), 841–881.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2018). *e1071: Misc functions of the department of statistics, probability theory group*. CRAN. R package version 1.7–0.
Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics, 11*(1), 110.
Patterson, E., & Sesia, M. (2018). *knockoff*. CRAN. R package version 0.3.2.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.
Phipson, B., & Smyth, G. (2010). Permutation *P*-values should never be zero: Calculating exact *P*-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, *9*(1).
Ramsey, J. D. (2014). A scalable conditional independence test for nonlinear, non-Gaussian data. *arXiv* preprint, arXiv:1401.5031
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).
Rinaldo, A., Wasserman, L., & G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *Annals of Statistics, 47*(6), 3438–3469.
Romano, Y., Sesia, M., Candès, E. (2020). Deep Knockoffs. *Journal of the American Statistical Association*, *115*(532) 1861–1872. https://doi.org/10.1080/01621459.2019.1660174
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225
Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory Series A, 13*(1), 145–147.
Scutari, M. (2010). Learning Bayesian networks with the bnlearnR package. *Journal of Statistical Software, 35*(3), 1–22.
Scutari, M., & Denis, J.-B. (2014). *Bayesian networks: With examples in R*. Chapman and Hall/CRC.
Sesia, M., Sabatti, C., & Candès, E. J. (2019). Gene hunting with hidden Markov model knockoffs. *Biometrika, 106*(1), 1–18.

Shah, R., & Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics, 48*(3), 1514–1538.
Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.
Shelah, S. (1972). A combinatorial problem: Stability and orders for models and theories in infinitariy languages. *Pacific Journal of Mathematics, 41*(1), 247–261.
Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning* (Vol. 70, pp. 3145–3153).
Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences, 100*(14), 8418–8423.
Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). The MIT Press.
Steinke, T., & Zakynthinou, L. (2020). Reasoning about generalization via conditional mutual information. In *Proceedings of the International Conference on Learning Theory* (pp. 3437–3452).
Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (statistical Methodology), 64*(3), 479–498.
Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*(1), 307.
Strobl, E. V., Zhang, K., & Visweswaran, S. (2018). Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, *7*(1), 20180017.
Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., & Gillette, M. A. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences, 102*(43), 15545–15550.
Tansey, W., Veitch, V., Zhang, H., Rabadan, R., & Blei, D.M. (2021). The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics*, 1–37.
Team, R. C. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B, 58*(1), 267–288.
Turner, N. C., & Reis-Filho, J. S. (2006). Basal-like breast cancer and the BRCA1 phenotype. *Oncogene, 25*, 5846.
van der Laan, M. J. (2006). Statistical inference for variable importance. *The International Journal of Biostatistics*, *2*(1).
van der Laan, M. J., & Rose, S. (Eds.). (2018). *Targeted learning in data science: Causal inference for complex longitudinal studies*. Springer.
Vapnik, V., & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies to their probabilities. *Theory of Probabability & Its Applications, 16*(2), 264–280.
Vejmelka, M., & Paluš, M. (2008). Inferring the directionality of coupling with conditional mutual information. *Physical Review E, 77*(2), 26214.
Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.
Verma, T., & Pearl, J. (1991). Equivalence and synthesis of causal models. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence* (pp. 255–270).
Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.
Wetzels, R., Raaijmakers, J. G. W., Jakab, E., Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian *t* test. *Psychonomic Bulletin & Review*, *16*(4), 752–760. https://doi.org/10.3758/PBR.16.4.752
Williamson, B. D., Gilbert, P. B., Carone, M., & Simon, N. (2021). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, *77*(1), 9–22. https://doi.org/10.1111/biom.13392
Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67–82.
Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1).
Wu, D., & Smyth, G. K. (2012). Camera: A competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research, 40*(17), e133.

Zhang, K., Peters, J., Janzing, D., & Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence* (pp. 804–813).