Ultra-marginal Feature Importance

Joseph Janssen  
Department of Earth, Ocean and Atmospheric Sciences  
University of British Columbia  
Vancouver  
joejanssen@eoas.ubc.ca

Vincent Guan  
Department of Mathematics  
University of British Columbia  
Vancouver  
vguan23@math.ubc.ca

Abstract

Scientists frequently prioritize learning from data rather than training the best possible model; however, research in machine learning often prioritizes the latter. Marginal feature importance methods, such as marginal contribution feature importance (MCI), attempt to break this trend by providing a useful framework for quantifying the relationships in data in an interpretable fashion. In this work, we generalize the framework of MCI while aiming to improve performance and runtime by introducing ultra-marginal feature importance (UMFI). To do so, we prove that UMFI can be computed directly by applying preprocessing methods from the AI fairness literature to remove dependencies in the feature set. We show on real and simulated data that UMFI performs at least as well as MCI, with significantly better performance in the presence of correlated interactions and unrelated features, while substantially reducing the exponential runtime of MCI to super-linear.

1 Introduction

Scientists often seek to determine the true relationships between a set of characteristics and some outcome of interest. These relationships are ideally determined by performing carefully controlled experiments so that causality can be established. However, experiments can be difficult and costly to pursue, unethical to perform, or impossible to control [70], leaving only observational data available. The relationships that are hidden within vast quantities of observational data are often difficult to determine, so statistical tools, such as feature importance, have been explored. Feature importance quantifies and, more importantly, ranks how related explanatory features are to a response.

There is no consensus on how to define feature importance, so choosing an appropriate method strongly depends on the questions that the user seeks to answer [66, 31, 14]. The most widely referenced distinction between feature importance methods is the conditional versus marginal divide, which defines the two extremes of feature importance methods [23]. This division is sometimes also framed as true to the model versus true to the data [16]. The distinction between these two types of methods is only evident in the presence of dependent features. Indeed, if all explanatory features in a dataset are independent, conditional and marginal importances are exactly the same [23, 50]. Conditional methods are typically used for feature selection, dimensionality reduction, and improving prediction performance. On the other hand, marginal methods are specifically developed to explain and interpret data [31], seeking to avoid the tendency of conditional methods to explain the model and give correlated features too little importance [16]. For example, if a scientist wants to build a database of genes associated with some disease, then a marginal approach could be preferred [23, 14, 16]. Alternatively, if an engineer wants to determine the smallest number of features to obtain good model predictions, a conditional importance metric would be more suitable [31, 14].

Recently, feature importance methods such as Shapely-values [54, 18, 45], SAGE [21], accumulated local effects (ALE) [4], permutation importance [10], and conditional permutation importance (CPI) [23] have been used in high-impact journal papers by scientists who want to explain the mechanisms
within data [2, 6, 57, 37, 53, 29]. However, these methods do not adequately account for correlated features and feature interactions. ALE can only easily show first order effects [49], and although CPI improves upon some limitations of permutation importance, CPI has the property that two perfectly correlated features with significant predictive power would both be deemed unimportant [21]. Further, only one model is trained in ALE, CPI, and permutation importance. Thus, correlated features, which can alter the model assembly process, could be given artificially low importance if the goal is to explain the data [32]. Developers of feature importance methods like SAGE and marginal contribution feature importance (MCI) have attempted to address these issues by evaluating the difference in accuracy between a model trained with the feature of interest and a model trained without it, across all feature subsets [15, 21]. In particular, MCI was shown to have better quality and robustness when compared to Shapely-values, SAGE, ablation, and bivariate methods [15]. However, subset-based feature importance methods usually require an exponential number of model trainings, which makes these methods ineffective for interpreting large datasets (e.g., gene expression studies).

In this paper, we introduce ultra-marginal feature importance (UMFI), a new subset-based marginal feature importance method that performs at least as well as previous methods while drastically reducing run-time. UMFI generalizes the framework of MCI, developed by Catav et al. [15], in order to overcome some key shortcomings. First, MCI is given by a maximization problem that requires searching across all subsets of the feature space $F$. Second, although it can handle complex feature interactions and data with correlated features, MCI underestimates the importance of correlated features that interact in the expression for the response variable. Third, MCI can give non-zero importance to features that are completely unrelated to the response variable.

The rest of this paper is organized as follows. A brief overview of MCI is given in Section 2. The theoretical framework for UMFI is then formally presented in Section 3 along with an algorithm to compute the UMFI of a feature $f$. In Section 4, we conduct experiments on simulated and real data to assess the quality, robustness, and time complexity of UMFI compared to MCI. Finally, an overview of the work, its limitations, and ideas for future work are discussed in Section 5.

Related work

This paper is greatly inspired by the work of Catav et al. [15], whose marginal contribution feature importance was developed to handle complex feature interactions and correlated features, so we build upon their strong framework. To do so, we generalize the central maximization problem for MCI to also consider transformations of the feature set $F$. We show that the new maximization problem, which maximizes over all “information subsets” of $F$, is attained by the largest information subset that is independent of the feature of interest $f$. Therefore, the maximizer can be computed immediately after preprocessing the data to remove dependencies on $f$. Ideally, this would be done in a manner that preserves as much information from the original features as possible.

Finding orthogonal predictors for resolving the controversies of feature importance in the presence of multicollinearity is not strictly new [28]. However, the discussion of orthogonal predictors for relative feature importance methods has been seemingly limited to applications to multiple linear regression, mostly in the domain of psychology [8, 71]. While the techniques for orthogonalizing predictors has been limited to fairly simple linear algebra, more advanced and more general dependency removal methods have seen great progress within the AI fairness and privacy literature. The need for these independent information preserving data representations for better feature importance methods was mentioned as future work in König et al. [40]. Some examples of these techniques include linear regression [9], optimal transport [56], neural networks [13, 55], convex optimization [13], and principal inertial components [63]. Linear regression and optimal transport were implemented for UMFI in this paper.

Besides MCI and SAGE [15, 21], other feature importance methods have recognized the difficulty in providing accurate scores when complex interactions are present and when predictors are dependent. Likely owing to the simplicity and popularity of permutation-based methods, the most common approach has been to retool permutation importance [10] and conditional permutation importance [58] to better handle these issues. For example, the conditional subgroup permutation has been proposed, such that the subgroups are constructed by decision trees, in which the distribution of the feature becomes less dependent on other features [50]. However, this method only performs well when the response is dependent on just a few features [50]. Despite their popularity and myriad of variations, a thorough overview by Hooker et al. [32] has shown that permutation-based methods are not suitable when trained on a feature set exhibiting statistical dependencies, as they can produce...
highly misleading results. Hooker et al. [32] instead advocate for feature importance methods that achieve their score by 1) permuting or generating new values of the feature using its distribution conditional on the remaining features, or 2) removing the feature and examining the drop in accuracy when a new model is trained. We note that MCI and UMFI both fall under the second category.

2 Marginal contribution feature importance

Let $F$ be the set of features used to predict the response variable, $Y$. Given an evaluation function $\nu : \mathcal{P}(F) \to \mathbb{R}^+$, Catav et al. [15] defined the marginal contribution feature importance (MCI) of a feature $f \in F$ as

$$I_\nu(f) = \max_{S \subseteq F} \nu(S \cup \{f\}) - \nu(S).$$

(1)

Let $\hat{y}$ be a constant predictor and $G(S)$ be the set of all predictive models restricted to using features in $S \subseteq F$. Given a loss function $l$, the "idealized" evaluation function satisfies

$$\nu(S) = \min_{\hat{y}} \mathbb{E}[l(\hat{y}, Y)] - \min_{g \in G(S)} \mathbb{E}[l(g(S), Y)].$$

(2)

With this definition, $\nu(\emptyset) = 0$ since both terms will cancel out, and $\nu$ is monotonically increasing since if $S_1 \subseteq S_2$ then $G(S_1) \subseteq G(S_2)$. Though these last two properties are not strictly guaranteed when faced with real data and an imperfect model, they reasonably follow the intuition that more features give a model more information about the response variable. In practice, one can for example use machine learning models such as elastic-net or random forests for $G$ and negative log-loss cross validation or out-of-bag accuracy for the loss function $l$ [15].

Given a monotonically increasing evaluation function $\nu$, satisfying $\nu(\emptyset) = 0$, Catav et al. [15] proved that $I_\nu(f)$ is the only function that satisfies three desirable axioms that they believe a method should have for explaining data properly. First, the marginal contribution axiom states that the importance of a feature should be at least as large as the increase in the evaluation function after adding the feature to the model containing all other features, i.e. $I_\nu(f) \geq \nu(F) - \nu(F \setminus \{f\})$. Second, the elimination axiom states that the importance of a feature can only decrease if features from the feature set are removed. Third, the minimalism axiom states that we should take the minimum feature importance value that satisfies the first two axioms, thus ensuring its uniqueness.

3 Ultra-marginal feature importance

We hypothesize that the maximization function formulated by Catav et al. [15] is too restrictive. Instead, we introduce ultra-marginal feature importance (UMFI) by proposing a maximization over information subsets of the feature set $F$

$$U_\nu(f) = \max_{S \in \mathcal{I}(F)} \nu(S \cup \{f\}) - \nu(S).$$

(3)

We define the space of information subsets of $F$ by $\mathcal{I}(F) = \{g(A) : A \subseteq F\}$ where $g$ is any real or vector-valued function that may act on $A$. We call these information subsets because $I(Y; g(F)) \leq I(Y; F)$ is a property of mutual information for any function $g$ [20, p. 33]. This definition allows us to consider functions that transform feature subsets of the form $B \cup f$ into a feature subset of size $|B|$ that is independent of the feature of interest $f$. In particular, we may use dependency removal methods from the AI fairness literature, which were developed to achieve unbiased predictions [55, 30]. Given this expanded framework, $U_\nu(f)$ obeys the marginal contribution and elimination axioms laid out by Catav et al. [15] (Theorem C.1).

In addition to $\nu(\emptyset) = 0$ and $\nu$ being monotonically increasing, we also assume that $\nu$ is supermodular under mutual independence and that $\nu$ is invariant under duplicate information, i.e. $\nu(A) = \nu(A \cup B)$ if $B \subseteq A$. Both of these conditions are satisfied when $\nu(S) = I(Y; S)$ [30]. Supplement A.1. The evaluation function is equivalent to mutual information in the idealized setting where we know the Bayes classifier and $\nu$ is defined via (2) with the cross entropy loss function [21]. In practice, the accuracy of the approximation $\nu(S) \approx I(Y; S)$ depends on the quality of the machine learning model and the specified loss function. Assuming these properties, we can solve directly for the maximizing information subset $S^*$ for Equation 5.
**Theorem 3.1.** The maximizing information subset $S^*$ for (3) is given by $S^* = \bigcup_{S \in I(F): S \perp f} S$, which is the largest information subset of $F$ such that $S^* \perp f$.

**Proof.** First, suppose for contradiction that $S^*$ solves (3) and that $S^* \notin f$. Let $\tilde{S}^*$ represent the remaining information of $S^*$ after having removed dependencies of $f$ from $S^*$. Then, $\nu(S^* \cup f) = \nu(\tilde{S}^* \cup f)$ from invariance under duplicate information, but $\nu(S^*) \geq \nu(\tilde{S}^*)$ from monotonicity. Hence, $\nu(\tilde{S}^* \cup f) - \nu(\tilde{S}^*) \geq \nu(S^* \cup f) - \nu(S^*)$, which contradicts the fact that $S^*$ is the maximizing information subset, so we may assume that $S^* \perp f$.

Now, suppose that $S^*, X, f$ are mutually independent information subsets of $F$. Then, by the supermodularity of $\nu$ under mutual independence, $X$ must be included in $S^*$. By invariance under duplicate information, the maximizer for (3) is satisfied by $S^* = \bigcup_{S \in I(F): S \perp f} S$. \qed

Note that the union of all information subsets that are independent of $f$ is equivalent to the information subset obtained by removing dependencies on $f$ from the feature set $F$ because of invariance under duplicate information. This immediately implies a fast algorithm for computing the ultra-marginal feature importance of a feature $f \in F$ (Algorithm 1).

**Algorithm 1: Algorithm for computing UMFI**

1. Let $Y$ be the response variable of the set of predictors $F$. Choose a feature $f \in F$.
2. Estimate $S^*$ by removing dependencies on $f$ from $F$ with minimal information loss.
3. Specify a model and evaluation function $\nu$.
4. Train a model using features $S^*$ to predict $Y$ and compute $\nu(S^*)$.
5. Train a model using features $(S^*, f)$ to predict $Y$ and compute $\nu(S^* \cup f)$.
6. **return** $U_\nu(f) = \nu(S^* \cup f) - \nu(S^*)$

4 Experiments

We perform experiments to compare UMFI and MCI with respect to quality, robustness, and time complexity. To implement UMFI, we consider optimal transport [40] (UMFI_OT) and linear regression [5] (UMFI_LR) as methods to remove dependencies from the data. A detailed overview of these implementations is shown in Supplement F and experiments comparing these methods appear in Supplement E. For all experiments, we use random forests’ out-of-bag accuracy ($R^2$ OOB-accuracy for regression tasks and OOB classification accuracy for classification tasks) as the evaluation metric $\nu$. We use the `ranger` R package to implement random forests with default hyperparameters and 100 for the number of trees [69]. All experiments were run in Microsoft R Open Version 4.0.2 [48]. Supplement G contains additional experiments comparing feature importance metrics. Code for all experiments can be found at https://github.com/joej1997/UMFI.

4.1 Experiments on simulated data

We run UMFI on simulated data with correlated interactions, correlated features, and non-linear interactions to verify that it performs well compared to MCI. The data in all simulation studies contains one response variable $Y$, four explanatory features $x_1, x_2, x_3, x_4$, and 500 randomly generated observations. Each of the simulation studies are repeated 100 times so that we can also test the stability of each method.

4.1.1 Correlated interactions

Interacting features are often correlated [33, 34], so it is vital to test if feature importance metrics work as desired in these settings. Let $A, B, C, D, E, G \sim N(0, 1)$. We consider

$$x_1 = A + B, \quad x_2 = B + C, \quad x_3 = D + E, \quad x_4 = E + G$$

$$Y = x_1 + x_2 + \text{sign}(x_1 * x_2) + x_3 + x_4$$
In this simulation, \( x_1 \) is correlated with \( x_2 \), and \( x_3 \) is correlated with \( x_4 \) in the same way. We would expect \( x_1 \) and \( x_2 \) to be more important than \( x_3 \) and \( x_4 \) because of the extra interaction term, \( \text{sign}(x_1 * x_2) \). The results in Figure 1a clearly show that UMFI provides better estimations of feature importance compared to MCI when correlated interactions are present. The variability of all estimates are approximately the same across methods. However, MCI estimates that all features have approximately the same feature importance scores, while both UMFI methods show significantly greater importance for \( x_1 \) and \( x_2 \) compared to \( x_3 \) and \( x_4 \). MCI fails in this experiment because \( x_2 \) contains most of the important information coming from \( x_1 \), which implies that \( I(Y; x_1, x_2) - I(Y; x_1) \leq I(Y; x_1) - I(Y; \emptyset) \). Therefore, when computing the importance for \( x_1 \), the maximizing subset that solves Equation 1 cannot contain \( x_2 \) if \( \nu(S) \sim I(Y; S) \). UMFI is able to detect this interaction because it can extract the information from \( x_2 \) that interacts with \( x_1 \) while keeping this extracted feature independent of \( x_1 \).

### 4.1.2 Correlations

Marginal feature importance methods such as MCI and UMFI should not change the measured importance of features in the presence of correlated variables. To test this, we implement a simulation study similar to the ones found in Catav et al. [15]. Let \( \epsilon \sim \mathcal{N}(0, 0.2) \). We consider

\[
\begin{align*}
  x_1, x_2, x_4 &\sim \mathcal{N}(0, 1), \ \ x_3 = x_1 + \epsilon \\
  Y & = x_1 + x_2
\end{align*}
\]

The addition of \( x_3 \) should not alter the importance of \( x_1 \), and \( x_1 \) should remain equally as important as \( x_2 \), since they have the same influence on the response \( Y \). The results shown in Figure 1b show that both MCI and UMFI work reasonably well. As with the previous simulation experiment, the variability is consistent across methods. As was desired, MCI and UMFI with linear regression show equal relative importance scores for \( x_1 \) and \( x_2 \). The importance given to \( x_2 \) was slightly greater than \( x_1 \) in the MCI and UMFI optimal transport scenario, but the difference is not significant. Interestingly, MCI assigns some importance to \( x_4 \), which was independent of the response, while both UMFI methods assign importance scores close to zero. Because of this, we conclude that UMFI with linear regression performs the best in this simulated scenario.

### 4.1.3 Non-linear interactions

Interaction effects are common in many scientific disciplines where assessing feature importance is prevalent, including hydrology [34, 2, 42], genomics [14, 65, 51], and finance [73, 59]. So, as was done in Catav et al. [15], we assess the ability of MCI and UMFI to detect non-linear interaction effects in the data. This simulation study provides an XOR-type causal relationship where individually, the explanatory features are unrelated with the response, but the response is influenced by the interaction of two features [46]. Let \( \epsilon \sim \text{Exp}(1/\sqrt{2}) \). We consider

\[
\begin{align*}
  x_1, x_2, x_3, x_4 &\sim \mathcal{N}(0, 1) \\
  Y & = \text{sign}(x_1 * x_2) * \epsilon
\end{align*}
\]

Ideally, the results of a feature importance metric would conclude that \( x_1 \) and \( x_2 \) have equally high importance while \( x_3 \) and \( x_4 \) have no importance. Figure 1c shows consistently good performance across all methods. Each method gave high relative importance scores to \( x_1 \) and \( x_2 \), while \( x_3 \) and \( x_4 \) received scores near zero. This is perfect as the response was only a function of \( x_1 \) and \( x_2 \), while it was independent of \( x_3 \) and \( x_4 \). The only difference across methods is that the variability of estimates was slightly greater when using UMFI with optimal transport compared with the other methods.

### 4.2 BRCA experiments

We use the same breast cancer (BRCA) classification dataset [62] used in previous feature importance studies including Catav et al. [15] and Covert et al. [21] to test the quality and robustness of UMFI on real data. The original data contains over 17,000 genes and 571 anonymous patients that have been diagnosed with one of 4 breast cancer sub-types. We consider the same subset of 50 genes as in Catav et al. [15] and Covert et al. [21] for easier computation and result visualization. Of the 50 selected genes, 10 are known to be associated with breast cancer, while the other 40 genes are randomly sampled. This data was downloaded from [https://github.com/TAU-MLwell/](https://github.com/TAU-MLwell/).
Figure 1: Results for the experiments on simulated data from Subsection 4.1. Feature importance scores are shown as a percentage of the total for each of $x_1$ to $x_4$. Results are shown for marginal contribution feature importance (MCI), ultra-marginal feature importance with linear regression (UMFI-LR), and ultra-marginal feature importance with pairwise optimal transport (UMFI-OT). Each box plot summarizes the results of the 100 replications of each simulation. The first and third quantiles (thin lines), the median (thick line), and outliers (points) of the 100 replications are shown.

In Catav et al. [15] and Covert et al. [21], these 40 randomly sampled genes are assumed to be unassociated with breast cancer. However, to ensure a more definitive ground truth, we also randomly permute the values of these 40 genes across their respective 571 observations to further reduce the chance that these genes have any association with breast cancer. Quality is then measured with the true positive and true negative rates: the 10 BRCA associated genes should have some non-zero importance (positive), and the other 40 genes should have exactly zero importance (negative). These experiments were run 200 times on different seeds and with a different random sample of 500 patients for each iteration. Robustness is measured using the standardized interquartile range (SIQR) of each feature’s importance from repeated experiments, which is calculated by dividing the average IQR across the 50 features by the average median.

We found that MCI and UMFI (UMFI_LR and UMFI_OT) correctly gave significant importance to the 10 genes that are known to be associated with breast cancer (Figure 2). However, MCI consistently gives non-zero importance to all features, while UMFI gives zero importance to the majority of the randomized genes as desired. Of the 40 randomized genes, the few that UMFI gives a non-zero score to have a score close to zero, and importantly, their scores are much smaller compared to the scores of any of the 10 BRCA-associated genes. During these experiments, we noticed that a greater proportion of permuted genes are given zero importance as the number of iterations grows. To confirm the convergence of these importance scores we ran the UMFI experiments 5000 times and noted that both UMFI methods have a perfect overall accuracy when distinguishing between important and permuted features (Supplement G.2.1). Interestingly, the ordering of important features was similar.
across methods, with BCL11A and SLC22A5 always being the most important and TEX14 always being the least important of the 10 BRCA-associated genes. Although UMFI scores have higher variability than MCI (Table 1), it is clear from the feature importance graph that UMFI separates the 10 associated genes from the 40 unassociated genes better than MCI does even after considering UMFI’s higher SIQR.

Even though UMFI should theoretically dominate MCI \( U_\nu(f) \geq I_\nu(f) \), we observed that virtually all of the genes’ MCI scores were larger than their UMFI scores. This could be due to a multitude of factors. First, we hypothesize that MCI can give incorrectly high scores when implemented with certain machine learning models, such as random forest and xgboost. In particular, random forests can be inaccurate when trained on only a few features. When trained on one or two features, the correlation between trees may increase dramatically, therefore giving a poor representation of an idealized evaluation function, \( \nu \) [10]. Because the maximizing subset in (1) may be very small or even empty, MCI is more susceptible to this fault compared to UMFI, which uses at least \(|F| - 1\) modified features for predictions when running Algorithm 1. Second, the evaluation function \( \nu(S) \) may not actually behave like \( I(Y:S) \) and hence may not obey supermodularity under mutual independence and invariance under duplicate information. This could lead to UMFI giving inaccurate results.

Figure 2: Median feature importance scores provided by (a) MCI, (b) UMFI with linear regression, and (c) UMFI with pairwise optimal transport, for each gene in the BRCA dataset after 200 iterations. Genes colored in blue are known to be associated with breast cancer while genes colored in grey are random permutations of randomly selected genes, which we assume to be unassociated with breast cancer. The first and third quantiles of the scores are visualized for each gene.
Table 1: The standardized interquartile range (SIQR), true positive rate (TPR), true negative rate (TNR), overall accuracy (OA), and the number of features for which feature importance can be calculated within 1, 15, and 60 minute(s) are displayed after running the methods on the BRCA data. The best results from each column are bolded.

| Method      | SIQR | TPR | TNR | OA  | @1min | @15min | @1hr |
|-------------|------|-----|-----|-----|-------|--------|------|
| MCI         | 6.6% | 1   | 0   | 0.20| 35    | 80     | 130  |
| UMFI (LR)   | 41.9%| 1   | 0.975| 0.98| 500   | 2000   | 4010 |
| UMFI (OT)   | 28.5%| 1   | 0.775| 0.82| 300   | 1500   | 3000 |

Figure 3: Computation time of MCI (dark red), MCI with the soft 2-size-submodularity assumption (pink), UMFI_OT (light blue), and UMFI_LR (dark blue) plotted against number of processed features from the BRCA data.

Third, our implementations of linear regression and optimal transport may have been inadequate for removing dependencies with minimal information loss. Experimental results in Supplement [F] indicate that optimal transport may have removed more information than necessary, and linear regression failed to completely remove dependencies.

4.3 Computational complexity

MCI must train and evaluate a model for each element of the power set of the feature set, which implies \(O(2^p)\) model trainings if there are \(p\) features. If the evaluation function \(\nu\) obeys soft \(k\)-size submodularity, then the maximizing subset has no more than \(k\) elements, which reduces the number of model trainings to \(O(p^{k+1})\) [15]. UMFI circumvents the exponential training time since we may immediately estimate the maximizing set of features after removing the dependencies of \(f\) from the feature set \(F\). To confirm the above statements, and to show that the extra model trainings required for MCI dominate the computation time for removing dependencies in UMFI, we ran a simple experiment. For a range of dataset sizes from the BRCA data, we evaluate the computation time for calculating the feature importance scores of all features using MCI and UMFI. We ran this experiment for a dataset with 5 features, and then slowly added features until our given time budget of 1 hour ran out. Once all 50 BRCA features were used, more features were randomly generated. All datasets had 571 observations. These experiments were run using an Intel Core i9-9980HK CPU 2.40GHz with 32GB of RAM. Code was parallelized in R, and 12 of the 16 available threads were used. Our experimental results are graphed in Figure [5] and indicate that UMFI is approximately superlinear, with UMFI_OT incurring more computational cost compared to UMFI_LR. Giving each method 1 hour to run, we found that MCI could process 19 features, MCI with the soft 2-size-submodularity assumption could process 130 features, UMFI_OT could process about 3000 features, and UMFI_LR could process about 4000 features (Table 1).
5 Conclusion

In this study, we built upon the framework laid out by marginal contribution feature importance (MCI) and introduced a new feature importance framework called ultra-marginal feature importance (UMFI). We showed that UMFI can be estimated directly for a feature $f$ by finding the largest information subset of the feature set $F$ that is independent of $f$. This signified the importance of removing dependencies for our method, which we explored using optimal transport and linear regression. When comparing UMFI and MCI, not only does UMFI satisfy the key axioms used to define MCI, but it also has the ability to obtain faster and more accurate estimates of marginal feature importance on real and simulated data, particularly in the presence of correlated interactions and unrelated features.

Throughout the work on this paper, several shortcomings appeared. First, we only considered two simple methods for removing dependencies, linear regression and pairwise optimal transport. Other methods certainly exist in the literature, including optimal transport with chaining [36], neural networks [13, 55], or principal inertial components [64]. Though our two methods performed fairly well on the real and simulated datasets in Section 4, optimal transport and linear regression failed to find representations of the data that were independent of the protected attribute when we tested the methods on a hydrology dataset with more shared information compared to BRCA [1] (Supplement G.4). However, neural nets or principal inertial components certainly could have given better results. Also, despite requiring significantly more computational cost, better methods for estimating the conditional CDF or using optimal transport with chaining should give better estimates for $S^*$ when implementing UMFI_OT. Second, UMFI scores are less robust than MCI since they have much higher variability, however, because of the significantly lower computational cost, UMFI can be run multiple times and averaged to increase robustness. Third, it is not clear precisely how close $\nu$ follows the invariance under duplicate information, monotonicity, and supermodularity under independence assumptions in practice. Finally, though UMFI can work for any arbitrary feature type, in this paper, we have only considered datasets with continuous explanatory variables.

In future work, we would like to test how well other methods, such as neural networks, pair with UMFI while further testing on a wider variety of random variable types such as binary, categorical, and ordinal features. Further, we would like to explore how well dependence can be removed and UMFI can be estimated on real data as the number of increases to sizes much larger than 50. It would also be beneficial to determine conditions that guarantee supermodularity of $\nu$ under mutual independence, as well as invariance under duplicate information, since these properties crucially enable the direct computation of UMFI via Algorithm 1. This can likely be done by investigating the relationship between the mutual information of a feature set and the prediction accuracy of a variety of models, so that we can quantify the approximation $\nu(S) \sim I(Y; S)$.

To reiterate, UMFI is a powerful tool to explain relationships in data using feature importance. We emphasise that UMFI is just a framework. A variety of other methods can be used to estimate the evaluation function including, but not limited to, XGBoost, neural networks, or Gaussian processes. Furthermore, new preprocessing techniques for dependence removal are still being developed in the AI fairness community, so these, in addition to other existing methods, can be used in future applications of UMFI for additional improvements.

We hope that UMFI will be a useful tool in a variety of disciplines including bioinformatics, ecology, earth sciences, and health science for discovering scientific processes and relationships hidden within data. Further, we hope our work provides further justification and encouragement for the development of better AI fairness preprocessing methods and better machine learning models that more closely resemble the idealized $\nu$. 


Appendix

A Mutual information

A.1 Properties of mutual information

We note that the mutual information $I(Y; S)$ for $S \in \mathcal{I}(F)$ satisfies all four assumptions that were made about the evaluation function $\nu : \mathcal{I}(F) \to \mathbb{R}^+$ in Section 3. First, we assumed that $\nu(S)$ obeys monotonicity: $\nu(S_1) \leq \nu(S_2)$ if $S_1 \subseteq S_2$. Indeed, $I(Y; S_1) \leq I(Y; S_2)$ if $S_1 \subseteq S_2$ [68]. Second, we assumed that $\nu(\emptyset) = 0$, which is satisfied by $\nu(\emptyset) = I(Y; \emptyset) = 0$ [61]. Third, we assumed that $\nu$ is invariant under duplicate information, meaning that $\nu(S_1) = \nu(S_1 \cup S_2)$ if $S_2 \subseteq S_1$. This is another property of mutual information [50]. Finally, we prove that $I(Y; :)$ obeys supermodularity under mutual independence.

**Theorem A.1.** (Supermodularity under mutual independence) Let $S, f, X$ be mutually independent random variables. Then, $I(Y; S, f, X) - I(Y; S, X) \geq I(Y; S, f) - I(Y; S)$ [71].

**Proof.**

\[
I(Y; S, f, X) - I(Y; S, X) = I(Y; S) + I(Y; X|S) + I(Y; f|S, X) - [I(Y; S) + I(Y; X|S)] \quad \text{(by chain rule)}
\]

\[
= I(Y; f|S, X) - I(Y; S, X; f) \quad \text{(by mutual independence)}
\]

\[
\geq I(Y; S; f) \quad \text{(by monotonicity of $I(\cdot; f)$)}
\]

\[
= I(Y; f|S) \quad \text{(by mutual independence)}
\]

\[
= I(Y; S, f) - I(Y; S) \quad \text{(by the chain rule for mutual information)}
\]

\[
\square
\]

A.2 Mutual information and feature importance

Let $F = \{f_1, \ldots, f_n\}$ be a set of features used to predict $Y$. As shown in Griffith and Koch [30], the mutual information $I(Y; F) = I(Y; f_1, \ldots, f_n)$ can be visualized using a partial information (PI) diagram [67]. We may interpret the mutual information shared between $Y$ and $F$ as a collection of non-negative pieces of information, whose sum forms $I(Y; F)$. Each of these pieces of information can be classified as unique, redundant, or synergistic (Figure 4). Unique information is the information about $Y$ that comes from only one feature and nowhere else. Redundant information is information about $Y$ that comes from a single feature, which can also be found elsewhere in $F$, i.e. from another feature or from a group of features. Synergistic information is information about $Y$ that cannot be extracted from a single feature, but is available when multiple features are considered.

PI-diagrams clarify how we would like a feature importance metric to behave when the goal is to explain the data in terms of the information that a predictive feature $f$ contains about the response $Y$. In order to accurately reflect the importance of $f$ towards predicting $Y$ in this sense, we submit that an accurate feature importance metric must properly gauge the unique, redundant and synergistic parts of $I(Y; F)$ that involve $f$.

1. The unique parts correspond to mutual information between $Y$ and the single feature $f$, which cannot be found elsewhere in $F$, either in other features or in any groups of features. For example, if $f, g \in F$ predict $Y = 3f + fg + g$, then the unique part for the feature importance of $f$ comes entirely from the term $3f$. Although the term $fg$ also helps predict $Y$, this arises from an interaction that also uses another feature $g$.

2. The redundant parts correspond to mutual information between $Y$ and $f$ that can also be found within other predictors or groups of predictors in $F$. For example, if $f, g \in F$ are highly correlated and help predict $Y$, then most of their respective feature importances would come from redundant information, as $f$ and $g$ share much of the same information.

3. The synergistic parts correspond to mutual information between $Y$ and $F$ that arise from interactions between $f$ and other features in $F$. For example, if $Y = fg$ where $f, g \in F$ are independent, then all of the feature importance of $f$ would come from synergistic information via the interaction $fg$. This would also hold for the feature importance of $g$. 

We note that the distinction between marginal and conditional methods for feature importance comes from their treatment of redundant information, i.e. their treatment of dependent features. A marginal method, like MCI or UMFI, should count all of the redundant information pertaining to \( f \) in \( I(Y; F) \) towards the feature importance of \( f \). Indeed, even though this information can be found elsewhere in the model, redundant information still constitutes part of the information that \( f \) shares about \( Y \) in the data. Conversely, a conditional approach, like conditional permutation importance (CPI), would count none of the redundant information towards the evaluation of a feature’s importance. This is because under a conditional framework, a feature’s importance is defined to be the additional predictive power supplied by \( f \) beyond that which is available from all other predictors.

Mutual information itself is a common choice in the context of feature selection. However, due to the computational cost and the limited number of observations available for the calculation of the high-dimensional joint probability density function, it is not practical to compute \( I(Y; S) \). For feature selection, users are only interested in the importance given to the top \( k \) features. Therefore, mutual information-based feature selection methods typically bypass the computation of \( I(Y; S) \) by instead studying the mutual information between the candidate feature and the response along with the mutual information between the candidate and the previously selected features. These methods are much less suitable for feature importance when the goal is to explain the data since interactions cannot be considered, which is why the prevalent approach is to train machine learning models to determine marginal feature importance.

Another connection between feature importance and mutual information comes from Louppe et al. who showed that when extremely randomized trees’ mean decrease in impurity (MDI) is used as a feature importance score, the MDI of a single feature converges to a quantity that is defined by conditional mutual information, as the number of trees and the number of observations goes to infinity. Also, the sum of the MDI scores across the feature set \( F \) converges to \( I(Y; F) \).

**A.3 Mutual information and machine learning evaluation functions**

The evaluation function \( \nu : \mathcal{P}(F) \to \mathbb{R}^+ \) for a machine learning model measures the usefulness of the features \( S \subseteq F \) towards predicting the response \( Y \). Thus, \( \nu \) should ideally mirror or at least closely approximate the mutual information \( I(Y; S) \). Direct relationships between mutual information and machine learning evaluation functions have been observed. For example, the Gini value is equivalent to the first order Taylor approximation of information entropy. The Gini impurity index is the central mechanism for choosing splits in random forests. Also, with some assumptions, mutual information and \( R^2 \) accuracy are related. Since mutual information can be expressed in terms of the linear correlation coefficient, if we assume the response and predictions are joint Gaussian and the
predictions are unbiased [20], we can approximate the mutual information between $Y$ and $F$ as:

$$I(Y; F) \geq I(Y; g(F)) = I(Y; \hat{Y}) = -\frac{1}{2} \log[1 - \rho^2(Y, \hat{Y})] = -\frac{1}{2} \log[1 - R^2].$$

Machine learning evaluation functions and mutual information have been equated many times in the feature importance literature. Covert et al. [21] demonstrated equivalence when the Bayes classifier is known and cross entropy loss is used. In a simple example, Catav et al. [14] used mutual information directly as the evaluation function. The connection between machine learning evaluation functions and mutual information was further used by Sutera et al. [61] to relate random forest feature importance with Shapely values.

## B Additional information about marginal contribution feature importance (MCI)

Two of the methods that are compared with MCI in Catav et al. [15] include ablation and bivariate association. Ablation methods determine feature importance based on the difference in accuracy between the full model and the full model without the feature of interest, i.e. $A_\nu(f) = \nu(F) - \nu(F \setminus f)$. Bivariate methods are among the most popular methods for genome-wide association studies [19, 25, 60]. In this case, the feature importance is given by the difference in the evaluation function of the model with just the feature of interest and the null model, i.e. $B_\nu(f) = \nu(f) - \nu(\emptyset)$. The three feature importance axioms proposed by Catav et al. [15] were partially motivated by the shortcomings of these two methods.

1. **Marginal contribution:** Ablation methods may underestimate the importance of features in cases where the correlation between features is high. In these scenarios, $\nu(F)$ may be approximately equal to $\nu(F \setminus f)$ even in cases where $f$ is highly related to the response. Because of this, the importance of a feature $I_{\nu}(f)$ should be at least as large as the importance given by ablation methods.

2. **Elimination:** Bivariate methods may underestimate the importance of features in cases where interactions exist between features. Many high-order interactions may be present in the data, so eliminating features from the feature set could prevent the detection of an important interaction. Thus, eliminating features from $F$ should only be able to decrease the feature importance $I_{\nu}(f)$.

3. **Minimalism:** Catav et al. [15] decided to impose the minimalism axiom so that MCI can be unique. If $I_{\nu}(f)$ satisfies the first two axioms, then multiplying $I_{\nu}(f)$ by any constant $\lambda > 1$ would not change this. The minimalism axiom helps disambiguate MCI from these trivial variations. We note that UMFI does not satisfy the minimalism axiom, but it is a non-trivial variation of MCI.

## C Additional information about ultra-marginal feature importance (UMFI)

Recall that ultra-marginal feature importance is given by

$$U_\nu(f) = \max_{S \in \mathcal{I}(F)} \nu(S \cup \{f\}) - \nu(S). \quad (4)$$

where an information subset $S \in \mathcal{I}(F)$ obeys the form $g(A)$ such that $A \subseteq F$.

**Theorem C.1.** $U_\nu$ satisfies the marginal contribution and elimination axioms from Catav et al. [15].

**Proof. Marginal Contribution:** Note that $F \setminus \{f\} \in \mathcal{I}(F)$. Hence,

$$U_\nu(f) = \max_{S \in \mathcal{I}(F)} \nu(S \cup \{f\}) - \nu(S) \geq \nu(F) - \nu(F \setminus \{f\}).$$

**Elimination:** Let $T \subseteq F$ and $\nu : F \setminus T \to \mathbb{R}^+$ be the restriction of $\nu$ on $F \setminus T$. Then, for $f \in F \setminus T$,

$$U_\nu(f) = \max_{S \in \mathcal{I}(F \setminus T)} \nu(S \cup \{f\}) - \nu(S) \leq \max_{S \in \mathcal{I}(F)} \nu(S \cup \{f\}) - \nu(S) = U_\nu(f)$$
We now present an alternative heuristic proof of Theorem 3.1 using PI diagrams.

**Theorem C.2.** The maximizing information subset $S^*$ for (4) is given by $S^* = \bigcup_{S \in \mathcal{I}(F) : S \in I^*} S$, which is the largest information subset of $F$ such that $S^* \in I^*$. 

**Proof.** For simplicity, we assume that $\nu(S) = I(Y; S)$. Therefore, we would like to find $S \in \mathcal{I}(F)$ that maximizes the expression $\nu(S \cup \{f\}) - \nu(S) = I(Y; S, f) - I(Y; S)$.

Let $f \in F$. To maximize the difference $I(Y; S, f) - I(Y; S)$, we may consider the PI diagrams for $I(Y; S, f)$ and $I(Y; S)$. It then becomes clear that $I(Y; S, f) - I(Y; S)$ corresponds precisely to the amount of unique and synergistic information that $f$ has about $Y$ when considering the partial model with predictors $(S, f)$. To maximize $f$’s unique information, $S$ should be chosen to be independent of $f$. Otherwise, $f$ would share redundant information with other features, which would take the place of $f$’s unique information. Then, $f$’s synergistic information is maximized when we maximize the number of possible interactions involving $f$. We conclude that $S$ must be chosen to be as large as possible, and that $S^*$ is the largest information subset of $F$ such that $S^* \in I^*$.

We emphasize that Theorem 3.1 relies on the use of a map $g$ such that the operation $g(F) = S^*$ optimally removes dependencies on the feature of interest $f$ from the feature set.

## D Additional information about other feature importance methods

Historically, feature importance methods were developed in the pursuit of scientific questions, but current research in this area typically focuses on model explainability or model optimization. Early forms of feature importance assessed the strength of the relationships between variables within animal biology or human psychology using methods such as the correlation coefficient [25], Spearman’s rank correlation coefficient [56], multiple linear regression [22], and partial correlation [70]. Although these methods are perfectly interpretable, they are inadequate for modelling and therefore explaining complex data, since they cannot quantify the unknown interactions between multiple features. To counteract this severe limitation, Breiman was instrumental with his introduction of variable importance within classification and regression trees [11]. Breiman seemed more concerned about the true strength of the relationships between the explanatory variables and the response, as he posited that a feature that is related to the response should be given some importance even if it does not appear in the final model [11]. However, starting with Breiman’s random forests, feature importance began to prioritize machine learning model explanation rather than data exploration. A good overview of the properties of some popular feature importance metrics is shown in Covert et al. [21].

### D.1 Disagreements with previous feature importance papers

**Claim #1:** Random forests permutation importance is a marginal method [31][58].

Refutation: Though the marginal and conditional divide is not as clear when considering random forests instead of linear regression [23], the experiments in Supplement [G.1] will show that random forest permutation importance is in the middle of the marginal and conditional framework. When a correlated feature is added to the data, the importance weights are divided equally among the correlated features. In marginal methods, the full shared importance is given to each correlated feature. In conditional methods (CPI), none of the shared importance is given to the correlated features. Thus, permutation importance is directly in the middle of these two methods.

**Claim #2:** Conditional feature importance is for causal studies [23].

Refutation: Contrary to the claim above, it has been argued by other authors that conditional methods are for prediction purposes while marginal methods are for explanatory or causal purposes [31]. While it is true that conditional importance measures can bare the same resemblance as conditional independence statements, which often form the basis of causal graphs, we agree more with the perspective of Grömping [31] for several reasons. First, any time that a feature importance metric is based on a single model with permuted values, the metric is focusing on interpreting the model rather than the data, and even interpreting the model can be misleading [32]. Second, discovering causes from observational data requires many assumptions that are hard to assume in large and complex datasets including no directed causal cycles, faithfulness, the response occurring before the predictors,
and no unmeasured confounders. Indeed neither marginal, nor conditional methods are well suited to make causal claims, but conditional methods such as conditional permutation importance certainly describe the model more than they describe the data.

Claim #3: A good feature importance method should incorporate some conditional and some marginal aspects [23, 31].

Refutation: To justify these claims, the authors cite Budescu [12] and Johnson and LeBreton [38]. In Johnson and LeBreton [38], they argue that individually analyzing marginal and conditional importance requires subjective interpretation, thus motivating the need for a single index that is descriptive of both perspectives. While this approach may be simpler to interpret, the separate consideration of marginal and conditional feature importance scores could give a user powerful information about the data. For example, by using both, one can quantify the redundant information that the feature set has about the response as well as each feature’s influence on the predictive power of the full model. If a third method is used that does not quantify interactions, perhaps the unique, redundant, and synergistic information from the partial information diagrams referenced in Section A.2 can each be quantified and compared. This may be an interesting project for future work.

Budescu [12] was concerned that current feature importance metrics depend on the subset of features used in a linear regression model. To counteract this, he introduced dominance analysis, which integrated both marginal and conditional importance. While the reasoning of Budescu [12] makes sense in the context of linear regression, the importance of features will certainly change depending on the other features in the feature set when interactions are considered. Thus, it is our position that even though the most suitable feature importance method usually depends on the problem the user seeks to answer, it is better to consider marginal importance, conditional importance, or both of them separately, than it is to seek a metric that mixes the marginal and conditional frameworks.

E Preprocessing methods for removing dependencies

Finding information preserving independent representations of our data is the central step of UMFI. These representations were first considered for AI fairness and privacy algorithms in order to give unbiased predictions in the face of sensitive attributes. For example, if one wants to remove the influence of race on recidivism likelihood predictions, preprocessing methods can be used to alter the original dataset such that the set of predictors are independent of race. In the following subsections, we discuss how optimal transport and linear regression can be used for finding these representations.

E.1 Optimal transport

Most of the results and methods explained in this section can be found in Johndrow and Lum [36]. To avoid confusion with the probability density function, we denote a feature \( X \) with respect to some \( F \) ∈ \( F \setminus Z \). We would like to remove the dependencies of \( Z \) from \( X \) with minimal information loss with respect to \( X \). To do so using optimal transport, we consider the Monge problem:

\[
 g_c(X, \tilde{X}) = \inf_{g : g(X) \sim \tilde{X}} \mathbb{E}[c(X, g(X))] = \inf_{g : g(X) \sim X} \int_{\mathbb{R}} c(x, g(x))d\mu(x). \tag{2.1.1}
\]

The quantity \( g_c(X, \tilde{X}) \) represents the transportation cost of moving \( X \) to \( \tilde{X} \) with respect to some cost function \( c \), and in our case, we desire \( \tilde{X} \bot \bot Z \). It is natural to use \( c(x, \tilde{x}) = d^2(x, \tilde{x}) \), where \( d \) is the Euclidean norm. The transportation cost is also given by the Wasserstein-\( q \) distance, \( g_c(X, \tilde{X}) = \mathcal{W}_q(X, \tilde{X}) \), defined below for one-dimensional distributions.

\[
 \mathcal{W}_q(X, \tilde{X}) = \int_0^1 |F^{\leftarrow}(p) - \tilde{F}^{\leftarrow}(p)|^q dp,
\]

where \( F \) and \( \tilde{F} \) are the cdfs of \( X \) and \( \tilde{X} \), and \( F^{\leftarrow}(p) = \sup_{x \in \mathbb{R}} F(x) \leq p \). It can be shown that given any continuous one dimensional distributions \( X \) and \( \tilde{X} \), the optimal transport map \( g : X \to \tilde{X} \) is given by \( g = \tilde{F}^{\leftarrow} \circ F \).

**Theorem E.1.** Let \( X \) be a r.v. with density \( f \) and cdf \( F \). Let \( \tilde{X} \) have cdf \( \tilde{F} \). Then \( g = \tilde{F}^{\leftarrow} \circ F \) is the minimizer to (2.1.1). Hence, \( g \) optimally transports \( X \) to \( \tilde{X} = \tilde{F}^{\leftarrow}(F(X)) \).
Proof. We show $\mathbb{E}[|X - g(X)|^q] = \int_0^1 |F^+(p) - \tilde{F}^+(p)|^q dp$ for $g = \tilde{F}^+ \circ F$

$$\mathbb{E}[|X - g(X)|^q] = \int_{-\infty}^{\infty} |x - \tilde{F}^+(F(x))|^q f(x) dx$$

$$= \int_{-\infty}^{\infty} |F^+(F(x)) - \tilde{F}^+(F(x))|^q f(x) dx = \int_0^1 |F^+(p) - \tilde{F}^+(p)|^q dp$$

We would ideally pick $\tilde{F}$ such that it minimizes the transportation cost $g_c(X, \tilde{X}) = g_c(X, \tilde{F}^+(F_{X|Z}(X)))$ across all cdfs $\tilde{F}$ in order to minimize information loss. However, in practice, the choice of $\tilde{F}$ does not matter much. In fact, as long as the support of $\tilde{F}$ is at least as large as the support of $F_X$, then any rank-based prediction rule, e.g. random forest, will be invariant to the choice of $\tilde{F}$. A standard choice for $\tilde{F}$ is $F_X$ so that we can recover the original quantiles of $X$.

Furthermore, $F_{X|Z}$ is not usually known and must be estimated from the data. For example, this can be done by splitting $Z$ into $N$ quantiles and using the empirical cdf $P(X \leq x_j | Z \in z_j$’s quantile). The ability of this method to remove dependencies on $Z$ from $X$ relies significantly on the accuracy of this estimate.

We may iterate Algorithm 2 over each feature in $F \setminus Z$ to obtain pairwise independence between the transported variables $\tilde{X}_j$ and $Z$. It is also possible to iterate Algorithm 2 via chaining to achieve mutual independence between the transformed variables $\tilde{X}_j$ and $Z$ [36] 2.4. However, this is computationally expensive, and pairwise independence should suffice for an accurate UMFI score, as will be explored further in Section E. Step 2 of Algorithm 1 can therefore be implemented with Algorithm 3.

Algorithm 3: Algorithm for estimating $S^*$ via pairwise optimal transport

Require: $Z = [z_1, ..., z_n], X_j = [x_{j1}, ..., x_{jn}]$ for $X_j$ in $F \setminus Z$

$S^* = \emptyset$

for $X_j$ in $F \setminus Z$

$\tilde{X}_j = \text{output of Algorithm 2 with } X_j$ and $Z$

add $\tilde{X}_j$ to $S^*$

end for

return $S^*$
In other words, we may estimate the largest information subset independent of $Z$ as:

$$S^* = \{F_{X|Z}^{-1}(F_{X|Z}(X)) : X \in F \setminus Z \} \in \mathcal{I}(F).$$

### E.2 Linear regression

The most basic method for removing dependencies is linear regression. Even though it is quite simple, it can be shown to be optimal with a few assumptions (Theorem E.3). This preprocessing technique is implemented in the popular Python package `fairlearn` [9, 47].

To reiterate, removing dependencies requires methods to make a feature or set of features $S$ independent of a protected attribute $f$, while keeping as much of the original information as possible. The overarching idea is that under the assumption that the residuals and the protected attribute are jointly Gaussian, we may show that the residuals can be utilized as a representation of $S$, which is independent of $f$.

**Theorem E.3.** Assuming no intercept term, if one specifies a simple linear regression model with

$$Y = \beta X + \epsilon$$

and $X$ and $\epsilon$ are joint normally distributed, then (1) $\epsilon \perp \perp X$ and (2) $\epsilon$ is correlated with $Y$ unless $Y$ can be completely predicted from $X$.

**Proof.** (1) From the normal equations, the definition of covariance, and the fact that $E[\epsilon] = 0$, it follows that

$$Cov(X, \epsilon) = E[X^T \epsilon] - E[\epsilon]E[X] = E[X^T \epsilon] = E[X^T(Y - X\beta)]$$

$$= E[X^T(Y - X(X^TX)^{-1}X^TY)] = E[X^TY - X^T X(X^TX)^{-1}X^TY] = E[X^TY - X^TY] = 0$$

Then, since $X$ and $\epsilon$ are jointly normal, $X \perp \perp \epsilon$.

(2) From the definition of the response variable $Y$ and the distributive property for covariances we know

$$Cov(Y, \epsilon) = Cov(X\beta + \epsilon, \epsilon) = \beta Cov(X, \epsilon) + Cov(\epsilon, \epsilon) = Var(\epsilon).$$

Thus, in step 2 the algorithm for UMFI (Algorithm [1]), we can estimate the largest information subset of $F$ that is independent of the feature $f$ as:

$$S^* = \{\epsilon_i = X_i - \beta_{0,i} - \beta_{1,i}f : X_i \in F \setminus f \}.$$
Figure 5: The relative mutual information $I_{rel}(f_i; F \setminus \{f_i\})$ between the $i$th feature in the BRCA dataset and all other features is plotted (black) for each $i \in \{1, 2, \ldots, 50\}$. The relative mutual information $I_{rel}(f_i; S^*)$ between the $i$th feature and all other features after preprocessing with linear regression (red) and optimal transport (blue) is also plotted. Relative mutual information is measured by random forest’s OOB-$R^2$.

Given $S^*$ \cite{55}. For example, if $I(f; S^*) = 0$, as is desired, then the optimal predictor of $f$ will have zero accuracy given $S^*$. If the opposite is true and $S^*$ contains all of the information from $f$, then an optimal predictor of $f$ should be able to perfectly predict $f$ from the given information in $S^*$. In the following experiments, we assume that random forests can form the optimal predictor of $f$ given $S^*$. We use the OOB-$R^2$ value coming from the random forest model to give a relative measure of the mutual information between $f$ and the transformed dataset $S^*$.

We used the BRCA dataset with 50 features to test the ability of optimal transport and linear regression to remove dependencies \cite{21, 14}. All 50 features are continuous and the response is categorical. For each individual feature, we first use random forest OOB-$R^2$ to give a relative measure of the mutual information $I(f; F \setminus \{f\})$ between the feature of interest $f$ and the other 49 features. We then consider the case where the 49 remaining features are preprocessed to have dependencies on $f$ removed via linear regression or pairwise optimal transport. Similarly, random forest’s OOB-$R^2$ is used to give a relative measure of $I(f; S^*)$.

The results are plotted in Figure 5. It is clear that the raw data (black line) shares considerable information across features. Most features can be predicted from the other untransformed features with an accuracy of $R^2 > 0.2$ and many can even be predicted with accuracies over 0.4. Since the data has extremely nonlinear dependencies between features, simple linear regression is unable to remove all the mutual information between the protected attributes and the rest of the features. Indeed, the data cannot be approximated with multivariate Gaussians. Conversely, pairwise optimal transport can successfully remove most of the mutual information present in the data. For all 50 features in the dataset, $f$ cannot be predicted successfully by random forest (OOB-$R^2 = 0$) from the other features after $F \setminus f$ is transformed with pairwise optimal transport.

F.2 Distortion

Not only do we require that the transformed features are independent of the feature of interest, but we also require that as much of the information present in the original data is preserved in the transformed data. To measure the amount of distortion imposed on the original data, we measure the dependence between the original and perturbed data using the maximal information coefficient \cite{39}. For each feature in the BRCA dataset with 50 features \cite{21, 14}, the information from the current feature is removed from all other features with either linear regression or pairwise optimal transport (Figure 6).
Figure 6: Cell \((i,j)\) indicates how similar the \(j^{th}\) variable in the BRCA dataset is compared to its transformation via pairwise optimal transport or linear regression with respect to feature \(i\). This is measured with the maximal information coefficient, which is comparable to \(R^2\). To make the plots more clear and accessible, only the first 15 features are shown.

Linear regression does not distort the transformed features in most cases. The dependence between the original and perturbed features usually remains near 1, though the dependence does go as low as 0.42 in one case (Figure 6). While linear regression transformed these features with minimal distortion, these results are moot since linear regression failed to remove the original dependencies in a significant way, which was the main goal of the method (Figure 5).

Compared to linear regression, pairwise optimal transport has a much more sizable effect on the distorted features, though this may have been necessary to completely remove dependence. The dependence between original and perturbed features mostly ranges from 0.6-0.9, though some are as low as 0.37 (Figure 6). While only the first 15 features are shown, the results are similar for the other 35 features.

G Further feature importance experiments

This section is comprised of additional experiments performed on the simulated data introduced in Section 4.1, the BRCA dataset with permuted random genes, the original BRCA dataset with unpermuted random genes \([62, 21, 15]\), and the CAMELS hydrology dataset \([1]\). MCI and UMFI used either random forests or extremely randomized trees \([10, 27]\). Both of these are implemented using the \texttt{ranger} R package \([69]\). Ablation, permutation importance, and conditional permutation importance used random forests. Ablation and permutation importance were implemented with the \texttt{ranger} R package \([69]\), while conditional permutation importance was implemented with the \texttt{randomForest} and \texttt{permimp} packages \([24, 43]\). All experiments were run in Microsoft R Open Version 4.0.2 \([48]\).

G.1 Experiments on simulated data

We repeat our previous experiments on simulated data from Section 4.1 to test how ablation, permutation importance (PI), and conditional permutation importance (CPI) behave in the presence of correlated interactions (Section G.1.1), correlations (Section G.1.2), and non-linear interactions (Section G.1.3). Further, we test how using extremely randomized trees instead of random forests for MCI and UMFI changes the results of the same simulation experiments. All results are plotted in Figure 7. Although boosting methods such as XGBoost \([17]\) could have been implemented for these experiments, XGBoost requires greater care when optimizing hyperparameters, so we chose to use extremely randomized trees instead, which is faster than random forests and provides similarly
good predictions [27]. Both random forests and extremely randomized trees are not sensitive to hyperparameters [52].

G.1.1 Correlated interactions

The first experiment considers the case where two correlated variables, \( x_1 \) and \( x_2 \), interact together in the response \( Y \). Thus, as explained in Section 4.1.1 we should expect \( x_1 \) and \( x_2 \) to have more importance compared to \( x_3 \) and \( x_4 \). Figure 7a shows that ablation, PI, and CPI all correctly weigh the importance of \( x_1 \) and \( x_2 \) higher relative to \( x_3 \) and \( x_4 \). The only notable difference is that the ablation method attributes an additional \( \sim 3\% \) importance to each of \( x_1 \) and \( x_2 \) compared to PI, CPI, MCI, and UMFI (Figure 7a).

When tested with extremely randomized trees instead of random forests, the correlated interaction simulation experiment results (Figure 7c) for MCI and UMFI are similar to the earlier results shown in Figure 1a. MCI gave slightly more importance to \( x_1 \) and \( x_2 \) compared to \( x_3 \) and \( x_4 \), though the differences are seemingly insignificant. On the other hand, both UMFI methods gave significantly more importance to \( x_1 \) and \( x_2 \) compared to \( x_3 \) and \( x_4 \), as expected.

G.1.2 Correlations

The second experiment tests how the metrics allocate importance to correlated features. As explained in Section 4.1.2, \( x_1 \) and \( x_2 \) should remain around the same relative importance, and \( x_3 = x_1 + \epsilon \), should have just slightly less importance compared to \( x_1 \) and \( x_2 \). Figure 7b indicates that CPI and ablation give near zero importance to the two heavily correlated features \( x_3 \) and \( x_3 \). This aligns with the discussion about conditional feature importance methods in Section 4.1.2 since these methods base their scores on the importance of a feature conditioned on all other variables. Ablation performs similarly to CPI in this test, albeit with slightly less drastic results. Finally, we see that PI splits the importance detected from \( x_1 \) and \( x_3 \) proportionally across both features. This shows that PI is in between the marginal and conditional approaches. The marginal approach allocates all of the redundant information to the feature. The conditional approach allocates none of the redundant information to the feature. PI splits the redundant information proportionally across the relevant correlated features.

When tested with extremely randomized trees, the correlation simulation experiment results (Figure 7d) for MCI and UMFI are slightly higher compared to the experiment with random forests in Figure 1b. MCI works well, though it still gives some non-zero importance to \( x_4 \). With random forests, the relative importance of \( x_1 \) was usually above 5\%, but with extremely randomized trees, the relative importance dropped below 5\%. The performance of UMFI with linear regression got slightly worse as now the importance of \( x_1 \) is slightly greater than that of \( x_2 \) on average. The performance of UMFI with optimal transport changed for the better and now the importance of \( x_1 \) and \( x_2 \) are almost identical which was not true before. In this experiment, UMFI_OT performed the best.

G.1.3 Non-linear interactions

The last experiment on simulated data handles the case where two variables, \( x_1 \) and \( x_2 \), interact in a non-linear way in the response \( Y \). As explained in Section 4.1.3, we should expect \( x_1 \) and \( x_2 \) to each contribute about half of the total importance, while \( x_3 \) and \( x_4 \) should be deemed completely irrelevant. Like in the first experiment about correlated interactions, Figure 7c shows that ablation, PI, and CPI all provide accurate scores.

When tested with extremely randomized trees, the non-linear interactions simulation experiment results for MCI and UMFI, shown in Figure 7c, remain mostly unchanged compared to the results from the experiment with random forests given in Figure 1c.

G.2 BRCA experiments with known ground-truth feature importance

The following experiments are performed on the BRCA dataset with 571 patients, each with one of four breast cancer subtypes, and 50 continuous predictor genes. The experiments use the same setting as in [4,2] where the 40 randomly chosen genes are also permuted so that the ground-truth feature importances are known. We observed that the overall classification accuracy of random forests for this dataset was 0.76.
G.2.1 Running 5000 iterations of UMFI

The original BRCA experiment conducted in Section 4.2 showed that UMFI_LR and UMFI_OT performed impressively on real data, providing significantly more accurate feature importance scores than MCI after 200 iterations of the experiment. Both UMFI_LR and UMFI_OT correctly gave high importance to the ten BRCA-associated genes, while giving zero median importance to about 80% of the unassociated genes. Additionally, in an overnight study spanning less than ten hours, UMFI_LR and UMFI_OT displayed ideal results after running 5000 iterations of the BRCA experiment. As shown in Figure 9, both implementations of UMFI achieve 100% overall accuracy by giving high importance to the ten BRCA-associated genes and zero median feature importance to all 40 unassociated genes. These results indicate that UMFI’s relatively low computational cost can be leveraged via aggregation to achieve superior performance on complex data within a reasonable time budget.

G.2.2 Ablation, PI, and CPI

We also test the quality and robustness of other feature importance metrics including ablation, PI, and CPI, by running 200 iterations of the BRCA experiment from Section 4.2 for each method. Results are shown in Figure 9. Ablation importance scores are small and have large uncertainties compared to its median importance scores, which makes the scores impractical to interpret. Eight of the ten important genes are identified by ablation, but all other genes are given exactly zero median importance. All ten important genes are given non-zero importance by CPI, however, some randomly permuted genes are given more importance than some genes known to be important, such as CDK6. PI gave more reliable and stable results compared to ablation and CPI in this experiment, exhibiting similar performance to UMFI_LR and UMFI_OT from the analogous experiment shown in Figure 2. We note that PI assigned zero importance to 29 of the 40 unassociated genes, making its TNR of 0.725 slightly lower than UMFI in the analogous experiment from Section 4.2.
Figure 8: Median feature importance scores provided by (a) UMFI with linear regression, and (b) UMFI with pairwise optimal transport, for each gene in the permuted BRCA dataset after 5000 iterations. Genes colored in blue are known to be associated with breast cancer while genes colored in grey are random permutations of randomly selected genes, which we assume to be unassociated with breast cancer subtype. The first and third quantiles of the scores are visualized for each gene.

G.3 Experiments on unpermuted BRCA data

Additional BRCA experiments were performed on the original randomized genes, as done in Covert et al. [21] and Catav et al. [15]. We observed that the overall classification accuracy of random forests for this dataset was 0.79.

Feature importance scores on this dataset were first computed with MCI, UMFI_LR, and UMFI_OT over 100 iterations, as shown in Figure [10]. The ordering of the BRCA associated genes is fairly similar across MCI and both UMFI methods. BCL11A and SLC22A5 are always the top two features and TEX14 is always the least important BRCA associated gene. While there are clear similarities in the results of all methods, the glaring difference is the number of features given zero importance. While MCI gives non-zero median importance to all 50 features, 14 features are given zero median importance by UMFI with linear regression, and ten features are given zero median importance by UMFI with pairwise optimal transport. It is unlikely that all 40 randomly selected genes, which have not shown any association with breast cancer in previous studies, share information about breast cancer, so in this respect, we conclude that UMFI performs better than MCI.

Feature importance scores on the unpermuted BRCA dataset were also computed with ablation, CPI, and PI over 100 iterations, as shown in Figure [11]. When also considering these results, we observe that MCI, UMFI, and PI give similar importance scores, while ablation and CPI performed significantly worse. Once again, ablation’s high variance hampers its interpretability. Meanwhile, CPI gave by far the highest importance to SLC25A1, which is not known to have any association with breast cancer. In the results of MCI, UMFI, and PI, BCL11A is the most important while CST9L is always among the most important non-BRCA associated genes. Contrary to this, ablation and CPI
Figure 9: Median feature importance scores provided by (a) ablation, (b) permutation importance, and (c) conditional permutation importance, for each gene in the permuted BRCA dataset after 200 iterations. Genes colored in blue are known to be associated with breast cancer while genes colored in grey are random permutations of randomly selected genes, which we assume to be unassociated with breast cancer subtype. The first and third quantiles of the scores are visualized for each gene.
Figure 10: Median feature importance scores provided by (a) MCI, (b) UMFI with linear regression, and (c) UMFI with pairwise optimal transport, for each gene in the unpermuted BRCA dataset after 100 iterations. Genes colored in blue are associated with breast cancer while genes colored in grey are randomly selected genes. The first and third quantiles of the scores are visualized for each gene.

G.3.1 Computational complexity

We compare the computational complexity of UMFI and MCI against the other feature importance methods that were explored in this section: ablation, PI, and CPI. To do so, we ran 10 iterations of the BRCA experiment, which has 50 features, each with 571 observations. We recorded the average time for each method to compute feature importance for 5, 10, 15, 20, 25, 30, 35, 40, 45, and 50 features. Figure 12 shows that PI is the fastest method, processing 50 features in 50 milliseconds on average, followed by ablation (50 features in 1.8 seconds), UMFI (50 features in 3 seconds when parallelized), CPI (50 features in 30 seconds), and finally MCI (50 features in 205 seconds).
Figure 11: Median feature importance scores provided by (a) ablation, (b) permutation importance, and (c) conditional permutation importance, for each gene in the unpermuted BRCA dataset after 100 iterations. Genes colored in blue are associated with breast cancer while genes colored in grey are randomly selected genes. The first and third quantiles of the scores are visualized for each gene.

G.4 Experiments on hydrology data

The final experiments for this study were conducted on a large-sample hydrology dataset called CAMELS [1]. This dataset records catchment averaged climate, soil, geology, topography, and land cover characteristics for 643 catchments across the contiguous United States. With these, there are 29 continuous explanatory variables. The response variable is averaged yearly streamflow, which is also continuous. Extremely randomized trees were used for these experiments for an additional comparison of MCI and UMFI on real data. The OOB-$R^2$ accuracy of extremely randomized trees on the full data is 0.91.

As can be seen in Figure 13, which is analogous to Figure 5 in Section F, linear regression and pairwise optimal transport both fail to completely remove dependencies from the CAMELS dataset. This could be due to the fact that each feature is almost completely dependent on the other explanatory features ($R^2 \geq 0.65$).
Figure 12: The average computation time for each method to process \( n \) features over 10 iterations of the original BRCA data is plotted for each \( n \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\} \).

Figure 13: The relative mutual information \( I_{rel}(f_i; F \setminus \{f_i\}) \) between the \( i \)th feature in the CAMELS dataset and all other features is plotted (black) for each \( i \in \{1, 2, \ldots, 30\} \). The relative mutual information \( I_{rel}(f_i; S^*_i) \) between the \( i \)th feature and all other features after preprocessing with linear regression (red) and optimal transport (blue) is also plotted. Relative mutual information is measured by random forest’s OOB-\( R^2 \).

The feature importance scores indicated in Figure 14 show that mean precipitation and aridity index are the features with the strongest relationships with mean annual streamflow. Geology and soil attributes such as bedrock permeability and soil porosity are always among the least important features. These conclusions are in line with previous studies [2, 35].
Figure 14: Median feature importance scores provided by (a) MCI, (b) UMFI with linear regression, and (c) UMFI with pairwise optimal transport, for each explanatory variable in the CAMELS dataset, taken after 100 iterations. The first and third quantiles of the scores are visualized for each feature.
References

[1] Nans Addor, Andrew J Newman, Naoki Mizukami, and Martyn P Clark. The camels data set: catchment attributes and meteorology for large-sample studies. Hydrology and Earth System Sciences, 21(10):5293–5313, 2017. [9] [18] [24]

[2] Nans Addor, Grey Nearing, Cristina Prieto, AJ Newman, Nataliya Le Vine, and Martyn P Clark. A ranking of hydrological signatures based on their predictability in space. Water Resources Research, 54(11):8792–8812, 2018. [2] [5] [25]

[3] Ahmed Al-Ani, Mohamed Deriche, and Jalel Chebil. A new mutual information based measure for feature selection. Intelligent Data Analysis, 7(1):43–57, 2003. [11]

[4] Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(4):1059–1086, 2020. [1]

[5] Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on neural networks, 5(4):537–550, 1994. [11]

[6] Adrián Bazaga, Dan Leggate, and Hendrik Weisser. Genome-wide investigation of gene-cancer associations for the prediction of novel therapeutic targets in oncology. Scientific reports, 10(1):1–10, 2020. [2]

[7] Mohamed Bennasar, Yulia Hicks, and Rossitza Setchi. Feature selection using joint mutual information maximisation. Expert Systems with Applications, 42(22):8520–8532, 2015. [11]

[8] Jian Bi. A review of statistical methods for determination of relative importance of correlated predictors and identification of drivers of consumer liking. Journal of Sensory Studies, 27(2):87–101, 2012. [2]

[9] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. Microsoft, Tech. Rep. MSR-TR-2020-32, 2020. [2] [4] [16]

[10] Leo Breiman. Random forests. Machine learning, 45(1):5–32, 2001. [1] [2] [4] [7] [18]

[11] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. Classification and regression trees. Routledge, 2017. [13]

[12] David V Budescu. Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. Psychological bulletin, 114(3):542, 1993. [11] [14]

[13] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. Advances in neural information processing systems, 30, 2017. [2] [9]

[14] Amnon Catav, Boyang Fu, Jason Ernst, Sriram Sankararaman, and Ran Gilad-Bachrach. Marginal contribution feature importance--an axiomatic approach for the natural case. arXiv preprint arXiv:2010.07910, 2020. [1] [5] [12] [17]

[15] Amnon Catav, Boyang Fu, Yazeed Zoabi, Ahuva Libi Weiss Meilik, Noam Shomron, Jason Ernst, Sriram Sankararaman, and Ran Gilad-Bachrach. Marginal contribution feature importance - an axiomatic approach for explaining data. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 1324–1335. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/catav21a.html [2] [3] [5] [6] [8] [12] [18] [21]

[16] Hugh Chen, Joseph D Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data? arXiv preprint arXiv:2006.16234, 2020. [1]

[17] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4):1–4, 2015. [18]
[18] Shay Cohen, Gideon Dror, and Eytan Ruppin. Feature selection via coalitional game theory. *Neural Computation*, 19(7):1939–1961, 2007. [1]

[19] Wellcome Trust Case Control Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661, 2007. [12]

[20] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006. ISBN 0471241954. [3][12]

[21] Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223, 2020. [1][2][3][5][6][12][15][17][18][21]

[22] Richard B Darlington. Multiple regression in psychological research and practice. *Psychological bulletin*, 69(3):161, 1968. [13]

[23] Dries Debeer and Carolin Strobl. Conditional permutation importance revisited. *BMC bioinformatics*, 21(1):1–30, 2020. [1][13][14]

[24] Dries Debeer, Torsten Hothorn, Carolin Strobl, and Maintainer Dries Debeer. Package ‘permimp’. 2021. [18]

[25] Douglas F Easton, Karen A Pooley, Alison M Dunning, Paul DP Pharoah, Deborah Thompson, Dennis G Ballinger, Jeffery P Struwing, Jonathan Morrison, Helen Field, Robert Luben, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148):1087–1093, 2007. [12]

[26] Francis Galton. I. co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 45(273-279):135–145, 1889. [13]

[27] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006. [18][19]

[28] WA Gibson. Orthogonal predictors: A possible resolution of the hoffman-ward controversy. *Psychological reports*, 11(1):32–34, 1962. [2]

[29] David A Gill, Michael B Mascia, Gabby N Ahmadia, Louise Glew, Sarah E Lester, Megan Barnes, Ian Craigie, Emily S Darling, Christopher M Free, Jonas Geldmann, et al. Capacity shortfalls hinder the performance of marine protected areas globally. *Nature*, 543(7647):665–669, 2017. [2]

[30] Virgil Griffith and Christof Koch. Quantifying synergistic mutual information. *arXiv preprint arXiv:1205.4265*, 2012. [3][10][11]

[31] Ulrike Grömping. Variable importance in regression models. *Wiley interdisciplinary reviews: Computational statistics*, 7(2):137–152, 2015. [1][13][14]

[32] Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6):1–16, 2021. [2][5][13]

[33] Aleks Jakulin and Ivan Bratko. Quantifying and visualizing attribute interactions: An approach based on entropy. 2003. [3]

[34] Joseph Janssen and Ali A Ameli. A hydrologic functional approach for improving large-sample hydrology performance in poorly gauged regions. *Water Resources Research*, 57(9):e2021WR030263, 2021. [4][5]

[35] Florian U Jehn, Konrad Bestian, Lutz Breuer, Philipp Kraft, and Tobias Houska. Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. *Hydrology and Earth System Sciences*, 24(3):1081–1100, 2020. [25]
[36] James E Johndrow and Kristian Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. The Annals of Applied Statistics, 13(1):189–220, 2019. [2][3][4][9][14][15]

[37] Pål V Johnsen, Signe Riemer-Sørensen, Andrew Thomas DeWan, Megan E Cahill, and Mette Langhaas. A new method for exploring gene–gene and gene–environment interactions in gwas with tree ensemble methods and shape values. BMC bioinformatics, 22(1):1–29, 2021. [2]

[38] Jeff W Johnson and James M LeBreton. History and use of relative importance indices in organizational research. Organizational research methods, 7(3):238–257, 2004. [14]

[39] Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. Proceedings of the National Academy of Sciences, 111(9):3354–3359, 2014. [17]

[40] Gunnar König, Timo Freiesleben, Bernd Bischl, Giuseppe Casalicchio, and Moritz Grosse-Wentrup. Decomposition of global feature importance into direct and associative components (dedact). arXiv preprint arXiv:2106.08086, 2021. [2]

[41] Ken Lau, Chandra Nair, and David Ng. A mutual information inequality and some applications. [10]

[42] Edward Le, Ali Ameli, Joseph Janssen, and John Hammond. Snow persistence explains stream high flow and low flow signatures with differing relationships by aridity and climatic seasonality. Hydrology and Earth System Sciences Discussions, pages 1–22, 2022. [5]

[43] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. R news, 2(3):18–22, 2002. [18]

[44] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees. Advances in neural information processing systems, 26, 2013. [11]

[45] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30, 2017. [1]

[46] Alexander Marx, Arthur Gretton, and Joris M Mooij. A weaker faithfulness assumption based on triple interactions. In Uncertainty in Artificial Intelligence, pages 451–460. PMLR, 2021. [5]

[47] Matthijs, Vincent Warmerdam, and ManyOthers. scikit-fairness. scikit-fairness. https://github.com/koaning/scikit-fairness, 2019. [16]

[48] R Core Team Microsoft. Microsoft R Open. Microsoft, Redmond, Washington, 2017. URL https://mran.microsoft.com/. [4][18]

[49] Christoph Molnar. Interpretable machine learning. Lulu.com, 2020. [2]

[50] Christoph Molnar, Gunnar König, Bernd Bischl, and Giuseppe Casalicchio. Model-agnostic feature importance and effects with dependent features—a conditional subgroup approach. arXiv preprint arXiv:2006.04628, 2020. [1][2]

[51] Alena Orlenko and Jason H Moore. A comparison of methods for interpreting random forest models of genetic association in the presence of non-additive interactions. BioData mining, 14(1):1–17, 2021. [5]

[52] Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. Tunability: importance of hyperparameters of machine learning algorithms. The Journal of Machine Learning Research, 20(1):1934–1965, 2019. [19]

[53] Lennart Schmidt, Falk Heße, Sabine Attinger, and Rohini Kumar. Challenges in applying machine learning models for hydrological inference: A case study for flooding events across germany. Water Resources Research, 56(5):e2019WR025924, 2020. [2]

[54] Lloyd S Shapley. A value for n-person games, contributions to the theory of games, 2, 307–317, 1953. [11]
[55] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 2164–2173. PMLR, 2019.

[56] Charles Spearman. "general intelligence" objectively determined and measured. 1961.

[57] Lina Stein, Martyn P Clark, Wouter JM Knoben, Francesca Pianosi, and Ross A Woods. How do climate and catchment attributes influence flood generating processes? a large-sample study for 671 catchments across the contiguous usa. Water Resources Research, 57(4):e2020WR028300, 2021.

[58] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. BMC bioinformatics, 9(1):1–11, 2008.

[59] Jiancheng Sun, Yunfan Hu, Zhinan Wu, Huimin Niu, and Si Chen. Task-oriented snapshot network construction of stock market. In International Conference on Intelligent Computing, pages 3–11. Springer, 2021.

[60] Shanwen Sun, Benzhi Dong, and Quan Zou. Revisiting genome-wide association studies from statistical modelling to machine learning. Briefings in Bioinformatics, 22(4):bbaa263, 2021.

[61] Antonio Sutera, Gilles Louppe, Van Anh Huynh-Thu, Louis Wehenkel, and Pierre Geurts. From global to local mdi variable importances for random forests and when they are shapley values. Advances in Neural Information Processing Systems, 34, 2021.

[62] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. Contemporary oncology, 19(1A):A68, 2015.

[63] Hao Wang and Flavio P Calmon. An estimation-theoretic view of privacy. In 2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pages 886–893. IEEE, 2017.

[64] Hao Wang, Lisa Vo, Flavio P Calmon, Muriel Médard, Ken R Duffy, and Mayank Varia. Privacy with estimation guarantees. IEEE Transactions on Information Theory, 65(12):8025–8042, 2019.

[65] Hui Wang, David A Bennett, Philip L De Jager, Qing-Ye Zhang, and Hong-Yu Zhang. Genome-wide epistasis analysis for alzheimer’s disease and implications for genetic risk prediction. Alzheimer’s research & therapy, 13(1):1–13, 2021.

[66] Pengfei Wei, Zhenzhou Lu, and Jingwen Song. Variable importance analysis: a comprehensive review. Reliability Engineering & System Safety, 142:399–432, 2015.

[67] Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. arXiv preprint arXiv:1004.2515, 2010.

[68] Edward Witten. A mini-introduction to information theory. arXiv preprint arXiv:1805.11965, 2018.

[69] Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. arXiv preprint arXiv:1508.04409, 2015.

[70] Sewall Wright. Correlation and causation. 1921.

[71] Lee H Wurm and Sebastiano A Fiscarco. What residualizing predictors in regression analyses does (and what it does not do). Journal of memory and language, 72:37–48, 2014.

[72] Jian-Bo Yang and Chong-Jin Ong. An effective feature selection method via mutual information estimation. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42(6):1550–1559, 2012.

[73] Biao Ye, Xu He, Shuo Sun, Rundong Wang, Rui Yin, Xu Chi, Chee Keong Kwoh, et al. Feature shapley: A general framework to discovering useful feature interactions. 2021.

[74] Ye Yuan, Liji Wu, and Xiangmin Zhang. Gini-impurity index analysis. IEEE Transactions on Information Forensics and Security, 16:3154–3169, 2021.