# Ultra-marginal Feature Importance

**Anonymous Author(s)**
Affiliation
Address
email

## 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 [44], 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 define feature importance, so choosing an appropriate method strongly depends on the questions that the user seeks to answer [42, 19, 9]. The most widely referenced distinction between feature importance methods is the conditional versus marginal divide, which defines the two extremes of feature importance methods [15]. This division is sometimes also framed as true to the model versus true to the data [11]. The distinction between these two types of methods is only evident in the presence of correlated features. Indeed, if all explanatory features in a dataset are independent, conditional and marginal importances are exactly the same [15, 30]. 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 [19], seeking to avoid the tendency of conditional methods to explain the model and give correlated features too little importance [11]. For example, if a scientist wants to build a database of genes associated with some disease, then a marginal approach could be preferred [15, 9, 11]. 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 [19, 9].

Recently, feature importance methods such as Shapely-values [33, 12, 26], SAGE [14], accumulated local effects (ALE) [3], permutation importance [7], and conditional permutation importance (CPI) [15] have been used in high-impact journal papers by scientists who want to explain the mechanisms within data [2, 4, 35, 24, 32, 17]. However, these methods do not adequately account for correlated

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

features and feature interactions. ALE can only easily show first order effects [29], 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 [14]. 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 [20]. 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 [10, 14]. In particular, MCI was shown to have better quality and robustness when compared to Shapely-values, SAGE, ablation, and bivariate methods [10]. 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. [10], 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. [10], 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 [16]. 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 [5, 45]. 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. Some examples of these techniques include linear regression [6], optimal transport [23], neural networks [8, 34], convex optimization [8], and principal inertial components [39]. Linear regression and optimal transport were implemented for UMFI in this paper.

Besides MCI and SAGE [10, 14], 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 [7] and conditional permutation importance [36] 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 [30]. However, this method only performs well when the response is dependent on just a few features [30]. Despite their popularity and myriad of variations, a thorough overview by Hooker et al. [20] 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. [20] 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. [10] defined the marginal contribution feature importance (MCI) of a feature $f \in F$ as

$$I_\nu(f) = \max_{S \subseteq F} \nu(S \cup \{f\}) - \nu(S). \tag{1}$$

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

$$\nu(S) = \min_{\hat{y}} \mathbb{E}[l(\hat{y}, Y)] - \min_{g \in G(S)} \mathbb{E}[l(g(S), Y)]. \tag{2}$$

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 properties are not strictly guaranteed when faced with real data and an imperfect model, they are reasonable and 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 $l$ [10].

Given a monotonically increasing evaluation function $\nu$, satisfying $\nu(\emptyset) = 0$, Catav et al. [10] 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. [10] is too restrictive. Instead, we introduce ultra-marginal feature importance (UMFI) by proposing a maximization over information subsets of the feature set $F$

$$U_\nu(f) = \max_{S \in \mathcal{I}(F)} \nu(S \cup \{f\}) - \nu(S). \tag{3}$$

We define the space of information subsets of $F$ by $\mathcal{I}(F) = \{g(A) : A \subseteq F \text{ and } g : \mathbb{R}^{|A|} \to \mathbb{R}^{|A|}\}$. We call these information subsets because $I(Y; g(F)) \leq I(Y; F)$ is a property of mutual information for any function $g$ [13, p. 33]. By considering transformations of feature subsets, we may use dependency removal methods from the AI fairness literature, which were developed to achieve unbiased predictions [34, 23]. Given this expanded framework, $U_\nu(f)$ obeys the marginal contribution and elimination axioms laid out by Catav et al. [10] (Supplement 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 $B \subset A$. Both of these conditions are satisfied when $\nu(S) = I(Y; S)$ ([18], 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 [14]. 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 3.

**Theorem 3.1.** *The maximizing information subset $S^*$ for* (3) *is given by* $S^* = \bigcup_{S \in \mathcal{I}(F): S \perp\!\!\!\perp F} S$*, which is the largest information subset of $F$ such that $S^* \perp\!\!\!\perp f$.*

*Proof.* First, suppose for contradiction that $S^*$ solves (3) and that $S^* \not\perp\!\!\!\perp f$. 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 $S^* \perp\!\!\!\perp f$.

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 $S^* = \bigcup_{S \in \mathcal{I}(F): S \perp\!\!\!\perp f} S$. □

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 [23] (UMFI_OT) and linear regression [6] (UMFI_LR) as methods to remove dependencies from the data. A detailed overview of these implementations is shown in Supplement D 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 $\nu$ [7]. We use the *ranger* R package to implement random forests with default hyperparameters and 100 for the number of trees [43]. All experiments were run in Microsoft R Open Version 4.0.2 [28]. Supplement G contains additional experiments comparing feature importance metrics.

## 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 [21, 22], so it is vital to test if feature importance metrics work as desired in these settings. Let $A, B, C, D, E, G \sim \mathcal{N}(0, 1)$. We consider

$$x_1 = A + B,\ x_2 = B + C,\ x_3 = D + E,\ x_4 = E + G$$
$$Y = x_1 + x_2 + sign(x_1 * x_2) + x_3 + x_4$$

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, $sign(x_1 * x_2)$. 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. [10]. Let $\epsilon \sim \mathcal{N}(0, 0.2)$. We consider

$$x_1, x_2, x_4 \sim \mathcal{N}(0, 1),\ x_3 = x_1 + \epsilon$$
$$Y = x_1 + x_2$$

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 $0$. Because of this, we conclude that UMFI with linear regression performs the best in this simulated scenario.

#### 4.1.3 Non-linear interactions

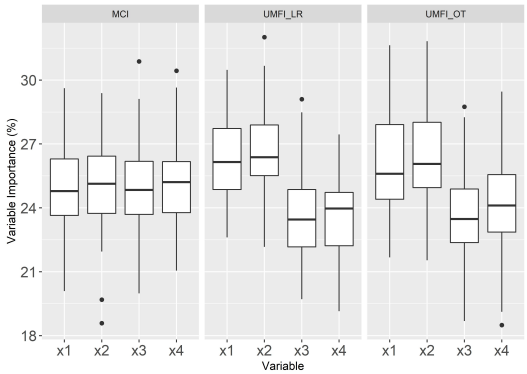
Interaction effects are common in many scientific disciplines where assessing feature importance is prevalent, including hydrology [22, 2, 25], genomics [9, 41, 31], and finance [46, 37]. So, as was done in Catav et al. [10], we assess the ability of MCI and UMFI to detect non-linear interaction effects 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 [27]. Let $\epsilon \sim Exp(1/\sqrt{2})$. We consider

$$x_1, x_2, x_3, x_4 \sim \mathcal{N}(0, 1)$$
$$Y = sign(x1 * x2) * \epsilon$$

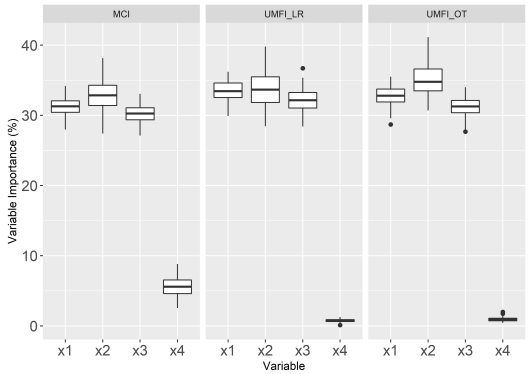
Ideally, the results of a feature importance metric would conclude that $x_1$ and $x_2$ have equally high importance while $x_3$ and $x_4$ have no importance. Figure 1c shows consistently good performance across all methods. Each method gave high relative importance scores to $x_1$ and $x_2$, while $x_3$ and $x_4$ received scores near $0$. 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

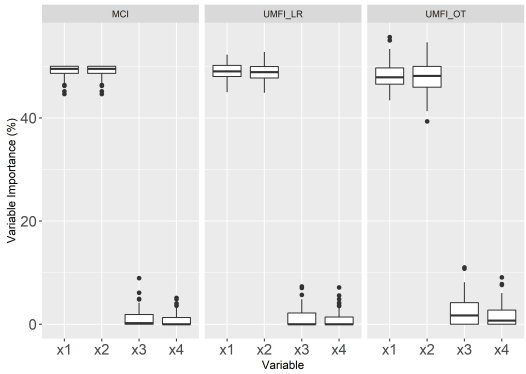
We use the same breast cancer (BRCA) classification dataset [38] used in previous feature importance studies including Catav et al. [10] and Covert et al. [14] to test the quality and robustness of UMFI on real data. The original data contains over $17,000$ genes and $571$ anonymous patients that have been diagnosed with one of $4$ breast cancer sub-types. We consider the same subset of $50$ genes as in Catav et al. [10] and Covert et al. [14] 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/Marginal-Contribution-Feature-Importance/tree/main/BRCA_dataset` (MIT License). In Catav et al. [10] and Covert et al. [14], 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




(a) Feature importance scores for 4.1.1




(b) Feature importance scores for 4.1.2




(c) Feature importance scores for 4.1.3

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.

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). Interestingly, the ordering of important features was similar across methods, with BCL11A and SLC22A5 always being the most important and TEX-14 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.

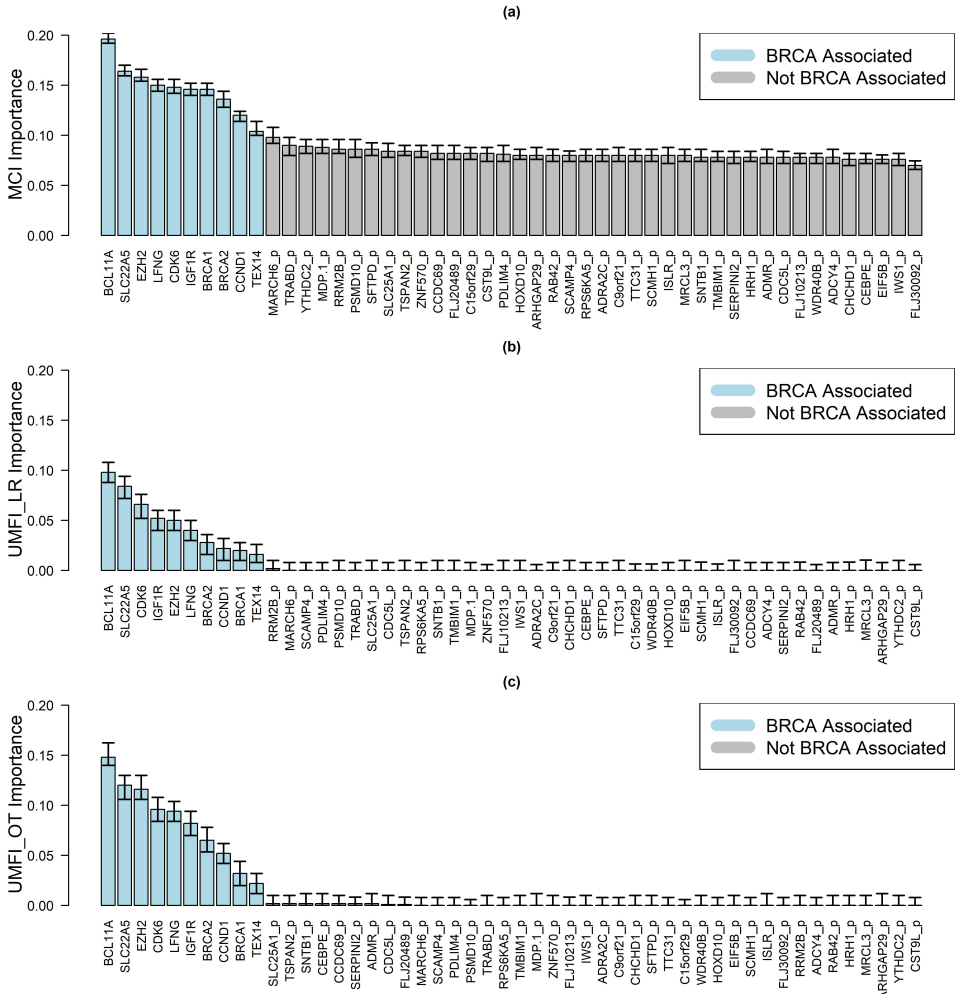



Figure 2: Median feature importance scores provided by (a) MCI, (b) UMFI with linear regression, and (c) UMFI with pairwise optimal transport, for each gene in the BRCA dataset after 200 iterations. Genes colored in blue are known to be associated with breast cancer while genes colored in grey are random 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.

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, $\nu$ [7]. 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. Third, our implementations of linear regression and optimal transport may have been inadequate for removing dependencies with minimal information loss. Experimental results in Supplement E indicate that optimal transport may have removed more information than necessary, and linear regression failed to completely remove dependencies.

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** | **0.83** | **0.98** | **500** | **2000** | **4010** |
| UMFI (OT) | 28.5% | **1** | 0.78 | 0.82 | 300 | 1500 | 3000 |

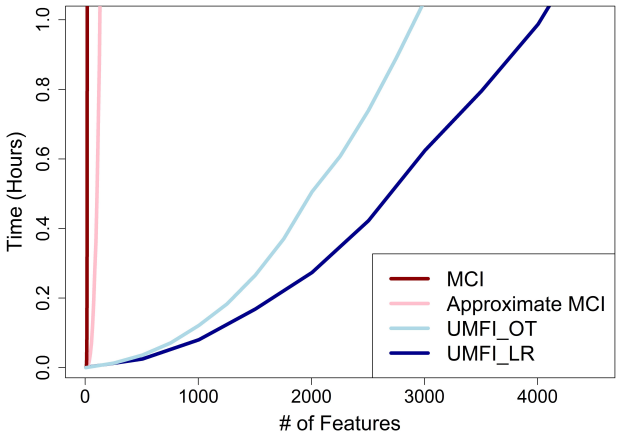



Figure 3: Computation time of MCI (dark red), MCI with the soft 3-size-submodularity assumption (pink), UMFI_OT (light blue), and UMFI_LR (dark blue) plotted against number of processed features from the BRCA data.

### 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 $O(p^{k+1})$ [10]. 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 3 and indicate that UMFI is approximately superlinear, with UMFI_OT incuring 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 3-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 [23], neural networks [8, 34], or principal inertial components [40]. 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 E). 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.

# Broader Impact

We hope that UMFI will be a useful tool in a variety of disciplines including bioinfomatics, 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 $\nu$. Though we think that our contributions can only lead to positive social and environmental impacts by aiding scientific discoveries in domains like earth science and bioinformatics, statistical methods, especially those that are aimed at genetics research, have historically been used to justify harmful and misleading claims. If such claims arise using our methods, then they should be dismissed since our work can only provide associative measures. Strong causal statements cannot be concluded after using our methods alone.

# 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.

[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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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`.

[11] 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.

[12] Shay Cohen, Gideon Dror, and Eytan Ruppin. Feature selection via coalitional game theory. *Neural Computation*, 19(7):1939–1961, 2007.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] Virgil Griffith and Christof Koch. Quantifying synergistic mutual information. *arXiv preprint arXiv:1205.4265*, 2012.

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

[20] 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.

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

[22] 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.

[23] 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.

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

[25] 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.

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

[27] 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.

[28] R Core Team Microsoft. *Microsoft R Open*. Microsoft, Redmond, Washington, 2017. URL `https://mran.microsoft.com/`.

[29] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[30] 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.

[31] 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.

[32] 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.

[33] Lloyd S Shapley. A value for n-person games, contributions to the theory of games, 2, 307–317, 1953.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

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

[43] 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.

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

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

[46] 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.

# Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] After reading the abstract and introduction please read Section 4 to see quantitative support for our claims.
   (b) Did you describe the limitations of your work? [Yes] Please see Section 5 for a detailed description of limitations.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Please see Section 3
   (b) Did you include complete proofs of all theoretical results? [Yes] Please see Section 3 and the Supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please see Section 4 and the Supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please see Section 4 and the Supplementary material.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] In most cases we did, see Section 4 and the Supplementary material, but for the computational complexity experiment, we did not because of the higher computational cost and the fact that repeated experiments would not significantly change the results (this was tested but not shown).

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please see Section 4 and the Supplementary material.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] Please see Section 4.2.

(b) Did you mention the license of the assets? [Yes] Please see Section 4.2. The BRCA dataset is available on Github via the MIT License.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Please see the supplemental material for code.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Please see Section 4.2.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Please see Section 4.2.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]