Supplementary Material for Ultra-marginal Feature Importance

Anonymous Author(s)
Affiliation
Address
email

A Mutual information

A.1 Properties of mutual information

Theorem A.1 (Supermodularity under mutual independence)**.** *Let S, X_1, X_2 be random variables such that S, X_1, X_2 are mutually independent. Then, $I(Y; S, X_1, X_2) - I(Y; S, X_2) \geq I(Y; S, X_1) - I(Y; S)$ [25, 33].*

Proof.

$$\begin{aligned} & I(Y; S, X_1, X_2) - I(Y; S, X_2) \\ & = I(Y; S) + I(Y; X_2|S) + I(Y; X_1|S, X_2) - [I(Y; S) + I(Y; X_2|S)] \quad \text{(by chain rule)} \\ & = I(Y; X_1|S, X_2) = I(Y, S, X_2; X_1) \quad \text{(by mutual independence)} \\ & \geq I(Y, S; X_1) \quad \text{(by monotonicity of } I(\cdot; f)) \\ & = I(Y; X_1|S) \quad \text{(by mutual independence)} \\ & = I(Y; S, X_1) - I(Y; S) \quad \text{(by the chain rule for mutual information)} \end{aligned}$$

□

Theorem A.2 (Data processing inequality)**.** *Let X, Y, Z be three random variables forming a Markov chain $X \to Y \to Z$, i.e. $X \perp\!\!\!\perp Z|Y$. Then, $I(X; Y) \geq I(X; Z)$.*

Proof. The proof can be found in Cover and Thomas [13, p. 32]. □

Theorem A.3. *Let F be a set of features used to predict the response Y. Then $I(Y; F) \geq I(Y; g(F))$ for any function g. If g is injective, then $I(Y; F) = I(Y; g(F))$.*

Proof. The first claim $I(Y; F) \geq I(Y; g(F))$ follows from the data processing inequality A.2 since $Y \to F \to g(F)$ forms a Markov chain.

If g is injective, then we may write $F = h(g(F))$ where $h : Im(g) \to F$ is the inverse of g restricted to the image of g. Hence, it follows that $Y \to g(F) \to F$ is a Markov chain. Note that $Y \perp\!\!\!\perp F|g(F)$ is equivalent to $Y \perp\!\!\!\perp h(g(F))|g(F)$, and therefore F is a constant given $g(F)$. By the data processing inequality, $I(Y; g(F)) \geq I(Y; F)$ and combining with the above inequality yields the desired claim, $I(Y; g(F)) = I(Y; F)$ when g is injective. □

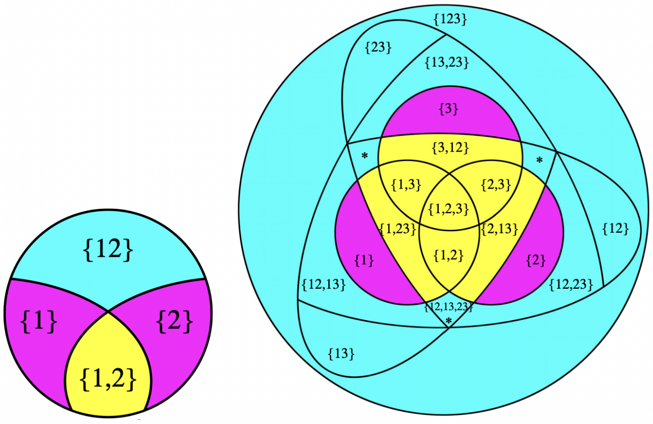
A.2 Mutual information and feature importance

Let $F = \{x_1, ..., x_p\}$ be a set of features used to predict Y. As shown in Griffith and Koch [20], the mutual information $I(Y; F) = I(Y; x_1, ..., x_p)$ can be visualized using a partial information (PI) diagram [38]. We may interpret the mutual information shared between Y and F as a collection of

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

non-negative pieces of information, whose sum forms $I(Y;F)$. Each of these pieces of information can be classified as (1) unique, (2) redundant, or (3) synergistic (Figure 4).

We note that the distinction between feature importance methods that seek to explain data vs. methods that seek to explain the model comes from their treatment of redundant information, i.e. their treatment of dependent features. A true-to-data method, like MCI or UMFI, should count all of the redundant information pertaining to x_i in $I(Y;F)$ towards the feature importance of x_i. Indeed, even though this information can be found elsewhere in the model, redundant information still constitutes part of the information that x_i shares about Y in the data. Conversely, a true-to-model approach, like conditional permutation importance (CPI), would count none of the redundant information towards the evaluation of a feature's importance, since this information is already found in another feature.



(a) $n = 2$ (b) $n = 3$

Figure 4: PI-diagrams taken from Griffith and Koch [20] for $I(Y;F)$ when $|F| = 2$ (left) and $|F| = 3$ (right). Magenta represents unique information, redundant information is colored with yellow, and synergistic information is in cyan. The starred regions represent a single region.

Mutual information itself is a common choice in the context of feature selection [4, 3, 41, 5]. 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 [5, 4]. 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 feature importance.

Another connection between feature importance and mutual information comes from Louppe et al. [27], 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 [27, Eq. 4], 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 for a machine learning model $\nu_f(S) : \mathcal{P}(S) \rightarrow \mathbb{R}_{\geq 0}$ measures how well the response Y can be predicted using the model f and the features $S \subseteq F$. Intuitively, $\nu_f(S)$ should ideally mirror or at least covary with mutual information $I(Y;S)$. Direct relationships between mutual information and machine learning evaluation functions have been observed in previous works. For example, the Gini value is equivalent to the first order Taylor approximation of information entropy [42]. The Gini impurity index is the central mechanism for choosing splits in random forests [39]. In the case of regression, one can also closely relate mutual information to the explained variance of a model. Indeed, with some assumptions, mutual information and R^2 accuracy are related.

If we assume the response and predictions are joint Gaussian and the predictions are unbiased [13], 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. [14] demonstrated equivalence when the Bayes classifier is known and cross entropy loss is used. In a simple example, Catav et al. [9] 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. [35] 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. [10] 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. $A_\nu(x_i) = \nu(F) - \nu(F \setminus \{x_i\})$. Bivariate methods are among the most popular methods for genome-wide association studies [12, 17, 34]. 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. $B_\nu(x_i) = \nu(x_i) - \nu(\emptyset)$. The three feature importance axioms proposed by Catav et al. [10] were partially motivated by the shortcomings of these two methods.

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

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 of x_i.

3. **Minimalism**: Catav et al. [10] decided to impose the minimalism axiom so that MCI can be unique. If $I_\nu(x_i)$ satisfies the first two axioms, then multiplying $I_\nu(x_i)$ by any constant $\lambda > 1$ would not change this. The minimalism axiom helps disambiguate MCI from these trivial variations.

We intentionally excluded some of the MCI axioms and properties included by Catav et al. [10] when proposing axioms for true-to-data feature importance methods in Section 2. Most importantly, the marginal contribution axiom is not included because it conflicts directly with the blood relation axiom. In the collider example presented by Harel et al. [21], they present the causal graph $Y \leftarrow S \rightarrow G \leftarrow E$, where S is unmeasured. Let $F = \{E, G\}$ be used to predict Y. Then, the marginal contribution axiom requires that feature E is given importance. Indeed, if we know G, then feature E can help predict the response, and thus, $I_\nu(E) \geq A_\nu(E) = \nu(\{E,G\}) - \nu(\{G\}) > 0$. However, as stated in Harel et al. [21], feature E has no relation to the response Y, so it would be more reasonable to give E zero importance. Indeed, E is given zero importance under the blood relation axiom, so the blood relation axiom is more reasonable and justified compared to the marginal contribution axiom. In contrast, G inherently contains information about Y via S, but this information is noised up by E. Therefore, although E can be used to denoise G and predict Y better, only G should be given importance when explaining the data, and indeed, G is blood related to Y. We note that UMFI obeys the blood relation axiom under some assumptions, and hence does not obey the marginal contribution axiom.

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

Theorem C.1 (Existence of optimal preprocessing $\hat{S}^F_{x_i}$ when all variables are jointly Gaussian)**.** *Let $x_i \in F$ and suppose that all features in the random vector F are joint normally distributed with mean 0 and that the preprocessed matrix $S^F_{x_i}$ is obtained via multiple linear regression with the model:*

$$F \setminus \{x_i\} = \beta x_i + \epsilon,$$

where $\epsilon = S^F_{x_i}$, x_i is a random variable in F, and β is the column vector of size $|F| - 1$ that minimizes the least squares error. Then, $S^F_{x_i}$ is an optimal preprocessing.

Proof. To show that $S^F_{x_i}$ is an optimal preprocessing (Definition 1), it suffices to show that $S^F_{x_i} \perp\!\!\!\perp x_i$ and that $I(Y; F) = I(Y; S^F_{x_i}, x_i)$, since $S^F_{x_i}$ is a function of F by construction.

From the normal equations and the definition of covariance, we know that $Cov(S^F_{x_i}, x_i) = 0$, as shown in the proof of Theorem E.3. Since $S^F_{x_i} = F \setminus \{x_i\} - \beta x_i$, and all features in F are joint normally distributed, it follows that $(S^F_{x_i}, x_i)$ is joint normally distributed as well, since $(S^F_{x_i}, x_i)$ can be obtained via the linear transformation $AF = (S^F_{x_i}, x_i)$, where the main diagonal entries of A are 1, the other $|F| - 1$ entries of the column corresponding to x_i are given by the entries of $-\beta$, and all other entries are 0. Without loss of generality, we may reorder the columns of the matrix such that the last column is attributed to feature x_i, and write

$$A = \begin{bmatrix} 1 & 0 & \dots & \dots & -\beta_1 \\ 0 & 1 & 0 & \dots & -\beta_2 \\ \vdots & & \ddots & & \\ 0 & 0 & \dots & \dots & 1 \end{bmatrix} \qquad A^{-1} = \begin{bmatrix} 1 & 0 & \dots & \dots & \beta_1 \\ 0 & 1 & 0 & \dots & \beta_2 \\ \vdots & & \ddots & & \\ 0 & 0 & \dots & \dots & 1 \end{bmatrix}.$$

Hence, $Cov(x_i, S^F_{x_i}) = 0 \implies S^F_{x_i} \perp\!\!\!\perp x_i$ from the properties of multivariate Gaussians.

To prove the second claim $I(Y; F) = I(Y; S^F_{x_i}, x_i)$, by Theorem A.3, it suffices to show that the map $h(F) = (S^F_{x_i}, x_i) = AF$ is injective. This is immediate from the fact that the matrix A, defined above, is invertible and thus bijective (injective and surjective).

□

For all subsequent proofs in this section, we assume that the evaluation function $\nu(S) = I(Y; S)$.

Theorem C.2 (Elimination axiom)**.** *Let $x_i \in F$ and $x_{p+1} \notin F$. When preprocessing is performed using optimal transport with chaining, $U^{F,Y}_\nu(x_i) \leq U^{F \cup \{x_{p+1}\},Y}_\nu(x_i)$.*

Proof. Let $S^{F \cup \{x_{p+1}\}}_{x_i}$ be the preprocessed version of $F \cup \{x_{p+1}\}$ relative to x_i and let $S^F_{x_i}$ be the preprocessed version of F relative to x_i. By optimal transport with chaining [23], we may assume that $S^{F \cup \{x_{p+1}\}}_{x_i}$ obeys the form $S^{F \cup \{x_{p+1}\}}_{x_i} = S^F_{x_i} \cup \tilde{x}$ and that $S^F_{x_i}, x_i, \tilde{x}$ are mutually independent. It follows from supermodularity of mutual information under mutual independence (Theorem A.1) that

$$\begin{aligned} &U^{F \cup \{x_{p+1}\},Y}_\nu(x_i) = I(Y; S^{F \cup \{x_{p+1}\}}_{x_i}, x_i) - I(Y; S^{F \cup \{x_{p+1}\}}_{x_i}) = I(Y; S^F_{x_i}, \tilde{x}, x_i) - I(Y; S^F_{x_i}, \tilde{x}) \\ &\geq I(Y; S^F_{x_i}, x_i) - I(Y; S^F_{x_i}) = U^{F,Y}_\nu(x_i). \end{aligned}$$

□

Theorem C.3 (Duplication invariance and symmetry axiom)**.** *Let $x_j \in F$, $\hat{x} \notin F$, and $\hat{x} = x_j$. Suppose that for all $x_i \in F$, the preprocessed feature sets $\hat{S}^F_{x_i}$ and $\hat{S}^{F \cup \{\hat{x}\}}_{x_i}$ are optimal preprocessings. Then, $\forall x_i \in F$, $U^{F,Y}_\nu(x_i) = U^{F \cup \{\hat{x}\},Y}_\nu(x_i)$ and $U^{F \cup \{\hat{x}\},Y}_\nu(\hat{x}) = U^{F \cup \{\hat{x}\},Y}_\nu(x_j)$.*

Proof. Recall that an optimal preprocessing relative to a feature set F and a feature of interest x_i are defined in Definition 1. To prove the claims, we first show that all optimal preprocessings $\hat{S}^F_{x_i}$ are also optimal preprocessings for $\hat{S}_{x_i}^{F\cup\{\hat{x}\}}$, and that all optimal preprocessings $\hat{S}_{x_i}^{F\cup\{\hat{x}\}}$ are also optimal preprocessings $\hat{S}_{x_i}^{F}$. We prove this for all $x_i \in F$.

The agreement of the first two properties in Definition 1 follows immediately from the fact that a function with repeated arguments can be defined to be equal to the same function without repeated arguments and the fact that both $\hat{S}_{x_i}^{F}$ and $\hat{S}_{x_i}^{F\cup\{\hat{x}\}}$ must be independent of x_i by definition. Then, since mutual information is invariant under duplicate information and since $\hat{S}_{x_i}^{F}$ and $\hat{S}_{x_i}^{F\cup\{\hat{x}\}}$ are optimal, we know that

$$I(Y;F,\hat{x}) = I(Y;\hat{S}_{x_i}^{F\cup\{\hat{x}\}},x_i) = I(Y;F) = I(Y;\hat{S}_{x_i}^{F},x_i)$$

Hence, we may assume that optimal preprocessings $\hat{S}_{x_i}^{F}$ and $\hat{S}_{x_i}^{F\cup\{\hat{x}\}}$ are interchangeable for all $x_i \in F$, and it follows that

$$U_\nu^{F,Y}(x_i) = I(Y;\hat{S}_{x_i}^{F},x_i) - I(Y;\hat{S}_{x_i}^{F}) = I(Y;\hat{S}_{x_i}^{F\cup\{\hat{x}\}},x_i) - I(Y;\hat{S}_{x_i}^{F\cup\{\hat{x}\}}) = U_\nu^{F\cup\{\hat{x}\},Y}(x_i).$$

Finally, since $x_j = \hat{x}$, we can see that $\hat{S}_{x_j}^{F\cup\{\hat{x}\}}$ and $\hat{S}_{\hat{x}}^{F\cup\{\hat{x}\}}$ are interchangeable, which proves the symmetry axiom

$$U_\nu^{F\cup\{\hat{x}\},Y}(\hat{x}) = U_\nu^{F\cup\{\hat{x}\},Y}(x_j).$$

□

We note that this proof holds when the preprocessings $S_{x_i}^{F}$ and $S_{x_i}^{F\cup\{\hat{x}\}}$, used to compute UMFI scores, are interchangeable. This fact does not require that the preprocessings must be optimal, and also holds when the removal of dependencies on a feature x_i is done in a pairwise fashion (see Algorithm 2) or via optimal transport with chaining [23].

Theorem C.4 (Blood relation axiom for Gaussian graphical model)**.** *Assuming the data is generated from a Gaussian graphical model obeying the global Markov property and faithfulness, and that the preprocessings $\hat{S}_{x_i}^{F}$ are optimally obtained via linear regression, $U_\nu^{F,Y}(x_i) > 0$ if and only if $x_i \in BR(Y)$.*

Proof. To start, we know that $U_\nu^{F,Y}(x_i) = I(Y;\hat{S}_{x_i}^{F},x_i) - I(Y;\hat{S}_{x_i}^{F}) = I(Y;x_i|\hat{S}_{x_i}^{F})$. And from the definition of conditional mutual information, we know $U_\nu^{F,Y}(x_i) = 0 \iff I(Y;x_i|\hat{S}_{x_i}^{F}) = 0 \iff Y \perp\!\!\!\perp x_i|\hat{S}_{x_i}^{F}$. Since we have a Gaussian graphical model, the features and the response (Y,F) are jointly normally distributed. Furthermore, because $S_{x_i}^{F}$ is obtained via linear regression, $(Y,x_i,S_{x_i}^{F})$ is also jointly Gaussian, since it can be expressed as a linear transformation of (Y,F), like as was shown in the proof of Theorem C.1. We may therefore write the conditional independence statement in terms of covariance block matrices [37, Prop. 2.3]

$$Y \perp\!\!\!\perp x_i|\hat{S}_{x_i}^{F} \iff \Sigma_{x_i,Y} - \Sigma_{x_i,\hat{S}_{x_i}^{F}}\Sigma^{-1}_{\hat{S}_{x_i}^{F},\hat{S}_{x_i}^{F}}\Sigma_{\hat{S}_{x_i}^{F},Y} = 0. \quad (1)$$

Since $x_i \perp\!\!\!\perp \hat{S}_{x_i}^{F}$, (1) reduces to

$$Y \perp\!\!\!\perp x_i|\hat{S}_{x_i}^{F} \iff \Sigma_{x_i,Y} = 0 \iff x_i \perp\!\!\!\perp Y, \quad (2)$$

where this last equivalence is due to the fact that x_i and Y are jointly Gaussian.

All that is left to prove is $x_i \perp\!\!\!\perp Y \iff x_i \notin BR(Y)$. First, if $x_i \notin BR(Y)$, then $x_i \perp\!\!\!\perp Y$ follows from the global Markov property and the fact that x_i and Y are d-separated by the empty set. Indeed, every path from x_i to Y must have at least one collider. We consider two cases. (1) The edge coming out of Y is outgoing. Then since x_i is not a descendent of Y, the path must reverse its orientation at some vertex before meeting x_i. That vertex is a collider. (2) The edge connecting to Y points towards Y. Then the path must reverse its orientation at some point since x_i is not an ancestor of Y. The path must then reverse another time because otherwise, x_i would share a common ancestor with Y (the vertex of the first reversal). The vertex with the second reversal is a collider.

Conversely, let $x_i \in BR(Y)$. By the faithfulness assumption, it suffices to show that x_i and Y are d-connected by the empty set. Since $x_i \in BR(Y)$, there are two possible cases: either there is a directed path between x_i and Y, or x_i and Y share a common ancestor. In the first case, we simply choose the directed path between x_i and Y and observe that there cannot be a collider. Similarly, in the second case, we may pick the path beginning at Y and trace it up to the common ancestor and then travel to x_i. There can be no colliders along the path since every vertex has at least one outgoing edge by construction. Also, the empty set cannot contain any non-colliders.

□

Theorem C.5 (Blood relation axiom in the absence of interactions)**.** *Suppose that there is no synergistic information $I_{syn}(Y; S^F_{x_i}, x_i)$ about Y between x_i and $S^F_{x_i}$ for all $x_i \in F$, and that $S^F_{x_i} \perp\!\!\!\perp x_i$. Then, if the graphical model obeys the global Markov property and faithfulness, $U^{F,Y}_\nu > 0$ if and only if $x_i \in BR(Y)$.*

Proof. As in the proof of Theorem C.4, it suffices to show that $I(Y; x_i|S^F_{x_i}) = 0$ if and only if $x_i \notin BR(Y)$.

We may further rewrite $I(Y; x_i|S^F_{x_i}) = 0$ as $I(Y; S^F_{x_i}, x_i) = I(Y; S^F_{x_i})$. Using partial information decomposition [38], and since $S^F_{x_i} \perp\!\!\!\perp x_i$, we may decompose

$$I(Y; S^F_{x_i}, f) = I(Y; x_i) + I(Y; S^F_{x_i}) + I_{syn}(Y; S^F_{x_i}, x_i).$$

where we note that $I(Y; x_i) = I_{uniq}(Y; x_i)$ and that $I(Y; S^F_{x_i})$ captures the unique information that $S^F_{x_i}$ shares with Y as well as synergistic information within the random vector $S^F_{x_i}$ that is shared with Y. As proven in Theorem C.4, $I(Y; x_i) = 0$ if $x_i \notin BR(Y)$ and $I(Y; x_i) > 0$ if $x_i \in BR(Y)$ by the global Markov property and faithfulness. Since $I_{syn}(Y; S^F_{x_i}, x_i) = 0$ by assumption, this gives us the desired statement $I(Y; S^F_{x_i}, x_i) = I(Y; S^F_{x_i})$ if and only if $x_i \notin BR(Y)$. □

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 [18], Spearman's rank correlation coefficient [32], multiple linear regression [15], and partial correlation [40]. 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 [8]. At that time, 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 [8]. 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. [14].

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 [23]. In this section, we denote features in the feature set F by X_j or X_i to emphasize that they are random variables, rather than the previously used x_j and x_i, where the former is used to denote observations

x_j sampled from X_j instead. To obtain a preprocessing $S^F_{X_i}$, we may remove the dependencies of x_i from each $X_j \in F \setminus \{X_i\}$ with minimal information loss with respect to X_j. To do so using optimal transport, we consider the Monge problem:

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

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

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

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

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 $\mathbb{E}[|X - g(X)|^q] = \int_0^1 |F^{\leftarrow}(p) - \tilde{F}^{\leftarrow}(p)|^q dp$ for $g = \tilde{F}^{\leftarrow} \circ F$

$$\begin{aligned}\mathbb{E}[|X - g(X)|^q] &= \int_{-\infty}^{\infty} |x - \tilde{F}^{\leftarrow}(F(x))|^q f(x) dx \\ &= \int_{-\infty}^{\infty} |F^{\leftarrow}(F(x)) - \tilde{F}^{\leftarrow}(F(x))|^q f(x) dx = \int_0^1 |F^{\leftarrow}(p) - \tilde{F}^{\leftarrow}(p)|^q dp\end{aligned}$$

□

Theorem E.2. *Let $F_{j|x_i}(x) = P(X_j \le x_j | X_i = x_i)$ denote the CDF of $X_j|\{X_i = x_i\}$. Then $g = \tilde{F}^{\leftarrow} \circ F_{j|x_i}$ optimally transports $X_j|\{X_i = x_i\}$ to $\tilde{X}_j \perp\!\!\!\perp X_i$ for any CDF $\tilde{F}$*

Proof. We apply Theorem E.1 on the random variable $X_j|\{X_i = x_i\}$ and note that $X_j|\{X_i = x_i\}$ is independent of X_i. In particular, $g(X_j|X_i = x_i) \perp\!\!\!\perp X_i$ for any choice of $\tilde{F}$. □

Theorem E.2 suggests an algorithm for transporting data $(x_{j1}, ..., x_{jn})$ sampled from X_j, to $(\tilde{x}_{j1}, ..., \tilde{x}_{jn}) \perp\!\!\!\perp (x_{i1}, ..., x_{in})$. Since x_{jk} is taken jointly with x_{ik}, as they are attributes coming from the kth sample in the dataset, then x_{jk} is a realization of the distribution $X_j|\{X_i = x_{ik}\}$. Consequently, for each $k = 1, ..., n$, we should transport x_{jk} to $\tilde{x}_{jk} = \tilde{F}^{\leftarrow}(F_{j|x_{ik}}(x_{jk}))$, where we may pick any CDF $\tilde{F}$. This procedure can also adapted for features sampled from discrete r.v's, as shown in Johndrow and Lum [23].

Algorithm 1: Algorithm for removing dependencies of X_i from X_j

```
Require: $X_j = [x_{j1}, ..., x_{jn}]$, $X_i = [x_{i1}, ..., x_{in}]$, $X_j|(X_i = x_{ik}) \sim F_{j|x_{ik}}$, $\tilde{F}$ is a CDF
  for $k = 1, ..., n$ do
    $\tilde{x}_{jk} = \tilde{F}^{\leftarrow}(F_{j|x_{ik}}(x_{jk}))$
  end for
  return $\tilde{X}_j = [\tilde{x}_{j1}, ..., \tilde{x}_{jn}]$
```

We denote the result of the algorithm by $\tilde{X}_j = \tilde{F}^{\leftarrow}(F_{j|X_i}(X_j))$ and would ideally pick $\tilde{F}$ such that it minimizes the transportation cost $g_c(X_j, \tilde{X}_j) = g_c(X_j, \tilde{F}^{\leftarrow}(F_{j|X_i}(X_j)))$ 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 a large as the support of F_j, the cdf of X_j, then any rank-based prediction rule, e.g. random forest, will be invariant to the choice of $\tilde{F}_j$ [23]. A standard choice for $\tilde{F}_j$ is F_j so that we can recover the original quantiles of X_j.

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

We may iterate Algorithm 1 over each feature in $F \setminus \{X_i\}$ to obtain pairwise independence between the transported variables $\tilde{X}_j$ and X_i. It is also possible to iterate Algorithm 1 via chaining to achieve mutual independence between the transformed variables $\tilde{X}_j$ and X_i [23, 2.4]. However, this is computationally expensive, and pairwise independence should suffice for an accurate UMFI score, as will be explored further in Section F. Step 2 of Algorithm 1 in the main paper can therefore be implemented with Algorithm 2.

```
**Algorithm 2:** Algorithm for estimating $S^F_{X_i}$ via pairwise optimal transport
**Require:** $X_i = [x_{i1}, ..., x_{in}]$, $X_j = [x_{j1}, ..., x_{jn}]$ for $X_j$ in $F \setminus X_i$
  $S^F_{X_i} = \emptyset$
  **for** $X_j$ in $F \setminus \{X_i\}$ **do**
    $\tilde{X}_j =$ output of Algorithm 1 with $X_j$ and $X_i$
    add $\tilde{X}_j$ to $S^F_Z$
  **end for**
  **return** $S^F_{X_i}$
```

In other words, we may estimate $S^F_{X_i}$ as:

$$S^F_{X_i} = \{F_j^{\leftarrow}(F_{j|X_i}(X_j)) : X_j \in F \setminus \{X_i\}\}.$$

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* [6, 28].

To reiterate, removing dependencies requires methods to make a feature or set of features S independent of a protected attribute x_i, 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 x_i.

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

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

and X and ϵ are joint normally distributed, then (1) $\epsilon \perp\!\!\!\perp X$ and (2) ϵ 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 $\mathbb{E}[\epsilon] = 0$, it follows that

$$Cov(X, \epsilon) = \mathbb{E}[X^T \epsilon] - \mathbb{E}[\epsilon]\mathbb{E}[X] = \mathbb{E}[X^T \epsilon] = \mathbb{E}[X^T(Y - X\beta)]$$
$$= \mathbb{E}[X^T(Y - X(X^TX)^{-1}X^TY))] = \mathbb{E}[X^TY - X^TX(X^TX)^{-1}X^TY] = \mathbb{E}[X^TY - X^TY] = 0$$

Then, since X and ϵ 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

$$S^F_{X_i} = \{\epsilon_j = X_j - \beta_{0,j} - \beta_{1,j} X_i : X_j \in F \setminus \{X_i\}\}$$

where $\beta_{0,j}$ is the intercept term of the linear regression model $X_i = \beta_{0,j} + \beta_{1,j} X_j + \epsilon_j$.

F Experiments comparing linear regression and optimal transport

In the following subsections, we compare the ability of linear regression and pairwise optimal transport to remove the information of a feature from data while distorting the original data as little as possible. It can be concluded that while linear regression works optimally when the data is jointly Gaussian, on real data, such as the BRCA dataset, pairwise optimal transport can find independent representations of the data, while linear regression fails (Section F.1).

To implement UMFI paired with linear regression, we only remove dependencies when the regression slope coefficient is statistically significant (p-value < 0.01). To implement UMFI paired with pairwise optimal transport, when removing dependencies on the feature X_i from the dataset, we estimate $F_{j|x_{ik}}$ by breaking up X_i into quantiles of size 150 and running linear regression on each quantile. The new orthogonal predictors are then given by the values of the inverse empirical CDF of the residuals from the mentioned linear regression model.

F.1 Removing dependencies

It is crucial for our linear regression and optimal transport preprocessing methods to remove the information associated with the feature of interest, x_i, from the rest of the dataset $F \setminus \{x_i\}$. Therefore, we would like the preprocessed dataset $S^F_{x_i}$ to share zero mutual information with x_i. The mutual information $I(x_i; S^F_{x_i})$ is difficult to calculate, but it is closely related to the optimal predictor of x_i given $S^F_{x_i}$ [31]. For example, if $I(x_i; S^F_{x_i}) = 0$, as is desired, then the optimal predictor of x_i will have zero accuracy given $S^F_{x_i}$. If the opposite is true and $S^F_{x_i}$ contains all of the information from x_i, then an optimal predictor of x_i should be able to perfectly predict x_i from the given information in $S^F_{x_i}$. In the following experiments, we assume that random forests can form the optimal predictor of x_i given $S^F_{x_i}$. We use the OOB-R^2 value coming from the random forest model to give a relative measure of the mutual information between x_i and the transformed dataset $S^F_{x_i}$.

We used the BRCA dataset with 50 features to test the ability of optimal transport and linear regression to remove dependencies [14, 9]. 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 $I_{rel}(x_i; F \setminus \{x_i\})$ of the mutual information $I(x_i; F \setminus \{x_i\})$ between the feature of interest x_i and the other 49 features. We then consider the case where the 49 remaining features are preprocessed to have dependencies on x_i removed via linear regression or pairwise optimal transport. Similarly, random forest's OOB-R^2 is used to give a relative measure $I_{rel}(x_i; S^F_{x_i})$of $I(x_i; S^F_{x_i})$.

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 certainly 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, x_i cannot be predicted successfully by random forest (OOB-$R^2 = 0$) from the other features after $F \setminus x_i$ 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 [24]. For each feature in the BRCA dataset with 50 features [14, 9], the information from the current feature is removed from all other features with either linear regression or pairwise optimal transport (Figure 6).

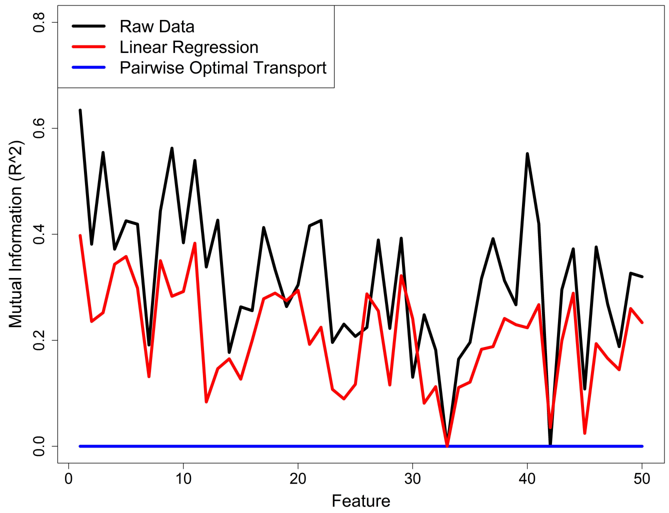
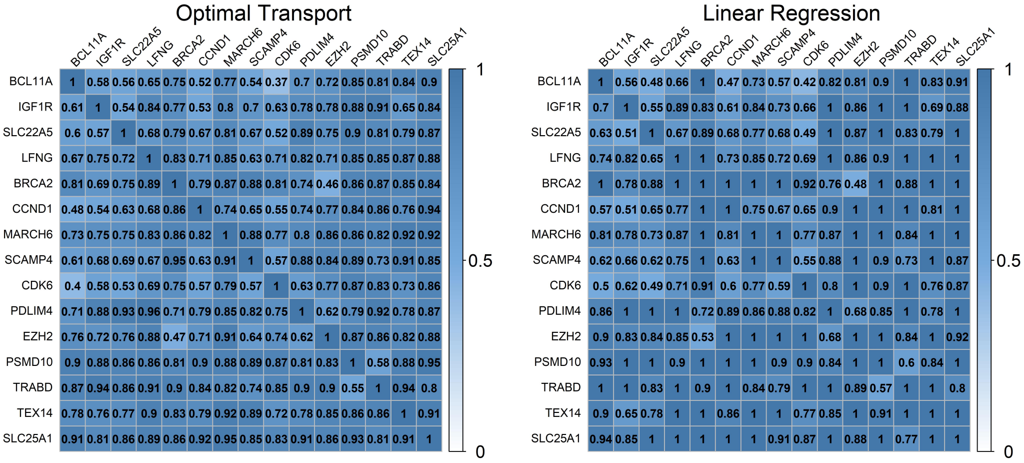


Figure 5: The relative mutual information $I_{rel}(x_i; F \setminus \{x_i\})$ between the ith feature in the BRCA dataset and all other features is plotted (black) for each $i \in \{1, 2, ...50\}$. The relative mutual information $I_{rel}(x_i; S^F_{x_i})$ between the ith feature and all other features 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.



Optimal Transport

	BCL11A	IGF1R	SLC22A5	LFNG	BRCA2	CCND1	MARCH6	SCAMP4	CDK6	PDLIM4	EZH2	PSMD10	TRABD	TEX14	SLC25A1
BCL11A	1	0.58	0.56	0.65	0.75	0.52	0.77	0.54	0.37	0.7	0.72	0.85	0.81	0.84	0.9
IGF1R	0.61	1	0.54	0.84	0.77	0.53	0.8	0.7	0.63	0.78	0.78	0.88	0.91	0.65	0.84
SLC22A5	0.6	0.57	1	0.68	0.79	0.67	0.81	0.67	0.52	0.89	0.75	0.9	0.81	0.79	0.87
LFNG	0.67	0.75	0.72	1	0.83	0.71	0.85	0.63	0.71	0.82	0.71	0.85	0.85	0.87	0.88
BRCA2	0.81	0.69	0.75	0.89	1	0.79	0.87	0.88	0.81	0.74	0.46	0.86	0.87	0.85	0.84
CCND1	0.48	0.54	0.63	0.68	0.86	1	0.74	0.65	0.55	0.74	0.77	0.84	0.86	0.76	0.94
MARCH6	0.73	0.75	0.75	0.83	0.86	0.82	1	0.88	0.77	0.8	0.86	0.86	0.82	0.92	0.92
SCAMP4	0.61	0.68	0.69	0.67	0.95	0.63	0.91	1	0.57	0.88	0.84	0.89	0.73	0.91	0.85
CDK6	0.4	0.58	0.53	0.69	0.75	0.57	0.79	0.57	1	0.63	0.77	0.87	0.83	0.73	0.86
PDLIM4	0.71	0.88	0.93	0.96	0.71	0.79	0.85	0.82	0.75	1	0.62	0.79	0.92	0.78	0.87
EZH2	0.76	0.72	0.76	0.88	0.47	0.71	0.91	0.64	0.74	0.62	1	0.87	0.86	0.82	0.88
PSMD10	0.9	0.88	0.86	0.86	0.81	0.9	0.88	0.89	0.87	0.81	0.83	1	0.58	0.88	0.95
TRABD	0.87	0.94	0.86	0.91	0.9	0.84	0.82	0.74	0.85	0.9	0.9	0.55	1	0.94	0.8
TEX14	0.78	0.76	0.77	0.9	0.83	0.79	0.92	0.89	0.72	0.78	0.85	0.86	0.86	1	0.91
SLC25A1	0.91	0.81	0.86	0.89	0.86	0.92	0.95	0.85	0.83	0.91	0.86	0.93	0.81	0.91	1

Linear Regression

	BCL11A	IGF1R	SLC22A5	LFNG	BRCA2	CCND1	MARCH6	SCAMP4	CDK6	PDLIM4	EZH2	PSMD10	TRABD	TEX14	SLC25A1
BCL11A	1	0.56	0.48	0.66	1	0.47	0.73	0.57	0.42	0.82	0.81	0.9	1	0.83	0.91
IGF1R	0.7	1	0.55	0.89	0.83	0.61	0.84	0.73	0.66	1	0.86	1	1	0.69	0.88
SLC22A5	0.63	0.51	1	0.67	0.89	0.68	0.77	0.68	0.49	1	0.87	1	0.83	0.79	1
LFNG	0.74	0.82	0.65	1	1	0.73	0.85	0.72	0.69	1	0.86	0.9	1	1	1
BRCA2	1	0.78	0.88	1	1	1	1	1	0.92	0.76	0.48	1	0.88	1	1
CCND1	0.57	0.51	0.65	0.77	1	1	0.75	0.67	0.65	0.9	1	1	1	0.81	1
MARCH6	0.81	0.78	0.73	0.87	1	0.81	1	1	0.77	0.87	1	1	0.84	1	1
SCAMP4	0.62	0.66	0.62	0.75	1	0.63	1	1	0.55	0.88	1	0.9	0.73	1	0.87
CDK6	0.5	0.62	0.49	0.71	0.91	0.6	0.77	0.59	1	0.8	1	0.9	1	0.76	0.87
PDLIM4	0.86	1	1	1	0.72	0.89	0.86	0.88	0.82	1	0.68	0.85	1	0.78	1
EZH2	0.9	0.83	0.84	0.85	0.53	1	1	1	1	0.68	1	1	0.84	1	0.92
PSMD10	0.93	1	1	0.9	1	1	1	0.9	0.9	0.84	1	1	0.6	0.84	1
TRABD	1	1	0.83	1	0.9	1	0.84	0.79	1	1	0.89	0.57	1	1	0.8
TEX14	0.9	0.65	0.78	1	1	0.86	1	1	0.77	0.85	1	0.91	1	1	1
SLC25A1	0.94	0.85	1	1	1	1	1	0.91	0.87	1	0.88	1	0.77	1	1

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 [36, 14, 10], and the CAMELS hydrology dataset [1]. MCI and UMFI used either random forests or extremely randomized trees [7, 19]. Both of these are implemented using the *ranger* R package [39]. Ablation, permutation importance, and conditional permutation importance used random forests. Ablation and permutation importance were implemented with the *ranger* R package [39], while conditional permutation importance was implemented with the *randomForest* and *permimp* packages [16, 26]. All experiments were run in Microsoft R Open Version 4.0.2 [29].

G.1 Extra 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 nonlinear interactions (Section G.1.1), correlated interactions (Section G.1.2), correlation (Section G.1.3), and blood and non-blood related features (Section G.1.4). Further, we test how using extremely randomized trees instead of random forests for MCI and UMFI changes the results of the same simulation experiments. Although other methods such as XGBoost [11] 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 [19]. Both random forests and extremely randomized trees are not sensitive to hyperparameters [30]. For these simulation studies, we also perturb the size of the quantiles used by UMFI_OT. We now use quantiles of size 30 instead of size 150. Quantiles of size 30 worked better on the hydrology data used in later experiments, so we test to see if the simulation results were sensitive to this choice in quantile size for dependency removal via optimal transport.

G.1.1 Nonlinear interactions

The first experiment on simulated data handles the case where two variables, x_1 and x_2, interact in a nonlinear way in the response Y. As explained in Section 4.1.1, we should expect x_1 and x_2 to contribute more than half of the total importance, while x_3 and x_4 should be important, but less important compared to x_1 and x_2. Figure 8a shows that ablation, PI, and CPI all provide accurate scores.

When tested with extremely randomized trees, the nonlinear interactions simulation experiment results for MCI and UMFI, shown in Figure 8e, remain mostly unchanged compared to the results from the experiment with random forests given in Figure 1a.

G.1.2 Correlated interactions

The second experiment considers the case where two correlated variables, x_1 and x_2, interact together in the response Y. Thus, as explained in Section 4.1.2, we should expect x_1 and x_2 to have more importance compared to x_3 and x_4. Figure 8b 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 8b).

When tested with extremely randomized trees instead of random forests, the correlated interaction simulation experiment results (Figure 8f) for MCI and UMFI are similar to the earlier results shown in Figure 1b. 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.1.3 Correlation

The third experiment tests how the metrics allocate importance to correlated features. As explained in Section 4.1.3, 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 8c indicates that CPI and ablation give near zero importance to the two heavily correlated features x_1 and x_3. This aligns with the discussion about true-to-model feature importance methods in Section A.2 since these methods base their scores on the importance of a feature conditioned on all other variables present in the model. 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 can be viewed as in between the true-to-data and true-to-model approaches. The true-to-data approaches (MCI and UMFI) allocate all of the redundant information to the feature. The true-to-model approaches (CPI and ablation) allocate none of the redundant information to the feature. PI evenly splits the redundant information across the relevant correlated features.

When tested with extremely randomized trees, the correlation simulation experiment results (Figure 8g) for MCI and UMFI change slightly compared to the experiment with random forests in Figure 1c. MCI works well, though it still gives some non-zero importance to x_4. With random forests, the relative importance of x_4 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.1.4 Blood relation

For the last simulation experiment, we revisit the blood relation experiment performed in Section 4.1.4 using data generated from the causal graph in Figure 7. The feature S is unobserved, so the only blood related features to Y in F are x_3 and x_4. x_3 and x_4 should therefore be given high importance while x_1 and x_2 should receive zero importance. When tested on ablation, CPI, and PI, we notice that all three metrics fail to capture the desired importance, since they each give significant importance to x_2, which is not blood related to Y. We also note that this experiment provides an explicit example of UMFI not satisfying the marginal contribution axiom, which states that feature importance metrics should allocate at least as much importance as attributed by the ablation metric. Indeed, as shown in Figure 1d, UMFI gives 0 importance to non-blood related features x_1 and x_2, whereas ablation gives significant importance to x_2.

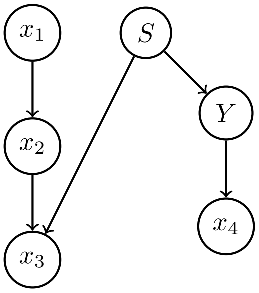


Figure 7: Causal graph which generates the data for the blood relation simulation experiment.

When the experiment was re-tested on MCI and both implementations of UMFI using extremely randomized trees instead of random forest, we observe that UMFI_LR and UMFI_OT both continue to give positive importance to the blood related features x_3 and x_4, while giving near-zero importance to the two remaining observed features (Figure 8h). However, we note that x_3 is given much more importance relative to x_4 when implemented with extremely randomized trees compared to random forests (Figure 1d). On the other hand, MCI gives positive importance to x_2 in this experiment. However, we note that it also correctly gave x_1 almost zero importance while giving x_3 and x_4 significantly more importance compared to the random forest implementation. Across most simulation studies, it appears MCI performs better using extremely randomized trees compared to random forests.

G.2 Extra 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 Section 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 10. 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.

G.3 Experiments on unpermuted BRCA data

Additional BRCA experiments were performed on the original randomized genes, as done in Covert et al. [14] and Catav et al. [10]. The observed 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 11. 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 10 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 12. 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 relative 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 give high importance to BRCA1, BRCA2, TEX14, EZH2, and IGF1R for BRCA associated genes, and SLC25A1 for non-BRCA associated genes.

G.3.1 Computational complexity

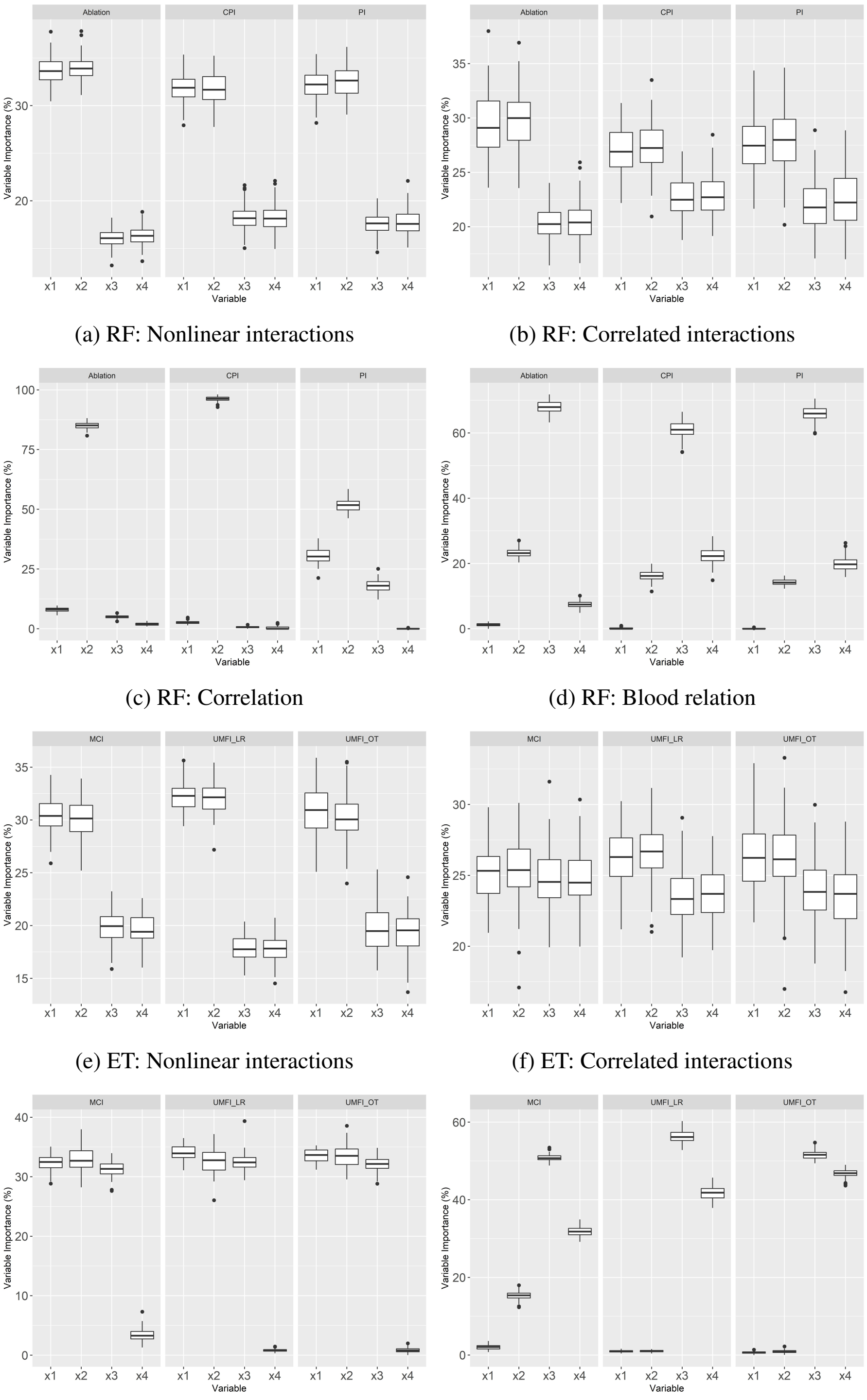
We compare the computational complexity of UMFI and MCI against the other feature importance methods that were explored in this section: ablation, PI, and CPI. To do so, we ran 10 iterations of the BRCA experiment, which has 50 features, each with 571 observations. We recorded the average time for each method to compute feature importance for $5, 10, 15, 20, 25, 30, 35, 40, 45$, and 50 features. Figure 13 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 with soft 2-size submodularity (50 features in 205 seconds).

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 in this experiment with an overall OOB-R^2 of 0.91.

Figure 14, which is analogous to Figure 5 in Appendix F, shows that both preprocessing methods fail to completely remove dependencies from the CAMELS dataset. This can likely be attributed to the fact that each feature is extremely dependent on the other explanatory features ($R^2 \geq 0.65$).

The feature importance scores indicated in Figure 15 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, 22], thus, even when dependencies can not be completely removed, UMFI can still provide reasonable measurements of feature importance.



(a) RF: Nonlinear interactions

(b) RF: Correlated interactions

(c) RF: Correlation

(d) RF: Blood relation

(e) ET: Nonlinear interactions

(f) ET: Correlated interactions

(g) ET: Correlation

(h) ET: Blood relation

Figure 8: Results for the experiments on simulated data from Subsection G.1. The results for ablation, conditional permutation importance (CPI), and permutation importance (PI) were implemented with random forest (RF), and are shown in Figures 8a, 8b, 8c, and 8d . The results for MCI, UMFI_LR, and UMFI_OT were implemented with extremely randomized trees (ET), and are shown in Figures 8e, 8f, 8g, and 8h. Feature importance scores are shown as a percentage of the total for each of x_1 to x_4 from 100 replications.

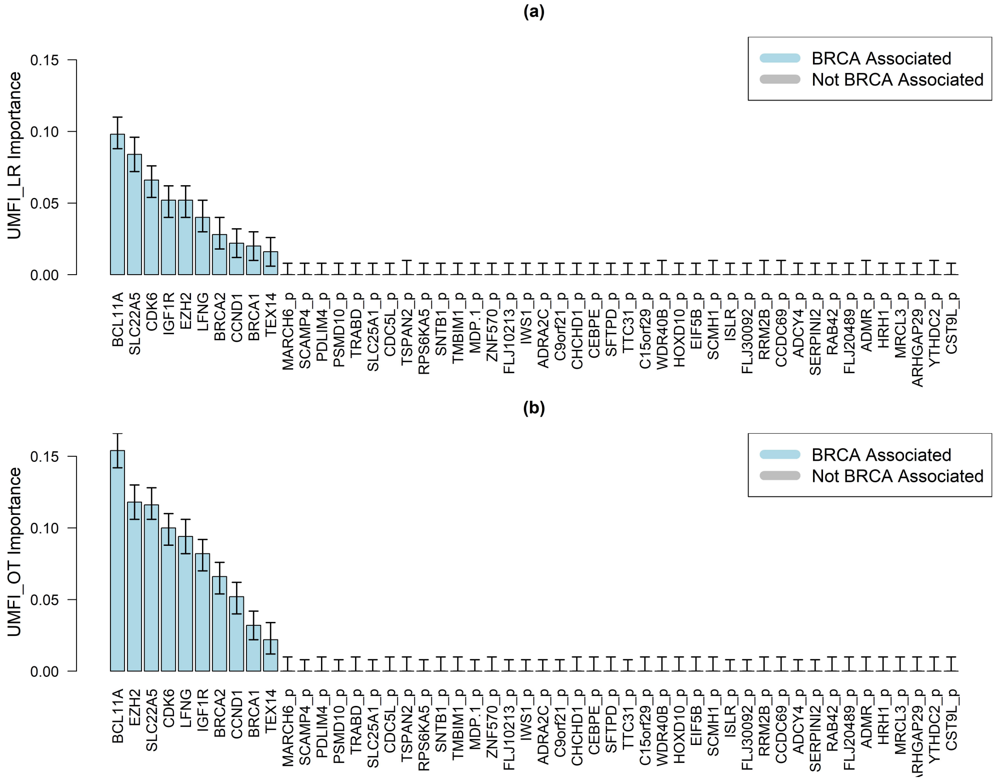


Figure 9: Median feature importance scores provided by (a) UMFI with linear regression, and (b) UMFI with pairwise optimal transport, for each gene in the permuted BRCA dataset after 5000 iterations. Genes colored in blue are known to be associated with breast cancer while genes colored in grey are random 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.

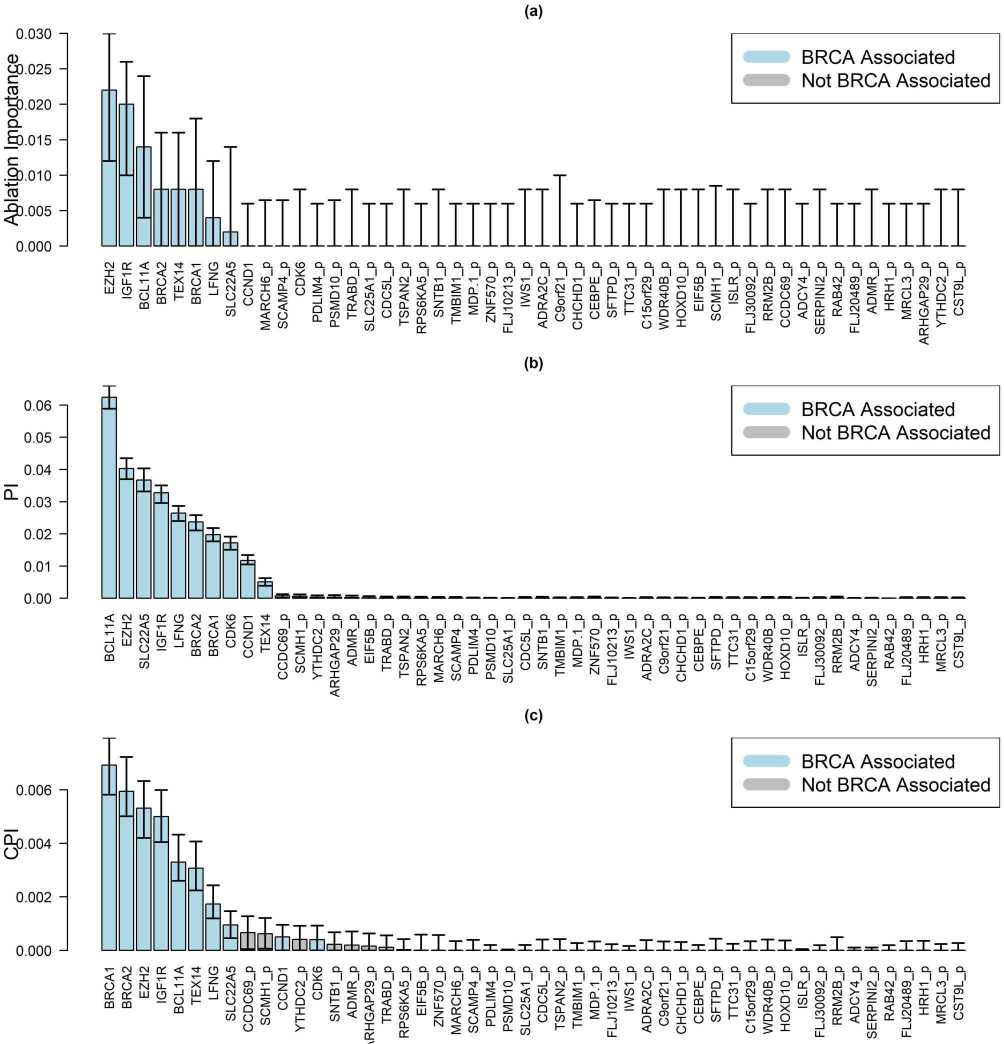


Figure 10: Median feature importance scores provided by (a) ablation, (b) permutation importance, and (c) conditional permutation importance, for each gene in the permuted BRCA dataset after 200 iterations. Genes colored in blue are known to be associated with breast cancer while genes colored in grey 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.

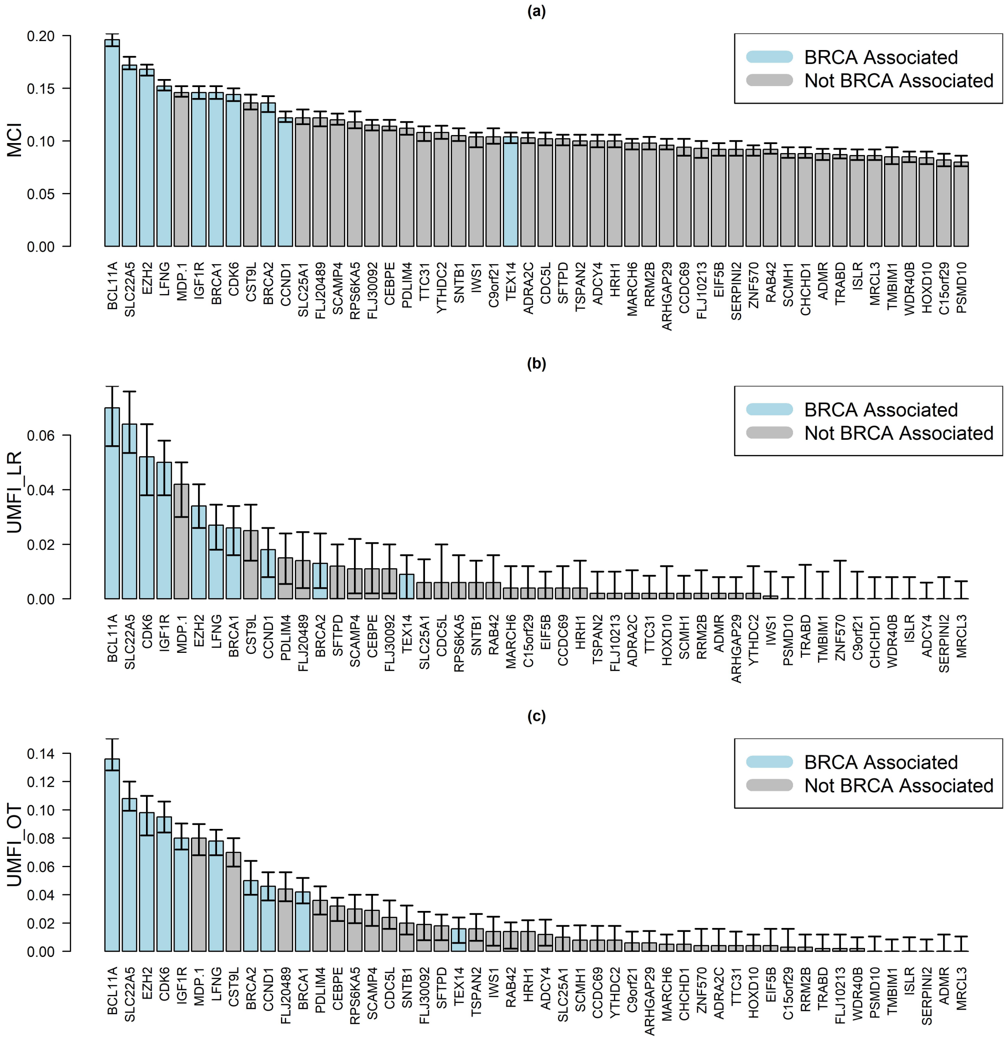


Figure 11: Median feature importance scores provided by (a) MCI, (b) UMFI with linear regression, and (c) UMFI with pairwise optimal transport, for each gene in the unpermuted BRCA dataset after 100 iterations. Genes colored in blue are associated with breast cancer while genes colored in grey are randomly selected genes. The first and third quantiles of the scores are visualized for each gene.

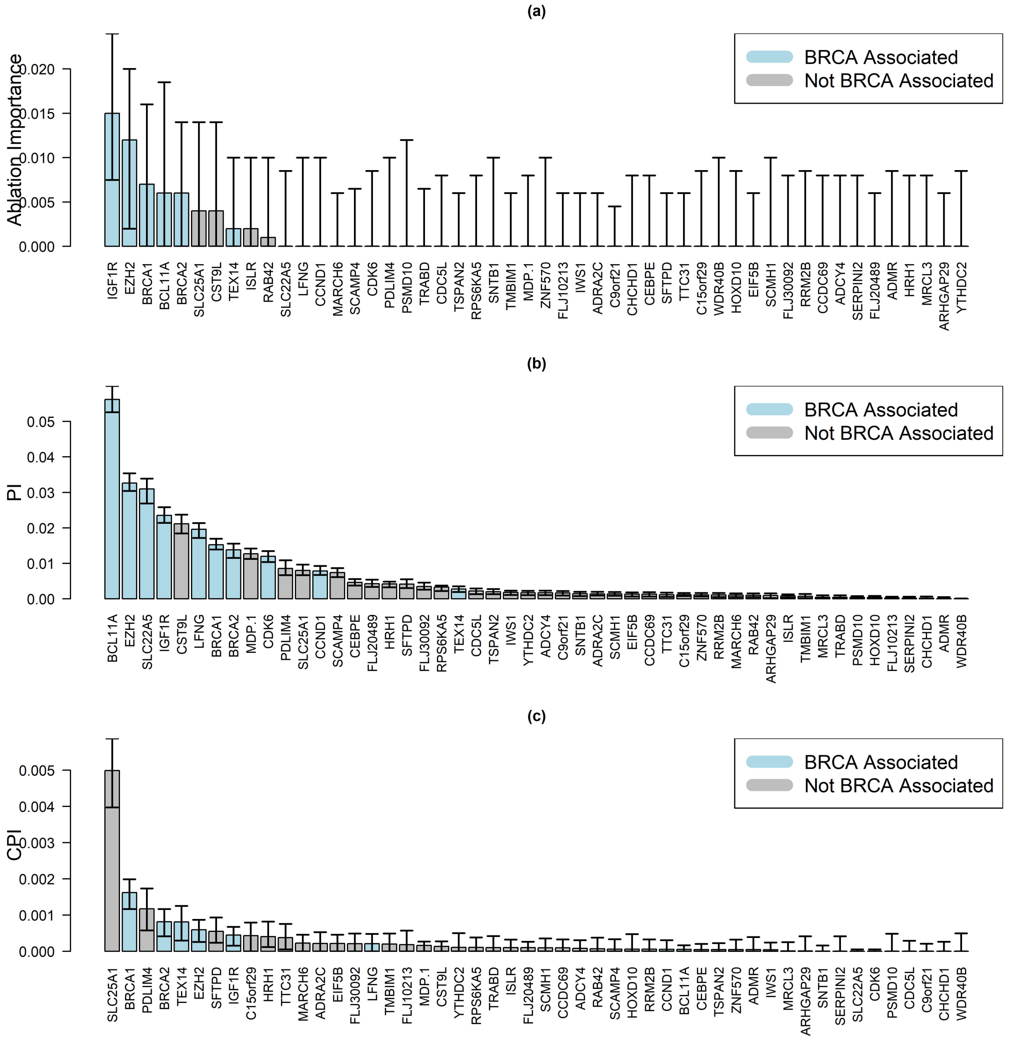


Figure 12: Median feature importance scores provided by (a) ablation, (b) permutation importance, and (c) conditional permutation importance, for each gene in the unpermuted BRCA dataset after 100 iterations. Genes colored in blue are associated with breast cancer while genes colored in grey are randomly selected genes. The first and third quantiles of the scores are visualized for each gene.

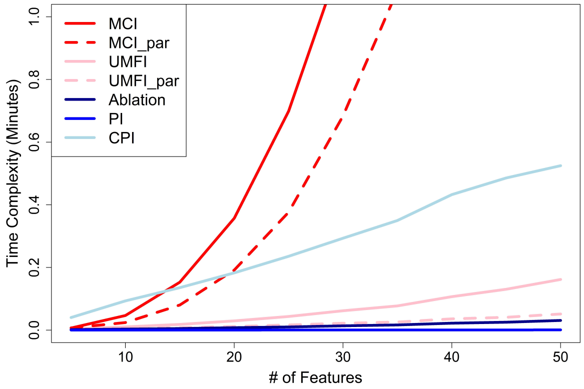


Figure 13: The average computation time for each method to process p features over 10 iterations of the original BRCA data is plotted for each $p \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$.

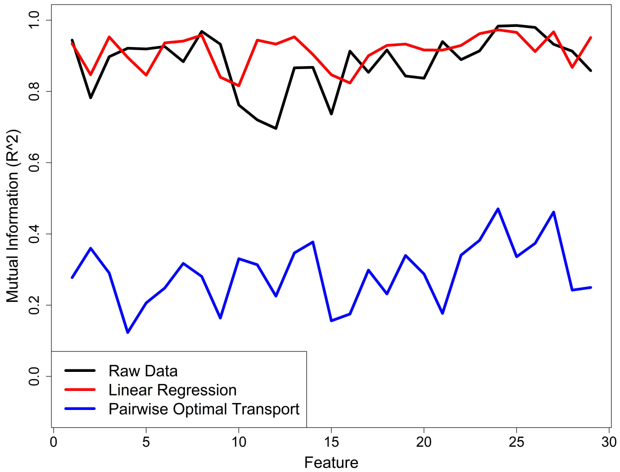


Figure 14: The relative mutual information $I_{rel}(x_i; F \setminus \{f_i\})$ between the ith feature in the CAMELS dataset and all other features is plotted (black) for each $i \in \{1, 2, ...30\}$. The relative mutual information $I_{rel}(x_i; S^F_{f_i})$ between the ith feature and all other features 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.

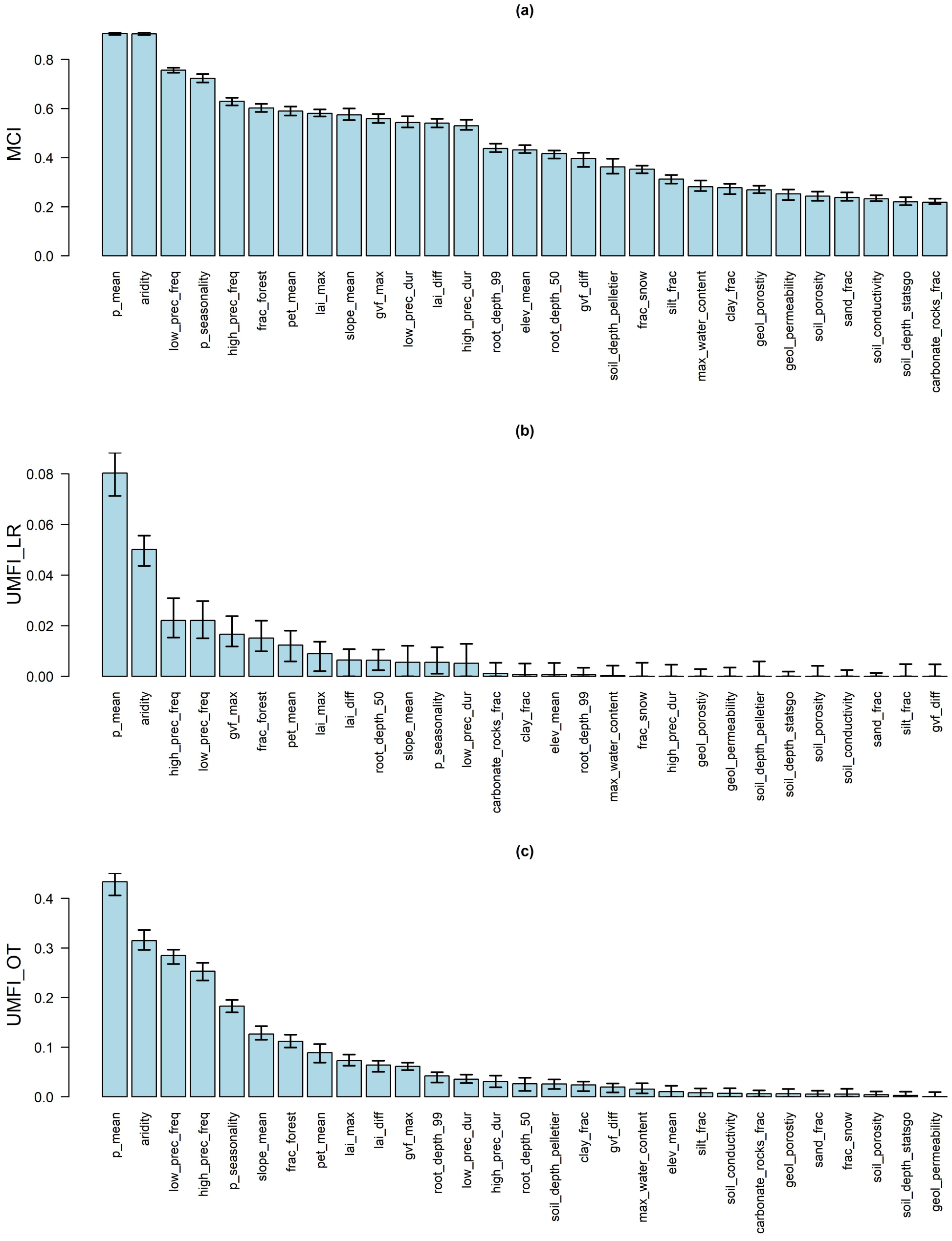


Figure 15: Median feature importance scores provided by (a) MCI, (b) UMFI with linear regression, and (c) UMFI with pairwise optimal transport, for each explanatory variable in the CAMELS dataset, taken after 100 iterations. The first and third quantiles of the scores are visualized for each feature.

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.

[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] 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.

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

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

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

[12] 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.

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

[16] Dries Debeer, Torsten Hothorn, Carolin Strobl, and Maintainer Dries Debeer. Package 'permimp'. 2021.

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

[18] 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.

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

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

[21] Nimrod Harel, Ran Gilad-Bachrach, and Uri Obolski. Inherent inconsistencies of feature importance. *arXiv preprint arXiv:2206.08204*, 2022.

[22] 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.

[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] 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.

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

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

[27] 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.

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

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

[30] 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.

[31] 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.

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

[33] Bastian Steudel and Nihat Ay. Information-theoretic inference of common ancestors. *Entropy*, 17(4):2304–2327, 2015.

[34] 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.

[35] 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.

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

[37] Caroline Uhler. Gaussian graphical models: An algebraic and geometric perspective. *arXiv preprint arXiv:1707.04345*, 2017.

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

[39] 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.

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

[41] 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.

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