

Disentangled Feature Importance

June 26, 2025

Abstract

Feature importance quantification in the presence of arbitrary dependencies remains a fundamental challenge in modern statistics. We introduce Disentangled Feature Importance (DFI), a nonparametric framework that transforms features into an independent latent space via optimal transport, computes importance for the disentangled features, and attributes importance back to original features through transport map sensitivities.

We establish that under squared-error loss, two major classes of feature importance measures—those based on response submodel refitting and those using conditional covariate resampling—target identical population functionals, revealing their shared vulnerability to correlation. Our disentangled importance functional provably eliminates this distortion in the latent space while preserving total predictive variability when attributed back to original features. We develop semiparametric theory for general transport maps and explicitly derive the efficient influence functions for Bures-Wasserstein map. Under mild conditions, we establish root- n consistency and asymptotic normality, with our estimators exhibiting double robustness: the error vanishes when the product of regression and transport map estimation errors is of smaller order than root- n .

DFI extends the classical R^2 decomposition from multiple linear regression to nonparametric settings, providing a principled framework for feature importance measures and uncertainty quantification under arbitrary dependencies. Unlike existing feature importance measures that require response submodel fitting or covariate conditional distribution resampling, DFI leverages the independence structure of transformed features to achieve computational efficiency while maintaining statistical validity.

1 Introduction

Feature importance quantification lies at the heart of interpretable machine learning and statistics, yet a fundamental challenge undermines its reliability: when predictors are correlated, standard importance measures can severely underestimate or misattribute the influence of key variables (Verdinelli and Wasserman, 2024a,b). This distortion affects critical applications across domains, from identifying disease-driving genes in genomics (Du et al., 2025b,a) to understanding large language models (Kang et al., 2025), where features naturally exhibit complex dependencies. As predictive models increasingly inform high-stakes decisions, the need for importance measures that accurately handle dependent predictors has become paramount.

Consider the regression task of predicting an outcome $Y \in \mathbb{R}$ from covariates $X \in \mathbb{R}^d$. The regression function $f(x) = \mathbb{E}[Y \mid X = x]$ captures the true predictive relationship, which we estimate from data using a fitted black-box model $\hat{f}(x) = \hat{\mathbb{E}}[Y \mid X = x]$. Feature importance methods aim to decompose the predictive signal of f (or $\hat{f}$) across the d input dimensions, revealing which features contribute most to the model’s predictions. Feature importance methods seek to decompose this predictive signal across the d input features. The literature distinguishes between *statistical feature importance*, which measures each feature’s role in the true function f , and *algorithmic feature importance*, which quantifies contributions to a specific model $\hat{f}$ without reference to the underlying distribution. We adopt the statistical perspective, focusing on population-level importance that reflects the true data-generating process.

Traditional feature importance methods—including Shapley values (Shapley, 1953), leave-one-covariate-out (LOCO) (Lei et al., 2018), and conditional permutation importance (CPI) (Chamma et al., 2023)—share a critical limitation: they dramatically underestimate the importance of features that depend on other predictors (Verdinelli and Wasserman, 2024b). This distortion arises from a fundamental assumption that removing or permuting a feature leaves the joint distribution of remaining features unchanged. When features are dependent, this assumption fails catastrophically, leading to severe underestimation or complete misidentification of important features, as we demonstrate in Examples 5 and 6.

Rather than patching existing methods with ad-hoc corrections, we propose a fundamentally different approach: transform the feature space to eliminate dependencies before computing importance. Our framework operates in three stages. First, we map the original features X to a disentangled representation $Z = T(X)$ with independent coordinates via a transport map $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$. Second, we compute feature importance in this latent space using resampling methods on the independent disentangled features. Third, we attribute importance back to the original features through the inverse transport map, yielding interpretable scores that properly account for feature dependencies.

This disentanglement strategy is motivated by a classical result in statistics: the decomposition of R^2 for linear regression with correlated covariates.

Example 1 (Linear regression with correlated covariates). Consider a multiple linear model:

$$Y = \beta^\top X + \epsilon \tag{1}$$

with $X \sim \mathcal{N}_d(0, \Sigma)$ for some covariance matrix $\Sigma \in \mathbb{S}_+^d$ and ϵ being an independent noise of X . Without loss of generality, suppose Y and X_j 's all have zero mean and unit variance. In this case, Σ is also the correlation matrix of the covariates, and we also have $\Sigma_{jj} = \sum_{i=1}^d (\Sigma^{\frac{1}{2}})_{ji}^2 = 1$ for $j = 1, \dots, d$. The data model can be characterized through only two parameters Σ and β , and the *coefficient of determination* is given by $R^2 = \beta^\top \Sigma \beta$. [Genizi \(1993\)](#) showed a decomposition

$$R^2 = \sum_{j=1}^d c_j^2 = \sum_{l=1}^d v_l^2,$$

where $c_j^2 = (e_j^\top \Sigma^{\frac{1}{2}} \beta)^2$ is the component assigned to the decorrelated feature Z_j and $v_l^2 = \sum_{j=1}^d (\Sigma^{\frac{1}{2}})_{jl}^2 c_j^2$ is the component assigned to feature X_l .

In the above example, R^2 also represents the intrinsic variation $\mathbb{V}[\beta^\top X]$. Hence, the decomposition attributes each feature an importance that sums up to the intrinsic variation. Our disentangled feature importance framework extends this classical decomposition to nonparametric settings, providing a principled way to measure feature importance in the presence of arbitrary dependencies.

1.1 Motivating applications

The need for correlation-aware feature importance extends across numerous domains:

Example 2 (Genomics and systems biology). In genetic perturbation experiments, researchers seek to understand how perturbing individual genes affects cellular behavior. Here, X represents baseline gene expression levels, A indicates which gene is perturbed, and Y denotes post-perturbation expression. Modern deep learning models and foundation models ([Du et al., 2022](#); [Moon et al., 2025](#); [Cui et al., 2024](#)) predict these outcomes, but genes often exhibit complex co-expression patterns. Understanding which genes truly drive predictions requires disentangling these correlations.

Example 3 (Natural language processing). In text generation tasks, X represents input tokens and Y represents output tokens. Understanding how each input token influences the output is crucial for model interpretability ([Kang et al., 2025](#)). However, tokens exhibit strong sequential dependencies and semantic correlations, making traditional importance measures unreliable.

Example 4 (Causal inference). When estimating conditional average treatment effects $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$ under the unconfoundedness assumption, understanding which covariates X are most important for treatment effect heterogeneity is essential ([Hines et al., 2022](#)). However, confounders are often correlated, and failing to account for these

correlations can lead to incorrect conclusions about which variables drive treatment effect variation.

1.2 Contributions

Our work makes several key contributions to the feature importance literature:

- **Critique and unification of existing measures:** In Section 2, we formally establish the equivalence between two major classes of feature importance measures under squared-error loss: LOCO, which requires refitting regression submodels after feature exclusion, and CPI, which requires resampling from conditional covariate distributions. We prove that both methods target the identical population functional—the expected conditional variance $\mathbb{E}[\mathbb{V}[\mu(X) \mid X_{-j}]]$. Despite this theoretical equivalence, we demonstrate through concrete examples that both approaches suffer from correlation/dependency distortion.
- **A new feature importance measure:** To address this core limitation, in Section 3, we introduce *Disentangled Feature Importance (DFI)*, a nonparametric framework that provides a principled way to attribute importance in the presence of arbitrary feature dependencies. DFI operates through a three-stage approach: (i) it transforms the original features into an independent latent space via a transport map; (ii) it computes importance in this disentangled representation; and (iii) it attributes importance back to the original features based on the sensitivities of the transport map. This process yields scores that reflect each feature’s intrinsic predictive contribution, accounting for the dependencies among features.
- **Semiparametric inferential theory:** In Section 4, we establish a semiparametric inference framework for DFI using general transport maps. Under the Bures-Wasserstein transport map, we derive the explicit form of efficient influence functions for both the latent and the attributed feature importance measures, prove their $\sqrt{n}$ -consistency and asymptotic normality under mild conditions, and demonstrate that our estimators possess a desirable double robustness property.
- **Nonparametric R^2 decomposition:** In Section 4.3, we show that DFI provides a direct nonparametric generalization of the classical R^2 decomposition for multiple linear models. This result establishes a clear theoretical lineage for our method, demonstrating that it preserves and meaningfully distributes the total predictive variability, $\mathbb{V}[\mathbb{E}[Y \mid X]]$, across all features, regardless of their dependency structure.
- **Computational efficiency:** Our DFI framework is designed for practical application. By leveraging the independence structure of the transformed features, our estimation procedure avoids both the computational burden of refitting numerous submodels (a drawback of LOCO) and the statistical challenge of estimating high-dimensional conditional distributions (a bottleneck for CPI). The practical utility of DFI is demonstrated through simulation studies in Section 5 and an analysis of a real-world HIV-1 dataset in Section 6.

1.3 Notation

For a vector $X \in \mathbb{R}^d$, $X_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$ denotes a sub-vector of X indexed by $\mathcal{S} \subseteq [d]$, and $X_{-j} := X_{[d] \setminus j}$. The j -th standard basis vector is denoted by e_j . The cardinality of a set $\mathcal{S}$ is denoted by $|\mathcal{S}|$. The indicator function is denoted by $\mathbb{1}\{\cdot\}$.

The partial derivative of a function f with respect to its j -th argument is denoted by $\partial_{x_j} f$. The gradient of f is denoted by ∇f . The space of twice continuously differentiable functions on $\mathbb{R}^d$ is denoted by $\mathcal{C}^2(\mathbb{R}^d)$. The Lipschitz constant for a function f is denoted by $\text{Lip}(f)$. For (potentially random) measurable functions f , we denote expectations with respect to the data-generating distribution of O alone by $\mathbb{P}f(O) = \int f d\mathbb{P}$, while $\mathbb{E}[f(O)]$ marginalizes out all randomness from both O and any nuisance functions f is dependent on. The empirical expectation over n samples is denoted by $\mathbb{P}_n f(O) = \frac{1}{n} \sum_{i=1}^n f(O_i)$. The L_2 norm of a function with respect to a measure $\mathbb{P}_X$ is denoted by $\|\cdot\|_{L_2(\mathbb{P}_X)}$.

For a symmetric matrix Σ , its (unique) positive semi-definite square root and its inverse are denoted by $\Sigma^{\frac{1}{2}}$ and $\Sigma^{-\frac{1}{2}}$, respectively. The set of symmetric and positive definite $d \times d$ matrices is denoted by $\mathbb{S}_+^d$. The Löwner order is denoted by $\preceq$; that is, $A \preceq B$ means that $B - A$ is positive semi-definite. The operator norm of a matrix is denoted by $\|\cdot\|_{\text{op}}$, and its trace is denoted by $\text{tr}(\cdot)$.

For a tuple of random vectors $O = (X, Y)$, the expectation and probability over the joint distribution $\mathbb{P}$ are denoted by $\mathbb{E}$ and $\mathbb{P}$, respectively. The marginal measure of X is denoted by $\mathbb{P}_X$, and its cumulative distribution function (CDF) is F_X . For a statistical estimand ϕ , we write $\phi(\mathbb{P})$ to emphasise its dependence on the underlying distribution $\mathbb{P}$. The population and empirical variances are denoted by $\mathbb{V}$ and $\mathbb{V}_n$. Statistical independence is denoted by $\perp\!\!\!\perp$. The pushforward of a measure $\mathbb{P}_X$ by a map T is denoted by $T_{\#}\mathbb{P}_X$. The d -dimensional multivariate normal distribution with mean μ and covariance Σ is denoted by $\mathcal{N}_d(\mu, \Sigma)$.

We use “ o ” and “ $\mathcal{O}$ ” to denote the little- o and big- $\mathcal{O}$ notations; “ $o_{\mathbb{P}}$ ” and “ $\mathcal{O}_{\mathbb{P}}$ ” are their probabilistic counterparts. For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if $a_n = \mathcal{O}(b_n)$; and $a_n \asymp b_n$ if $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$. Convergence in distribution is denoted by “ $\xrightarrow{d}$ ”.

2 Feature importance measure

Feature importance has been explored from several complementary perspectives that differ in both their inferential targets and the assumptions they impose on the data-generating process. Broadly speaking, *algorithmic* approaches, including Shapley value and its fast approximations (Shapley, 1953; Lundberg and Lee, 2017), attribute the prediction of a fixed (possibly black-box) model f to its input coordinates; the distribution of (X, Y) appears only through the empirical sample on which f was trained. In contrast, *statistical* or *population* approaches regard f as the true regression function and seek functionals of the joint law of (X, Y) that decompose predictive risk.

Classical representatives include leave-one-covariate-out (LOCO) and its Shapley-value reinterpretation (Verdinelli and Wasserman, 2024b), Conditional Permutation Importance (CPI) and its block- or Sobol-style refinements (Chamma et al., 2023; Reyero Lobo et al., 2025; Chamma et al., 2024), as well as doubly robust, influence-function based estimators of variable importance (VIM, SPVIM) under general loss functions (Williamson et al., 2021,

2023; Williamson and Feng, 2020). For causal inference, extensions to local feature importance recast these ideas, either through LOCO-type refitting (Hines et al., 2022; Dai et al., 2024; Lundborg et al., 2024) or CPI-type permutation schemes that circumvent nuisance estimation (Paillard et al., 2025). A parallel literature in global sensitivity analysis Sobol’ (1990) distributes variance via functional ANOVA or Sobol indices of independent inputs and has recently been connected to partial dependence (Owen and Prieur, 2017; Herbringer et al., 2024). For a more detailed review of the related literature, the readers are referred to Hooker et al. (2021). These seemingly disparate importance measures coincide only under restrictive independence assumptions; however, strong feature correlations can induce severe *correlation distortion* whereby genuinely influential variables receive negligible scores.

Recognising these limitations, we focus on three conceptually simple yet widely used importance measures, LOCO, CPI, and Shapley value, that serve as building blocks for many of the above methods. The remainder of this section formalizes their definitions and highlights their connections.

Consider a regression problem when $Y \in \mathbb{R}$ is the response and $X \in \mathbb{R}^d$ is the feature. Let $\mu(x) = \mathbb{E}[Y \mid X = x]$ denote the regression function using all features, and $\mu_{\mathcal{S}}(x_{\mathcal{S}}) := \mathbb{E}[Y \mid X_{\mathcal{S}} = x_{\mathcal{S}}]$ denote the regression function using a subset of features $X_{\mathcal{S}}$ for any $\mathcal{S} \subseteq [d]$. In particular, for a given coordinate $j \in [d]$, define the regression function of Y given X_{-j} as $\mu_{-j}(x_{-j}) = \mathbb{E}[Y \mid X_{-j} = x_{-j}]$.

2.1 LOCO

Leave-One-Covariate-Out (LOCO) metric and assumption-lean inference framework such as conformal prediction for LOCO have been studied by Lei et al. (2018); Rinaldo et al. (2019). Consider a loss function ℓ (e.g., the negative squared error loss, classification loss, etc.), the LOCO parameter for X_j is defined as

$$\psi_{X_j}^{\text{LOCO}} = \mathbb{E}[\ell(\mu(X), Y) - \ell(\mu_{-j}(X_{-j}), Y)],$$

which quantifies the change of expected loss when dropping X_j from the full model. Up to scaling, ψ_{loco} is a nonparametric version of the usual R^2 from standard regression.

Consider fitting estimated regression function $\hat{\mu}$ and $\hat{\mu}_{-j}$ from m independent and identically distributed samples of (X, Y) . We then compute an estimate of the LOCO parameter on n additional samples of (X, Y) :

$$\hat{\psi}_{X_j}^{\text{LOCO}} = \mathbb{P}_n\{\ell(\hat{\mu}(X), Y) - \ell(\hat{\mu}_{-j}(X_{-j}), Y)\}.$$

For LOCO, there are two nuisance functions (μ, μ_{-j}) to be estimated from data.

2.2 CPI

The traditional permutation importance approach ignores the correlation among features and may lead to uncontrolled type-1 errors. To mitigate this issue, conditional permutation importance has been proposed in different contexts: for random forests (Strobl et al., 2008) and for general predictors under ℓ_2 -loss (Chamma et al., 2023) and general loss functions (Reyero Lobo et al., 2025). The key idea is to sample from the conditional distribution

$p(X_j \mid X_{-j})$. Thus, it can also be interpreted as replace-one-covariate, or swap-one-covariate feature importance.

Define a random vector $X^{(j)}$ such that $X_{-j}^{(j)} = X_{-j}$ and $X_j^{(j)} \sim p(X_j \mid X_{-j})$ is an independent copy of X_j given all the other covariate X_{-j} , which is also independent of the outcome Y . The CPI parameter is then defined as:

$$\psi_{X_j}^{\text{CPI}} = \frac{1}{2} \mathbb{E}[\ell(\mu(X), Y) - \ell(\mu(X^{(j)}), Y)],$$

Similarly, we can compute an estimate of the CPI parameter as:

$$\hat{\psi}_{X_j}^{\text{CPI}} = \frac{1}{2} \mathbb{P}_n \{ \ell(\hat{\mu}(X), Y) - \ell(\hat{\mu}(X^{(j)}), Y) \},$$

with $\hat{X}_j^{(j)} \sim \hat{p}(X_j \mid X_{-j})$.

Instead of estimating the conditional density function $\hat{p}(X_j \mid X_{-j})$ directly, we can sample $X_j^{(j)}$ by estimating the regression function $\mathbb{E}[X_j \mid X_{-j}]$ and bootstrapping the residuals. This approach requires that X_j can be written as $X_j = \nu_j(X_{-j}) + \epsilon_j$ where $\epsilon_j \perp X_{-j}$ and $\mathbb{E}[\epsilon_j] = 0$ (Reyero Lobo et al., 2025, Lemma 3.2). Under this additive structure, one draws an independent sample W of X and sets $\hat{X}_j^{(j)} = \hat{\nu}_j(X_{-j}) + (W_j - \hat{\nu}_j(W_{-j}))$. While this independence assumption is restrictive in general, it holds naturally when features are independent or when covariates follow a multivariate Gaussian distribution. For CPI, this approach requires estimating two nuisance functions: the regression functions of the response and the j th covariate (μ, ν_j) .

2.3 Shapley value

Originating from cooperative game theory (Shapley, 1953), the *Shapley value* attributes a model's predictive ability to individual covariates. Let $[d] = \{1, \dots, d\}$ index the features and let $v(\mathcal{S})$ be a value function that measures the expected utility obtained when only the subset $\mathcal{S} \subseteq [d]$ is available. Throughout this paper, we adopt the squared error as the value function:

$$v(\mathcal{S}) := \mathbb{E}[\ell(\mu_{\mathcal{S}}(X_{\mathcal{S}}), Y)], \quad \ell(\hat{y}, y) = -(\hat{y} - y)^2,$$

where $\mu_{\mathcal{S}}(x_{\mathcal{S}}) = \mathbb{E}[Y \mid X_{\mathcal{S}} = x_{\mathcal{S}}]$ is the oracle regression trained on $\mathcal{S}$ alone. The Shapley importance of feature j is

$$\psi_{X_j}^{\text{SHAP}} = \sum_{\mathcal{S} \subseteq [d] \setminus \{j\}} \frac{|\mathcal{S}|! (d - |\mathcal{S}| - 1)!}{d!} \{v(\mathcal{S} \cup \{j\}) - v(\mathcal{S})\},$$

or, equivalently, the expectation over a random permutation π of the incremental gain achieved when j enters the model immediately after its predecessors in π . The vector $(\psi_{X_1}^{\text{SHAP}}, \dots, \psi_{X_d}^{\text{SHAP}})$ is the unique attribution satisfying the four canonical Shapley axioms (efficiency, symmetry, dummy, additivity).

As shown by Verdinelli and Wasserman (2024b), each term $v(\mathcal{S} \cup \{j\}) - v(\mathcal{S})$ coincides with LOCO for subset $\mathcal{S}$, i.e., $\psi_{X_j}^{\text{LOCO}}(\mathcal{S} \cup \{j\}) := \mathbb{E}[(\mu_{\mathcal{S} \cup \{j\}}(X_{\mathcal{S} \cup \{j\}}) - \mu_{\mathcal{S}}(X_{\mathcal{S}}))^2]$ under ℓ_2 -loss.

Hence, the Shapley value is a weighted average of LOCO scores across all 2^{d-1} submodels:

$$\psi_{X_j}^{\text{SHAP}} = \sum_{\mathcal{S} \subseteq [d] \setminus \{j\}} w_{\mathcal{S}} \psi_{X_j}^{\text{LOCO}}(\mathcal{S} \cup \{j\}), \quad w_{\mathcal{S}} := \frac{(|\mathcal{S}| + 1)! (d - |\mathcal{S}|)!}{d!}.$$

Even under linear models, this averaging yields a highly non-linear functional of the joint distribution of (X, Y) .

Exact evaluation of $\psi_{X_j}^{\text{SHAP}}$ is infeasible beyond very small d ; practical implementation relies on submodel subsampling (Williamson and Feng, 2020). Because the value function itself must be estimated for every sampled coalition, the resulting estimator is both statistically and computationally intensive, complicating further tasks such as variance estimation and formal inference. These difficulties motivate the search for importance measures that permit computationally tractable estimation and valid uncertainty quantification. For these reasons, we henceforth focus primarily on analyzing the properties of LOCO and CPI in the following subsection, while Shapley values are retained as a benchmark in the simulation study.

2.4 Properties of LOCO and CPI

Throughout the rest of the paper, we focus on the ℓ_2 -loss $\ell(\hat{y}, y) = -(\hat{y} - y)^2$. We begin by establishing two fundamental properties of LOCO and CPI that will motivate our proposed framework.

First, both measures correctly assign zero importance to truly irrelevant features, as formalized below.

Lemma 2.1 (Null features). Under null hypothesis $\mathcal{H}_{0j} : X_j \perp\!\!\!\perp Y \mid X_{-j}$, it holds that $\psi_{X_j}^{\text{CPI}} = \psi_{X_j}^{\text{LOCO}} = 0$.

Second, LOCO and CPI are equivalent under squared-error loss and can be expressed in terms of conditional variance.

Lemma 2.2 (Equivalence under ℓ_2 -loss). For ℓ_2 -loss $\ell(\hat{y}, y) = -(\hat{y} - y)^2$, the two importance measures coincide: $\psi_{X_j}^{\text{LOCO}} = \psi_{X_j}^{\text{CPI}} = \mathbb{E}[\mathbb{V}[\mu(X) \mid X_{-j}]]$.

The conditional variance $\mathbb{E}[\mathbb{V}[\mu(X) \mid X_{-j}]]$ is known as the upper Sobol index (Sobol', 1990; Owen and Prieur, 2017), which measures how much uncertainty in the prediction $\mu(X)$ remains after observing all features except X_j . This connection to sensitivity analysis provides additional interpretability to these importance measures. While Hooker et al. (2021, Theorem 2) established a similar equivalence result¹ using functional analysis, our proof relies only on elementary probability theory. Specifically, we exploit the alternative variance representation $\mathbb{V}[W] = \mathbb{E}[(W - \mathbb{E}[W])^2] = \mathbb{E}[(W - W')^2]/2$, where W' is an independent copy of W , for any random variable W . This representation reveals a fundamental equivalence under ℓ_2 -loss: LOCO's approach of refitting the response submodel $Y \mid X_{-j}$ yields the same importance measure as CPI's approach of resampling from the conditional covariate

¹We note that the definition of "conditional variable importance" in Hooker et al. (2021) appears to have a factor of 1/2 missing compared to CPI formulation.

distribution $X_j \mid X_{-j}$, with CPI having an additional factor of $1/2$ to account for the doubled variability from comparing two random realizations. While this equivalence is specific to squared-error loss, the two approaches may yield different importance measures under other loss functions.

Despite these appealing properties, both LOCO and CPI suffer from a fundamental limitation known as correlation distortion. As noted by [Verdinelli and Wasserman \(2024b\)](#), *“its value depends not just on how strongly $\mu(X)$ depends on X_j , but also on the correlation between X_j and the other features. In the extreme case of perfect dependence, it is 0. This could lead users to erroneously conclude that some features are irrelevant even when $\mu(X)$ strongly depends on X_j .”* The following examples illustrate this critical issue.

Example 5 (Dependent features with a bijective mapping). Consider a simple linear model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon \quad (2)$$

with $X_1 = X_2 = U$. Under the ℓ_2 -loss $\ell(\hat{y}, y) = -(\hat{y} - y)^2/2$,

$$\psi_{X_1}^{\text{LOCO}} = \psi_{X_2}^{\text{LOCO}} = \psi_{X_1}^{\text{CPI}} = \psi_{X_2}^{\text{CPI}} = 0.$$

Thus, LOCO/CPI consistently declare both unimportant, regardless of the value of (β_1, β_2) .

Example 6 (Dependent features with an injective mapping). Now suppose $X_1 = X_2^2 = U^2$ with U symmetric in the same model (2). Then

$$\psi_{X_1}^{\text{LOCO}} = \psi_{X_1}^{\text{CPI}} = 0, \quad \psi_{X_2}^{\text{LOCO}} = \psi_{X_2}^{\text{CPI}} = \beta_2^2 \mathbb{V}(U).$$

Thus, LOCO/CPI down-weights X_1 entirely: when $\beta_1 \neq 0$ but $\beta_2 = 0$, neither of the features is viewed as important by LOCO/CPI. Furthermore, although X_2 is assigned nonzero importance, the assigned value only captures the linear contribution $\beta_2^2 \mathbb{V}(U)$, which underestimates the total predictive signal $\mathbb{V}(\mathbb{E}[Y \mid U]) = \mathbb{V}(\beta_1 U^2 + \beta_2 U)$. Thus, even when only one feature is treated as important, the full predictive variability is still not accurately reflected.

These examples highlight a fundamental limitation of LOCO and CPI: when features exhibit strong dependencies, these methods can either completely miss important variables (assigning zero importance to genuinely influential features) or severely underestimate their predictive contribution. This also undermines the interpretability and reliability of these importance measures in real-world applications where feature dependencies are the norm rather than the exception. This motivates the need for a new feature importance notion that is robust to feature dependencies. These failures underscore the need for a new notion of feature importance that is robust to feature dependencies.

3 Disentangled feature importance

3.1 Disentangled representations

As illustrated by the previous examples, existing measures such as LOCO and CPI can be severely distorted when strong dependencies exist among input features. To overcome this limitation, we propose to *transform the features into a disentangled space* before computing feature importance.

Specifically, we seek a transformation $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that

$$Z = T(X)$$

has *independent* coordinates, ideally matching a simple reference distribution such as a standard multivariate Gaussian $\mathcal{N}_d(0, I)$. Without loss of generality, we assume Z_j has zero mean and unit variance for all $j = 1, \dots, d$. This approach provides a principled, nonparametric method for decorrelating features while preserving the essential structure of the data.

Once the disentangled representation $Z = (Z_1, \dots, Z_d)$ is obtained, we define the feature importance for each coordinate Z_j as follows. Let $\mu(x) = \mathbb{E}[Y \mid X = x]$ and $\eta(z) := \mu(T^{-1}(z))$ denote the regression functions in the two feature spaces. For each feature $j \in [d]$, we define the feature importance measure under ℓ_2 -loss:

$$\phi_{Z_j}(\mathbb{P}) := \mathbb{E}[\mathbb{V}(\eta(Z) \mid Z_{-j})]. \quad (3)$$

From Lemma 2.2, this measure coincides with both LOCO and CPI, except it is defined for the latent feature Z instead of the raw feature X .

Since the coordinates of Z are independent (or at least weakly dependent), the correlation distortion that affected traditional importance measures on X is mitigated. Importantly, the feature importance measure ϕ_{Z_j} defined in (3) accurately reflects the intrinsic predictive contribution of each feature.

This property is formalized in the following proposition. Thus, in the disentangled space, the importance of a latent feature faithfully identifies whether it truly influences the prediction function, without being confounded by dependency among features.

Proposition 3.1 (Free of correlation distortion). For ℓ_2 -loss, $\phi_{Z_j} = 0$ if and only if $\mathbb{E}[Y \mid Z]$ does not depend on Z_j almost surely.

3.2 Importance attribution for raw features

After computing feature importance for the disentangled features Z , our goal is to attribute importance back to the original features X . To define a meaningful importance score ϕ_{X_l} for each original feature X_l , we seek to satisfy the following properties:

- (1) Tight statistical error control: If the outcome Y does not depend on X_l conditionally, and X_l is independent of the other covariates, that is, $\mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid X_{-l}]$ and $X_l \perp\!\!\!\perp X_{-l}$, then the importance score satisfies $\phi_{X_l} = 0$.

- (2) Power to detect relevant covariates: If X_l affects the outcome, either directly or through its dependence on other covariates—formally, if $\mathbb{E}[Y \mid X] \neq \mathbb{E}[Y \mid X_{-l}]$ or $X_l \not\perp\!\!\!\perp X_{-l}$, then $\phi_{X_l} > 0$.

In other words, an ideal importance measure should be both *parsimonious* (assigning zero to truly irrelevant features) and *exhaustive* (detecting all forms of feature influence, whether direct or combinatorial).

Towards this goal, we define the feature importance score ϕ_{X_l} for each original feature X_l by transferring importance from the disentangled features Z back to X through the sensitivity of Z_j with respect to X_l . Formally, we define the disentangled feature importance (DFI) measure as

$$\phi_{X_l}(\mathbb{P}) := \sum_{j=1}^d \mathbb{E} \left[\mathbb{V} \left(\mathbb{E}[Y \mid Z] \mid Z_{-j} \right) \left(\frac{\partial X_l}{\partial Z_j} \right)^2 \right], \quad (4)$$

where $\frac{\partial X_l}{\partial Z_j}$ denotes the partial derivative of $X_l = e_l^\top T^{-1}(Z)$ with respect to Z_j .

The inner term $\mathbb{V}(\mathbb{E}[Y \mid Z] \mid Z_{-j})$ is the first-order Sobol index of Z_j , representing the “intrinsic” predictive signal uniquely attributable to that disentangled direction. Multiplying by $(\partial_{Z_j} X_l)^2$ gauges how strongly fluctuations in Z_j are expressed through X_l ; integrating over the data distribution averages these local sensitivities into a global importance score. Thus, ϕ_{X_l} quantifies how much of the irreducible signal carried by all latent directions is channelled through X_l .

3.3 Properties

We now analyze the properties of the proposed feature importance measure, focusing on its behavior with null and dependent features.

On one hand, the null features get assigned zero importance by DFI. Specifically, for any $\mathcal{S} \subseteq [d]$, if the map factorizes as $T = (T_{\mathcal{S}}, T_{\mathcal{S}^c})$, then $X_{\mathcal{S}}$ is null features (i.e., $\mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid X_{\mathcal{S}}]$ and $X_{\mathcal{S}} \perp\!\!\!\perp X_{\mathcal{S}^c}$), if and only if $\phi_{Z_j} = 0$ for all $j \in \mathcal{S}^c$, if and only if $\phi_{X_j} = 0$ for all $j \in \mathcal{S}^c$. The factorization of the map naturally holds for the optimal map on product measures on $(X_{\mathcal{S}}, X_{\mathcal{S}^c})$.

On the other hand, DFI captures truly important features among dependent features. To see this, we revisit the examples of perfectly correlated features from Examples 5 and 6 to demonstrate how the proposed measures handle feature redundancy and allocate importance in the presence of dependence.

Example 5 (Continued). For $Z_1 = X_1 = X_2 = U$ and $Z_2 \sim \mathcal{N}(0, 1)$ independent of Z_1 , we have

$$\phi_{X_1} = \phi_{X_2} = \phi_{Z_1} = (\beta_1 + \beta_2)^2 \mathbb{V}(U), \quad \phi_{Z_2} = 0.$$

Because both X_1 and X_2 have importance measures equal to the total variation of the regression function $\mathbb{V}(\mathbb{E}[Y \mid X_1, X_2]) = \mathbb{V}(\mathbb{E}[Y \mid Z])$ using Z alone, we can either include only X_1 or X_2 into a prediction model instead of both.

Example 6 (Continued). For $Z_1 = X_2 = U$ and $Z_2 \sim \mathcal{N}(0, 1)$ independent of Z_1 , we have

$$\phi_{X_1} = (\beta_1^2 \mathbb{E}[U^2] + \beta_2)/4, \quad \phi_{X_2} = \phi_{Z_1} = \mathbb{V}(\beta_1 U^2 + \beta_2 U), \quad \phi_{Z_2} = 0.$$

Because $\phi_{X_1} < \phi_{X_2} = \phi_Z$, it suggests that one can only include X_2 in the prediction model.

Compared to LOCO and CPI, DFI attributes the total predictive variability $\mathbb{V}(\mathbb{E}[Y | X])$ of the combined predictive power to each dependent feature, as shown in the above examples.

3.4 Disentangle transformation

The key component of DFI is the transform map T . Depending on the distributions of features and/or latent variables, some common choices of transport maps include:

- (1) Bures-Wasserstein map (Givens and Shortt, 1984): If $X \sim \mathcal{N}(0, \Sigma)$ and $Z \sim \mathcal{N}_d(0_d, I_d)$, the optimal map is the whitening transform $T(x) = \Sigma^{-\frac{1}{2}}x$.
- (2) Knothe-Rosenblatt transport (Rosenblatt, 1952): For any X with joint distribution function F and $Z \sim \text{Unif}[0, 1]^{\otimes d}$, the triangular Knothe-Rosenblatt transport map is given by $T(x) = (F_{X_1}(x_1), F_{X_2|X_1}(x_2 | x_1), \dots, F_{X_d|X_{[d-1]}}(x_d | x_{[d-1]}))$, which pushes X to $Z \sim \text{Unif}[0, 1]^{\otimes d}$.

For our theoretical analysis, we focus on the optimal transport map, which provides a principled method for transforming one probability distribution into another by minimizing a specified transportation cost. Consider two probability measures $\mathbb{P}_X$ and $\mathbb{P}_Z$ on $\mathbb{R}^d$. Given a cost function $c(x, z)$, the Monge problem of optimal transport seeks a measurable map $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ solving:

$$\inf_{T_{\#}\mathbb{P}_X = \mathbb{P}_Z} \int_{\mathbb{R}^d} c(x, T(x)) d\mathbb{P}_X(x), \quad (5)$$

where $T_{\#}\mathbb{P}_X = \mathbb{P}_Z$ indicates that T pushes forward $\mathbb{P}_X$ to $\mathbb{P}_Z$, i.e., $Z = T(X) \sim \mathbb{P}_Z$ whenever $X \sim \mathbb{P}_X$. In this work, we focus on the squared Euclidean distance:

$$c(x, z) = \|x - z\|^2.$$

Under this quadratic cost, optimal transport maps exhibit desirable mathematical properties, leading naturally to disentangled representations of the original random feature X .

To guarantee the existence and uniqueness of the optimal transport map, we impose the following standard assumption:

Assumption 3.1 (Absolute continuity). *The measure $\mathbb{P}_X$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$.*

Under Assumption 3.1, Brenier’s theorem (Brenier, 1991) guarantees the existence of a unique (almost everywhere) optimal transport map, represented as the gradient of a convex function:

Theorem 3.2 (Brenier’s theorem). Under Assumption 3.1, there exists a convex potential function $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ such that the optimal transport map is uniquely determined $\mathbb{P}_X$ -almost surely by:

$$T(x) = \nabla \Phi(x).$$

Consequently, the random vector $Z = T(X)$, obtained via an optimal transport map, provides a unique (up to a set of measure zero) disentangled representation of the original data X . By ensuring Z has independent coordinates, we effectively mitigate issues associated with feature correlation when attributing feature importance. In the next section, we analyze the statistical properties for estimation and inference of the target estimands defined in (3) and (4) under this setup.

4 Statistical estimation and inference

4.1 Estimation of disentangled feature importance

Suppose we are given estimators $\hat{\mu}$ for the conditional mean function μ and $\hat{T}$ for the transport map T . We define the corresponding estimator for feature importance as:

$$\hat{\phi}_{Z_j}(\mathbb{P}) = \frac{1}{2} \mathbb{P}_n \left[(Y - \mathbb{P}_{Z_j^{(j)}, m} \{ \hat{\eta}(Z^{(j)}) \})^2 - (Y - \hat{\eta}(Z))^2 \right], \quad (6)$$

where Z is the disentangled representation and $Z^{(j)}$ is a modified version of Z in which the j -th coordinate Z_j is replaced by an independent copy $Z'_j \sim Z_j$ (independently of everything else), while the other coordinates remain unchanged, $Z_{-j}^{(j)} = Z_{-j}$. Here, $\mathbb{P}_n$ denotes the empirical average over n samples of (Z, X, Y) and $\mathbb{P}_{Z_j^{(j)}, m}$ denotes the empirical average over m samples of $Z_j^{(j)}$.

From Lemma 2.2, we know that both the LOCO and CPI estimators target the same underlying feature importance functional. In this work, we adopt the CPI-type estimator for the transformed features for several reasons. First, it avoids the need to retrain a regression model for $\eta_{-j}(Z_{-j}) = \mathbb{E}[Y \mid Z_{-j}]$ after each perturbation of the latent feature; only the full regression model η needs to be estimated once, which depends on both the original regression model μ and the transport map T . Second, because the coordinates of Z are independent, sampling Z_j conditional on Z_{-j} (as required in the standard CPI procedure) simplifies to sampling Z_j independently. This simplification remains valid even when Z is estimated by $\hat{Z}$. Finally, under our construction, the conditional covariate distribution required by CPI simplifies due to independence.

To present our main results on the statistical properties of the proposed estimator, we first introduce technical assumptions. Apart from Assumption 3.1, we also require convexity of the Brenier potential.

Assumption 4.1 (Convexity). *The Brenier potential Φ is a convex function such that $\Phi \in \mathcal{C}^2(\mathbb{R}^d)$ and $\lambda^{-1}I_d \preceq \nabla^2 \Phi(x) \preceq \lambda I_d$ for all $x \in \mathbb{R}^d$ and for some universal constant $\lambda > 1$.*

Additionally, we require the following assumptions on the distributions of the covariate and response, as well as the regression function.

Assumption 4.2 (Data model). *There exists a constant $C > 0$ such that the following holds with probability tending to one as sample size tends to infinity:*

- (1) *Covariate: the random vector X is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^d$ with zero mean and bounded second moment: $\mathbb{E}[X] = 0$ and $\|X\|_{L_2} \leq C$;*
- (2) *Response: the response Y has zero mean and bounded variance: $\mathbb{E}[Y] = 0$ and $\mathbb{E}[Y^2] \leq C$;*
- (3) *Regression function: the ranges of μ and $\hat{\mu}$ are contained uniformly in $(-C, +C)$, and μ is Lipschitz continuous with Lipschitz constant $\text{Lip}(\mu) = L_\mu$.*

We can decompose the estimation error into two primary sources, namely, the error in estimating the conditional expectation μ and the error in estimating the transport map T .

Theorem 4.1 (Estimation error decomposition). Under Assumptions 3.1–4.2, if the nuisance estimator $(\hat{\mu}, \hat{T})$ of (μ, T) is independent of n observations of $O = (Z, X, Y)$, such that $\varepsilon_\mu := \|\hat{\mu} - \mu\|_{L_2(\mathbb{P}_X)} < 1/\sqrt{2}$. Denote $\Delta_j(z) := \eta(z) - \eta_{-j}(z)$. Then the correlated feature importance estimator (6) satisfies that

$$\hat{\phi}_{Z_j}(\mathbb{P}) - \phi_{Z_j}(\mathbb{P}) = (\mathbb{P}_n - \mathbb{P})\{\varphi_{Z_j}(O; \mathbb{P})\} + \mathcal{O}(n^{-1/2}\mathcal{E}_n + \mathcal{E}_n^2),$$

with probability at least $1 - \varepsilon_\mu^2$. Here, the influence function is given by

$$\varphi_{Z_j}(O; \mathbb{P}) = 2(Y - \eta(Z))\Delta_j(Z) + \Delta_j(Z)^2 - \phi_{Z_j}(\mathbb{P}),$$

and the remainder term is given by $\mathcal{E}_n := \|\hat{\mu} - \mu\|_{L_2(\mathbb{P}_X)} + \|\hat{T} - T\|_{L_2(\mathbb{P}_X)}$.

From Theorem 4.4, because the error term is second order in the estimation error of the regression function μ and the transport map T . This error term vanishes as long as the product of the two estimation errors decays faster than $n^{-1/2}$, exhibiting doubly robustness. On the other hand, as the first term of the decomposition is a centered empirical process, it converges to a standard normal distribution. These observations give rise to the following asymptotic normality, which one can use to perform inference.

Corollary 4.2 (Statistical inference). Under the same setup as in Theorem 4.1, if $\phi_{Z_j} > 0$, if $\|\hat{\mu} - \mu\|_{L_2(\mathbb{P}_X)} = o(n^{-1/4})$ and $\|\hat{T} - T\|_{L_2(\mathbb{P}_X)} = o(n^{-1/4})$, then

$$\sqrt{n}(\hat{\phi}_{Z_j}(\mathbb{P}) - \phi_{Z_j}(\mathbb{P})) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}[\varphi_{Z_j}(O; \mathbb{P})]). \quad (7)$$

The influence function can be estimated via

$$\hat{\varphi}_{Z_j}(O; \hat{\mathbb{P}}) = 2(Y - \hat{\eta}(Z)) \left(\hat{\eta}(Z) - \mathbb{E}_{Z^{(j)}}[\hat{\eta}(Z^{(j)})] \right) + \left(\hat{\eta}(Z) - \mathbb{E}_{Z^{(j)}}[\hat{\eta}(Z^{(j)})] \right)^2 - \hat{\phi}_{Z_j}$$

where $\hat{\eta} = \hat{\mu} \circ \hat{T}^{-1}$. Consequently, the asymptotic variance $\mathbb{V}[\varphi_{Z_j}(O; \mathbb{P})]/n$ can be estimated with $\hat{\sigma}^2 = \mathbb{V}_n[\hat{\varphi}(O; \hat{\mathbb{P}})]/n$.

Remark 4.1 (Confidence intervals near the null). When $\phi_{Z_j} > 0$ is independent of the sample size n , Equation (7) provides valid confidence interval $\mathcal{C}_{n,\alpha} = (\phi_{Z_j} - z_{\frac{\alpha}{2}}\widehat{\sigma}, \phi_{Z_j} + z_{\frac{\alpha}{2}}\widehat{\sigma})$ such that $\mathbb{P}(\widehat{\phi}_{Z_j} \in \mathcal{C}_{n,\alpha}) \geq 1 - \alpha$. When $\phi_{Z_j} = 0$, the influence function vanishes and asymptotic normality does not hold anymore. In more general cases when ϕ_{Z_j} is a function of n and converges to zero, finding a uniformly covered confidence interval is unsolved, due to the behavior of quadratic functionals. Though one way to mitigate such an issue is to expand the confidence interval by $\mathcal{O}(n^{-1/2})$ to maintain validity at the expense of efficiency at the null; see [Verdinelli and Wasserman \(2024b, Section 6\)](#).

4.2 Estimation of attributed feature importance

Proposition 4.3 (Efficient influence function of $\phi_{X_l}(\mathbb{P})$). Let $O = (X, Y) \sim \mathbb{P}$ with $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$. Let $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a transport map and write $Z = T(X)$. Fix $j, l \in [d]$ and set

$$\Delta_j(z) := \eta(z) - \eta_{-j}(z), \quad H_{jl}(x) := (\partial_{z_j}(T^{-1})_l(z))^2.$$

Define the parameter

$$\phi_{jl}(\mathbb{P}) := \mathbb{P}\{\Delta_j(Z)^2 H_{jl}(X)\}, \quad \theta_{jl}(\mathbb{P}) := \mathbb{P}\{H_{jl}(X)\}.$$

Assume a non-parametric model $\mathcal{M}$ in which $Y \mid X = x$ has finite second moment for $\mathbb{P}$ -a.e. x , and the efficient influence function $\varphi_{H_{jl}}$ of $\theta(\mathbb{P})$ exists. Then, ϕ_{jl} is pathwise differentiable at every $\mathbb{P} \in \mathcal{M}$ with efficient influence function

$$\begin{aligned} \varphi_{\Delta_j^2 H_{jl}}(O; \mathbb{P}) &= H_{jl}(X)\Delta_j(Z)^2 - \phi_{jl}(\mathbb{P}) + 2H_{jl}(X)[Y - \eta(Z)]\Delta_j(Z) \\ &\quad + (\varphi_{H_{jl}}(O; \mathbb{P}) - H_{jl}(X) + \theta_j(\mathbb{P}))\Delta_j(Z)^2. \end{aligned} \quad (8)$$

In general, obtaining a closed-form expression for the efficient influence function of a functional $\theta_{jl}(\mathbb{P})$ is challenging. For the class of transport maps introduced in Section 3.1, this task becomes more tractable. Below, we restrict attention to the Gaussian setting.

Assumption 4.3 (Covariate distribution). $X \sim \mathcal{N}(0, \Sigma)$ and $\lambda^{-1}I_d \preceq \Sigma \preceq \lambda I_d$ for some constant $\lambda > 1$.

Under Assumption 4.3, the optimal Brenier map is linear and therefore admits an explicit form. By Theorem 3.2, this map is $\mathbb{P}_X$ -almost surely unique, ensuring that the associated feature-importance measures ϕ_{Z_j} and ϕ_{X_l} are likewise well defined and unique. We have the following result.

Theorem 4.4 (Estimation error of $\phi_{X_l}(\mathbb{P})$ under Bures-Wasserstein transport). Under Assumptions 3.1–4.3, consider the Bures-Wasserstein transport map $Z = T(X) = L^{-1}X$ where $L := \Sigma^{\frac{1}{2}}$, and the nuisance estimator $(\widehat{\mu}, \widehat{L})$ of (μ, L) is independent of n observations of

$O = (Z, X, Y)$. The attributed feature importance estimator

$$\hat{\phi}_{X_l}(\mathbb{P}) := \frac{1}{2} \sum_{j=1}^d \mathbb{P}_n \left\{ \hat{L}_{jl}^2 [(Y - \hat{\eta}(Z^{(j)}))^2 - (Y - \hat{\eta}(Z))^2] \right\}, \quad (9)$$

satisfies that:

$$\hat{\phi}_{X_l}(\mathbb{P}) - \phi_{X_l}(\mathbb{P}) = (\mathbb{P}_n - \mathbb{P})\{\varphi_{X_l}(O; \mathbb{P})\} + \mathcal{O}_{\mathbb{P}}(n^{-1/2}\varepsilon_n + \varepsilon_n^2),$$

where the efficient influence function is given by

$$\varphi_{X_l}(O; \mathbb{P}) = \sum_{j=1}^d L_{jl}^2 [2(Y - \eta(Z))\Delta_j(Z) + \Delta_j(Z)^2] - \phi_{X_l}(\mathbb{P}),$$

and the remaining term satisfies that $\varepsilon_n := \|L_{\cdot l}\|_2 \|\hat{\mu} - \mu\|_{L_2} + \text{tr}(\Sigma)^{1/2} \|\hat{L} - L\|_{\text{op}}$.

Several comments concerning the scope and implications of Theorem 4.4 follow.

Firstly, the normality assumption on X (Assumption 4.3) is a sufficient but not necessary condition. The conclusions of Theorem 4.4 hold more generally for any invertible linear transport map whose singular values are bounded away from zero and infinity, thereby ensuring the stability of the transformation.

Secondly, the proof strategy is not confined to the optimal transport map. It can be readily extended to settings involving nonlinear or non-optimal maps, which may be motivated by computational efficiency. The principal requirement for such an extension is the derivation of the efficient influence function for $\theta(\mathbb{P})$ under the new map, as demonstrated for the Knothe–Rosenblatt map in Appendix B.6.

Thirdly, the linear expansion of ϕ_{X_l} possesses a remainder term of the same order as that for ϕ_{Z_j} in Theorem 4.1. This term is second-order in the estimation errors of the two nuisance functions, the regression function μ and the transport map L . This structure implies a form of double robustness, which is highly desirable for statistical estimation.

Finally, the feature importance ϕ_{X_l} is constructed as a weighted average of the disentangled importances $\{\phi_{Z_j}\}_{j=1}^d$. Its influence function inherits this same weighted-average structure. While individual components in the expansion of ϕ_{X_l} might appear to depend on the dimension d , the weights sum to the feature’s variance, $\mathbb{V}(X_l)$. This result frames our measure as a direct nonparametric extension of the classical variance decomposition in Example 1, as we will discuss in Section 4.3.

Remark 4.2 (Estimation of linear transformation L). *To construct the linear transformation L and the disentangled features $Z = L^{-1}X$, one must estimate both Σ and its inverse. However, the error bound in Theorem 4.4 depends only on the accuracy of $\hat{L}$. Under Assumption 4.3, the perturbation bound for matrix inversion guarantees that $\|\hat{L}^{-1} - L^{-1}\|_{\text{op}}$ is of the same order as $\|\hat{L} - L\|_{\text{op}}$, so controlling the latter suffices.*

When d is fixed and $n \rightarrow \infty$,

$$\|\hat{\Sigma} - \Sigma\|_{\text{op}} = \mathcal{O}_{\mathbb{P}}(d^{1/2}n^{-1/2}), \quad \|\hat{\Sigma}^{-1} - \Sigma^{-1}\|_{\text{op}} \leq \frac{\kappa(\Sigma)^2 \|\hat{\Sigma} - \Sigma\|_{\text{op}}}{1 - \kappa(\Sigma) \|\hat{\Sigma} - \Sigma\|_{\text{op}}},$$

where $\kappa(\Sigma) = \|\Sigma^{-1}\|_{\text{op}}\|\Sigma\|_{\text{op}}$ is the condition number. Hence $\|\hat{L} - L\|_{\text{op}}$ and $\|\hat{L}^{-1} - L^{-1}\|_{\text{op}}$ share the same $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$ rate up to $\kappa(\Sigma)^2$.

If $d \gg n$, one can combine (a) thresholding or tapering for Σ and (b) a sparse precision estimator such as the graphical Lasso

$$\hat{\Theta} = \underset{\Theta \succ 0}{\operatorname{argmin}} \left\{ -\log \det \Theta + \operatorname{tr}(\hat{\Sigma}\Theta) + \lambda \sum_{i \neq j} |\Theta_{ij}| \right\},$$

with $\lambda \asymp \sqrt{\log d / n}$ to obtain $\hat{L}^2 = \hat{\Theta}$. Under the usual sparsity and eigenvalue conditions (Wainwright, 2019, Sec. 11.2),

$$\|\hat{L}^2 - L^2\|_{\text{op}} = \|\hat{\Theta} - \Sigma^{-1}\|_{\text{op}} = \mathcal{O}_{\mathbb{P}}\left(s_{\text{eff}} \sqrt{\frac{\log d}{n}}\right),$$

where s_{eff} bounds the number of non-zeros per row/column of Σ^{-1} . These results could possibly be used to extend Theorem 4.4 to a high-dimensional setting.

4.3 Connection to R^2 decomposition of correlated linear regression

In the special case where the transport map T is a linear transformation, the relationship between the importances in the X -space and the Z -space simplifies. Specifically, if T is linear, we expect that $\phi_{X_l}(\mathbb{P})$ reduces to linear combination of $\phi_{Z_j}(\mathbb{P})$'s:

$$\phi_{X_l}(\mathbb{P}) = \sum_{j=1}^d L_{jl}^2 \phi_{Z_j}(\mathbb{P}), \quad (10)$$

where the weights depend on the entries of the linear map. Thus, in the linear case, the feature importance for X_l can be interpreted as a weighted combination of the importances of the disentangled features Z_j .

Proposition 4.5 (Decomposition of feature importance). Under the same settings as Theorem 4.4, it holds that

$$\sum_{j=1}^d \Sigma_{jj} \phi_{Z_j}(\mathbb{P}) = \sum_{l=1}^d \phi_{X_l}(\mathbb{P}). \quad (11)$$

A consequence of Proposition 4.5 is that for standardized features, where $\mathbb{V}[X_l] = 1$ for all $l = 1, \dots, d$, the total feature importance is conserved under the optimal transport map: $\sum_{j=1}^d \phi_{Z_j}(\mathbb{P}) = \sum_{l=1}^d \phi_{X_l}(\mathbb{P})$. This equality provides a crucial link to classical regression analysis. For a multiple linear regression model, our measure precisely recovers the terms of the R^2 decomposition discussed in Example 1. Therefore, this result establishes our framework as a nonparametric generalization of R^2 decomposition (Genizi, 1993).

Example 1 (Continued). Consider the multiple linear model (1), for the disentangled feature $Z = \Sigma^{-\frac{1}{2}}X \sim \mathcal{N}_d(0, I)$, one can show that $\phi_{Z_j} = (e_j^\top \Sigma^{\frac{1}{2}}\beta)^2$. Then, we attribute importance to original features $\phi_{X_l} = \sum_{j=1}^d (\Sigma^{\frac{1}{2}})_{jl}^2 \phi_{Z_j}$ based on (10), which coincides with the decomposition of the coefficient of determination R^2 for linear regression with correlated regressors (Genizi, 1993).

More generally, for any positive definite covariance matrix Σ , one has $\phi_{Z_j} = \Sigma_{jj}^{-1/2} (e_j^\top \Sigma^{\frac{1}{2}}\beta)^2$, while the attribution relationship (10) remains the same.

5 Simulation

5.1 Comparative analysis of importance attribution

In the first simulation study, we present the following four simulated examples. In each case, we plot the estimated importance of all features (averaged over 100 simulations) with a different value of correlation ρ between the features. In each case, the sample size is $n = 1000$.

- (M1) Linear Gaussian model: $Y = 5X_1 + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$ and $X \sim \mathcal{N}_{10}(0, \Sigma)$ with $\text{diag}(\Sigma) = 1$, $\Sigma_{12} = \Sigma_{21} = \rho$ and zero otherwise.
- (M2) Non-linear model: $Y = 5 \cos(X_1) + 5 \cos(X_2) + \varepsilon$, with the same covariate and noise as (M1).
- (M3) Piecewise interaction with indicators: $Y = 1.5X_1X_2 \mathbf{1}_{\{X_3 > 0\}} + X_4X_5 \mathbf{1}_{\{X_3 < 0\}} + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$ and $X \sim \mathcal{N}_5(0, \Sigma)$, $\Sigma_{jj} = 1$, $\Sigma_{12} = \Sigma_{21} = \Sigma_{45} = \Sigma_{54} = \rho$, and 0 otherwise.
- (M4) Low-density model: $Y = 5X_1 + \varepsilon$, where $X_{-2} \sim \mathcal{N}_9(0, 1)$, $X_2 = 3X_1^2 + \delta$ and $(\delta, \varepsilon) \sim \mathcal{N}_2(0, 1)$.

In our analysis, we denote the feature importance measures derived from our framework as DFI (ϕ_{X_l}) and DFI-Z (ϕ_{Z_j}). We compare their performance with several established methods: LOCO, nLOCO (normalized LOCO; Verdinelli and Wasserman, 2024b), dLOCO (decorrelated LOCO; Verdinelli and Wasserman, 2024a), and the Shapley value (Shapley, 1953). For dLOCO and Shapley values, we use sampling approximation as described in Verdinelli and Wasserman (2024b), with 1000 and 100, samples respectively, to reduce computational burden. All methods employ two-fold cross-fitting, where one fold estimates the nuisance functions (regression models and transport maps) and the other fold computes the importance measures. For nuisance function estimation, we use a random forest with 500 trees and a minimum of five samples per leaf. The comparative results for the four simulation models are presented in Figures 1–4.

In the linear model (M1), as shown in Figure 1, all methods perform well under weak correlation ($\rho = 0.2$), correctly identifying X_1 as the important feature. However, under strong correlation ($\rho = 0.8$), the importance value of LOCO and nLOCO diminishes for X_1 in the presence of a correlated feature X_2 , illustrating the correlation distortion issue noted by

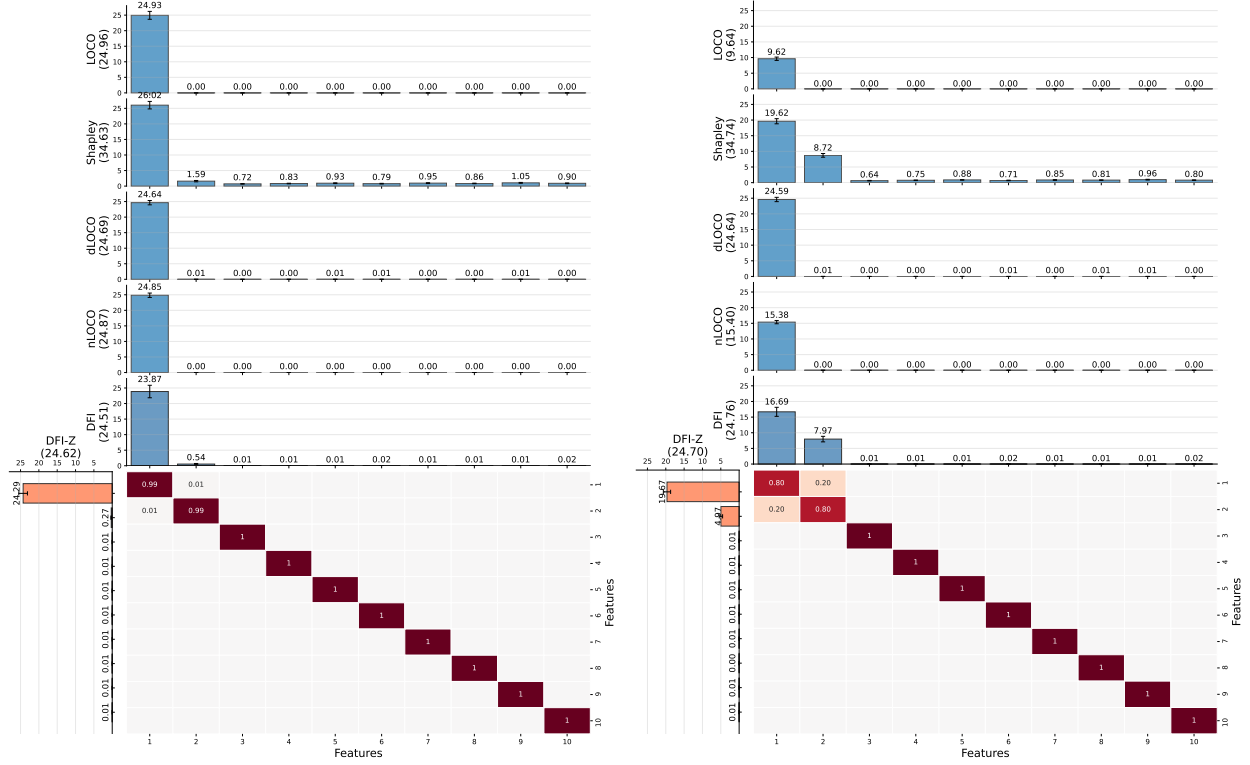


Figure 1: Simulation results under (M1). (a) weak correlation ($\rho = 0.2$); (b) strong correlation ($\rho = 0.8$). For every method, the number shown in parentheses is the total importance; this sum should converge to the signal variance $\mathbb{V}[\mathbb{E}[Y | X]] = 25$ when the covariates are independent ($\rho = 0$). Bars give the mean over 100 random seeds, and the error bars indicate the corresponding standard deviations. The heatmap on the right visualizes the weights $(\Sigma^{\frac{1}{2}})_{jl}^2$ that transfer importance from the latent coordinates to the observed features.

Verdinelli and Wasserman (2024a). In contrast, dLOCO, Shapley, and DFI correctly assign a larger importance value (≥ 16) to X_1 , while only the latter two measures acknowledge X_2 to be informative, though less important than X_1 . In addition, only the total importance calculated by DFI is approximately equal to the total predictive variability, $\mathbb{V}(\mathbb{E}[Y | X])$.

For the nonlinear model (M2), presented in Figure 2, all methods correctly identify the two equally important features, X_1 and X_2 . However, the LOCO variants and the Shapley value substantially underestimate their importance. Only DFI accurately captures the total predictive variability and distributes it equally between the two features. Similar results are observed for the piecewise interaction model (M3) in Figure 3, where LOCO-based methods again underestimate the importance of the correlated features.

Finally, we consider model (M4), where X_1 and X_2 exhibit a non-linear relationship but are linearly uncorrelated. This represents a misspecified setting for our method, as DFI seeks the best linear projection to explain the data generating process. As shown in Figure 4, the LOCO variants assign almost all importance to X_1 . In contrast, both Shapley and DFI recognize X_2 's relevance. Although the true model depends solely on X_1 , knowing X_2 reduces uncertainty about the magnitude of X_1 , which may be valuable for prediction.

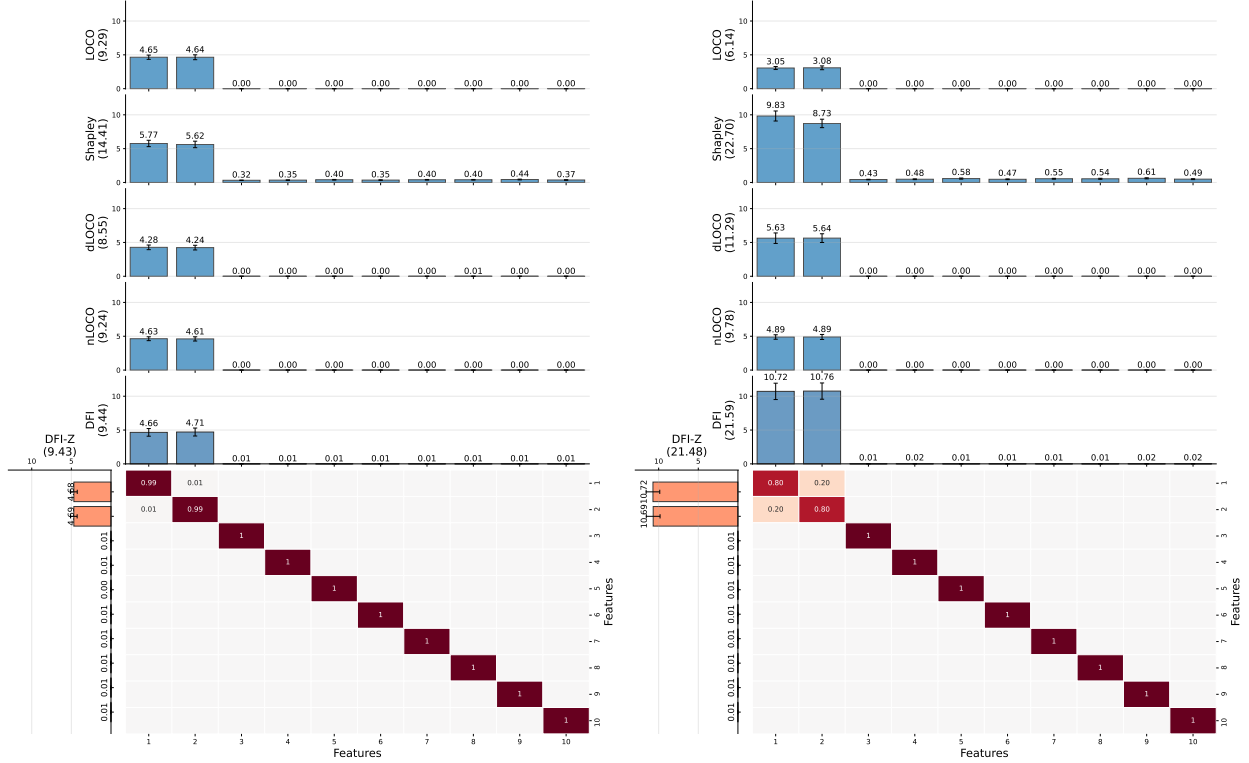


Figure 2: Simulation results for model (M2). (a) weak correlation ($\rho = 0.2$); (b) strong correlation ($\rho = 0.8$). The interpretation of the bars, error bars, and heat-map is identical to Figure 1. The number in parentheses is the total estimated importance, which should be close to the signal variance $\mathbb{V}[\mathbb{E}[Y | X]] = 25 + 25e^{-2} - 100e^{-1} + 50e^{-1} \cosh(\rho)$ (which is ≈ 10 when $\rho = 0$, and ≈ 20 when $\rho = 1$).

With the optimal nonlinear transport map, this discrepancy would vanish. However, under the linear transport constraint, DFI’s partial attribution to X_2 represents the best linear approximation to the true disentangled representations under model misspecification.

Across all scenarios, our proposed DFI framework provides a more consistent and interpretable assessment of feature importance than the compared methods, especially in the presence of feature correlations and non-linearities.

5.2 Inferential validity and computational performance

Beyond providing accurate point estimates, a practical feature importance measure must allow for valid statistical inference and be computationally tractable. In our second simulation study, we evaluate these quantitative aspects of DFI and its competitors. We assess inferential validity by examining the empirical coverage rates of the confidence intervals constructed for the importance measures. As shown in Table 1, DFI, along with the other methods, achieves average coverage at or above the nominal 95% level, confirming the validity of our asymptotic theory.

However, a key practical advantage of DFI lies in its computational efficiency. As illus-

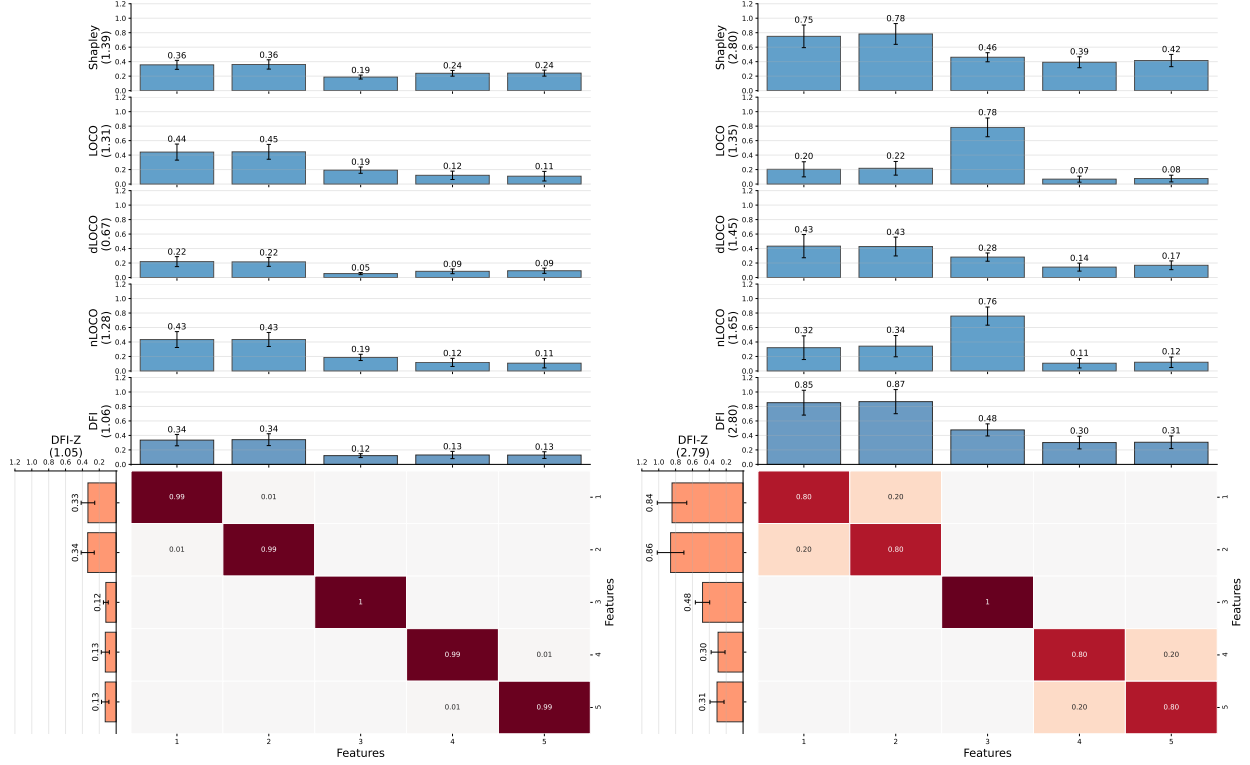


Figure 3: Simulation results for model (M3). (a) weak correlation ($\rho = 0.2$); (b) strong correlation ($\rho = 0.8$). The interpretation of the bars, error bars, and heat-map is identical to Figure 1. The number in parentheses is the total estimated importance, which should be close to the signal variance $\mathbb{V}[\mathbb{E}[Y \mid X]] = (26 + 27\rho^2)/16$.

ρ	0				0.4				0.8			
n	100	250	500	750	100	250	500	750	100	250	500	750
DFI	0.944	0.969	0.988	0.991	0.941	0.981	0.988	1.000	0.962	0.975	0.994	1.000
DFI-Z	0.959	0.969	0.984	0.975	0.972	0.984	0.984	0.997	0.984	0.975	0.994	0.991
LOCO	0.984	0.956	0.941	0.959	0.988	0.966	0.953	0.944	0.941	0.972	0.966	0.956
dLOCO	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	1.000
nLOCO	0.981	0.956	0.941	0.956	0.991	0.966	0.953	0.941	0.944	0.975	0.966	0.953

Table 1: Coverage of different importance measures for the null features ($X_3, \dots, X_{10}$) under model (M1) with a nominal level of $\alpha = 0.1$.

trated in Figure 5, DFI is significantly faster than competing methods that also account for feature correlations, such as dLOCO and sampling-based approximations of the Shapley value. This efficiency stems from DFI’s design, which avoids both the need to refit multiple submodels (a drawback of LOCO-based approaches) and the direct estimation of complex conditional distributions (a challenge for standard CPI). This combination of statistical robustness, inferential validity, and computational efficiency makes DFI a practical tool for researchers analyzing complex, high-dimensional datasets.

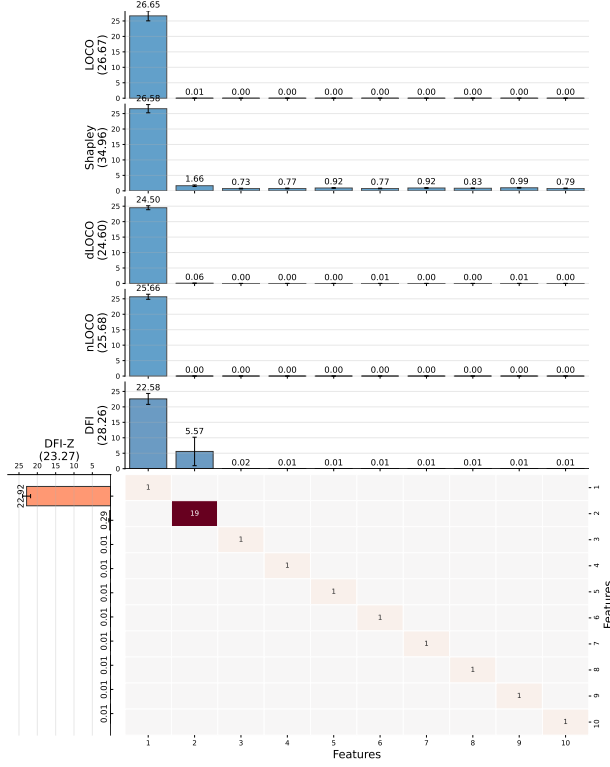


Figure 4: Simulation results for model (M4). The interpretation of the bars, error bars, and heat-map is identical to Figure 1. The number in parentheses is the total estimated importance, which should be close to the signal variance $\mathbb{V}[\mathbb{E}[Y | X]] = 25$.

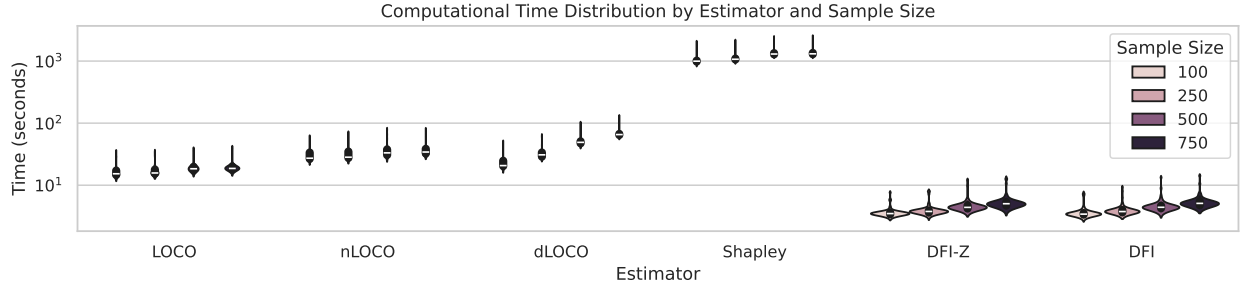


Figure 5: Computational time of different feature importance measures in the logarithmic scale.

6 Real data

In this section, we demonstrate the application of DFI to a real-world dataset from HIV-1 research, aiming to identify viral envelope features that predict resistance to the broadly neutralizing antibody VRC01. This is a critical step toward designing effective HIV-1 vaccines and antibody-based therapies. The dataset, introduced by Montefiori (2009) and previously analyzed by Williamson et al. (2023), contains highly correlated genomic features, presenting a suitable and challenging test case for our method. While Williamson et al. (2023) address feature correlation by assessing marginal importance, DFI offers a principled framework for

directly attributing the predictive signal.

Following the preprocessing steps of [Williamson et al. \(2023\)](#), with additional standardization of features and the outcome, the data comprises 611 samples and 832 features organized into 14 groups. The outcome is a binary indicator of whether the 50% inhibitory concentration (IC_{50}) was right-censored, where high IC_{50} values signify viral resistance to neutralization ([Montefiori, 2009](#)). The additive structure of the DFI measure (4) facilitates the assessment of group importance for a feature set $X_{\mathcal{G}}$ with $\mathcal{G} \subseteq [d]$, defined as

$$\phi_{X_{\mathcal{G}}}(\mathbb{P}) := \sum_{l \in \mathcal{G}} \sum_{j=1}^d \mathbb{E} \left[\mathbb{V} \left(\mathbb{E}[Y | Z] \mid Z_{-j} \right) \left(\frac{\partial X_l}{\partial Z_j} \right)^2 \right].$$

The statistical estimation and inference framework from Section 4 extends directly to this group setting. To ensure stable estimates, we compute the regression function and covariance matrix on the full dataset, treating them as fixed for an algorithmic feature importance analysis. To address the issue of inference near the null noted in Remark 4.1, we inflate the estimated variance by $n^{-1/2}/z_{1-\alpha}$ when computing p-values at a significance level of α .

For $\alpha = 0.05$ (with Bonferroni correction for one-sided tests), DFI identifies 218 of 832 features as significant in the disentangled space and 726 of 832 in the original feature space. The group-level importance results are summarized in Figure 6. The analysis reveals that a few feature groups are significantly associated with VRC01 neutralization resistance. The three most influential groups are Group 5 (Sites with residues covarying with the VRC01 binding footprint), Group 6 (Sites linked to VRC01-specific potential N-linked glycosylation effects), and Group 1 (VRC01 binding footprint). These groups exhibit the highest importance and Z-scores, indicating strong statistical significance that far exceeds the Bonferroni-corrected threshold. This finding suggests that the interplay between the VRC01 binding footprint and glycosylation patterns is a primary determinant of neutralization sensitivity.

Groups 2 (CD4 binding sites), 3 (Sites with sufficient exposed surface area), 8 (Sites for indicating N-linked glycosylation), and 4 (Sites identified as important for glycosylation) also demonstrate significant, though comparatively lesser, importance. The remaining groups, which include those related to viral geometry and geographic confounders, show minor importance, with several failing to reach statistical significance. These findings are largely consistent with those of [Williamson et al. \(2023\)](#), which used different importance measures, though our analysis provides a slightly different ranking of the most important groups.

7 Discussion

In this paper, we introduced disentangled feature importance (DFI), a framework that addresses the fundamental challenge of measuring feature importance when predictors are correlated. Traditional methods like LOCO, CPI, and Shapley values suffer from correlation distortion, where a feature may appear unimportant simply because its predictive signal is captured by correlated variables. Our approach overcomes this limitation by first transforming features into a disentangled space where they are independent, measuring importance in this space, and then attributing importance back to the original features. For Bures-Wasserstein transport map, we develop statistical theories including asymptotic normality

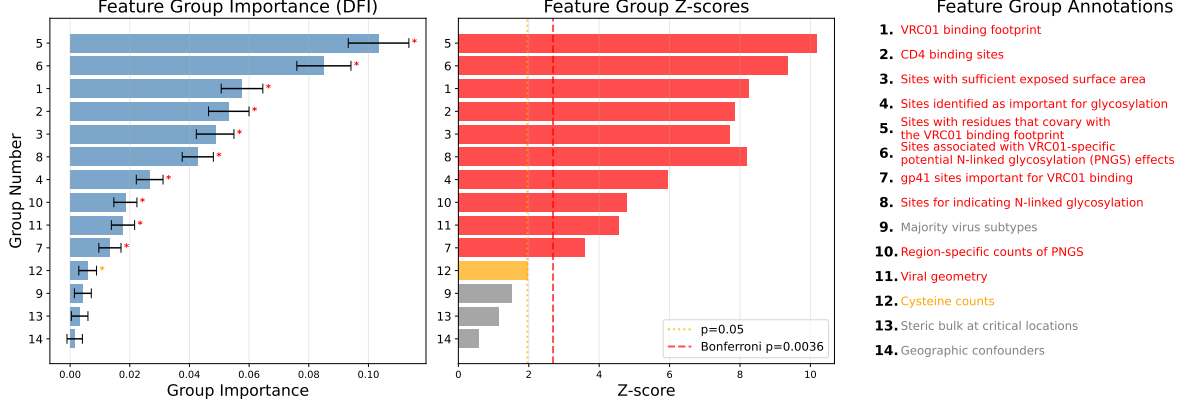


Figure 6: Feature importance of antibody against HIV-1 infection. (a) The bars show the point estimate, while the error bars indicate the values within one estimated standard deviation. (b) The corresponding Z-score for one-sided tests of $\mathcal{H}_{0l} : \phi_{X_l}(\mathbb{P}) \leq 0$. (c) Feature group description. Stars * and * denote importance deemed statistically significantly different from zero at the 0.05 and 0.0036 (0.05/14) levels, respectively.

and efficient influence functions. Notably, our method recovers the classical R^2 decomposition for linear models, positioning DFI as a natural nonparametric extension of this fundamental concept.

The practical advantages of DFI extend beyond its theoretical properties. Unlike Shapley values, which require fitting exponentially many submodels, or CPI, which necessitates learning conditional distributions, DFI leverages the independence structure of disentangled features to avoid both computational bottlenecks. Our simulation studies demonstrate that DFI consistently captures the total predictive variability and distributes it appropriately among features, even under strong correlations where traditional methods fail. The method performs well in nonlinear and misspecified settings and correctly identifies relevant features that other approaches miss due to correlation effects.

Despite these strengths, several limitations warrant consideration. Our theoretical framework assumes absolutely continuous covariate distributions, which may not hold for discrete or mixed-type data commonly encountered in practice. While our empirical results suggest reasonable performance in various settings, formal extensions to such cases require further development. Additionally, like other methods based on quadratic functionals, statistical inference near the null hypothesis presents challenges. When true feature importance is zero or very small, standard confidence intervals may not maintain proper coverage, though this issue can be mitigated by expanding intervals at the cost of reduced efficiency.

Several promising research directions emerge from this work. First, developing computationally efficient methods for high-dimensional transport map estimation represents a key practical challenge. Modern generative models, such as variational autoencoders or diffusion models, could provide scalable alternatives for learning disentangled representations. Second, extending DFI beyond squared-error loss to classification and other prediction tasks would significantly broaden its applicability. Third, developing local versions of DFI for individual-level feature importance could bridge the gap between global and local interpretability methods. Finally, incorporating domain knowledge such as feature hierarchies or

network structures into the disentanglement process could yield more powerful and contextually relevant importance measures.

Supplement

The appendix includes proof of all the theoretical results.

Appendix

The appendix includes the proof for all the theorems, computational details, and extra experimental results. The structure of the appendix is listed below:

Outline. The structure of the appendix is listed below:

Appendix		Content
Appendix A	Appendix A.1	Proof of Lemma 2.1.
	Appendix A.2	Proof of Lemma 2.2.
Appendix B	Appendix B.1	Proof of Proposition 3.1.
	Appendix B.2	Proof of Theorem 4.1.
	Appendix B.3	Proof of Proposition 4.3.
	Appendix B.4	Proof of Theorem 4.4.
	Appendix B.5	Proof of Proposition 4.5.
	Appendix B.6	Extension to Knothe-Rosenblatt transport map.
Appendix C	Appendix C.1	Basic properties of estimators: Lemma C.1, Proposition C.2.
	Appendix C.2	Basic properties of transport maps: Lemmas C.3 and C.4.
	Appendix C.3	Lemmas related to LOCO estimation: Lemmas C.5 and C.6.
	Appendix C.4	Lemmas related to the disentangled feature importance estimator: Lemma C.7, Proposition C.8, and Lemma C.9.

A Proof in Section 2

A.1 Proof of Lemma 2.1

Proof of Lemma 2.1. Note that $X^{(j)} \perp\!\!\!\perp Y \mid X_{-j}$, because $X_j^{(j)}$ is constructed without using Y (i.e., $X_j^{(j)} \perp\!\!\!\perp Y \mid X_{-j}$). From the construction of $X^{(j)}$, we have that $X^{(j)} \stackrel{\text{i.i.d.}}{\sim} X$. Under the null that $X_j \perp\!\!\!\perp Y \mid X_{-j}$, we further have $(X^{(j)}, Y) \stackrel{\text{i.i.d.}}{\sim} (X, Y)$ and $\mu(X) = \mu_{-j}(X_{-j}) = \mu(X^{(j)})$. Consequently, $\psi_{X_j}^{\text{CPI}} = \psi_{X_j}^{\text{LOCO}} = 0$. This finishes the proof. $\square$

A.2 Proof of Lemma 2.2

Proof of Lemma 2.2. For any random variable W , its variance can be expressed as $\mathbb{V}[W] = \mathbb{E}[(W - \mathbb{E}[W])^2] = \mathbb{E}[(W - W')^2]/2$ where W' is an independent copy of W . Conditioned on X_{-j} , it follows that

$$\begin{aligned}\mathbb{V}[\mu(X) \mid X_{-j}] &= \mathbb{E}[(\mu(X) - \mathbb{E}[\mu(X) \mid X_{-j}])^2 \mid X_{-j}] = \mathbb{E}[(\mu(X) - \mu_{-j}(X_{-j}))^2 \mid X_{-j}] \\ \mathbb{V}[\mu(X) \mid X_{-j}] &= \frac{1}{2}\mathbb{E}[(\mu(X) - \mathbb{E}[\mu(X^{(j)}) \mid X_{-j}])^2 \mid X_{-j}] = \frac{1}{2}\mathbb{E}[(\mu(X) - \mu(X^{(j)}))^2 \mid X_{-j}].\end{aligned}$$

On the other hand, for ℓ_2 -loss, we have

$$\begin{aligned}\psi_{X_j}^{\text{LOCO}} &= \mathbb{E}[(Y - \mu_{-j}(X_{-j}))^2] - \mathbb{E}[(Y - \mu(X))^2] \\ &= \mathbb{E}[(\mu(X) - \mu_{-j}(X_{-j}))^2] - 2\mathbb{E}[(\mu(X) - Y)(\mu(X) - \mu_{-j}(X_{-j}))] \\ &= \mathbb{E}[(\mu(X) - \mu_{-j}(X_{-j}))^2],\end{aligned}$$

because $\mathbb{E}[Y(\mu(X) - \mu_{-j}(X_{-j}))] = \mathbb{E}[\mathbb{E}[Y(\mu(X) - \mu_{-j}(X_{-j})) \mid X]] = \mathbb{E}[\mu(X)(\mu(X) - \mu_{-j}(X_{-j}))]$ by iterative expectation. Analogously, one has

$$\begin{aligned}2\psi_{X_j}^{\text{CPI}} &= \mathbb{E}[(Y - \mu(X^{(j)}))^2] - \mathbb{E}[(Y - \mu(X))^2] \\ &= \mathbb{E}[(\mu(X) - \mu(X^{(j)}))^2] - 2\mathbb{E}[(\mu(X) - Y)(\mu(X) - \mu(X^{(j)}))] \\ &= \mathbb{E}[(\mu(X) - \mu(X^{(j)}))^2].\end{aligned}$$

Combining the above results finishes the proof. $\square$

B Proof of main results

B.1 Proof of Proposition 3.1

Proof of Proposition 3.1. From Lemma 2.2, we have $\phi_{Z_j} = \mathbb{E}[\mathbb{V}[\mu(W) \mid Z_{-j}]]$. Because $\mathbb{V}[\mu(W) \mid Z_{-j}] \geq 0$, we have that $\phi_{Z_j} = 0$ if and only if $\mathbb{V}[\mu(W) \mid Z_{-j}] = 0$ almost surely. This is equivalent to $\mu(W) = \mathbb{E}[\mu(W) \mid Z_{-j}]$ almost surely. Also note that $\mathbb{E}[Y \mid Z] = \mathbb{E}[\mathbb{E}[Y \mid W] \mid Z] = \mathbb{E}[\mu(W) \mid Z] = \mu(W)$ because of $W \sim p(W \mid Z)$ and tower property of conditional expectation. Then, the condition is equivalent to $\mathbb{E}[Y \mid Z] = \mathbb{E}[\mu(W) \mid Z_{-j}]$ almost surely. By the tower property, we also have $\mathbb{E}[\mu(W) \mid Z_{-j}] = \mathbb{E}[\mathbb{E}[Y \mid Z] \mid Z_{-j}] = \mathbb{E}[Y \mid Z_{-j}]$. These imply that $\phi_{Z_j} = 0$ if and only if $\mathbb{E}[Y \mid Z] = \mathbb{E}[Y \mid Z_{-j}]$. Because Z_j and Z_{-j} are independent, $\mathbb{E}[Y \mid Z] = \mathbb{E}[Y \mid Z_{-j}]$ is equivalent to the event that $\mathbb{E}[Y \mid Z]$ does not depend on Z_j almost surely. $\square$

B.2 Proof of Theorem 4.1

Proof of Theorem 4.1. We split the proof into three steps.

Step 1: Error decomposition for the LOCO functional. From Williamson et al. (2021, Lemma 1), the efficient influence function for $\phi_{Z_j} = \mathbb{E}[(\eta(Z) - \eta_{-j}(Z_{-j}))^2]$ is $\varphi(O; \mathbb{P}) = 2(Y - \eta(Z))(\eta(Z) - \eta_{-j}(Z_{-j})) + (\eta(Z) - \eta_{-j}(Z_{-j}))^2 - \phi_{Z_j}$. Furthermore, for any given estimator $\hat{\mathbb{P}}$ (with associated estimators $\hat{\eta}$ and $\hat{\eta}_{-j}$) of the nonparametric model $\mathbb{P}$, the LOCO estimator

$$\hat{\phi}_{Z_j}^{\text{LOCO}}(\mathbb{P}; \hat{\eta}, \hat{\eta}_{-j}) := \mathbb{P}_n \left[(Y - \hat{\eta}_{-j}(Z_{-j}))^2 - (Y - \hat{\eta}(Z))^2 \right], \quad (12)$$

admits the following error decomposition (see the proof of Williamson et al. (2021, Theorem 1)):

$$\hat{\phi}_{Z_j}^{\text{LOCO}}(\mathbb{P}; \hat{\eta}, \hat{\eta}_{-j}) - \phi_{Z_j} = (\mathbb{P}_n - \mathbb{P})\{\varphi(O; \mathbb{P})\} + R(\hat{\mathbb{P}}, \mathbb{P}) + H(\hat{\mathbb{P}}, \mathbb{P}),$$

where the first term scaled by $\sqrt{n}$ is asymptotically normal by the Central Limit Theorem (since φ is mean-zero and square-integrable), and the remaining terms read as

$$R(\hat{\mathbb{P}}, \mathbb{P}) = \|\hat{\eta}_{-j} - \eta_{-j}\|_{L_2}^2 - \|\hat{\eta} - \eta\|_{L_2}^2 \quad (13)$$

$$H(\hat{\mathbb{P}}, \mathbb{P}) = (\mathbb{P}_n - \mathbb{P})\{\varphi(O; \hat{\mathbb{P}}) - \varphi(O; \mathbb{P})\}. \quad (14)$$

Define the estimation errors:

$$\delta(Z) := \hat{\eta}(Z) - \eta(Z), \quad \delta_{-j}(Z_{-j}) := \hat{\eta}_{-j}(Z_{-j}) - \eta_{-j}(Z_{-j}).$$

By the Cauchy–Schwarz inequality and the fact that $\mathbb{P}_n - \mathbb{P}$ is an average of n i.i.d. mean-zero variables,

$$|H(\hat{\mathbb{P}}, \mathbb{P})| \leq \|\varphi_{\hat{\mathbb{P}}} - \varphi_{\mathbb{P}}\|_{L_2} \sqrt{\frac{\text{Var}[\varphi_{\hat{\mathbb{P}}} - \varphi_{\mathbb{P}}]}{n}} = \mathcal{O}_{\mathbb{P}} \left(\frac{\|\varphi_{\hat{\mathbb{P}}} - \varphi_{\mathbb{P}}\|_{L_2}}{\sqrt{n}} \right). \quad (15)$$

Further by Lemma C.5, we have

$$|H(\widehat{\mathbb{P}}, \mathbb{P})| = \mathcal{O}_{\mathbb{P}} \left(\frac{1}{\sqrt{n}} (\|\delta\|_{L_2} + \|\delta_{-j}\|_{L_2}) + \frac{1}{\sqrt{n}} (\|\delta\|_{L_2} + \|\delta_{-j}\|_{L_2})^2 \right).$$

Consequently, we have

$$\widehat{\phi}_{Z_j}^{\text{LOCO}}(\mathbb{P}; \widehat{\eta}, \widehat{\eta}_{-j}) - \phi_{Z_j} = (\mathbb{P}_n - \mathbb{P})\{\varphi(O; \mathbb{P})\} + \mathcal{O}_{\mathbb{P}}(\|\delta\|_{L_2}^2 + \|\delta_{-j}\|_{L_2}^2) + \frac{1}{\sqrt{n}} \mathcal{O}_{\mathbb{P}}(\|\delta\|_{L_2} + \|\delta_{-j}\|_{L_2}). \quad (16)$$

Step 2: Error decomposition for the CPI functional. Next, we relate the CPI functional to the LOCO functional. Recall that our estimator is defined as

$$\widehat{\phi}_{Z_j} := \frac{1}{2} \mathbb{P}_n [(Y - \widehat{\eta}(Z^{(j)}))^2 - (Y - \widehat{\eta}(Z))^2].$$

Note that for $\widehat{\eta}_{-j} = \eta_{-j}$, $\delta_{-j} = 0$. We decompose the estimation error:

$$\begin{aligned} \widehat{\phi}_{Z_j} - \phi_{Z_j} &= [\widehat{\phi}_{Z_j} - \widehat{\phi}_{Z_j}^{\text{LOCO}}(\mathbb{P}; \widehat{\eta}, \eta_{-j})] + [\widehat{\phi}_{Z_j}^{\text{LOCO}}(\mathbb{P}; \widehat{\eta}, \eta_{-j}) - \phi_{Z_j}] \\ &= \frac{1}{2} \mathbb{P}_n [(Y - \widehat{\eta}(Z^{(j)}))^2 - (Y - \eta_{-j}(Z_{-j}))^2] \\ &\quad + \frac{1}{2} \mathbb{P}_n [(Y - \widehat{\eta}(Z))^2 - (Y - \eta_{-j}(Z_{-j}))^2] \\ &\quad + (\mathbb{P}_n - \mathbb{P})\{\varphi(O; \mathbb{P})\} + \mathcal{O}_{\mathbb{P}}(n^{-1/2} \|\delta\|_{L_2} + \|\delta\|_{L_2}^2) \\ &= \frac{1}{2} R_e + \frac{1}{2} R + (\mathbb{P}_n - \mathbb{P})\{\varphi(O; \mathbb{P})\} + \mathcal{O}_{\mathbb{P}}(n^{-1/2} \|\delta\|_{L_2} + \|\delta\|_{L_2}^2), \end{aligned} \quad (17)$$

where the empirical process terms R_e and remaining bias terms R_b are defined as:

$$\begin{aligned} R_e &= (\mathbb{P}_n - \mathbb{P})[\widehat{\phi}_{Z_j} - \widehat{\phi}_{Z_j}^{\text{LOCO}}(\mathbb{P}; \widehat{\eta}, \eta_{-j})] \\ R_b &= \mathbb{P}[\widehat{\phi}_{Z_j} - \widehat{\phi}_{Z_j}^{\text{LOCO}}(\mathbb{P}; \widehat{\eta}, \eta_{-j})]. \end{aligned}$$

Before analyzing each of these terms, we introduce

$$\zeta := \eta(Z) - \eta_{-j}(Z_{-j}), \quad \zeta_j := \eta(Z^{(j)}) - \eta_{-j}(Z_{-j}), \quad \delta_j := \widehat{\eta}(Z^{(j)}) - \eta(Z^{(j)}),$$

and write $\varepsilon := Y - \eta$, $U := \varepsilon + \zeta$, $U_j := \varepsilon + \zeta_j$. Because

$$\begin{aligned} (Y - \widehat{\eta}(Z))^2 - (Y - \eta_{-j}(Z_{-j}))^2 &= (U - \delta)^2 - U^2 = \delta^2 - 2U\delta, \\ (Y - \widehat{\eta}(Z^{(j)}))^2 - (Y - \eta_{-j}(Z_{-j}))^2 &= (U_j - \delta_j)^2 - U^2 = \delta_j^2 - 2U_j\delta_j, \end{aligned}$$

we have

$$\begin{aligned} R_e &= \frac{1}{2} (\mathbb{P}_n - \mathbb{P})[(Y - \widehat{\eta}(Z^{(j)}))^2 - (Y - \eta_{-j}(Z_{-j}))^2] + \frac{1}{2} (\mathbb{P}_n - \mathbb{P})[(Y - \widehat{\eta}(Z))^2 - (Y - \eta_{-j}(Z_{-j}))^2] \\ &= \frac{1}{2} (\mathbb{P}_n - \mathbb{P})[\delta^2 - 2U\delta + \delta_j^2 - 2U_j\delta_j]. \end{aligned}$$

Because $\|\delta\|_{L_\infty} \leq 2C$ from Assumption 4.2 (3) and $\|U\|_{L_2} \leq C$ from Lemma C.1, we have

$$\begin{aligned}\|\delta^2\|_{L_2} &\leq 2C \|\delta\|_{L_2} \\ \|U\delta\|_{L_2} &\leq \|U\|_{L_2} \|\delta\|_{L_2} \leq 3C \|\delta\|_{L_2}.\end{aligned}$$

Similarly, we have $2C \|\delta_j^2\|_{L_2} \leq 2C \|\delta_j\|_{L_2}$ and $\|U\delta_j\|_{L_2} \leq 3C \|\delta_j\|_{L_2}$. Additionally,

$$\|\delta_j\|_{L_2}^2 = \mathbb{E}[\mathbb{E}[(\hat{\eta}(Z^{(j)}) - \eta(Z^{(j)}))^2 \mid Z_{-j}]] = \mathbb{E}[(\hat{\eta}(Z) - \eta(Z))^2] = \|\delta\|_{L_2}^2,$$

Hence, it follows that

$$R_e = \mathcal{O}_{\mathbb{P}}(n^{-1/2} \|\delta^2 - 2U\delta + \delta_j^2 - 2U_j\delta_j\|_{L_2}) = \mathcal{O}_{\mathbb{P}}(n^{-1/2} \|\delta\|_{L_2}). \quad (18)$$

For the bias term, note that

$$R_b = \frac{1}{2} \mathbb{E}[(Y - \hat{\eta}(Z^{(j)}))^2 - (Y - \eta_{-j}(Z_{-j}))^2] + \frac{1}{2} \mathbb{E}[(Y - \hat{\eta}(Z))^2 - (Y - \eta_{-j}(Z_{-j}))^2] =: A_1 + A_2.$$

Using $\hat{\eta}(Z^{(j)}) - \eta_{-j}(Z_{-j}) = \zeta_j + \delta_j$ and $\mathbb{E}[Y - \eta_{-j}(Z_{-j}) \mid Z^{(j)}, Z_{-j}] = 0$, the two terms of R can be written as:

$$\begin{aligned}A_1 &= \frac{1}{2} \mathbb{E}[(Y - \eta(Z^{(j)}) + \eta(Z^{(j)}) - \hat{\eta}(Z^{(j)}))^2 - (Y - \eta_{-j}(Z_{-j}))^2] \\ &= \frac{1}{2} \mathbb{E}[(Y - \eta(Z^{(j)}))^2 + (\eta(Z^{(j)}) - \hat{\eta}(Z^{(j)}))^2 - (Y - \eta_{-j}(Z_{-j}))^2] \\ &= -\frac{1}{2} \phi_{Z_j} + \frac{1}{2} \mathbb{E}[(\eta(Z^{(j)}) - \hat{\eta}(Z^{(j)}))^2] \\ &= -\frac{1}{2} \phi_{Z_j} + \frac{1}{2} \|\delta\|_{L_2}^2 \\ A_2 &= \frac{1}{2} \mathbb{E}[(Y - \eta(Z) + \eta(Z) - \hat{\eta}(Z))^2 - (Y - \eta_{-j}(Z_{-j}))^2] \\ &= \frac{1}{2} \mathbb{E}[(Y - \eta(Z))^2 + (\eta(Z) - \hat{\eta}(Z))^2 - (Y - \eta_{-j}(Z_{-j}))^2] \\ &= \frac{1}{2} \phi_{Z_j} + \frac{1}{2} \|\delta\|_{L_2}^2,\end{aligned}$$

Hence, one has

$$R = \|\delta\|_{L_2}^2 \quad (19)$$

Combining (17), (18) and (19) yields that

$$\hat{\phi}_{Z_j} - \phi_{Z_j} = (\mathbb{P}_n - \mathbb{P})\{\varphi(O; \mathbb{P})\} + \mathcal{O}_{\mathbb{P}}(n^{-1/2}(\|\delta\|_{L_2} + \|\delta\|_{L_2}^2)). \quad (20)$$

Step 3: bounding $\|\hat{\eta} - \eta\|_{L_2(\mathbb{P}_Z)}$. Recall that $\eta = \mu \circ T^{-1}$, $\hat{\eta} = \hat{\mu} \circ \hat{T}^{-1}$, and decompose for every $z \in \mathcal{Z}$

$$\hat{\eta}(z) - \eta(z) = \hat{\mu}(\hat{T}^{-1}(z)) - \hat{\mu}(T^{-1}(z)) + \hat{\mu}(T^{-1}(z)) - \mu(T^{-1}(z)) =: B_1 + B_2. \quad (21)$$

Term B_1 is the transport error. Set $X := T^{-1}(Z) \sim \mathbb{P}_X$ and $X' := \widehat{T}^{-1}(Z) \sim \mathbb{P}_{X'}$, and define

$$\varepsilon_\mu := \|\widehat{\mu} - \mu\|_{L_2(\mathbb{P}_X)}, \quad \varepsilon_{T^{-1}} := \|\widehat{T}^{-1} - T^{-1}\|_{L_2(\mathbb{P}_Z)}, \quad \varepsilon_T := \|\widehat{T} - T\|_{L_2(\mathbb{P}_X)}.$$

Introduce the truncation event $E := \{\|Z\| \leq R_{\varepsilon_\mu}/\lambda\}$ with $R_{\varepsilon_\mu} = \lambda\sqrt{d + 2\log(1/\varepsilon_\mu)}$, for which $\mathbb{P}(E) \geq 1 - \varepsilon_\mu^2$ by a Gaussian tail bound. On E both X and X' lie in the ball $B := \{x \in \mathbb{R}^d : \|x\| \leq R_{\varepsilon_\mu}\}$. By Proposition C.2, with probability at least $1 - \varepsilon_\mu^2$,

$$L_{\mu, \varepsilon_\mu} := \text{Lip}(\widehat{\mu}; B) \leq L_\mu + C L_\mu^{1/2} \lambda^{-1/2} [d + 2\log(1/\varepsilon_\mu)]^{-1/4} \varepsilon_\mu^{1/2}, \quad (22)$$

where $C = 4\sqrt{2}$ is the universal constant from the proposition. Event (22) is contained in E , so it also occurs with probability at least $1 - \varepsilon_\mu^2$. At this event, we have

$$|B_1| = |\widehat{\mu}(X') - \widehat{\mu}(X)| \leq L_{\mu, \varepsilon_\mu} \|X' - X\|, \quad \mathbb{E}_Z[B_1^2] \leq L_{\mu, \varepsilon_\mu}^2 \varepsilon_{T^{-1}}^2 \leq \lambda^2 L_{\mu, \varepsilon_\mu}^2 \varepsilon_T^2,$$

where the last inequality is from Assumption 4.1.

Term B_2 is the regression error. By definition,

$$\mathbb{E}_Z[B_2^2] = \|\widehat{\mu} - \mu\|_{L_2(\mathbb{P}_X)}^2 =: \varepsilon_\mu^2.$$

Use the inequality $(a + b)^2 \leq 2a^2 + 2b^2$ with $a = B_1$ and $b = B_2$:

$$\|\widehat{\eta} - \eta\|_{L_2(\mathbb{P}_Z)}^2 = \mathbb{E}_Z[(B_1 + B_2)^2] \leq 2\mathbb{E}_Z[B_1^2] + 2\mathbb{E}_Z[B_2^2] \leq 2\lambda^2 L_{\mu, \varepsilon_\mu}^2 \varepsilon_T^2 + 2\varepsilon_\mu^2.$$

Taking square roots and keeping only the leading factor $\sqrt{2}$ yields

$$\|\widehat{\eta} - \eta\|_{L_2(\mathbb{P}_Z)} \leq \sqrt{2}\lambda L_{\mu, \varepsilon_\mu} \varepsilon_T + \sqrt{2}\varepsilon_\mu, \quad (23)$$

with probability at least $1 - \varepsilon_\mu^2$.

Finally, combining (20) with (23), we further have

$$\begin{aligned} \widehat{\phi}_{Z_j} - \phi_{Z_j} &= (\mathbb{P}_n - \mathbb{P})\varphi(O; \mathbb{P}) + \mathcal{O}(n^{-1/2}(\|\widehat{\mu} - \mu\|_{L_2(\mathbb{P}_X)}^2 + \|\widehat{T} - T\|_{L_2(\mathbb{P}_X)}^2)^{1/2} \\ &\quad + \|\widehat{\mu} - \mu\|_{L_2(\mathbb{P}_X)}^2 + \|\widehat{T} - T\|_{L_2(\mathbb{P}_X)}^2), \end{aligned}$$

with probability at least $1 - \|\widehat{\mu} - \mu\|_{L_2(\mathbb{P}_X)}^2$. This completes the proof. $\square$

B.3 Proof of Proposition 4.3

Proof of Proposition 4.3. Throughout the proof, we drop the subscript j for simplicity and put

$$A(O) := \Delta(Z)^2, \quad B(O) := H(X), \quad \psi_A(\mathbb{P}) := \mathbb{P}[A(O)], \quad \psi_B(\mathbb{P}) := \theta(\mathbb{P}) = \mathbb{P}[B(O)].$$

With this notation, the target parameter is the expectation of a product, $\phi(\mathbb{P}) = \mathbb{P}[A(O)B(O)]$.

Step 1: EIF of A . From Williamson et al. (2021, Lemma 1), the efficient influence function (EIF) of $\psi_A(\mathbb{P}) = \mathbb{E}[\Delta(Z)^2]$ is

$$\varphi_A(O; \mathbb{P}) = 2[Y - \eta(Z)] \Delta(Z) + \Delta(Z)^2 - \psi_A(\mathbb{P}).$$

Writing $A_c(O) = A(O) - \psi_A(\mathbb{P})$ we therefore have the decomposition

$$\varphi_A(O; \mathbb{P}) = A_c(O) + \alpha_A(O), \quad \alpha_A(O) := 2[Y - \eta(Z)] \Delta(Z).$$

Step 2: EIF of B . By assumption the (non-degenerate) parameter $\theta(\mathbb{P}) = \mathbb{E}[H(X)]$ is pathwise differentiable with EIF $\varphi_H(O; \mathbb{P})$. Writing $B_c(O) = B(O) - \theta(\mathbb{P})$ we set

$$\alpha_B(O) := \varphi_H(O; \mathbb{P}) - B_c(O) = \varphi_H(O; \mathbb{P}) - \{H(X) - \theta(\mathbb{P})\}.$$

Step 3: EIF of the product functional. Apply Lemma C.7 with A, B, α_A, α_B defined above, it follows that

$$\varphi_{\Delta^2 H}(O; \mathbb{P}) = A(O)B(O) - \phi(\mathbb{P}) + B(O)\alpha_A(O) + A(O)\alpha_B(O).$$

Substituting the explicit expressions gives

$$\begin{aligned} \varphi_{\Delta^2 H}(O; \mathbb{P}) &= H(X) \Delta(Z)^2 - \phi(\mathbb{P}) + H(X) \{2[Y - \eta(Z)] \Delta(Z)\} \\ &\quad + \Delta(Z)^2 \{\varphi_H(O; \mathbb{P}) - H(X) + \theta(\mathbb{P})\}. \end{aligned}$$

Re-arranging the terms completes the proof. □

B.4 Proof of Theorem 4.4

Proof of Theorem 4.4. Before we present the proof, we will introduce some notation for clarity. Fix a coordinate $l \in [d]$. For each $j \in [d]$, define

$$H_{j,l}(x) := (\partial_{x_l} T_j(x))^2, \quad \Delta_j(z) := \eta(z) - \eta_{-j}(z_{-j}), \quad \phi_{j,l}(\mathbb{P}) := \mathbb{E}[H_{j,l}(X) \Delta_j(Z)^2].$$

Under the Bures–Wasserstein (BW) map $T(x) = Lx$ we have $H_{j,l}(x) = L_{jl}^2$, a constant that depends only on $L := \Sigma^{-\frac{1}{2}}$. The target parameter is the sum $\phi_{X_l} = \sum_{j=1}^d \phi_{j,l}(\mathbb{P})$.

Step 1: Influence function expansion. From Propositions 4.3 and C.8, the efficient influence function of ϕ_{X_l} is given by

$$\varphi_{X_l}(O; \mathbb{P}) = \sum_{j=1}^d \{L_{jl}^2 [2(Y - \eta(Z)) \Delta_j(Z) + \Delta_j(Z)^2] - L_{jl}(Z_j Z_l - \delta_{jl}) \Delta(Z)^2\} - \phi_{X_l}.$$

Let $\{(X_i, Z_i, Y_i)\}_{i=1}^n$ be an i.i.d. sample independent of the nuisance estimates $(\hat{\eta}, \hat{\eta}_{-j}, \hat{T})$, $\hat{H}_{j,l}(x) = (\partial_{x_l} \hat{T}_j(x))^2$ and $\hat{\Delta}_j(z) = \hat{\eta}(z) - \hat{\eta}_{-j}(z_{-j})$.

Recall that the one-step estimator is defined as

$$\begin{aligned} \hat{\phi}_{X_l}^{\text{LOCO}}(\mathbb{P}) &= \phi_{X_l}(\hat{\mathbb{P}}) + \mathbb{P}_n\{\varphi(O; \hat{\mathbb{P}})\} \\ &= \frac{1}{2} \mathbb{P}_n \left\{ \sum_{j=1}^d \hat{L}_{jl}^2 [(Y - \hat{\eta}(Z^{(j)}))^2 - (Y - \hat{\eta}(Z))^2] \right\} + \mathcal{O}_{\mathbb{P}}(n^{-1/2} \|\delta\|_{L_2} + \|\delta\|_{L_2}^2) \\ &= \hat{\phi}_{X_l}(\mathbb{P}) + \mathcal{O}_{\mathbb{P}}(n^{-1/2} \|\delta\|_{L_2} + \|\delta\|_{L_2}^2), \end{aligned} \quad (24)$$

where the second equality follows similarly from Step 2 in the proof of Theorem 4.1 and the remainder term corresponds to the difference of the LOCO estimator $\hat{\phi}_{X_l}^{\text{LOCO}}(\mathbb{P})$ and the CPI estimator $\hat{\phi}_{X_l}(\mathbb{P})$, weighted by $\hat{L}_{jl}^2$'s. Note that $\hat{\eta}_{-j}(Z_{-j}) = \mathbb{E}[\hat{\eta}(Z) \mid Z_{-j}]$ and $\hat{T}(X) = \hat{L}X$ and $\partial_{x_l} \hat{T}_j(X) = \hat{L}_{jl}$; hence, we only need to estimate two nuisance functions, essentially.

Because the nuisance functions are estimated from independent samples other than $\mathbb{P}_n$, the classical von Mises expansion (see, e.g., Du et al. (2025b), Equation (2.2)) reads

$$\hat{\phi}_{X_l}^{\text{LOCO}} - \phi_{X_l} = (\mathbb{P}_n - \mathbb{P})\{\varphi_{X_l}(O; \mathbb{P})\} + \underbrace{(\mathbb{P}_n - \mathbb{P})[\varphi_{X_l}(O; \hat{\mathbb{P}}) - \varphi_{X_l}(O; \mathbb{P})]}_{E_n} + \underbrace{\mathbb{P}[\varphi_{X_l}(O; \hat{\mathbb{P}}) - \varphi_{X_l}(O; \mathbb{P})]}_{R_n}. \quad (25)$$

It therefore suffices to show $|E_n| = \mathcal{O}_{\mathbb{P}}(n^{-1/2} \varepsilon_n)$, and $|R_n| = \mathcal{O}_{\mathbb{P}}(\varepsilon_n^2)$.

Step 2: Bounding the empirical process term E_n . Conditioning on the auxiliary split, the summands $\varphi_{X_l}(O; \hat{\mathbb{P}}) - \varphi_{X_l}(O; \mathbb{P})$ are i.i.d. with mean 0, so that

$$|E_n| = |(\mathbb{P}_n - \mathbb{P})[\varphi_{X_l}(O; \hat{\mathbb{P}}) - \varphi_{X_l}(O; \mathbb{P})]| \leq n^{-1/2} \|\varphi_{X_l}(O; \hat{\mathbb{P}}) - \varphi_{X_l}(O; \mathbb{P})\|_{L_2} = \mathcal{O}_{\mathbb{P}}(n^{-1/2} \varepsilon_n)$$

where the last equality is from Lemma C.9.

Step 3: Bounding the bias remainder R_n . From Lemma C.9, the bias term is bounded by $|R_n| = \mathcal{O}_{\mathbb{P}}(\varepsilon_n^2)$.

Finally, combining the bounds for E_n and R_n with (24) and (25) completes the proof. $\square$

B.5 Proof of Proposition 4.5

Proof of Proposition 4.5. Recall that

$$\phi_{X_l}(\mathbb{P}) = \sum_{j=1}^d L_{lj}^2 \phi_{Z_j}(\mathbb{P}).$$

Summing the above display over l and exchanging the order of summation gives

$$\sum_{l=1}^d \phi_{X_l}(\mathbb{P}) = \sum_{j=1}^d \phi_{Z_j}(\mathbb{P}) \left(\sum_{l=1}^d L_{lj}^2 \right) = \sum_{j=1}^d w_j \phi_{Z_j}(\mathbb{P}). \quad (26)$$

Because $L = L^\top$ and $L^2 = \Sigma^{-1}$, we have

$$w_j = \sum_{l=1}^d L_{lj}^2 = (L^\top L)_{jj} = \Sigma_{jj}. \quad (27)$$

Substituting (27) into (26) yields that

$$\sum_{j=1}^d \Sigma_{jj} \phi_{Z_j}(\mathbb{P}) = \sum_{l=1}^d \phi_{X_l}(\mathbb{P}),$$

which completes the proof. $\square$

B.6 Knothe-Rosenblatt transport

The classical Brenier map is the unique solution of the quadratic optimal-transport problem $c(x, y) = \|x - y\|_2^2$ (Brenier, 1991). A series of papers by Carlier et al. (2010); Bonnotte (2013) shows that the triangular Knothe–Rosenblatt rearrangement T_{KR} may be recovered from Brenier maps by anisotropically rescaling the cost:

$$c_\varepsilon(x, y) = \sum_{j=1}^d \lambda_j(\varepsilon) (x_j - y_j)^2,$$

where $\lambda_1(\varepsilon) \gg \lambda_2(\varepsilon) \gg \dots \gg \lambda_d(\varepsilon)$ and $\lambda_{j+1}(\varepsilon) / \lambda_j(\varepsilon) \rightarrow 0$ as $\varepsilon \rightarrow 0$. Typical weights are $\lambda_j(\varepsilon) = \varepsilon^{j-1}$. Let T_ε be the $\mathcal{W}_2$ -optimal map for c_ε . Then, whenever the source measure $\mathbb{P}$ is absolutely continuous and the target $\mathbb{Q}$ is non-degenerate, Carlier et al. (2010, Theorem 3.1) implies that

$$T_\varepsilon \longrightarrow T_{\text{KR}} \quad \text{in } L^2(\mathbb{P}) \quad \text{as } \varepsilon \rightarrow 0.$$

Hence T_{KR} inherits a genuine optimal-transport interpretation: it is the map one obtains when infinitely prioritising the first coordinate over the second, the second over the third, and so on. The resulting continuation path $\varepsilon \mapsto T_\varepsilon$ provides a homotopy strategy for numerical solvers that bridges the inexpensive T_{KR} and the fully isotropic Brenier map (Carlier et al., 2010).

Below, we recall a self-contained existence lemma.

Lemma B.1 (Knothe–Rosenblatt (K–R) transport). For two Borel probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathbb{R}^d$, consider the transport map $T : X \mapsto Z$ such that $T_\# \mathbb{P} = \mathbb{Q}$:

$$T(x_1, \dots, x_d) = (T_1(x_1), T_2(x_1, x_2), \dots, T_d(x_1, \dots, x_d)) \quad \text{with} \quad T_\# \mathbb{P} = \mathbb{Q},$$

where each coordinate is given by:

$$\begin{aligned} T_1(x_1) &= F_{Z_1}^{-1}(F_{X_1}(x_1)), \\ T_j(x_{1:j}) &= F_{Z_j|Z_{1:(j-1)}}^{-1}\left(F_{X_j|X_{1:(j-1)}}(x_j \mid x_{1:(j-1)}) \mid z_{1:(j-1)}\right), \quad 2 \leq j \leq d, \end{aligned}$$

and $z_{1:(j-1)} := T_{1:(j-1)}(x_{1:(j-1)})$.

If (i) $\mathbb{P}$ -a.s. within the class of triangular, coordinate-wise non-decreasing maps, if $\mathbb{P}$ is absolutely continuous and (ii) each one-dimensional *conditional* cdf of $\mathbb{Q}$ is continuous and strictly increasing, then T is well-defined and unique within the monotone triangular class $\mathbb{P}$ -almost surely.

Proof. We split the proof into 2 steps.

Step 1: Existence. The absolute continuity of $\mathbb{P}$ possesses a joint density $f(x)$; in particular every conditional distribution $F_{X_j|X_{1:j-1}}(\cdot \mid x_{1:j-1})$ is well defined. Thus, we may factorise $\mathbb{P}$ sequentially:

$$d\mathbb{P}(x) = f_{X_1}(x_1) f_{X_2|1}(x_2 \mid x_1) \cdots f_{X_d|1:(d-1)}(x_d \mid x_{1:d-1}) dx.$$

Choose coordinates of T by matching conditional quantiles:

$$T_1(x_1) = F_{Z_1}^{-1}(F_{X_1}(x_1)),$$

$$T_j(x_{1:j}) = F_{Z_j|Z_{1:(j-1)}}^{-1}\left(F_{X_j|X_{1:(j-1)}}(x_j \mid x_{1:(j-1)}) \mid z_{1:(j-1)}\right), \quad 2 \leq j \leq d,$$

where $z_{1:(j-1)} := T_{1:(j-1)}(x_{1:(j-1)})$. Each step is well-defined because the inner cdf is continuous and the outer inverse cdf exists and is unique by the second assumption. Induction shows $T_{\sharp}\mathbb{P} = \mathbb{Q}$, establishing the existence.

Step 2: Uniqueness (within the monotone triangular class). Restrict to maps satisfying $x_j \mapsto T_j(x_{1:j})$ is non-decreasing for every fixed $x_{1:(j-1)}$. Let T and $\tilde{T}$ be two such transports. Because both first coordinates push X_1 to the same measure $\mathbb{Q}_1$ and both are monotone, we must have $T_1(x_1) = \tilde{T}_1(x_1)$ $\mathbb{P}$ -almost surely. By induction, we have $T_j = \tilde{T}_j$ for $j = 2, \dots, d$ $\mathbb{P}$ -almost surely. Hence, uniqueness holds in this class. $\square$

When $\mathbb{Q} = \text{Unif}[0, 1]^{\otimes d}$, the following lemma provides efficient influence function for $\theta_{jl}(\mathbb{P}) = \mathbb{E}[(\partial_{z_l} S_j(Z))^2]$, where $S := T^{-1}$.

Lemma B.2 (Efficient influence function of $\theta_{jl}(\mathbb{P})$ under the triangular (Knothe–Rosenblatt) transport). Let $X = (X_1, \dots, X_d)^\top$ have a joint distribution $\mathbb{P}$ that admits a Lebesgue density and let

$$T(X) = (F_1(X_1), F_{2|1}(X_2 \mid X_1), \dots, F_{d|1:(d-1)}(X_d \mid X_{1:(d-1)})) =: Z \sim \text{Unif}(0, 1)^{\otimes d}.$$

Write $S := T^{-1}$. For any pair of indices $(j, l) \in [d]^2$ define $H_{jl}(X) := \{\partial_{z_l} S_j(Z)\}^2$ and $\theta_{jl}(\mathbb{P}) := \mathbb{E}_{\mathbb{P}}[H_{jl}(X)]$.

- (i) If $l > j$ then $H_{jl}(X) \equiv 0$, so $\theta_{jl}(\mathbb{P}) = 0$ and the efficient influence function is identically 0.
- (ii) If $l \leq j$, the parameter is non-degenerate and its efficient influence function (EIF) is

$$\varphi_{jl}(x) = H_{jl}(x) - \theta_{jl} - 2 H_{jl}(x) \left\{ \mathbb{1}\{X_l \leq x_l\} - F_{l|1:(l-1)}(x_l \mid x_{1:(l-1)}) \right\}, \quad (28)$$

which attains the semiparametric efficiency bound.

Proof of Lemma B.1. If $l > j$, because the inverse K–R map is lower-triangular, S_j depends only on $(z_1, \dots, z_j)$, hence $\partial_{z_l} S_j \equiv 0$ whenever $l > j$; all assertions are immediate.

Next, we analyze the second case in two steps. First, we introduce some notation. Fix (j, l) with $l \leq j$. Denote the lower-triangular Jacobian $J_T(X) := [\partial_{x_k} T_j(X)]_{j,k \leq d}$ and

$$d_j(X) := J_T(X)_{jj} = f_{j|1:(j-1)}(X_j \mid X_{1:(j-1)}), \quad b_{jk}(X) := J_T(X)_{jk} = \partial_{x_k} F_{j|1:(j-1)}(X_j \mid X_{1:(j-1)}), \quad k < j.$$

Because $J_T(X)$ is unit-triangular up to the positive diagonal $\{d_j\}$, its inverse $J_S(Z) = J_T^{-1}(X) = [\gamma_{jk}(X)]_{j,k \leq d}$ is also lower-triangular.

Set

$$W := X_{1:(l-1)}, \quad V := X_l, \quad U := X_{(l+1):d},$$

so that $X = (W, V, U)$ and the factorisation of $\mathbb{P}$ reads $f_X = f_W f_{V|W} f_{U|W,V}$. Because S_j depends only on $X_{1:j}$, H_{jl} may be viewed as a measurable functional of $(W, V, U_{1:(j-l)})$; its exact algebraic form, is multiplicatively proportional to $d_l(X)^{-2}$:

$$H_{jl}(X) = \gamma_{jl}(X)^2 = R_{jl}(X) d_l(X)^{-2}, \quad R_{jl}(X) = \{d_l(X) \gamma_{jl}(X)\}^2. \quad (29)$$

Alternatively, because the inverse-Jacobian entries satisfy

$$\gamma_{ul}(X) = \frac{1}{d_l(X)}, \quad \gamma_{jl}(X) = -\frac{1}{d_j(X)} \sum_{k=l}^{j-1} b_{jk}(X) \gamma_{kl}(X),$$

the weight R_{jl} can be computed recursively:

$$R_{ll}(X) = 1, \quad R_{jl}(X) = \left[-\sum_{k=l}^{j-1} \frac{b_{jk}(X)}{d_j(X)} \sqrt{R_{kl}(X)} \right]^2, \quad l < j.$$

The factor R_{jl} depends on lower-order conditional CDFs only through the values of (W, V, U) and therefore behaves as a fixed, bounded weight in the influence-function calculation.

Pathwise differentiability. Let $\{\mathbb{P}_t : |t| < \varepsilon\}$ be a regular sub-model with score $\dot{\ell}(X)$. Write $f_{V|W,t}$ for the conditional density of V given W under $\mathbb{P}_t$ and set $\dot{f}_{V|W} := \partial_t f_{V|W,t}|_{t=0}$. A simple differentiation gives

$$\dot{f}_{V|W}(v | w) = f_{V|W}(v | w) \left\{ \dot{\ell}(w, v, U) - \mathbb{E}[\dot{\ell}(X) | W = w] \right\}.$$

Hence, differentiating $\theta_{jl}(t) := \mathbb{E}_{\mathbb{P}_t}[H_{jl}(X)]$ and using (29),

$$\dot{\theta}_{jl} = \mathbb{E} \left[\dot{H}_{jl}(X) + H_{jl}(X) \dot{\ell}(X) \right] = \mathbb{E} \left[\{-2H_{jl}(X) \eta(X) + H_{jl}(X) - \theta_{jl}\} \dot{\ell}(X) \right],$$

where $\eta(X) := \dot{\ell}(X) - \mathbb{E}[\dot{\ell}(X) | W] = \dot{\ell}(X) - \mathbb{E}[\dot{\ell}(X) | X_{1:(l-1)}]$. Now note the well-known representation $\eta(X) = \frac{\partial}{\partial v} \left\{ \mathbb{1}\{V \leq v\} - F_{V|W}(v | W) \right\} \Big|_{v=V}$; integrating once in v yields $\eta(X) = f_{V|W}(V | W) \left\{ \mathbb{1}\{V \leq x_l\} - F_{V|W}(x_l | W) \right\}$, so that $-2H_{jl}(X)\eta(X) = -2H_{jl}(X)\{\mathbb{1}\{X_l \leq x_l\} - F_{l|1:(l-1)}(x_l | x_{1:(l-1)})\}$. Substituting back proves that (28) is a centered influence function.

Efficiency. Because φ_{jl} lies in the tangent space and reproduces the pathwise derivative for every sub-model, it is the unique canonical gradient. Square integrability follows from $\mathbb{E}[R_{jl}(X)] < \infty$ and $\mathbb{E}[f_{V|W}^{-4}(V | W)] < \infty$ (which holds whenever the joint density is locally bounded and strictly positive on compact sets). $\square$

Some comments on Lemma B.1 follow.

Firstly, the proof of Lemma B.1 hinges only on the multiplicative decomposition (29); it does not require an explicit closed form for R_{jl} . Hence, the lemma extends to any smooth functional of the K–R map whose density–dependence enters solely through $f_{l[1:(l-1)]}^{-2}$.

Secondly, the EIF is identically zero for $l > j$, reflecting the fact that the triangular Rosenblatt inverse does not respond to perturbations in later coordinates when reconstructing an earlier component.

C Supporting lemmas

C.1 Basic properties of regression estimators

Lemma C.1 (L_2 boundedness). Let (Y, Z) be a pair of random variables with $Z = (Z_1, \dots, Z_d)$ and Z_{-j} denoting all components except Z_j . Define the conditional expectations

$$\eta(Z) := \mathbb{E}[Y \mid Z], \quad \eta_{-j}(Z_{-j}) := \mathbb{E}[Y \mid Z_{-j}].$$

If $\mathbb{E}[Y^2] \leq C$, then it holds that

- (1) $\mathbb{E}[\eta_{-j}(Z_{-j})^2] \leq \mathbb{E}[\eta(Z)^2] \leq C$;
- (2) $\mathbb{E}[(Y - \eta(Z))^2] \leq C$ and $\mathbb{E}[(\eta(Z) - \eta_{-j}(Z_{-j}))^2] \leq C$.

Proof of Lemma C.1. For (1), by Jensen's inequality,

$$\mathbb{E}[\eta(Z)^2] \leq \mathbb{E}[Y^2] < \infty, \quad \mathbb{E}[\eta_{-j}(Z_{-j})^2] \leq \mathbb{E}[Y^2] \leq C.$$

For (2), by the law of total variance,

$$\mathbb{V}(Y) = \mathbb{E}[\mathbb{V}(Y \mid Z)] + \mathbb{V}(\mathbb{E}[Y \mid Z]).$$

Since $\mathbb{E}[Y^2] < \infty$, both terms on the right-hand side are finite. In particular,

$$\mathbb{E}[(Y - \eta(Z))^2] = \mathbb{E}[\mathbb{V}(Y \mid Z)] \leq \mathbb{V}[Y] \leq \mathbb{E}[Y^2] \leq C.$$

Next, define the function $\delta(Z) := \eta(Z) - \eta_{-j}(Z_{-j})$. Then,

$$\mathbb{E}[\delta(Z)^2] = \mathbb{E}[(\eta(Z) - \eta_{-j}(Z_{-j}))^2] \leq 2(\mathbb{E}[\eta(Z)^2] + \mathbb{E}[\eta_{-j}(Z_{-j})^2]).$$

From (1), we have

$$\mathbb{E}[(\eta(Z) - \eta_{-j}(Z_{-j}))^2] \leq \mathbb{E}[\eta(Z)^2] \leq C.$$

This completes the proof. $\square$

Proposition C.2 (High-probability Lipschitz bound for the regression estimator). Under Assumptions 3.1–4.2, assume that the estimator $\hat{\mu}$ satisfies the mean-squared error $\varepsilon := \|\hat{\mu} - \mu\|_{L_2(\mathbb{P}_X)} \in (0, 1/\sqrt{2})$. Define the data-dependent radius

$$R_\varepsilon := \lambda \sqrt{d + 2 \log(1/\varepsilon)}, \tag{30}$$

and the ball

$$B := \{x \in \mathbb{R}^d : \|x\| \leq R_\varepsilon\}.$$

Then there exists a universal constant $C = 4\sqrt{2}$, such that

$$\mathbb{P}_Z(\text{Lip}(\hat{\mu}; B) \leq L_\mu + C L_\mu^{1/2} \lambda^{-1/2} [d + 2 \log(1/\varepsilon)]^{-1/4} \varepsilon^{1/2}) \geq 1 - \varepsilon^2.$$

Proof of Proposition C.2. Throughout the proof, we rely on two arguments, truncation and a covering-number bound. We split the proof into multiple steps.

Step 1: Truncation of the input domain. For $Z \sim \mathcal{N}(0, I_d)$ the standard tail bound $\mathbb{P}(\|Z\| > t) \leq \exp(-t^2/2)$ implies

$$\mathbb{P}(\|Z\| > R_\varepsilon/\lambda) \leq \exp(-\tfrac{1}{2}[d + 2\log(1/\varepsilon)]) = \varepsilon^2.$$

Because $X = T^{-1}(Z)$ and T^{-1} is λ -Lipschitz, $\|X\| \leq \lambda\|Z\|$; hence, with probability at least $1 - \varepsilon^2$, $X \in B$.

Step 2: A finite h -net of the truncated ball. Conditioned on event $X \in B$, and fix a mesh size $h > 0$. The ball B admits an h -net $\mathcal{N}_h = \{x_1, \dots, x_{N_h}\}$ of cardinality $N_h \leq (1 + 2R_\varepsilon/h)^d$. Write $\Delta_\mu := \hat{\mu} - \mu$. By Cauchy-Schwarz,

$$\max_{x_i \in \mathcal{N}_h} |\Delta_\mu(x_i)| \leq \frac{\|\Delta_\mu\|_{L_2(\mathbb{P}_X)}}{\sqrt{\mathbb{P}_X(B)}} \leq \frac{\varepsilon}{\sqrt{1 - \varepsilon^2}} \leq 2\varepsilon. \quad (31)$$

Step 3: Lipschitz bound on the ball B . For arbitrary $x, y \in B$ choose $x_i, x_j \in \mathcal{N}_h$ with $\|x - x_i\| \leq h$ and $\|y - x_j\| \leq h$. Insert and remove μ :

$$\begin{aligned} |\hat{\mu}(x) - \hat{\mu}(y)| &\leq |\hat{\mu}(x) - \hat{\mu}(x_i)| + |\hat{\mu}(x_i) - \hat{\mu}(x_j)| + |\hat{\mu}(x_j) - \hat{\mu}(y)| \\ &\leq L_\mu(\|x - x_i\| + \|x_j - y\|) + |\Delta_\mu(x_i)| + |\Delta_\mu(x_j)| \\ &\leq L_\mu(\|x - y\| + 2h) + 4\varepsilon, \end{aligned}$$

where the second line uses the L_μ -Lipschitzness of μ and the third invokes (31). Dividing by $\|x - y\|$ and taking the supremum over $x, y \in B$ yields

$$\text{Lip}(\hat{\mu}; B) \leq L_\mu + 2L_\mu h/R_\varepsilon + 4\varepsilon/h. \quad (32)$$

Step 4: Optimizing the mesh size. Minimizing the right-hand side of (32) over $h > 0$ gives the choice

$$h^* = \sqrt{\frac{2\varepsilon R_\varepsilon}{L_\mu}}.$$

Substituting it into (32) yields that

$$\begin{aligned} \text{Lip}(\hat{\mu}; B) &\leq L_\mu + 4\sqrt{2}\sqrt{\frac{L_\mu \varepsilon}{R_\varepsilon}} \\ &= L_\mu + C L_\mu^{1/2} \lambda^{-1/2} [d + 2\log(1/\varepsilon)]^{-1/4} \varepsilon^{1/2}, \end{aligned} \quad (33)$$

where the last equality is from the definition of R_ε in (30) and $C = 4\sqrt{2}$. This finishes the proof. $\square$

C.2 Basic properties of transport maps

Lemma C.3 (Bi-Lipschitz $\implies$ Bijective). Let $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ satisfy the *global bi-Lipschitz* bounds

$$c \|x - y\| \leq \|T(x) - T(y)\| \leq C \|x - y\|, \quad \forall x, y \in \mathbb{R}^d,$$

for some constants $0 < c \leq C < \infty$. Then T is a bijection and a bi-Lipschitz homeomorphism of $\mathbb{R}^d$ onto itself.

Proof of Lemma C.3. We split the proof into two steps.

Step 1: Bijection. We verify the properties of T one by one:

- (a) *Injectivity.* If $T(x) = T(y)$ then $c\|x - y\| \leq 0$, so $x = y$. Hence, T is injective
- (b) *Properness.* From the lower bound, $\|T(x)\| \geq c\|x\| - \|T(0)\|$, hence $\|x\| \rightarrow \infty \implies \|T(x)\| \rightarrow \infty$. Thus inverse images of bounded sets are bounded (and closed by continuity), hence compact.
- (c) *Openness.* For any ball $B(x, r)$ one has $B(T(x), cr) \subset T(B(x, r)) \subset B(T(x), Cr)$, so T maps open sets to open sets.
- (d) *Surjectivity.* The image $T(\mathbb{R}^d)$ is non-empty, open (3) and closed (properness + continuity). Since $\mathbb{R}^d$ is connected, the only such subset is $\mathbb{R}^d$ itself.

Consequently, T is a bijection.

Step 2: Lipschitzness of inverse mapping. For any two points $u, v \in \mathbb{R}^d$ and set $x := T^{-1}(u)$, $y := T^{-1}(v)$. This gives

$$c\|x - y\| \leq \|T(x) - T(y)\| = \|u - v\|.$$

Rearranging yields

$$\|T^{-1}(u) - T^{-1}(v)\| = \|x - y\| \leq \frac{1}{c} \|u - v\|, \quad \forall u, v \in \mathbb{R}^d.$$

Hence T^{-1} is globally Lipschitz with

$$\text{Lip}(T^{-1}) \leq 1/c.$$

Consequently, T is a bijection and a bi-Lipschitz homeomorphism of $\mathbb{R}^d$ onto itself. $\square$

Lemma C.4 (Bi-Lipschitz gradient $\iff$ uniform Hessian bounds). Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a C^2 , convex function and set $T = \nabla \Phi$. The following are equivalent:

- (a) Uniform Hessian bounds: There exist constants $0 < m \leq M < \infty$ such that, for all $x \in \mathbb{R}^d$,

$$m I_d \preceq \nabla^2 \Phi(x) \preceq M I_d. \tag{34}$$

- (b) Global bi-Lipschitz gradient: There exist constants $0 < m \leq M < \infty$ such that, for all

$$x, y \in \mathbb{R}^d,$$

$$m\|x - y\| \leq \|\nabla\Phi(x) - \nabla\Phi(y)\| \leq M\|x - y\|. \quad (35)$$

Proof of Lemma C.4. We split the proof into different steps.

Step 1: (a) $\Rightarrow$ (b). For $x, y \in \mathbb{R}^d$ define $\gamma(t) = y + t(x - y)$, $t \in [0, 1]$. By the fundamental theorem of calculus,

$$\nabla\Phi(x) - \nabla\Phi(y) = \int_0^1 \nabla^2\Phi(\gamma(t)) (x - y) dt.$$

Applying the bounds (34) inside the integral gives

$$m\|x - y\| \leq \|\nabla\Phi(x) - \nabla\Phi(y)\| \leq M\|x - y\|,$$

establishing (35).

Step 2: (b) $\Rightarrow$ (a). Fix $x \in \mathbb{R}^d$ and a unit vector v . For $t > 0$ let $y = x + tv$. Using (35),

$$mt \leq \|\nabla\Phi(x + tv) - \nabla\Phi(x)\| \leq Mt.$$

Divide by t and send $t \downarrow 0$ to obtain $m \leq \|\nabla^2\Phi(x)v\| \leq M$. Since v is arbitrary, this yields (34). $\square$

C.3 LOCO estimation

Lemma C.5 (Estimated influence function for LOCO). Assume that $\hat{\eta}$ and η are bounded in L_∞ with probability tending to one. Under Assumption 4.2, the influence function estimation error for the LOCO estimator (12) satisfies that

$$\|\varphi_{Z_j}(O; \hat{\mathbb{P}}) - \varphi_{Z_j}(O; \mathbb{P})\|_{L_2} = \mathcal{O}_{\mathbb{P}}(\|\hat{\eta} - \eta\|_{L_2} + \|\hat{\eta}_{-j} - \eta_{-j}\|_{L_2} + \|\hat{\eta} - \eta\|_{L_2}^2 + \|\hat{\eta}_{-j} - \eta_{-j}\|_{L_2}^2),$$

Proof of Lemma C.5. For the proof below, we drop the subscript Z_j and the dependency in (Z, Z_{-j}) for simplicity. Recall the efficient influence function

$$\varphi(O; \mathbb{P}) = 2(Y - \eta)(\eta - \eta_{-j}) + (\eta - \eta_{-j})^2 - \phi, \quad \eta := \mathbb{E}[Y | Z], \quad \eta_{-j} := \mathbb{E}[Y | Z_{-j}].$$

For the estimated distribution $\hat{\mathbb{P}}$ we plug in $\hat{\eta}, \hat{\eta}_{-j}$:

$$\varphi(O; \hat{\mathbb{P}}) = 2(Y - \hat{\eta})(\hat{\eta} - \hat{\eta}_{-j}) + (\hat{\eta} - \hat{\eta}_{-j})^2 - \hat{\phi}_{Z_j}.$$

Denote $\delta := \hat{\eta} - \eta$, $\delta_{-j} := \hat{\eta}_{-j} - \eta_{-j}$ and $V := \delta - \delta_{-j}$. Observe $\hat{\eta} - \hat{\eta}_{-j} = (\eta - \eta_{-j}) + V$. Then, we have

$$\varphi(O; \hat{\mathbb{P}}) - \varphi(O; \mathbb{P}) = 2[(Y - \hat{\eta})(\hat{\eta} - \hat{\eta}_{-j}) - (Y - \eta)(\eta - \eta_{-j})] + [(\hat{\eta} - \hat{\eta}_{-j})^2 - (\eta - \eta_{-j})^2] + (\phi - \hat{\phi}).$$

Insert $Y - \hat{\eta} = (Y - \eta) - \delta$ and $\hat{\eta} - \hat{\eta}_{-j} = (\eta - \eta_{-j}) + V$, it follows that

$$\varphi(O; \hat{\mathbb{P}}) - \varphi(O; \mathbb{P}) = 2(Y - \eta)V - 2\delta(\eta - \eta_{-j}) - 2\delta V + 2(\eta - \eta_{-j})V + V^2 + (\phi - \hat{\phi}).$$

Because $(\eta - \eta_{-j})V - \delta(\eta - \eta_{-j}) = -(\eta - \eta_{-j})\delta_{-j}$, we further have

$$\varphi(O; \hat{\mathbb{P}}) - \varphi(O; \mathbb{P}) = 2(Y - \eta)V - 2(\eta - \eta_{-j})\delta_{-j} - 2\delta V + V^2 + (\phi - \hat{\phi}).$$

Under the boundedness assumption $\mathbb{E}[Y^2] \leq C$ with Lemma C.1 and using Cauchy–Schwarz:

$$\begin{aligned} \|\varphi(O; \hat{\mathbb{P}}) - \varphi(O; \mathbb{P})\|_{L_2} &\leq 2\|Y - \eta\|_{L_2}\|V\|_{L_2} + 2\|\eta - \eta_{-j}\|_{L_2}\|\delta_{-j}\|_{L_2} + 2\|\delta\|_{L_2}\|V\|_{L_2} + \|V\|_{L_2}^2 + \|\phi - \hat{\phi}\|_{L_2} \\ &\leq 4C\|V\|_{L_2} + 2C\|\delta_{-j}\|_{L_2} + \|V\|_{L_2}^2 + (\phi - \hat{\phi}). \end{aligned}$$

Because $\hat{\phi}$ is uniformly bounded in probability and $\hat{\phi} - \phi = o_{\mathbb{P}}(1)$, by dominate convergence theorem, we have that $\|\phi - \hat{\phi}\|_{L_2} = o_{\mathbb{P}}(1)$. Since $\|V\|_{L_2} \leq \|\delta\|_{L_2} + \|\delta_{-j}\|_{L_2}$, the right-hand side is

$$\mathcal{O}_{\mathbb{P}}(\|\hat{\eta} - \eta\|_{L_2} + \|\hat{\eta}_{-j} - \eta_{-j}\|_{L_2} + \|\hat{\eta} - \eta\|_{L_2}^2 + \|\hat{\eta}_{-j} - \eta_{-j}\|_{L_2}^2),$$

which completes the proof. $\square$

Lemma C.6 (L_2 -bound for LOO estimator). Let $\hat{\eta}$ be an estimator of η , estimated from independent samples. Suppose $Z_j \perp\!\!\!\perp Z_{-j}$ and define the leave- j -out estimator

$$\hat{\eta}_{-j}(z_{-j}) = \mathbb{E}[\hat{\eta}(Z) | \hat{\eta}, Z_{-j} = z_{-j}] = \mathbb{E}[\hat{\eta}(Z^{(j)}) | \hat{\eta}, Z_{-j} = z_{-j}],$$

for $\eta_{-j}(z_{-j}) = \mathbb{E}[\eta(Z) \mid Z_j = z_j]$. Then

$$\|\widehat{\eta}_{-j}(Z_{-j}) - \eta_{-j}(Z_{-j})\|_{L_2(P_{Z_{-j}})} \leq \|\widehat{\eta}(Z) - \eta(Z)\|_{L_2(\mathbb{P}_Z)}.$$

Proof of Lemma C.6. Since $Z_j \perp\!\!\!\perp Z_{-j}$, the joint law of (Z_{-j}, Z_j) (the same holds for $(Z_{-j}, Z_j^{(j)})$) factors and by Jensen's inequality

$$\begin{aligned} \mathbb{E}_{Z_{-j}}[(\widehat{\eta}_{-j}(Z_{-j}) - \eta_{-j}(Z_{-j}))^2] &= \mathbb{E}_{Z_{-j}}[(\int [\widehat{\eta}(Z_{-j}, z_j) - \eta(Z_{-j}, z_j)] dF_{Z_j}(z_j))^2] \\ &\leq \mathbb{E}_{Z_{-j}}[\int [\widehat{\eta}(Z_{-j}, z_j) - \eta(Z_{-j}, z_j)]^2 dF_{Z_j}(z_j)] \\ &= \mathbb{E}_Z[(\widehat{\eta}(Z) - \eta(Z))^2]. \end{aligned}$$

Taking square roots yields the result. □

C.4 Disentangled feature importance

Lemma C.7 (EIF of the expected product). Let the data be $O \sim P$ and suppose that the (possibly distribution-dependent) functions A and B satisfy

$$\psi_A(P) = \mathbb{E}_P[A(O)], \quad \psi_B(P) = \mathbb{E}_P[B(O)].$$

Denote by $\varphi_A(O; P)$ and $\varphi_B(O; P)$ their efficient influence functions (EIFs), normalised so that

$$\mathbb{E}_P[\varphi_A(O; P)] = \mathbb{E}_P[\varphi_B(O; P)] = 0.$$

Define the centered functions $A_c(O) = A(O) - \psi_A(P)$ and $B_c(O) = B(O) - \psi_B(P)$, and the remainder term

$$\alpha_A(O) := \varphi_A(O; P) - A_c(O), \quad \alpha_B(O) := \varphi_B(O; P) - B_c(O),$$

where the remainder terms α_A, α_B collect the pathwise-derivative contributions coming from the dependence of A and B on P . Then, the EIF of $\psi_{AB}(P) := \mathbb{E}_P[A(O)B(O)]$ is given by:

$$\varphi_{AB}(O; P) = A(O)B(O) - \psi_{AB}(P) + B(O)\alpha_A(O) + A(O)\alpha_B(O). \quad (36)$$

Proof of Lemma C.7. Let $P_\varepsilon = (1 + \varepsilon h)P$ be a regular, mean-zero submodel with score h . Differentiating $\psi_{AB}(P_\varepsilon) = \int A_{P_\varepsilon} B_{P_\varepsilon} dP_\varepsilon$ at $\varepsilon = 0$ gives

$$\left. \frac{d}{d\varepsilon} \psi_{AB}(P_\varepsilon) \right|_{\varepsilon=0} = \mathbb{E}_P[h(O) A(O)B(O)] + \mathbb{E}_P[B(O) \dot{A}_h(O)] + \mathbb{E}_P[A(O) \dot{B}_h(O)], \quad (37)$$

where $\dot{A}_h, \dot{B}_h$ are the pathwise derivatives of $A_{P_\varepsilon}(O)$ and $B_{P_\varepsilon}(O)$ along h .

The defining property of φ_A and φ_B is

$$\mathbb{E}_P[h(O) \varphi_A(O; P)] = \mathbb{E}_P[h(O) A(O)] + \mathbb{E}_P[\dot{A}_h(O)], \quad (38)$$

and similarly for B . This implies

$$\mathbb{E}_P[\dot{A}_h(O)] = \mathbb{E}_P[h(O) \alpha_A(O)], \quad \mathbb{E}_P[\dot{B}_h(O)] = \mathbb{E}_P[h(O) \alpha_B(O)]. \quad (39)$$

Define

$$\varphi_{AB}(O; P) := A(O)B(O) - \psi_{AB}(P) + B(O) \alpha_A(O) + A(O) \alpha_B(O). \quad (40)$$

We verify that (40) satisfies the EIF equation. Using (37)-(39), we have

$$\mathbb{E}_P[h(O) \varphi_{AB}(O; P)] = \left. \frac{d}{d\varepsilon} \psi_{AB}(P_\varepsilon) \right|_{\varepsilon=0},$$

and $\mathbb{E}_P[\varphi_{AB}] = 0$ by construction, so (40) is indeed the efficient influence function of ψ_{AB} . This completes the proof. $\square$

Remark C.1 (Special cases of Lemma C.7). When A is distribution-dependent, one has $\alpha_A = 0$ and (36) becomes $\varphi_{AB} = A\varphi_B$. When neither A nor B depends on P , one has $\alpha_A = \alpha_B = 0$, $\varphi_A = A_c$, $\varphi_B = B_c$, and (36) reduces to $\varphi_{AB} = AB - \mathbb{E}[AB]$. This can also happen if A and B are linear functions. For instance, for the covariance parameter $\psi_{AB}(P) = \mathbb{E}_P[(A - \mathbb{E}_P[A])(B - \mathbb{E}_P[B])]$, one has $X(O) = X$ and $\varphi_X(O; P) = X_c(O) = X - \mathbb{E}_P[X]$ for $X \in \{A, B\}$. Hence, $\alpha_A = \alpha_B = 0$ and the last two terms in (36) vanish and (36) reduces to $\varphi_{AB}(O; P) = (A - \mathbb{E}_P[A])(B - \mathbb{E}_P[B]) - \psi_{AB}(P)$.

Proposition C.8 (EIF of $\theta(\mathbb{P})$ under Bures-Wasserstein transport). Let the observable $O = X \in \mathbb{R}^d$ satisfy

$$\mathbb{P}[X] = 0, \quad \Sigma = \mathbb{P}[XX^\top] \succ 0, \quad L := \Sigma^{-1/2}.$$

For fixed indices $j, l \in [d]$ define

$$T(x) := Lx, \quad H_{jl}(X) := \partial_{x_l} T_j(X) = L_{jl}.$$

The efficient influence function of the parameter

$$\theta_{jl}(\mathbb{P}) := \mathbb{P}[H_{jl}(X)^2] = L_{jl}^2,$$

is given by

$$\varphi_{jl}(O; \mathbb{P}) = -L_{jl}(X^\top L e_j e_l^\top L X - \delta_{jl}) = -L_{jl}(Z_j Z_l - \delta_{jl}).$$

Proof of Proposition C.8. Take an arbitrary regular one-dimensional sub-model $\{\mathbb{P}_\varepsilon : \varepsilon \in (-\varepsilon_0, \varepsilon_0)\}$ embedded in the non-parametric model such that $\mathbb{P}_{\varepsilon=0} = \mathbb{P}$ and with score $s_\varepsilon(O) = \partial_\varepsilon \log p_\varepsilon(O)|_{\varepsilon=0}$. The usual regularity conditions give $\mathbb{P}[s(O)] = 0$ and $\mathbb{P}[Xs(O)] = 0$ (the latter enforces the constraint $\mathbb{E}[X] = 0$ along the path).

Write $\Sigma(\varepsilon) := \mathbb{E}_{\mathbb{P}_\varepsilon}[XX^\top]$. Differentiating at $\varepsilon = 0$,

$$\partial_\varepsilon \Sigma(\varepsilon)|_{\varepsilon=0} = \mathbb{P}[XX^\top s(O)] = \mathbb{P}[\{XX^\top - \Sigma\} s(O)],$$

because $\mathbb{P}[s(O)] = 0$. Hence, the influence function for Σ is

$$U_\Sigma(O) := XX^\top - \Sigma, \quad \mathbb{P}[U_\Sigma(O)] = 0.$$

Treat θ_{jl} as the composition $\theta_{jl} = h \circ \Sigma$ with $h(M) = (M^{-1/2})_{jl}^2$. A matrix calculus identity gives the Fréchet derivative of $M \mapsto M^{-1/2}$:

$$dL = d(\Sigma^{-\frac{1}{2}}) = -\frac{1}{2} L (d\Sigma) L.$$

Because $\langle A, B \rangle_F := \text{tr}(A^\top B)$, we can rewrite the last term as an inner product:

$$d\theta_{jl} = \langle \nabla_\Sigma \theta_{jl}(\Sigma), d\Sigma \rangle_F, \quad \nabla_\Sigma \theta_{jl}(\Sigma) = -\frac{L_{jl}}{2} (L e_j e_l^\top L + L e_l e_j^\top L).$$

Insert $d\Sigma = \partial_\varepsilon \Sigma(\varepsilon)|_{\varepsilon=0} = \mathbb{P}[U_\Sigma(O) s(O)]$ from Step 1:

$$\partial_\varepsilon \theta_{jl}(\mathbb{P}_\varepsilon) \Big|_{\varepsilon=0} = \langle \nabla_\Sigma \theta_{jl}, \mathbb{P}[U_\Sigma(O) s(O)] \rangle_F = \mathbb{P}[\langle \nabla_\Sigma \theta_{jl}, U_\Sigma(O) \rangle_F s(O)].$$

Therefore, the efficient influence function φ_{jl} in the tangent space satisfies that $\partial_\varepsilon \theta_{jl}(\mathbb{P}_\varepsilon)|_{\varepsilon=0} = \mathbb{P}[\varphi_{jl}(O) s(O)]$ for every score s is

$$\varphi_{jl}(O; \mathbb{P}) = \langle \nabla_\Sigma \theta_{jl}(\Sigma), U_\Sigma(O) \rangle_F = -L_{jl} \{X^\top L e_j e_l^\top L X - \delta_{jl}\}.$$

Centring holds automatically because $\mathbb{P}[X^\top L e_j e_l^\top L X] = \delta_{jl}$, so $\mathbb{P}[\varphi_{jl}(O; \mathbb{P})] = 0$. $\square$

Lemma C.9 (Estimated influence function for ϕ_{X_l}). Under the assumptions of Theorem 4.4, let

$$\varphi_{X_l}(O; \mathbb{P}) = \sum_{j=1}^d \{L_{jl}^2 [2(Y - \eta(Z)) \Delta_j(Z) + \Delta_j(Z)^2] - L_{jl}(Z_j Z_l - \delta_{jl}) \Delta(Z)^2\} - \phi_{X_l}(\mathbb{P}),$$

and let $\varphi_{X_l}(O; \widehat{\mathbb{P}})$ be the same expression with (η, L) replaced by $(\widehat{\eta}, \widehat{L})$ and η_{-j} replaced with $\widehat{\eta}_{-j} = \mathbb{E}_{Z_j}[\widehat{\eta}(Z) | Z_{-j}]$. Then

$$\|\varphi_{X_l}(O; \widehat{\mathbb{P}}) - \varphi_{X_l}(O; \mathbb{P})\|_{L_2} = \mathcal{O}_{\mathbb{P}}(\varepsilon_n), \quad \mathbb{P}\{\varphi_{X_l}(O; \widehat{\mathbb{P}}) - \varphi_{X_l}(O; \mathbb{P})\} = \mathcal{O}_{\mathbb{P}}(\varepsilon_n^2),$$

where $\varepsilon_n = \|L_{\cdot l}\|_2 \|\widehat{\mu} - \mu\|_{L_2} + \text{tr}(\Sigma)^{1/2} \|\widehat{L} - L\|_{\text{op}}$.

Proof of Lemma C.9. For simplicity, we will not explicitly indicate the dependency on Z . For $l \in [d]$ fixed and for every $j \in [d]$, we use the following notation $\delta = \widehat{\eta} - \eta$, $\delta_{-j} = \widehat{\eta}_{-j} - \eta_{-j}$, $\vartheta_j = \delta - \delta_{-j}$, $\Delta_j = \eta - \eta_{-j}$, $\widehat{\Delta}_j = \widehat{\eta} - \widehat{\eta}_{-j}$, $D_{jl} := \widehat{L}_{jl}^2 - L_{jl}^2$, $U = Y - \eta$, $\widehat{U} = Y - \widehat{\eta}$.

Step 1: Decomposition. Put

$$A_j := L_{jl}^2 [2U \Delta_j + \Delta_j^2], \quad B_j := L_{jl} (Z_j Z_l - \delta_{jl}) \Delta_j^2,$$

and denote by $\widehat{A}_j, \widehat{B}_j$ the corresponding quantities with (L_{jl}, η, Δ_j) replaced by $(\widehat{L}_{jl}, \widehat{\eta}, \widehat{\Delta}_j)$. Then

$$\varphi_{X_l}(O; \widehat{\mathbb{P}}) - \varphi_{X_l}(O; \mathbb{P}) = \sum_{j=1}^d (E_{1,j} + E_{2,j}) - [\phi_{X_l}(\widehat{\mathbb{P}}) - \phi_{X_l}(\mathbb{P})], \quad (41)$$

where $E_{1,j} := \widehat{A}_j - A_j$ and $E_{2,j} := -\widehat{B}_j + B_j$. Using $\widehat{\Delta}_j = \Delta_j + \vartheta_j$ and $U - \delta = Y - \widehat{\eta}$ we obtain

$$E_{1,j} = D_j^{(2)} [2U \Delta_j + \Delta_j^2] + \widehat{L}_{jl}^2 [2(U - \delta) \vartheta_j + \vartheta_j^2 - 2\delta_{-j} \Delta_j], \quad (42)$$

$$E_{2,j} = -D_j^{(1)} (Z_j Z_l - \delta_{jl}) \Delta_j^2 - \widehat{L}_{jl} (Z_j Z_l - \delta_{jl}) [2\Delta_j \vartheta_j + \vartheta_j^2], \quad (43)$$

with $D_j^{(2)} := \widehat{L}_{jl}^2 - L_{jl}^2$, $D_j^{(1)} := \widehat{L}_{jl} - L_{jl}$.

Because the plug-in estimator of ϕ_{X_l} employs the empirical mean of $\widehat{\Delta}_j^2$,

$$\phi_{X_l}(\widehat{\mathbb{P}}) - \phi_{X_l}(\mathbb{P}) = \sum_{j=1}^d \{D_j^{(2)} \mathbb{P}[\Delta_j^2] + \widehat{L}_{jl}^2 \mathbb{P}[2\Delta_j \vartheta_j + \vartheta_j^2] + \widehat{L}_{jl} (\mathbb{P}_n - \mathbb{P})[\widehat{\Delta}_j^2]\}. \quad (44)$$

Substituting (42)–(44) into (41) and collecting like terms yields

$$\varphi_{X_l}(O; \widehat{\mathbb{P}}) - \varphi_{X_l}(O; \mathbb{P}) = T_{1l} + T_{2l} + T_{3l} + T_{4l}, \quad (45)$$

where

$$\begin{aligned} T_{1l} &:= \sum_j D_j^{(2)} [2U\Delta_j + \Delta_j^2 - \mathbb{P}[\Delta_j^2]], \\ T_{2l} &:= \sum_j \widehat{L}_{jl}^2 \{2(U - \delta)\vartheta_j + \vartheta_j^2 - 2\delta_{-j}\Delta_j - \mathbb{P}[2\Delta_j\vartheta_j + \vartheta_j^2]\}, \\ T_{3l} &:= \sum_j [-D_j^{(1)}(Z_j Z_l - \delta_{jl}) \Delta_j^2 - \widehat{L}_{jl}(Z_j Z_l - \delta_{jl})(2\Delta_j\vartheta_j + \vartheta_j^2)], \\ T_{4l} &:= -\sum_j \widehat{L}_{jl}^2 (\mathbb{P}_n - \mathbb{P})[\widehat{\Delta}_j^2]. \end{aligned}$$

Step 2: L_2 -norm bound. We bound $\|T_{kl}\|_{L_2}$ for $k = 1, \dots, 4$, term by term. Define $e_{L,l} := \|\widehat{L}_{\cdot l} - L_{\cdot l}\|_2$, $e_\eta = \|\widehat{\eta} - \eta\|_{L_2}$, and $e_n = e_{L,l} + \|L_{\cdot l}\|_2 e_\eta$.

(i) *Term T_{1l} .* Because $|D_j^{(2)}| \leq 2C|D_j^{(1)}|$ and $\|2U\Delta_j + \Delta_j^2 - \mathbb{P}[\Delta_j^2]\|_{L_2} \leq 10C^2$,

$$\|T_{1l}\|_{L_2} \leq 20C^3 \sum_j |D_j^{(1)}| \leq 20C^3 e_{L,l} = \mathcal{O}_{\mathbb{P}}(e_n).$$

(ii) *Term T_{2l} .* Each factor in the curly braces of T_{2l} contains at least one of ϑ_j or δ_{-j} :

$$\|\vartheta_j\|_{L_2} \leq 2(\|\widehat{\eta} - \eta\|_{L_2} + \|\widehat{\eta}_{-j} - \eta_{-j}\|_{L_2}), \quad \|\delta_{-j}\|_{L_2} \leq \|\widehat{\eta}_{-j} - \eta_{-j}\|_{L_2}.$$

Hence, from Cauchy-Schwartz inequality and Lemma C.6, it follows that

$$\|T_{2l}\|_{L_2} \leq \|\widehat{L}_{\cdot l}\|_{L_2} (\|\widehat{\eta} - \eta\|_{L_2} + \|\widehat{\eta}_{-j} - \eta_{-j}\|_{L_2}) = \mathcal{O}_{\mathbb{P}}(e_n),$$

by noting that $\|\widehat{L}_{\cdot l}\|_{L_2} \leq \|L_{\cdot l}\|_2 + e_{L,l}$.

(iii) *Term T_{3l} .* For the first piece in T_{3l} , $\|(Z_j Z_l - \delta_{jl})\Delta_j^2\|_{L_2} \leq 4C^3$, so $\|D_j^{(1)}(\cdot)\|_{L_2} \leq 4C^3|D_j^{(1)}|$. Summation gives $\mathcal{O}_{\mathbb{P}}(e_{L,l})$. For the second piece, note that $\|Z_j Z_l - \delta_{jl}\|_{L_2} \leq 2C^2$ and $\|2\Delta_j\vartheta_j + \vartheta_j^2\|_{L_2} \lesssim \|\widehat{\eta} - \eta\|_{L_2} + \|\widehat{\eta}_{-j} - \eta_{-j}\|_{L_2} \leq 2e_\eta$, from Lemma C.6. Hence, by Cauchy-Schwarz inequality, we have

$$\|T_{3l}\|_{L_2} = \mathcal{O}_{\mathbb{P}}(e_n).$$

(iv) *Term T_{4l} .* Conditional on the nuisance estimates, $\mathbb{V}\{(\mathbb{P}_n - \mathbb{P})[\widehat{\Delta}_j^2]\} \leq \frac{16C^4}{n}$, so $\|(\mathbb{P}_n - \mathbb{P})[\widehat{\Delta}_j^2]\|_{L_2} = \mathcal{O}(n^{-1/2})$. With $|\widehat{L}_{jl}| \leq C$, $\|T_{4l}\|_{L_2} = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$.

Putting the four bounds together, we have

$$\|\varphi_{X_l}(O; \widehat{\mathbb{P}}) - \varphi_{X_l}(O; \mathbb{P})\|_{L_2} \leq \|T_{1l}\|_{L_2} + \|T_{2l}\|_{L_2} + \|T_{3l}\|_{L_2} + \|T_{4l}\|_{L_2} = \mathcal{O}_{\mathbb{P}}(e_n + n^{-1/2}).$$

Step 3: Expectation bound. We again bound the expectation term by term.

(i) *Term T_{1l} .* T_{1l} is of the form $D_j^{(2)}\{2U\Delta_j + \Delta_j^2 - \mathbb{E}[\Delta_j^2]\}$. Conditional on the (independent) nuisance estimates, the factor $D_j^{(2)}$ is fixed, while $\mathbb{E}[2U\Delta_j | Z] = 0$ and $\mathbb{E}[\Delta_j^2] = \mathbb{E}[\Delta_j^2]$ by definition, so the centered bracket has mean 0. Thus, $\mathbb{E}[T_{1l}] = 0$.

(ii) *Term T_{2l} .* Each term inside the braces of T_{2l} contains at least one factor $\vartheta_j = \delta - \delta_{-j}$ or δ_{-j} . Because $\|\delta_{-j}\|_{L_2} \leq \|\delta\|_{L_2} = e_\eta$, Cauchy–Schwarz inequality gives

$$|\mathbb{E}[(U - \delta)\vartheta_j]| = |\mathbb{E}[\delta\vartheta_j]| \leq \|\delta\|_{L_2}\|\vartheta_j\|_{L_2} \lesssim e_\eta^2,$$

and similarly $|\mathbb{E}[\delta_{-j}\Delta_j]| \lesssim e_\eta^2$. All other raw moments in the bracket are of this same order. Therefore, we have

$$|\mathbb{E}[T_{2l}]| \lesssim \|\widehat{L}_{\cdot l}\|_{L_2}^2 e_\eta^2 = \mathcal{O}_\mathbb{P}(e_n^2).$$

(iii) *Term T_{3l} .* Because $\mathbb{E}[Z_j Z_l - \delta_{jl}] = 0$, we have

$$\mathbb{P}[T_{3l}] = - \sum_j \widehat{L}_{jl} \mathbb{E}[(Z_j Z_l - \delta_{jl})(2\Delta_j \vartheta_j + \vartheta_j^2)].$$

Every summand contains at least one factor ϑ_j ; by Cauchy–Schwarz,

$$|\mathbb{E}[(Z_j Z_l - \delta_{jl})(2\Delta_j \vartheta_j + \vartheta_j^2)]| \leq 3C^2 \|\vartheta_j\|_{L_2} \max\{2\|\Delta_j\|_\infty, \|\vartheta_j\|_{L_2}\} = \mathcal{O}_\mathbb{P}(e_\eta^2).$$

Multiplying by $|\widehat{L}_{jl}| \leq C$ and summing gives

$$|\mathbb{E}[T_{3l}]| \lesssim \|\widehat{L}_{\cdot l}\|_{L_2} e_\eta^2 = \mathcal{O}_\mathbb{P}(e_n^2).$$

(iv) *Term T_{4l} .* Because $\mathbb{E}[(\mathbb{P}_n - \mathbb{P})f] = 0$ for any square-integrable f , we have $\mathbb{E}[T_{4l}] = 0$.

Combining the above results yields that $\mathbb{P}\{\varphi_{X_l}(O; \widehat{\mathbb{P}}) - \varphi_{X_l}(O; \mathbb{P})\} = \mathcal{O}_\mathbb{P}(e_n^2)$. Finally, we will show that $e_n \lesssim \varepsilon_n$. From (23), we have that

$$\|\widehat{\eta} - \eta\|_{L_2} \lesssim \|\widehat{\mu} - \mu\|_{L_2} + \|\widehat{T} - T\|_{L_2}.$$

Let $E := \widehat{L} - L$. Using the cyclic property of the trace,

$$\|\widehat{T} - T\|_{L_2(\mathbb{P}_X)}^2 = \mathbb{E}[\|EX\|_2^2] = \text{tr}(E\Sigma E^\top) = \text{tr}(E^\top E\Sigma).$$

For any two positive-semidefinite matrices $A, B \succeq 0$, $\text{tr}(AB) \leq \lambda_{\max}(A) \text{tr}(B)$. Applying this with $A = E^\top E$ and $B = \Sigma$ yields

$$\|\widehat{T} - T\|_{L_2(\mathbb{P}_X)}^2 \leq \lambda_{\max}(E^\top E) \text{tr}(\Sigma) = \|E\|_{\text{op}}^2 \text{tr}(\Sigma) = \text{tr}(\Sigma) \|\widehat{L} - L\|_{\text{op}}^2.$$

Furthermore, the column-wise error can also be bounded by:

$$\max_{l \in [d]} \|\widehat{L}_{\cdot l} - L_{\cdot l}\|_2 \leq \|\widehat{L} - L\|_{\text{op}}.$$

Thus, we further have $e_n \lesssim \|\widehat{L}_{\cdot l} - L_{\cdot l}\|_2 + \|L_{\cdot l}\|_2 [\|\widehat{\mu} - \mu\|_{L_2} + \text{tr}(\Sigma)^{1/2} \|\widehat{L} - L\|_{\text{op}}] \lesssim \varepsilon_n$, which completes the proof. $\square$

References

- Bonnotte, N. (2013). From knothe’s rearrangement to brenier’s optimal transport map. *SIAM Journal on Mathematical Analysis*, 45(1):64–87.
- Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417.
- Carlier, G., Galichon, A., and Santambrogio, F. (2010). From knothe’s transport to brenier’s map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576.
- Chamma, A., Engemann, D. A., and Thirion, B. (2023). Statistically valid variable importance assessment through conditional permutations. *Advances in Neural Information Processing Systems*, 36:67662–67685.
- Chamma, A., Thirion, B., and Engemann, D. (2024). Variable importance in high-dimensional settings requires grouping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38 (10), pages 11195–11203.
- Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. (2024). scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480.
- Dai, G., Shao, L., and Chen, J. (2024). Moving beyond population variable importance: concept, theory and applications of individual variable importance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae115.
- Du, J.-H., Cai, Z., and Roeder, K. (2022). Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scvaeit. *Proceedings of the National Academy of Sciences*, 119(49):e2214414119.
- Du, J.-H., Shen, M., Mathys, H., and Roeder, K. (2025a). Causal differential expression analysis under unmeasured confounders with causarray. *bioRxiv*, pages 2025–01.
- Du, J.-H., Zeng, Z., Kennedy, E. H., Wasserman, L., and Roeder, K. (2025b). Causal inference for genomic data with multiple heterogeneous outcomes. *Journal of the American Statistical Association*, pages 1–24.
- Genizi, A. (1993). Decomposition of R^2 in multiple regression with correlated regressors. *Statistica Sinica*, pages 407–420.
- Givens, C. R. and Shortt, R. M. (1984). A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240.
- Herbinger, J., Wright, M. N., Nagler, T., Bischl, B., and Casalicchio, G. (2024). Decomposing global feature effects based on feature interactions. *Journal of Machine Learning Research*, 25(381):1–65.

- Hines, O., Diaz-Ordaz, K., and Vansteelandt, S. (2022). Variable importance measures for heterogeneous causal effects. *arXiv preprint arXiv:2204.06030*.
- Hooker, G., Mentch, L., and Zhou, S. (2021). Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31:1–16.
- Kang, J. S., Butler, L., Agarwal, A., Erginbas, Y. E., Pedarsani, R., Ramchandran, K., and Yu, B. (2025). Spex: Scaling feature interaction explanations for llms. *arXiv preprint arXiv:2502.13870*.
- Lei, J., G’Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Lundborg, A. R., Kim, I., Shah, R. D., and Samworth, R. J. (2024). The projected covariance measure for assumption-lean variable significance testing. *The Annals of Statistics*, 52(6):2851–2878.
- Montefiori, D. C. (2009). Measuring hiv neutralization in a luciferase reporter gene assay. *HIV protocols*, pages 395–405.
- Moon, H., Du, J.-H., Lei, J., and Roeder, K. (2025). Augmented doubly robust post-imputation inference for proteomic data. *bioRxiv*, pages 2024–03.
- Owen, A. B. and Prieur, C. (2017). On shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002.
- Paillard, J., Lobo, A. R., Kolodyazhniy, V., Thirion, B., and Engemann, D. A. (2025). Measuring variable importance in heterogeneous treatment effects with confidence. *arXiv preprint arXiv:2408.13002*.
- Reyero Lobo, A., Neuvial, P., and Thirion, B. (2025). Sobol-cpi: a doubly robust conditional permutation importance statistic. *arXiv e-prints*, pages arXiv–2501.
- Rinaldo, A., Wasserman, L., and G’Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438–3469.
- Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472.
- Shapley, L. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307.
- Sobol’, I. M. (1990). On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2(1):112–118.

- Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, 9:1–11.
- Verdinelli, I. and Wasserman, L. (2024a). Decorrelated variable importance. *Journal of Machine Learning Research*, 25(7):1–27.
- Verdinelli, I. and Wasserman, L. (2024b). Feature importance: A closer look at shapley values and loco. *Statistical Science*, 39(4):623–636.
- Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.
- Williamson, B. and Feng, J. (2020). Efficient nonparametric statistical inference on population feature importance using shapley values. In *International conference on machine learning*, pages 10282–10291. PMLR.
- Williamson, B. D., Gilbert, P. B., Carone, M., and Simon, N. (2021). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22.
- Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. (2023). A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658.