Mathematical Theory of Collinearity Effects on Machine Learning Variable Importance Measures

Anonymous authors
Paper under double-blind review

Abstract

In many machine learning problems, understanding variable importance is a central concern. Two common approaches are Permute-and-Predict (PaP), which randomly permutes a feature in a validation set, and Leave-One-Covariate-Out (LOCO), which retrains models after permuting a training feature. Both methods deem a variable important if predictions with the original data substantially outperform those with permutations. In linear regression, empirical studies have linked PaP to regression coefficients and LOCO to t-statistics, but a formal theory has been lacking. We derive closed-form expressions for both measures, expressed using square-root transformations. PaP is shown to be proportional to the coefficient and predictor variability: $\mathrm{PaP}_i = |\hat{\beta}_i|\sqrt{2\,\mathrm{Var}(\mathbf{x}_i^v)}$, while LOCO is proportional to the coefficient but dampened by collinearity (captured by Δ): $\mathrm{LOCO}_i = |\hat{\beta}_i|(1-\Delta)\sqrt{1+c}$. These derivations explain why PaP is largely unaffected by multicollinearity, whereas LOCO is highly sensitive to it. Monte Carlo simulations confirm these findings across varying levels of collinearity. Although derived for linear regression, we also show that these results provide reasonable approximations for models like Random Forests. Overall, this work establishes a theoretical basis for two widely used importance measures, helping analysts understand how they are affected by the true coefficients, dimension, and covariance structure. This work bridges empirical evidence and theory, enhancing the interpretability and application of variable importance measures.

1 Introduction

In regression problems, and in supervised learning more broadly, understanding the contribution of different variables to a predictive model is a longstanding challenge. In many cases, assessing the relative importance of variables to a prediction can be as vital to the analysis as the accuracy of the predictions themselves. Consequently, methods for measuring variable importance have been employed across a wide range of disciplines, including social sciences (Blalock Jr, 1961; Darlington, 1968), medicine/bioinformatics (Strobl et al., 2008; Nicodemus, 2011; Goldstein et al., 2017), ecology (Cutler et al., 2007; Murray & Conner, 2009), and economics (Varian, 2014).

Some variable importance methods have become so commonplace in predictive modeling that their role is often taken for granted. In standard linear regression problems, when interactions and higher order effects are absent, beta coefficients and t-statistics provide a concise and largely comprehensive summary of the influence of individual variables. However in more complicated black box models, the role of a predictor is dispersed and no single model parameter can quantify its importance. Leo Breiman introduced a groundbreaking approach for assessing feature importance in Random Forests, commonly known as the Permute-and-Predict (PaP) technique, which compares model accuracy with and without permuting a variable (Breiman, 2001). This idea was later generalized into a model-agnostic framework known as model reliance, extending the same principle across a wide range of machine learning models (Fisher et al., 2019). As PaP gained widespread use, however, critics highlighted its potential limitations (Hooker et al., 2021; Gregorutti et al., 2017; Strobl et al., 2008), prompting the development of alternative methods. One such approach is Leave-One-Covariate-Out (LOCO), an idea with roots in classical regression that was later formalized for machine learning (Lei et al.,

2018; Molnar, 2025). LOCO involves dropping a variable, refitting, and comparing the resulting accuracy to that of the original model.

Although the idea of variable importance may seem straightforward, the term conceals fundamental ambiguity. Specifically, we must distinguish between the importance of a variable to a given predictive model and its importance to the prediction task itself (Chen et al., 2020). Consider a feature that is strongly correlated with the response and thus receives a large regression coefficient, but is also highly collinear with another predictive feature. From a predictive-sensitivity perspective, the variable is important: perturbing it substantially alters predictions in a detrimental way. From a model-necessity perspective, however, the variable is dispensable: removing it causes little loss in performance because the correlated predictor compensates for its removal. This ambiguity underlies an important difference in classical regression metrics. Beta coefficients emphasize predictive sensitivity, whereas t-statistics emphasize model necessity. The sensitivity of the t-statistic to collinearity is well established and reflects its reduced statistical power in multiple regression (Hastie et al., 2009; Draper & Smith, 1998; Freedman, 1983). The same distinction appears in machine learning metrics: PaP measures emphasize predictive sensitivity, while LOCO measures emphasize model necessity.

LOCO's sensitivity arises because it retrains the model after removing or permuting the variable of interest. Thus, regardless of its strength in the initial model, if a variable's predictive information is already captured by correlated predictors, then its removal has little effect on model errors, and LOCO deems it relatively unimportant. Through our derivations, we show that LOCO is approximately linearly (and inversely) related to the off-diagonal covariance structure of a variable transformation matrix. By contrast, PaP is unaffected by collinearity. Because the model is not retrained, permuting a variable inflates prediction error according to the magnitude of its coefficient, even when that variable is highly collinear with others. This distinction is not new. It is central to Chen et al. (2020), a fundamental component of Fisher et al. Fisher et al. (2019), and closely related to the categorizations in Gregorutti et al. (2017) and Strobl et al. (2008) of marginal versus conditional importance. Within that nomenclature, our example feature is marginally important but not conditionally important, as its relationship to the response disappears when conditioned on collinear variables.

One might ask why Strobl et al. (2007) observed inflated importance for correlated predictors in Random Forests. The answer lies in the mechanics of tree-based methods, which differ from regression. Consider two highly correlated variables and two independent variables. The regression model could be

$$\begin{aligned} y &= 2Cor_1 + 2Cor_2 + 2Ind_3 + 2Ind_4 + \epsilon \\ &\approx 4Cor_i + 2Ind_3 + 2Ind_4 + \epsilon \text{ where } i \in \{1, 2\}. \end{aligned}$$

In linear regression, this redundancy poses no problem as the effects of Cor_1 and Cor_2 are estimated jointly. In axis-aligned tree-based methods like Random Forests, however, splits are made greedily on a single variable at a time. Although random feature subsampling helps reduce this greediness, collinear predictors still gain elevated priority because their shared effect makes them strong candidates for early splits. This tendency to overstate importance also holds for strong negative correlations and coefficients. In contrast, when coefficients and correlations have opposite signs, tree-based methods understate the importance of correlated predictors (Bladen, 2022).

In this paper we do not provide any arguments supporting the preference of one notion of variable importance over another. Instead, our primary motivation is to provide a clearer understanding of the precise differences in PaP and LOCO, particularly regarding their distinct sensitivities to collinearity. Specifically, we aim to provide a mathematical rationale for previously observed empirical findings by Bladen & Cutler (2024): within a regression framework, PaP values exhibit a strong alignment with the regression coefficients, while LOCO values more closely resembled the t-statistics. To accomplish this, we introduce a highly general linear regression framework and show that, under certain assumptions, these empirical relationships necessarily follow from how each measure is calculated. We further support these derivations through a range of Monte Carlo simulations demonstrating their close alignment with empirical observation across a variety of conditions. Finally, although our derivations specifically quantify how PaP and LOCO operate within linear regression models, we conduct additional simulations to examine how well this theory generalizes to

other models, such as Random Forests. Overall, by clarifying which data characteristics drive these variable importance measures, our results offer a clearer lens for interpreting variable importance in applied settings.

2 Methods and Derivations

In this section, we provide derivations for the mathematical relationships between PaP, LOCO, t, and data attributes such as slope ($\boldsymbol{\beta}$), dimension (p), size (n), and collinearity (driven by Δ).

2.1 Latent Variable Framework

We begin by introducing a simple, but highly generalizable data structure that we can readily use in completing our derivations. Specifically, consider a traditional regression setup where the response variable $\mathbf{y}$ is expressed as a linear function of $\mathbf{X}$ and some error $\boldsymbol{\epsilon} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 = 0.1)$. To represent collinearity in the predictors, assume that $\mathbf{X}$ itself arises as a linear transformation of iid latent variables $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ via a correlation-inducing matrix $\mathbf{A}$, such that $\mathbf{X} = \mathbf{ZA}$. Substituting this structure into the regression model gives

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times p}{\mathbf{X}} \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}} = \underset{n\times p}{\mathbf{Z}} \underset{p\times p}{\mathbf{A}} \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}, \tag{1}$$

where

$$\underset{p\times p}{\mathbf{A}} = \begin{bmatrix} 1 & \Delta & \cdots & \Delta \\ \Delta & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Delta \\ \Delta & \cdots & \Delta & 1 \end{bmatrix}.$$

For subsequent derivations, we can use the following formulation for $\mathbf{A}$: $\mathbf{A} = \Delta\mathbf{J} + (1-\Delta)\mathbf{I}$, where $\mathbf{J}$ is a matrix of ones, $\mathbf{I}$ is the identity matrix, and Δ is a fixed real value in the range $[0, 1)$. Given $\mathbf{X} = \mathbf{ZA}$, $\mathbf{X^T X} = \mathbf{A^T Z^T Z A}$. Since $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{Z^T Z} \sim \mathcal{W}_p(n, \mathbf{I})$. This implies that the cross-product matrix satisfies $\mathbf{X^T X} \sim \mathcal{W}_p(n, \mathbf{A^T A})$, which has $E\left[\mathbf{X^T X}\right] = n\mathbf{A^T A}$. This means that since $E\left[\mathbf{X}\right] = 0$, the expected sample covariance reduces to

$$\begin{aligned} E\left[cov(\mathbf{X})\right] &= E\left[\frac{(\mathbf{X}-\bar{\mathbf{X}})^T(\mathbf{X}-\bar{\mathbf{X}})}{n}\right] = E\left[\frac{\mathbf{X^T X} - n\bar{\mathbf{X}}\bar{\mathbf{X}}^\mathbf{T}}{n}\right] = \frac{E\left[\mathbf{X^T X}\right] - E\left[n\bar{\mathbf{X}}\bar{\mathbf{X}}^\mathbf{T}\right]}{n} \\ &= \frac{n\mathbf{A^T A} - \mathbf{A^T A}}{n} = \frac{n-1}{n}\mathbf{A^T A} = \frac{n-1}{n}\left[(2\Delta + (p-2)\Delta^2)\mathbf{J} + (1-\Delta)^2\mathbf{I}\right]. \end{aligned} \tag{2}$$

This implies that the expected sample variance reduces to

$$E\left[var(\mathbf{X}_i)\right] = \frac{n-1}{n}(2\Delta + (p-2)\Delta^2 + (1-\Delta)^2) = \frac{n-1}{n}(1 + (p-1)\Delta^2). \tag{3}$$

2.2 Derivation of Variable Importance for PaP

We begin by deriving the theoretical value for PaP.

Definition 1 (PaP)**.** *The Permute-and-Predict variable importance measure can be defined as*

$$PaP_i = \sqrt{\mathscr{L}(\hat{f}(\mathbf{X}^v_{\pi(i)}), \mathbf{y}^v) - \mathscr{L}(\hat{f}(\mathbf{X}^v), \mathbf{y}^v)}, \tag{4}$$

where $\mathscr{L}$ is a selected loss function, $\hat{f}$ is a learned model, $\mathbf{X}^v$ is a validation predictor set, $\mathbf{y}^v$ is a validation response variable, $\pi(i)$ denotes a permutation applied to the ith variable in the data set, and the square

root function is a monotonic transformation on the original definition (Breiman, 2001; Fisher et al., 2019; Hooker et al., 2021; Bladen & Cutler, 2024).

In plain terms, we are calculating the PaP importance for a given variable by permuting it in the validation set, generating new predictions, and comparing this new loss to the original loss. While many versions of PaP compute simple loss differences or compare root losses, we instead apply a square-root transformation to the raw difference, a monotonic adjustment motivated by observations in Bladen & Cutler (2024).

Now consider the classical regression loss function MSE, which is minimized in linear regression and commonly used in the calculation of PaP. Here, $E[.]$ denotes expectation over the data-generating process, and for vector-valued quantities is understood to operate on the average squared error, i.e., $E\left[(\mathbf{y}-\hat{\mathbf{y}})^2\right] = E\left[\frac{1}{n}(\mathbf{y}-\hat{\mathbf{y}})^2\right]$.

$$\begin{aligned} MSE(\mathbf{y}^v, \hat{f}(\mathbf{X}^v)) &= E\left[(\mathbf{y}^v - \hat{\mathbf{y}}^v)^2\right] = E\left[(\mathbf{X}^v\boldsymbol{\beta} + \boldsymbol{\epsilon} - \mathbf{X}^v\hat{\boldsymbol{\beta}})^2\right] \\ &= E\left[(\mathbf{X}^v(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + \boldsymbol{\epsilon})^2\right] = E\left[(\mathbf{X}^v(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}))^2\right] + E\left[(\boldsymbol{\epsilon})^2\right] \\ &= E\left[(\mathbf{X}^v(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}))^2\right] + \sigma^2. \end{aligned} \tag{5}$$

Next, suppose we implement a permutation to convert $\mathbf{X}^v$ to $\mathbf{X}^v_{\pi(i)}$. In this case, the learned coefficients $\hat{\boldsymbol{\beta}}$ are unchanged. Because only the ith variable has been altered, $\hat{f}(\mathbf{X}^v_{\pi(i)})$ can be defined as

$$\hat{f}(\mathbf{X}^v_{\pi(i)}) = \mathbf{X}^v\hat{\boldsymbol{\beta}} - \hat{\beta}_i\mathbf{x}^v_i + \hat{\beta}_i\mathbf{x}^v_{\pi(i)},$$

where $\mathbf{x}^v_{\pi(i)}$ denotes a permutation of only the ith variable. This yields

$$\begin{aligned} MSE(\mathbf{y}^v, \hat{f}(\mathbf{X}^v_{\pi(i)})) &= E\left[(\mathbf{y}^v - \hat{f}(\mathbf{X}^v_{\pi(i)}))^2\right] \\ &= E\left[(\mathbf{X}^v\boldsymbol{\beta} + \boldsymbol{\epsilon} - (\mathbf{X}^v\hat{\boldsymbol{\beta}} - \hat{\beta}_i\mathbf{x}^v_i + \hat{\beta}_i\mathbf{x}^v_{\pi(i)}))^2\right] \\ &= E\left[(\mathbf{X}^v(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + \hat{\beta}_i\mathbf{x}^v_i - \hat{\beta}_i\mathbf{x}^v_{\pi(i)} + \boldsymbol{\epsilon})^2\right] \\ &= E\left[(\mathbf{X}^v(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}))^2\right] + E\left[(\hat{\beta}_i\mathbf{x}^v_i)^2\right] + E\left[(\hat{\beta}_i\mathbf{x}^v_{\pi(i)})^2\right] + E\left[(\boldsymbol{\epsilon})^2\right] \qquad \text{(cross terms } = 0) \\ &= E\left[(\mathbf{X}^v(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}))^2\right] + \hat{\beta}_i^2\left(E\left[(\mathbf{x}^v_i)^2\right] + E\left[(\mathbf{x}^v_{\pi(i)})^2\right]\right) + \sigma^2 \\ &= E\left[(\mathbf{X}^v(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}))^2\right] + 2\hat{\beta}_i^2 E[\mathrm{Var}(\mathbf{x}^v_i)] + \sigma^2. \end{aligned} \tag{6}$$

where $E\left[(\mathbf{x}^v_i)^2\right] = E[\mathrm{Var}(\mathbf{x}^v_i)]$ since $E[\mathbf{x}^v_i] = 0$. Expected values for cross terms involving $\mathbf{x}^v_{\pi(i)}$ vanish due to independence from all other variables, while cross terms involving $\boldsymbol{\epsilon}$ vanish due to its zero-mean assumption and independence from the predictors. The remaining cross terms involving $\mathbf{x}^v_i$ vanish by linearity of expectation and the centering of the predictors. Substituting Equation 5 and 6 into Equation 4 yields the closed-form solution in Theorem 1.

Theorem 1. *Under the latent variable framework in Equation 1, the PaP importance has the closed-form*

$$PaP_i = |\hat{\beta}_i|\sqrt{2E[\mathrm{Var}(\mathbf{x}^v_i)]} = |\hat{\beta}_i|\sqrt{2\frac{n-1}{n}(1+(p-1)\Delta^2)} \tag{7}$$

where $E[\mathrm{Var}(\mathbf{x}^v_i)]$ is provided in Equation 3. □

Thus, in the proposed regression framework, PaP_i depends only on the magnitude of the estimated coefficient $\hat{\beta}_i$ and the variance of the feature. Importantly, the degree of collinearity (defined by Δ) has no direct effect on PaP_i aside from its influence on $\mathrm{Var}(\mathbf{x}^v_i)$. Consequently, PaP_i is independent of collinearity as long as there are no co-occurring effects on variable scale. This independence can also be guaranteed when variables

are standardized (S.T. $\text{Var}(\mathbf{x}_i^v) = 1, \text{for all } i$) (Hooker et al., 2021). We also note that a Drop-and-Predict technique admits an essentially identical theoretical formulation to PaP, differing only by a multiplicative factor of 2 (Bladen & Cutler, 2024).

2.3 Derivation of Variable Importance for LOCO

Similar to the expression for PaP, the leave-one-covariate-out variable importance measure can be expressed in terms of loss functions.

Definition 2 (*LOCO*)**.** *The leave-one-covariate-out variable importance measure can be defined as*

$$LOCO_i = \sqrt{\mathscr{L}(\hat{f}^t_{\pi(\mathbf{i})}(\mathbf{X}^v), \mathbf{y}^v) - \mathscr{L}(\hat{f}(\mathbf{X}^v), \mathbf{y}^v)}, \tag{8}$$

where the new notation ($\hat{f}^t_{\pi(\mathbf{i})}$) represents a model retrained on the training data with the ith variable permuted, and the square root function again is a monotonic transformation on the original definition: $\mathbf{X}^t_{\pi(i)}$ (Lei et al., 2018; Hooker et al., 2021; Bladen & Cutler, 2024).

We note that the original LOCO measure was defined by retraining the model after removing the ith variable from the training data (Lei et al., 2018). In contrast, we permute the variable within the training set prior to refitting. This modification preserves the dimensionality of the design matrix and therefore yields coefficient vectors $\hat{\beta}$ and $\hat{\beta}_{\pi(i)}$ that lie in the same parameter space. Maintaining equal dimensionality is essential for directly analyzing how the estimated coefficients redistribute under collinearity. Moreover, using permutation rather than deletion aligns the LOCO construction with PaP, providing a unified perturbation framework. This standardization facilitates principled comparison between LOCO and PaP while preserving interpretability of coefficient-level effects.

In other words, we calculate the LOCO importance for a given variable by permuting it in the training set, refitting the model, generating new validation predictions, and comparing this new loss to the original loss. We again take the square root to place this on a scale familiar to regression metrics (Bladen & Cutler, 2024).

When a training variable is permuted and a model is refit, the result is a different set of coefficients $\hat{f}^t_{\pi(\mathbf{i})} = \hat{\boldsymbol{\beta}}_{\pi(\mathbf{i})}$. Under the latent construction $\mathbf{X} = \mathbf{ZA}$, the expected Gram matrix satisfies $E\left[\mathbf{X}^T\mathbf{X}\right] = n[(2\Delta + (p-2)\Delta^2)\mathbf{J} + (1-\Delta)^2\mathbf{I}]$ which has compound-symmetric form $a\mathbf{I} + b\mathbf{J}$ (Muirhead, 1982). Matrices of this form are invariant under simultaneous permutations of rows and columns, and their inverses retain the same structure (Searle, 1982). Consequently, predictors are exchangeable in expectation.

This exchangeability of predictors implies that, in expectation, the induced change in coefficients is symmetric across all predictors $j \neq i$. Accordingly, we model the expected change in coefficients as

$$E[\delta_j^{(i)}] = \begin{cases} -\hat{\beta}_i, & j = i \\ \gamma_i, & j \neq i \end{cases} \tag{9}$$

where γ_i is constant across $j \neq i$ and represents the average redistributed effect assigned to each predictor.

Motivated by this symmetry, we approximate this expected symmetric behavior with a deterministic redistribution of the permuted coefficient across the remaining predictors. Specifically, we assume the redistributed effect is proportional to the original coefficient magnitude $\hat{\beta}_i$, and set $\gamma_i = \hat{\beta}_i * c$.

$$\hat{\boldsymbol{\beta}}_{\pi(\mathbf{i})} = \begin{bmatrix} \hat{\beta}'_1 \\ \vdots \\ \hat{\beta}'_{i-1} \\ \hat{\beta}'_i = 0 \\ \hat{\beta}'_{i+1} \\ \vdots \\ \hat{\beta}'_p \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 + \hat{\beta}_i * c \\ \vdots \\ \hat{\beta}_{i-1} + \hat{\beta}_i * c \\ \hat{\beta}_i - \hat{\beta}_i \\ \hat{\beta}_{i+1} + \hat{\beta}_i * c \\ \vdots \\ \hat{\beta}_p + \hat{\beta}_i * c \end{bmatrix} = \hat{\boldsymbol{\beta}} + \boldsymbol{\delta}^{(i)}, \tag{10}$$

where c is a coefficient controlling how much of the original magnitude is absorbed (hereafter called the coefficient of absorption) and $\boldsymbol{\delta}^{(i)} = \left[\hat{\beta}_i * c, \ldots, \hat{\beta}_i * c, -\hat{\beta}_i, \hat{\beta}_i * c, \ldots, \hat{\beta}_i * c\right]^T$. Note that the symmetric redistribution of the permuted coefficient ($\gamma_i = \hat{\beta}_i * c$) is supported by the empirical results in Section 3.2.1 and Appendix D. Also note that the coefficient of absorption, which we analyze in greater detail later on, is heavily dependent on the degree of collinearity in the data.

Now we use the results of Equation 10 to derive the new loss function:

$$\begin{aligned}
MSE(\mathbf{y}^v, \hat{f}^t_{\pi(i)}(\mathbf{X}^v)) &= E\left[(\mathbf{y}^v - \hat{f}^t_{\pi(i)}(\mathbf{X}^v))^2\right] \\
&= E\left[(\mathbf{X}^v\boldsymbol{\beta} + \boldsymbol{\epsilon} - \mathbf{X}^v(\hat{\boldsymbol{\beta}} + \boldsymbol{\delta}^{(i)}))^2\right] \\
&= E\left[(\mathbf{X}^v(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) - \mathbf{X}^v\boldsymbol{\delta}^{(i)} + \boldsymbol{\epsilon})^2\right] \\
&= E\left[(\mathbf{X}^v(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}))^2\right] + E\left[(\mathbf{X}^v\boldsymbol{\delta}^{(i)})^2\right] + E\left[(\boldsymbol{\epsilon})^2\right] \qquad \text{(cross terms } = 0) \\
&= E\left[(\mathbf{X}^v(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}))^2\right] + E\left[(\mathbf{Z}^v\mathbf{A}\boldsymbol{\delta}^{(i)})^2\right] + \sigma^2 \\
&= E\left[(\mathbf{X}^v(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}))^2\right] + E\left[(\boldsymbol{\delta}^{(i)})^T\mathbf{A}^T\mathbf{Z}^{vT}\mathbf{Z}^v\mathbf{A}\boldsymbol{\delta}^{(i)}\right] + \sigma^2 \\
&= E\left[(\mathbf{X}^v(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}))^2\right] + (\boldsymbol{\delta}^{(i)})^T\mathbf{A}^T\mathbf{A}\boldsymbol{\delta}^{(i)} + \sigma^2.
\end{aligned} \tag{11}$$

Since $\boldsymbol{\epsilon}$ and $\mathbf{X}^v$ are zero-mean and independent of each other, their means and cross terms vanish. Additionally, because the coefficient perturbation $\boldsymbol{\delta}^{(i)}$ is fixed conditional on the training sample, the cross term has expectation zero.

Substituting Equation 2 for $\mathbf{A}^T\mathbf{A}$ in Equation 11 gives

$$\begin{aligned}
MSE(\hat{f}^t_{\pi(i)}(\mathbf{X}^v), \mathbf{y}^v) =& E\left[(\mathbf{X}^v(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}))^2\right] + \sigma^2 \\
&+ \boldsymbol{\delta}^{(i)T}\left((2\Delta + (p-2)\Delta^2)\mathbf{J} + (1-\Delta)^2\mathbf{I}\right)\boldsymbol{\delta}^{(i)} \\
=& E\left[(\mathbf{X}^v(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}))^2\right] + \sigma^2 \\
&+ \hat{\beta}_i^2\Big[(2\Delta + (p-2)\Delta^2)((p-1)c-1)^2 + (1-\Delta)^2(1+(p-1)c^2)\Big].
\end{aligned} \tag{12}$$

Substituting the results of Equation 5 and 12 into Equation 8 yields a closed-form solution for LOCO (stated in Theorem 2).

Theorem 2. *Under the framework in Equation 1, the LOCO importance has the closed-form*

$$LOCO_i = |\hat{\beta}_i|\sqrt{(1-\Delta)^2(1+(p-1)c^2) + (2\Delta + (p-2)\Delta^2)((p-1)c-1)^2}. \quad \square \tag{13}$$

While c was initially unknown, being learned from model metrics rather than a fixed parameter, it follows that theoretical c should be a function of Δ and p.

Conjecture 3 (c)**.** *Based on empirical observations, we postulate that c takes the following form:*

$$c(\Delta, p) = \frac{2\Delta + (p-2)\Delta^2}{1 - 4\Delta + 2p\Delta + p^2\Delta^2 - 3p\Delta^2 + 3\Delta^2}. \tag{14}$$

Note that Conjecture 3 is formulated by assuming the relationship subsequently shown in Theorem 6 and then deriving the exact c value that yields this relationship. However, we will show that (1) empirical c values fit this conjecture very closely (see Section 3.2.1), and (2) if this conjecture is true, then Theorem 4 and 6 necessarily follow.

Equation 14 can be easily evaluated at the boundary points of Δ. If $\Delta = 0$, then all variables in $\mathbf{X}$ are independent and $c = 0$, meaning there is no absorption of the permuted coefficient. In this case, substituting

$\Delta = 0$ and $c = 0$ into Equation 13 yields $\text{LOCO}_i = |\hat{\beta}_i|$. If $\Delta = 1$, then all variables in $\mathbf{X}$ are perfectly collinear and $c = \frac{1}{p-1}$. In this situation, substituting these into Equation 13 yields $\text{LOCO}_i = 0$.

By substituting Equation 14 into Equation 13 and simplifying (see Appendix A), we obtain the following representation of LOCO.

Theorem 4. *Given Equation 13, LOCO simplifies via Conjecture 3 to*

$$LOCO_i = |\hat{\beta}_i|(1-\Delta)\sqrt{\frac{(1+(p-1)\Delta)^2}{(1+(p-2)\Delta)^2+(p-1)\Delta^2}} = |\hat{\beta}_i|(1-\Delta)\sqrt{1+c}. \quad \square \tag{15}$$

A very natural approximation for this importance is

$$\text{LOCO}_i \approx |\hat{\beta}_i|(1-\Delta), \tag{16}$$

since $1 < \sqrt{1+c} < \frac{p}{p-1}$. Put simply, these formulations show that collinearity systematically shrinks LOCO, in contrast to PaP, which depends only on predictor variability.

While the theoretical value of LOCO is well defined in Equation 15, Bladen & Cutler (2024) showed that it has a close connection to regression t-statistics.

Definition 3 (t-statistic)**.** *The t-statistic for coefficient $\hat{\beta}_i$ is defined as*

$$t_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\sqrt{\text{Var}(\epsilon)(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}_{ii}}}. \tag{17}$$

Theorem 5 (t-statistic as function of Δ)**.** *Under the framework in Equation 1, the t-statistic of Equation 17 can be expressed as*

$$t_i = \hat{\beta}_i(1-\Delta)\sqrt{\frac{(n-p-1)(1+(p-1)\Delta)^2}{\text{Var}(\epsilon)[(1+(p-2)\Delta)^2+(p-1)\Delta^2]}}. \quad \square \tag{18}$$

A complete mathematical proof deriving Theorem 5 from Equation 17 is available in Appendix B.

Theorem 6. *Given Equation 15, Equation 18, and Conjecture 3, the precise relationship between the t-statistic and LOCO can be mathematically quantified as*

$$t_i = \pm LOCO_i\sqrt{\frac{n-1}{\text{Var}(\epsilon)}}. \quad \square \tag{19}$$

2.4 Simulation Design

We now offer a system of simulations to show the correctness and relevance of our resulting derivations and to validate our assumptions. The general setup aligns with the latent variable framework introduced in subsection 2.1.

2.4.1 Data Generation

We simulate $\mathbf{X}$ as a set of p predictor variables with n observations. These are constructed by applying the linear transformation $\mathbf{A} = \Delta\mathbf{J} + (1-\Delta)\mathbf{I}$ to a matrix $\mathbf{Z}$ of independent standard normal variables, where $\mathbf{J}$ is the $p \times p$ matrix of ones and $\Delta \in [0, 0.99)$ controls the correlation structure. The response variable $\mathbf{y}$ is defined as a linear function of $\mathbf{X}$ with identical coefficients $\boldsymbol{\beta}$ and additive noise $\epsilon \overset{\text{iid}}{\sim} \mathcal{N}(\text{Mean} = 0, \text{Var} = 0.1)$.

We vary $\Delta \in \{0, 0.11, 0.22, \ldots, 0.88, 0.99\}$, $p \in \{3, 6, 9, 12\}$, and $n \in \{20, 63, 200, 632, 2000\}$ systematically. For each parameter combination, we generate 100 Monte Carlo iterations using identically structured training and validation sets.

2.4.2 Linear Model Implementation

For each simulation, we first fit a linear regression model to the training set and evaluate the metrics PaP_i, LOCO_i, t_i, and c on the validation set. These empirically observed quantities are then compared to their corresponding theoretical or assumed expressions in Equation 7, 15, and 14 to assess the validity of our derivations under the assumed linear data-generating process. Comparisons are made by computing the relative difference: $\frac{\text{empirical}-\text{theory}}{\text{theory}}$. Functional relationships are also assessed by plotting both the empirical and theoretical trends across Δ.

2.4.3 Random Forest Implementation

To assess the robustness of our results under model misspecification and nonlinearity, we repeat the full set of simulations using a Random Forest model in place of the linear predictor. The data generation and evaluation settings remain identical. For each simulation, we fit a Random Forest on the training set and compute the same set of metrics on the validation set. We fix the number of candidate variables at each split to $mtry = \max\left(2,\ \text{floor}(\frac{p}{3})\right)$ to maintain consistency across values of p. This nonparametric modeling approach allows us to test whether the theoretical insights derived under linear assumptions retain utility in more flexible model classes.

Comparisons are again made by computing the relative difference for PaP ($\frac{\text{empirical}-\text{theory}}{\text{theory}}$) or the absolute difference for LOCO (empirical $-$ theory) to accommodate the noise associated with Random Forest importances and difficulty of performing a relative comparison for small theoretical values. Functional relationships are also assessed by plotting the empirical and theoretical trends across Δ.

3 Results

We employ `R` version 4.4.0 (R Core Team, 2025) and the `tidyverse` (Wickham et al., 2019) to carry out the simulations defined in subsection 2.4. Random Forest analyses are implemented using the `randomForest` package (Liaw & Wiener, 2002). For visualization, we extract a meaningful subset of our results and plot comparisons using `ggplot2` (Wickham, 2016) together with `ggh4x` (van den Brand, 2025). Plots for the full simulation range are in the supplement in Section 1.

In subsection 3.1, we provide plots comparing the means and spreads of empirical linear regression metrics with the theoretical values derived in Equation 7 and 15 (see Appendix C for empirically testing Equation 19). We also offer visualizations of the functional relationship between these metrics and Δ. These visualizations are also applied to show how empirical Random Forest metrics compare to the theoretical values and to validate our assumptions for c in Equation 14.

3.1 Linear Model Importance Results

We begin by conducting parity comparisons for importance in linear models. The initial question tested by the simulation is how well the derived theoretical expressions match the empirical results for both LOCO and PaP as a function of Δ. Figure 1 illustrates the average empirical values computed for each measure with 2SE error bars as a function of Δ for sample values of n and p. Recall from Equation 7 that $\text{PaP}_i = |\hat{\beta}_i|\sqrt{2\,\text{Var}(\mathbf{x}_i^v)}$, and for our simulations, $\text{Var}(\mathbf{x}_i^v) = 1 + (p-1)\Delta^2$. As a result, both the theoretical and empirical values of PaP increase with Δ under this covariance structure found in Figure 1. In contrast, LOCO decreases nearly linearly with increasing Δ, closely matching the relationship identified by Equation 15.

3.2 Random Forest Importance Results

Figure 2 illustrates how the distribution of empirical Random Forest importance values, the theoretical expectations, and the relationship between them behave as functions of Δ. Theoretical values are again derived from Equation 7 and 15. When compared to the near-perfect agreement between empirical and theoretical values for the linear model in Figure 1 (aside from some small-sample error), the Random Forest results in Figure 2 are notably less consistent.

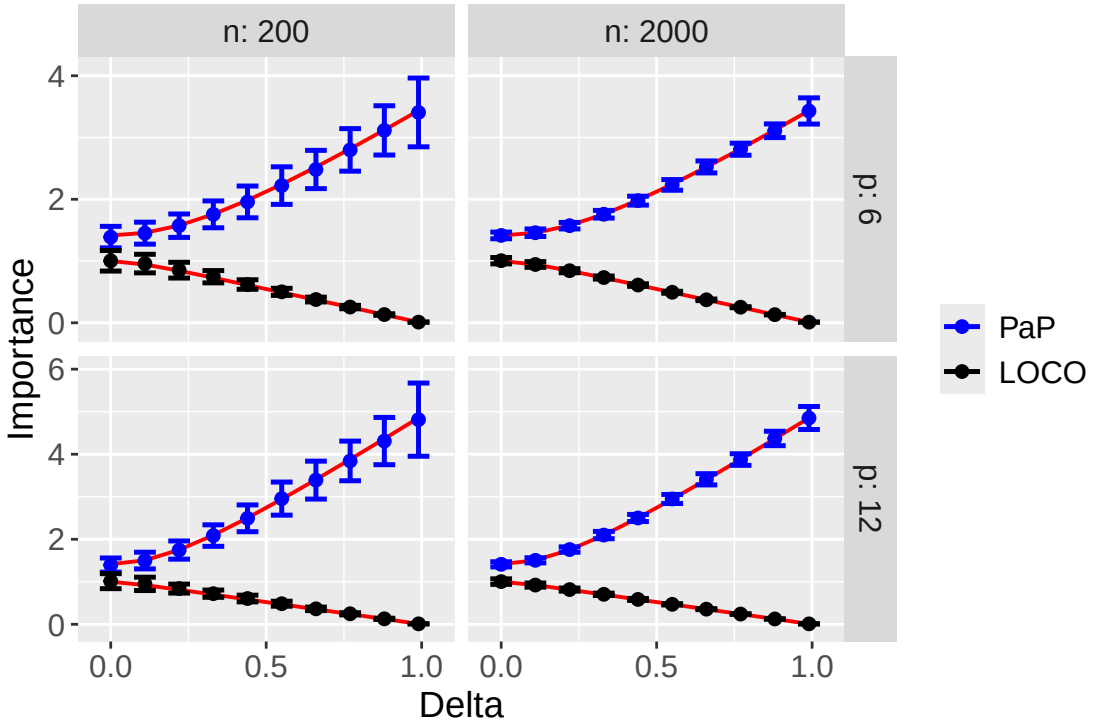

Figure 1: Mean and 2 Standard deviation error bars for assessing the relationship between importance values and Δ (see Equation 7 and 15) for data defined in subsection 2.4.1. Red lines denote theoretical projections. Plots show importance values for raw data and are faceted by p and n dimensions. Plots show that PaP increases with Δ, while LOCO has a nearly linear negative relationship with Δ (See Equation 16).

Empirical patterns for both PaP and LOCO generally align with the theoretical curves, though some systematic deviations emerge. For PaP, most differences are reasonably modest, except at $\Delta = 0$, where empirical values fall notably below the theoretical expectation. A similar anomaly is observed for LOCO at this point. Furthermore, LOCO exhibits a more consistent bias pattern: empirical values tend to lie below the theory for small Δ and above it for larger Δ. Overall, despite some of these deviations, the trends show that both PaP and LOCO align closely with our theoretical predictions, particularly for larger n. This is indicative that the collinearity effects quantified in our framework extend and provide meaningful insight into variable importance for Random Forests and likely other complex machine learning regression models.

3.2.1 c Results

Figure 3 illustrates the relationship between Δ, p, and the functional value of c in a linear regression model. It also visualizes the relative differences between the empirical distribution and the proposed value of c, postulated in Conjecture 3. As shown, c increases monotonically with Δ, though the increase is clearly sublinear and converges toward an upper bound of $\frac{1}{p-1}$. This limit is highly intuitive: when one variable is removed or permuted, the remaining $p-1$ variables absorb the effect, and if that influence is distributed uniformly, each receives a share of $\frac{1}{p-1}$. Figure 3 shows an intuitive decrease in variance as n increases and a more surprising decrease in variance with increasing p and Δ. Importantly, no observable bias is present in the plot, visually suggesting that Equation 14 provides an unbiased estimate of c and is therefore highly relevant for the LOCO derivations presented in subsection 2.3.

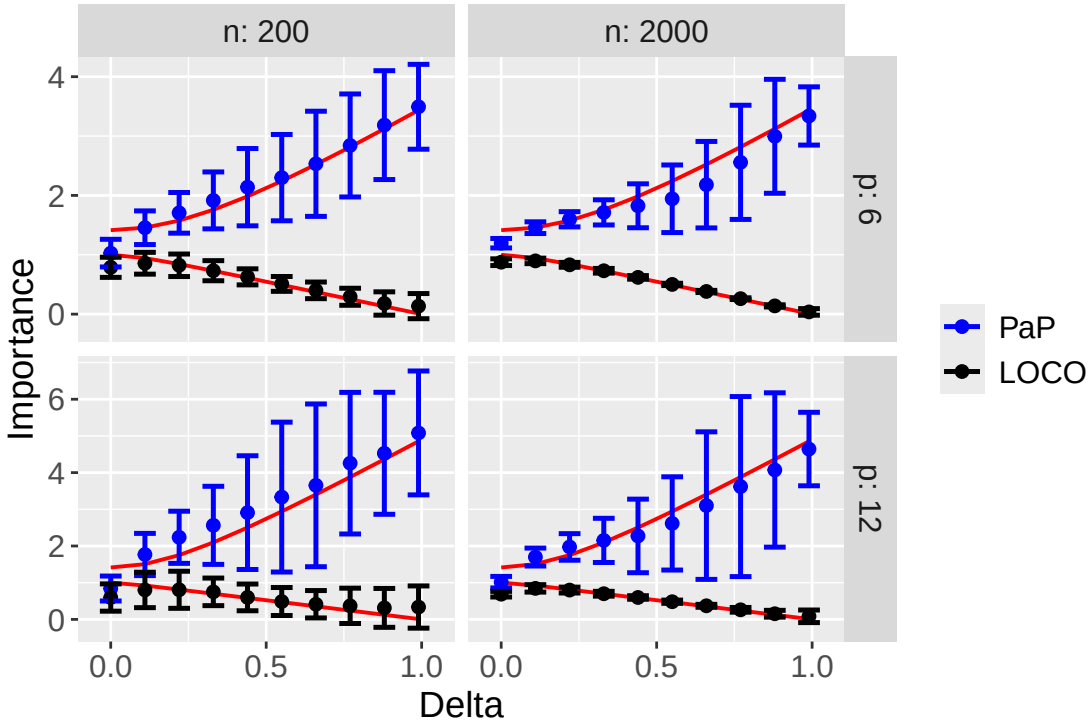

Figure 2: Mean and 2 Standard deviation error bars for assessing the relationship between Random Forest importance values and Δ for data defined in subsection 2.4.1. Red lines denote theoretical projections for linear data. Plots show importance values for raw data and are faceted by p and n dimensions. Plot trends appear similar to Figure 1. However, there are clear anomalies regarding bias within the plot, especially for $\Delta = 0$.

3.2.2 Small Sample Bias

Figure 4 illustrates the relative differences between empirical and theoretical importance values for the linear regression model. As expected, the variance decreases with increasing n and decreasing p. A subtle small-sample bias can be observed for the raw PaP metrics, where theoretical values slightly exceed empirical estimates. Meanwhile, raw LOCO values exhibit a clear upward bias, with empirical values exceeding theory—most notably when n is small and p is large.

To address these small-sample biases, we apply a sample size correction to our evaluation metrics. For the PaP metric, we adopt Bessel's correction — a widely used adjustment in statistical estimation — most notably in the unbiased estimation of sample variance (Kenney & Keeping, 1951; Stigler, 1973; Reichel, 2025). This correction typically takes the form $\frac{n}{n-1}$, or its inverse $\frac{n-1}{n}$, depending on the formulation. Incorporating this adjustment into Equation 7 yields:

$$\text{PaP}_i = |\hat{\beta}_i| \sqrt{2\,\text{Var}(\mathbf{x}_i^v)} \times \frac{n-1}{n}. \tag{20}$$

LOCO is addressed with a different, yet similarly well-established, correction drawn from regression modeling. This finite-sample bias correction factor, often referred to as a degrees-of-freedom adjustment, takes the form $\sqrt{\frac{n}{n-p}}$ or its inverse $\sqrt{\frac{n-p}{n}}$, depending on the desired scale and interpretation (Efron, 1986; Shao, 1993). This adjustment especially compensates for overfitting and bias that can occur when the number of predictors p

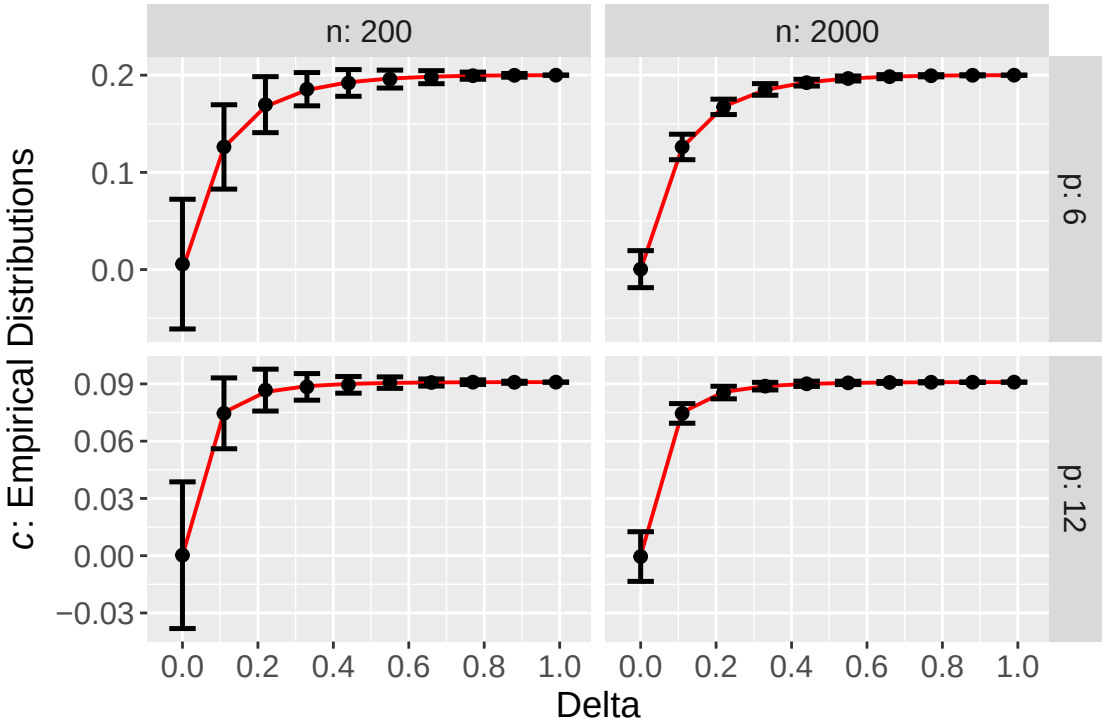

Figure 3: Mean and 2 Standard deviation error bars for assessing the relationship between c values and Δ (see Conjecture 3) for data defined in subsection 2.4.1. Red lines denote theoretical projections. Plots are faceted by n and p dimensions. Plots show increased stability as n increases, Δ increases, and p decreases. Plots also show sublinear growth converging toward values of $\frac{1}{p-1}$. Plots are free of any clear bias.

approaches the sample size n. Incorporating this term into Equation 15 yields:

$$\text{LOCO}_i = |\hat{\beta}_i|(1-\Delta)\sqrt{1+c} \times \sqrt{\frac{n}{n-p}}. \tag{21}$$

Figure 4 shows that implementing these adjustments retains the general variance patterns of the raw metrics. However, small-sample biases are effectively eliminated, so that empirical distributions are centered closely around their theoretical counterparts.

4 Discussion and Conclusion

Our results provide new theoretical insights into two widely used variable importance measures: Permute-and-Predict and Leave-One-Covariate-Out. By deriving formal mathematical expressions for each, we show how they are precisely linked to closed-form regression metrics and data characteristics of dimension (p), collinearity (Δ), and effect size ($\hat{\beta}_i$). This provides a more grounded understanding of what these measures reveal about the structure of the data itself, especially where collinearity is concerned. Empirical simulations show that the proposed framework provides a very precise model for how these variable importance measures react to collinearity in linear regression and generalizes very well to other models such as Random Forests. Placed in the context of prior work, our results confirm earlier assessments of importance methods by Hooker et al. (2021); Chen et al. (2020), extend prior comparisons to regression metrics by Bladen & Cutler (2024), and solidify the theoretical foundations for interpreting variable importance in relation to data characteristics.

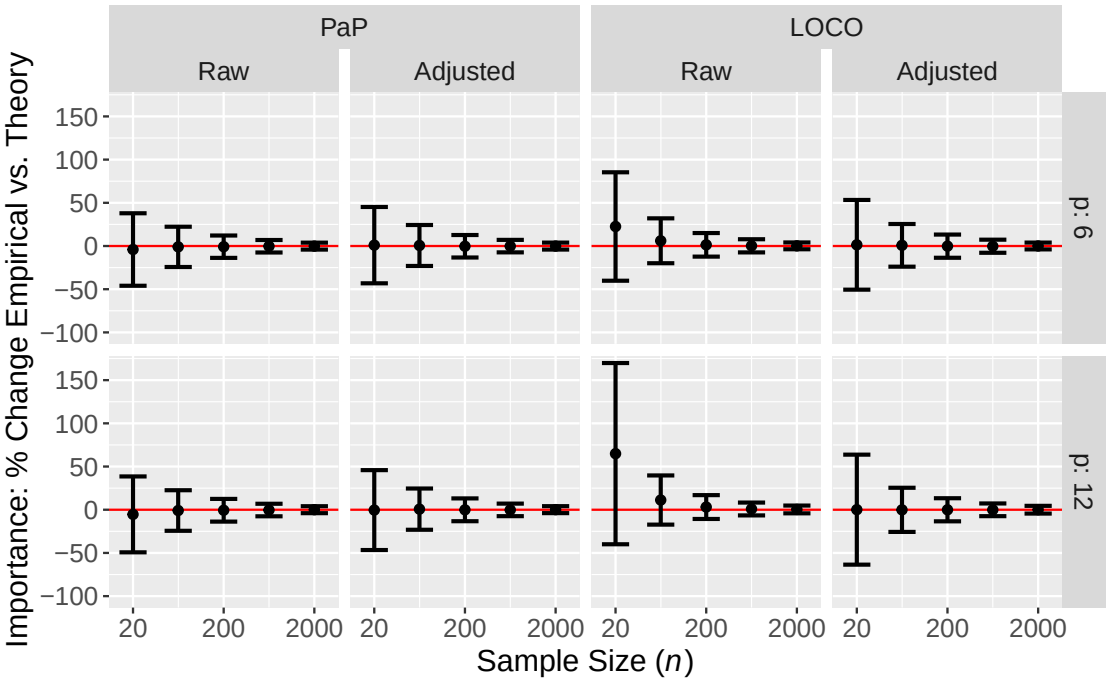

Figure 4: Mean and 2 Standard deviation error bars for comparing empirical importance values to the derived theoretical importances (see Equation 7 and 15) for data defined in subsection 2.4.1. Plots show sample size (n) on the x-axis and are faceted by dimensions (p), importance metric, and whether a sample sample adjustment was applied to the theoretical values. Plots show increased stability as n increases and as p decreases. Plots also show a small sample bias for both raw metrics, but particularly for LOCO. This bias is especially noticeable for small n and large p combinations. Conversely, the adjusted metrics are free of any clear visual bias.

The conceptual interpretation of the derived relationships can be summarized as follows. If a variable captures information about the response that could also be represented by other predictors, the PaP measure assigns it importance regardless of how little unique information it contributes. In contrast, LOCO assigns little importance to such variables, since retraining allows that information to be absorbed by other predictors. In this sense, PaP conceptually resembles a marginal association measure, since it attributes importance to predictors even when their information is redundant with other variables. Conversely, LOCO behaves more like a conditional or partial association measure, since importance is retained only to the extent that a predictor contributes information not recoverable from the remaining covariates. Consequently, PaP is particularly effective when meaningful correlated predictors jointly encode relevant information, whereas LOCO is better suited for identifying predictors that explain the response independently of the others.

In the framework of this paper, the degree of shared information is captured by collinearity, and both PaP and LOCO can be expressed precisely in terms of the magnitude of coefficient $\hat{\beta}_i$ and, for LOCO, the degree of collinearity (Δ). In more general settings where variables may share nonlinear mutual information, these derivations are not directly applicable. Nevertheless, nonlinear relationships among predictors would likely produce similar effects on these measures, provided that the regression model has sufficient capacity to map the different representations of information to the response.

4.1 Limitations and Considerations

Despite these contributions, several limitations should be acknowledged. Our derivations are framed in standard regression settings and, while they provide valuable and legitimate insights, theoretical extensions

to more complex or nonlinear data structures remain an important area for future work; see Appendix E for preliminary empirical exploration of these trends. Note that while we tested these derivations in the context of Random Forests, the experiments still assumed the underlying linear relationships and collinearity structure defined in subsection 2.1. Moreover, while Random Forests are an illustrative test case to explore these derivations for non-parametric models, further evaluation in additional contexts is warranted before generalizing broader conclusions.

Additionally, the finite-sample corrections considered are grounded in classical regression theory, so their applicability to high-dimensional data or strongly regularized models requires further investigation. Our empirical analysis also has some limitations as it is necessarily constrained by the chosen simulation designs, which reflect particular distributions and dimensionalities. While no single study can capture every possible condition, these results establish a solid foundation that future work can build on to explore broader settings and applications. We are optimistic that these contributions will not only clarify existing methods but also inspire new directions in the development of interpretable machine learning tools and theories.

4.2 Extensions and Applications

Although our analysis was carried out in an idealized regression framework, the results naturally extend to other common data structures. One clear example involves K independent blocks of correlated variables. When applying our theory to such cases, the effective dimensionality (p_k) corresponds to the correlated block which a variable belongs to. We note that prior research uses a dataset where one such block contains variables that are independent of each other as well as the rest of the blocks (Strobl et al., 2008). In this case, our theoretical results apply to each block, with the independent block characterized by $\Delta_{i,j} = 0$ where i or j is in the indices of block K. These extensions follow directly from the logic of our derivations and illustrate the generality of the framework. Time-series data structures represent a more challenging setting. While the exact results presented here would not hold, the general trends remain informative: both PaP and LOCO are shaped by the degree of collinearity in the data, and so the insights provided here should still serve as a reasonable approximation. Exploring this connection more formally offers an interesting avenue for future work.

Additionally, extending these derivations to alternative importance frameworks (e.g., Global Shap values (Lundberg & Lee, 2017), Partial Dependence Plot importances (Greenwell et al., 2018)) could provide further clarity and help unify interpretability methods across modeling paradigms (Verdinelli & Wasserman, 2024). More extensive simulation studies and applications to real-world datasets would further strengthen the validation of the theoretical derivations of PaP and LOCO, particularly in modern machine learning techniques. Beyond theoretical contributions, these results offer practical guidance by enabling importance measures to be understood and interpreted with greater confidence. Taken together, this work provides theoretical derivation bridged with empirical validation, laying a strong foundation for more principled and reliable use of variable importance in applied machine learning.

Code Access

The code used to generate the data and figures in this article is available, and we will include a link after acceptance of the manuscript.

References

Kelvyn K Bladen. *Contributions to Random Forest Variable Importance with Applications in R*. All Graduate Theses and Dissertations. 8587, 2022.

Kelvyn K Bladen and D Richard Cutler. Assessing agreement between permutation and dropout variable importance methods for regression and random forest models. *Electronic Research Archive*, 32(7):4495–4514, 2024.

Hubert M Blalock Jr. Evaluating the relative importance of variables. *American Sociological Review*, pp. 866–874, 1961.

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

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

D Richard Cutler, Thomas C Edwards Jr, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson, and Joshua J Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.

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

Norman R. Draper and Harry Smith. *Applied Regression Analysis*. Wiley, New York, 3rd edition, 1998. ISBN 978-0-471-17082-2.

Bradley Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470, 1986.

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

David A. Freedman. A note on screening regression equations. *The American Statistician*, 37(2):152–155, 1983. doi: 10.1080/00031305.1983.10482727.

Benjamin A. Goldstein, Ann Marie Navar, Michael J. Pencina, and John P.A. Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health record data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208, 2017. doi: 10.1093/jamia/ocw042.

Brandon M Greenwell, Bradley C Boehmke, and Andrew J McCarthy. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*, 2018.

Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, 27:659–678, 2017. doi: 10.1007/s11222-016-9646-1.

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–318, 1986.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, 2 edition, 2009. doi: 10.1007/978-0-387-84858-7. See Section 3.2.2, pp. 50–52 for discussion of collinearity and variance inflation.

Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31:1–16, 2021.

J. F. Kenney and E. S. Keeping. *Mathematics of Statistics, Part One*. Van Nostrand, Princeton, NJ, 2nd edition, 1951.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

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

Christoph Molnar. *Interpretable Machine Learning*. Molnar, Christoph, 3 edition, 2025. ISBN 978-3-911578-03-5. See Chapter 24 on Leave-One-Feature-Out (LOFO) importance.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory.* John Wiley & Sons, New York, 1982. ISBN 9780471094426.

Kim Murray and Mary M Conner. Methods to quantify variable importance: implications for the analysis of noisy ecological data. *Ecology*, 90(2):348–355, 2009.

Kristin K Nicodemus. On the stability and ranking of predictors from random forest variable importance measures. *Briefings in bioinformatics*, 12(4):369–373, 2011.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2025. URL `https://www.R-project.org/`.

Felix Reichel. On bessel's correction: Unbiased sample variance, the bariance, and a novel runtime-optimized estimator. *arXiv preprint arXiv:2503.22333*, 2025.

Shayle R. Searle. *Matrix Algebra Useful for Statistics.* John Wiley & Sons, New York, 1982. ISBN 9780471866818.

Jun Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88 (422):486–494, 1993.

Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix. *Annals of Mathematical Statistics*, 20 (4):621–622, 1949. doi: 10.1214/aoms/1177729959.

Stephen M. Stigler. The idea of the normal distribution. *American Statistician*, 27(1):34–38, 1973.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):1–11, 2008. URL `https://doi.org/10.1186/1471-2105-9-307`.

Teun van den Brand. *ggh4x: Hacks for 'ggplot2'*, 2025. URL `https://CRAN.R-project.org/package=ggh4x`. R package version 0.3.4.

Hal R. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28, 2014. doi: 10.1257/jep.28.2.3.

Isabella Verdinelli and Larry Wasserman. Feature importance: A closer look at shapley values and loco. *Statistical Science*, 39(4):623–636, 2024.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org/`.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

A Derivation of LOCO Variable Importance

We now offer a mathematical proof for the simplified version of LOCO shown in Theorem 4.

Proof. To prove Theorem 4 by substituting the assumed value of c from Equation 14 into Equation 13, we begin by recalling and restating their mathematical expressions:

$$\text{LOCO}_i = |\hat{\beta}_i|\sqrt{(1-\Delta)^2(1+(p-1)c^2) + (2\Delta + (p-2)\Delta^2)((p-1)c-1)^2} \qquad \text{(Equation 13)}$$

and

$$c = \frac{2\Delta + (p-2)\Delta^2}{1 - 4\Delta + 2p\Delta + p^2\Delta^2 - 3p\Delta^2 + 3\Delta^2}. \qquad \text{(Equation 14)}$$

For simplicity, we define the denominator as $d = 1 - 4\Delta + 2p\Delta + p^2\Delta^2 - 3p\Delta^2 + 3\Delta^2$. We then split the summand of Equation 13 into terms: $s_1 = (1-\Delta)^2(1+(p-1)c^2)$ and $s_2 = (2\Delta + (p-2)\Delta^2)((p-1)c-1)^2$. We will further split these into factors containing c: $s_1' = (p-1)c^2 + 1$ and $s_2' = (p-1)c - 1$.

We'll simplify both of the factors, starting with s_2':

$$\begin{aligned} s_2' = (p-1)c - 1 &= \frac{(p-1)(2\Delta + (p-2)\Delta^2) - d}{d} \\ &= \frac{2p\Delta + p^2\Delta^2 - 2p\Delta^2 - 2\Delta - p\Delta^2 + 2\Delta^2 - 1 + 4\Delta - 2p\Delta - p^2\Delta^2 + 3p\Delta^2 - 3\Delta^2}{d} \\ &= \frac{-1 + 4\Delta - 2\Delta + 2\Delta^2 - 3\Delta^2 + 2p\Delta - 2p\Delta + 3p\Delta^2 - 2p\Delta^2 - p\Delta^2 + p^2\Delta^2 - p^2\Delta^2}{d} \\ &= \frac{-1 + 2\Delta - \Delta^2}{d} = \frac{-(1-\Delta)^2}{d}. \end{aligned} \tag{22}$$

Squaring s_2' and multiplying it by $2\Delta + (p-2)\Delta^2$ yields the second term:

$$s_2 = \frac{(1-\Delta)^4(2\Delta + (p-2)\Delta^2)}{d^2} = \frac{(1-\Delta)^4 c}{d}. \tag{23}$$

Now we alter s_1', using the relationship found in Equation 22:

$$\begin{aligned} s_1' &= (p-1)c^2 + 1 \\ &= (p-1)c^2 - c + c + 1 \\ &= ((p-1)c - 1)c + 1 + c \\ &= \frac{-(1-\Delta)^2 c}{d} + 1 + c. \end{aligned} \tag{24}$$

Multiplying s_1' by $(1-\Delta)^2$ yields:

$$s_1 = \frac{-(1-\Delta)^4 c}{d} + (1-\Delta)^2(1+c). \tag{25}$$

Now, we have each of the terms for the summand in Equation 13. Summing Equation 25 and 23 and simplifying yields:

$$s_1 + s_2 = \frac{(1-\Delta)^4 c}{d} + \frac{-(1-\Delta)^4 c}{d} + (1-\Delta)^2(1+c) = (1-\Delta)^2(1+c). \tag{26}$$

Finally, inserting Equation 26 back into Equation 13 yields:

$$\mathbf{LOCO_i} = |\hat{\beta}_i|\sqrt{s_1 + s_2} = |\hat{\beta}_i|\sqrt{(1-\Delta)^2(1+c)} = |\hat{\beta}_i|(1-\Delta)\sqrt{1+c}. \qquad \text{(Equation 15)}$$

□

B Derivation of t-statistic Formula

We now offer a mathematical proof for Theorem 5 where the t-statistic is expressed as a function of $\hat{\beta}_i$, p, and Δ. Proving Theorem 5 is fundamental to proving Theorem 6.

Proof. Proving Theorem 5 begins with the definition for the t-statistic:

$$t_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\sqrt{\mathrm{Var}(\epsilon)(\mathbf{X^T X})^{-1}_{ii}}}. \tag{Equation 17}$$

The main bottleneck here is $\mathbf{X^T X}^{-1}_{ii}$. Given that $\mathbf{X^T X} \sim \mathcal{W}_p(n, \mathbf{A^T A})$, this implies that $\mathbf{X^T X}^{-1} \sim \mathcal{W}_p^{-1}(n, \mathbf{A^T A}^{-1})$, which has $E\left[\mathbf{X^T X}^{-1}\right] = \frac{\mathbf{A^T A}^{-1}}{n-p-1}$.

Recall from Equation 2 (found in subsection 2.1) that $\mathbf{A^T A} = (2\Delta + (p-2)\Delta^2)\mathbf{J} + (1-\Delta)^2\mathbf{I}$. If we let $\alpha = \frac{2\Delta+(p-2)\Delta^2}{(1-\Delta)^2}$, then we can combine definitions with a simplification of Equation 2 to show that

$$\mathbf{A A^T} = (1-\Delta)^2[\mathbf{I} + \alpha\mathbf{J}]. \tag{27}$$

To solve for $\mathbf{A^T A}^{-1}_{ii}$, we use the Sherman–Morrison formula (Sherman & Morrison, 1949) which states:

$$(\mathbf{M} + uv^T)^{-1} = \mathbf{M}^{-1} - \frac{\mathbf{M}^{-1}uv^T\mathbf{M}^{-1}}{1 + v^T\mathbf{M}^{-1}u}. \tag{28}$$

Since $\alpha\mathbf{J}$ in Equation 27 is rank-one, we can express it as a multiplication of vectors filled with 1s: $\alpha\mathbf{1}\mathbf{1}^T$. Now, let $\mathbf{I} = \mathbf{M} = \mathbf{M}^{-1}$ and let $u = v = \sqrt{\alpha}\mathbf{1}$ be vectors of length p with all entries $\sqrt{\alpha}$. Then we can find $(\mathbf{I} + \alpha\mathbf{J})^{-1}$ via the Sherman Morrison formula with some variable substitution and simplification which yields

$$(\mathbf{I} + \alpha\mathbf{J})^{-1} = (\mathbf{I} + uv^T)^{-1} = \mathbf{I} - \frac{uv^T}{1 + v^T u} = \mathbf{I} - \frac{\alpha\mathbf{J}}{1 + \alpha p}. \tag{29}$$

Therefore,

$$\begin{aligned} \mathbf{A^T A}^{-1} &= ((1-\Delta)^2[\mathbf{I} + \alpha\mathbf{J}])^{-1} \\ &= \frac{1}{(1-\Delta)^2}\left[\mathbf{I} - \frac{\alpha\mathbf{J}}{1+\alpha p}\right]. \end{aligned} \tag{30}$$

Substituting $\mathbf{I}_{ii} = \mathbf{J}_{ii} = 1$ into Equation 30 and simplifying yields:

$$\begin{aligned} (\mathbf{A^T A})^{-1}_{ii} &= \frac{1}{(1-\Delta)^2}\left[1 - \frac{\alpha}{1+\alpha p}\right] \\ &= \frac{1}{(1-\Delta)^2}\left[\frac{1+\alpha(p-1)}{1+\alpha p}\right]. \end{aligned} \tag{31}$$

Substituting α into the latter expression of Equation 31 and simplifying yields:

$$\begin{aligned}
\left[\frac{1+(p-1)\alpha}{1+p\alpha}\right] &= \left[\frac{1+(p-1)\frac{2\Delta+(p-2)\Delta^2}{(1-\Delta)^2}}{1+p\frac{2\Delta+(p-2)\Delta^2}{(1-\Delta)^2}}\right] \\
&= \left[\frac{\frac{(1-\Delta)^2+2(p-1)\Delta+(p-1)(p-2)\Delta^2}{(1-\Delta)^2}}{\frac{(1-\Delta)^2+2p\Delta+p(p-2)\Delta^2}{(1-\Delta)^2}}\right] \\
&= \left[\frac{(1-\Delta)^2+2(p-1)\Delta+(p-1)(p-2)\Delta^2}{(1-\Delta)^2+2p\Delta+p(p-2)\Delta^2}\right] \\
&= \left[\frac{1-2\Delta+\Delta^2+2p\Delta-2\Delta+p^2\Delta^2-3p\Delta^2+2\Delta^2}{1-2\Delta+\Delta^2+2p\Delta+p^2\Delta^2-2p\Delta^2}\right] \\
&= \left[\frac{1+2p\Delta-4\Delta+p^2\Delta^2-3p\Delta^2+3\Delta^2}{1+2p\Delta-2\Delta+p^2\Delta^2-2p\Delta^2+\Delta^2}\right] \\
&= \left[\frac{1+2p\Delta-4\Delta+p^2\Delta^2-4p\Delta^2+4\Delta^2+(p-1)\Delta^2}{1+2(p-1)\Delta+(p-1)^2\Delta^2}\right] \\
&= \left[\frac{1+2(p-2)\Delta+(p-2)\Delta^2+(p-1)\Delta^2}{(1+(p-1)\Delta)^2}\right] \\
&= \left[\frac{(1+(p-2)\Delta)^2+(p-1)\Delta^2}{(1+(p-1)\Delta)^2}\right].
\end{aligned} \tag{32}$$

Therefore,

$$\begin{aligned}
(\mathbf{A}^{\mathbf{T}}\mathbf{A})^{-1}_{ii} &= \frac{1}{(1-\Delta)^2}\left[\frac{1+\alpha(p-1)}{1+\alpha p}\right] \\
&= \frac{(1+(p-2)\Delta)^2+(p-1)\Delta^2}{(1-\Delta)^2(1+(p-1)\Delta)^2}.
\end{aligned} \tag{33}$$

This means,

$$E\left[\mathbf{X}^{\mathbf{T}}\mathbf{X}^{-1}\right] = \frac{\mathbf{A}^{\mathbf{T}}\mathbf{A}^{-1}}{n-p-1} = \frac{(1+(p-2)\Delta)^2+(p-1)\Delta^2}{(n-p-1)(1-\Delta)^2(1+(p-1)\Delta)^2}. \tag{34}$$

Placing the inverse Wishart expectation from Equation 34 into Equation 17, and taking $\beta = \hat{\beta}$ for simplicity, we obtain a population-level approximation of the t-statistic:

$$t_i \approx \hat{\beta}_i(1-\Delta)\sqrt{\frac{(n-p-1)(1+(p-1)\Delta)^2}{\mathrm{Var}(\epsilon)[(1+(p-2)\Delta)^2+(p-1)\Delta^2]}}. \tag{Equation 18}$$

□

C Empirical Relationship between LOCO and t-statistic

In similar fashion to subsection 3.1, we offer Figure 5 for comparing empirical t-statistics to empirical LOCO metrics and validating our derivation of Equation 19. Since both metrics are empirical, we employ a parity scatter plots to display this comparison. We apply a small sample bias correction from subsection 3.2.2 to LOCO. Conversely, no small sample bias correction is applied to the t-statistics since this is already accounted for in their calculation. Noticeable clusters are monotonically driven by Δ, where empirical values near 1 correspond to low Δ and values near 0 correspond to high Δ.

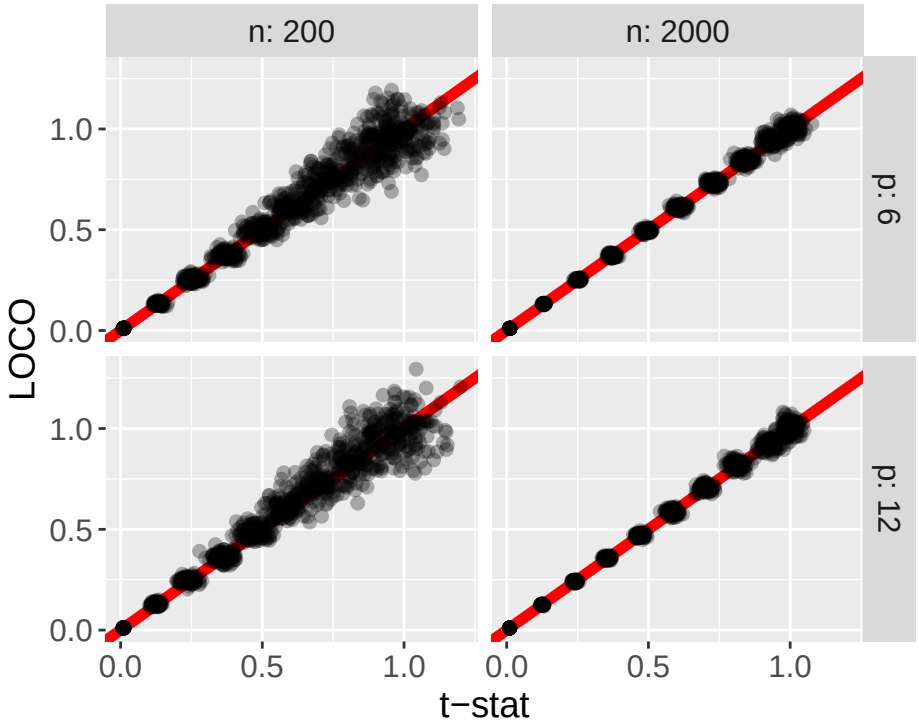

Figure 5: Parity scatter plots for assessing the relationship between empirical t-statistics and empirical LOCO values for data defined in subsection 2.4.1. Red lines denote a slope of 1. Plots are faceted by p and n dimensions. We used Equation 19 to ensure a slope of 1 by transforming the t-statistics. We also applied a small sample bias correction from subsection 3.2.2 to LOCO. No small sample bias correction was applied to the t-statistics since this is already accounted for in their calculation.

Since LOCO values naturally ranged from 0 to 1, we use and reverse Equation 19 to multiply the t-statistics by $\sqrt{\frac{\text{Var}(\epsilon)}{n-1}}$, with the understanding that the scatter plots should have a slope of 1 if Equation 19 is correct. Plots for the full simulation range are available in the supplement in Section 1.

From the comparisons displayed in Figure 5, Equation 19 appears to be correct, as the resulting points visually fit a line with a slope of 1 extremely well. This offers empirical evidence supporting the work in subsection 2.3 and Appendix B.

D Uniform Coefficient Absorption

Here we provide an additional empirical demonstration showing that regression coefficients tend to redistribute additively across correlated predictors. In this example, we intentionally vary the beta coefficient magnitudes to investigate whether absorption occurs uniformly or whether larger coefficients absorb disproportionately more signal.

In this case, we simulate $\mathbf{X}$ as a set of 4 predictor variables with 2000 observations. The predictors are generated using the construction described in Section 2.4.1, in which the linear transformation $\mathbf{A}$ is applied to a matrix $\mathbf{Z}$ of independent standard normal variables. The response variable is defined as

$$\mathbf{y} = 9x_1 + 6x_2 + 3x_3 + 0x_4 + \epsilon, \tag{35}$$

where $\epsilon \overset{\text{iid}}{\sim} \mathcal{N}(\text{Mean} = 0, \text{Var} = 0.25)$.

We systematically vary $\Delta \in \{0, 0.3, 0.6, 0.9\}$ and perform 100 Monte Carlo simulations using identically structured training and validation sets.

Figure 6 shows the change in all regression coefficients after each predictor is permuted and the model is refit under varying levels of collinearity. When $\Delta = 0$, permuting a predictor produces essentially no redistribution to the remaining coefficients. As Δ increases, however, the permuted coefficient is redistributed approximately uniformly across the remaining predictors. At $\Delta = 0.9$, permuting x_1 yields nearly complete redistribution of β_1 across β_2, β_3, and β_4, despite their differing original magnitudes. Similar behavior occurs for each predictor in turn. Intermediate values of Δ exhibit the same qualitative pattern with weaker redistribution, where approximately 83% of the coefficient magnitude is redistributed when $\Delta = 0.3$ and approximately 97.5% is redistributed when $\Delta = 0.6$. Importantly, redistribution appears approximately independent of the original coefficient magnitude, supporting the symmetric redistribution assumption introduced in Section 2.3.

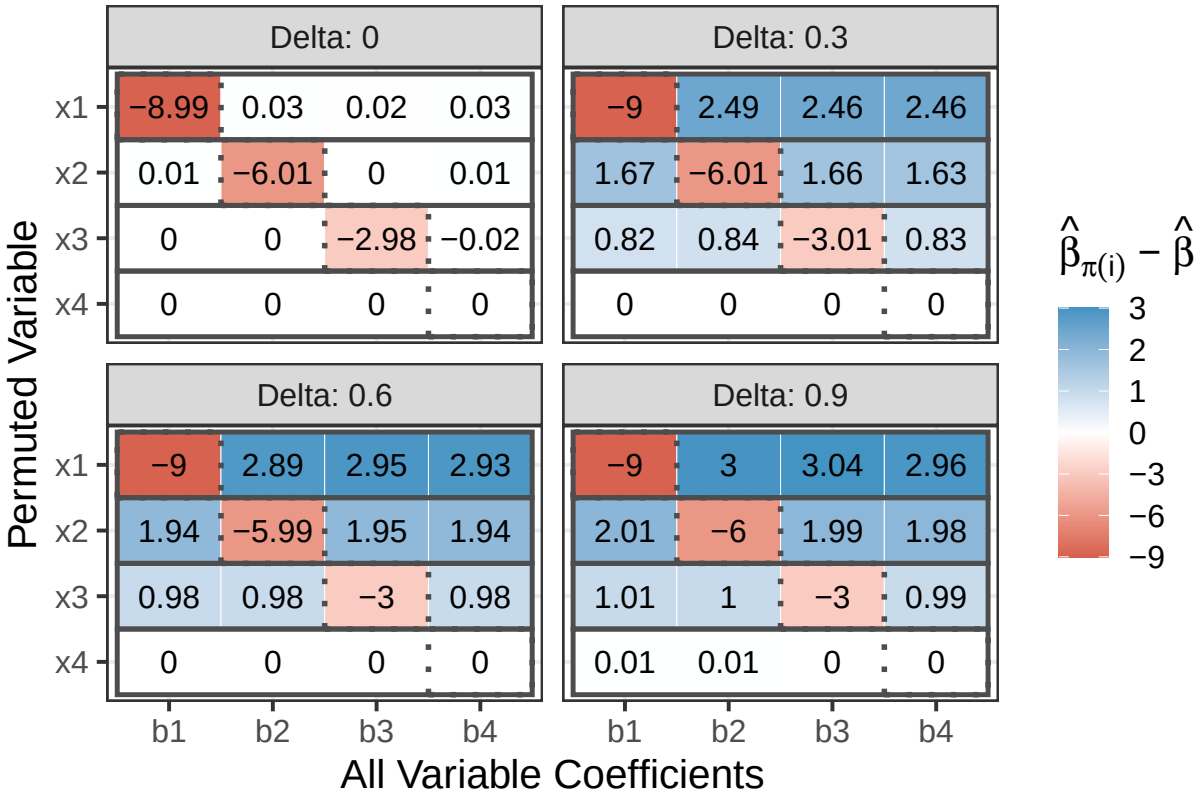

Figure 6: Heatmaps showing beta coefficient redistribution after permuting each predictor and refitting a linear model under increasing collinearity levels Δ. Rows indicate the permuted variable, columns indicate the affected coefficient, and colors/entries represent changes in the four coefficients when a variable is permuted ($\hat{\beta}_{\pi(i)} - \hat{\beta}$). As collinearity increases, redistribution becomes increasingly complete. Each lost coefficient is redistributed approximately uniformly across the remaining predictors regardless of their original magnitudes.

E Nonlinear Response Functions

While our theoretical derivations rely on linearity assumptions, it is worth considering whether the resulting variable importance trends generalize at some level to nonlinear settings. Although nonlinear theoretical

derivations are beyond the scope of this manuscript, we include two empirical studies comparing nonlinear feature importance behavior to the theoretical trends derived under linear response functions.

In both datasets, predictors are generated using the latent-variable construction described in Section 2.4.1, where collinearity is controlled through the parameter Δ. The resulting Gaussian predictors were transformed to approximately uniform predictor distributions while preserving the dependence structure induced by Δ. Note, rescaling the predictor variables in this manner prevents sparse observations in the tails of the distribution from disproportionately impacting the results. For the sinusoidal setting, predictors are scaled to the interval $[-3\pi/2, 3\pi/2]$ to provide coverage across two full oscillations of the sine function. We vary $\Delta \in \{0, 0.11, \ldots, 0.88, 0.99\}$ and compare empirical PaP and LOCO importance values across generalized additive models (GAMs) (Hastie & Tibshirani, 1986) and Random Forests.

In the first setting, the response is generated using a quadratic signal,

$$y = \sum_{j=1}^{p} x_j^2 + \epsilon, \tag{36}$$

while in the second setting the response follows a sinusoidal structure,

$$y = \sum_{j=1}^{p} \sin(x_j) + \epsilon, \tag{37}$$

where $\epsilon \overset{\text{iid}}{\sim} \mathcal{N}(\text{Mean} = 0, \text{Var} = 0.01)$. We consider combinations of $n \in \{200, 2000\}$ and $p \in \{6, 12\}$ to evaluate how these relationships behave under varying sample sizes and dimensionalities.

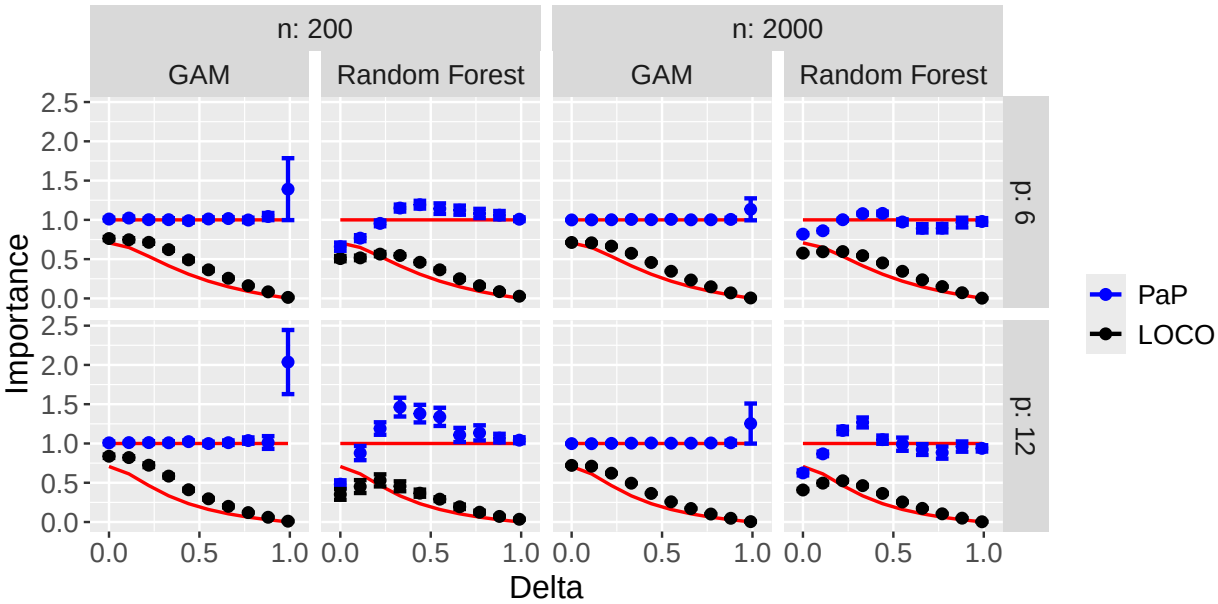

Figure 7: Mean and 2 Standard error bars for assessing the relationship between empirical PaP and LOCO importance values and Δ under Equation 36. Results are shown for GAMs and Random Forests under different sample sizes and predictor dimensions. Red curves denote the theoretical linear-regression expressions derived in Section 2. Although Equation 36 violates the assumptions underlying the theory, the empirical trends remain qualitatively similar.

Figures 7 and 8 compare the resulting empirical importance trends against the theoretical linear-regression expressions derived in Section 2 (shown as red curves). Because the underlying response functions are nonlinear, the empirical values are not expected to match the linear theory exactly. Nevertheless, both PaP and LOCO continue to exhibit qualitatively similar dependence on collinearity as in the linear setting. In particular, PaP importance generally remains flat or oscillates with Δ, while LOCO importance decreases as correlated predictors increasingly absorb shared signal. These results suggest that the broad effects of collinearity on variable importance measures described in this paper hold even for some non-linear settings. In particular, across both nonlinear response functions, increasing collinearity among predictors systematically deflates LOCO importance, while PaP importance remains comparatively stable.

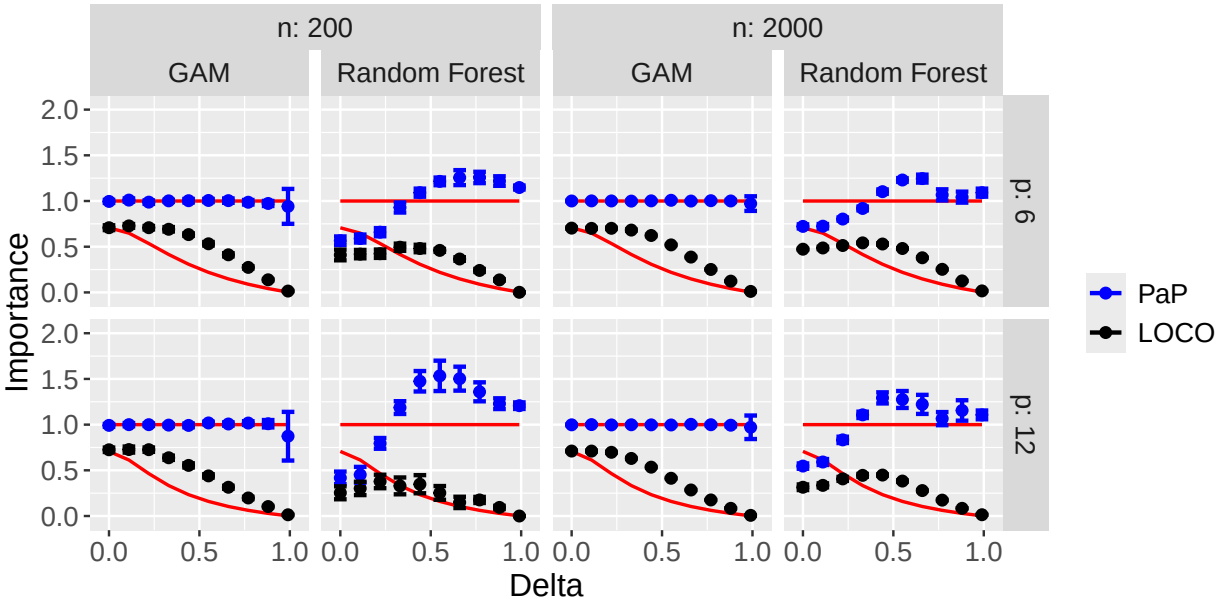

Figure 8: Mean and 2 Standard error bars for assessing the relationship between empirical PaP and LOCO importance values and Δ under Equation 37. Results are shown for GAMs and Random Forests under different sample sizes and predictor dimensions. Red curves denote the theoretical linear-regression expressions derived in Section 2. While the empirical results do not exactly match the those of Figure 7 or follow the linear theory, the overall dependence on collinearity remains directionally consistent.