Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 32(7): 4495–4514.
DOI: 10.3934/era.2024203
Received: 31 December 2023
Revised: 01 June 2024
Accepted: 08 July 2024
Published: 22 July 2024

Research article

Assessing agreement between permutation and dropout variable importance methods for regression and random forest models

Kelvyn Bladen*and D. Richard Cutler

Department of Mathematics and Statistics, Utah State University, 3900 Old Main Hill, Logan, UT 84322, USA

* **Correspondence:** Email: kelvyn.bladen@usu.edu; Tel: +1-435-258-7782.

Abstract: Permutation techniques have been used extensively in machine learning algorithms for evaluating variable importance. In ordinary regression, however, variables are often removed to gauge their importance. In this paper, we compared the results for permuting variables to removing variables in regression to assess relations between these two methods. We compared permute-and-predict (PaP) methods with leave-one-covariate-out (LOCO) techniques. We also compared these results with conventional metrics such as regression coefficient estimates, t-statistics, and random forest out-of-bag (OOB) PaP importance. Our results indicate that permutation importance metrics are practically equivalent to those obtained from removing variables in a regression setting. We demonstrate a strong association between the PaP metrics, true coefficients, and regression-estimated coefficients. We also show a strong relation between the LOCO metrics and the regression t-statistics. Finally, we illustrate that manual PaP methods are not equivalent to the OOB PaP technique and suggest prioritizing the use of manual PaP methods on validation data.

Keywords: permutation; variable importance; random forest; variable selection; regression; machine learning

1. Introduction

Variable importance and model selection are nuanced concepts that are relevant in statistics, data science, and many other areas of scientific literature (see Kruskal et al. [1]). Perhaps the simplest example of a metric for variable importance in regression may be found in introductory textbooks (see Achen [2]). When all variables have been standardized, the magnitude of the regression coefficients are considered measures of importance of the associated variables. A slew of variable importance measures have been developed over the years, including t-statistics and stepwise elimination of variables on the basis of statistical significance or measures like AIC. Most of these have been superseded by the

LASSO [3] and variations, including the elastic net [4].

A different and axiomatic approach was taken by Pratt [5]. Starting with exchangeable, standardized predictor variables, Pratt showed that the importance of the j^{th} predictor variable may be defined as:

$$VarImp_j = \hat{\beta}_j \times r_{y,x_j} \tag{1.1}$$

where $\hat{\beta}_j$ is the regression coefficient for the j^{th} predictor variable and r_{y,x_j} is the marginal (Pearson) correlation between the response (y) and the j^{th} predictor variable, x_j.

In regression, the approach of using t-statistics for assessing variable importance may be shown to be equivalent to removing a variable and looking at the difference in mean squared errors for the models with and without the variable of interest. Complex machine learning methods, such as random forests, do not employ regression variable diagnostics such as estimated coefficients or t-statistics. Leo Breiman [6] introduced a variation of this idea for random forests in which the so-called out-of-bag (OOB) values on a variable of interest are permuted-and-predicted (PaP) so that change in accuracy can be observed and aggregated over all observations. This approach and related permutation methods have become a standard method in machine learning over the last 20 years; but see Strobl et al. [7] and Bladen [8] for a discussion of the impact of collinearity on permutation variable importance. Hooker et al. [9] pointed out that the variable permutation of Breiman's [6] original algorithm leads to a form of potentially problematic extrapolation and suggested that re-learning additional models is required to handle this problem. They pointed to the work done by Lei et al. [10] involving the technique of leave-one-covariate-out (LOCO) of the dataset and re-learning the model of interest. Barber et al. [11] and Candès et al. [12] chose to handle this via a technique they call knockoffs, which involves switching original variables for random replacements that are sampled conditionally on the remaining variables. Each of the LOCO and knockoff techniques involves fitting a new model after removing, permuting, or otherwise altering a training variable and comparing the new model to the original untainted model.

Variable importance assessments for random forests have been explored and furthered in various ways. In their work, Hooker et al. [9], expanded upon the LOCO technique and utilized a similar concept for work with random forests. Ye et al. [13] created the SOIL technique designed for sparse linear modeling and high-dimensional regression and showed the proficiency it has in model selection for random forests. Strobl et al. [7] attempted to navigate feature importance issues caused by collinearity via a conditional variable importance algorithm. Many researchers have developed plots relating changes in the predicted response to changes in a particular variable while holding other variables constant. Apley et al. [14] contibuted with the concept and visualization known as accumulated local effects (ALE) plots. Goldstein et al. [15] and Greenwell et al. [16] developed individual conditional expectation (ICE) plots and averaged partial dependence plots (PDPs), respectively. PDPs, in particular, have been used to assess the importance of a variable by taking the standard deviation of the response predictions from the PDP for the desired variable [16].

In this paper, we highlight how some variations of the PaP and LOCO methods for assessing variable importance in linear modeling and machine learning relate to regression metrics. By doing so, we hope to illustrate which variable importance methods should be employed to best approximate specific regression metrics of interest.

In the remainder of this section, we introduce, define, and provide notations for standard regression metrics and several machine learning variable importance computations. In Section 2, we outline our simulated data for comparing these importance values. In Section 3, we show comparative plots of our

importance metrics and comment on observed results within them. Finally, in Section 4, we offer our conclusions about the different variable importance metrics we analyze.

1.1. Regression metrics

In ordinary regression, there are three different metrics provided in standard summary tables: estimated coefficients, t-statistics, and p-values. If variables have been standardized, then higher magnitudes for coefficients are often interpreted to imply greater importance [2]. The same should be true for t-statistics, which are robust to standardization. While p-values communicate similar information to t-statistics, we recognize that p-values are rarely optimal for interpreting variable importance because they have an inverse and nonlinear relationship with t-statistics.

If variables are orthogonal, then t-statistics are proportional to the estimated coefficients. Let $\hat{\beta}_j$ be the estimate for the j^{th} regression coefficient. These two metrics have the following relationship for the j^{th} variable:

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \propto \hat{\beta}_j \tag{1.2}$$

if the predictor variables have been standardized so that the standard error $SE(\hat{\beta}_j)$ is irrespective of which variable we are considering. However, these relationship conditions rarely hold true, which suggests a potential disconnect between the estimated coefficients and the t-statistics. A common issue occurs when variables are collinear. The standard errors of their estimated coefficients are then larger, which diminishes their t-statistics. Thus, the effect or relationship between a predictor and the response remains the same, but we are less certain of this relationship.

1.2. Permute-and-predict (PaP) metrics

In more complex machine learning algorithms, a technique for assessing variable importance was introduced by Breiman [6]. Commonly called the permute-and-predict (PaP) method, it involves randomly scrambling the values of a chosen variable and accumulating predictions for the newly altered data. This method was originally conceived for random forests and was performed on out-of-bag (OOB) data, a type of validation dataset for individual trees. We will use the OOB PaP technique when assessing random forest importance metrics, but we will also perform the procedure manually on a separate validation set to compare to regression metrics and the OOB PaP.

Our research is motivated by earlier work we performed assessing how random forest variable importance metrics depend on collinearity and `mtry`, a tunable hyper-parameter of the number of randomly selected variables to consider for a given split for a tree in the forest [8]. In that work, we derived a heuristic suggesting a relationship between the regression t-statistics and random forest variable importance metrics. This heuristic submits that $\text{importance}_j \propto t_j^2$, which implies that a square root transformation is helpful for relating random forest importances back to regression metrics. In this paper, we will explore the relationship between regression metrics and PaP importances. We will also look at the relationships these metrics have with model refitting (LOCO importances) and with the true equation coefficients.

If the predictions using the raw data and the permuted data yield similar accuracies or errors, then that variable is not particularly important. Alternatively, if the original predictions clearly outperform the permutation predictions, then the variable is important.

Mathematically, our permutation variable importance may be expressed using syntax where the index indicates the validation data that we generate predictions for. Thus,

$$\text{PaP}_j = \sqrt{\frac{\text{vMSE}_{\sim j} - \text{MSE}}{\text{MSE}}}, \tag{1.3}$$

where MSE is the validation mean squared error from the original model and $\text{vMSE}_{\sim j}$ is the MSE generated from that same model when the validation values of the j^{th} predictor have been permuted.

Alternatively, the variable importance can be assessed with a drop-and-predict (DaP) method. In our work, we accomplish this dropout by first rescaling each variable so that the center of the distribution is 0 and then setting all values of a given variable to the central value of 0. The structure is fundamentally equivalent to Eq (1.3), with one simple difference:

$$\text{DaP}_j = \sqrt{\frac{\text{vMSE}_{-j} - \text{MSE}}{\text{MSE}}}, \tag{1.4}$$

where vMSE_{-j} is the MSE generated from the original model when the validation values of the j^{th} predictor have all been set to the central value of 0.

These importance metrics are therefore equivalent if $\text{vMSE}_{\sim j} = \text{vMSE}_{-j}$.

1.3. Leave-one-covariate-out (LOCO) metrics

Another method for assessing variable importance is called re-learning (see Hooker et al. [9]) or LOCO (see Lei et al. [10]). When performing this technique, we randomly permute the values of a chosen variable in the training data rather than the validation data. We then build a new regression model using the training data containing the permuted feature. The difference of MSEs between the permuted data model and the original model is again computed. Just as before, if the original predictions clearly outperform the permutation predictions, then the variable is important.

The LOCO variable importance will be expressed in similar fashion to the PaP and DaP methods in Section 1.2. However, in this syntax, the index provides information about the model used to extract prediction and the training data that generated that model. Thus,

$$\text{perm\_LOCO}_j = \sqrt{\frac{\text{tMSE}_{\sim j} - \text{MSE}}{\text{MSE}}}, \tag{1.5}$$

where MSE is the validation mean squared error from the original model, just as in Section 1.2. The only difference between this and Eq (1.3) is the substitution of $\text{vMSE}_{\sim j}$ with $\text{tMSE}_{\sim j}$. Here, $\text{tMSE}_{\sim j}$ is the MSE generated with the original validation data but from a new regression model where training values of the j^{th} predictor have been permuted.

Just like DaP techniques, LOCO variable importances can be assessed by rescaling each variable so the distribution center is 0 and then setting the values of the variable to 0. The structure is identical to Eq (1.5), with the exception of substituting notation for dropping a variable rather than permuting it:

$$\text{drop\_LOCO}_j = \sqrt{\frac{\text{tMSE}_{-j} - \text{MSE}}{\text{MSE}}}, \tag{1.6}$$

where tMSE_{-j} is the MSE generated with the original validation data but from a new regression model where training values of the j^{th} predictor have all been set to the central value of 0.

Similar to Section 1.2, these importance metrics are equivalent if $\text{tMSE}_{\sim j} = \text{tMSE}_{-j}$.

2. Methods

Using the standard regression metrics and the machine learning importance definitions listed above, we designed several simulations to assess the relationships among these metrics. The general architecture is to build both a training and validation dataset of identical sizes and structures with established coefficients and variable relations. In our subsequent methods and analyzes, we use only the training data for model creation, while we use the validation data to assess variable importance. We then collect regression metrics and compute PaP and LOCO metrics. Finally, we plot pairwise scatterplots and correlations amongst these importance values to assess the relationship between them. Higher correlations between metrics provide an empirical foundation for showing proportionality and agreement between variable importance assessments. We utilize the definitions in Sections 1.2 and 1.3 when computing permutation and dropout variable importance metrics [8].

2.1. Orthogonal data

For our first simulation, we let six predictor variables be independent and identically distributed (iid) from a $\mathcal{N}(0, 1)$ distribution. Here, iid implies that all of the features are orthogonal to each other. We generated 1000 training observations and 1000 validation observations for each of these variables and then created the response variable values using the following equation:

$$y = 5v_1 + 4v_2 + 3v_3 + 2v_4 + 1v_5 + 0v_6 + \epsilon. \tag{2.1}$$

where initially $\epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and then in a second iteration $\epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, 10)$.

2.2. Collinear data

For our next simulation, we utilize a data structure from Strobl et al. [7] that allows for high collinearity among some of the predictor variables. In this architecture, we use a linear equation with twelve predictor variables and 1000 observations in the training and validation sets.

The regression coefficients for the predictor variables are chosen as follows:

$$y = 5Cor_1 + 5Cor_2 + 2Cor_3 + 0Cor_4 + 5v_5 + 5v_6 + 2v_7 + 0v_{8-12} + \epsilon, \tag{2.2}$$

where initially $\epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and then in a second iteration $\epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, 10)$.

The predictor variables are sampled from a multivariate normal distribution:

$$Cor_1, ..., Cor_4, v_5, ..., v_{12} \sim \mathcal{N}(0, \Sigma).$$

The covariance structure Σ is chosen such that all variables have unit variance $\sigma_{i,i} = 1$ and the first four predictor variables are block-correlated with $\sigma_{i,j} = 0.95$ for $i \neq j \leq 4$, while the rest are orthogonal with $\sigma_{i,j} = 0$.

2.3. Nonlinear squared data

We expand the previous simulation to create nonlinear datasets. We again use 12 predictor variables and 1000 observations. We initially sample the predictors from the exact multivariate normal distribution described in Section 2.2 to impose correlations. We then take the cumulative distribution function

of each predictor to convert it to a standard uniform distribution, $\mathcal{U}(0, 1)$. We multiply these values by 2 and subtract 1 from them to provide a new distribution of $\mathcal{U}(-1, 1)$ for each predictor.

The regression equation matches the one found in Section 2.2, except we square each variable as follows:

$$y = 5Cor_1^2 + 5Cor_2^2 + 2Cor_3^2 + 0Cor_4^2 + 5v_5^2 + 5v_6^2 + 2v_7^2 + 0v_{8-12}^2 + \epsilon, \tag{2.3}$$

where $\epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.1)$.

2.4. Nonlinear cosine data

Leaning heavily on the architecture in Section 2.3, we now create a response using cosine relations with the predictors. We sample the predictors from the same technique described in Section 2.3 to impose correlations and create standard uniform distributions, $\mathcal{U}(0, 1)$. Instead of converting this to $\mathcal{U}(-1, 1)$, we convert it to $\mathcal{U}(-4\pi, 4\pi)$ by multiplying the values by 8π and subtracting 4π from them.

The regression equation matches that found in Section 2.2, except we remove the error term and take the cosine of each variable as follows:

$$\begin{aligned} y = {} & 5cos(Cor_1) + 5cos(Cor_2) + 2cos(Cor_3) + 0cos(Cor_4) + \\ & 5cos(v_5) + 5cos(v_6) + 2cos(v_7) + 0cos(v_{8-12}). \end{aligned} \tag{2.4}$$

2.5. Interaction data

Finally, we generate interaction datasets of 1000 observations for 8 predictor variables and an error (ϵ) term. Each of the predictors is orthogonal and identically distributed from a $\mathcal{N}(0, 1)$ distribution. We then create the response with this equation:

$$y = 4v_1v_2 + 2v_3v_4 + 1v_5v_6 + 0v_7v_8 + \epsilon, \tag{2.5}$$

where $\epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.1)$.

2.6. Regression modeling and variable importance

For each linear equation, we fit a linear regression model on the training data. The regression model summary provides estimates for the true equation coefficients and t-statistics. We collect and compare these to several machine learning assessments of importance and to the true coefficients.

For all datasets, except the orthogonal data, we fit a random forest regression model on the training data. In this case, we compare machine learning variable importances to the true coefficients, regression metrics (if the equation is linear), and the default random forest variable importance technique of OOB PaP discussed in Section 1.2.

The machine learning assessments that we will collect are the PaP, DaP, perm_LOCO, and drop_LOCO metrics discussed in Sections 1.2 and 1.3. For each metric, large values suggest higher variable importance, while small values indicate low variable importance.

3. Results

We now provide pairwise scatterplots and correlations between each of the variable importance metrics calculated for our simulated datasets. We also show results relating to random forest models

and compare their default importance metric to other PaP metrics as functions of the dominant hyperparameter `mtry`.

3.1. Orthogonal comparison

We begin with our dataset where variables are all orthogonal. The importance values and their pairwise plots are provided in Figure 1. For this analysis, the regression $R^2 \approx 0.98$.

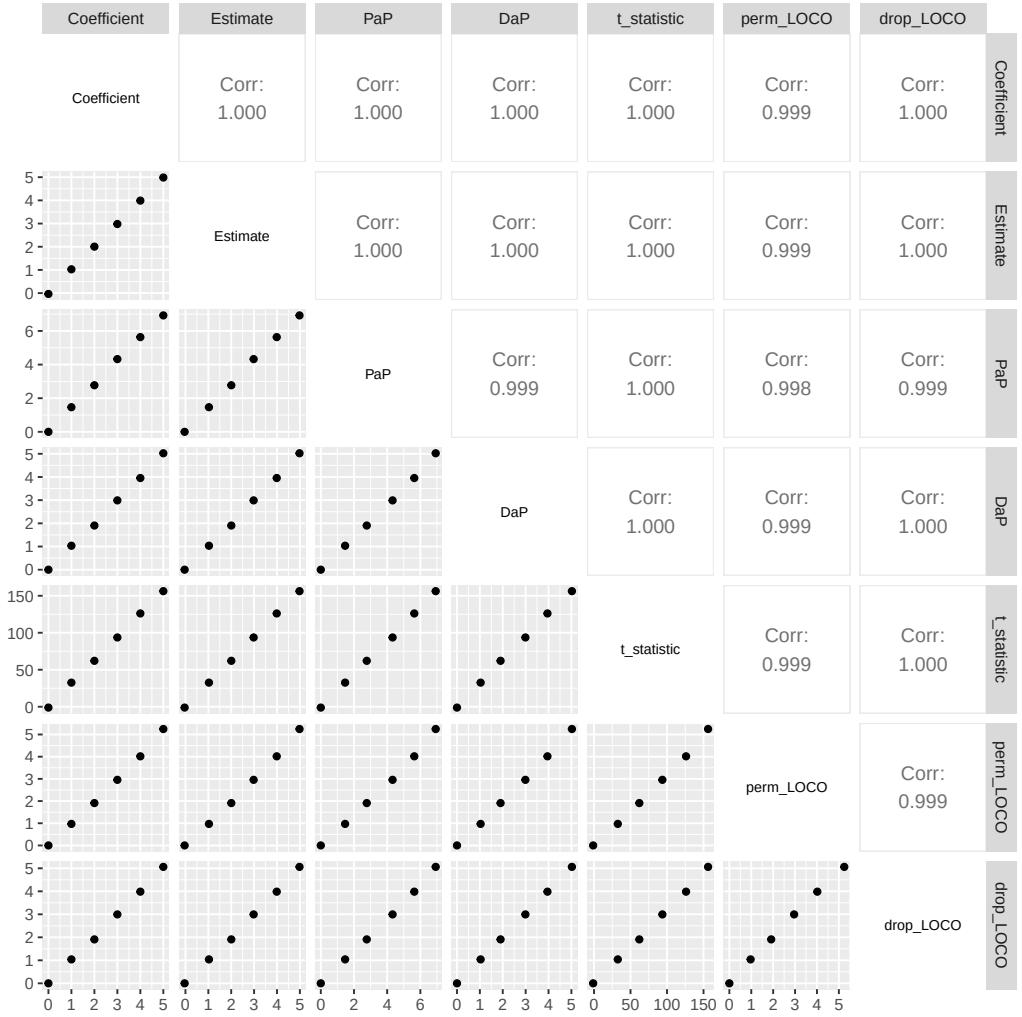

Figure 1. Comparative plots of importance metrics for the linear regression model. The equation for the response is $y = 5v_1 + 4v_2 + 3v_3 + 2v_4 + v_5 + 0v_6 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. See also Eq (2.1).

From Figure 1, we observe the following when variables are orthogonal:

- All variable importance metrics have a near perfect correlation with an average of 1.00.
- While this example is trivial, these results do not deviate from expectations and illustrate agreement between all importance metrics for orthogonal predictors such as principal components.

We will now use an identical regression structure but increase the standard deviation of the noise so that $\epsilon \sim \mathcal{N}(0, 10)$. This yields a massive drop in the regression $R^2 \approx 0.35$. The plots of these results are in Figure 2.

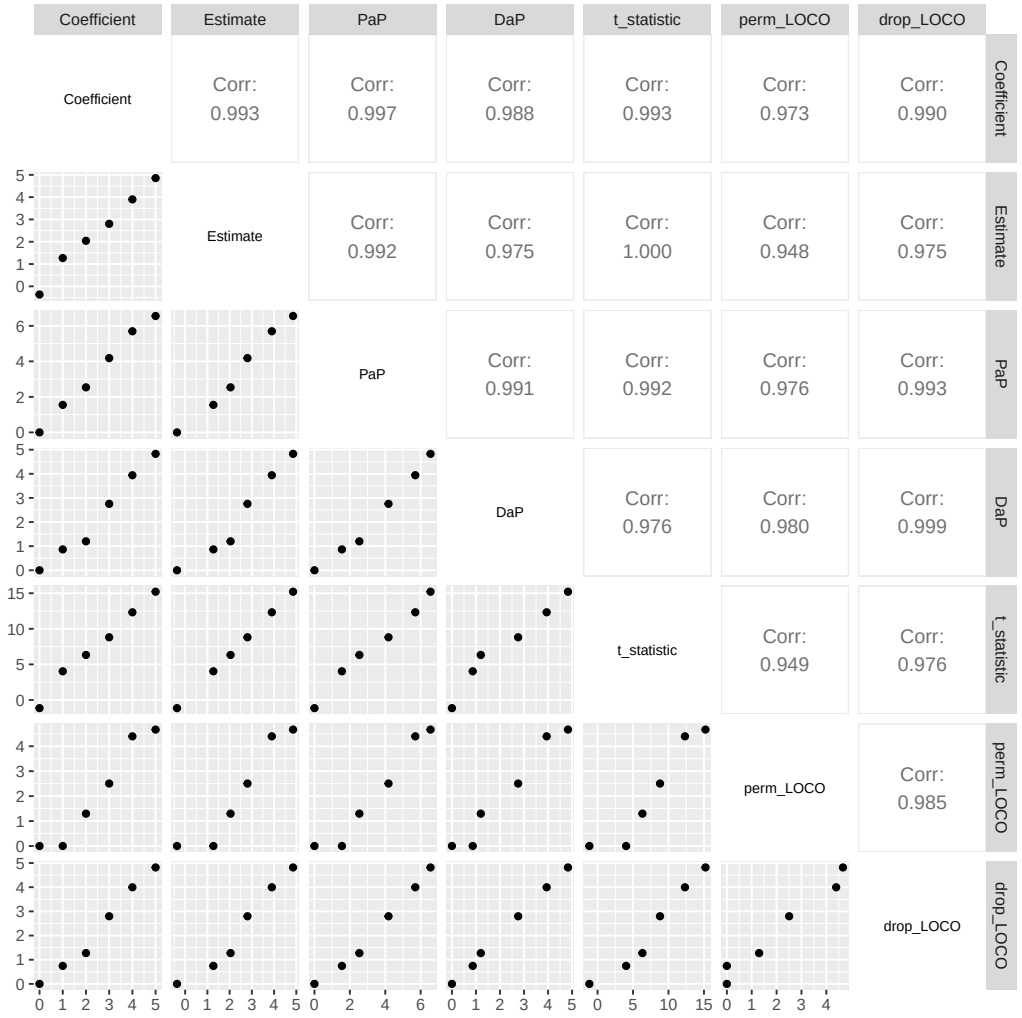

Figure 2. Comparative plots of importance metrics for the linear regression model. The equation for the response is $y = 5v_1 + 4v_2 + 3v_3 + 2v_4 + v_5 + 0v_6 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 10)$. See also Eq (2.1).

From Figure 2, we observe the following when increasing the error for the response variable:

- All variable importance metrics have a strong association with an average correlation of 0.98.
- These results illustrate agreement between all importance metrics for orthogonal predictors.

3.2. Collinearity comparison

We move to the dataset where some variables possess high collinearity. The pairwise plots of importance assessments for a linear regression model are provided in Figure 3. For this analysis, $\epsilon \sim \mathcal{N}(0, 1)$ and the regression $R^2 \approx 0.995$.

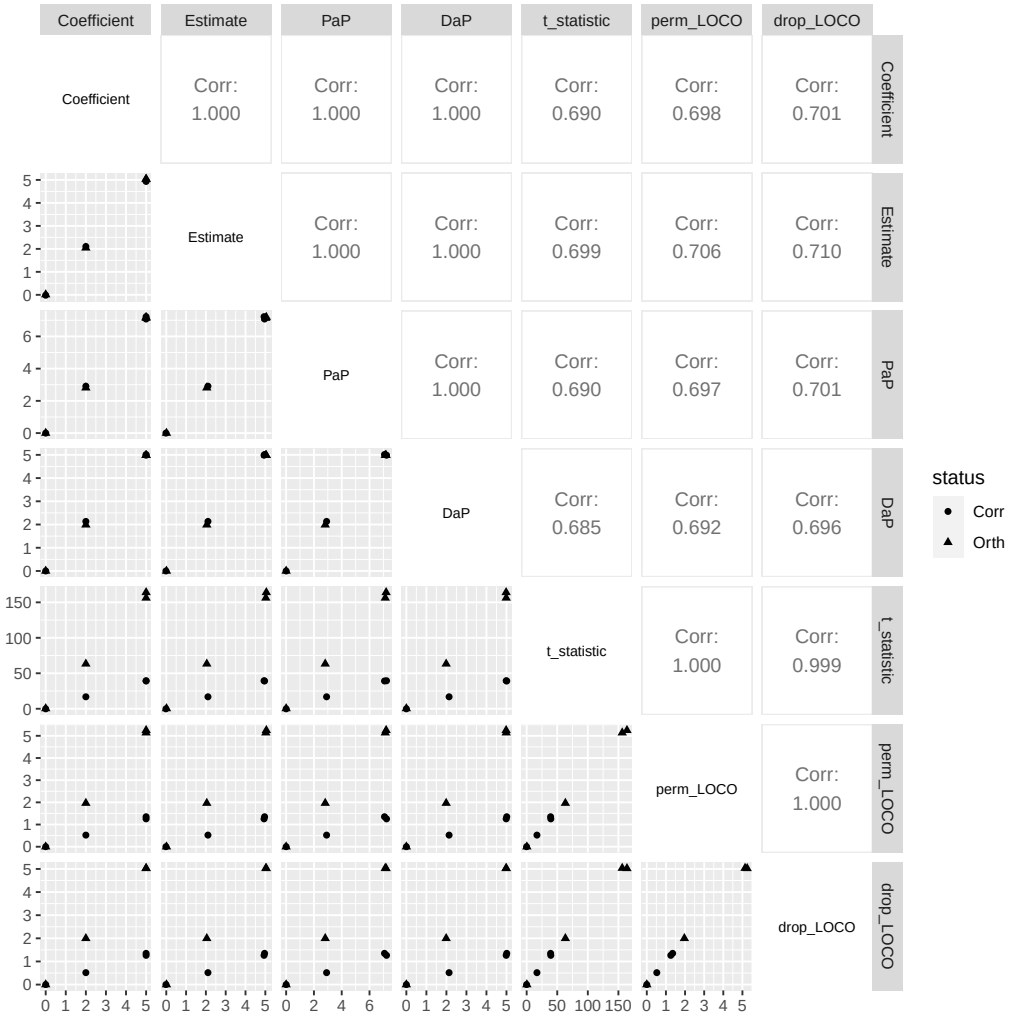

Figure 3. Comparative plots of importance metrics taken from or performed on the linear regression model. Plot symbols denote the status of a given variable: whether it was part of the correlated or orthogonal group of features. The equation for the response is $y = 5Cor_1 + 5Cor_2 + 2Cor_3 + 0Cor_4 + 5v_5 + 5v_6 + 2v_7 + 0v_8 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. See also Eq (2.2).

From Figure 3, we observe the following when variables are highly collinear:

- Coefficients have perfect correlation with estimates but only moderate correlation with t-statistics.
- There are two blocks of high similarity metrics.
- The first block involving the true coefficients, the regression estimates, and PaP techniques has an average correlation of 1.00.
- The second block, involving t-statistics and LOCO techniques, has an average correlation of 1.00.
- In both blocks, the respective permute and drop methods have a perfect correlation of 1.00.

We now analyze this same dataset utilizing a random forest model. The pairwise plots of importance are provided in Figure 4. We keep the linear regression estimates and t-statistics from Figure 3, but now we perform the PaP and LOCO methods on the random forest model.

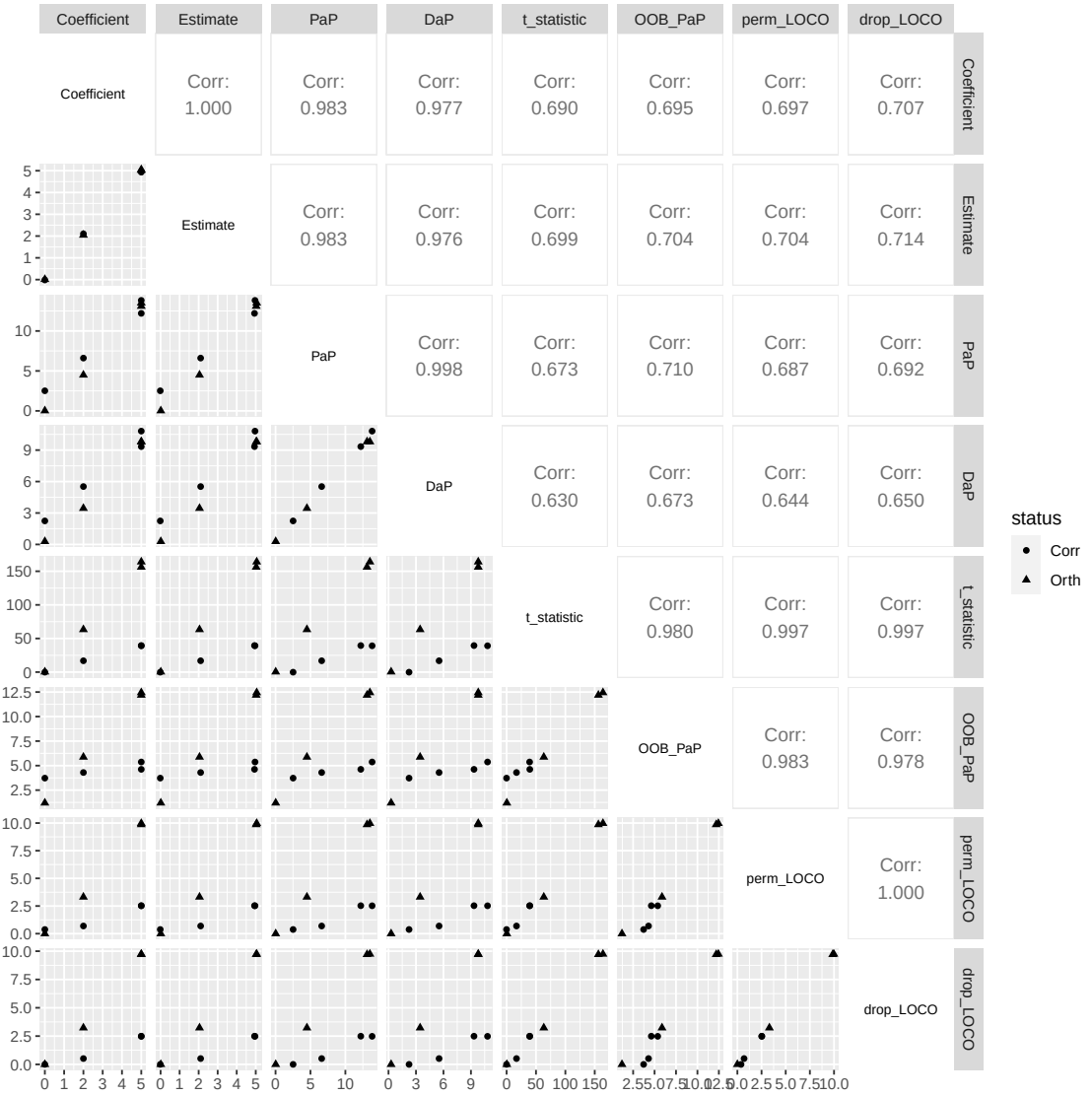

Figure 4. Comparative plots of importance metrics: regression model estimates and t-statistics and PaP and LOCO methods performed on a random forest model. Plot symbols denote the status of a given variable: whether it was part of the correlated or orthogonal group of features. The equation for the response is $y = 5Cor_1 + 5Cor_2 + 2Cor_3 + 0Cor_4 + 5v_5 + 5v_6 + 2v_7 + 0v_8 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. See also Eq (2.2).

From Figure 4, we observe the following for high-collinearity data:

- The same two blocks of high-similarity metrics found in Figure 3 are observed here.
- The first block, the true coefficient and PaP techniques, has an average correlation of 0.99.
- The second block involving the regression t-statistics, the OOB PaP, and both LOCO techniques has an average correlation of 0.99.
- In both blocks, the permute and drop methods have a near perfect correlation of 1.00.
- The default importance method for random forests, OOB PaP, aligns with the t-statistics, while the other PaP methods align with the coefficients.

We now use an identical regression structure, but increase the error so $\epsilon \sim \mathcal{N}(0, 10)$. This yields a large drop in the regression $R^2 \approx 0.67$. These results are plotted in Figure 5.

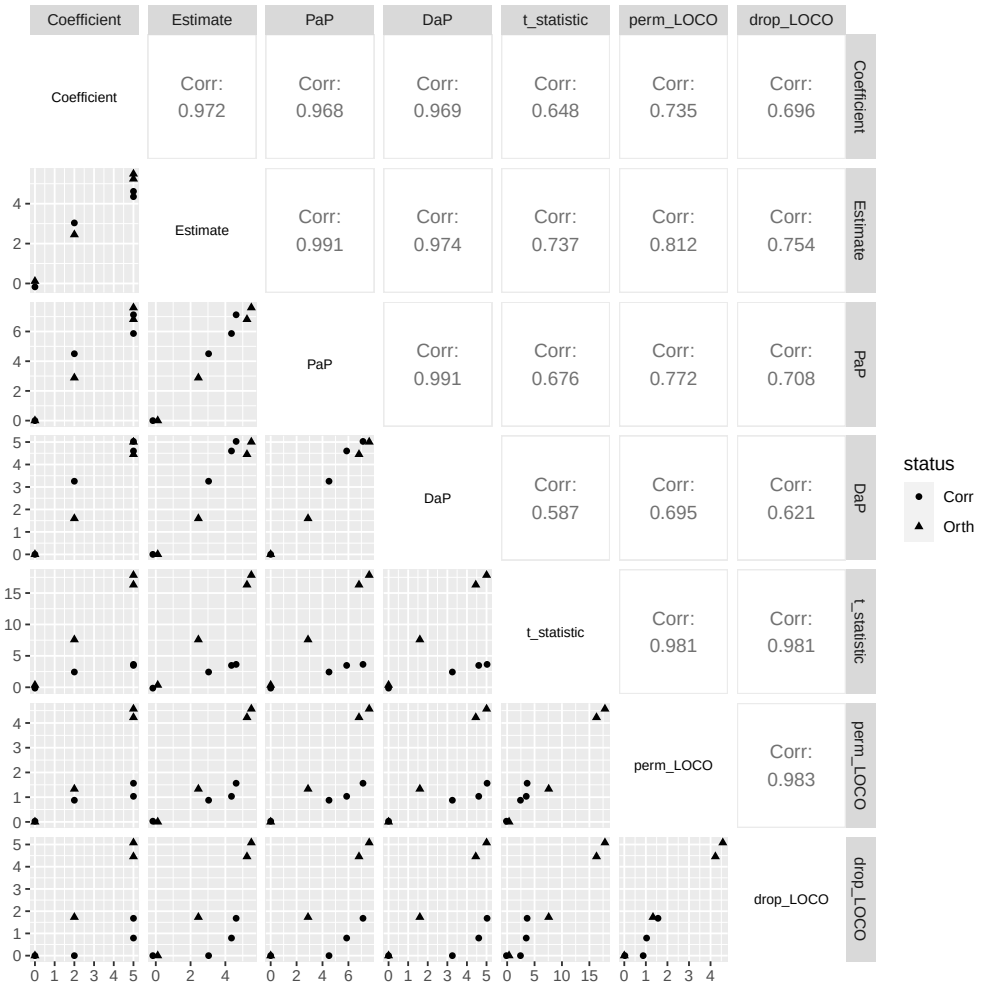

Figure 5. Comparative plots of importance metrics taken from or performed on the linear regression model. Plot symbols denote the status of a given variable: whether it was part of the correlated or orthogonal group of features. The equation for the response is $y = 5Cor_1 + 5Cor_2 + 2Cor_3 + 0Cor_4 + 5v_5 + 5v_6 + 2v_7 + 0v_8 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 10)$. See also Eq (2.2).

From Figure 5, we observe the following when increasing the error for the response variable:

- High and low correlation patterns align with those identified in Figure 3.
- The first block involving the true coefficients, the regression estimates, and both PaP techniques has an average correlation of 0.98.
- The second block, involving t-statistics and LOCO techniques, has an average correlation of 0.98.
- In each block, the permutation and dropout methods have nearly perfect correlations ≈ 0.99.

We now analyze this same dataset utilizing a random forest model. The pairwise plots of importance

metrics are provided in Figure 6. We keep the linear regression estimates and t-statistics from Figure 5, but now we perform the PaP and LOCO methods on the random forest model.

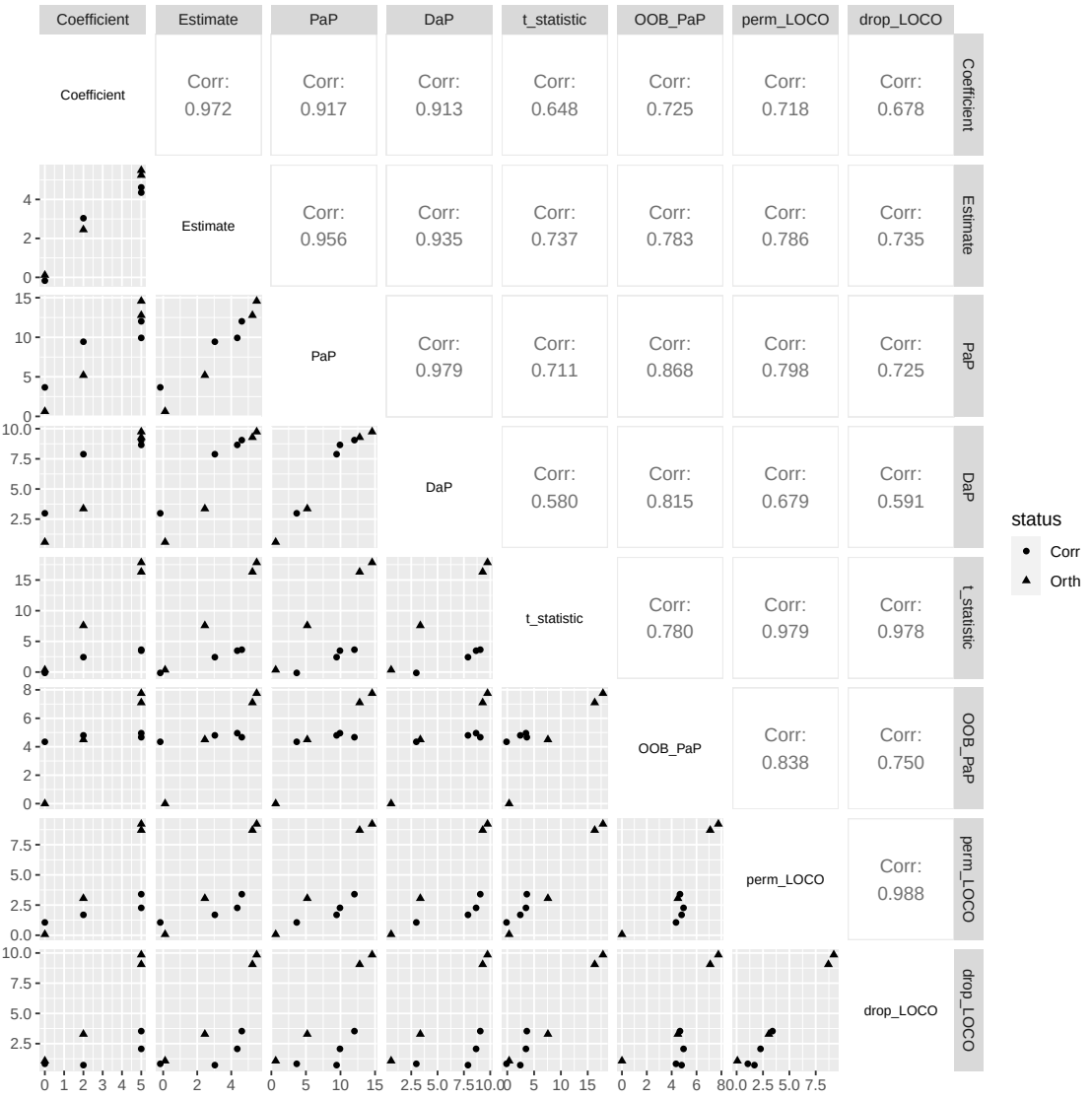

Figure 6. Comparative plots of importance metrics: regression model estimates and t-statistics and PaP and LOCO methods performed on a random forest model. Plot symbols denote the status of a given variable: whether it was part of the correlated or orthogonal group of features. The equation for the response is $y = 5Cor_1 + 5Cor_2 + 2Cor_3 + 0Cor_4 + 5v_5 + 5v_6 + 2v_7 + 0v_8 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 10)$. See also Eq (2.2).

From Figure 6, we observe the following for the high-collinearity data:

- The same blocks of high-similarity metrics are observed again.
- The first block involving the true coefficients, the regression estimates, and both PaP techniques has an average correlation of 0.95.
- The second block involving the t-statistics, the OOB PaP, and both LOCO techniques has an

average correlation of 0.89, but it moves up to 0.98 when OOB PaP is removed.

- In both blocks, the respective permutation and dropout methods have a near perfect correlation of 0.98.
- The default importance method for random forests, OOB PaP, is the least stable metric in terms of alignment with other metrics.

3.3. *Collinearity results*

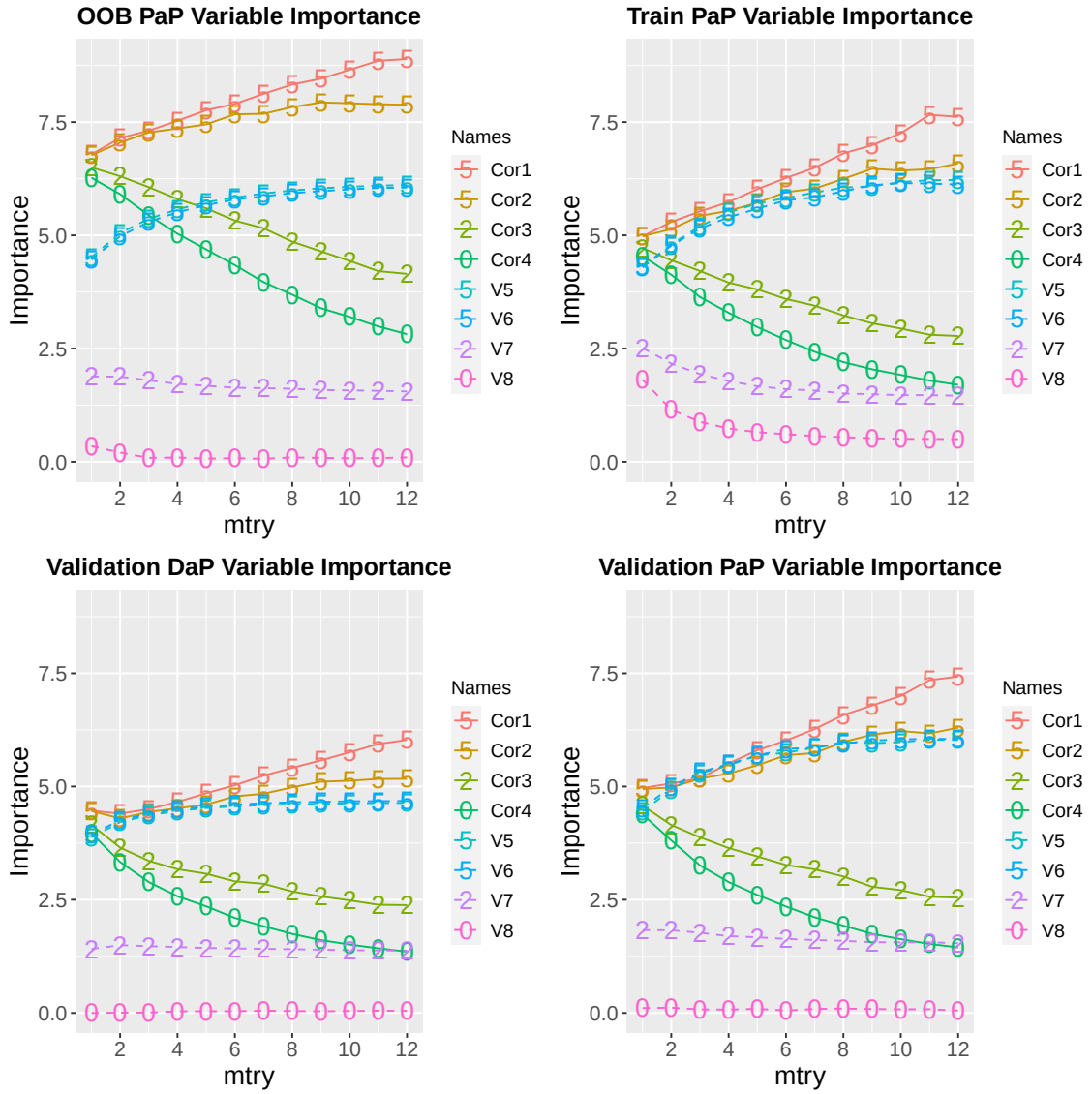

Figure 7. Panel of plots comparing four different variable importance metrics across `mtry`. The equation for the response is $y = 5Cor_1 + 5Cor_2 + 2Cor_3 + 0Cor_4 + 5v_5 + 5v_6 + 2v_7 + 0v_8 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. See also Eq (2.2).

We now offer results exclusively focused on our random forest models. In Section 1.2, we mention the hyper-parameter `mtry` and prior research assessing how variable importance fluctuates based on `mtry` [8]. Moving forward, we discuss plots that show how a few importance metrics change across `mtry`. These metrics include the default OOB PaP, the PaP and DaP metrics (Section 1.2), and a variation of PaP done on the training data instead of our validation data.

We start with our simulation from Section 2.2, with $\epsilon \sim \mathcal{N}(0, 1)$. In Figure 4, we show importance metrics for a default random forest, where `mtry` = 4. Here, we utilize a Monte Carlo simulation of 20 replicates to build random forests for `mtry` = 1, 2, ..., 12.

From Figure 7, we observe the following:

- The OOB plot experiences a distinct shift or bias between the variables possessing coefficients of 5, for all `mtry`.
- The Train PaP experiences a distinct bias such that the importance of a pure noise variable, V8, is well above 0 for all `mtry`, especially for lower values of `mtry`.
- As `mtry` increases, the plots generally trend toward importances that are more proportional with the true coefficients.
- The Validation DaP spread is a bit smaller than the Validation PaP, suggesting slightly more variance in the PaP technique.
- The Validation DaP and PaP plots appear superior to the others due to their lack of bias and shifts in the importance values.

3.4. Squared results

We repeat this assessment for our nonlinear datasets, beginning with the simulation where each feature has a quadratic relation with the response. The simulation is described in Section 2.3. We again utilize a Monte Carlo simulation of 20 replicates to build random forests for all values of `mtry`.

From Figure 8, we observe the following:

- Each plot experiences a shift between the variables possessing coefficients of 5, for all `mtry`.
- The OOB metrics contain a much stronger shift between the variables possessing coefficients of 5 than the other importance metrics.
- The Train PaP again experiences a distinct bias such that the pure noise variable, V8, is well above 0 for all `mtry`.
- The plots again trend toward importances that are more proportional with the true coefficients as `mtry` increases.
- The Validation DaP spread is again slightly smaller than the Validation PaP, suggesting slightly more variance in the PaP technique.
- The Validation DaP and PaP plots again appear superior to the others due to their reduced bias and shifts in the importance values.

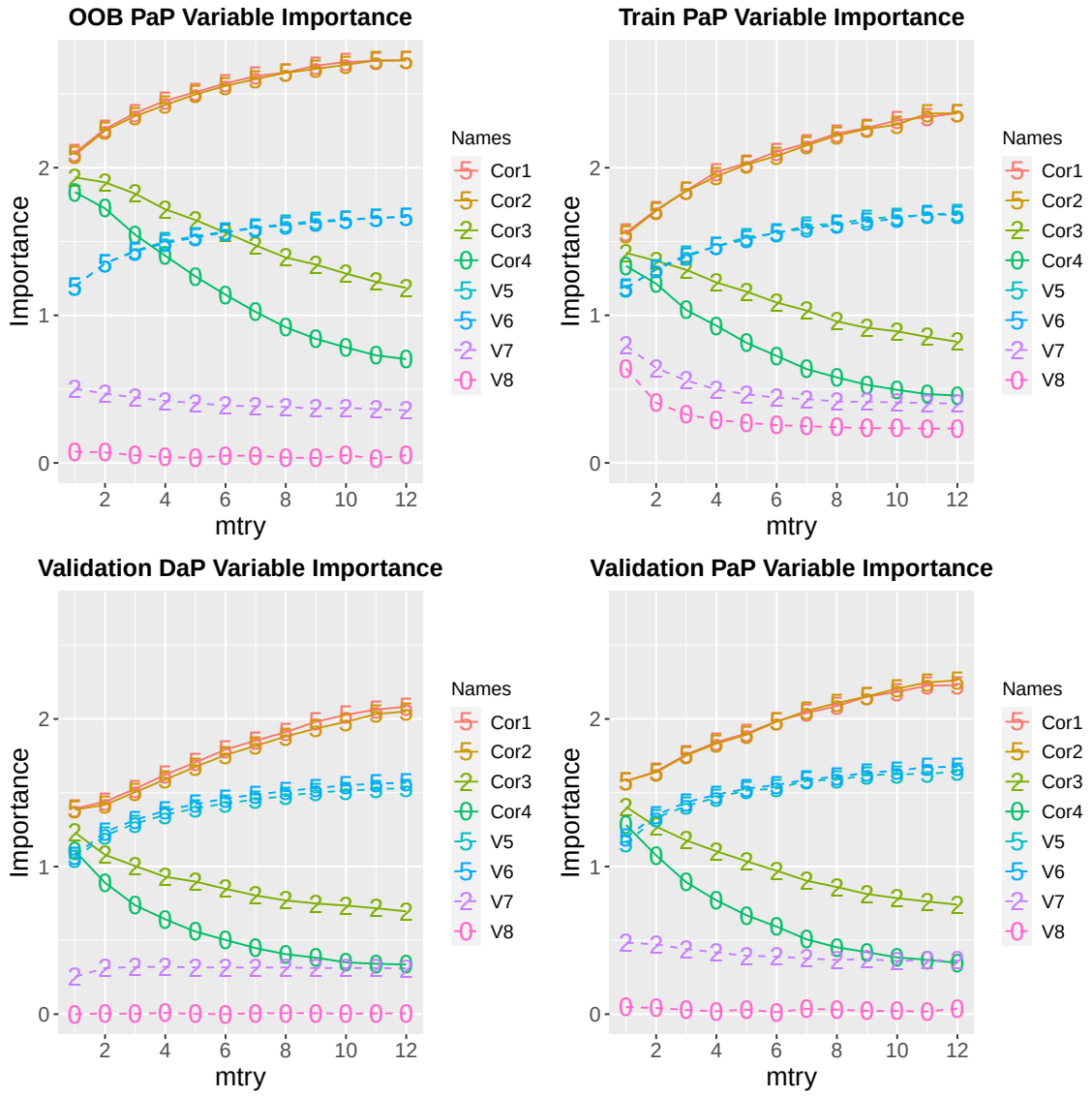

Figure 8. Panel of plots comparing four different variable importance metrics across `mtry`. The equation for the response is $y = 5Cor_1^2 + 5Cor_2^2 + 2Cor_3^2 + 0Cor_4^2 + 5v_5^2 + 5v_6^2 + 2v_7^2 + 0v_{8-12}^2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1)$. See also Eq (2.3).

3.5. Cosine results

We then expand to the structure where each feature has a cosine relation with the response, as described in Section 2.4.

From Figure 9, we observe the following:

- All plots experience a shift between the correlated variables and the orthogonal variables across `mtry`.
- The OOB metrics contain a much stronger shift between the correlated and orthogonal variables than the other importance metrics.

- The Train PaP experiences an enormous positive bias in the importance of V8.
- Each of the four importance metrics has at least a slight bias in the importance of V8, but the Validation DaP consistently has the lowest bias.
- The Validation DaP metric appears superior, especially when compared with the OOB and Train PaP.

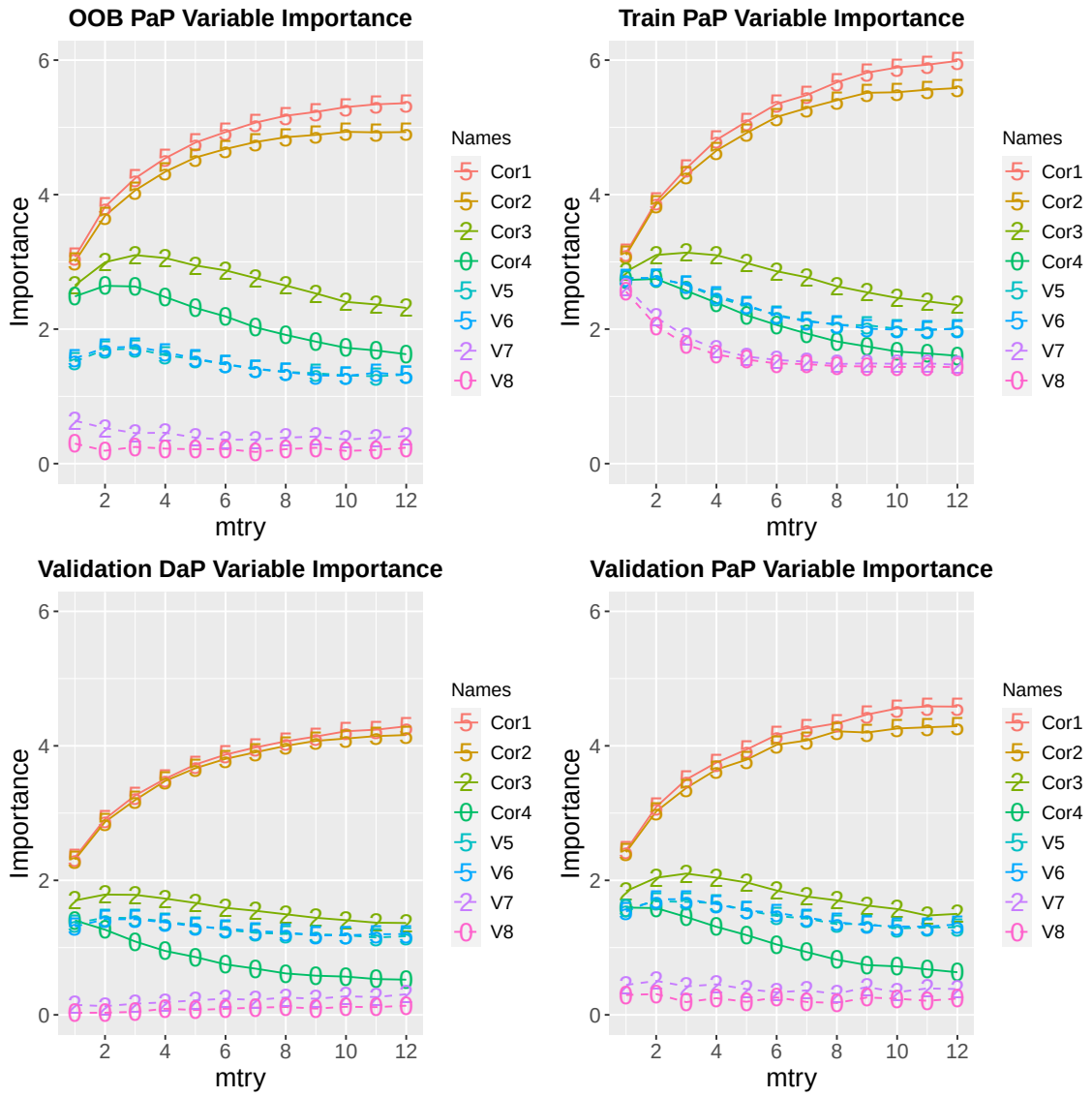

Figure 9. Panel of plots comparing four different variable importance metrics across `mtry`. The equation for the response is $y = 5cos(Cor_1) + 5cos(Cor_2) + 2cos(Cor_3) + 0cos(Cor_4) + 5cos(v_5) + 5cos(v_6) + 2cos(v_7) + 0cos(v_{8-12})$. See also Eq (2.4).

3.6. Interaction results

Finally, we explore the situation where each variable relates to the response through an interaction. The data generation process is shown in Section 2.5.

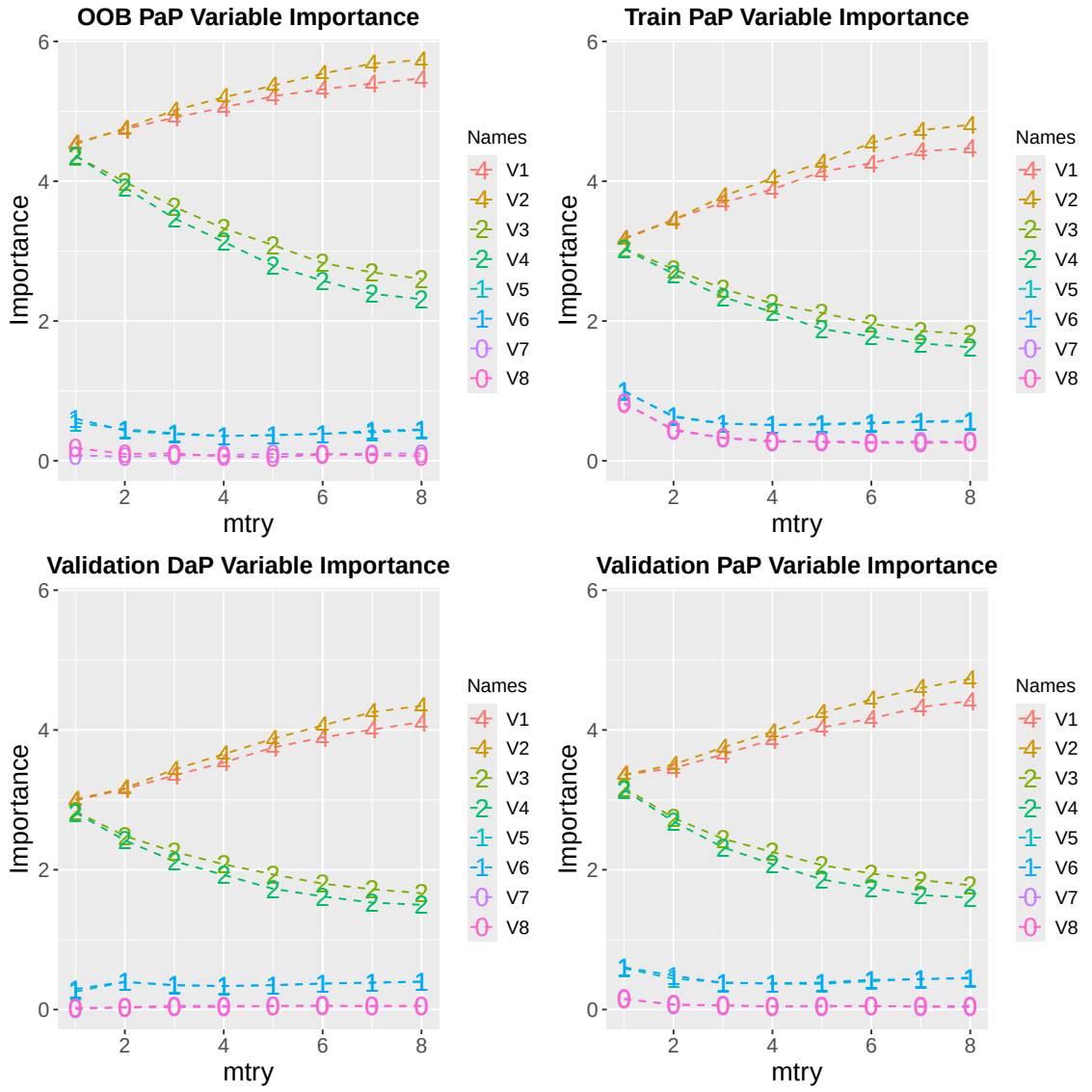

Figure 10. Panel of plots comparing four different variable importance metrics across `mtry`. The equation for the response is $y = 4v_1v_2 + 2v_3v_4 + 1v_5v_6 + 0v_7v_8 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1)$. See also Eq (2.5).

From Figure 10, we observe the following:

- The OOB plot experiences the greatest spread in importance values.
- The Train PaP experiences a distinct positive bias in the importance values of V7 and V8.
- As `mtry` increases, each of the plots generally trends toward importances that are more proportional with the true coefficients.
- The gap between coefficients of 2 and 1 is much larger for the OOB plot than the other plots.
- The Validation DaP spread is just slightly smaller than the Validation PaP.

4. Discussion

Our results highlight some powerful discoveries about permutation and variable deletion techniques. First and most notably, they provide valuable empirical evidence of a very high association between permuting a variable and dropping that variable if the structures are fundamentally equivalent.

This work suggests that t-statistics have high alignment with the LOCO technique for permutation and dropout. Mathematically, $t_j \approx \text{perm\_LOCO}_j \approx \text{drop\_LOCO}_j$. If surrogate t-statistic importances are desired in machine learning, then utilizing a LOCO technique appears to be a reasonable approach to approximate them. However, it is noteworthy that this can often be computationally expensive, especially for large datasets.

Meanwhile, the true coefficients have strong alignment with the regression estimated coefficients and the PaP and DaP methods. This can be expressed as $\beta_j \approx \hat{\beta}_j \approx \text{PaP}_j \approx \text{DaP}_j$. If interest is predominately in predictor relationships with the response rather than marginal predictive capacity, then these metrics would be preferred. Recognizing this relationship, PaP or DaP techniques could be employed to obtain meaningful surrogates to the functional equation coefficients. We especially note from our results that these should be calculated using validation data over training data due to biases in variables with lower importance. If no reasonable validation set is available, the same process might be done using a k-fold cross-validation technique.

Our work also suggests a clear difference between the random forest OOB importance and the validation PaP and DaP importances. This contradicts common intuition that they should be equal or proportional. Further consideration led us to realize that OOB importance assesses the variables in the individual trees and aggregates them, while Validation PaP and DaP assess the variables in the entire forest. If precision of estimating importance values matters, then the Validation PaP or DaP might be preferred over the OOB PaP.

In summary, we have shown that permutation techniques and dropout techniques are approximately equal. We also illustrate that manual PaP and DaP methods are not equal to the OOB PaP. Even more noteworthy, our work shows that equation coefficients and regression estimates have strong associations with manual PaP metrics, while t-statistics have strong associations with LOCO metrics. This implies that some excellent future theoretical work may be available to relate these approaches to mathematically well-defined regression metrics. Ultimately, if a machine learning project desires a certain regression metric, our work suggests which variable importance techniques should be used to obtain an appropriate surrogate.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflicts of interest.

Code access

The code used to generate the data and figures in this report can be found at:
https://github.com/KelvynBladen/reg_Permute_Importance

References

1. W. Kruskal, R. Majors, Concepts of relative importance in recent scientific literature, *Am. Stat.*, **43** (1989), 2–6.
2. C. Achen, *Interpreting and Using Regression*, Sage, **29** (1982).
3. R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. B*, **58** (1996), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
4. H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. B*, **67** (2005), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
5. J. Pratt, Dividing the indivisible: using simple symmetry to partition variance explained, in *Proceedings of the Second International Tampere Conference in Statistics*, (1987), 245–260.
6. L. Breiman, Random forests, *Mach. Learn.*, **45** (2001), 5–32. https://doi.org/10.1023/A:1010933404324
7. C. Strobl, A. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, Conditional variable importance for random forests, *BMC Bioinf.*, **9** (2008), 1–11. https://doi.org/10.1186/1471-2105-9-307
8. K. Bladen, *Contributions to Random Forest Variable Importance with Applications in R*, MS thesis, Utah State University, 2022.
9. G. Hooker, L. Mentch, S. Zhou, Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance, *Stat. Comput.*, **31** (2021), 1–16. https://doi.org/10.1007/s11222-021-10057-z
10. J. Lei, M. G'Sell, A. Rinaldo, R. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression, *J. Am. Stat. Assoc.*, **113** (2018), 1094–1111. https://doi.org/10.1080/01621459.2017.1307116
11. R. Barber, E. Candès, Controlling the false discovery rate via knockoffs, *Ann. Stat.*, **43** (2015), 2055–2085.
12. E. Candès, Y. Fan, L. Janson, J. Lv, Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection, *J. R. Stat. Soc. B*, **80** (2018), 551–577. https://doi.org/10.1111/rssb.12265
13. C. Ye, Y. Yang, Y. Yang, Sparsity oriented importance learning for high-dimensional linear regression, *J. Am. Stat. Assoc.*, **113** (2018), 1797–1812. https://doi.org/10.1080/01621459.2017.1377080
14. D. Apley, J. Zhu, Visualizing the effects of predictor variables in black box supervised learning models, *J. R. Stat. Soc. B*, **82** (2020), 1059–1086. https://doi.org/10.1111/rssb.12377
15. A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin, Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation, *J. Comput. Graphical Stat.*, **24** (2015), 44–65. https://doi.org/10.1080/10618600.2014.907095

16. B. Greenwell, B. Boehmke, A. McCarthy, A simple and effective model-based variable importance measure, preprint, arXiv:1805.04755, 2018.

AIMS Press