Journal of Data Science 23 (1), 243–263
January 2025
DOI: 10.6339/24-JDS1152
Statistical Data Science

Variable Importance Measures for Multivariate Random Forests

Sharmistha Sikdar[1,*], Giles Hooker[2,†], and Vrinda Kadiyali[3,†]

[1]*Tuck School of Business at Dartmouth, Marketing Department, Hanover, NH,USA*
[2] *Wharton School of Business, Department of Statistics and Data Science, U. Pennsylvania, Philadephia, PA, USA*
[3]*SC Johnson College of Business, Marketing Department, Cornell University, Ithaca, NY, USA*

Abstract

Multivariate random forests (or MVRFs) are an extension of tree-based ensembles to examine multivariate responses. MVRF can be particularly helpful where some of the responses exhibit sparse (e.g., zero-inflated) distributions, making borrowing strength from correlated features attractive. Tree-based algorithms select features using variable importance measures (VIMs) that score each covariate based on the strength of dependence of the model on that variable. In this paper, we develop and propose new VIMs for MVRFs. Specifically, we focus on the variable's ability to achieve split improvement, i.e., the difference in the responses between the left and right nodes obtained after splitting the parent node, for a multivariate response. Our proposed VIMs are an improvement over the default naïve VIM in existing software and allow us to investigate the strength of dependence both globally and on a per-response basis. Our simulation studies show that our proposed VIM recovers the true predictors better than naïve measures. We demonstrate usage of the VIMs for variable selection in two empirical applications; the first is on Amazon Marketplace data to predict Buy Box prices of multiple brands in a category, and the second is on ecology data to predict co-occurrence of multiple, rare bird species. A feature of both data sets is that some outcomes are sparse — exhibiting a substantial proportion of zeros or fixed values. In both cases, the proposed VIMs when used for variable screening give superior predictive accuracy over naïve measures.

Keywords *multivariate response problems; multivariate tree-based ensembles; split improvement; variable selection*

1 Introduction

Multivariate random forest or MVRF developed by Segal and Xiao (2011) is a class of tree-based ensembles developed from multivariate regression tree or MVT (Segal, 1992) that can model multivariate responses. A multivariate response is a vector of measurements taken across K (> 1) different variables that are jointly associated with a vector of explanatory variables or covariates (Joe, 1997). Because the joint modeling of multiple response variables accounts for the covariation or co-occurrence observed in the responses across the K different variables, multivariate models determine the covariates or predictors of interest that are jointly associated with the multiple response variables. Examples include ecological studies on geographic co-existence

*Corresponding author. Email: sharmistha.sikdar@tuck.dartmouth.edu.

†This work is part of the first author's dissertation research with the second and third authors as dissertation advisors.

Received October 31, 2023; Accepted August 24, 2024

of multiple species (De'Ath, 2002; Adler et al., 2017), psychological studies on joint measurement of multiple sub-scales of psychological well-being (Miller et al., 2016), and multivariate models of customer behavior in marketing, e.g., page views across multiple websites (Danaher, 2007), website visit duration and purchase spend (Danaher and Smith, 2011) etc. When there is high class imbalance or sparsity (e.g., zero inflation) in some of the correlated responses, multivariate response modeling can be especially useful. In such situations, by jointly modeling multiple responses one can borrow explanatory strength from the less sparse outcomes. Recent research has shown that MVRFs when used to model correlated multivariate responses yield higher predictive accuracy over random forests (or RFs), and other machine learning (ML) methods such as Elastic Net and Kernelized Bayesian Multi-Task Learning (Rahman et al., 2017; Pierdzioch and Risse, 2020).

A critical factor to improving predictive accuracy is to be able to identify predictors and understand their interactions or associations with the response variable (Breiman, 2001). For tree-based ensembles such as RFs and MVRFs, variable importance measures (henceforth, VIMs) can be used to identify a variable's predictive ability and therefore used as a tool for variable selection (Strobl et al., 2007; Ishwaran, 2007). In RFs, the commonly used VIMs are permutation importance based on mean decrease in accuracy (Breiman, 2001) and Gini importance based on mean decrease in node impurity (Friedman, 2001). However, in MVRF, the VIMs in existing software use the naïve measures based on the average frequency with which a variable is used in a tree. Global measures such as permutation importance (Breiman, 2001) or aggregates of local explanations (Ribeiro et al., 2016; Covert et al., 2020) also do not take multivariate responses into account and for which specific extensions need to be estimated. For the purposes of this paper, we restrict our interest to variable importance methods specific to multivariate tree-based ensembles and where naïve metrics remain the only established importance measures.

In this paper, we develop new VIMs for multivariate ensemble methods and specifically for MVRFs. We propose new methods to measure variable importance based on two different split improvement (SI) criteria. Each of the two proposed VIMs scores a variable by first summing the magnitude of SI measured by the respective SI criterion across all node splits where the variable is used within a tree, and then averaging across the forest ensemble. The first SI criterion measures the difference in the mean structure between parent and children nodes. This is a multivariate generalization of least squares where the magnitude of SI is the difference between the sum of squared errors at the parent splitting node and those at the children nodes. We call the measure that uses this SI criterion the mean structure-based SI VIM. The second criterion measures the magnitude of difference in outcomes between left and right children nodes of each response variable at each splitting node that the variable has been used. We call the measure that uses this second SI criterion the outcome difference-based SI VIM. Using the outcome difference SI, a variable can be scored differently in its ability to split the multiple response variables. The outcome difference SI thus generates a vector of importance measures for each variable. Our implementation of MVRF uses the R package '*MultivariateRandomForest*'. The core idea of our project is that a good VIM will more accurately identify the true predictors and thus give a more accurate prediction of the multivariate outcome that we model. We benchmark our proposed VIMs against the naïve measures: the average incidence and average frequency (as used in this R package) with which a variable is used across an ensemble.

To demonstrate the variable selection ability of the proposed VIMs we use a recursive feature elimination (RFE) strategy to eliminate the least important variables (Guyon et al., 2002). The RFE strategy iteratively builds MVRF using bootstrapped sub-samples (Mentch and Hooker, 2016) computes the importance of each variable, and discards the lowest-scored variables. To

generate a baseline score of VIM for variable elimination, we introduce a Gaussian probe or pseudo-variable in the training set at the start of each iteration. The algorithm discards all variables with a VIM score lower than that of the pseudo-variable at the end of the iteration.

We demonstrate the validity of the proposed VIMs in recovering true covariates under four simulated data scenarios that have varying conditions of error correlation and zero inflation in the response. Under the simulated non-sparse scenarios, the proposed methods of variable importance can recover the ranking of the true covariates more accurately than the naïve measures. Under sparsity, both proposed and naïve VIMs show deteriorated performance in identifying true covariates.

We test our proposed VIMs in two distinct empirical applications (marketing and ecology) that require modeling multivariate correlated response outcomes with varying degrees of sparsity in some of the outcomes. Our first application uses Amazon Marketplace data from five product categories to predict the default product prices quoted in the Add to Cart section, called "Buy Box". We jointly model the default product prices of multiple brands within each category to identify a common set of price predictors per category. Our second application is on ecology (e-bird) data provided by the Cornell Lab of Ornithology on self-reported sightings of migrant bird species from amateur birdwatchers as part of the e-bird citizen science program. We predict co-occurrence or joint sightings of a set of five migrant species as a multivariate outcome. In both cases, we use MVRF to leverage anticipated commonalities between response dependence on features. Both the Amazon Marketplace and ecology data sets exhibit responses with sparsity (for Amazon this means for some brands default prices are steady with a near-zero variance; for e-bird, this means bird sightings for some species are close to zero). We compare the comparative ability of the proposed and naïve VIMs to screen features for predictive performance.

In both applications, we implement and test the predictive accuracy of the proposed and extant VIMs of MVRF using our RFE strategy as outlined earlier. We find that our proposed measures of variable importance when applied as a variable selection tool outperform the naïve measures in their prediction accuracy (in terms of mean squared errors or MSEs) and provide a more stable method of variable pruning. Furthermore, we demonstrate uncertainty quantification procedures to determine the stability of the importance scores. The SI-based VIMs developed in this paper make important contributions to research on multivariate models and in particular to multivariate random forests. We have developed an R-package called *MulvariateRandomForestVarImp* (Sikdar et al., 2021), that can be used in conjunction with the *MultivariateRandomForest* R-package to calculate the two proposed SI-based variable importance scores for MVRFs.

The outline of this paper is as follows. In section 2, we discuss the multivariate extension of regression trees and random forests using sub-bagging procedure. In section 3, we discuss the proposed variable importance measures using the SI criteria for the multivariate case. In section 4, we briefly discuss the RFE strategy for variable selection and propose the application of infinitesimal jackknife variance estimator (Wager et al., 2014) to examine the distributional properties of the proposed VIMs for retained features. We discuss the results of the simulation studies in section 5. In section 6, we discuss the robustness of the proposed VIMs using the variable selection procedure on the two data sets. We also suggest uncertainty quantification studies using the proposed VIMs and the corresponding implications. We conclude with limitations and scope for future work.

2 Multivariate Regression Tree and Multivariate Random Forest

In regression analysis, random forests can be applied to build trees where the tree predictor takes on numerical values rather than class labels (Breiman, 2001). An important decision element associated with a tree-based algorithm is determining the split function. The split function at each splitting node of a multivariate regression tree exploits the between-node heterogeneity using mean and covariance for continuous outcomes (Segal, 1992) and entropy for binary response (Zhang, 1998).

As noted in Segal (1992), under the assumptions that the multivariate response vector is continuous with no missing components in the response, the split function of the multivariate regression tree is a generalization of the least squares split function of the univariate case. For the multivariate case, the mean structure-based split function explores the node heterogeneity by using the difference in a generalized sum of squares between the parent node and the children nodes. The covariance structure based split function replaces the sum of squares at each node with the norm of the difference between the sample covariance and hypothesized covariance matrices. Like in the construction of forests for univariate response outcomes, in the multivariate case individual trees are grown and combined to give the multivariate forest prediction.

The multivariate regression tree (MVT) method for panel data is developed as follows: suppose there are K outcome variables observed over N time periods denoted by the matrix $\mathbf{Y} = \{y_1, y_2, \ldots, y_K\}$, where y_k is the $N \times 1$ vector of observations for the kth outcome in the panel. Further, we assume there are P features or covariates in the covariate set $\mathbf{X} = \{X_1, X_2, \ldots, X_P\}$. A tree algorithm proceeds using a two-step approach. At each node of the tree, the algorithm first draws a random subset $L \leqslant P$ of covariates or predictors and examines every allowable split (s) on each predictor variable $(X_l, l = 1, 2, .., L)$. Second, it determines the best predictor-split combination $(X_l, s(X_l))$ and splits the node into left and right children nodes according to whether $X_l < s(X_l)$. In the case of multivariate outcomes, the covariate used in each node split identifies a cluster of homogeneous multiple outcomes. This algorithm proceeds at each child node and continues until a desired tree size has been grown. The covariates can be either continuous or categorical. For continuous variables, each split divides the data into a lower left and a higher right group, and the best split finds the best division between lower and higher data points. For ordered categorical variables, a split divides the categories into two groups, where the covariate values in one group are larger than those in the other. In case of unordered categorical variables, the split divides the two nodes into disjoint sets of categories.

2.1 Multivariate Random Forests (MVRF) Using Sub-Bagging

In our development of MVRFs, we assume a K-dimensional outcome vector denoted by $\mathbf{Y} = \{y_1, y_2, \ldots, y_K\}$ and P features or predictors denoted by the vector $(X_l, l = 1, 2, .., P)$. We split the data into training and testing sets. We use the sub-bagging algorithm (Andonova et al., 2002; Mentch and Hooker, 2016) to bootstrap subsamples of the full training set. The sub-bagging procedure has been found in many applications to outperform traditional bagging (Zaman and Hirose, 2009). We build the multivariate regression trees on the bootstrapped subsamples using the *build_single_tree* function in the *MultivariateRandomForest* R package. The tree prediction is obtained using the *single_tree_prediction* function (see Online Supplement Algorithm A.1).

3 Variable Importance Measures in Multivariate Random Forests

Existing methods for variable importance in MVRF provide only summaries of the use of features: reporting *Incidence*, or the percentage of trees in which that feature occurs at least once, or *Frequency*, the average number of splits that make use of the feature in a tree. The metrics we study below are generalizations of split improvement methods (Friedman, 2001) designed to measure the importance of each split for prediction, where we are both interested in aggregate importance across outcome variables and in understanding relationships on a per-outcome basis. These are the main reference methods for our study. Other alternative methods are model-agnostic tools such as permutation importance (Breiman, 2001), local explanations such as LIME (Ribeiro et al., 2016), and SHAP (Covert et al., 2020); however, these currently have no multivariate analogs. The most immediately applicable measure is permutation importance (or mean decrease in accuracy) but this measure has a different estimand to what we examine. Additionally, the permutation accuracy measure has been critically reviewed for exhibiting substantial statistical bias (Hooker et al., 2021; Verdinelli and Wasserman, 2023). For these reasons, we restrict our examination to metrics that, in common with *Incidence* and *Frequency*, exploit the structure of the trees in a random forest. The split-based metrics we examine here can also be biased based on feature complexity, but this can be corrected by sample splitting (Zhou and Hooker, 2021) which we employ here.

3.1 Extant Variable Importance Measures in MVRF Packages

In canned MVRF packages, the variable importances are measured by either the *Incidence* or *Frequency* of the use of the features in building the multivariate forest. The *Incidence* based VIM calculates the percentage of trees that used a feature in building the ensemble. More formally, the incidence-based VIM for feature m is calculated as

$$\text{Incidence VIM}_m = \frac{B_m}{B}, \tag{1}$$

where B_m is the number of trees that uses the feature m and B is the total number of trees in the ensemble. Therefore, using the incidence-based VIM, the feature that has been used for the highest percentage of trees in an ensemble build gets the top rank. We note here that this measure is fairly coarse – it will not distinguish between features that are reliably used in all trees and can assign importance to features that get randomly included in very deep trees.

The *Frequency* based VIM first calculates the frequency of feature use in a given tree, and then calculates the average frequency across the ensemble. Formally, the frequency based VIM for feature m is given as

$$\text{Frequency VIM}_m = \frac{\sum_{b=1}^{B} F_{m,b}}{B}, \tag{2}$$

where $F_{m,b}$ is the number of nodes in which the decision is based on feature m in tree b and B is the total number of trees in the ensemble build. Thus, compared to the incidence-based VIM, the frequency-based VIM gives higher importance to a feature used more frequently on average across the ensemble. While this provides some more resolution than incidence-based measures, it does advantage features (e.g., continuous variables) with more potential splits: for example, binary features can only be used once on any path between the root node and leaves of a tree.

3.2 Split Improvement Criterion

We develop VIMs for the multivariate case based on the split improvement (SI) criterion, i.e., the objective of maximizing either within-node homogeneity or between-node heterogeneity at each split. This implies that a variable that achieves a higher magnitude of either within-node homogeneity or between-node heterogeneity at a split gets a higher importance. We develop variable importance measures based on SI criterion in two ways: compute the difference in mean structure (i.e., the change in Gaussian likelihood) of parent and children nodes following Segal (1992) and compute the absolute difference in mean outcomes between nodes.

The general procedure to construct the VIMs is as follows. We build an ensemble of trees on the subsamples drawn from the training sample. We overlay the testing set on each tree and calculate the SI at each node split using the test sample. The importance assigned to a variable is equal to the magnitude of the SI obtained at a node split. If a variable is used at multiple splitting nodes in a given tree, the SI at each node is added up across all such splitting nodes to get the importance measure of the variable for that tree. The overall importance measure for the variable is then simply the average of the ensemble.

Algorithm 1 Computing SI based variable importance measures.

Inputs: training and testing sets, x and x^*, subsample size l_N, number of subsamples r_N
for b in 1 to r_N **do**
  Select subsample of size l_N from training set x
  Build tree on subsample b with number of splitting nodes Q_b
  Use tree to predict on testing set x^*
  Initialize VIM vector of dimension $P \times 1$ for tree b as $VIM_0^b = 0$
  for j in 1 to Q_b **do**
    Calculate magnitude of SI for split j in tree b as SI_{bj}
    for m in 1 to P **do**
      if feature m is used for split j in tree b **then**
        $VIM_{0m}^b = VIM_{0m}^b + SI_{bj}$
      end if
    end for
  end for
end for
Average the r_N predictions to obtain final estimate ($\hat{Y}_{N,l_N,r_N}$)
Average the r_N calculations of VIM vector VIM_{N,l_N,r_N}^b to get VIM_{N,l_N,r_N}^*

3.3 Mean Structure Based SI

Segal (1992) defines the mean structure-based split function $\phi_m(s, g)$ as the difference between the within parent node (g) sum of squares and the within children nodes ($g_d, d = L, R$) sum of squares. That is,

$$\phi_m(s, g) = SS(g) - SS(g_L) - SS(g_R), \tag{3}$$

in which

$$SS(g) = \sum_{i=1}^{N} (y - \mu(g))^T V(\theta, g)^{-1} (y - \mu(g)), \tag{4}$$

$$SS(g_d) = \sum_{i=1}^{N} (y_d - \mu(g_d))^T V(\theta_d, g_d)^{-1} (y_d - \mu(g_d)), d = L, R, \tag{5}$$

where g is the parent node and g_d, $d = L, R$ are the children nodes. The multivariate outcome vectors are denoted by y and y_d for parent and children nodes respectively. We define $SS(g)$ and $SS(g_d)$ as the corresponding within-node sum of squares. Further, $\mu(g)$ and $\mu(g_d)$ denote the vectors of mean response outcomes for the parent and children nodes respectively. The covariance matrices at the parent and children nodes are denoted by $V(\theta, g)$ and $V(\theta_d, g_d)$ respectively. The parameters are respectively denoted by θ and θ_d, $d = L, R$. The best split is thus determined as $s^* = \text{argmax}\phi_m(s, g)$. To ensure that $\phi_m(s, g)$ is non-negative, the method restricts the covariance structures as $V(\theta, g) = V(\theta_L, g_L) = V(\theta_R, g_R)$.

To derive the mean structure-based SI importance, we use the formulation as given in equations 3–5 above to quantify the SI contributed by a variable used for a node split. At a given node split of a tree, the corresponding values of the out-of-bag (OOB) sample outcome vectors y^*, y_d^*, $d = L, R$, the OOB sample mean vectors $\hat{\mu}(g)$, $\hat{\mu}(g_d)$, $d = L, R$, and the covariance matrix of the overall OOB residual error $\hat{V}$ are used. We note that $\hat{V}$ will be independent of g only asymptotically since $\hat{\mu}(g)$ is used to calculate the OOB sample residuals from which $\hat{V}$ is derived.

The equivalent OOB sample sum of squares at the parent and children nodes is

$$\hat{SS}(g) = \sum_{i=1}^{N} (y^* - \hat{\mu}(g))^T \hat{V}^{-1} (y^* - \hat{\mu}(g)), \tag{6}$$

$$\hat{SS}(g_d) = \sum_{i=1}^{N} (y_d^* - \hat{\mu}(g_d))^T \hat{V}^{-1} (y_d^* - \hat{\mu}(g_d)), d = L, R. \tag{7}$$

Letting m denote the covariate used in the node split, its corresponding importance measure is then computed from the mean structure-based SI as

$$\text{Mean Structure VIM}_m(g) = \hat{SS}(g) - \hat{SS}(g_L) - \hat{SS}(g_R). \tag{8}$$

3.4 Outcome Difference Based SI

In this method, the SI is defined as the absolute difference in mean outcomes between the left and right children nodes of a split. With a multivariate outcome, this measure results in a vector of absolute difference of the same dimension as the outcome vector $\mathbf{y}$. Similar to the mean structure-based SI in the prior sub-section, we estimate the magnitude of SI on the OOB sample. The importance attributed to the variable m on splitting the k^{th} outcome at the splitting node g is computed as the absolute difference in the corresponding testing sample mean outcomes between left and right nodes as

$$\text{Outcome Difference VIM}_{m,k}(g) = |\hat{\mu}_k(g_L) - \hat{\mu}_k(g_R)|. \tag{9}$$

This metric is given on a per-response basis that captures a differentiation among variables important to different responses.

3.5 SI with Significance Testing of Node Splits

Both the proposed VIMs in sections 3.3 and 3.4 are subject to variance, with many small splits potentially inflating importance scores. We thus propose the following refinement to reduce noise from less reliable splits. We perform a test of significance of each node split and include the SI in the VIM calculation only for the statistically significant ones. Extant measures that are based on SI in the univariate random forest case, e.g., Friedman (2001), do not have this additional step. However, we anticipate that this step may help to separate genuine signal from noise when we are using importance measures as a screening tool.

The Hotelling's T-squared used to choose splits, is given by

$$T^2 = \frac{n_L n_R}{(n_L + n_R)}((\hat{\mu}(g_L) - \hat{\mu}(g_R))^T \hat{V}^{-1}(\hat{\mu}(g_L) - \hat{\mu}(g_R))), \tag{10}$$

where n_d is the number of test samples in daughter node $d = 1, 2$. When the response data follows a multivariate normal distribution, the T^2 statistic is transformed into an F statistic as follows

$$F = \frac{n_L + n_R - K - 1}{K(n_L + n_R - 2)} T^2 \sim F_{K, n_L + n_R - K - 1}. \tag{11}$$

Under multivariate Gaussian assumptions on the response vector, for the null hypothesis $H_0 : \boldsymbol{\mu}(g_L) = \boldsymbol{\mu}(g_R)$ the F statistic given in equation 11 follows an F distribution with K and $n_L + n_R - K - 1$ degrees of freedom. In practice, this distribution is used to test multivariate hypotheses with an asymptotic justification that only requires finite second moments. In the context of our proposed VIMs, evalution using out-of-bag data ensures that $\hat{V}$ is also independent of $(\hat{\mu}(g_l), \hat{\mu}(g_R))$.

For the modifications in the SI-based importance measures discussed above, we include the SIs, as given by equations 8 and 9 only for the splits that are significant using the two-sample F test. For the node splits, where H_0 is not rejected, the importance measure for the corresponding splitting variable takes the value 0. With this modification the general algorithm for the variable importance measure is modified to include the significance testing at each node split (see Online Supplement Algorithm A.2. for pseudo-code). Note that while the outcome difference VIM is given on a per-outcome basis, we threshold based on a test across all outcomes. This is a form of borrowing-strength; by assessing the global importance of a split, we include differences for sparse outcomes if there is evidence from other outcomes that the split is important.

A note of caution here is that some genuine signals might be lost in the process by incorporating only significant splits, thus undervaluing an otherwise important feature. We leave it to the discretion of the researcher to select among the alternative SI-based VIMs, with or without F-test, based on the respective predictive performance in their specific application.

4 Variable Selection and Uncertainty Quantification

4.1 Variable Selection Using Recursive Feature Elimination Strategy

Our proposed RFE strategy is an iterative procedure of forest build, feature or variable elimination based on importance scores and recording of predictive performance (see Online Supplement Algorithm A.3 for pseudo-code). The result of this algorithm is a stochastic backwards elimination; features that do not contribute to predictions will be removed with approximately 50% probability each iteration. Important features have much lower probability of being eliminated

before any given round. To provide a benchmark for feature elimination, we introduce a Gaussian random noise term in the variables list at each iteration of the forest and compute its score. All variables that have scores lower than that of the random noise are dropped at the end of each iteration. The iterative process eventually reaches a steady state where no further pruning happens. In each iteration, we record the test set predictive performance (mean squared error or MSE) of the MVRF with the selected features from the prior round. The optimal iteration is chosen as the one where the process is both at a steady state and the test set predictive performance shows improvement, this latter check acts as a form of regularization.

4.2 Uncertainty Quantification

In addition to the iterative feature elimination, we may also be interested in examining the reliability of the proposed importance measures in variable selection. This can be done by examining the distributional properties of the importance scores of the retained features. The importance measure for a feature can be viewed as a random variable that follows a distribution with mean and variance parameters. We estimate the variance in the tree-wise importance measures for each feature using the Infinitesimal Jackknife (IJ) estimate of variance (Efron, 2014; Wager et al., 2014). As noted in the literature, the IJ estimate is a consistent estimator of the variance parameter. The IJ variance estimate of the importance measure for the mth feature is

$$\hat{V}_m^{IJ} = \sum_{i=1}^{N} Cov(I^b_{i,l_N,r_N}, \text{VIM}^b_{N,l_N,r_N,m})^2, \tag{12}$$

where I^b_{i,l_N,r_N} is the number of times the ith training sample is used in the b^{th} bootstrap subsample of size l_N when r_N subsamples are drawn from the training data of size N. The expression $\text{VIM}^b_{N,l_N,r_N,m}$ is the importance measure of the mth feature computed from the tree generated by the corresponding bootstrap subsample. Like the average importance score for each feature, we compute the IJ variance in the tree-wise importance measure for each retained feature.

5 Simulation Studies

We study the robustness of the proposed VIMs under four simulation scenarios. The study settings differ in terms of assumptions made on the correlation of errors in the multivariate response generation and sparsity, or zero-inflation, in the responses. In all four simulation studies, we construct a $(K \times 1)$ multivariate response vector $\mathbf{y}$ from a specified data generating model; where $K = 4$ and the data generating process has $M' = 5$ explanatory variables. We generate $M'' = 10$ spurious or nonsense covariates as additional columns in the simulated data matrix. Therefore, the first five columns of the overall data matrix X contain the true explanatory variables used in generating the response vector. The variables in the data matrix X consist of binomial (e.g., X_1, X_5), uniform (e.g., X_2, X_4), and Poisson (e.g., X_3) variables. The simulation design for the full list of variables (explanatory and spurious) for the non-sparse and sparse cases is provided in the Online Supplement Table A.1.

The purpose of the simulation studies is to test the variable ranking properties of the proposed importance measures. We view ranking as the most salient metric since it is what is most commonly presented to the user. We would not necessarily expect variable importance measures for random forests to exhibit selection consistency by themselves; which is before the RFE procedure in Section 4.1, and do not emphasize variable selection here.

We generate a training dataset of size $N = 300$. We build $numforest = 10$ multivariate random forests each with $r_N = 3000$ trees. For each forest all 15 ($M'+M''$) variables are then scored based on both the proposed and naïve measures of variable importance. For the proposed VIMs - mean structure-based SI (with and without F test) and outcome difference SI (with and without F test), we compute the measures as given by equations 8 and 9 using the OOB sample. We compute a second set of SI measures using the actual splits made on the training trees. The naïve measures of *Incidence* and *Frequency* are computed based on the individual training trees built within an ensemble. The scores of each VIM are then averaged across the $numforest = 10$ forests. The ranks of the variables are computed based on the average scores for each of the importance measures. The test of robustness of a VIM is provided by the ability to recover the rank ordering of the features, i.e., true explanatory variables should get the highest importance measures. For brevity, we provide detailed results for two of the simulation scenarios. The results of the remaining two simulations are in the Online Supplement, Tables A.2 and A.3.

5.1 Scenario 1: Linear Model with No Sparsity and Uncorrelated Errors

We consider the following data generating process (DGP)

$$y_k = \sum_{m=1}^{5} a_{km} X_m + \epsilon_k, \tag{13}$$

where $\epsilon_k \sim N(0, (var(\sum_{m=1}^{5} a_{km} X_m))/10)$; $k = 1, 2, 3, 4$ and $m = 1, 2, .., 5$. The variance of the error term is chosen so that the signal to noise ratio is 10.

The coefficients of the explanatory variables are given by

$$A = \begin{bmatrix} 1.85 & 0.95 & -0.05 & 0.95 & -0.85 \\ 1.3 & 0.9 & 0.08 & 0.8 & -0.75 \\ 2.45 & 0.8 & 0.09 & 0.95 & -0.9 \\ 1.01 & 0.9 & -0.09 & 0.8 & 0.75 \end{bmatrix},$$

where row k represents the coefficients associated with response y_k and column m represents the contribution of X_m.

We report the rank ordering of the true variables as retrieved by the proposed and extant VIMs, true positive rate (TPR) and false positive rate (FPR) in Table 1. For the proposed measures, we compute the variable rankings as given by the VIMs using the training (i.e., the actual tree splits) and the OOB samples. For the extant measures of frequency and incidence, we compute the ranking based on the training data. The TPR of variable identification, i.e., true explanatory variables in the top 5 ranks, is 100% for all the VIMs. The outcome difference-based VIMs (without and with F-test) is able to best recover the variable ranking, especially on the training data. We note that one of the naïve measures, incidence-based VIM allocates the same rank to three of the five explanatory variables. That is, it fails to distinguish the rank ordering among the explanatory variables.

5.2 Scenario 2: Non-linear Model with Sparse Data and Uncorrelated Errors

We consider a non-linear DGP to create a sparse data scenario designed to mimic the structure of the responses observed in our two case studies. This is specified by $y_k = I_k * \exp(1)$, where I_k is an indicator function generated from the binomial model: $I_k = Binomial(1, P(Logistic(\sum_{m=1}^{5} a_{km}$

Table 1: Variable ranking by naive and proposed VIMs under scenario 1.

				Mean Struc.		.. w/ F-test		Outcome Diff.		.. w/ F-test	
Var.	True rank	Freq.	Incid.	Train	OOB	Train	OOB	Train	OOB	Train	OOB
X_1	1	2	1	2	3	2	3	1	1	1	1
X_2	2	3	4	3	4	3	4	2	3	2	2
X_3	4	1	1	1	2	1	2	3	2	3	3
X_4	3	5	5	5	5	5	5	4	5	4	5
X_5	5	4	1	4	1	4	1	5	4	5	4
TPR		100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FPR		0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Abbreviations: Freq. = Frequency based VIM, Incid. = Incidence based VIM, Mean Struc. VIM = Mean Structure based VIM, Outcome Diff. = Outcome Difference based VIM.

Table 2: Variable ranking by naive and proposed VIMs under scenario 2.

				Mean Struc.		.. w/ F-test		Outcm. Diff.		.. w/ F-test	
Var.	True rank	Freq.	Incid.	Train	OOB	Train	OOB	Train	OOB	Train	OOB
X_1	1	1	1	1	1	1	1	1	1	1	1
X_2	2	2	2	2	**9**	2	**9**	2	**10**	2	**9**
X_3	4	4	**6**	3	5	3	3	5	**6**	5	3
X_4	3	**6**	4	**7**	**8**	**7**	**8**	4	**8**	4	**8**
X_5	5	5	**8**	4	3	4	5	**6**	3	**6**	5
TPR		80%	60%	80%	60%	80%	60%	80%	40%	80%	60%
FPR		10%	20%	10%	20%	10%	20%	10%	30%	10%	20%

Bolded numbers indicate ranks that are lower than those for the spurious covariates.

$X_m + \epsilon_k)))$. The coefficients associated with the explanatory variables under the sparse condition are given by

$$B = \begin{bmatrix} 4.85 & 1.5 & -0.1 & 1.45 & -0.09 \\ 5.3 & 2.01 & -0.08 & 1.02 & -0.07 \\ 4.45 & 1.24 & -0.09 & 1.02 & -0.08 \\ 3.01 & 1.05 & -0.09 & 1.02 & 0.075 \end{bmatrix},$$

where row k represents the coefficients associated with I_k and column m represents the contribution of X_m. Further, like scenario 1, $\epsilon_k \sim N(0, (var(\sum_{m=1}^{5} a_{km} X_m))/10); k = 1, 2, 3, 4$ and $m = 1, 2, .., 5$. All the covariates with the exception of $X_4 \sim Binomial(1, 0.5)$ are generated identically as Scenario 1. We report the results in Table 2.

Under the sparse response scenario, with uncorrelated errors, performance of all the VIMs deteriorate in terms of variable rank ordering. As expected, the performance of the proposed measures is weaker when using OOB samples. However, the comparison is on par with the extant VIMs on the training set. For simulation scenarios 3 and 4 (results not shown here), we

replicate the DGP of scenarios 1 and 2 respectively under correlated errors. We find that the error correlation does not alter the performance results of the VIMs from scenarios 1 and 2.

6 Empirical Application

6.1 Application on Amazon Marketplace Data

Our first empirical application solves a marketing problem using Amazon Marketplace data. Amazon operates on its marketplace both directly as a seller and as a platform owner where it allows other independent sellers (also called third-party or 3p sellers) to sell identical or similar items, e.g., same brand and stock-keeping unit (SKU). On the product page, there is a default price and seller option under a section called "Add to Cart" otherwise known as "Buy Box". The other sellers of the product appear at the bottom of the Buy Box section under "Other Sellers". Thus, on Amazon's marketplace, different sellers can fulfill a customer order, though the Buy Box is the recommended or default option selected by Amazon. For each item, Amazon's proprietary algorithms choose a seller (either itself or a 3p) as the featured seller on its Buy Box. More than 80% of a product's sales are attributed to the Buy Box (Chen et al., 2016). Some of the factors that determine a seller's likelihood to win the Buy Box are price and seller reputation (Chen et al., 2016; Á Gómez-Losada and Duch-Brown, 2019). Amazon and 3p sellers change prices dynamically to win the Buy Box. The Buy Box price can thus get adjusted as the "winning" seller changes or if the winning seller changes its offer price. Amazon and 3p sellers' price changes on a brand are often associated with those of rival brands within a category and with other observed variables on the marketplace, e.g., seller rating, count of answered questions, number of product reviews, etc. (Sikdar et al., 2022). Therefore, the Buy Box prices of multiple brands within a category can be correlated with these observed Marketplace factors. Furthermore, Amazon monitors the prices of rival e-commerce platforms, e.g., Walmart, (Amazon - Price Matching, 2021) with the claim that Amazon strives to maintain low and competitive prices on everything they carry.

We thus use MVRF to jointly model the Buy Box prices of multiple brands within a category on Amazon as a function of the observed factors on its own Marketplace and those of rival Walmart. We compare the performance of the proposed and extant VIMs based on the predictive performance of MVRF when using the respective VIMs to identify the most relevant predictors. We obtained data including Buy Box prices from the product pages of five categories on Amazon – Luggage, Cookware, Video Games, Office Supplies (specifically, printer paper) and Home Cleaning. We scraped data on a six-hourly interval for the period from June 2020 − August 2021. We performed identical scraping for the same period from Walmart. For each category on Amazon, we select a set of brands (multiple SKUs per brand identified by Amazon Standard Identification Number or ASIN) whose Buy Box prices are likely to be correlated (e.g., top selling brands as identified by Amazon Choice or sales rank). Our unit of time is every scraping instance of the SKUs selected for analysis. We model the Buy Box prices observed on each scraping instance of this set of SKUs/brands as a multivariate outcome.

In Table 3, we provide the summary statistics of the Buy Box prices (across the data period) of the selected brands/ SKUs in the multivariate outcome vector for each category. For instance, in the Cookware category, we model a 2×1 outcome vector of Buy Box prices. We examine the frequency of price changes over the study period to determine across and within categories the sparse versus non-sparse price changes. Modeling steady prices is equivalent to modeling sparsity since there is limited variation in the outcome data. The Buy Box prices are steady or change least frequently in Luggage followed by Office Supplies and Cookware. In addition, the

Table 3: Buy box price statistics of representative brands/SKUs.

Category	Brands	Buy Box Price Statistics				
		Mean ($)	SD ($)	Min. ($)	Max. ($)	Price Chgs. (%)
Cookware	Hamilton Beach	35.07	5.71	28.51	71.28	24.7
	Crock Pot	47.24	4.58	44.97	70.00	17.4
Office Supplies	Hammermill	9.61	2.45	5.00	22.40	10.1
	Amazon Basics	12.30	0.86	10.40	14.00	32.6
Luggage	Amer. Tourister	72.60	4.10	70.00	102.00	3.1
	Amazon Basics	59.20	7.59	54.40	75.00	3.4
	Rockland	95.10	7.03	79.00	116.00	6.9
Home Cleaning	Clorox (SKU1)	19.21	1.94	16.94	26.67	51.8
	Clorox (SKU2)	22.47	4.24	16.45	40.00	76.3
	Lysol	14.13	1.27	12.67	18.90	42.4
Video Games	PS (SKU1)	20.10	1.46	18.75	24.99	25.0
	PS (SKU2)	50.45	4.67	39.99	59.99	51.8
	2K (SKU1)	29.78	0.49	26.96	29.99	30.4
	2K (SKU2)	30.45	2.01	29.00	39.99	21.4
	Electronic Arts	28.59	3.75	19.88	37.98	23.2

1. For most brands, we selected only 1 SKU per brand in the outcome vector. For brands in Home Cleaning and Video Games, there were at least two SKUs for some brands that satisfied the criteria for representativeness. In such cases, we modeled the Buy Box prices of both SKUs. The SKUs are denoted in parentheses.
2. The last column records the proportion of days in the tracking period when the Buy Box price of the SKU changed.

Luggage category has a high sparsity of Buy Box price variation across all outcomes. In contrast, the Home Cleaning and Video Games categories have the greatest frequency of Buy Box price changes. However, within each of these categories, the individual outcomes vary in relative Buy Box price change frequency. For instance, in Video Games, PlayStation (SKU2) Buy Box prices change most frequently ($\sim$ 52%) while that of 2K (SKU2) and Electronic Arts have the least frequency of change ($\sim$ 21% and $\sim$ 23% respectively), i.e., sparse cases within the category.

For predictors, we include all observed product-specific characteristics like product rating, sales rank, count of reviews, number of answered queries, whether the item is a best seller, whether the item is labeled Amazon's Choice, and whether it is Prime eligible and in-stock. From the Walmart data, we use comparable variables, e.g., prices of similar or identical SKUs, including star rating, pack size, delivery fee, cut-off for free shipping, etc. For a given daily level data, we summarize these characteristics by taking maximum, minimum, mean, and standard deviation across all past days until the focal day. In our data scraping, we have multiple SKUs scraped for a given brand. For all SKUs of a brand irrespective of their selection in the outcome vector, we use their summary statistics of these above characteristics as predictors in our model.

For each category, we sample 60% of the observed panel as training data and the remaining

40% for testing. We bootstrap 3000 samples and run 20 iterations of the RFE algorithm as discussed in Section 4.1. The predictions on the test set in terms of mean squared error (MSE) from the optimal iteration (i.e., when the process hits steady state) are in Table 4 below.

Table 4: Test set mean squared errors for buy box price prediction (all categories).

Category	Brands	Freq.	Incid.	Mean Struc.	.. w/F-test	Outcm. Diff.	.. w/F-test
Cookware	Hamilton Beach	7.89	8.00	5.82	5.88	5.58	6.05
		(1.54)	(1.53)	(1.39)	(1.35)	(1.35)	(1.42)
	Crock Pot	10.22	10.34	9.81	9.95	9.80	9.96
		(1.59)	(1.60)	(1.62)	(1.62)	(1.63)	(1.62)
Office Supplies	Hammermill	0.62	0.61	0.62	0.64	0.65	0.63
		(0.22)	(0.22)	(0.22)	(0.22)	(0.23)	(0.22)
	Amazon Basics	0.13	0.13	0.13	0.13	0.13	0.13
		(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Luggage	American Tourister	7.69	7.73	7.48	7.40	7.43	7.37
		(4.76)	(4.81)	(4.79)	(4.79)	(4.83)	(4.77)
	Amazon Basics	2.60	2.67	2.17	2.26	2.10	2.09
		(0.54)	(0.55)	(0.52)	(0.52)	(0.53)	(0.53)
	Rockland	11.60	11.61	10.95	10.77	10.97	10.86
		(2.81)	(2.78)	(2.75)	(2.79)	(2.80)	(2.80)
Home Cleaning	Clorox (SKU1)	1.38	1.37	1.41	1.39	1.42	1.36
		(0.15)	(0.16)	(0.15)	(0.16)	(0.15)	(0.16)
	Clorox (SKU2)	11.18	11.25	10.96	11.18	11.24	11.07
		(4.30)	(4.35)	(4.19)	(4.31)	(4.31)	(4.37)
	Lysol	1.31	1.33	1.32	1.32	1.31	1.35
		(0.53)	(0.54)	(0.52)	(0.53)	(0.52)	(0.54)
Video Games	PS (SKU1)	1.09	1.06	1.06	1.04	1.07	1.03
		(0.37)	(0.35)	(0.35)	(0.35)	(0.36)	(0.34)
	PS (SKU2)	27.40	27.55	27.18	27.79	27.38	27.35
		(8.29)	(8.34)	(8.26)	(8.60)	(8.33)	(8.46)
	2K (SKU1)	0.10	0.10	0.09	0.09	0.10	0.10
		(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	2K (SKU2)	0.91	0.88	0.75	0.61	0.67	0.63
		(0.12)	(0.11)	(0.09)	(0.12)	(0.10)	(0.11)
	Electronic Arts	11.19	11.19	11.03	10.73	11.10	11.15
		(4.36)	(4.34)	(4.36)	(4.24)	(4.37)	(4.35)

1) Abbreviations used: Freq. = Frequency based VIM, Incid. = Incidence based VIM, Mean Struc. VIM = Mean Structure based VIM, Outcm. Diff. = Outcome Difference based VIM.
2) Standard Errors reported in parentheses.

The Home Cleaning category has the most non-sparse outcomes across categories. Here, we find our proposed VIMs perform better than or on par with the extant ones, e.g., for Clorox (SKU1), outcome difference with F-test is the top performer and for Clorox (SKU2) mean structure-based VIM. Our proposed VIMs, i.e., mean structure and outcome difference-based VIMs outperform the extant ones in predicting Buy Box prices in the Luggage, Cookware, and Video Game categories. We recall that Luggage has sparsity across multiple outcomes (and highest sparsity across all categories), while Video Games category has a mix of non-sparse, e.g., PlayStation (SKU2) and sparse, e.g., 2K (SKU2) and Electronic Arts, cases in its mix. For both these categories, our proposed VIMs, especially when modified with F-test, perform the best for the sparse outcomes.

6.2 Application on e-Bird Data

Our second empirical application uses an ecology data set provided by the Cornell Lab of Ornithology on observer sightings of migrant bird species. These data were collected as part of the e-bird citizen science program (Fink et al., 2021) which collects data from amateur bird watchers across the globe. These data are then paired with geographic information obtained from satellite imagery (Sullivan et al., 2009; Fink et al., 2020) to produce summaries of topography, land use, and land cover in the local region. The resulting data has been used to monitor biodiversity (Johnston et al., 2021), study inter-specifies competition (Chen et al., 2022), migration responses to climate (Coleman et al., 2020), large-scale changes in avian biomass (Rosenberg et al., 2019), and migratory responses to resource production (Ng et al., 2022). Much of these analyses have been based on models produced on a per-species level (Fink et al., 2010). Here we examine the joint modeling of related species of songbirds to improve predictive performance and our understanding of shared responses to geographic and ecological features.

This data set contains sightings of 25 neo-tropical migrant bird species (warblers and vireos) in the North-East US for the monthly period of June 2016. This data contains 235,036 observer group row entries. Each row entry in the data set contains the count of sightings of each bird species within 0.25 km of the search distance and 0.25 search hours by an observer group. Since our primary objective is to model co-occurrence of multiple species, we remove all row entries that report zero sightings across all 25 species. This reduces the data set size to 27,873 observer group entries. We reduced the number of bird species to include only those that have similar habitat preferences, which allow for co-occurrence of sightings. This gives us a set of five bird species *Setaphaga Americana*, *Setaphaga Petechia*, *Vireo Gilvus*, *Vireo Olivaceus* and *Vireo Solitarius* to model the multivariate co-occurrence outcome.

We define the multivariate response as a 5×1 vector of the count of sightings made by an observer group. We provide the count of sightings and sightings as the percentage of observer entries of the selected species in the reduced data set (27,873 entries) in Table 5. Each observer group entry records a set of observer-specific features, and temporal and ecological factors associated with the sightings. These are the predictors used to model the count of sightings. We have a total of 85 predictor variables in the data set.

For model training, we sample 50% observer entries (14,073) and retain the rest for testing (13,836 entries). From the training set, we bootstrap $r_N = 500$ subsamples of size $l_N = 500$. From the holdout data, we sample 500 entries to construct the testing set. We compare the predictive accuracy of the proposed VIMs by modeling co-occurrence of multiple species. Like in the Amazon application, we use the iterative RFE procedure to build multivariate trees and aggregate into an MVRF. We perform 20 iterations of the RFE algorithm for each of the VIMs

Table 5: Distribution of sightings count by species.

Species	No. Sightings	% Sightings
Vireo Solitarius	1,788	6.4
Setophaga Americana	1,813	6.5
Vireo Gilvus	3,775	13.5
Setophaga Petechia	11,219	40.3
Vireo Olivaceus	14,579	52.3

Table 6: Test set mean squared error for sightings count prediction.

Species	Freq.	Incid.	Mean Struc.	.. w/F-test	Outcm diff.	.. w/F-test
V. Solitarius	0.08	0.08	0.08	0.08	0.08	0.08
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
S. Americana	0.10	0.09	0.08	0.08	0.09	0.08
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
V. Gilvus	0.21	0.22	0.21	0.21	0.20	0.20
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
S. Petechia	1.38	1.39	1.30	1.32	1.28	1.29
	(0.55)	(0.52)	(0.51)	(0.52)	(0.51)	(0.51)
V. Olivaceus	0.65	0.73	0.69	0.70	0.62	0.63
	(0.10)	(0.10)	(0.10)	(0.10)	(0.09)	(0.09)

and record the test set predictions (MSEs) for each iteration. In Table 6, we report the MSE of the optimal iteration.

While the predictive performance of the proposed VIMs in the e-bird data is less powerful than in the Amazon Buy Box data application, we make a few important observations. The e-bird study is equivalent to simulation scenario 2 with a mix of sparse and non-sparse outcomes. We recall that in the simulation scenario 2, both proposed and extant VIMs are equally likely to make erroneous variable selections and therefore may have similar predictive performance for some of the outcomes. First, for the three rare species *V. Solitarius*, *S. Americana* and *V. Gilvus*, the accuracy of sighting predictions is on par for both proposed and naive measures. Second, the predictive accuracy for the remaining two species, *S. Petechia* and *V. Olivaceus* is higher than naive ones using our proposed VIMs, specifically the outcome difference method. This indicates that at their worst, our proposed SI-VIMs perform on par with the naive ones, and at their best, they outperform. These results from the second empirical study provide encouraging validation of the robustness of the proposed SI-based importance measures.

From two different applications, we thus have evidence that when using MVRFs to model multivariate responses, our proposed SI-based measures are likely to outperform the extant naive ones. This suggests that our proposed VIMs can better leverage the multivariate structure of MVRF to "borrow strength" from the associations observed between covariates and non-sparse outcomes to identify variables associated with the sparse ones. This, in turn, implies that our proposed SI-VIMs are more effective than naive ones in variable selection for high-dimensional data when using MVRF.

6.3 Uncertainty Quantification of the Proposed VIMs

In this section, we employ uncertainty quantification methods discussed in Section 4.2 to demonstrate the stability of the proposed SI importance measures in variable selection. For purposes of demonstration, we use the results from the Amazon Marketplace example in the Luggage category. In Table 7, we summarize the top five features across all brands retained by the mean structure-based VIM and the brand-specific top 5 features using the outcome difference measures respectively using the RFE procedure. We apply the IJ estimator for variance as given in Section 4.2 to estimate the variance of the variable importance scores and construct boxplots and confidence intervals (CIs). For brevity, we report these in the Online Supplement Figures A.1 through A.4.

We recall that the mean structure-based SI VIM calculates the difference in the generalized sum of squares among the nodes (parent and children nodes). Thus, a variable assigned a higher score using this measure has a higher ability to split among multiple response outcomes (in this case, Buy Box price predictions of multiple brands). The outcome difference-based SI VIM calculates the outcome-specific absolute difference between the children nodes. Thus, a variable assigned a higher score for a specific outcome is better able to separate responses associated with that outcome. In this example, the top-ranked features identified by the mean structure based VIM include those of competitor brands (e.g., Osprey) not examined as part of the multivariate Buy Box price outcome. In contrast, the outcome-difference identifies own past period price changes as the top-ranked features for the Buy box price prediction of the brand (e.g., for American Tourister's Buy Box price prediction, highest rank features are its own lagged price changes).

Second, the spread or inter-quartile (IQ) range of the box plots and the width of the CIs are determinants of the reliability of the importance measure and rank ordering produced by it. As an overall goal, the variable selection procedure using the proposed measures will be reliable if we can recover the same set of high ranked features and preferably in the same rank order using different samples from the population. A lower IQ range will indicate lower

Table 7: Top five features.

		Outcome Difference		
Rank	Mean Structure	American Tourister	Amazon Basics	Rockland
1	No. answered queries Osprey(1-pd. lag)	Price Am. Tourister (1-pd. lag)	No. answered queries Osprey (1-pd. lag)	Price Rockland (1-pd. lag)
2	Sales rank Osprey (1-pd. lag)	Price Am. Tourister (2-pd. lag)	Price Amazon Basics (1-pd. lag)	Price Rockland (3-pd. lag)
3	Price Amazon Basics (3-pd lag)	Price Am. Tourister (3-pd. lag)	Price Amazon Basics (2-pd. lag)	Price Rockland (2-pd. lag)
4	Price Amazon Basics (1-pd. lag)	Max. Walmart reviews of Am Tourister (1-pd. lag)	No.Amazon Basics reviews (1-pd. lag)	No. answered queries Coolife (1-pd. lag)
5	Price Amazon Basics (2-pd. lag)	Price Rockland (1-pd. lag)	No. Osprey reviews (1-pd. lag)	No. Osprey reviews (1-pd. lag)

variability and higher stability of the importance score of a feature across multiple samples. Shorter CIs will indicate higher precision of the importance scores assigned to a feature. For two closely ranked features we would want the CIs to be non-overlapping to ensure the rank ordering is preserved under multiple sampling scenarios. We find that the mean structure-based SI VIM has shorter IQ range for four out of the top five features in comparison to the outcome difference SI VIM. However, the CIs produced by the mean structure-based SI VIM are broader and less differentiated (see Online Supplement Figure A.1.). For example, the features (*Price of Amazon Basics 1-pd lagged* and *Price of Amazon Basics 2-pds lagged* have overlapping intervals. In contrast, the IQ ranges of the features selected by the outcome difference VIM are higher indicating higher variability. However, the CIs of the outcome specific top five features are more differentiated, i.e., non-overlapping, indicating higher precision (see Online Supplement Figures A.2. through A.4.). This indicates that the variable rank ordering is more likely to create distinct ranks across competing variables when using the outcome difference SI measure.

In conclusion, we propose a set of SI-based VIMs for MVRFs. These proposed VIMs are better than the naive measures available in statistical software, in variable selection ability especially when some outcomes of the multivariate response are sparse. Based on evidence from two different empirical applications, we recommend that if the research goal is to identify features that jointly explain the multivariate response outcome one could use the mean structure-based SI VIM. On the other hand, if the goal is to identify features specific to a response outcome while modeling for a multivariate response, one can employ the outcome difference-based SI VIM. Further, the reliability of the variable ranking may differ based on the measure used. We find that the outcome difference measure gives a more differentiated rank ordering of the features. For interpretation of the underlying relationship between features and outcome, the RFE procedure using either of the proposed importance measures can be used as a pre-processing step in high-dimensional multivariate problems to extract high-ranked features. The extracted features can then be used in standard parametric or non-parametric multivariate regression analysis to investigate the nature of interaction, linearity of relationship, and significance of coefficients in parametric specifications.

7 Conclusion

This paper proposes and examines novel methods of measuring variable importance for variable selection in multivariate random forests. Our proposed methods exploit the split improvement criterion and node heterogeneity in determining the importance scores. We proposed two variable importance measures based on split improvement: mean structure and outcome difference. We demonstrate using two different empirical applications (marketing and ecology) that these proposed measures when used as tools for variable selection give higher predictive accuracy than the naïve measures currently available in canned statistical software like R. Furthermore, we examine the distributional properties of the importance measures developed and discuss the reliability of variable ranking produced by the proposed measures. We propose that the choice of the importance measure will depend on the research goal. The mean structure-based SI VIM isolates predictors that jointly determine the multivariate response. Though more reliable in feature ranking, the outcome difference-based SI VIM isolates outcome specific predictors from a multivariate response model. The proposed measures and the variable selection procedure (RFE strategy) can be applied to reduce features in high-dimensional multivariate response problems. Highly ranked features can then be examined using standard parametric or non-parametric

multivariate regression settings to examine the underlying nature of the relationship between outcomes and predictors.

Future research directions include developing the theoretical properties of these proposed importance measures in the context of multivariate regression trees and ensemble methods. An important methodological extension is to cases where the response vectors can have missing entries. While Segal (1992) explores the modifications to the split function for these exceptions, we do not test the implications of missing data on the proposed importance measures. Another avenue of future research is to examine the variable selection performance of the proposed split improvement-based importance measures for multivariate extensions of other tree-based ensembles like gradient-boosted trees (Friedman, 2001).

Our proposed importance measures for feature extraction in multivariate response models will be useful to researchers in ecology, marketing, economics, computational biology, genomics, and biological statistics. We hope that scholars will continue investigating these multivariate extensions of variable importance measures.

Supplementary Material

In our Online Supplement, we have included pseudo-codes on the MVRF ensemble build using sub-bagging procedure, proposed SI-based VIMs with significant splits, and the proposed RFE strategy of our iterative variable selection method. We have also included the variable choices in the simulation design; box plots and confidence intervals of top features selected by our proposed VIMs from the Amazon application on Luggage category.

Acknowledgement

We thank the Cornell Lab of Ornithology for the e-bird data. We thank our excellent RA Nika Dogonadze for the e-commerce data scraping effort and for helping create and maintain the R-package. We thank the Tuck School of Business for financial support with data scraping.

References

Adler P, Kleinhesselink AR, Hooker G, Teller BJ, Ellner S, Taylor JB (2017). Weak interspecific interactions in a sagebrush steppe: Evidence from observations, models, and experiments. In: *2017 ESA Annual Meeting (August 6–11)*.

Amazon - Price Matching (no date). Amazon - price matching. https://www.amazon.com/gp/help/customer/display.html?nodeId=G9EAYKPV5YYDB8P7. Accessed: 25 August 2021.

Andonova S, Elisseeff A, Evgeniou T, Pontil M (2002). A simple algorithm for learning stable machines. In: *ECAI*.

Breiman L (2001). Random forests. *Machine Learning*, 45: 5–32.

Chen KH, Lin WL, Lin SM (2022). Competition between the black-winged kite and Eurasian kestrel led to population turnover at a subtropical sympatric site. *Journal of Avian Biology*, 10: e03040.

Chen L, Mislove A, Wilson C (2016). An empirical analysis of algorithmic pricing on Amazon marketplace. In: *Proceedings of the 25th International Conference on World Wide Web*.

Coleman T, Mentch L, Fink D, Sorte F, Hooker G, Hochachka W, et al. (2020). Statistical inference on tree swallow migrations with random forests. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 69(4): 973–989.

Covert I, Lundberg SM, Lee SI (2020). Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33: 17212–17223.

Danaher PJ (2007). Modeling page views across multiple websites with an application to Internet reach and frequency prediction. *Marketing Science*, 26(3): 422–437.

Danaher PJ, Smith MS (2011). Modeling multivariate distributions using copulas: Applications in marketing. *Marketing Science*.

De'Ath G (2002). Multivariate regression trees: A new technique for modeling species–environment relationships. *Ecology*, 83(4): 1105–1117.

Efron B (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507): 991–1007.

Fink D, Auer T, Johnston A, Ruiz-Gutierrez V, Hochachka WM, Kelling S (2020). Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications*, 30(3): e02056.

Fink D, Auer T, Johnston A, Strimas-Mackey M, Iliff M, Kelling S (2021). ebird status and trends. cornell lab of ornithology, ithaca, new york.

Fink D, Hochachka WM, Zuckerberg B, Winkler DW, Shaby B, Munson MA, et al. (2010). Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications*.

Friedman JH (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 1189–1232.

Á Gómez-Losada , Duch-Brown N (2019). Competing for Amazon's buy box: A machine-learning approach. In: *Business Information Systems Workshops: BIS 2019 International Workshops, Seville, Spain, June 26–28, 2019, Revised Papers 22* (W Abramowicz, R Corchuelo, eds.), 445–456. Springer.

Guyon I, Weston J, Barnhill S, Vapnik V (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46: 389–422. https://doi.org/10.1023/A:1012487302797

Hooker G, Mentch L, Zhou S (2021). Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31: 1–16. https://doi.org/10.1007/s11222-021-10057-z

Ishwaran H (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 519–537.

Joe H (1997). *Multivariate Models and Multivariate Dependence Concepts*. CRC press, Florida.

Johnston A, Hochachka WM, Strimas-Mackey ME, Ruiz Gutierrez V, Robinson OJ, Miller ET, et al. (2021). Analytical guidelines to increase the value of community science data: An example using ebird data to estimate species distributions. *Diversity and Distributions*, 27(7): 1265–1277. https://doi.org/10.1111/ddi.13271

Mentch L, Hooker G (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(1): 841–881.

Miller PJ, Lubke GH, McArtor DB, Bergeman C (2016). Finding structure in data using multivariate tree boosting. *Psychological Methods*, 21(4): 583. https://doi.org/10.1037/met0000087

Ng WH, Fink D, LaSorte FA, Auer T, Hochachka WM, Johnston A, et al. (2022). Continental-scale biomass redistribution by migratory birds in response to seasonal variation in productivity. *Global Ecology and Biogeography*, 31(4): 727–739. https://doi.org/10.1111/geb.13460

Pierdzioch C, Risse M (2020). Forecasting precious metal returns with multivariate random forests. *Empirical Economics*, 58(3): 1167–1184. https://doi.org/10.1007/s00181-018-1558-9

Rahman R, Otridge J, Pal R (2017). Integratedmrf: Random forest-based framework for integrating prediction from different data types. *Bioinformatics*, 33(9): 1407–1410. https://doi.org/10.1093/bioinformatics/btw765

Ribeiro MT, Singh S, Guestrin C (2016). "why should I trust you?" explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Rosenberg KV, Dokter AM, Blancher PJ, Sauer JR, Smith AC, Smith PA, et al. (2019). Decline of the North American avifauna. *Science*, 366(6461): 120–124. https://doi.org/10.1126/science.aaw1313

Segal M (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, 87(418): 407–418.

Segal M, Xiao Y (2011). Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1): 80–87.

Sikdar S, Hooker G, Kadiyali V (2021). Multivariate random forest variable importance measures r package. https://github.com/Megatvini/VIM/.

Sikdar S, Kadiyali V, Hooker G (2022). Price dynamics on amazon marketplace: A multivariate random forest variable selection approach. Tuck School of Business Working Paper, (3518690).

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1): 1–21. https://doi.org/10.1186/1471-2105-8-1

Sullivan BL, Wood CL, Iliff MJ, Bonney RE, Fink D, Kelling S (2009). ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10): 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006

Verdinelli I, Wasserman L (2023). Feature importance: A closer look at shapley values and loco. arXiv preprint: https://arxiv.org/abs/2303.05981.

Wager S, Hastie T, Efron B (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15(1): 1625–1651.

Zaman F, Hirose H (2009). Effect of subsampling rate on subbagging and related ensembles of stable classifiers. In: *Pattern Recognition and Machine Intelligence: Third International Conference, PReMI 2009 New Delhi, India, December 16-20, 2009 Proceedings 3* (S Chaudhury, S Mitra, CA Murthy, PS Sastry, SK Pal, eds.), 44–49. Springer.

Zhang H (1998). Classification trees for multiple binary responses. *Journal of the American Statistical Association*, 93(441): 180–193. https://doi.org/10.1080/01621459.1998.10474100

Zhou Z, Hooker G (2021). Unbiased measurement of feature importance in tree-based methods. *ACM Transactions on Knowledge Discovery from Data*, 15(2): 1–21. https://doi.org/10.1145/3425637