2015

A Copula Based Approach for Design of Multivariate Random Forests for Drug Sensitivity Prediction

Saad Haider

Raziur Rahman

Souparno Ghosh

Ranadip Pal

PLOS | ONE

RESEARCH ARTICLE

A Copula Based Approach for Design of Multivariate Random Forests for Drug Sensitivity Prediction

Saad Haider[1], Raziur Rahman[1], Souparno Ghosh[2], Ranadip Pal[1]*

1 Department of Electrical and Computer Engineering, Texas Tech University, Lubbock, Texas, United States of America, **2** Department of Mathematics and Statistics, Texas Tech University, Lubbock, Texas, United States of America

* ranadip.pal@ttu.edu

Citation: Haider S, Rahman R, Ghosh S, Pal R (2015) A Copula Based Approach for Design of Multivariate Random Forests for Drug Sensitivity Prediction. PLoS ONE 10(12): e0144490. doi:10.1371/journal.pone.0144490

Editor: Attila Gursoy, Koc University, TURKEY

Received: March 13, 2015

Accepted: November 19, 2015

Published: December 10, 2015

Data Availability Statement: The data used in the manuscript is available from http://www.cancerrxgene.org/. The Matlab implementation of the algorithms can be downloaded from https://github.com/sahaider/copulamrf.

Funding: This work was supported by National Science Foundation Grant CCF 0953366. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Modeling sensitivity to drugs based on genetic characterizations is a significant challenge in the area of systems medicine. Ensemble based approaches such as Random Forests have been shown to perform well in both individual sensitivity prediction studies and team science based prediction challenges. However, Random Forests generate a deterministic predictive model for each drug based on the genetic characterization of the cell lines and ignores the relationship between different drug sensitivities during model generation. This application motivates the need for generation of multivariate ensemble learning techniques that can increase prediction accuracy and improve variable importance ranking by incorporating the relationships between different output responses. In this article, we propose a novel cost criterion that captures the dissimilarity in the output response structure between the training data and node samples as the difference in the two empirical copulas. We illustrate that copulas are suitable for capturing the multivariate structure of output responses independent of the marginal distributions and the copula based multivariate random forest framework can provide higher accuracy prediction and improved variable selection. The proposed framework has been validated on genomics of drug sensitivity for cancer and cancer cell line encyclopedia database.

Introduction

An important goal of systems medicine is to generate genomics informed personalized therapeutic regimes with higher efficacy. The ability of inferred models to accurately predict sensitivity of an individual tumor to a drug or drug combination can assist in designing personalized cancer therapy treatments with expected effectiveness significantly higher than current standard of care approaches. A variety of techniques have been proposed for drug sensitivity prediction based on genetic characterizations. A common approach is to consider a training set of cell lines with experimentally measured genomic characterizations (RNA expression, Protein Expression, Methylation, SNPs etc.) and response to different drugs, and design

supervised predictive models for each individual drug based on one or more genomic characterizations. For instance, statistical tests have been used to show that genetic mutations can be predictive of the drug sensitivity in non-small cell lung cancers [1]. In [2], gene expression profiles have been used to predict the binarized efficacy of a drug over a cell line with an accuracy ranging from 64% to 92%. In [3], a co-expression extrapolation (COXEN) approach was used to predict the drug sensitivity for samples outside the training set with an accuracy of around 82% and 75% in predicting the binarized sensitivity of bladder and breast cancer cell lines respectively. Tumor sensitivity prediction has also been considered as (a) a drug-induced topology alteration [4] using phosphor-proteomic signals and prior biological knowledge of generic pathway and (b) a molecular tumor profile based prediction [1, 5]. Drug sensitivity prediction using an elastic net regression analysis [6] over more than 100,000 genomic features (RNA expression, Mutational status of specific genes and SNPs) was considered in [7]. The correlation coefficients between the predicted and actual sensitivity over 450 cell lines using 10 fold cross validation ranged from 0.08 to 0.76 for different targeted drugs. [8] used a Random Forest (RF) based approach to tumor prediction in the NCI 60 cell lines with performance exceeding multiple existing approaches.

The motivation towards Multivariate random forests originated from our participation in a recent community based effort organized by Dialogue on Reverse Engineering Assessment and Methods (DREAM) project [9] and National Cancer Institute (NCI) that explored multiple different drug sensitivity prediction algorithms applied to a common dataset. More than 40 different approaches were applied and our submission based on Random Forests (RF) that considered the generation of individual models for each drug was a top performer in the challenge [10]. However, sets of drugs can have common targets or paths of action resulting in correlated responses in their sensitivities, which can possibly be utilized to improve the accuracy of the supervised predictive model. Note that the best performing approach in this challenge considered the relationships in the output responses in the form of Bayesian multitask learning [9] and the details of this multi-output regression approach is available at [11]. Multi-drug model has also been pursued by [12] where they have used multi-output regression using neural networks on the Genomics of Drug Sensitivity for Cancer (GDSC) dataset. Since our top performing RF model was ignoring the multi-drug response dependencies, we investigated the extension of the RF framework to Multivariate Random Forests (MRF) that incorporates the relationships between the output sensitivities. The objective of the MRF framework is to generate predictions that minimize error and have a multivariate structure similar to the relationships in the original training output responses. To generate individual multivariate regression trees for the construction of MRF, we altered the node cost function to consist of the weighted sum of the squared differences from the mean (similar to univariate regression tree cost function) and a penalty term to capture the difference between the multivariate relationship in the output responses at the node and the multivariate relationship observed in the original training data. Our initial choice for creating the regression tree node cost was to use Mahalanobis distance square [13], which improved our results as compared to RF approach [14]. The Mahalanobis distance square, being based on the covariance of the output responses, is suitable for scenarios where the relationships between the drug sensitivities is linear but can fail to capture non-linear relationships with low correlation coefficients. With this consideration, this article explores the design of multivariate regression trees that can capture all types of relationship in the output responses.

To capture the multivariate structure present in the output responses, we consider the use of copulas as they can deconstruct a multivariate distribution into its marginal distributions and underlying relationships that are represented by copula functions. We expect that the multivariate distribution of the sensitivities to a drug set will change based on the type of cell lines they

are being applied to but the relationship structure separated from the marginal sensitivity distributions will remain similar. As an example, consider two drugs Gefitinib and Lapatinib that might have higher sensitivities when applied to breast cancer cell lines but lower sensitivities when applied to brain tumor cell lines. Thus, the multivariate distribution representing the sensitivities to the two drugs will appear to be skewed towards higher values for Breast cancer cell lines and skewed towards lower values for Brain tumor cell lines. However, we might observe similar correlation coefficients between the sensitivities for the Breast cancer cell lines and the Brain tumor cell lines as the primary target of both the drugs (EGFR) maintains the relationship. The correlation coefficient is one of the measures of the multivariate structure that will mostly capture linear relationships. However, incorporating the ability to separate the marginal distributions from the multivariate distribution will provide us with a more detailed representation of the underlying associations.

In this article, we discuss the appropriateness of copulas for capturing the multivariate structure in output responses and subsequently propose a cost function utilizing copulas for evaluating multivariate regression tree node splits. The cost function is a weighted combination of (a) the sum of squares of the differences between the node and mean responses and (b) the difference in the empirical copula observed at the node and the copula representing the training samples. We also demonstrate the suitability of the framework in variable selection where it provides higher importance to biologically relevant features as compared to competing approaches.

Note that the generation of the node cost function based on copulas presented in this paper can be considered as a generalization of the Multivariate Random Forest framework based on the square of Mahalanobis distances [13]. The presented approach can be applied to any predictive modeling scenario with multiple interrelated output responses.

The paper is organized as follows: The *Methods* section provides a description of the Random Forest framework with proposed extensions to copula based Multivariate Random Forests including design of the node cost function and an illustrative example. The *Results* section contains the performance of the proposed approach when applied to Genomics of Drug Sensitivity in Cancer database. The *Conclusions* section presents the conclusions of the current study and discusses future directions.

Methods

We first present a description of Random Forest regression followed by extension to Multivariate Random Forest regression utilizing the covariance structure of the data. Subsequently, the concept of Copulas is introduced along with their application in designing node splits for multivariate regression trees.

Random Forest Regression

Random Forest (RF) regression refers to ensembles of regression trees [15] where a set of T unpruned regression trees are generated based on bootstrap sampling from the original training data. For each node, the optimal node splitting feature is selected from a set of m features that are picked randomly from the total M features. For $m \ll M$, the selection of the node splitting feature from a random set of features decreases the correlation between different trees and thus, the average response of multiple regression trees is expected to have lower variance than individual regression trees. Larger m can improve the predictive capability of individual trees but can also increase the correlation between trees and void any gains from averaging multiple predictions. The bootstrap resampling of the data for training each tree also increases the variation between the trees.

Process of splitting a node

Let $x_{tr}(i, j)$ and $y(i)$ $(i = 1, \cdots, n; j = 1, \cdots, M)$ denote the training predictor features and output response samples respectively. At any node η_P, we aim to select a feature j_s from a random set of m features and a threshold z to partition the node into two child nodes η_L (left node with samples satisfying $x_{tr}(I \in \eta_P, j_s) \leq z$) and η_R (right node with samples satisfying $x_{tr}(i \in \eta_P, j_s) > z$).

We consider the node cost as sum of square differences:

$$D(\eta_P) = \sum_{i \in \eta_P} (y(i) - \mu(\eta_P))^2 \tag{1}$$

where $\mu(\eta_P)$ is the expected value of $y(i)$ in node η_P. Thus the reduction in cost for partition γ at node η_P is

$$C(\gamma, \eta_P) = D(\eta_P) - D(\eta_L) - D(\eta_R) \tag{2}$$

The partition γ^* that maximizes $C(\gamma, \eta_P)$ for all possible partitions is selected for node η_P. Note that for a continuous feature with n samples, a total of n partitions needs to be checked. Thus, the computational complexity of each node split is $O(mn)$. During the tree generation process, a node with less than n_{size} training samples is not partitioned any further.

Forest Prediction

Using the randomized feature selection process, we fit the tree based on the bootstrap sample $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ generated from the training data.

Let us consider the prediction based on a test sample $\mathbf{x}$ for the tree Θ. Let $\eta(\mathbf{x}, \mathbf{\Theta})$ be the partition containing $\mathbf{x}$, the tree response takes the form [15–17]:

$$y(\mathbf{x}, \Theta) = \sum_{i=1}^{n} w_i(\mathbf{x}, \Theta) y(i) \tag{3}$$

where the weights $w_i(\mathbf{x}, \Theta)$ are given by

$$w_i(\mathbf{x}, \Theta) = \frac{\mathbf{1}_{\{\mathbf{x}_{tr}(i) \in \eta(\mathbf{x}, \Theta)\}}}{\#\{r : \mathbf{x}_{tr}(i) \in \eta(\mathbf{x}_{tr}(r), \Theta)\}} \tag{4}$$

Let the T trees of the Random forest be denoted by $\Theta_1, \cdots, \Theta_T$ and let $w_i(\mathbf{x})$ denote the average weights over the forest i.e.

$$w_i(\mathbf{x}) = \frac{1}{T} \sum_{j=1}^{T} w_i(\mathbf{x}, \Theta_j). \tag{5}$$

The Random Forest prediction for the test sample $\mathbf{x}$ is then given by

$$\bar{y}(\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) y(i) \tag{6}$$

Multivariate Random Forest (MRF)

Let us now consider the multiple response scenario with output $y(i, k)(i = 1, \cdots, n; k = 1, \cdots, r)$. The primary difference between MRF and RF is in generation of the trees with different node costs $D_m(\eta)$ and $D(\eta)$ [13].

The node cost $D(\eta_P) = \Sigma_{i \in \eta_P}(y(i) - \mu(\psi_P))^2$ for the univariate case is the sum of squares of the differences between the output response and the mean output response for the node. For multivariate case, we would like to use a multivariate node cost that calculates the difference between a sample point and the multivariate mean distribution. One possible measure is the sum of the squares of Mahalanobis Distances [18] as shown next:

$$D_m(\eta_P) = \sum_{i \in \eta_P} (\mathbf{y}(i) - \boldsymbol{\mu}(\eta_P))\Lambda^{-1}(\mathbf{y}(i) - \boldsymbol{\mu}(\eta_P))^T \tag{7}$$

where Λ is the covariance matrix, $\mathbf{y}(i)$ is the row vector $(y(i, 1), \cdots, y(i, r))$ and $\boldsymbol{\mu}(\eta_P)$ is the row vector denoting the mean of $\mathbf{y}(i)$ in node η_P. The inverse covariance matrix (Λ^{-1}) is a precision matrix [19] which is helpful to test conditional dependence between multiple random variables. The Mahalanobis distance square normalizes the output responses by their standard deviations and in case of Λ being diagonal, it represents the normalized Euclidean distance. For bivariate case with covariance Λ, the node cost is increased when the deviations of the two output responses from the mean responses are in opposite directions.

Since, the Mahalanobis distance captures the distance of the sample point from the mean of the node along the principal component axes, it might be unable to capture nonlinear relationships that produces a closer to diagonal Covariance matrix. Thus, our objective is to introduce Copulas to capture the nonlinear multivariate structure.

Copula Description

Copulas can represent the dependence between multiple random variables independent of the marginal distributions. A copula function [20] is used to map the joint cumulative probability distribution in terms of the marginal cumulative probability distributions. Let $\Psi_1, \Psi_2 \ldots \Psi_N$ represent N real valued random variables uniformly distributed on [0, 1]. Copula C: $[0, 1]^N \rightarrow [0, 1]$ with parameter θ is defined as:

$$C_\theta(u_1, u_2 \ldots u_N) = P(\Psi_1 \leq u_1, \Psi_2 \leq u_2 \ldots \Psi_N \leq u_N) \tag{8}$$

The multivariate cumulative probability distribution $F_X(x_1, x_2, \ldots x_N)$ and the marginal cumulative probability distributions $F_i(x_i)$ for ($i \in \{1, 2, \ldots N\}$ are related by Sklar's theorem [20] as follows:

$$F_X(x_1, x_2, \ldots x_N) = C(F_1(x_1), F_2(x_2) \ldots F_N(x_N)) \tag{9}$$

If the marginal cumulative distributions ($F_i(x)$) are continuous, copula C is unique [20].

Some copulas can be parameterized using few parameters, for instance, the clayton copula [21] for bivariate distribution is defined as follows using parameter ξ:

$$C(u_1, u_2; \xi) = (u_1^{-\xi} + u_2^{-\xi} - 1)^{-1/\xi} \quad ; \xi \in (0, \infty) \tag{10}$$

Similarly, the copula characterizing two independent variables will have the form $C(u_1, u_2) = u_1 u_2$. Some other common forms of parameterized copulas include Gaussian Copula [22], Frank Copula [23], student's t-copula [24] and Gumbel copula [25]. However, the standard forms of parameterized copulas may not capture all forms of relationships. We can consider the use of empirical copulas that are estimated directly from the cumulative multivariate distribution. Note that the calculation of empirical copulas will have higher computational complexity than parameterized copulas but they can capture a broad range of relationships. We utilize empirical copulas to represent our multivariate structures.

Node Split Criteria using Copula

As described earlier, the regression tree generation process involves partition of a node into two branches based on optimizing a cost criterion. The node cost for univariate regression trees is given by Eq 1 and the node cost for multivariate regression trees utilizing Mahalanobis distance is shown in Eq 7. The feature and threshold that results in maximum cost reduction for that node is selected for splitting.

We next discuss the design of node cost function based on copulas to capture the output dependencies. We expect that the dependency structure among the samples in a node should be similar to the dependency structure observed in the original training data. Consider the node η_P with N_P samples and let Ψ denote the integral of the difference in the empirical copulas observed at node η_P and the root node (this is same as the empirical copula for the training data). We design the node cost $D_C(\eta_P)$ for a copula based multivariate regression tree as follows:

$$D_C(\eta_P) = D_1 + \alpha D_2 \qquad \text{where} \tag{11}$$

$$D_1 = 6N_P\Psi \qquad \text{and}$$
$$D_2 = \sum_{j=1}^{r} \frac{\sum_{i\in\eta_{P,j}}[y(i,j) - \mu(\eta_{P,j})]^2}{\sigma_j^2} \tag{12}$$

where α denotes a scaling factor determining the relative weight of the two components of the node cost, $\eta_{P,j}$ for $j \in \{1, \cdots, r\}$ denotes the set of jth output responses at node η_P and σ_j^2 for $j \in \{1, \cdots, r\}$ denotes the variance of the jth output response at root node.

We next present the motivation for selecting the weight 6 for the integral of copula distance along with approaches to select the scaling factor α. For maintaining D_1 and D_2 in the same range, we analyzed the range of Ψ as compared to D_2.

Hereafter, the MRF approach that uses copula based node splitting criteria (based on Eq 11) will be termed as *CMRF* and the MRF approach using covaraince based node splitting criteria (based on Eq 7) will be termed as *VMRF*.

Analyzing integral of differences in copulas

We first analyze the upperbound on the integral of the difference between two bivariate copulas and subsequently explore further multivariate copulas. Based on Frechet-Hoeffding bounds [26], any bivariate copula $C(u, v)$ is bounded by the following:

$$C_L(u,v) = max[u + v - 1, 0] \leq C(u,v) \leq min[u,v] = C_U(u,v)$$

Thus for any two copulas $C_1(u, v)$ and $C_2(u, v)$, we have

$$|C_1(u,v) - C_2(u,v)| \leq C_U(u,v) - C_L(u,v) \quad \forall u, v \in [0 \;\; 1]$$

Consequently,

$$\int_{v=0}^{1}\int_{u=0}^{1} |C_1(u,v) - C_2(u,v)|\mathrm{d}u\mathrm{d}v \leq \int_{v=0}^{1}\int_{u=0}^{1} [C_U(u,v) - C_L(u,v)]\mathrm{d}u\mathrm{d}v$$

Using the two diagonals in the unit square ($u = v$ and $u + v = 1$), we can divide the unit square into four triangles where the values of $C_U(u, v)$ and $C_L(u, v)$ are simple functions of u

and v. For region 1, we have $u > v$ and $u + v > 1$ and

$$C_U(u,v) - C_L(u,v) = v - (u + v - 1) = 1 - u$$

For region 2, we have $u > v$ and $u + v \leq 1$ and

$$C_U(u,v) - C_L(u,v) = v - 0 = v$$

Similarly for region 3, we have $u \leq v$ and $u + v \leq 1$ and

$$C_U(u,v) - C_L(u,v) = u - 0 = u$$

And for region 4, we have $u \leq v$ and $u + v > 1$ and

$$C_U(u,v) - C_L(u,v) = u - (u + v - 1) = 1 - v$$

The integral over region 1 is as follows:

$$\iint_{Region1} (C_U(u,v) - C_L(u,v))\, du\, dv$$
$$= \int_{v=0.5}^{1} \int_{u=v}^{1} (1-u) \mathrm{d}u\mathrm{d}v + \int_{v=0}^{0.5} \int_{u=1-v}^{1} (1-u) \mathrm{d}u\mathrm{d}v = 1/24$$

We can likewise show that the value of the integral for each of the there other regions is $\frac{1}{24}$. Thus $\int_{v=0}^{1} \int_{u=0}^{1} (C_U(u,v) - C_L(u,v)) \mathrm{d}u\mathrm{d}v = 1/6$ Thus, the upper bound on the surface integral of the difference between any two bivariate copulas is 1/6. Similarly, if we consider the independent copula $C_I(u, v) = uv$, then we can show that $\int_{v=0}^{1} \int_{u=0}^{1} (C_U(u,v) - C_I(u,v)) \mathrm{d}u\mathrm{d}v = 1/12$. For $n > 2$, we conducted simulations to estimate the value of the integrals which is shown in Table 1.

Thus, since the upper bound of $\int (C_U - C_L)$ lies in the range of 1/6 to 0.21, $D_1 = 6N_P\,\Psi$ will be upper bounded by N_P for a bivariate copula. If the regression tree is unable to reduce the initial variance in each output response, the value of D_2 will be in the range of rN_P as the jth numerator term will be close to $N_P\sigma_j^2$. However, since the regression tree will likely reduce the variance in the output response at nodes further away from the root, the value of D_2 will be much lower than rN_P.

Selection of α

Our previous analysis of Ψ provided a range of the integral difference between two copulas but was unable to provide a weight factor for combining D_1 and D_2 that is optimal in terms of predictive performance. We expect that the behavior of D_1 and D_2 will change significantly for different training datasets and thus we select α based on each training dataset. We next describe two techniques to select the weight factor α to achieve higher prediction accuracy.

Method 1: Evaluating and selecting from a set of α's. This is a straightforward approach where different values of α (we considered 10 values of α spaced between 0.1 to 10) are

Table 1. Integral of Copula Differences for different dimensions.

Dimensions	$\int C_U - C_L$	$\int C_U - C_I$	$\int C_U$
2	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{3}$
3	0.2093	0.1255	0.252
4	0.197	0.1425	0.2072

doi:10.1371/journal.pone.0144490.t001

evaluated and the one with best predictive performance selected. The original training data is sub-divided into secondary training and secondary testing samples. The secondary training samples are used to create MRF models that are used to find prediction of secondary testing samples. This process is repeated for a set of possible values of α. The correlation coefficient between predicted secondary testing samples and original secondary testing samples are recorded and the α corresponding to highest correlation coefficient is selected (α_S). This α_S is then used to create MRF model using the original training samples and tested on original testing samples. For our examples, we have applied 10 fold CV on the original data and thus for each fold of training data, we may select different α. However, for each specific fold, α will be fixed for all the trees generated. The above method increases the computational complexity due to the evaluation of multiple values of α. We next present another approach that attempts to reduce the evaluation of numerous values of α.

Method 2: Pareto Frontier Approach to select α. In this approach, we consider the node cost function minimization from a multi-objective optimization problem perspective where we aim to jointly minimize both D_1 and D_2. From the multi-objective perspective, if we plot the D_2 vs D_1 for all possible feature and threshold combinations for a specific node (if we have n samples at a specific node and m features, the number of partitions to be evaluated is mn), we should select a feature and threshold combination that lies in the Pareto frontier. In other words, we look for solutions that are not dominated by any other solution: for instance if the D_1 and D_2 values for w different feature and threshold combinations are denoted by $\{\epsilon_1(i), \epsilon_2(i)\}$ for $i \in \{1, \cdots, w\}$, a combination i is considered dominated by j if either (a) $\epsilon_1(i) > \epsilon_1(j)$ and $\epsilon_2(i) \geq \epsilon_2(j)$ or (b) $\epsilon_1(i) \geq \epsilon_1(j)$ and $\epsilon_2(i) > \epsilon_2(j)$ is valid. The feature and threshold combinations that are not dominated by any of the other $w - 1$ combinations form the Pareto Frontier. For instance, Fig 1(a) shows an example Pareto frontier (red circles) for the left child node for the first split of a specific tree (the D_1 and D_2 values are denoted by D_{1L} and D_{2L} respectively) generated from a synthetic example described in next section. Similarly, Fig 2(a) shows the Pareto frontier (red circles) for the right child node for the first split of a specific tree (the D_1 and D_2 values are denoted by D_{1R} and D_{2R} respectively).

Our idea is to approximate the Pareto frontier using straight lines and utilize the slope of the lines to design α. Figs 1(b) and 2(b) shows that the Pareto frontier can be approximated by two straight lines: one with slope greater than 1 and another with slope less than 1. Consequently, the value of α can be approximated by the following equation.

$$\alpha = -1/\rho$$

where ρ denotes the slope of the straight line fitted to the Pareto frontier. Thus, we have 2 possible values of α from the very first split of a specific tree. If we prepare scatter-plots for the first split of all the trees for $\alpha > 1$ and $\alpha < 1$, we arrive at plots similar to Fig 3(a) and 3(b) respectively. In this method, we calculate the predictive performance of the two average α's and select the one with the best performance to design the overall forest.

An example to illustrate the appropriateness of Copula based node cost for design of Multivariate Regression Trees

We have observed that multivariate random forests incorporating covariance (Mahalanobis distance square) between output responses is more suitable for predicting output responses with linear relationships as compared to output responses with non-linear relationships. Copula presents a methodology to capture the non-linear dependence relationships between multiple variables; and we anticipate that copula will be suitable for predicting output drug responses with non-linear relationships between them. We next present a synthetic example

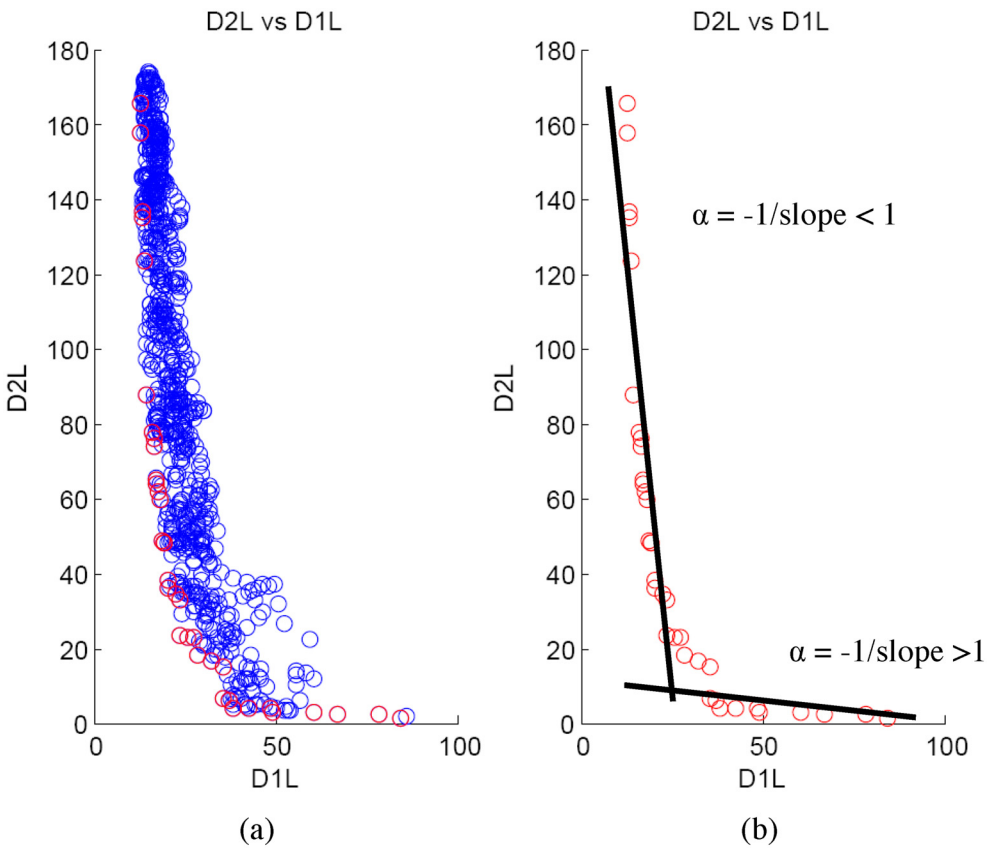


Fig 1. D_2 vs D_1 for left node. (a) shows an example Pareto frontier (red circles) for the left child node for the first split of a specific tree (the D_1 and D_2 values are denoted by D_{1L} and D_{2L} respectively). (b) shows that the Pareto frontier can be approximated by two straight lines: one with slope greater than 1 and another with slope less than 1.

doi:10.1371/journal.pone.0144490.g001

with non-linear relationship to investigate the performance of the proposed approach as compared to covariance based MRF design.

We consider a 50 × 10 input data matrix (50 samples and 10 features) denoted by $\mathbf{X}$ that was created randomly from a normal distribution $\mathcal{N}(0,1)$. We next generated two output responses $\mathbf{Y}_1$ and $\mathbf{Y}_2$ based on functions of the input features. Let column vectors $\mathbf{x}_i (i = 1, 2, \cdots, 10)$ denote the 10 features and the output responses $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are defined as follows:

$$\mathbf{Y}_1 = 2\mathbf{x}_1 + 5\mathbf{x}_2 - 1.5\mathbf{x}_3 + \mathbf{x}_4 \tag{13}$$

$$\mathbf{Y}_2 = (\mathbf{Y}_1 - \mathbb{E}(\mathbf{Y}_1))^2 \tag{14}$$

Note that the output responses are dependent on only 4 features out of the 10 possible input features. Based on the relative weights, $\mathbf{x}_2$ is the most weighted feature and should play a critical role while growing the trees at the beginning. Note that Y_1 and Y_2 has a quadratic relationship.

We consider two multivariate regression trees trained on the same input $\mathbf{X}$ and same output responses $[\mathbf{Y}_1,\mathbf{Y}_2]$ but different node splitting criterion. The regression trees denoted by *Tree* $\{\mathbf{V}, [\mathbf{Y}_1,\mathbf{Y}_2]\}$ and *Tree*$\{\mathbf{C}, [\mathbf{Y}_1,\mathbf{Y}_2]\}$ are inferred using the covariance (Eq 7) and copula (Eq 11) based node cost functions respectively. For this example, all the features were considered at

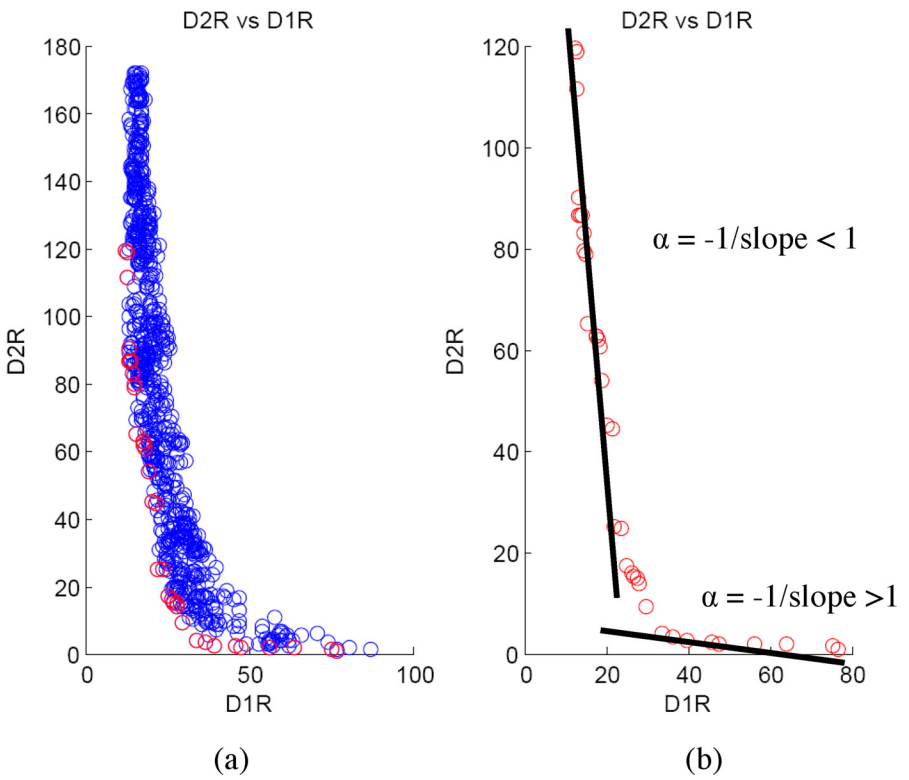


Fig 2. D_2 vs D_1 for right node. (a) shows an example Pareto frontier (red circles) for the right child node for the first split of a specific tree (the D_1 and D_2 values are denoted by D_{1R} and D_{2R} respectively). (b) shows that the Pareto frontier can be approximated by two straight lines: one with slope greater than 1 and another with slope less than 1.

doi:10.1371/journal.pone.0144490.g002

each node i.e. $m = M$ and randomly chosen 80% of 50 samples with bootstrapping were used for generating the regression trees.

Fig 4 shows two multivariate regression trees generated using copula (Eq 11) and covariance (Eq 7) based node cost functions. Fig 4 illustrates that the splitting process for each tree is dependent on different features at each node, which eventually leads to two totally dissimilar trees. The empty circles denote leaf nodes; the circles enclosing a number signify a split node and the number inside the circle indicate the featured selected on that node for splitting.

We expect that the regression tree generation based on copula as compared to covariance will be able to better capture the non-linear relationship between Y_1 and Y_2. Fig 4 demonstrates that the copula based $Tree\{\mathbf{C}, [\mathbf{Y}_1,\mathbf{Y}_2]\}$ has selected the most significant features (features 2 and 1 that have the highest weights during generation of Y_1 and Y_2) while generating the multivariate regression tree. On the hand, the covariance based tree $Tree\{\mathbf{V}, [\mathbf{Y}_1,\mathbf{Y}_2]\}$ trained on the same data selected a spurious feature 7 which was not involved in the generation of either Y_1 or Y_2.

To visually compare the multivariate structure during regression tree splits, we plotted the cumulative distribution functions (CDFs) for the original data and after splitting using copula and covariance based node cost functions. Fig 5 shows the original CDF and the CDFs at the left and right child nodes when the node split is based on Eq 11 (CMRF). Likewise, Fig 6 shows the original CDF and the CDFs at the left and right child nodes when the node split is based on Eq 7 (VMRF). We observe that the node split using copula based node cost better maintains the CDF observed in the original data (Fig 5(b) and 5(c) are similar to (a)) as compared to the split using covariance based node cost (Fig 6(b) is significantly different from (a)).

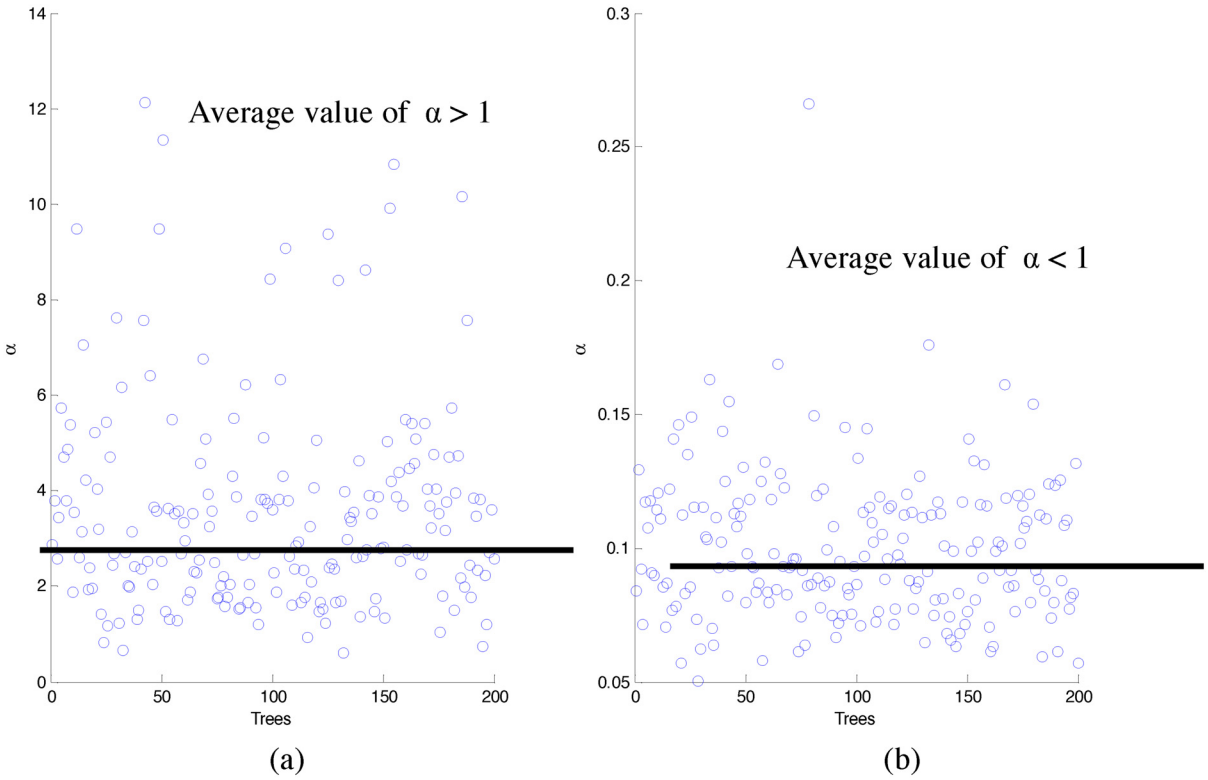


Fig 3. Scatter plot of *α*'s across the trees. (a) and (b) are scatter-plots for the first split of all the trees for $\alpha > 1$ and $\alpha < 1$ respectively.

doi:10.1371/journal.pone.0144490.g003

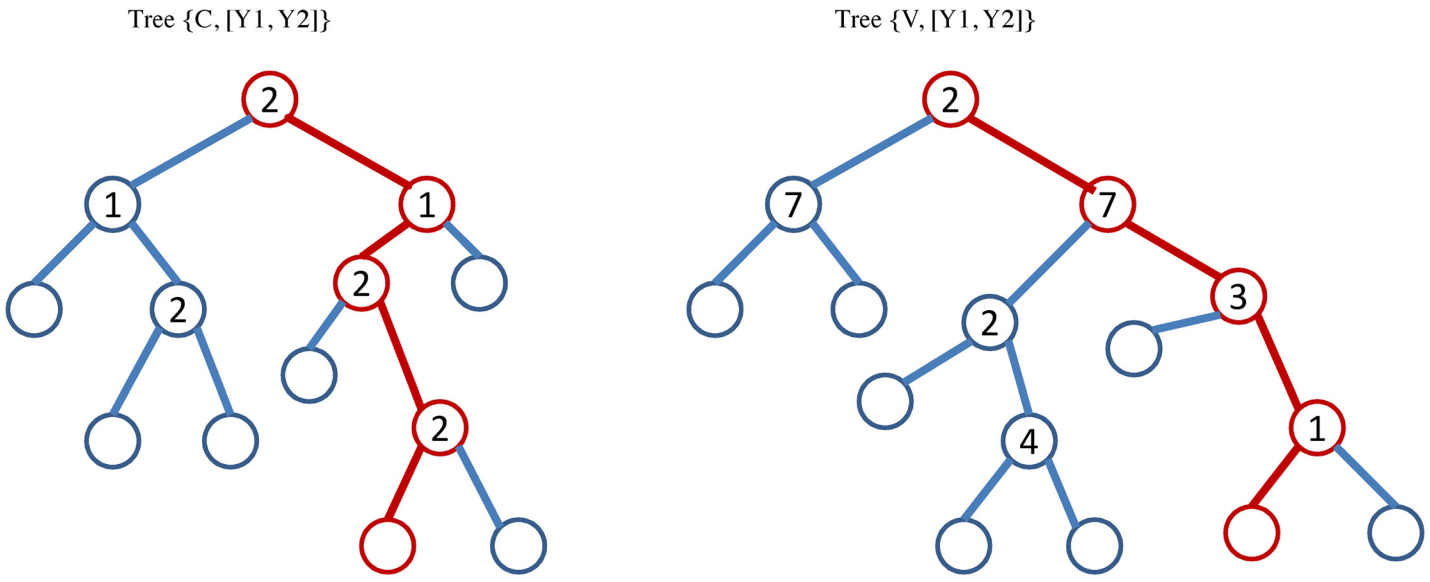


Fig 4. Two multivariate regression trees trained on the same input X and same output responses $[Y_1, Y_2]$ but the node cost criteria being copula based (*Tree*{C, $[Y_1, Y_2]$}) and covariance based (*Tree*{V, $[Y_1, Y_2]$}) respectively. The empty circles represent leaf nodes and the circles enclosing a number signifies a split node; the number inside the circle indicates the featured selected on that node for splitting.

doi:10.1371/journal.pone.0144490.g004

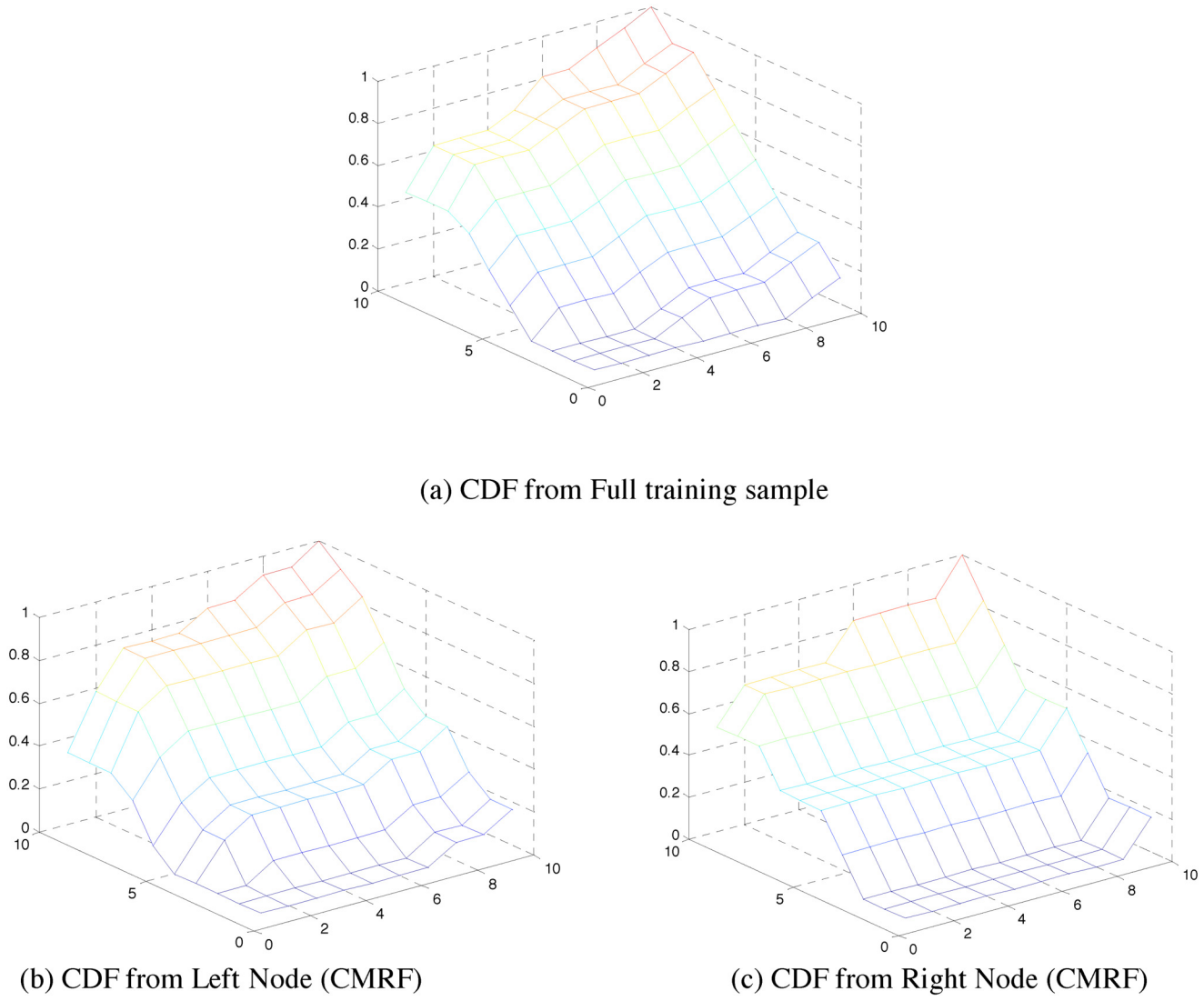
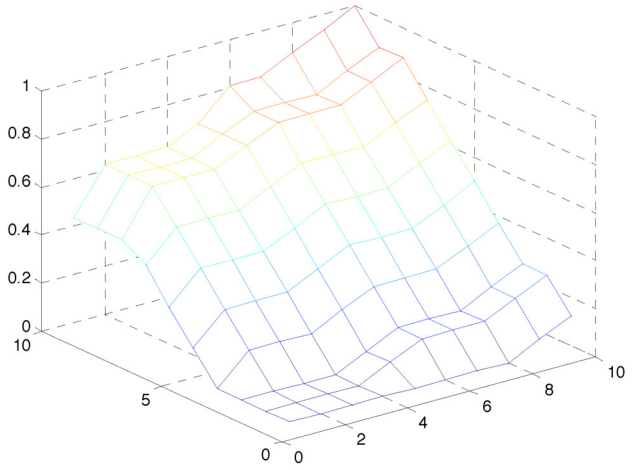


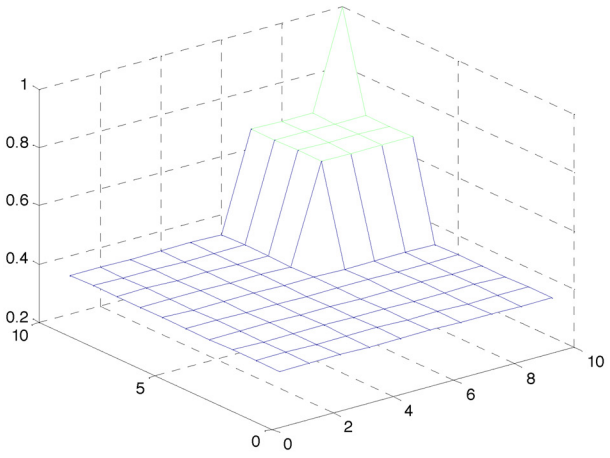
Fig 5. CDF created from left and right child node for a single split using CMRF. It is compared visually with the original CDF created from the training samples.

doi:10.1371/journal.pone.0144490.g005

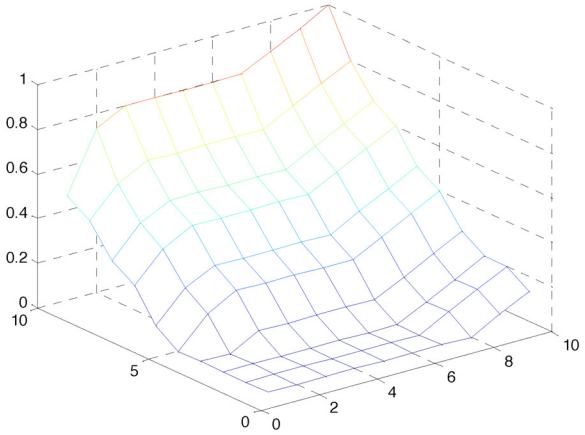
Variable Importance Measure (VIM). In this section, we consider the issue of feature selection for MRFs. We would like to generate and compare the Variable Importance Measure (VIM) for CMRF and VMRF. We expect that CMRF will have higher feature scores for the significant features as compared to VMRF. Typical variable importance measure for random forest considers the frequency of feature selection, out of bag error or permutation measures [27]. We consider the basic approach of calculating the number of times each feature gets selected and the VIM for each forest will be the sum of these frequencies across all trees normalized to the range between 0 and 1. Based on the synthetic data, we generated 100 Multivariate Regression Trees using CMRF (with fixed α) and VMRF with output responses Y_1 and Y_2 and generated the variable importance of the 10 input features. The normalized variable importance scores reported in Table 2 illustrate that the top four features selected by CMRF (X_2, X_1, X_3, X_4) are the same as the four features that were used to generate Y_1 and Y_2 using Eqs 13 and 14 respectively. Furthermore, the ordering of the scores $VIM(X_2) > VIM(X_1) > VIM(X_3) > VIM(X_4)$ is same as



(a) CDF from Full training sample



(b) CDF from Left Node (VMRF)



(c) CDF from Right Node (VMRF)

Fig 6. CDF created from left and right child node for a single split using VMRF. It is compared visually with the original CDF created from the training samples.

doi:10.1371/journal.pone.0144490.g006

the ordering of the absolute weights of the four features in generation of Y_1 where X_2 has the largest weight followed by X_1, X_3 and X_4. On the other hand, the top four features selected by *VMRF* X_2, X_1, X_3, X_6 fails to pick X_4 and includes a spurious feature X_6 that was not involved in the generation of output response Y_1. Thus, the example supports that copula based MRF might be better suitable to select top features as compared to covariance based MRF.

Table 2. Variable importance measure calculated using $CMRF_{Y_1, Y2}$ and $VMRF_{Y_1, Y2}$.

Feature	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
$CMRF_{Y_1, Y2}$	0.2440	0.3720	0.0952	0.0744	0.0179	0.0625	0.0238	0.0417	0.0327	0.0357
$VMRF_{Y_1, Y2}$	0.2340	0.3700	0.0836	0.0529	0.0056	0.0729	0.0418	0.0418	0.0474	0.0501

doi:10.1371/journal.pone.0144490.t002

Results

For analyzing the prediction capabilities of our framework, we considered two different datasets: GDSC and CCLE. Both include genomic characterization of numerous cell lines and different drug responses for each cell line. For the current analysis, we consider the gene expression data as the genomic characterization information for both datasets. Area under the Curves (AUC) is used as representation of drug responses for both GDSC and CCLE. Both datasets are high dimensional in the number of features (gene expressions). For all performance comparison results presented in this article, a prior feature selection method (RELIEFF [28]) is applied to reduce the number of features to be used for training. A performance comparison of random forest approaches with and without the application of prior feature selection is shown for GDSC dataset in Table A in S1 File.

For performance comparison purposes, we report results of Copula based MRF (CMRF) along with univariate RF (denoted by RF), Covariance based MRF (VMRF) and Kernelized Bayesian Multitask Learning (KBMTL) [11] approaches. KBMTL is Bayesian formulation that combines kernel based non-linear dimensionality reduction and regression in a *multitask learning* framework, that tries to solve distinct but related tasks jointly to improve overall generalization performance. We have implemented KBMTL using the algorithmic code provided in [11]. Based on the parameters used in [11], we have considered 200 iterations and gamma prior values (both α and β) of 1. Subspace dimensionality has been considered to be 20 and the standard deviation of hidden representations and weight parameters are selected to be the default 0.1 and 1 respectively.

Results on GDSC Dataset

The GDSC gene expression and drug sensitivity dataset was downloaded from Cancerrxgene.org [29]. The dataset has 789 cell lines with gene expression data and 714 cell lines with drug response data. We considered the intersection of cell lines that had both drug response and gene expression data.

For our experiments, we consider four sets of drug pairs where three of them have common primary targets and the remaining pair has no common target. We expect that the drug pairs with common primary targets will have some form of relationship among their sensitivities and CMRF should perform better than VMRF and both should perform better than RF approach. On the other hand, the drug pair without any common targets is expected to have minimal relationship among the drug sensitivities and thus RF is expected to outperform CMRF and VMRF. We also present results on 3 drug set and 138 drug set for GDSC as Tables C, D and E in S1 File.

Initially each cell line has 22277 features (probeset) as gene expressions. We have reduced it to 500 for each drug response using RELIEFF [28] and used a union of the 500 features in each of the four sets of drugs.

The first selected set S_{C2} consisting of {*Erlotinib*, *Lapatinib*} has common target *EGFR*[30–32]. The second set S_{C3} consisting of {*AZD-0530*, *TAE-684*} has common target *ABL1*[32]. The third set S_{C1} was {*AZD6244*, *PD-0325901*} with common target *MEK*[32–34]. The fourth set S_U consisting of {*17-AAG, Erlotinib*} has no common target.

As mentioned earlier, each drug has some missing responses across the 714 cell lines. The drug sets S_{C1}, S_{C2}, S_{C3} and S_U have drug responses in both drugs for 316, 349, 645 and 300 cell lines respectively. To report our results, we compared 5 fold cross-validated Pearson correlation coefficients, Mean Absolute Error (MAE) and Normalized Root Mean Square Error (NRMSE) between predicted and experimental responses for RF, VMRF, CMRF and KBMTL.

Table 3. 5 fold CV results for GDSC Dataset drug sensitivity prediction for four drug sets in the form of correlation coefficients. *VMRF*, *CMRF* represent Multivariate Random Forest using Covariance and Copula respectively. *KBMTL* represents Kernelized Bayesian multitask learning (Parameters considered are 200 iterations, $\alpha = \beta = 1$ and subspace dimensionality = 20).

			Correlation Co-efficients			
Drug Set	**Common Target**	**Drug Name**	***RF***	***VMRF***	***CMRF***	***KBMTL***
S_{C1}	EGFR	Erlotinib	0.5156	0.5193	0.5301	0.2500
		Lapatinib	0.5544	0.5742	0.5699	0.1132
S_{C2}	ABL1	AZD-0530	0.3553	0.3810	0.3990	0.3181
		TAE-684	0.4060	0.4100	0.4338	0.2420
S_{C3}	MEK	AZD6244	0.4625	0.4508	0.4590	0.0950
		PD-0325901	0.5890	0.6022	0.6016	0.3236
S_U	None	17-AAG	0.6304	0.6244	0.6167	0.4375
		Erlotinib	0.5859	0.5906	0.5708	0.4081

doi:10.1371/journal.pone.0144490.t003

NRMSE of drug m can be calculated as [11]:

$$NRMSE_m = \sqrt{\frac{(y_m - \hat{y}_m)^T (y_m - \hat{y}_m)}{(y_m - \mathbf{1} \cdot \mathbf{E}(y_m))^T (y_m - \mathbf{1} \cdot \mathbf{E}(y_m))}} \quad (15)$$

where y_m and $\hat{y}_m$ denote the vector of actual and predicted drug sensitivities respectively and $\mathbf{E}(y_m)$ denote mean of vector y_m. For both VMRF and CMRF, we set the minimum size of samples in each leaf to $n_{size} = 5$, the number of trees in the forest to $T = 150$ and the splitting in each node considers $m = 10$ random features.

The correlation coefficients using 5 fold cross validation error estimation are illustrated for each drug set in Table 3. The corresponding MAE and NRMSE behaviors are illustrated in Table 4.

For CMRF, results with scaling factor α selected using *Method-1* discussed earlier has been used. The robustness analysis of α using synthetic data is conducted using *Method-2* and is shown as Tables H and I in S1 File. Table 3 shows that CMRF outperformed (in terms of correlation coefficients) VMRF, RF and KBMTL for the related drug pairs S_{C1}, S_{C2}, S_{C3} whereas CMRF is outperformed by the other approaches for the unrelated drug pair S_U. Table 4 shows that *CMRF* outperforms VMRF, KBMTL and RF in terms of average NRMSE for the related pairs of drugs S_{c_1}, S_{c_2} and S_{c_3}. For the unrelated pair S_U, univariate RF outperforms the

Table 4. 5 fold CV results for GDSC Dataset drug sensitivity prediction for four drug sets in the form of MAE and NRMSE for RF, VMRF, CMRF and KBMTL approaches.

			MAE				NRMSE			
Drug Set	**Common Target**	**Drug Name**	***RF***	***VMRF***	***CMRF***	***KBMTL***	***RF***	***VMRF***	***CMRF***	***KBMTL***
S_{C1}	EGFR	Erlotinib	0.0319	0.0322	0.0314	0.0503	0.8733	0.8749	0.8719	1.3365
		Lapatinib	0.0292	0.0294	0.0286	0.0488	0.8516	0.8459	0.8486	1.3538
S_{C2}	ABL1	AZD-0530	0.0446	0.0448	0.0442	0.0613	0.9407	0.9378	0.9291	1.2344
		TAE-684	0.0829	0.0829	0.0821	0.1159	0.9285	0.9299	0.9195	1.3698
S_{C3}	MEK	AZD6244	0.0584	0.0590	0.0584	0.1138	0.8949	0.9034	0.8962	1.8016
		PD-0325901	0.0723	0.0727	0.0717	0.1199	0.8263	0.8230	0.8193	1.4028
S_U	None	17-AAG	0.0584	0.0590	0.0584	0.1198	0.7840	0.7894	0.7955	1.1624
		Erlotinib	0.0723	0.0727	0.0717	0.0410	0.8335	0.8441	0.8505	1.1013

doi:10.1371/journal.pone.0144490.t004

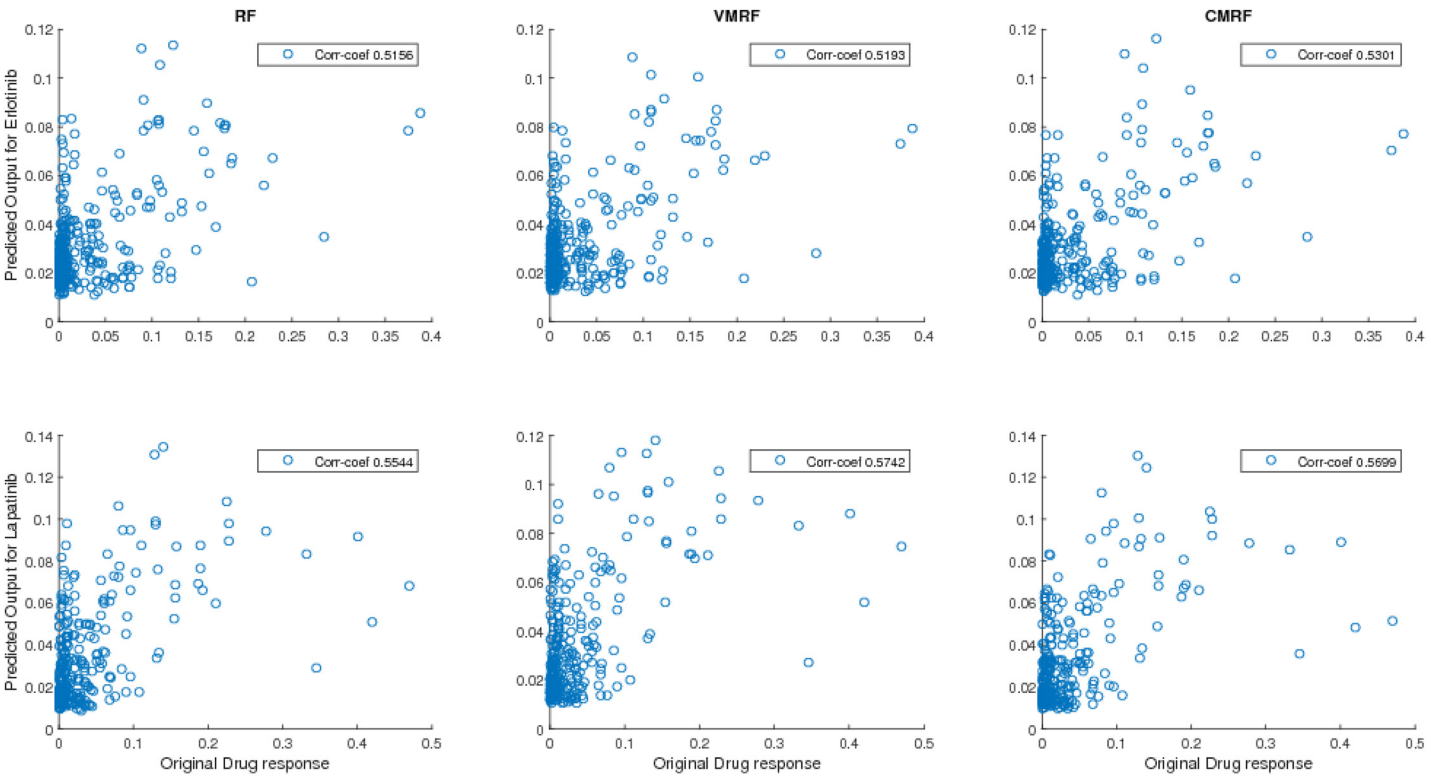


Fig 7. Scatter plots of predicted response vs original response for Erlotinib and Lapatinib (GDSC). Here corr-coef stands for correlation coefficient between predicted response and output response.

doi:10.1371/journal.pone.0144490.g007

multivariate approaches for both average correlation coefficients and NRMSE. The scatter plots of predicted response vs original response for drugset S_{C1} using RF, VMRF, CMRF are shown in Fig 7.

Results on CCLE Dataset

The CCLE [35] database includes genomic characterization for 1037 cell lines and drug responses over 24 drugs for over 480 cell lines. For the purpose of predicting responses, 4 sets of drugs were selected. The first set S_{C1} = {*Erlotinib*, *Lapatinib*} has *EGFR* as a common target [30, 31], the second set S_{C2} = {*PF-02341066(Crizotinib), PHA-665752*} has *MET* as a common target [36, 37], the third set S_{C3} = {*ZD6474(Vandetanib), AZD0530(Saracatinib)*} has *EGFR* as a common target [38, 39] and the fourth set S_4 = {*17-AAG, Erlotinib*} has no common target. We also include results on 4 drug set and 24 drug sets for CCLE as Tables A, F and G in S1 File.

Initially, each cell line had 18,988 features (probeset) as gene expressions. We reduced it to 500 for each drug response using RELIEFF [28] feature selection and considered a union of the 500 features in each of the four sets of drugs. We have used first 300 cell lines that have gene expression and drug responses for specific pairs of drugs. To report our results, we compared 5 fold cross-validated Pearson correlation coefficients, MAE and NRMSE between predicted and experimental responses for RF, VMRF, CMRF and KBMTL. For both VMRF and CMRF, we set the minimum size of samples in each leaf to $n_{size} = 5$, the number of trees in the forest to $T = 150$ and the splitting in each node considers $m = 10$ random features.

The correlation coefficients using 5 fold cross validation error estimation are illustrated for each drug set in Table 5. The corresponding MAE and NRMSE behaviors are illustrated in

Table 5. 5 fold CV results for CCLE Dataset drug sensitivity prediction for four drug sets in the form of correlation coefficients for *RF*, *VMRF*, *CMRF* and *KBMTL*.

			Correlation Co-efficients			
Drug Set	**Common Target**	**Drug Name**	***RF***	***VMRF***	***CMRF***	***KBMTL***
S_{C1}	EGFR	Erlotinib	0.3916	0.3980	0.3927	0.3457
		Lapatinib	0.4460	0.4468	0.4673	0.2609
S_{C2}	MET	Crizotinib	0.4813	0.4719	0.4882	0.4519
		PHA-665752	0.3547	0.3587	0.3746	0.2250
S_{C3}	EGFR	ZD-6474	0.2355	0.2535	0.2627	0.1304
		AZD-0530	0.1990	0.1844	0.1957	0.1973
S_U	None	17-AAG	0.3620	0.3337	0.3255	0.4100
		Erlotinib	0.3818	0.3852	0.3718	0.2828

doi:10.1371/journal.pone.0144490.t005

Table 6. For CMRF, results with scaling factor α selected using *Method-1* discussed earlier has been used. Tables 5 and 6 shows that CMRF performed better than VMRF, KBMTL and RF in terms of correlation coefficients and NRMSE for the related drug pairs S_{C1}, S_{C2}, and S_{C3}. When there is no relationship in the drug pair as in S_U, univariate RF performs better than the multivariate approaches on an average. The scatter plots of predicted response vs original response for drug-set S_{C2} using RF, VMRF, CMRF are shown in Fig 8.

Results of Variable Importance Analysis

We have examined the variable importance measure for GDSC data using VMRF and CMRF in terms of protein interaction network enrichment analysis. In this section, we will primarily provide the detailed results for S_{C1} in GDSC. To avoid any bias due to feature selection in variable importance, we consider the full set of probe set ids without application of RELIEFF for this analysis.

In both VMRF and CMRF, the 50 top ranked probesets were generated separately. It should be noted that multiple probeset IDs can map to a single Gene Symbol of a protein. This mapping was done in Genome Medicine Database of Japan (GeMDBJ) ID conversion tool (https://gemdbj.nibio.go.jp/dgdb/ConvertOperation.do). Based on this mapping, we arrived at 58 top ranked proteins for VMRF and 70 top ranked proteins for CMRF. These proteins were

Table 6. 5 fold CV results for CCLE Dataset drug sensitivity prediction for four drug sets in the form of MAE and NRMSE for *RF*, *VMRF*, *CMRF* and *KBMTL*.

			MAE				NRMSE			
Drug Set	**Common Target**	**Drug Name**	***RF***	***VMRF***	***CMRF***	***KBMTL***	***RF***	***VMRF***	***CMRF***	***KBMTL***
S_{C1}	EGFR	Erlotinib	0.0522	0.0520	0.0515	0.0612	0.9223	0.9210	0.9218	1.0593
		Lapatinib	0.0513	0.0520	0.0509	0.0654	0.8976	0.8977	0.8895	1.1398
S_{C2}	MET	Crizotinib	0.0484	0.0483	0.0477	0.0546	0.8836	0.8921	0.8828	0.9674
		PHA-665752	0.0492	0.0496	0.0489	0.0614	0.9367	0.9367	0.9307	1.1573
S_{C3}	EGFR	ZD-6474	0.0660	0.0659	0.0656	0.0876	0.9721	0.9674	0.9650	1.3037
		AZD-0530	0.0728	0.0728	0.0727	0.0866	0.9801	0.9834	0.9810	1.2188
S_U	None	17-AAG	0.1003	0.1005	0.1008	0.0997	0.9553	0.9614	0.9644	0.9740
		Erlotinib	0.0517	0.0519	0.0520	0.0612	0.9258	0.9260	0.9311	1.0957

doi:10.1371/journal.pone.0144490.t006

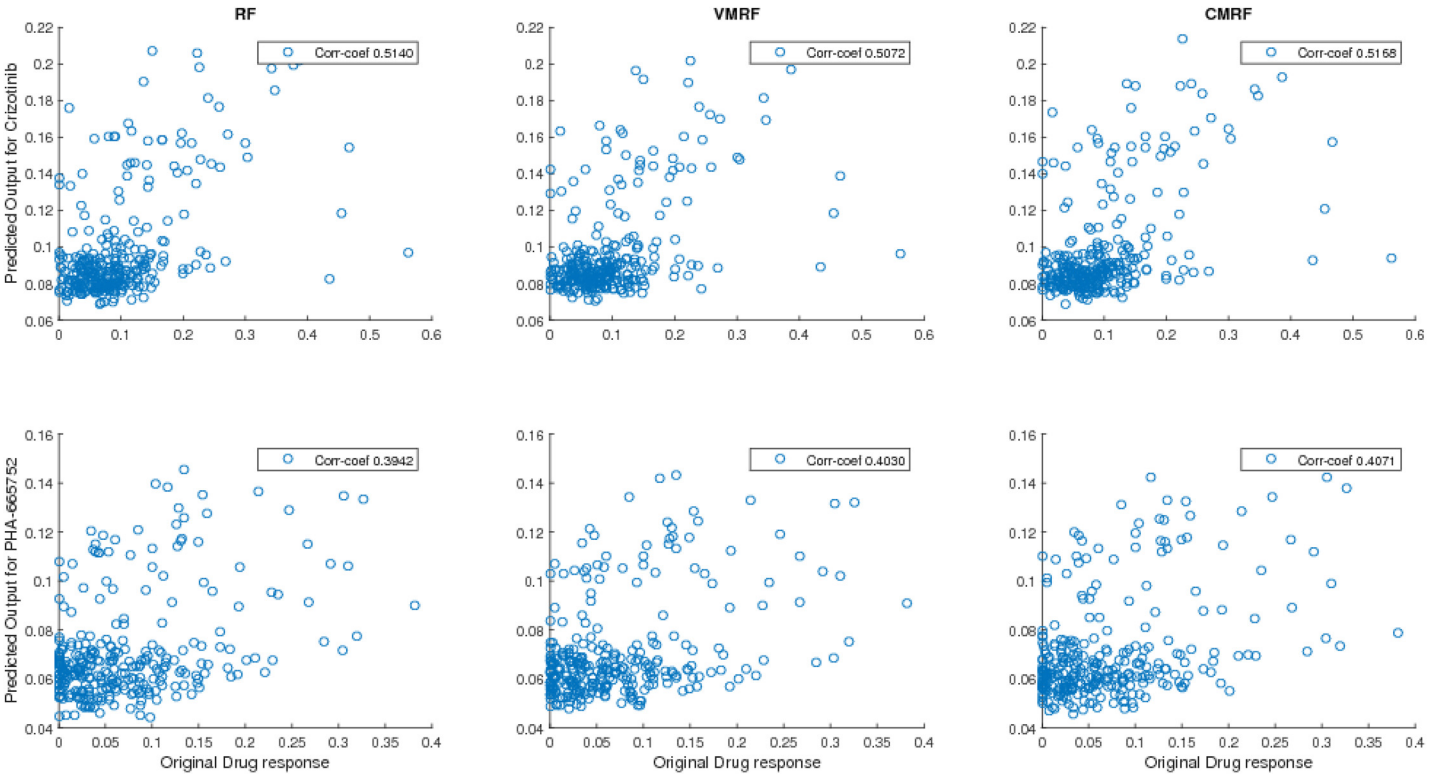


Fig 8. Scatter plots of predicted response vs original response for Crizotinib and PHA-665752 (CCLE). Here corr-coef stands for correlation coefficient between predicted response and output response.

doi:10.1371/journal.pone.0144490.g008

provided as inputs to the string-db database (http://string-db.org/) for known protein-protein interactions. The protein-protein interaction (PPI) networks for top proteins using CMRF and VMRF are shown in Figs 9 and 10 respectively. The enrichment analysis for both the networks are shown alongside each network. We observe that the network generated using CMRF is more enriched in connectivity than the network generated using VMRF. 18 interactions with a p-value of 0.132 were observed for the VMRF PPI network whereas a total of 35 interactions with a p-value of 0.00775 were observed for the CMRF network. Moreover, the common target EGFR is picked in the top 50 targets and is well connected to other targets of CMRF whereas EGFR is not selected even in top 150 targets of VMRF.

Similarly, in drugset S_{C2} of GDSC (network not shown), there are 42 interactions with 51 proteins in CMRF and 25 interactions with 54 proteins in VMRF.

Conclusions

In this article, we presented an approach to extend ensemble learning using regression trees to multivariate ensemble learning. We utilized the concept of copulas to represent the relationship between different drug sensitivities and incorporated them in the design of multivariate regression tree cost function. We designed the node cost function as a combination of (a) the sum of square of the differences from the mean and (b) a measure of the difference in the multivariate structure at the node compared to the original training data. The difference in the multivariate structure was captured as the integral of the absolute difference in the copulas observed at the node and the original training data. Two approaches were presented based on enumeration

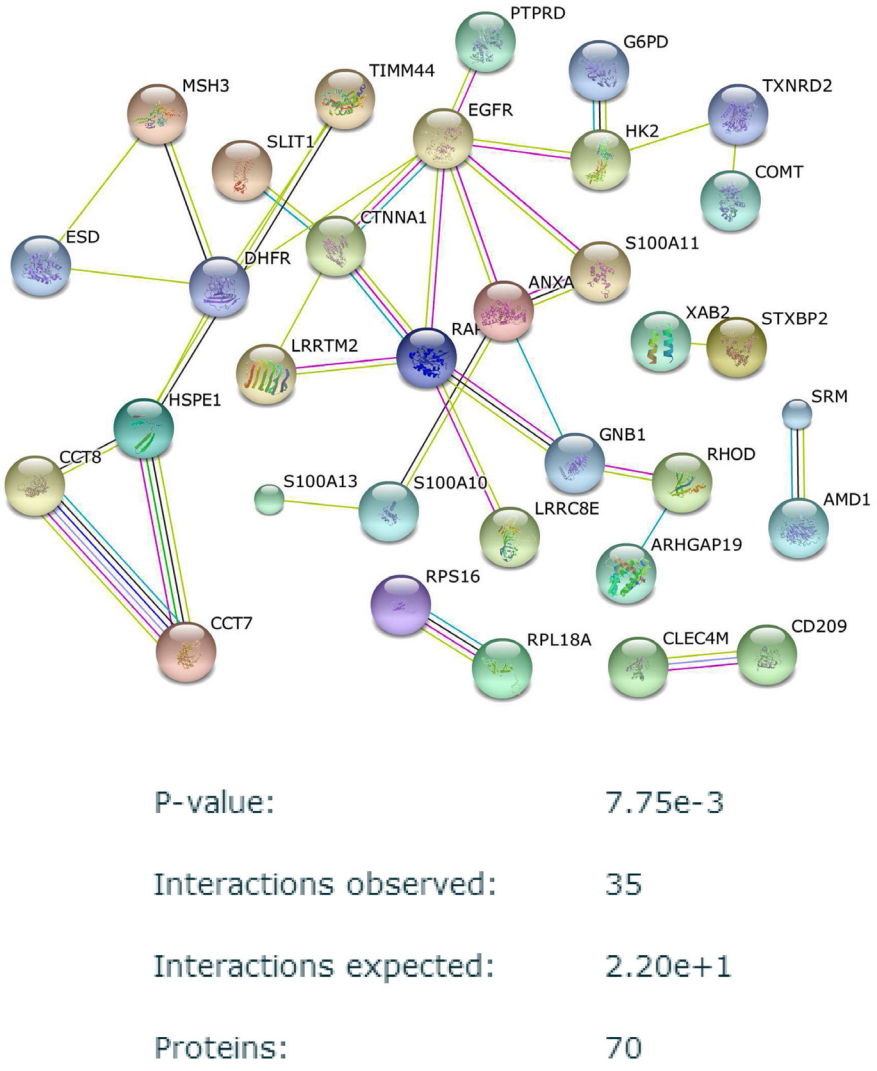


Fig 9. Protein-protein interaction network observed between top regulators found from CMRF in GDSC dataset S_{C1}. Disconnected nodes are hidden.

doi:10.1371/journal.pone.0144490.g009

and Pareto frontier to design the weights of the two parts of the cost function. Utilizing synthetic and biological data, we showed that the proposed copula based approach could increase the prediction accuracy as compared to univariate random forests or multivariate random forests based on covariance based node cost. As compared to RF, the gain in the correlation coefficient between predicted and experimental values was observed in scenarios where there exists a relationship between the drug pair sensitivities. The examples were also able to illustrate that CMRF is better suited for selecting the relevant features as compared to VMRF. The proposed methodology provides a novel technique to design multivariate regression trees for scenarios where there are nonlinear relationships between output responses. The presented research can be extended in multiple directions. One such direction will involve extending the concept of maintaining the multivariate structure in the design of weights of individual trees. Another direction consists of analyzing the detailed bias and variance relationship of the proposed technique and designing confidence intervals for the predictions.

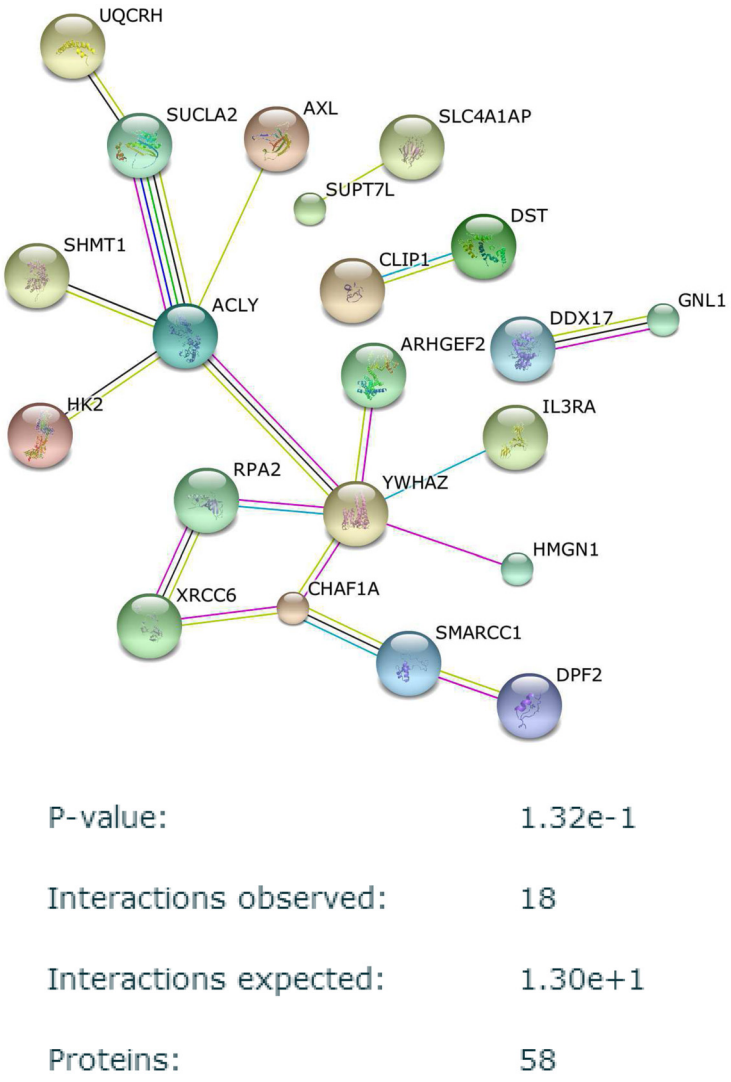


Fig 10. Protein-protein interaction network observed between top regulators found from VMRF in GDSC dataset S_{C1}. Disconnected nodes are hidden.

doi:10.1371/journal.pone.0144490.g010

Supporting Information

S1 File. Supporting Information for Article: *A copula based approach for design of multivariate random forests for drug sensitivity prediction*. RF, VMRF, CMRF results (5 fold cross validation) with and without prior feature selection (**Table A**). Results for CCLE Dataset drug sensitivity prediction for a drugset with 4 drugs in the form of correlation coefficients for *RF*, *VMRF*, *CMRF* and *KBMTL* approaches (**Table B**). Results for GDSC Dataset drug sensitivity prediction for a drugset with 3 drugs in the form of correlation coefficients for *RF*, *VMRF*, *CMRF* and *KBMTL* approaches (**Table C**). Results for GDSC Dataset drug sensitivity prediction for a drugset with 140 drugs in the form of correlation coefficients is shown (only 15 drugs that are common with CCLE are shown in detail while the average represents the average of all 140 drugs) (**Table D**). Results for GDSC Dataset drug sensitivity prediction for a drugset with 140 drugs in the form of NRMSE is shown (only 15 drugs that are common with CCLE are shown in detail while the average represents the average of all 140 drugs) (**Table E**). Results for CCLE Dataset

drug sensitivity prediction for the combined set of 24 drugs in the form of correlation coefficients (**Table F**). Results for CCLE Dataset drug sensitivity prediction for the combined set of 24 drugs in the form of Normalized Root Mean Square Error (**Table G**). Comparison of α for different sets of synthetic data with and without noise added to the drug response (**Table H**). Comparison of α for different amount of random subset of the original samples in a specific synthetic data. Original number of samples were 350 in this specific example (**Table I**). Simulation time for different drug-sets in GDSC data. The reported simulation times are the time needed to generate complete result for all drugs in a drug set for 5 fold cross validation (**Table J**). Simulation time for different drug-sets in GDSC data. The reported simulation times are the time needed to generate complete result for all drugs in a drug set for 30–70 case (**Table K**). Simulation time for different methods for all drugs of GDSC dataset (140) and CCLE dataset (24). The reported simulation times are the time (in seconds) needed to generate complete result for all drugs for 30–70 case (**Table L**).
(PDF)

Author Contributions

Conceived and designed the experiments: SH RP. Performed the experiments: SH RR. Analyzed the data: SH RR SG RP. Contributed reagents/materials/analysis tools: SH RR SG RP. Wrote the paper: SH RR RP.

References

1. Sos ML, et al. Predicting drug susceptibility of non-small cell lung cancers based on genetic lesions. The Journal of clinical investigation. 2009; 119(6):1727–1740. doi: 10.1172/JCI37127 PMID: 19451690
2. Staunton JE, et al. Chemosensitivity prediction by transcriptional profiling. Proceedings of The National Academy of Sciences. 2001; 98:10787–10792. doi: 10.1073/pnas.191368598
3. Lee JK, et al. A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. Proceedings of the National Academy of Sciences. 2007 Aug; 104(32):13086–13091. doi: 10.1073/pnas.0610292104
4. Mitsos A, et al. Identifying Drug Effects via Pathway Alterations using an Integer Linear Programming Optimization Formulation on Phosphoproteomic Data. PLoS Comput Biol. 2009; 5(12). doi: 10.1371/journal.pcbi.1000591 PMID: 19997482
5. Walther Z, Sklar J. Molecular tumor profiling for prediction of response to anticancer therapies. Cancer J. 2011; 17(2):71–9. PMID: 21427550
6. Zou H, Hastie T. Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society, Series B. 2005; 67:301–320. doi: 10.1111/j.1467-9868.2005.00527.x
7. Barretina J, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012 Mar; 483(7391):603–607. Available from: http://dx.doi.org/10.1038/nature11003. doi: 10.1038/nature11003 PMID: 22460905
8. Riddick G, et al. Predicting in vitro drug sensitivity using Random Forests. Bioinformatics. 2011 Jan; 27 (2):220–224. doi: 10.1093/bioinformatics/btq628 PMID: 21134890
9. Costello JC, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nature Biotechnology. 2014;p. doi: 10.1038/nbt.2877
10. Wan Q, Pal R. An ensemble based top performing approach for NCI-DREAM drug sensitivity prediction challenge. PLOS One. 2014; 9(6):e101183. doi: 10.1371/journal.pone.0101183 PMID: 24978814
11. Gonen M, Margolin AA. Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning. Bioinformatics. 2014; 30(17):i556–i563. doi: 10.1093/bioinformatics/btu464 PMID: 25161247
12. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, et al. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. PLoS One. 2013; 8(4). doi: 10.1371/journal.pone.0061318 PMID: 23646105
13. Segal M XY. Multivariate Random Forests. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011; 1:80–87.
14. Wan Q, Pal R. A multivariate random forest based framework for drug sensitivity prediction. In: International Workshop on Genomic Signal Processing and Statistics (GENSIPS); 2013. p. 53.

15. Breiman L. Random Forests. Machine learning. 2001; 45:5–32. doi: 10.1023/A:1010933404324
16. Meinshausen N. Quantile Regression Forests. Journal of Machine Learning Research. 2006; 7:983–999.
17. Biau G. Analysis of a random forests model. The Journal of Machine Learning Research. 2012; 13:1063–1095.
18. Mahalanobis PC. On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India. 1936; 2(1):49–55.
19. Sim KC, Gales M. Precision matrix modelling for large vocabulary continuous speech recognition. Cambridge University Engineering Department. June 2004;p. Appendix B1.
20. Sklar A. Fonctions de répartition á n dimensions et leurs marges. Publ Inst Statist Univ Paris. 1959; 8:229–231.
21. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. International Statistics Reviews. 1978; 65:141–151.
22. Lee L. Generalized Econometric Models with Selectivity. Econometrica. 1983; 51:507–512. doi: 10.2307/1912003
23. Frank MJ. On the simultaneous associativity of F(x,y) and x+y – F(x,y). Aequationes Math. 1979; 19:194–226. doi: 10.1007/BF02189866
24. Demarta S, McNeil AJ. The t copula and related copulas. International Statistical Review. 2005; 73:111–129. doi: 10.1111/j.1751-5823.2005.tb00254.x
25. Gumbel EJ. Distributions des Valeurs Extremes en Plusieurs Dimensions. Publications de l'Institute de Statistique de l'Université de Paris. 1960; 9:171–173.
26. Kouros Owzar PKS. Copulas: concepts and novel applications. International Journal of Statistics. 2003; LXI (3):323–353.
27. Archer KJ, Kimes RV. Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis. 2008; 52 (4):2249–2260. doi: 10.1016/j.csda.2007.08.015
28. Robnik-Sikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning. 2003; 53:23–69. doi: 10.1023/A:1025667309714
29. Yang W, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic acids research. 2013; 41(D1):D955–D961. doi: 10.1093/nar/gks1111 PMID: 23180760
30. Ling YH, Li T, Yuan Z, Haigentz M, Weber TK, Perez-Soler R. Erlotinib, an effective epidermal growth factor receptor tyrosine kinase inhibitor, induces p27KIP1 up-regulation and nuclear translocation in association with cell growth inhibition and G1/S phase arrest in human non-small-cell lung cancer cell lines. Molecular pharmacology. 2007; 72(2):248–258. doi: 10.1124/mol.107.034827 PMID: 17456787
31. Johnston SR, Leary A. Lapatinib: a novel EGFR/HER2 tyrosine kinase inhibitor for cancer. Drugs Today (Barc). 2006; 42(7):441–453. doi: 10.1358/dot.2006.42.7.985637
32. Super Target;. Http://bioinf-apache.charite.de/supertarget_v2/index.php?site = home.
33. Falchook GS, Lewis KD, Infante JR, Gordon MS, Vogelzang NJ, DeMarini DJ, et al. Activity of the oral MEK inhibitor trametinib in patients with advanced melanoma: a phase 1 dose-escalation trial. The lancet oncology; 13(8):782–789. doi: 10.1016/S1470-2045(12)70269-3 PMID: 22805292
34. Ciuffreda L, Del Bufalo D, Desideri M, Di Sanza C, Stoppacciaro A, Ricciardi MR, et al. Growth-inhibitory and antiangiogenic activity of the MEK inhibitor PD0325901 in malignant melanoma with or without BRAF mutations. Neoplasia (New York, NY). 2009; 11(8):720. doi: 10.1593/neo.09398
35. Broad-Novartis Cancer Cell Line Encyclopedia http://www.broadinstitute.org/ccle/home;. Genetic and pharmacologic characterization of a large panel of human cancer cell lines.
36. Tanizaki J, Okamoto I, Okamoto K, Takezawa K, Kuwata K, Yamaguchi H, et al. MET tyrosine kinase inhibitor crizotinib (PF-02341066) shows differential antitumor effects in non-small cell lung cancer according to MET alterations. Journal of Thoracic Oncology. 2011; 6(10):1624–1631. PMID: 21716144
37. Ma PC, Schaefer E, Christensen JG, Salgia R. A selective small molecule c-MET Inhibitor, PHA665752, cooperates with rapamycin. Clinical cancer research. 2005; 11(6):2312–2319. doi: 10.1158/1078-0432.CCR-04-1708 PMID: 15788682
38. Morabito A, Piccirillo MC, Falasconi F, De Feo G, Del Giudice A, Bryce J, et al. Vandetanib (ZD6474), a dual inhibitor of vascular endothelial growth factor receptor (VEGFR) and epidermal growth factor receptor (EGFR) tyrosine kinases: current status and future directions. The oncologist. 2009; 14 (4):378–390. doi: 10.1634/theoncologist.2008-0261 PMID: 19349511
39. Larsen AB, Stockhausen MT, Poulsen HS. Cell adhesion and EGFR activation regulate EphA2 expression in cancer. Cellular signalling. 2010; 22(4):636–644. doi: 10.1016/j.cellsig.2009.11.018 PMID: 19948216