2016

Design of Probabilistic Random Forests with Applications to Anticancer Drug Sensitivity Prediction- 2016

Raziur Rahman

Saad Haider

Souparno Ghosh

Ranadip Pal

Design of Probabilistic Random Forests with Applications to Anticancer Drug Sensitivity Prediction

Libertas Academica
FREEDOM TO RESEARCH

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

[1]Department of Electrical and Computer Engineering, Texas Tech University, Lubbock, TX, USA. [2]Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX, USA.

Supplementary Issue: Statistical Systems Theory in Cancer Modeling, Diagnosis, and Therapy

ABSTRACT: Random forests consisting of an ensemble of regression trees with equal weights are frequently used for design of predictive models. In this article, we consider an extension of the methodology by representing the regression trees in the form of probabilistic trees and analyzing the nature of heteroscedasticity. The probabilistic tree representation allows for analytical computation of confidence intervals (CIs), and the tree weight optimization is expected to provide stricter CIs with comparable performance in mean error. We approached the ensemble of probabilistic trees' prediction from the perspectives of a mixture distribution and as a weighted sum of correlated random variables. We applied our methodology to the drug sensitivity prediction problem on synthetic and cancer cell line encyclopedia dataset and illustrated that tree weights can be selected to reduce the average length of the CI without increase in mean error.

KEYWORDS: probabilistic random forests, drug sensitivity prediction, variance analysis of random forests, heteroscedasticity

SUPPLEMENT: Statistical Systems Theory in Cancer Modeling, Diagnosis, and Therapy

CITATION: Rahman et al. Design of Probabilistic Random Forests with Applications to Anticancer Drug Sensitivity Prediction. *Cancer Informatics* 2015:14(S5) 57–73 doi: 10.4137/CIN.S30794.

TYPE: Original Research

RECEIVED: December 10, 2015. **RESUBMITTED:** February 03, 2016. **ACCEPTED FOR PUBLICATION:** February 07, 2016.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Four peer reviewers contributed to the peer review report. Reviewers' reports totaled 1079 words, excluding any confidential comments to the academic editor.

FUNDING: This work was supported by a contract from NCI/Leidos Biomedical 15X073. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: ranadip.pal@ttu.edu

Introduction

The prediction of an output response Y based on supervised training of a predictor X has been approached using numerous methodologies, such as elastic net,[1] support vector regression, and random forests (RFs),[2,3] where majority of the techniques provide point prediction estimates of the output. In this article, we consider the generation of prediction confidence intervals (CIs) for RFs, which is a commonly used prediction model in diverse scenarios.[3,4] The generation of input-dependent prediction probability distribution provides an estimate of the heteroscedasticity or the change in error variance for different predictor samples. RF regression[2] consists of an ensemble of regression trees where the prediction output of the forest is based on the average prediction of individual regression trees. We utilize the concept of probabilistic regression trees[5,6] to convert the point estimate of individual trees to probability distributions and further consider the optimization of the weights of the ensemble of probabilistic regression trees that can provide stricter CIs.

The ensemble of probabilistic regression trees is considered from two different perspectives.

First, we consider the ensemble as a mixture distribution for each prediction sample X_i. Consider an ensemble of T trees where the tree j produces the predicted output probability density function $P(Y_j|X_i)$. The probability density function $P(Y|X_i)$ of the ensemble of the T regression trees with weights $\alpha_1,\cdots,\alpha_T$ is then given by $P(Y|Xi) = \sum_{j=1}^{T} \alpha_j P(Y_j|X_i)$. This approach considers that based on the weights α_j, a tree k will be selected and the prediction will be decided based on $P(Y_k|X_i)$.

For the second perspective, we consider the output of the ensemble to be the random variable Z where $Z = \sum_{j=1}^{T} \alpha_j Z_j$ is a weighted sum of T random variables $Z_1,\cdots,Z_T$ with Z_j denoting a random variable with probability density function $P(Z_j|X_i)$ based on tree j. This scenario is equivalent to analyzing a weighted sum of random variables with different probability density functions, ie, we model the weighted sum of the realizations of the random variables rather than the weighted sum of their distributions as was considered in the first case.

Note that the use of equal weights (ie, $\alpha_j = \frac{1}{T}$ for $j = 1, \cdots, T$) for the regression trees is supposed to work well in terms of reducing the variance of prediction when the generated

trees are uncorrelated. However, some of the generated trees can often be correlated to each other, and in such a scenario, we can potentially optimize the weights of the trees to reduce the variance of ensemble prediction. Based on this idea, we analyze the variance of the prediction based on a weighted sum of random variables scenario for different forms of tree covariance matrices. For the mixture distribution scenario, we use maximum-likelihood estimation (MLE) to generate the weights of the regression trees and analyze the effect of estimated weights on the mean and variance of the error distributions. We applied our methodology over synthetic and experimental cancer cell line encyclopedia (CCLE) dataset and illustrated a reduction in variance with comparable mean error following the application of MLE-optimized tree weights.

Background

A probabilistic theory for classification has been developed for some time that can provide bounds on the probability of misclassification.[7] For instance, binary minimax probability machine classification algorithm[8] computes a bound on the probability of misclassification, using only estimates of the covariance matrix and mean for each class, as obtained from the training data. Probability estimation trees (PETs)[9] are introduced in Ref. 10 as classification trees[11] with a class probability distribution at each leaf instead of single-class label. Similar to classification trees, the PETs can be used for classifying examples, and this is simply done by assigning the most probable class according to the PET. They can also be used for ranking examples, and this is done by ordering the examples according to their likelihood of belonging to some particular class as estimated by the PET. Probabilistic RF for classification has been introduced in Ref. 12 with the perspective of providing an estimate of the probability of misclassification for each data point, without detailed probability distribution assumptions or resorting to density modeling. Probabilistic RF for classification is based on two existing algorithms: minimax probability machine classification[8] and RFs.[2]

It has been noted in Ref. 13 that classification trees are not equally successful in labeling all instances. This simple observation led to the idea that use of selected trees in classification can potentially increase accuracy. The selection of trees based on their performance on similar instances had limited success. Further refinement of this idea led to the concept of weighted voting. Mishina et al.[14] proposed a boosted RF model where a boosting algorithm is integrated with a conventional RF approach. The boosted RF maintains a high classification performance, even with fewer decision trees, based on constructing complementary classifiers through sequential training by boosting.

For our relevant purpose of regression using ensemble approaches, there have been limited studies on the probabilistic behavior of ensemble of regression trees. Theoretical analysis of RF models has usually focused on the consistency and rate of convergence of the design procedure.[15] Probabilistic decision and regression trees have been considered in Ref. 5 but the ensemble of probabilistic regression trees in the context of altering the variance of prediction error has not been explored. A weighted random forest (wRF) for regression approach has been proposed,[16] where the weight of each tree has been calculated based on the prediction accuracy of out-of-bag samples for that tree. wRF considers the empirical out-of-bag errors for estimating the regression tree weights, whereas this article considers an analytical approach where parametric distributions are estimated to specify a probabilistic representation of each regression tree and the sample-dependent probability distributions are utilized to generate the tree weights.

Methods

RF regression. RF regression refers to ensembles of regression trees,[2] 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 $m = M$, 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 and 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 $x_{tr}(I \in \eta_P, j_s) \le z$) and η_R (right node with samples satisfying $x_{tr}(i \in \eta_P, j_s) > z$). We consider the node cost as the sum of square differences:

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

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

$$C(\gamma, \eta_P) = D(\eta_P) - D(\eta_L) - D(\eta_R) \quad (2)$$

The partition γ^* that maximizes $C(\gamma, \eta_P)$ for all possible partitions is selected for node η_P.

Forest prediction. Using the randomized feature selection process, we fit the tree based on the bootstrap sample $\{(X_1, Y_1),..., (X_N, Y_N)\}$ generated from the training data. Let $\hat{Y}_i(x)$ denote the regression tree prediction for input response x corresponding to tree i. The prediction for the RF consisting of T trees denoted by $\hat{Y}(x)$ is given by $\hat{Y}_i(x) = \frac{1}{T} \sum_{i=1}^{T} \hat{Y}_i(x)$.

Weighted RF. For comparison purposes, we will also consider the wRF methodology proposed in Ref. 16 that uses empirical values to calculate the weight of the trees. The prediction error of a tree denoted by tPE_j is calculated based on the out-of-bag samples for that tree, and the weight of that tree is estimated as $\omega_j = x_j / \sum_{j=1}^{ntree} x_j$, where $x_j = 1 - tPE_j$.

Probabilistic regression trees. Let us consider the generation of regression trees from a probabilistic perspective, which will allow us to utilize well-known concepts of parameter estimations for statistical models. Estimation of regression trees using probability models has been explored in Refs. 5,6. For a regression tree, our goal is to generate the conditional density of the form $P(y|x, \varphi)$ where y and x refers to the output and input responses, respectively, and φ denotes the collection of parameters for the tree. The tree splits can be modeled by probabilistic decisions that are conditional on the input x and previous node decisions. As an example, consider the two-level tree shown in Figure 1.

The first decision is based on probability $P(\omega_1|x, \eta)$ where ω_1 is the event signifying partition toward the left of the root node and η denotes a parameter vector $\eta = [\eta_1\ T_1]$.

Note that if we consider

$$P(\omega_1|x, \eta) = \frac{e^{\eta_1^T T1}}{e^{\eta_1^T X} + e^{\eta_1^T T1}} + \frac{1}{1 + e^{\eta_1^T (x - T1)}} \tag{3}$$

with large η_1, the split will be close to a sharp linear decision boundary similar to a regression tree. Similarly, we will have

$$P(\omega_2|x, \eta) = 1 - P(\omega_1|x, \eta) = \frac{1}{1 + e^{\eta_1^T (T1 - x)}} \tag{4}$$

If we consider all the branches of the tree as shown in Figure 1, the corresponding distribution of y conditional on x and tree parameters $\phi = \{\eta, \lambda_1, \lambda_2, \vartheta_{11}, \vartheta_{12}, \vartheta_{21}, \vartheta_{22}\}$ will be given by Eq. 5.

$$P(y|x, \phi) = \sum_{i=1}^{2}\sum_{j=1}^{2} P(y|x, \omega_{ij}, \lambda_i, \omega_i, \eta) P(\omega_{ij}|\lambda_i, \omega_i, \eta, x) P(\omega_i|\eta, x) \tag{5}$$

For larger number of branches in the tree, the above technique can be extended to obtain $P(y|x, \phi)$ for a tree with parameter set ϕ. In this article, we consider that the tree parameters φ are generated based on the standard RF node generation criteria given in Eq. 2. The probability distribution at any leaf node is approximated by a Gaussian distribution with mean and variance equal to the mean and variance of the samples at the leaf node. Some examples of empirical distributions fitted to normal approximations are shown in Figure 1.

Consequently, an ensemble of T trees generated by RF regression can be represented by the T tree parameters $\phi_1, \phi_2, \cdots, \phi_T$ with each producing the conditional distribution $P(y|x, \phi_i)$ for $i = 1, \cdots, T$.

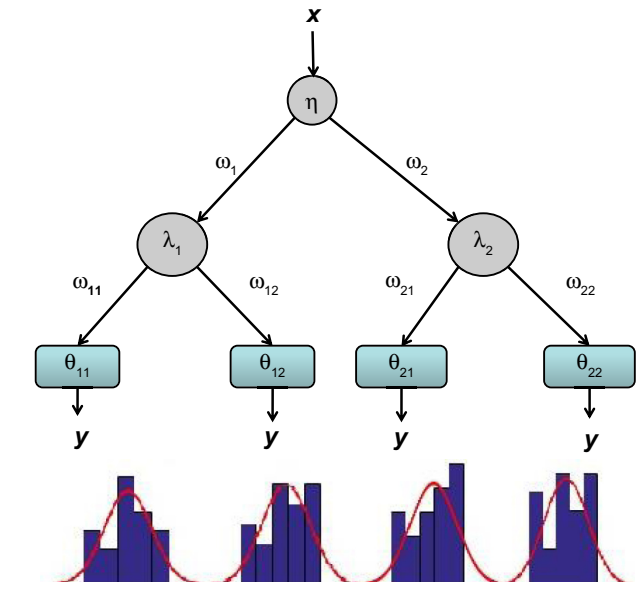

Figure 1. Example of probabilistic decision tree.

Probabilistic RFs

Mixture distribution. As discussed earlier, we consider the prediction ensemble as a mixture distribution for each sample X_i. Consider an ensemble of T trees where tree j produces the predicted output probability density function $P(Y_j|X_i)$. The predicted distribution for each tree is based on the estimated probabilistic regression tree model described in the previous section. The probability density function (pdf) $P(Y|X_i)$ of the forest of T regression trees with weights $\alpha_1, \cdots, \alpha_T$ with $\alpha_j \geq 0$ and $\sum_{j-1}^{T} \alpha_j = 1$ is given by $P(Y|X_i) = \sum_{j=1}^{T} \alpha_j P(Y_j|X_i)$. This approach considers that based on the weights α_j, a tree k will be selected, and the prediction will be decided based on $P(Y_k|X_i)$.

The mean (μ) of the mixture distribution will be equal to the weighted sum of the distribution means (μ_i) of the trees as shown in Eq. 6.

$$E[Y] = \mu = \sum_{i=1}^{T} \alpha_i \mu_i \tag{6}$$

The variance of the mixture distribution (σ^2) is given by Eq. 7.

$$E[(Y - \mu)^2] = \sigma^2 = \sum_{i=1}^{T} \alpha_i ((\mu_i - \mu)^2 + \sigma_i^2) \tag{7}$$

Weighted sum of random variables. The mixture distribution approach selects a tree based on the tree weights and then selects a sample output according to the pdf of the tree. Another potential is to consider the weighted sum of realizations from each tree. As discussed earlier, this will be equivalent to considering the output of the forest to be a random variable Z where $Z = \sum_{j=1}^{T} \alpha_j Z_j$ is a weighted sum of T random

variables $Z_1,\cdots,Z_T$ where Z_j denotes a random variable with pdf $P(Z_j|X_i)$ based on tree j. The distribution of a sum of random variables can be computed as the convolution of the individual distributions. This scenario is equivalent to analyzing a weighted sum of random variables with different probability density functions, ie, we model the weighted sum of the realizations of the random variables rather than the weighted sum of their distributions as was considered in the first case.

Example for weighted sum of two uncorrelated Gaussian distributions. Consider two independent random variables X_1 and X_2 that are normally distributed with pdfs $\mathbb{N}(\mu x_1, \sigma^2_{X_1})$ and $\mathbb{N}(\mu x_2, \sigma^2_{X_2})$, respectively. The distributions of $X_{1\alpha}= \alpha_1 X_1$ and $X_{2\alpha}= \alpha_2 X_2$ are given by Eq. 8:

$$\begin{aligned} fx_{1\alpha}(x_{1\alpha}) &= \frac{1}{\sqrt{2\pi\sigma^2_{X1}\alpha_1^2}} \exp^{-\frac{(x_{1\alpha}-\alpha_1\mu x_1)^2}{2\alpha_1^2\alpha^2_{X_1}}} \\ fx_{2\alpha}(x_{2\alpha}) &= \frac{1}{\sqrt{2\pi\sigma^2_{X2}\alpha_2^2}} \exp^{-\frac{(x_{2\alpha}-\alpha_2\mu x_2)^2}{2\alpha_2^2\alpha^2_{X_2}}} \end{aligned} \tag{8}$$

Based on the idea of derived distributions, the pdf of random variable $Z = \alpha_1 X_1 + \alpha_2 X_2 = X_{1\alpha} + X_{2\alpha}$ is given by the convolution of the pdfs of $X_{1\alpha}$ and $X_{2\alpha}$ and is given by Eq. 9:

$$fz(z) = \int_{-\infty}^{\infty} fx_{1\alpha}(z-x)\, fx_{2\alpha}(x)dx \tag{9}$$

Substituting Eq. 8 in Eq. 9, we arrive at:

$$\begin{aligned} fz(z) = \int_{-\infty}^{\infty} & \frac{1}{\sqrt{2\pi\sigma^2_{X_1}\alpha_1^2}} \exp^{-\frac{(z-x_{1\alpha_1}-\alpha_1\mu x_1)^2}{2\sigma^2_{X_1}\alpha_1^2}} \\ & \frac{1}{\sqrt{2\pi\sigma^2_{X_2}\alpha_2^2}} \exp^{-\frac{(x_{2\alpha_2}-\alpha_2\mu x_2)^2}{2\sigma^2_{X_2}\alpha_2^2}} dx \end{aligned} \tag{10}$$

$$= \frac{1}{\sqrt{2\pi(\sigma^2_{X_1}\alpha_1^2+\sigma^2_{X_2}\alpha_2^2)}} \exp\left[-\frac{(z-(\alpha_1\mu x_1+\alpha_2\mu x_2))^2}{2(\sigma^2_{X_1}\alpha_1^2+\sigma^2_{X_2}\alpha_2^2)}\right] \tag{11}$$

Eq. 11 represents the sum of two independent Gaussian random variables. For T independent Gaussian random variables X_1, X_T with pdfs $\mathbb{N}(\mu_1,\sigma_1),\cdots,\mathbb{N}(\mu_T,\sigma_T)$ representing the distribution at the T leaf nodes of the forest, the distribution of the random variable Z representing their weighted sum with weights $\alpha_1,\cdots,\alpha_T$ is given by Eq. 12 (derived based on multiple convolutions).

$$f_z(z) = \frac{1}{\sqrt{2\pi(\sum_{i=1}^{T}\sigma^2_{X_i}\alpha_i^2)}} \exp\left[-\frac{(z-(\sum_{i=1}^{T}\alpha_i\mu_{X_i}))^2}{2(\sum_{i=1}^{T}\sigma^2_{X_i}\alpha_i^2)}\right] \tag{12}$$

Thus, Z has a normal distribution with mean $= \sum_{i=1}^{T}\alpha_i\mu_{X_i}$ and variance $= \sum_{i=1}^{T}\alpha_i^2\sigma^2_{X_i}$.

However, if the random variables are correlated, ie, the covariance between different tree outputs are nonzero, the mean and variance of Z are given as follows:

$$\text{Mean}\left(\sum_{i=1}^{T}\alpha_i X_i\right) = \sum_{i=1}^{T}\alpha_i\mu_{X_i} \tag{13}$$

$$\begin{aligned} \text{Var}\left(\sum_{i=1}^{T}\alpha_i X_i\right) &= \sum_{i=1}^{T}\sum_{i=1}^{T}\text{Cov}\left(\alpha_i X_i, \alpha_j X_j\right) \\ &= \sum_{i=1}^{T}\alpha_i^2\text{Var}\left(X_i\right) + 2\sum_{1\le i<j\le T}\alpha_i\alpha_j\text{Cov}\left(X_i, X_j\right) \end{aligned} \tag{14}$$

If the vector $C = [\alpha_1,\cdots,\alpha_T]'$ represent the weight vector and Σ represent the $T\times T$ covariance matrix, the variance of Z can be represented concisely as

$$\text{Var}(Z) = |C'\Sigma C \tag{15}$$

where C' represents the transpose of C.

Note that the mean of Z denotes the weighted sum of the means of individual trees in the forest and the prediction is same as regular RF when the tree weights are equal. The mean of Z remains the same irrespective of whether the trees are correlated or not, whereas the variance of Z is directly related to the covariance of the trees using Eq. 15. In the following sections, we will attempt to estimate the covariance among the trees in a forest and analyze the effect of change in C on the variance of Z.

Empirical measure of correlation between probabilistic trees. The covariance between the trees will be estimated using empirical approaches to arrive at the covariance matrix Σ. The i,j position element of Σ denotes the covariance between the predictions of ith and jth tree represented by random variables Y_i and Y_j, respectively, ie, $\Sigma(i,j) = E[(Y_i - E(Y_i))(Y_j - E(Y_j))]$.

For each input sample X_i, tree j will produce a pdf $P(Y_j|X_i)$, which will be used to select an output prediction realization y_j. We perform this for all the other trees to arrive at a joint realization of the trees for sample X_i. This is repeated for N input training samples to produce N joint realizations of the random variables $Y_1,\cdots,Y_T$, which are used to calculate the sample covariance matrix shown in Eq. 34.

$$V = \begin{bmatrix} \frac{1}{N-1}\sum_{i=1}^{N}(Y_1(i)-E(Y_1))^2 & \cdots & \frac{1}{N-1}\sum_{i=1}^{N}(Y_1(i)-E(Y_1))(Y_T(i)-E(Y_T)) \\ \vdots & \ddots & \vdots \\ \frac{1}{N-1}\sum_{i=1}^{N}(Y_T(i)-E(Y_T))(Y_1(i)-E(Y_1)) & \cdots & \frac{1}{N-1}\sum_{i=1}^{N}(Y_T(i)-E(Y_T))^2 \end{bmatrix} \tag{16}$$

Effect of tree weight on variance. In this section, we will attempt to generate the lower and upper bounds on CVC where $C=[\alpha_1,\cdots,\alpha_T]'$ represents the tree weight vector. Assuming V is a Hermitian positive definite matrix (note that the covariance matrix V is always positive semidefinite[17]), we can generate the Cholesky decomposition[18] of $V=LL^T$, where L is a lower triangular matrix with real and positive diagonal entries. Let the variance of the prediction of a specific forest be given by the function $f(C)$. We have

$$\begin{aligned} f(C) &= C'VC \\ &= C'LL'C \\ &= (L'C)'L'C \\ &= A'A \end{aligned} \tag{17}$$

where $A=L'C$.

Let us analyze the minimum and maximum value of $f(C)$.

$$f(C)=A'A=\|A\|_2^2=\|L'C\|_2^2\le\|L'\|_2^2\|C\|_2^2 \tag{18}$$

Since

$$\|(L')^{-1}L'C\|_2^2\le\|(L')^{-1}\|_2^2\|L'C\|_2^2 \tag{19}$$

and

$$\|L'\|_2^2=\text{maximum eigenvalue of } LL'=\max\text{eig}(V) \tag{20}$$

$$\|(L')^{-1}\|_2^2=\frac{1}{\text{minimum eigenvalue of}(LL')}=\frac{1}{\min\text{eig}(V)} \tag{21}$$

We have,

$$\min\text{eig}(V)\|C\|_2^2\le f(C)\le\max\text{eig}(V)\|C\|_2^2 \tag{22}$$

The minimum for $f(C)$ under the constraint $C'e=1$ where $e=[1,1,\cdots,1]'$ is given by $f(C)=\frac{1}{e'V^{-1}e}$. Note that this does not preclude solutions with entries of C being less than zero. The details of the derivation using Lagrange multipliers is included in the Appendix. The weight vector C achieving the minimum is given by $C=\frac{V^{-1}e}{e'V^{-1}e}$. The computational complexity of estimating the weight vector is $\mathbb{O}(n^{2.376})$ based on the complexity of matrix inversion using Coppersmith–Winograd algorithm.

Diagonal elements of covariance matrix equal. In the conventional RF model, it is assumed that the trees are uncorrelated. Thus, nondiagonal elements of the covariance matrix (which shows the covariance between two different trees) are infinitesimal compared with the diagonal elements (reflecting the variance in the tree). If we ignore the small nondiagonal values and replace them with zeroes, then the covariance matrix (V) is a diagonal matrix. If the variance of the trees are equal ($\text{Var}[X_1]=\text{Var}[X_2]=\ldots=\text{Var}[X_T]=\sigma^2$), then the covariance matrix (Eq. 34) is a diagonal matrix.

Since the covariance matrix is diagonal with each diagonal entry equal to σ^2, all the T eigenvalues will also be equal to σ^2.

From Eq. 22,

$$\sigma^2\|C\|_2^2\le f(C)\le\sigma^2\|C\|_2^2 \tag{23}$$

Since $||C||_1=1$ and each entry of C is nonnegative,

$$\min\|C\|_2^2=\frac{T}{T^2}=\frac{1}{T} \tag{24}$$

Thus from Eq. 23

$$\frac{\sigma^2}{T}\le f(C) \tag{25}$$

When $C'=[1/T,1/T,\cdots,1/T]$ as in a conventional RF scenario, the variance is given by:

$$\begin{bmatrix}\frac{1}{T} & \frac{1}{T}\cdots\frac{1}{T}\end{bmatrix}\begin{bmatrix}\sigma^2 & 0 & \cdots & 0\\ 0 & \sigma^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma^2\end{bmatrix}\begin{bmatrix}\frac{1}{T}\\ \vdots\\ \frac{1}{T}\end{bmatrix}=T*\frac{1}{T^2}\sigma^2=\frac{\sigma^2}{T} \tag{26}$$

Comparing Eqs. 25 and 26, we observe that $C'=[1/T,1/T,\cdots,1/T]$ achieves the minimum variance for uncorrelated trees with equal variance.

Diagonal elements of covariance matrix unequal. Consider the case where the covariance matrix is a diagonal matrix and the variance of the trees are not equal as shown in Eq. 27.

$$V=\begin{bmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma_T^2\end{bmatrix} \tag{27}$$

where σ_i^2 is the variance of the ith tree. Without loss of generality, assume $\sigma_1^2\le\sigma_2^2\le\cdots\le\sigma_T^2$ Based on Eq. 22, we have

$$\sigma_1^2\|C\|_2^2\le f(C)\le\sigma_T^2\|C\|_2^2 \tag{28}$$

Since $||C||_1=1$ and each entry of $C\ge 0$,

$$\frac{1}{T}\le\|C\|_2^2\le 1 \tag{29}$$

$$\frac{\sigma_1^2}{T} \le f(C) \le \sigma_T^2 \tag{30}$$

When the weights of the trees are equal (ie, $C' = [1/T, 1/T, \cdots, 1/T]$ we have

$$f(C_e) = \frac{\sum_{i=1}^{T} \sigma_i^2}{T^2} \tag{31}$$

Thus, there is always a possibility that for some C,

$$f(C) \le f(C_e) \frac{\sigma_1^2}{T} \le \frac{\sum_{i=1}^{T} \sigma_i^2}{T^2} \tag{32}$$

We can show that the minimum $f(C_{\min})$ in such a scenario is

$$f(C_{\min}) = \frac{1}{\sum_{j=1}^{T} \sigma_j^2} \tag{33}$$

where $C_{\min} = \left[\frac{1}{\sum_{j=1}^{T} \frac{\sigma_1^2}{\sigma_j^2}}, \cdots, \frac{1}{\sum_{j=1}^{T} \frac{\sigma_T^2}{\sigma_j^2}} \right]'$. The derivation is included in the Appendix.

Forest with correlated trees. If we consider scenarios where trees are correlated (ie, covariance matrix is not diagonal), placing higher weights on uncorrelated trees will result in lower variance. We illustrate this idea intuitively for a forest consisting of three trees.

Consider the covariance matrix for a three-tree forest as

$$\begin{bmatrix} a_1 & b_2 & c_2 \\ b_1 & a_2 & d_2 \\ c_1 & d_1 & a_3 \end{bmatrix} \tag{34}$$

where $a_1 \cong a_2 \cong a_3 \cong a, b_1 \cong b_2 \cong b, c_1 \cong c_2 \cong c$, and $d_1 \cong d_2 \cong d$. Consider that the first two trees have high correlation between themselves, while the third tree has little correlation with the other two. Thus, $c_1 \cong c_2 \cong c \cong 0$ and $d_1 \cong d_2 \cong d \cong 0$.

The minimum variance will be achieved for $C = \left[\frac{a}{3a+b}, \frac{a}{3a+b}, \frac{a+b}{3a+b} \right]'$. The details of the derivation are included in the Appendix. Based on numerical weights, we next illustrate the effect of placing higher weight for the third tree on the variance.

Consider $C = \left[\frac{a}{3a+b}, \frac{a}{3a+b}, \frac{a+b}{3a+b} \right]'$ (note that the third tree that is uncorrelated to the other trees has higher weight), then the variance of the forest is given by

$$\begin{bmatrix} \frac{a}{3a+b} & \frac{a}{3a+b} & \frac{a+b}{3a+b} \end{bmatrix} \begin{bmatrix} a & b & 0 \\ b & a & 0 \\ 0 & 0 & a \end{bmatrix} \begin{bmatrix} \frac{a}{3a+b} \\ \frac{a}{3a+b} \\ \frac{a+b}{3a+b} \end{bmatrix} = \frac{a(a+b)}{3a+b} \tag{35}$$

For a regular RF scenario with equal weights $C = \left[\frac{1}{3} \frac{1}{3} \frac{1}{3} \right]'$, the variance of the forest is given by

$$\begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} a & b & 0 \\ b & a & 0 \\ 0 & 0 & a \end{bmatrix} \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix} = \frac{(3a+2b)}{9} \tag{36}$$

We note that $\frac{(3a+2b)}{9} - \frac{a(a+b)}{3a+b} = \frac{2b^2}{9(3a+b)} > 0$ when $b > 0$. Weights for achieving minimum variance for few more scenarios are derived in the Appendix.

Regression Forest Weight Optimization

In this section, we discuss two approaches to select the weights for the ensemble of trees based on MLE and incorporation of tree correlations.

MLE for mixture model. Consider N independent and identically distributed samples (x_i, y_i) for $i = \{1, \cdots, N\}$ used for the generation of the T trees. Let α_1, α_T denote the weights of the trees, then the likelihood (conditional) will be given by:

$$\mathfrak{L}(x_1, y_1, \cdots, x_N, y_N) = \prod_{i=1}^{N} \left(\sum_{j=1}^{T} \alpha_j \left(P(y_1 \mid x_i \phi_j \right) \right) \tag{37}$$

To ensure that $\sum_{j=1}^{T} \alpha_j P(y_i \mid x_i, \phi_j)$ represent a valid probability density function, the weights has to satisfy the following constraints $\alpha_i \ge 0$ for $i = 1, \cdots, T$ and $\sum_{j=1}^{T} \alpha_j = 1$.

If we denote $P(y_i \mid x_i, \phi_j)$ by $\xi_{i,j}$ for $i = 1, 2, \cdots, N$ samples and $j = 1, \cdots, T$ trees, a compact form of representation of the likelihood of the samples as an N length vector $f = [f_1, \cdots, f_N]^T$ is $\mathbf{f} = \boldsymbol{\xi\alpha}$:

$$\begin{bmatrix} f_1 \\ . \\ . \\ f_N \end{bmatrix} = \begin{vmatrix} \xi_{1,1} & \cdots & \xi_{1,K} \\ & \cdots & \\ & \cdots & \\ \xi_{N,1} & \cdots & \xi_{N,K} \end{vmatrix} \begin{bmatrix} \alpha_1 \\ . \\ . \\ \alpha_K \end{bmatrix}$$

The goal is to maximize the product $\prod_{i=1}^{N} f_i$ with constraints $\alpha \ge 0$ and $\sum_{j=1}^{K} \alpha_j - 1 = 0$.

We solve this optimization problem using Matlab *fmincon* function that utilizes an interior point approach to find the minimum of a constrained nonlinear function.

Weight distribution based on correlation of trees for weighted sum of random variables model. Among T trees in the forest, consider that some of the trees can have higher correlation between themselves which can be clustered as groups with high correlations among the trees in a group but have limited correlation between trees in different groups. The purpose is to provide higher weight to the uncorrelated trees as compared with the correlated trees. The algorithmic pseudo code is shown as Algorithm 1.

Algorithm 1. Algorithmic representation of weight selection.

STEP 1: Cluster Trees Based on Correlations

STEP 2: Let the k clusters be $[\alpha_1,\cdots,\alpha_{\rho_1}],.,[\alpha_{T-\rho_k+1}\cdots,\alpha_T]$

STEP 3: Assign equal weight $\frac{1}{k}$ to each cluster

ie, assign weight $\frac{1}{\rho_r^k}$ to each tree in cluster r

To achieve the clustering of the trees, we have applied *hierarchical clustering* with inverse of the covariance between trees as the distance criteria and linkages between clusters decided based on the minimum distance among pairs belonging to the two clusters (single-linkage clustering). The pair of trees that have the smallest distance among all pairs is linked first followed by the next pair and so on. An example of hierarchical ordering with six trees is shown in Figure 2. To generate the final clusters, we have applied a threshold for the inverse covariance and all links below the threshold are considered as separate clusters. The threshold has been taken to be 60% of the average variance of the trees or in other words threshold $= \frac{0.6}{T} tr(\Sigma)$ where $tr(\Sigma)$ denotes the trace of the covariance matrix. As an example, consider the hierarchical ordering in Figure 2 with a threshold of 3.8, which will result in four separate cluster of trees [2,3], [4,5], [1], and [6].

Results

Synthetic dataset. *ML estimate for mixture model.* To evaluate the performance of our algorithm as compared with competing methodologies, we created a synthetic dataset consisting of 100×10 size predictor matrix $\mathbf{X}$ and 100×1 size response vector $\mathbf{Y}$. The predictor variables are randomly generated based on a [0 1] uniform distribution. The response variable is generated based on the predictor variables using Eq. 38 where $\mathbf{N}_1$ denotes a random noise with pdf $\mathbb{N}(0,1)$.

$$\mathbf{Y}_1 = 5\mathbf{x}_3 + 5\mathbf{x}_9 - 1.5\mathbf{x}_1^3 + \mathbf{x}_2 * \mathbf{x}_4 + 2\mathbf{x}_5 + 10\mathbf{x}_6 - 1.5\mathbf{x}_7^2 + \mathbf{x}_6 * \mathbf{x}_8 + 2\mathbf{N}_1 \quad (38)$$

One of our objectives is to check if we are able to reduce the mean square error (MSE) in prediction along with lowering the width of the CI by using MLE of the tree weights. From henceforth, the probabilistic RF with tree weights generated by MLE will be termed as PRF and the probabilistic RF with equal tree weights will be denoted as RF. The weighted random forest approach[16] will be denoted by wRF.

To report our results, we have used 75 samples (75%) for training and the remaining 25 samples (25%) for testing (holdout validation) and compared Pearson correlation coefficients, mean absolute error (MAE), MSE, normalized root mean square error (NRMSE), and width of the CI between predicted and experimental responses for RF, wRF, and PRF models. NRMSE of output response can be calculated as:

$$\text{NRMSE} = \sqrt{\frac{(y-\hat{y})^T(y-\hat{y})}{(y-1\cdot\mathbb{E}(y)^T(y-1\cdot\mathbb{E}(y))}} \quad (39)$$

where y and $\hat{y}$ denote the vector of actual and predicted drug sensitivities, respectively, and $\mathbb{E}(y)$ denote the expectation of vector y. We have considered different number of trees to build the models (RF, wRF, and PRF) and the change in MAE, NRMSE, and correlation coefficients between actual and predicted values for these models are shown in Table 1. Table 1 shows that the prediction errors measured in terms of MAE and NRMSE are smaller with PRF as compared with wRF and RF for different number of trees in the forest. We observe analogous behavior based on a similarity measure with PRF correlation coefficient between predicted and experimental values to be higher than wRF and RF.

The previous measures are based on the mean of the predicted pdf and actual observation. We also consider a probabilistic measure to capture where an actual observation lies in comparison with the predicted pdf. Similar to P-value for doubled tailed event, we considered the probability $\eta(y_i)$ of observing results more extreme than y_i when our prediction probability density function is given by the pdf of $\hat{Y}$.

$$\eta(y_i) = 2\times\min\{\Pr(\hat{Y} \le y_i), \Pr(\hat{Y} \ge y_i)\} \quad (40)$$

A higher value of $\eta(y_i)$ will denote that we cannot reject the hypothesis that the observed responses are from the predicted distributions. A higher value of $\mathbb{E}(\eta(y_i))$, where the

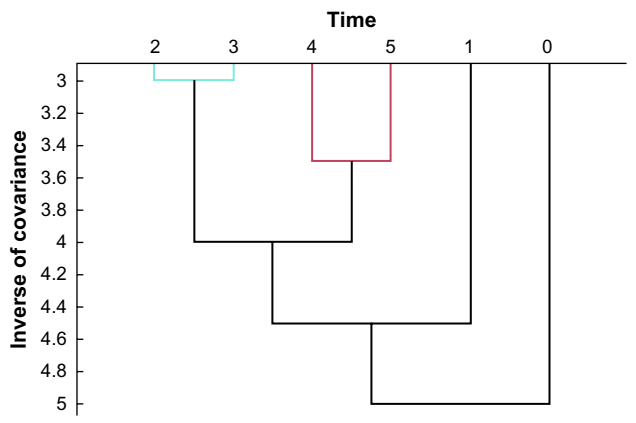

Figure 2. Example of hierarchical clustering.

Table 1. MAE, NRMSE, and correlation between actual and predicted responses for 100 samples for different number of trees in the forest.

TREE	MAE			NRMSE			CORRELATION		
	RF	wRF	PRF	RF	wRF	PRF	RF	wRF	PRF
5	0.1437	0.1431	0.1357	0.7785	0.7759	0.7565	0.6560	0.6600	0.6681
10	0.1126	0.1129	0.1044	0.7623	0.7645	0.7326	0.6535	0.6507	0.6848
20	0.0937	0.0943	0.0876	0.7103	0.7127	0.6747	0.7435	0.7402	0.7627
30	0.1319	0.1318	0.1195	0.6383	0.6371	0.6051	0.8169	0.8188	0.8217
50	0.1240	0.1239	0.1109	0.6737	0.6738	0.6394	0.8151	0.8169	0.8499

Note: Minimum leaf size is 3 and 5 features considered for each split.

expectation is over the testing samples, will be preferable for model comparisons. Table 2 shows $\mathbb{E}(\eta(y_i))$ for different number of trees for RF, wRF, and PRF. Based on Table 2, we observe that $\mathbb{E}(\eta(y_i))$ for PRF is higher than both wRF and RF.

We report the percentage difference in the width of the CI at different confidence levels (CLs) for PRF as compared with RF and wRF in Table 3. An *M%* change in the width of the CI denotes that on an average, PRF generated CI is *M%* lower than the RF generated CI. We observe that the average width of the CI for PRF is lower than RF and wRF for all CLs and different number of trees.

Table 4 shows the NRMSE and correlation coefficient between actual and predicted responses for RF, wRF, and PRF and percentage change in the width of CI with PRF as compared with RF for a simulation with 250 samples. Similar to the previous results with 100 samples, we observe improvement with PRF as compared with both RF and wRF with respect to NRMSE and correlation coefficient between actual and predicted responses. The results also show that the NRMSE has decreased for all the approaches when the sample size has been increased to 250 samples as compared to 100 samples. The absolute difference in performance for PRF as compared with RF and wRF is better for 100 samples, but the percentage improvement in performance is similar for both the sample scenarios.

Weighted sum of random variables. In this section, we consider the effect of tree weights on the MSE and prediction variance for the weighted sum of random variables scenario. We generated a synthetic feature matrix of 500 samples and 1000 features based on a uniform probability distribution [0 1]. The output response has been generated based on Eq. 38 where the output response is dependent on nine of the input features. A random set of 300 of these 500 samples have been used for training, while the remaining 200 samples have been used for testing. We have used the filter feature selection approach *RReliefF*[19] to reduce the initial set of 1000 features to 100 (10 among these 100 are randomly considered for each node splitting) for training the regression trees.

We have considered five trees for the generation of the RF model, and the covariance matrix for the five trees based on the training samples is given by Eq. 41. We have used smaller number of trees for easy visualization of the covariance matrix along with concise analysis of the inferred weights.

$$V = \begin{bmatrix} 0.0322 & 0.0146 & 0.0137 & 0.0126 & 0.0166 \\ 0.0151 & 0.0396 & 0.0202 & 0.0178 & 0.0231 \\ 0.0135 & 0.0202 & 0.0357 & 0.0133 & 0.0174 \\ 0.0126 & 0.0178 & 0.0133 & 0.0336 & 0.0183 \\ 0.0170 & 0.0238 & 0.0177 & 0.0178 & 0.0390 \end{bmatrix} \quad (41)$$

In Eq. 41, the diagonal elements are the variance of each tree with itself, while the nondiagonal elements are covariance between different trees. We note that the covariance between trees 2 and 5 is high compared with the other covariances. By applying hierarchical clustering with inverse covariance as

Table 2. $\mathbb{E}(\eta(y_i))$ for different number of trees in the forest.

TREE	$E(\eta(y_i))$		
	RF	wRF	PRF
5	0.5596	0.5586	0.5617
10	0.6467	0.6467	0.6655
20	0.7049	0.7034	0.7226
30	0.6328	0.6335	0.6281
50	0.6043	0.6042	0.6303

Note: Minimum leaf size is 3 and 5 features considered for each split.

Table 3. Change in CI width for different CLs between RF and PRF and wRF and PRF model for 100 samples for different number of trees in the forest.

TREE	% DECREASE IN MEAN CI COMPARED TO RF					% DECREASE IN MEAN CI COMPARED TO wRF				
	50% CL	70% CL	80% CL	95% CL	99% CL	50% CL	70% CL	80% CL	95% CL	99% CL
5	7.52	9.29	10.13	8.99	6.31	6.49	8.06	8.57	8.20	5.79
10	0.14	0.63	0.78	1.55	1.98	0.72	0.81	1.00	1.55	1.92
20	3.16	2.43	1.93	0.73	0.50	3.16	2.34	1.78	0.64	0.34
30	12.78	13.45	12.53	9.65	7.54	12.92	13.36	13.01	9.65	7.56
50	3.94	2.61	2.55	1.41	1.65	3.79	2.43	2.48	1.32	1.74

Note: Minimum leaf size is 3 and 5 features considered for each split.

Table 4. NRMSE, correlation between actual and predicted output, and change in CI width for different CLs for 250 samples for different number of trees in the forest.

TREE	NRMSE			CORRELATION			% DECREASE IN MEAN CI WITH PRF COMPARED TO RF DIFFERENT CONFIDENCE LEVEL				
	RF	wRF	PRF	RF	wRF	PRF	50%	70%	80%	95%	99%
5	0.5880	0.5871	0.5710	0.8177	0.8182	0.8240	4.66	4.36	4.63	4.61	4.02
10	0.5503	0.5498	0.5299	0.8396	0.8400	0.8529	4.19	4.45	3.92	3.26	2.87
20	0.5746	0.5750	0.5681	0.8243	0.8241	0.8302	3.86	3.07	2.77	2.60	2.79
30	0.5830	0.5832	0.5809	0.8180	0.8178	0.8185	1.36	1.34	1.51	1.35	1.26
50	0.5821	0.5826	0.5626	0.8259	0.8258	0.8380	4.19	3.15	2.90	2.73	2.89

Note: Minimum leaf size is 3 and 5 features considered for each split.

the distance measure and 60% of average variance as threshold, we arrive at four clusters: [2, 5], [1], [3], [4]. We assign equal weights to each cluster (0.25), and where there is more than one tree in a cluster, the weight is equally divided among the trees in the cluster. Thus, we arrive at the following weight vector for PRF model $C = [0.25\ 0.125\ 0.25\ 0.25\ 0.125]$.

Since we considered holdout validation for variance comparison, we generated the covariance among the trees for the testing samples (denoted by Σ) which is shown in Eq. 42.

$$\Sigma = \begin{bmatrix} 0.0363 & 0.0033 & -0.0025 & 0.0029 & 0.0045 \\ 0.0036 & 0.0366 & 0.0008 & 0.0030 & 0.0132 \\ -0.0033 & 0.0006 & 0.0348 & 0.0008 & -0.0008 \\ 0.0030 & 0.0032 & 0.0018 & 0.0330 & 0.0001 \\ 0.0043 & 0.0120 & -0.0010 & 0.0002 & 0.0379 \end{bmatrix} \quad (42)$$

Consequently, the variance of the forest with equal tree weights is given by

$$[1/5\ \ 1/5\ \ 1/5\ \ 1/5\ \ 1/5] * \Sigma * [1/5\ \ 1/5\ \ 1/5\ \ 1/5\ \ 1/5]' = 0.092 \quad (43)$$

whereas the variance of the forest based on our weight selection is given by

$$C * \Sigma * C' = 0.089 \quad (44)$$

The above results illustrate that for the weighted sum of random variables scenario, the variance of the forest prediction can be reduced by generating the weight of the trees based on tree clusters as compared with using equal weights for all trees.

ML estimate of mixture model applied to CCLE dataset. CCLE dataset has been downloaded from http://www.broadinstitute.org/ccle/home. CCLE dataset has two types of genetic characterization information: (i) gene expression and (ii) single-nucleotide polymorphism (SNP6). Gene expression has been downloaded from *CCLE_Expression_Entrez_2012-09-29.gct*. In this dataset, there are 18,988 gene features with no missing values for 1037 cell lines. The SNP6 dataset has been extracted from *CCLE_copynumber_byGene_2013-12-03.txt*. For 1043 cell lines, there are 23,316 features. For our experiments, we have selected 1012 cell lines that are common to both gene expression and SNP6 dataset.

The drug sensitivity data has been downloaded from the addendum published by Barretina et al.[20] The data provide 24 drug responses for 504 cell lines. Drug sensitivity data of the *area under the curve* have been collected from Act Area and normalized to [0 1]. The SNP6 and gene expression data integrated model was constructed based on individual RF models combined with a linear regression stacking approach.[3]

CI and variance. For the calculation of the CI, we have considered 15 drugs in the CCLE database and considered samples with drug sensitivity higher than 0.1 so as to have noticeable variance among the output responses. The number of samples used for the experiments for the 15 drugs varies from 70 to 395. We have used fivefold cross-validation for all our computations, where the data samples are randomly partitioned into five equal parts and four parts are used for training and the remaining part used for testing and the process repeated five times corresponding to the five different testing partitions.

Based on the model inferred from the training samples, the mean and variance of the output of the leaf node for the testing set has been calculated. Thus, for a testing set of 20 samples and 10 trees, we have a matrix of mean and variance of size 20×10. Based on the calculated means and variances, a Gaussian mixture distribution has been derived. Cumulative distribution function has been eventually derived from this distribution to calculate the CIs for different CLs.

To analyze the estimated CIs, we have considered the ratio of the number of experimental testing responses contained in the predicted CI to the total number of testing samples. We will term the ratio as the coverage probability of the CI. Note that we are calculating the coverage probability from cross-validation data as compared with resubstitution data, and thus, there can be significant differences from the CI level for limited samples.

Table 5. Coverage probabilities for four CIs (CL) for PRF and RF predictions for different drugs.

DRUG	COVERAGE PROBABILITY							
	50% CL		70% CL		80% CL		95% CL	
	RF	PRF	RF	PRF	RF	PRF	RF	PRF
17-AAG	0.686	0.650	0.911	0.883	0.977	0.959	1	1
AZD0530	0.829	0.764	0.934	0.929	0.949	0.959	0.994	0.989
AZD6244	0.743	0.712	0.920	0.893	0.951	0.938	0.991	0.986
Erlotinib	0.844	0.836	0.931	0.939	0.982	0.991	0.991	1
Lapatinib	0.838	0.788	0.932	0.898	0.974	0.949	1	1
Nilotinib	0.795	0.742	0.913	0.881	0.956	0.913	0.986	0.989
Nutlin-3	0.872	0.825	0.941	0.953	0.953	0.965	1	1
Paclitaxel	0.707	0.671	0.909	0.886	0.969	0.959	1	0.997
PD-0325901	0.686	0.665	0.893	0.872	0.965	0.944	0.996	0.996
PD-0332991	0.849	0.831	0.973	0.929	0.991	0.964	1	0.991
PF2341066	0.842	0.808	0.931	0.938	0.972	0.965	0.993	1
PHA-665752	0.855	0.842	0.973	0.960	1	1	1	1
PLX4720	0.9	0.785	0.971	0.942	0.985	0.957	0.985	0.985
Sorafenib	0.901	0.862	0.950	0.950	0.980	0.970	0.99	0.99
TAE684	0.816	0.771	0.955	0.948	0.982	0.965	0.996	0.996

Note: We have used $T = 10$ trees and the following constraints for the weights of the trees for PRF model $\frac{1}{3T} \leq \alpha_i \leq 1$ and $\sum_{j=1}^{T} \alpha_j = 1$.

The coverage probability for different CLs for all the 15 drugs is shown in Table 5. We observe that the RF and PRF coverage probabilities are quite similar and PRF coverage probability is closer to the actual CL than the RF coverage probability. As expected, the coverage probability is increasing with the increase in CL for both RF and PRF model.

For the results shown in Table 5, we also calculated the *P*-values of paired *t*-test between PRF and RF predictions and actual responses. The *P*-values of paired *t*-test between (a) PRF prediction and actual responses turned out to be 0.6172 and between (b) RF prediction and actual responses turned out to be 0.6052. A higher value for the PRF scenario represents that the PRF predictions are closer to the actual responses as compared with the RF predictions.

The change in coverage probability with the number of trees (*T*) for drug 17-AAG is shown in Table 6. We observe that the coverage probabilities are closer to the actual CLs with lower number of trees. However, the increase in the number of trees in the forest produces lower variance and higher prediction accuracy.

From Tables 5 and 6, we observe that both RF and PRF provide similar coverage probabilities for the generated CIs.

We next analyzed the error in prediction using different error metrics (MSE, MAE, and NRMSE) and the length of the CIs for PRF in comparison with RF and wRF.[16]

We first explored whether PRF in comparison with RF and wRF can reduce prediction error (as measured by different metrics) while decreasing the CI in majority of the cases. The ratio of the number of testing samples, where the PRF model-generated CI is lower than the RF model-generated CI, to all samples is defined as PRF CI ratio. For example, a PRF CI ratio of 0.60 will denote that for 60% of the testing samples, PRF model-generated CI is lower than RF model-generated CI.

Table 6. Coverage probabilities for four CIs for different number of trees (from 2 to 100) for drug 17-AAG with 395 samples.

NO. OF TREES	COVERAGE PROBABILITY							
	50% CL		70% CL		80% CL		95% CL	
	RF	PRF	RF	PRF	RF	PRF	RF	PRF
T = 2	0.6532	0.6228	0.8228	0.8101	0.8911	0.8886	0.9848	0.9873
T = 5	0.6937	0.6658	0.9038	0.8861	0.9570	0.9418	0.9975	0.9949
T = 10	0.686	0.650	0.911	0.883	0.977	0.959	1	1
T = 20	0.7089	0.6886	0.9139	0.9089	0.9747	0.9722	1	1
T = 100	0.7291	0.7241	0.9342	0.9266	0.9823	0.9772	1	1

Note: Results for both RF and PRF models show similar type of behavior.

Table 7. Performance of all the drugs in terms of MSE, MAE, and NRMSE.

DRUG	MSE			MAE			NRMSE		
	RF	wRF	PRF	RF	wRF	PRF	RF	wRF	PRF
17-AAG	0.0175	0.0167	0.0185	0.1075	0.1051	0.1108	1.0055	0.9828	1.0363
AZD0530	0.0071	0.0066	0.0078	0.0628	0.0605	0.0650	1.0023	0.9637	1.0502
AZD6244	0.0157	0.0160	0.0169	0.0983	0.1018	0.1011	0.9567	0.9642	0.9946
Erlotinib	0.0047	0.0049	0.0057	0.0513	0.0528	0.0573	0.9956	1.0021	1.0894
Lapatinib	0.0070	0.0073	0.0079	0.0629	0.0616	0.0649	0.9799	0.9992	1.0418
Nilotinib	0.0241	0.0226	0.0230	0.1015	0.0931	0.0974	1.0326	1.0021	1.0116
Nutlin-3	0.0034	0.0038	0.0037	0.0435	0.0449	0.0438	0.9762	1.0415	1.0296
Paclitaxel	0.0237	0.0236	0.0243	0.1226	0.1240	0.1257	0.9229	0.9205	0.9354
PD-0325901	0.0259	0.0254	0.0279	0.1312	0.1306	0.1364	0.9534	0.9446	0.9873
PD-0332991	0.0053	0.0045	0.0058	0.0573	0.0524	0.0609	0.9825	0.9139	1.0379
PF2341066	0.0075	0.0074	0.0060	0.0646	0.0603	0.0564	1.0578	1.0458	0.9519
PHA-665752	0.0039	0.0039	0.0039	0.0509	0.0497	0.0488	1.0667	1.0700	1.0616
PLX4720	0.0100	0.0107	0.0106	0.0730	0.0775	0.0720	1.0011	1.0270	1.0283
Sorafenib	0.0072	0.0069	0.0063	0.0568	0.0533	0.0505	1.0471	1.0309	0.9923
TAE684	0.0087	0.0075	0.0089	0.0696	0.0644	0.0707	0.9682	0.8957	0.9759
Average	0.0114	0.0112	0.0118	0.0769	0.0755	0.0774	0.9966	0.9869	1.0149

Note: We have used $T = 10$ and the following constraints for the weight of the trees for the PRF model $\frac{1}{4T} \le \alpha_i \le 1$ and $\sum_{j=1}^{T} \alpha_j = 1$.

The MSE, MAE, and NRMSE for different drugs are shown in Table 7, while Table 8 shows the PRF CI ratio in comparison with RF and wRF for different CLs. Tables 7 and 8 show that the average errors for PRF in comparison with RF and wRF is similar based on multiple error metrics, whereas the PRF CI ratio is >0.5 (between 0.54 and 0.6) for all CLs. Thus, the results support the idea that as compared with using equal weights for all trees, weight optimization using MLE

Table 8. Performance of all the drugs in terms of CI.

DRUG	% DECREASE IN CI COMPARED TO RF					% DECREASE IN CI COMPARED TO wRF				
	50% CL	70% CL	80% CL	95% CL	99% CL	50% CL	70% CL	80% CL	95% CL	99% CL
17-AAG	2.43	2.39	2.11	2.03	2.28	2.41	2.34	2.11	2.00	2.29
AZD0530	3.04	2.20	2.18	2.56	3.15	3.09	2.20	2.17	2.56	3.15
AZD6244	4.6	4.57	4.48	4.76	5.10	4.60	4.61	4.50	4.77	5.12
Erlotinib	–1.92	–2.13	–2.02	–1.34	1.20	–1.81	–2.26	–1.98	–1.27	1.24
Lapatinib	4.25	4.77	4.70	4.04	4.62	4.21	4.89	4.64	3.94	4.65
Nilotinib	7.87	9.07	11.76	13.12	9.60	7.39	9.20	11.87	13.29	9.84
Nutlin-3	8.76	6.76	5.80	5.66	7.73	8.94	6.87	5.92	5.81	7.76
Paclitaxel	3.02	2.97	3.07	3.26	3.35	2.96	2.93	3.12	3.25	3.37
PD-0325901	1.10	1.84	2.07	2.11	2.12	1.10	1.75	2.09	2.05	2.08
PD-0332991	2.23	1.22	0.36	0.74	1.52	2.24	0.93	0.35	0.64	1.46
PF2341066	3.69	4.63	4.26	3.45	3.75	3.56	4.55	4.21	3.43	3.69
PHA-665752	9.65	9.11	9.06	8.79	8.35	9.47	9.01	8.98	8.70	8.37
PLX4720	13.12	11.08	11.85	12.85	11.41	13.13	11.55	12.03	13.12	11.51
Sorafenib	–2.46	–3.15	–3.65	–1.50	2.34	–2.39	–3.23	–3.48	–1.47	2.31
TAE684	4.08	3.90	3.64	3.59	3.62	4.17	3.91	3.65	3.52	3.63

Notes: PRF CI ratio denotes the ratio of samples where PRF CI is lower than RF CI or wRF CI. We have used $T = 10$ and the following constraints for the weight of the trees for the PRF model $\frac{1}{4T} \le \alpha_i \le 1$ and $\sum_{j=1}^{T} \alpha_j = 1$.

can potentially predict drug sensitivity with higher confidence while maintaining similar error. Figure 3 represents two example pdfs generated by PRF and RF, which shows that the PRF predicted distribution has lower variance as compared with RF, while maintaining similar mean.

The percentage decreases in mean CI with PRF as compared with RF and wRF are shown in Table 9. We note that the average CI for PRF is lower than RF and wRF in an overwhelming majority of cases.

We also compared our approach with quantile regression forests (QRFs)[21] that uses nonparametric empirical distributions to model the distributions at the leaf nodes. We observed (results not included) that QRFs can produce smaller CIs than RF and PRF but the coverage probability of PRF is significantly lower. It appears that the empirical distributions based on a few samples can provide smaller variance but has limited coverage that defeats the purpose of designing the CIs.

Prior feature selection. In this experiment, we have used filter feature selection algorithm RRelieff[19] to reduce the initial set of features used for training the RF, PRF, and wRF models. We have considered the CCLE cell lines that are common to all 15 drugs resulting in 396 samples. Features election has been used to reduce the number of features to 50 for each dataset.

Table 10 shows the average errors in terms of MSE and MAE for the 15 drugs with 50 selected features for RF, wRF, and PRF. We observe that PRF performs better in comparison with RF and wRF in terms of both average MSE and MAE.

The prediction performance can also be measured in terms of the bias and variance of the error distributions produced by different predictive models. The bias will be an inverse measure of accuracy, and variance will be an inverse measure of precision. Table 11 shows the bias and variance for RF, wRF, and PRF for different drugs. We note that the average absolute bias (measure of inaccuracy) is lower for PRF (0.0014) as compared with RF (0.0020) and wRF (0.0025). Similarly, the variance (measure of imprecision) for PRF (0.0087) is smaller than variances for RF (0.0096) and wRF (0.0090).

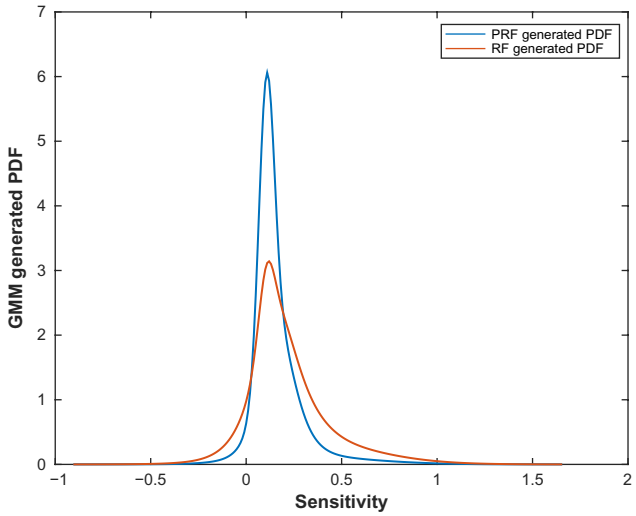

Figure 3. RF generated PDF is more spread out than PRF generated pdf, which implies that the CI of RF generated pdf is higher than PRF generated pdf.

Table 9. Percentage decrease in mean CI with PRF as compared with RF and wRF for 15 drugs of CCLE dataset.

DRUG	% DECREASE IN CI COMPARED TO RF					% DECREASE IN CI COMPARED TO wRF				
	50% CL	70% CL	80% CL	95% CL	99% CL	50% CL	70% CL	80% CL	95% CL	99% CL
17-AAG	2.43	2.39	2.11	2.03	2.28	2.41	2.34	2.11	2.00	2.29
AZD0530	3.04	2.20	2.18	2.56	3.15	3.09	2.20	2.17	2.56	3.15
AZD6244	4.6	4.57	4.48	4.76	5.10	4.60	4.61	4.50	4.77	5.12
Erlotinib	–1.92	–2.13	–2.20	–1.34	1.20	–1.84	–2.26	–1.98	–1.27	1.24
Lapatinib	4.25	4.77	4.70	4.04	4.62	4.21	4.89	4.64	3.94	4.65
Nilotinib	7.87	9.07	11.76	13.12	9.60	7.39	9.20	11.87	13.29	9.84
Nutlin-3	8.76	6.76	5.80	5.66	7.73	8.94	6.87	5.92	5.81	7.76
Paclitaxel	3.02	2.97	3.07	3.26	3.35	2.96	2.93	3.12	3.25	3.37
PD-0325901	1.10	1.84	2.07	2.11	2.12	1.10	1.75	2.09	2.05	2.08
PD-0332991	2.23	1.22	0.36	0.74	1.52	2.24	0.93	0.35	0.64	1.46
PF2341066	3.69	4.63	4.26	3.45	3.75	3.56	4.55	4.21	3.43	3.69
PHA-665752	9.65	9.11	9.06	8.79	8.35	9.47	9.01	8.98	8.70	8.37
PLX4720	13.12	11.08	11.85	12.85	11.41	13.13	11.55	12.03	13.12	11.51
Sorafenib	–2.46	–3.15	–3.65	–1.50	2.34	–2.39	–3.23	–3.48	–1.47	2.31
TAE684	4.08	3.90	3.64	3.59	3.62	4.17	3.91	3.65	3.52	3.63

Notes: Number of features used for each split is 10, minimum number of samples in a leaf node = 5, $T = 10$ and PRF constraints $\frac{1}{4T} \leq \alpha_j \leq 1$ and $\sum_{j=1}^{T} \alpha_j = 1$.

Table 10. Performance of all the drugs in terms of MSE and MAE for PRF compared to RF and wRF.

DRUG	MSE			MAE		
	RF	wRF	PRF	RF	wRF	PRF
17-AAG	0.0179	0.0186	0.0171	0.1088	0.1138	0.1044
AZD0530	0.0103	0.0102	0.0079	0.0839	0.0739	0.0737
AZD6244	0.0162	0.0128	0.0133	0.1037	0.0920	0.0927
Erlotinib	0.0042	0.0043	0.0045	0.0502	0.0524	0.0503
Lapatinib	0.0052	0.0052	0.0058	0.0513	0.0512	0.0565
Nilotinib	0.0056	0.0069	0.0042	0.0527	0.0520	0.0490
Nutlin-3	0.0042	0.0035	0.0030	0.0436	0.0441	0.0450
Paclitaxel	0.0192	0.0219	0.0182	0.1122	0.1180	0.1098
PD-0325901	0.0244	0.0231	0.0230	0.1313	0.1233	0.1218
PD-0332991	0.0050	0.0039	0.0046	0.0535	0.0532	0.0501
PF2341066	0.0046	0.0041	0.0068	0.0445	0.0440	0.0536
PHA-665752	0.0043	0.0046	0.0030	0.0453	0.0458	0.0427
PLX4720	0.0042	0.0030	0.0068	0.0445	0.0413	0.0512
Sorafenib	0.0055	0.0034	0.0031	0.0460	0.0439	0.0443
TAE684	0.0120	0.0093	0.0089	0.0850	0.0758	0.0764
Average	0.0095	0.0090	0.0087	0.0704	0.0683	0.0681

Notes: Number of features used for building the model is 50, and the number of trees considered is 40. $\frac{1}{4T} \le \alpha_i \le 1$ and $\sum_{j=1}^{T} \alpha_j = 1$.

The results given in Tables 10 and 11 show that the PRF provides improvement in terms of average error, accuracy, and precision. We next consider the length of the CI with PRF as compared with RF and wRF. Table 12 shows that the percentage decrease in average CI for PRF when compared with RF and wRF is positive for majority of the drugs.

Conclusions

In this article, we considered the probabilistic analysis of RFs by representing an RF as an ensemble of probabilistic regression trees. The two perspectives that we presented in the manuscript are based on how we would like to treat a probabilistic ensemble of regression trees. We can consider that we would like to select one tree from the available trees conditional on the weights and predict the output response based on the tree distribution resulting in the mixture distribution scenario. The second scenario is where the output response is considered as the weighted average of all the realizations of the trees similar to the averaging of responses from different trees as considered in conventional RF. Thus, if individual trees have large biases (measure of inaccuracy) that are both positive and negative, considering a weighted sum of random variables can provide a better representation. If we consider the mixture distribution approach for this case, selecting an individual tree for each prediction might be unable to remove the bias. However, the mixture distribution approach is reasonable in selecting tree weights to reduce the CIs, while maintaining coverage and MSE as shown in the results presented in this article.

The probabilistic representation presented in this article allowed us to generate and analyze the CIs of individual predictions. We explored various structures of covariance matrices representing the relationships between the generated probabilistic regression trees and the corresponding tree weights that will optimally reduce the variance of prediction. We studied

Table 11. Performance of all the drugs in terms of bias and variance for PRF compared with RF and wRF.

DRUG	BIAS			VARIANCE		
	RF	wRF	PRF	RF	wRF	PRF
17-AAG	0.0023	0.0087	0.0020	0.0182	0.0187	0.0173
AZD0530	−0.0055	−0.0150	−0.0005	0.0104	0.0101	0.0080
AZD6244	−0.0133	0.0190	0.0037	0.0162	0.0126	0.0135
Erlotinib	0.0054	0.0110	−0.0054	0.0042	0.0042	0.0045
Lapatinib	−0.0116	−0.0080	−0.0142	0.0052	0.0052	0.0057
Nilotinib	−0.0062	0.0002	0.0012	0.0057	0.0070	0.0043
Nutlin-3	−0.0102	0.0020	−0.0057	0.0042	0.0036	0.0031
Paclitaxel	0.0040	0.0192	0.0351	0.0194	0.0217	0.0171
PD-0325901	0.0086	0.0025	−0.0048	0.0246	0.0234	0.0232
PD-0332991	0.0118	0.0064	0.0083	0.0050	0.0039	0.0047
PF2341066	0.0012	−0.0053	−0.0003	0.0047	0.0041	0.0069
PHA-665752	−0.0045	−0.0108	0.0058	0.0044	0.0046	0.0030
PLX4720	−0.0027	0.0030	−0.0038	0.0042	0.0030	0.0069
Sorafenib	−0.0081	0.0005	−0.0028	0.0055	0.0034	0.0032
TAE684	−0.0009	0.0048	0.0030	0.0122	0.0094	0.0090
Average	−0.0020	0.0025	0.0014	0.0096	0.0090	0.0087

Notes: Number of features used for building the model is 50, and the number of trees considered is 40. $\frac{1}{4T} \le \alpha_i \le 1$ and $\sum_{j=1}^{T} \alpha_j = 1$.

Table 12. Performance of all the drugs in terms of % decrease in mean CI with PRF as compared with RF and wRF.

DRUG	% DECREASE IN CI COMPARED TO RF					% DECREASE IN CI COMPARED TO wRF				
	50% CL	70% CL	80% CL	95% CL	99% CL	50% CL	70% CL	80% CL	95% CL	99% CL
17-AAG	2.79	2.37	2.48	2.59	2.54	2.87	2.38	2.52	2.61	2.54
AZD0530	3.10	2.47	2.86	2.52	3.40	3.11	2.35	2.98	2.58	3.37
AZD6244	1.52	1.68	1.88	0.87	0.70	1.53	1.83	1.79	0.93	0.79
Erlotinib	3.02	2.13	1.75	1.37	1.63	3.03	2.17	1.76	1.39	1.64
Lapatinib	1.49	1.34	1.29	1.02	1.97	1.37	1.31	1.29	0.96	1.95
Nilotinib	1.62	0.64	0.49	2.12	4.50	1.44	0.57	0.44	2.13	4.49
Nutlin-3	0.01	−0.61	−1.01	−1.18	0.06	0.07	−0.65	−1.01	−1.14	0.08
Paclitaxel	3.92	2.91	2.38	1.92	2.03	3.90	2.79	2.48	1.91	2.01
PD-0325901	−0.09	−0.07	−0.03	−0.49	−0.40	0.00	−0.01	−0.06	−0.43	−0.41
PD-0332991	3.38	3.13	2.58	1.65	2.35	3.39	3.13	2.55	1.57	2.29
PF2341066	5.76	6.09	5.08	5.06	5.88	5.82	6.02	5.15	5.05	5.80
PHA-665752	1.83	0.97	0.37	−0.59	−0.23	1.76	0.93	0.37	−0.66	−0.23
PLX4720	0.80	0.59	0.97	−0.23	0.04	0.80	0.64	0.94	−0.20	0.01
Sorafenib	3.95	4.16	3.57	2.81	4.33	3.95	4.29	3.55	2.87	4.32
TAE684	0.61	0.40	0.02	−0.68	−0.33	0.70	0.44	0.02	−0.62	−0.33

Notes: Number of features used for building the model is 50, and the number of trees considered is $T = 40$. $\frac{1}{4T} \leq \alpha_j \leq 1$ and $\sum_{j=1}^{T} \alpha_j = 1$.

the effect of tree weights generated using MLE on different error measures and prediction CIs. The application of the maximum likelihood estimates of tree weights on the CCLE drug sensitivity prediction problem illustrated the average reduction in CI, while maintaining or lowering MSE. Future research will consider the generation of a probabilistic framework for multivariate RFs along with generation of sufficiency conditions for reduction in CI by optimizing tree weights.

Software Availability

Matlab implementation can be downloaded from https://github.com/razrahman/PRF_codes.git.

Author Contributions

Conceived and designed the experiments: RR, SH, SG, RP. Analyzed the data: RR, SH, RP. Wrote the first draft of the manuscript: RR, RP. Contributed to the writing of the manuscript: RR, SH, RP. Agreed with manuscript results and conclusions: RR, SH, SG, RP. Jointly developed the structure and arguments for the paper: RR, RP. Made critical revisions and approved the final version: RR, RP. All the authors reviewed and approved the final manuscript.

REFERENCES

1. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;67:301–20.
2. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
3. Wan Q, Pal R. An ensemble based top performing approach for nci-dream drug sensitivity prediction challenge. *PLoS One*. 2014;9:e101183.
4. Riddick G, Song H, Ahn S, et al. Predicting *in vitro* drug sensitivity using random forests. *Bioinformatics*. 2011;27:220–24.
5. Jordan M. A statistical approach to decision tree modeling. In: Warmuth M, ed. Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory. ACM Press, New Brunswick, NJ, USA; 1994:13–20.
6. Jordan MI, Jacobs RA. Hierarchical mixtures of experts and the em algorithm. *Neural Comput*. 1994;6:181–214.
7. Devroye L, Gyorfi L, Lugosi G. *A Probabilistic Theory of Pattern Recognition*. Berlin: Springer; 1997.
8. Lanckriet GR, Ghaoui LE, Bhattacharyya C, Jordan MI. A robust minimax approach to classification. *Mach Learn Res*. 2003;3:555–82.
9. Provost F, Domingos P. Tree induction for probability-based ranking. *Mach Learn*. 2003;52:199–215.
10. Boström H. Estimating class probabilities in random forests. In: Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on. IEEE, Cincinnati, Ohio; 2007:211–6.
11. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and Regression Trees*. CRC press, Taylor Francis Ltd, United States; 1984.
12. Breitenbach M, Nielsen R, Grudic GZ. Probabilistic random forests: predicting data point specific misclassification probabilities. Univ. of Colorado at Boulder, Tech. Rep. CU-CS-954-03; 2003.
13. Robnik-Sikonja M. Improving random forests. Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, Dino Pedreschi, editors. In: *Machine Learning: ECML 2004*. Springer, Pisa, Italy; 2004:359–70.
14. Mishina Y, Tsuchiya M, Fujiyoshi H. Computer Vision Theory and Applications (VISAPP), 2014 International Conference on January 5–8, 2014. Lisbon, Portugal, Publisher: IEEE *Boosted Random Forest*. 2:594–8.
15. Biau G. Analysis of a random forests model. *J Mach Learn Res*. 2012;13:1063–95.
16. Winham SJ, Freimuth RR, Biernacka JM. A weighted random forests approach to improve predictive performance. *Stat Anal Data Min*. 2013;6:496–505.
17. Newey WK, West KD. A Simple, Positive Semi-Definite, Heteroskedasticity and Auto-Correlation Consistent Covariance Matrix; 1986. *Econometrica*. 55(3), May 1987:703–8.
18. Higham NJ. Analysis of the Cholesky Decomposition of a Semi-Definite Matrix; Reliable Numerical Computation. University Press; 1990:161–85.
19. Robnik-Sikonja M, Kononenko I. An adaptation of relief for attribute estimation in regression. In: Proceedings of the Fourteenth International Conference on Machine Learning ICML '97. San Francisco, CA: Morgan Kaufmann Publishers Inc; 1997:296–304.
20. Barretina J, Caponigro G, Stransky N, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483:603–7.
21. Meinshausen N. Quantile regression forests. *J Mach Learn Res*. 2006;7:983–99.

Appendix

Effect of tree weight on variance. Consider the incorporation of the constraint of $C\left(\sum_{i=1}^{T} C=1\right)$ in f(c), then according to the Lagrange multiplier equations $f(c)$ will be given by

$$\begin{aligned} f(C) &= C^T V C + \lambda\left(\sum_{i=1}^{T} C - 1\right) \\ &= \sum_{i=1}^{T}\sum_{j=1}^{T} \alpha_i V_{ij} \alpha_j + \lambda\left(\sum_{i=1}^{T} \alpha_i - 1\right) \end{aligned} \tag{45}$$

Taking the derivative of Eq. 45 with respect to each weight α_k:

$$\frac{\partial f}{\partial \alpha_k} = \sum_{j=1}^{T} 2\alpha_j V_{kj} + \lambda = 0 \tag{46}$$

where $V_{kj} = V_{jk}$.

Thus,

$$-\frac{\lambda}{2} = \sum_{j=1}^{T} \alpha_j V_{kj} \tag{47}$$

and

$$C = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_T \end{bmatrix} = -\frac{\lambda}{2} V^{-1} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{48}$$

Since, $e'C = 1$ where $e = [1,1,\ldots 1]'$, we have

$$1 = -\frac{\lambda}{2} e' V^{-1} e \tag{49}$$

By solving Eq. 49, we arrive at the value of λ to be substituted in Eq. 45 to generate the optimized C.

Diagonal covariance matrix with unequal variances. Here, we consider the specific case where the covariance matrix is a diagonal matrix with unequal variances. Then, the covariance matrix will be similar to Eq. 50.

$$V = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_T^2 \end{bmatrix} \tag{50}$$

where σ_i^2 is the variance of the ith tree. Without loss of generality, assume $\sigma_1^2 \le \sigma_2^2 \le \cdots \le \sigma_T^2$. The Lagrange multiplier equations will be:

$$f(C) = \sum_{i=1}^{T} \sigma_i^2 \alpha_i^2 + \lambda\left(\sum_{i=1}^{T} \alpha_i - 1\right) \tag{51}$$

Differentiating $f(C)$ with respect to α_i,

$$\frac{\partial f(C)}{\partial \alpha_i} = 2\alpha_i \sigma_i^2 + \lambda = 0 \tag{52}$$

$$\lambda = -2\alpha_i \sigma_i^2 \tag{53}$$

$f(C)$ will achieve the minimum when $\alpha_1\sigma_1^2 = \alpha_2\sigma_2^2 = \cdots = \alpha_T\sigma_T^2 = \gamma$, where γ is some constant.

$$f(c)_{\min} = \sum_{i=1}^{T} \sigma_i^2 \alpha_i^2 = \sum_{i=1}^{T} \alpha_i \gamma = \gamma \tag{54}$$

as $\sum_{i=1}^{T} \alpha_i = 1$. Now, $\alpha_i = \frac{\gamma}{\sigma_i^2}$, so

$$\sum_{i=1}^{T} C_i = \gamma \sum_{j=1}^{T} \frac{1}{\sigma_j^2} = 1 \tag{55}$$

$$\gamma = \frac{1}{\sum_{j=1}^{T} \sigma_j^2} \tag{56}$$

Substituting the value of γ in $\alpha_i = \frac{\gamma}{\sigma_i^2}$, we have

$$\alpha_i = \frac{1}{\sum_{j=1}^{T} \frac{\sigma_i^2}{\sigma_j^2}} \tag{57}$$

Eq. 57 provides the weight of the trees that produce the lowest variance for the forest.

Correlated trees. In this section, we consider various forms of covariance structures between the trees.

At first, consider a case where the first two trees are correlated among themselves, while there is limited correlation between the other trees and variance of all trees are the same. The covariance matrix (V) for such a scenario will be as follows:

$$V = \begin{bmatrix} \sigma^2 & \rho & \cdots & 0 \\ \rho & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \tag{58}$$

The variance will be given by the following equation:

$$C'VC = \sigma^2 \sum_{i=1}^{T} \alpha_i^2 + 2\rho\alpha_1\alpha_2 \tag{59}$$

Our objective function to minimize is provided next:

$$f(C) = \sigma^2 \sum_{i=1}^{T} \alpha_i^2 + 2\rho\alpha_1\alpha_2 + \lambda\left(\sum_{i=1}^{T} \alpha_i - 1\right) \tag{60}$$

where λ is the Lagrange's multiplier. α_1 and α_2 are the weights of the trees that are correlated, while α_i for $3 \leq i \leq T$ are the weights of the other trees that are not correlated.

Differentiating Eq. 60 with respect to α_1, α_2, and α_i for $i > 2$, we have

$$\frac{\partial f}{\partial \alpha_1} = 2\alpha_1\sigma^2 + 2\rho\alpha_2 + \lambda = 0 \tag{61}$$

$$\frac{\partial f}{\partial \alpha_2} = 2\alpha_2\sigma^2 + 2\rho\alpha_1 + \lambda = 0 \tag{62}$$

$$\frac{\partial f}{\partial \alpha_i} = 2\alpha_i\sigma^2 + \lambda = 0;\ i > 2 \tag{63}$$

Rearranging Eqs. 61, 62, and 63,

$$\alpha_1 = \frac{-\lambda}{2(\sigma^2 + \rho)} = \alpha_2 \tag{64}$$

$$\alpha_i = \frac{-\lambda}{2\sigma^2};\ i > 2 \tag{65}$$

$$\sum_{i=1}^{T} \alpha_i = \frac{-\lambda}{2}\left[\frac{T-2}{\sigma^2} + \frac{2}{\sigma^2 + \rho}\right] = 1 \tag{66}$$

From Eq. 66,

$$-\lambda = \frac{2\sigma^2(\sigma^2 + \rho)}{T\sigma^2 + \rho(T-2)} \tag{67}$$

Substituting the value of Lagrange's multiplier into Eqs. 64 and 65,

$$\alpha_1 = \frac{\sigma^2}{T\sigma^2 + \rho(T-2)} = \alpha_2 \tag{68}$$

$$\alpha_i = \frac{\sigma^2 + \rho}{T\sigma^2 + \rho(T-2)};\ i > 2 \tag{69}$$

As an example, if $\rho = \sigma^2$ and the number of trees is 3, then $C_1 = C_2 = 1/4$ and $C_3 = 1/2$ will generate the minimum variance. For general case with covariance matrix as in Eq. 58, $C_1 = C_2 = \frac{1}{2(T-1)}$ and $C_i = \frac{1}{(T-1)}$ where $i > 2$.

Let us consider another general case with equal variance for all trees and the first two and last two trees of the forest are correlated among themselves, and the remaining trees have limited correlations among each other. The covariance matrix (V) will have the structure as follows:

$$V = \begin{bmatrix} \sigma^2 & \rho_1 & \cdots & 0 & 0 \\ \rho_1 & \sigma^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 & \rho_2 \\ 0 & 0 & \cdots & \rho_2 & \sigma^2 \end{bmatrix} \tag{70}$$

The variance is given by

$$C'VC = \sigma^2 \sum_{i=1}^{T} \alpha_i^2 + 2\rho_1\alpha_1\alpha_2 + 2\rho_2\alpha_{T-1}\alpha_T \tag{71}$$

and the function to be minimized is

$$f(C) = \sigma^2 \sum_{i=1}^{T} \alpha_i^2 + 2\rho_1\alpha_1\alpha_2 + 2\rho_2\alpha_{T-1}\alpha_T + \lambda\left(\sum_{i=1}^{T} \alpha_{i-1}\right) \tag{72}$$

where λ is the Lagrange's multiplier. Differentiating Eq. 72 with respect to α_1, α_2, α_{T-1}, α_T, and α_i ($i \neq 1,2,T-1,T$)

$$\frac{\partial f}{\partial \alpha_1} = 2\alpha_1\sigma^2 + 2\rho_1\alpha_2 + \lambda = 0 \tag{73}$$

$$\frac{\partial f}{\partial \alpha_2} = 2\alpha_2\sigma^2 + 2\rho_1\alpha_1 + \lambda = 0 \tag{74}$$

$$\frac{\partial f}{\partial \alpha_{T-1}} = 2\alpha_{T-1}\sigma^2 + 2\rho_2\alpha_T + \lambda = 0 \tag{75}$$

$$\frac{\partial f}{\partial \alpha_T} = 2\alpha_T\sigma^2 + 2\rho_2\alpha_{T-1} + \lambda = 0 \tag{76}$$

$$\frac{\partial f}{\partial \alpha_i} = 2\alpha_i\sigma^2 + \lambda = 0;\ i \neq 1,2,T-1,T \tag{77}$$

Based on Eqs. 73 and 74, if $\sigma^2 \neq \rho_1$, then $\alpha_1 = \alpha_2$. While based on Eqs. 75 and 76, if $\sigma^2 \neq \rho_2$, then $\alpha_{T-1} = \alpha_T$. Applying these results in the above equation, we arrive at

$$\alpha_1 = \frac{-\lambda}{2(\sigma^2 + \rho_1)} = C_2 \tag{78}$$

$$\alpha_{T-1} = \frac{-\lambda}{2(\sigma^2 + \rho_2)} = C_T \tag{79}$$

$$\alpha_i = \frac{-\lambda}{2\sigma^2};\ i \neq 1,2,T-1,T \tag{80}$$

$$\sum_{i=1}^{T} \alpha_i = -\lambda\left[\frac{T-4}{2\sigma^2} + \frac{1}{\sigma^2 + \rho_1} + \frac{1}{\sigma^2 + \rho_2}\right] = 1 \tag{81}$$

If $\rho_1 = \rho_2 = \rho$, then solving Eq. 81 provides the value of λ.

$$-\lambda = \frac{2\sigma^2(\sigma^2 + \rho)}{(T-4)(\sigma^2 + \rho) + 4\sigma^2} \tag{82}$$

Substituting the value of Lagrange's multiplier in Eqs. 78, 79, and 80,

$$\alpha_1 = \alpha_2 = \alpha_{T-1} = \alpha_T = \frac{\sigma^2}{(T-4)(\sigma^2+\rho)+4\sigma^2} \tag{83}$$

$$\alpha_i = \frac{\sigma^2+\rho}{(T-4)(\sigma^2+\rho)+4\sigma^2}; i \neq 1,2,T-1,T \tag{84}$$

The minimization solutions show that the general trend is to increase the weight of trees that are uncorrelated to other trees and reduce the weight of the trees that are correlated with other trees.