Alakus *et al. BMC Bioinformatics* (2023) 24:258
https://doi.org/10.1186/s12859-023-05377-y

BMC Bioinformatics

RESEARCH Open Access

# Covariance regression with random forests

Cansu Alakus[1*], Denis Larocque[1] and Aurélie Labbe[1]

*Correspondence: cansu.alakus@hec.ca

[1] Department of Decision Sciences, HEC Montréal, Montréal, Canada

## Abstract

Capturing the conditional covariances or correlations among the elements of a multivariate response vector based on covariates is important to various fields including neuroscience, epidemiology and biomedicine. We propose a new method called Covariance Regression with Random Forests (CovRegRF) to estimate the covariance matrix of a multivariate response given a set of covariates, using a random forest framework. Random forest trees are built with a splitting rule specially designed to maximize the difference between the sample covariance matrix estimates of the child nodes. We also propose a significance test for the partial effect of a subset of covariates. We evaluate the performance of the proposed method and significance test through a simulation study which shows that the proposed method provides accurate covariance matrix estimates and that the Type-1 error is well controlled. An application of the proposed method to thyroid disease data is also presented. `CovRegRF` is implemented in a freely available R package on CRAN.

**Keywords:** Covariance regression, Multivariate response, Random forests, Variable importance

## Introduction

Most existing multivariate regression analyses focus on estimating the conditional mean of the response variable given its covariates. For example, in traditional regression analysis, the expectation of the response variables is related to linear combinations of covariates. While estimating the conditional covariances or correlations among multiple responses based on covariates is also important, it is a less studied problem. For example, functional brain connectivity focuses on the exploration of the co-occurrence of brain activity in different brain regions, and this co-variability can be explained as a function of covariates [1]. As another example, human biomarkers such as glucose, cholesterol, iron, albumin, and so on, are important for biomedical research and the covariance of these biomarkers is influenced by age [2]. In microbiome studies, the changes in the co-occurrence patterns among taxa with respect to the covariates have been studied [3, 4]. In tasks of cognitive and physical performance, a research question is whether the correlation between speed and accuracy is influenced by other covariates, such as sustained attention or age [5]. In neuroscience, the associations of functional criticality with intelligence can be affected by age [6]. In all these examples, the main goal could be just to estimate the conditional covariance

between multiple responses, e.g. for microbiome data, the goal is to estimate network changes with respect to a set of covariates. Another interesting application of estimating covariance matrices based on covariates is to verify the homoscedasticity assumption in classical multivariate regression. In cases where testing the effect of covariates on the covariability of the response variables leads to the rejection of the null hypothesis, conditional estimates of the covariance matrix can be used to have valid inference, for example to build multivariate confidence or prediction regions.

In general terms, let $\mathbf{Y}_{n\times q}$ be a matrix of $q$ response variables measured on $n$ observations, where $\mathbf{y}_i^\top$ represents the $i$th row of $\mathbf{Y}$. Similarly, let $\mathbf{X}_{n\times p}$ be a matrix of $p$ covariates available for all $n$ observations, where $\mathbf{x}_i^\top$ represents the $i$th row of $\mathbf{X}$. For an observation with covariates $\mathbf{x}_i$ and responses $\mathbf{y}_i$, the goal is to estimate the conditional covariance of the response variables $\mathrm{Cov}[\mathbf{y}_\mathrm{i}|\mathbf{x}_\mathrm{i}] \triangleq \Sigma_{\mathbf{x}_\mathrm{i}}$, which is a measurable matrix function of covariates $\mathbf{x}_i$, and to analyze how this conditional covariance matrix varies with respect to the covariates. For this problem, [7] use a kernel estimator to estimate the conditional covariance matrix for a single continuous covariate. However, it is not clear how to extend this approach to situations with multiple covariates. [8] propose a linear covariance regression model

$$\mathbf{y}_i = (\mathbf{A} + \gamma_i \mathbf{B}) \begin{pmatrix} 1 \\ \mathbf{x}_i \end{pmatrix} + \boldsymbol{\epsilon}_i,$$

where the mean and covariance of the multivariate response is parameterized as functions of covariates. This model can also be interpreted as a special random-effects model where $\mathbf{A}_{q\times(p+1)}$ and $\mathbf{B}_{q\times(p+1)}$ characterize the fixed and random parts of the model, respectively. The scalar $\gamma_i$ can be interpreted as an individual-level variability in addition to the random error $\boldsymbol{\epsilon}_i$. The rows of $\mathbf{B}$ indicate how much this additional variability affects $\mathbf{y}_i$. The vector $\boldsymbol{\epsilon}_i$ is of dimension $q \times 1$ and is assumed to be normally distributed. In this framework, they assume that $E[\gamma_i] = 0$, $E[\boldsymbol{\epsilon}_i] = 0$, $E[\gamma_i \boldsymbol{\epsilon}_i] = 0$, $Var[\gamma_i] = 1$, $Var[\boldsymbol{\epsilon}_i] = \boldsymbol{\Psi}$, leading to the following covariance matrix

$$\boldsymbol{\Sigma}_{\mathbf{x}_i} = \boldsymbol{\Psi} + \mathbf{B} \begin{pmatrix} 1 \\ \mathbf{x}_i \end{pmatrix} \begin{pmatrix} 1 \\ \mathbf{x}_i \end{pmatrix}^\top \mathbf{B}^\top.$$

[9] illustrate an application of this model with a four-dimensional health outcome. [10] propose a Bayesian nonparametric model for covariance regression within a high-dimensional response context. Their approach relates the high-dimensional multivariate response set to a lower-dimensional subspace through covariate-dependent factor loadings obtained with a latent factor model. The conditional covariance matrix is a quadratic function of these factor loadings. The method is limited to data sets with smaller sample sizes. [11] proposes a parametric Bayesian model for high-dimensional responses. In this model, the conditional covariance matrices vary with continuous covariates. [12] propose another covariance regression model where the covariance matrix is linked to the linear combination of similarity matrices of covariates. [13] propose a covariance regression method called Covariate Assisted Principal Regression (CAPR). Unlike the other covariance regression methods described in this section, the CAPR aims to find a linear projection of the multivariate response data such that the covariates can best describe

the data variation in the projected space. The model assumes that in the eigendecomposition of covariance matrices, all covariance matrices in the sample are diagonalized by the same orthogonal matrix which results in a restrictive covariance matrix form.

In this study, we propose a nonparametric covariance regression method for estimating the covariance matrix of a multivariate response given a set of covariates, using a random forest framework. The above-mentioned methods are very useful in modeling covariance matrix but compared to them the proposed method offers higher flexibility in estimating the covariance matrix given the set of covariates. For example, with the proposed method, we can estimate the conditional covariance matrix for a set of covariates including multiple continuous and categorical variables, and the proposed method can be used to capture complex interaction patterns with the set of covariates. Moreover, the proposed method is nonparametric and needs less computational time compared to the parametric models, and can be applied to data sets with larger sample sizes.

Random forest [14] is an ensemble tree-based algorithm involving many decision trees, and can also be seen as an adaptive nearest neighbour predictor [15–21]. In the proposed random forest framework, we grow each tree with a splitting rule specially designed to maximize the difference in the sample covariance of $\mathbf{Y}$ between child nodes. For a new observation $\mathbf{y}^*$ with covariates $\mathbf{x}^*$, the proposed random forest finds the set of nearest neighbour observations among the out-of-bag (OOB) observations that are not used in the tree growing process. This set of nearest neighbour observations is then used to estimate the conditional covariance matrix of $\mathbf{y}^*$ given $\mathbf{x}^*$. In each tree built in the proposed random forest framework, the set of covariates is used to find subgroups of observations with similar conditional covariance matrices, assuming that they are related to conditional covariance matrices. We propose a hypothesis test to evaluate the effect of a subset of covariates on the estimated covariance matrices while controlling for the others. We investigate two particular cases, the global effect of the covariates and the partial effect of a single covariate.

This paper is organized as follows. In Section Method, we give the details of the proposed method, significance test and variable importance measure. The simulation study results for accuracy evaluation, global and partial effects of covariates, and variable importance are presented in Section Simulations. We provide a real data example in Section Real data example, and conclude with some remarks in Section Concluding remarks.

## Method

Let $\mathbf{\Sigma}_{\mathbf{x}_i}$ be the true conditional covariance matrix of $\mathbf{y}_i$ based on covariates $\mathbf{x}_i$, and $\mathbf{\Sigma}_{\mathbf{X}}$ be the collection of all conditional covariance matrices for $n$ observations, $\mathbf{\Sigma}_{\mathbf{X}} = \{\mathbf{\Sigma}_{\mathbf{x}_i} : i = 1, \ldots, n\}$. Similarly, let $\hat{\mathbf{\Sigma}}_{\mathbf{x}_i}$ be the estimated conditional covariance matrix of $\mathbf{y}_i$ based on covariates $\mathbf{x}_i$, and $\hat{\mathbf{\Sigma}}_{\mathbf{X}}$ be the collection of all estimated conditional covariance matrices for $n$ observations, $\hat{\mathbf{\Sigma}}_{\mathbf{X}} = \{\hat{\mathbf{\Sigma}}_{\mathbf{x}_i} : i = 1, \ldots, n\}$. In this section, we describe the proposed method in detail.

### Tree growing process and estimation of covariance matrices for new observations with random forests

We aim to train a random forest with the set of covariates $\mathbf{X}$ to find subgroups of observations with similar covariance matrices of $\mathbf{Y}$, based on many unsupervised decision trees built with a specialized splitting criterion. The tree growing process follows the

CART approach [22]. The basic idea of the CART algorithm is to select the best split at each parent node among all possible splits, all evaluated with a selected splitting criterion, to obtain the purest child nodes. The algorithm evaluates all possible splits to determine the split variable and split point. Instead of considering all possible splits at each parent node, the best split search in random forests is confined to a randomly chosen subset of covariates that varies from node to node. The splitting process continues until all nodes are terminal.

Our goal is to obtain subgroups of observations with distinct covariance matrices. Hence, we propose a customized splitting rule that will seek to increase the difference in covariance matrices between two child nodes in the tree [17, 20, 21, 23]. We define $\mathbf{\Sigma}^L$ as the sample covariance matrix estimate of the left node as follows:

$$\mathbf{\Sigma}^L = \frac{1}{n_L - 1} \sum_{i \in t_L} \left(\mathbf{y}_i - \bar{\mathbf{Y}}_L\right)\left(\mathbf{y}_i - \bar{\mathbf{Y}}_L\right)^\top,$$

where $t_L$ is the set of indices of the observations in the left node, $n_L$ is the left node size and $\bar{\mathbf{Y}}_L = \frac{1}{n_L} \sum_{i \in t_L} \mathbf{y}_i$. The sample covariance matrix estimate of the right node, $\mathbf{\Sigma}^R$, is computed in the same way, where $n_R$ is the right node size. The proposed splitting criterion is

$$\sqrt{n_L n_R} * d(\mathbf{\Sigma}^L, \mathbf{\Sigma}^R), \tag{1}$$

where $d(\mathbf{\Sigma}^L, \mathbf{\Sigma}^R)$ is the Euclidean distance between the upper triangular part of the two matrices and computed as follows:

$$d(\mathbf{D}, \mathbf{E}) = \sqrt{\sum_{i=1}^{q} \sum_{j=i}^{q} \left(\mathbf{D}_{ij} - \mathbf{E}_{ij}\right)^2}, \tag{2}$$

where $\mathbf{D}_{q \times q}$ and $\mathbf{E}_{q \times q}$ are symmetric matrices. The best split among all possible splits is the one that maximizes (1).

The final covariance matrices are estimated based on the random forest. For a new observation, we use the nearest neighbour observations to estimate the final covariance matrix. The idea of finding the nearest neighbour observations, a concept very similar to the 'nearest neighbour forest weights' [15, 16], was introduced in [17] and later used in [18–21]. [19] called this set of observations the Bag of Observations for Prediction (BOP).

For a new observation $\mathbf{x}^*$, we form the set of nearest neighbour observations with the out-of-bag (OOB) observations [24, 25]. We can define the $BOP_{oob}$ for a new observation as

$$BOP_{oob}(\mathbf{x}^*) = \bigcup_{b=1}^{B} O_b(\mathbf{x}^*),$$

where $B$ is the number of trees and $O_b(\mathbf{x}^*)$ is the set of OOB observations in the same terminal node as $\mathbf{x}^*$ in the $b$th tree. Each tree is built with a selected random sub-sample instead of a bootstrap sample, i.e. in-bag observations ($I_b$), which has 63.2 percent distinct observations from the original sample. The remaining training observations,

namely $O_b$, are OOB observations for that tree and are not used to build the $b$th tree. $BOP_{oob}$ is slightly different than the nearest neighbour sets in the previous papers who use in-bag observations to form BOP. Since the OOB observations are not used in the tree building process, for the trees where they are OOB, they act as new observations. Therefore, OOB observations represent a new observation better than in-bag observations. Using OOB observations for neighbourhood construction is similar to the idea of honesty in the context of forests. An honest double-sample tree splits the training subsample into two parts: one part for tree growing and another part for estimating the desired response [26]. We use the nearest neighbour construction idea to estimate the covariance matrices for the new observations. Algorithm 1 describes how to estimate the covariance matrix with OOB observations for a new or training observation. After training the random forest with the specialized splitting criterion, for a new observation $\mathbf{x}^*$, we form $BOP_{oob}(\mathbf{x}^*)$ and then we estimate the covariance matrix by computing the sample covariance matrix of the observations in $BOP_{oob}(\mathbf{x}^*)$. See the Supplementary figures 1 and 2 in the Additional file 1 for the results of the simulation study comparing different ways of estimating the final covariance matrix.

### `nodesize` tuning

The number of observations in the nodes decreases as we progress down the tree during the tree-building process. The `nodesize` parameter is the target average size for the terminal nodes. Lowering this parameter results in deeper trees, which means more splits until the terminal nodes. Tuning the `nodesize` parameter can potentially improve the prediction performance [16].

In typical supervised problems where the target is the observed true response, random forests search for the optimal level of the `nodesize` parameter by using out-of-bag (OOB) prediction errors computed using the true responses and OOB predictions. The `nodesize` value with the smallest OOB error is chosen. However, in our problem, the target is the conditional covariance matrix which is unknown. Therefore, we propose a heuristic method for tuning the `nodesize` parameter. For `nodesize` tuning, we use the OOB covariance matrix estimates, as described in Algorithm 1.

**Algorithm 1** Estimation of covariance matrix for a new or training observation

**Input:** A forest built with the proposed method
1: $BOP_{oob}(\mathbf{x}_i) = \emptyset$
2: **for** b=1,...,B **do**
3: **if** $\mathbf{x}_i$ is a new observation **or** ($\mathbf{x}_i$ is a training observation **and** $\mathbf{x}_i \in O_b$) **then**
4: Find the terminal node of $\mathbf{x}_i$ at tree $b$, say $d$
5: $BOP_{oob}(\mathbf{x}_i) = BOP_{oob}(\mathbf{x}_i) \cup O_b^d(\mathbf{x}_i)$ (where $O_b^d(\mathbf{x}_i)$ is the set of OOB observations in the same terminal node $d$ as $\mathbf{x}_i$, excluding $\mathbf{x}_i$ itself when $\mathbf{x}_i$ is a training observation)
6: **end if**
7: **end for**
8: Compute sample covariance matrix with the observations in $BOP_{oob}(\mathbf{x}_i)$

The general idea of the `nodesize` tuning method is to find the `nodesize` level where the average difference between OOB covariance matrix predictions at two consecutive `nodesize` levels is the smallest among the set of `nodesize` values. We first train separate random forests for a set of `nodesize` values (see the Parameter settings section in simulation study). Then, we compute the OOB covariance matrix estimates as described

in Algorithm 1 for each random forest. Define $MAD(\mathbf{D}, \mathbf{E}) = \frac{2}{q(q+1)} \sum_{i=1}^{q} \sum_{j=i}^{q} |\mathbf{D}_{ij} - \mathbf{E}_{ij}|$. Let $\hat{\mathbf{\Sigma}}_{\mathbf{x}_i}^{s}$ be the estimated covariance matrix for observation $i$ when `nodesize`$= s$. Let $s(1) < \ldots < s(M)$ be a set of increasing node sizes. For $j = \{1, \ldots M-1\}$, let

$$MAD_j = \frac{1}{n} \sum_{i=1}^{n} MAD\left(\hat{\mathbf{\Sigma}}_{\mathbf{x}_i}^{s(j)}, \hat{\mathbf{\Sigma}}_{\mathbf{x}_i}^{s(j+1)}\right).$$

Then we select $s(j)$ that corresponds to the value $j$ for which $MAD_j$ is the minimum among $\{MAD_1, \ldots, MAD_M\}$. See the Additional file 2 for the results of a `nodesize` tuning experiment and the illustration of the process with an example.

When a node sample size $n_d$ is smaller than the number of responses $q$, the sample covariance matrix becomes highly variable. In fact, if $n_d - 1 < q$, the estimate is singular and hence non-invertible. Therefore, the tuning set of `nodesize` levels should be larger than $q$. In fact, we need more than $q$ distinct values, so we use sub-sampling instead of bootstrap resampling for each tree building step of the proposed method to guarantee distinctness, assuming the observations in the original sample are distinct.

### Significance test

The proposed method uses covariates to find groups of observations with similar covariance matrices with the assumption that the set of covariates is important to distinguish between these covariance matrices. However, some (or all) covariates might not be relevant. In this paper, we propose a hypothesis test to evaluate the effect of a subset of covariates on the covariance matrix estimates, while controlling for the other covariates.

If a subset of covariates has an effect on the covariance matrix estimates obtained with the proposed method, then the conditional covariance matrix estimates given all covariates should be significantly different from the conditional covariance matrix estimates given the controlling set of covariates. We propose a hypothesis test to evaluate the effect of a subset of covariates on the covariance matrix estimates for the null hypothesis

$$H_0 : \mathbf{\Sigma}_{\mathbf{X}} = \mathbf{\Sigma}_{\mathbf{X}^c}, \tag{3}$$

where $\mathbf{\Sigma}_{\mathbf{X}}$ is the conditional covariance matrix of $\mathbf{Y}$ given all $X$ variables, and $\mathbf{\Sigma}_{\mathbf{X}^c}$ is the conditional covariance matrix of $\mathbf{Y}$ given only the set of controlling $X$ variables. The proposed significance test is described in Algorithm 2. After computing the covariance matrix estimates for all covariates and control variables only, we compute the test statistic with

$$T = \frac{1}{n} \sum_{i=1}^{n} d\left(\hat{\mathbf{\Sigma}}_{\mathbf{x}_i}, \hat{\mathbf{\Sigma}}_{\mathbf{x}_i^c}\right), \tag{4}$$

where $d(.,.)$ is computed as (2). The test statistic specifies how much the covariance matrix estimates given all covariates differ from the estimates given only the controlling set of covariates. As $T$ becomes larger, we have more evidence against $H_0$.

We conduct a permutation test under the null hypothesis (3) by randomly permuting rows of $\mathbf{X}$. Let $R$ be the total number of permutations and $T_r$ be the global test statistic (4) computed for the $r$th permuted $\mathbf{X}$. We estimate the test $p$-value with

$$p = \frac{1}{R}\sum_{r=1}^{R} I(T_r > T), \tag{5}$$

and we reject the null hypothesis (3) at a pre-specified level $\alpha$ if the $p$-value is less than $\alpha$.

**Algorithm 2** Permutation test for a subset of covariates effect

Train RF with $\mathbf{X}$ and $\mathbf{Y}$, estimate covariance matrices as described in Algorithm 1, say $\hat{\mathbf{\Sigma}}_{\mathbf{x}_i}$ $\forall i = \{1, \ldots, n\}$
Train RF with $\mathbf{X}^c$ and $\mathbf{Y}$, and estimate covariance matrices as described in Algorithm 1, say $\hat{\mathbf{\Sigma}}^c_{\mathbf{x}_i}$ $\forall i = \{1, \ldots, n\}$
Compute test statistic with $T = \frac{1}{n}\sum_{i=1}^{n} d\left(\hat{\mathbf{\Sigma}}_{\mathbf{x}_i}, \hat{\mathbf{\Sigma}}^c_{\mathbf{x}_i}\right)$, where $d(.,.)$ is computed as (2)
**for** $r = 1 : R$ **do**
    Permute rows of $\mathbf{X}$, say $\mathbf{X}_r$
    Train RF with $\mathbf{X}_r$ and $\mathbf{Y}$
    Estimate covariance matrices as described in Algorithm 1, say $\hat{\mathbf{\Sigma}}'_{\mathbf{x}_i}$ $\forall i = \{1, \ldots, n\}$
    Train RF with $\mathbf{X}^c_r$ and $\mathbf{Y}$
    Estimate covariance matrices as described in Algorithm 1, say $\hat{\mathbf{\Sigma}}^{c'}_{\mathbf{x}_i}$ $\forall i = \{1, \ldots, n\}$
    Compute test statistic with $T_r = \frac{1}{n}\sum_{i=1}^{n} d\left(\hat{\mathbf{\Sigma}}'_{\mathbf{x}_i}, \hat{\mathbf{\Sigma}}^{c'}_{\mathbf{x}_i}\right)$
**end for**
Approximate the permutation $p$-value with $p = \frac{1}{R}\sum_{r=1}^{R} I(T_r > T)$
Reject the null hypothesis when $p < \alpha$. Otherwise, do not reject the null hypothesis.

In the significance test described above, we need to apply the proposed method many times: for the original data with *(i)* all covariates and *(ii)* the set of control covariates, and at each permutation for the permuted data with *(iii)* all covariates and *(iv)* the set of control covariates. The proposed method applies a `nodesize` tuning as described in the previous section. Since tuning the `nodesize` parameter can be computationally demanding, we tune the `nodesize` for the original data with all covariates and with the set of control covariates only and use those tuned values for their corresponding permutation steps.

The proposed significance test has two particular cases of interest. The first is to evaluate the global effect of the covariates on the conditional covariance estimates. If $\mathbf{X}$ has a global effect on the covariance matrix estimates obtained with the proposed method, then the conditional estimates $\mathbf{\Sigma}_{\mathbf{X}}$ should be significantly different from the unconditional covariance matrix estimate $\mathbf{\Sigma}_{root}$ which is computed as the sample covariance matrix of $\mathbf{Y}$. The null hypothesis (3) becomes

$$H_0 : \mathbf{\Sigma}_{\mathbf{X}} = \mathbf{\Sigma}_{root}. \tag{6}$$

See the Supplementary Algorithm 1, Additional file 3 for the details of the global significance test. The second case is to evaluate the effect of a single covariate when the other covariates are in the model. In that particular case, the null hypothesis (3) remains. The only difference between the global and partial significance tests is the number of forests we need to train. In the partial significance test, we need to train two random forests per sample, one for all covariates and one for the controlling variables, which makes a total $2R + 2$ random forests. However, when we test for the global effect, we need to train only one random forest per sample (in total $R + 1$ random forests) since we do not need to build a random forest for the root node.

### Variable importance

For traditional regression tree problems, we can get the variable importance (VIMP) measures by computing the average change in prediction accuracy using the OOB samples. However, the covariance regression problem does not have an observed target. We can compute the VIMP measures by using the *fit-the-fit* approach which has been applied to enhance interpretability of the covariates on the response [21, 27–31]. In the univariate response case, we get the importance measures by fitting a regression forest to re-predict the predicted values. However, in covariance regression, we have a predicted covariance matrix for each observation and not a single value. Therefore, we use a multivariate splitting rule based on the Mahalanobis distance [32] to re-predict the predicted covariance matrices. We begin by applying the proposed method using the original covariates and responses and estimate the covariance matrices as described in Algorithm 1. Next, we train a random forest with the original covariates and the vector of upper-triangular estimated covariance matrix elements as a multivariate response. VIMP measures are obtained from this random forest. Covariates with higher VIMP measures indicate higher importance for the estimation of covariance matrices. The proposed VIMP computation is described in Supplementary Algorithm 2 in the Additional file 4.

### Software

We have developed an R package called `CovRegRF`. We used the custom splitting feature of the `randomForestSRC` package [33] to implement our specially designed splitting criterion in the tree building process. The package is available on CRAN, https://CRAN.R-project.org/package=CovRegRF.

## Simulations

In this section, we perform a simulation study to demonstrate the performance of the proposed method, validate the proposed significance test with two particular cases-global and partial significance tests-and evaluate the variable importance estimations of the covariates.

### Data generating process

We carry out a simulation study using four Data Generating Processes (DGPs). The details of the DGPs are given in the Additional file 5. The first two DGPs are variations of the first simulated data set used in [8]. Both DGPs include one covariate and two response variables. The covariate $x$ is generated uniformly on $[-1, 1]$. In DGP1, the covariance matrix for the observation $x_i$ is $\mathbf{\Sigma}_{\mathbf{x}_i} = \mathbf{\Psi} + \mathbf{B}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{B}^\top$ where $\mathbf{x}_i^\top = (1, x_i)^\top$. DGP2 is similar to DGP1, except that we add a quadratic term to the covariance matrix equation such as $\mathbf{\Sigma}_{\mathbf{x_i}} = \mathbf{\Psi} + \mathbf{B}\dot{\mathbf{x}}_i\dot{\mathbf{x}}_i^\top\mathbf{B}^\top$ where $\dot{\mathbf{x}}_i^\top = (1, (x_i + x_i^2))^\top$.

In DGP3, the vector of covariates includes seven independent variables generated from the standard normal distribution. For the covariance structure, we use an AR(1) structure with heterogeneous variances. The correlations are generated with all seven covariates according to a tree model with a depth of three and eight terminal nodes. The variances are functions of the generated correlations. In DGP4, the covariance

matrix has a compound symmetry structure with heterogeneous variances. Both variances and correlations are functions of covariates. The covariates are generated from the standard normal distribution. The correlations are generated with a logit model and the variances are functions of these generated correlations. The number of covariates and response variables varies depending on the simulation settings. For all DGPs, after generating $\mathbf{\Sigma}_{\mathbf{x}_i}$, $\mathbf{y}_i$ is generated from a multivariate normal distribution $N(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{x}_i})$.

## Simulation design

### *Accuracy evaluation*

We perform a simulation study based on the four DGPs described above to evaluate the accuracy of the proposed method for estimating the covariance matrices. For DGP3 and DGP4, we consider five response variables. For each DGP, we use several values of the training sample size $n_{train} = \{50, 100, 200, 500, 1000\}$, which generates a total of 20 settings (4 DGPs $\times$ 5 training sample sizes). We repeat each setting 100 times. In each run of the simulations, we generate an independent test set of new observations with $n_{test} = 1000$.

We evaluate the performance of the covariance matrix estimates using the mean absolute errors (MAE) computed for both the estimated correlations and standard deviations separately. For the estimated correlations, we compute the MAE between the upper triangular (off-diagonal) matrices of the true and estimated correlations over all observations as follows:

$$MAE^{cor}(\hat{\mathbf{C}}_{\mathbf{X}}, \mathbf{C}_{\mathbf{X}}) = \frac{2}{q(q-1)n_{test}} \sum_{i=1}^{n_{test}} \sum_{j=1}^{q} \sum_{k=j+1}^{q} |\hat{\rho}_{ijk} - \rho_{ijk}|,$$

where $\mathbf{C}_{\mathbf{X}}$ and $\hat{\mathbf{C}}_{\mathbf{X}}$ are the collection of all correlation matrices corresponding to $\Sigma_{\mathbf{X}}$ and $\hat{\Sigma}_{\mathbf{X}}$, respectively. The values $\rho_{ijk}$ and $\hat{\rho}_{ijk}$ represent the correlations in row $j$ and column $k$ of $\mathbf{C}_{\mathbf{x}_i}$ and $\hat{\mathbf{C}}_{\mathbf{x}_i}$, respectively.

For the estimated standard deviations, we compute the normalized MAE between the true and estimated standard deviations over all observations as follows:

$$MAE^{sd}(\hat{\mathbf{\Sigma}}_{\mathbf{X}}, \mathbf{\Sigma}_{\mathbf{X}}) = \frac{1}{qn_{test}} \sum_{i=1}^{n_{test}} \sum_{j=1}^{q} \left| \frac{\hat{\sigma}_{ij} - \sigma_{ij}}{\sigma_{ij}} \right|.$$

The values $\sigma_{ij}^2$ and $\hat{\sigma}_{ij}^2$ represent the $j$th diagonal element of $\mathbf{\Sigma}_{\mathbf{x}_i}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{x}_i}$, respectively.

Smaller values of $MAE^{cor}$ and $MAE^{sd}$ indicate better performance. We compare our proposed method with the original Gaussian-based covariance regression model `covreg` developed in [8] which was presented in the Introduction. This method is currently available in the `covreg` R package [34]. Moreover, as a simple benchmark method, we compute the sample covariance matrix without covariates, which is then used as the covariance matrix estimate for all new observations from the test set.

### *Variable importance*

For the variable importance evaluation simulations, we use DGP3 and DGP4 in which we add five noise variables $X$ to the covariates set. As above, we consider several

values for the training sample sizes $n_{train} = \{50, 100, 200, 500, 1000\}$, for a total of 10 scenarios studied. We examine whether the estimated VIMP measures tend to rank the important variables first. The variable with the highest VIMP measure has a rank of 1. For each scenario, we compute the average rank for the important variables group and for the noise variables group.

#### *Evaluating the power of the global significance test*

We studied four scenarios to evaluate the global effect of the covariates, two of which are under the null hypothesis (6) and the other two under the alternative hypothesis. We generate the data sets for these scenarios as follows:

1. $H_0$ (case 1): we generate 5 $Y$ with a constant population covariance matrix and 10 $X$ variables which are all independent following a standard normal distribution. In this case, the covariance of $Y$ is independent of $X$ and we are therefore under the null hypothesis.
2. $H_0$ (case 2): we first generate 7 $X$ and 5 $Y$ under DGP3. Then, we replace the $\mathbf{X}$ matrix with 10 independent $X$ variables generated from a standard normal distribution. In this case, the covariance of $\mathbf{Y}$ varies with some of the $X$ variables but those $X$ variables are not available in the training set. Therefore, we are again under the null hypothesis.
3. $H_1$ (without noise): we generate 7 $X$ and 5 $Y$ under DGP3, and the covariates are available in the training set. In this case, the covariance of $\mathbf{Y}$ varies with all $X$ variables.
4. $H_1$ (with noise): we generate 7 $X$ and 5 $Y$ under DGP3 and we add 3 independent $X$ variables to the covariates' training set. In this case, the covariance of $\mathbf{Y}$ varies with some of the $X$ variables but not all.

#### *Evaluating the power of the partial significance test*

We can consider three scenarios to evaluate the effect of a single covariate, where one is under the null hypothesis (3) and the other two under the alternative hypothesis. We generate the data sets for these scenarios as follows:

1. $H_0$: We first generate 2 $X$ and 5 $Y$ with DGP4 and we add 1 independent $X$ variable to the covariates' training set. In this case, the covariance of $\mathbf{Y}$ varies only with the first two $X$ variables. The control set of variables is $\{X_1, X_2\}$ and we evaluate the effect of the $X_3$ variable. Therefore, we are under the null hypothesis.
2. $H_1(weakest)$: We generate 3 $X$ and 5 $Y$ with DGP4. In this case, the covariance of $\mathbf{Y}$ varies with all $X$ variables. The control set of variables is $\{X_1, X_2\}$ and we evaluate the effect of $X_3$, which has the weakest effect on the covariance matrix.
3. $H_1(strongest)$: We generate 3 $X$ and 5 $Y$ with DGP4. In this case, the covariance of $\mathbf{Y}$ again varies with all $X$ variables. But now the control set of variables is $\{X_2, X_3\}$ and we evaluate the effect of $X_1$, which has the strongest effect on the covariance matrix.

For both the global and partial significance test simulations, we use training sample sizes of $n_{train} = \{50, 100, 200, 300, 500\}$. The number of permutations and the number of replications for each scenario are set to 500. We estimate the type-1 error as the proportion of rejection in the scenarios simulated under $H_0$ and the power as the proportion of rejection in the scenarios simulated under $H_1$. We estimate a $p$-value for each replication and we reject the null hypothesis if the $p$-value is less than the significance level $\alpha = 0.05$. Finally, we compute the proportion of rejection over 500 replications.

### Parameter settings

For the simulations, we use the following parameters for the proposed method. We set the number of trees to 1000. Letting $p$ be the number of covariates, then the number of covariates to randomly split at each node, `mtry`, is set to $\lceil p/3 \rceil$. The number of random splits for splitting a covariate at each node, `nsplit`, is set to $\max\{n_{train}/50, 10\}$. We tune the `nodesize` parameter with the set of `nodesize`$= \{[sampsize \times (2^{-1}, 2^{-2}, 2^{-3}, \ldots)] > q\}$ where $q$ is the number of responses and `sampsize`$= 0.632 n_{train}$. In each replication, `covreg` is run in four independent chains for 8000 iterations, with the first half taken as burn-in.

## Results

### *Accuracy evaluation*

Figures 1 and 2 present the accuracy results for 100 repetitions. For each method, we can see the change in $MAE^{cor}$ and $MAE^{sd}$ computed for 100 repetitions with an increasing training sample size. As demonstrated in Fig. 1, for DGP1 and DGP2 when $n_{train} = 50$, the proposed method and `covreg` both have a similar performance with respect to the correlation estimation, with a slight advantage for `covreg`. For DGP1, `covreg` performs better for both the correlation and standard deviation compared to the proposed method as the sample size increases. This is expected since DGP1 is generated exactly under the `covreg` model. However, the proposed method still remains competitive. For DGP2, in which a quadratic term is added, the proposed method performs better for the correlation than `covreg` with increasing sample size. `covreg` shows better standard deviation estimation performance for smaller sample sizes, but after $n_{train} = 500$ the proposed method performs slightly better. As demonstrated in Fig. 2, for DGP3, the proposed method shows a significantly smaller $MAE^{cor}$ and $MAE^{sd}$ than `covreg` for all sample sizes. Moreover, for the smaller sample sizes, the proposed method has considerably lower variance in MAE. For DGP4, both methods improve with increasing sample size, but the proposed method shows smaller or equal MAEs for both correlation and standard deviation estimations. For DGP3 and DGP4, these results are expected, since the proposed method can capture a nonlinear effect. Supplementary figures 5 and 6 in the Additional file 6 present the difference in MAE between the proposed method and `covreg`. Moreover, we evaluate the accuracy with Stein's loss which is the Kullback–Leibler divergence between the estimated and true covariance matrices. The conclusions remain the same. See Supplementary Figure 7, Additional file 6.

For the `nodesize` tuning, we compare the accuracy results for different levels of `nodesize` along with the proposed tuning method. Supplementary figures 3 and 4 in

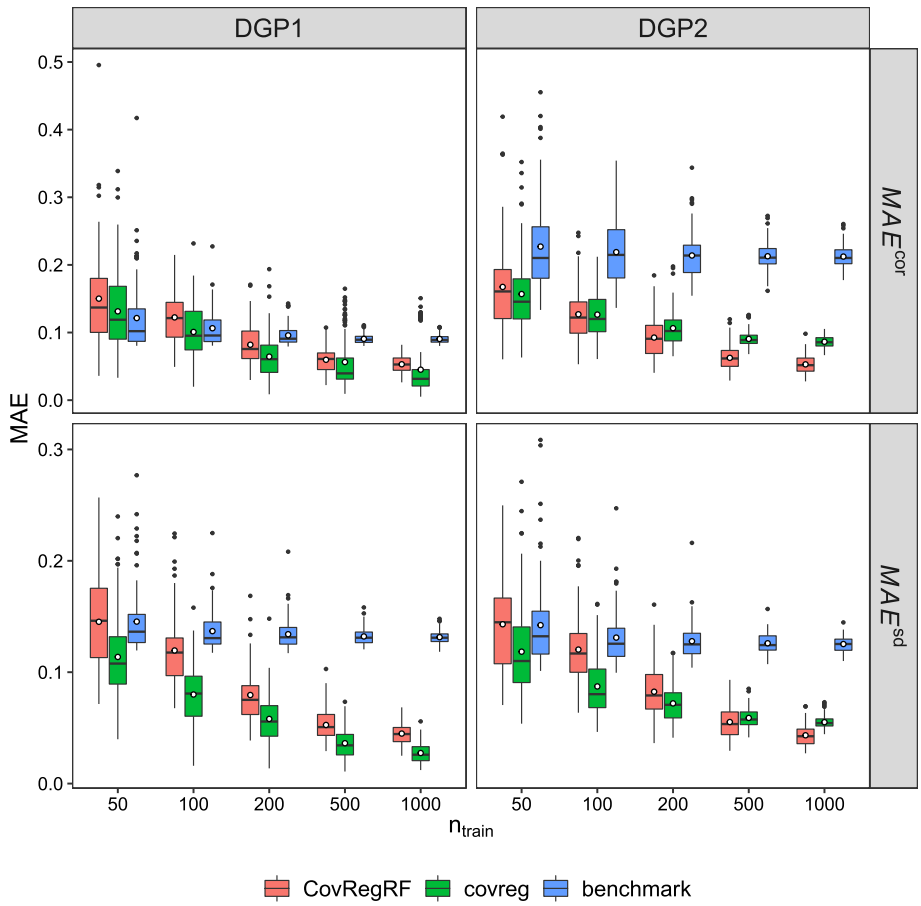


**Fig. 1** Accuracy evaluation results for DGP1 and DGP2. Smaller values of $MAE^{cor}$ and $MAE^{sd}$ are better

the Additional file 2 present the MAE results for all DGPs which show that the tuning method works well.

### *Variable importance*

Supplementary Figure 8 in the Additional file 4 presents the average ranks of the VIMP measures for both the important and noise sets of variables for DGP3 and DGP4. In all scenarios, the important variables have smaller average ranks than noise variables. As the sample size increases, the difference between the average ranks of important and noise variables increases, as expected.

### *Global significance test*

The left plot in Fig. 3 presents the estimated type-1 error and power for different training sample sizes for the two $H_0$ scenarios and two $H_1$ scenarios, respectively. We expect the type-1 error to be close to the significance level ($\alpha = 0.05$) and we can see that it is well controlled in both cases studied. In both $H_1$ scenarios, the power increases with the sample size. When the sample size is small, adding noise covariates slightly decreases the power, but this effect disappears as the sample size increases.

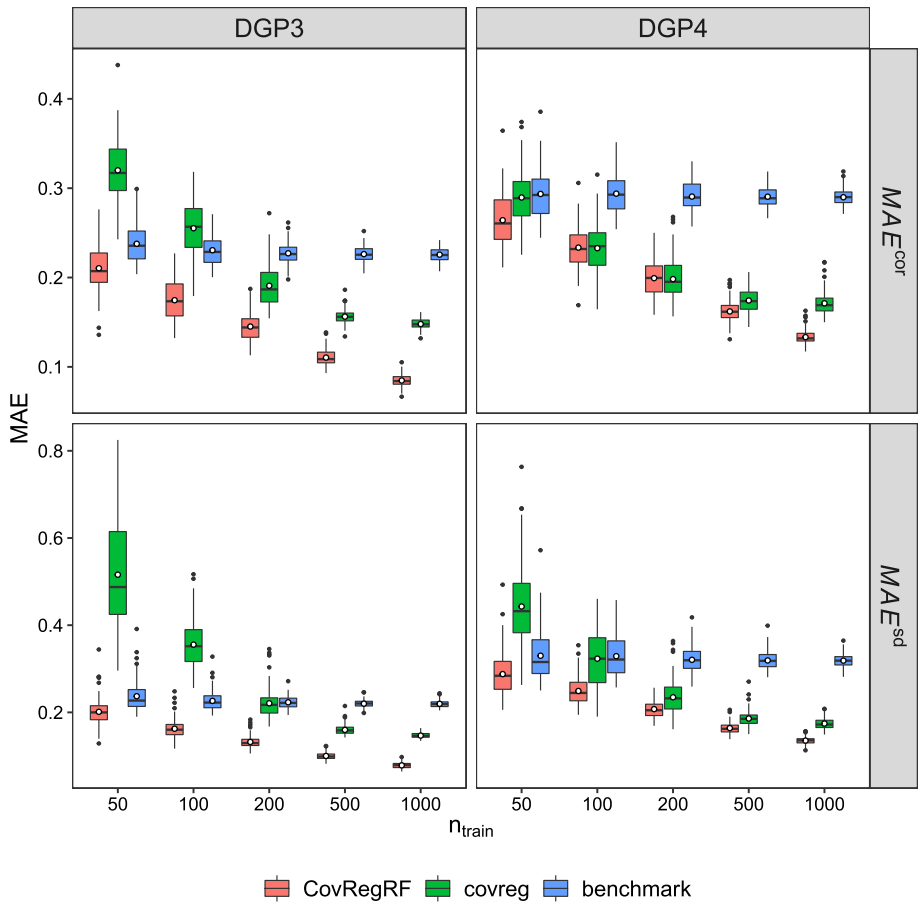


**Fig. 2** Accuracy evaluation results for DGP3 and DGP4. Smaller values of $MAE^{cor}$ and $MAE^{sd}$ are better

#### *Partial significance test*

The right plot in Fig. 3 presents the estimated type-1 error and power for different training sample sizes for the $H_0$ scenario and two $H_1$ scenarios, respectively. As can be seen from the $H_0$ line, the type-1 error is close to the significance level ($\alpha = 0.05$). In both $H_1$ scenarios, the power increases with the sample size as expected. However, the power is much smaller when one tests the weakest covariate compared to the strongest covariate.

## Real data example

Thyroid hormone, the collective name for two hormones, is widely known for regulating several body processes, including growth and metabolism [35, 36]. The main hormones produced by the thyroid gland are triiodothyronine (T3) and thyroxine (T4). The synthesis and secretion of these hormones are primarily regulated by thyroid stimulating hormone (TSH), which is produced by the pituitary gland. Primary hypothyroidism is a condition that occurs when the thyroid gland is underactive and the thyroid hormone produced is insufficient to meet the body's requirements, which leads to an increase of TSH. Contrarily, when the thyroid gland produces levels of thyroid hormones that are too high, leading to decreased levels of TSH, the resulting condition is hyperthyroidism.

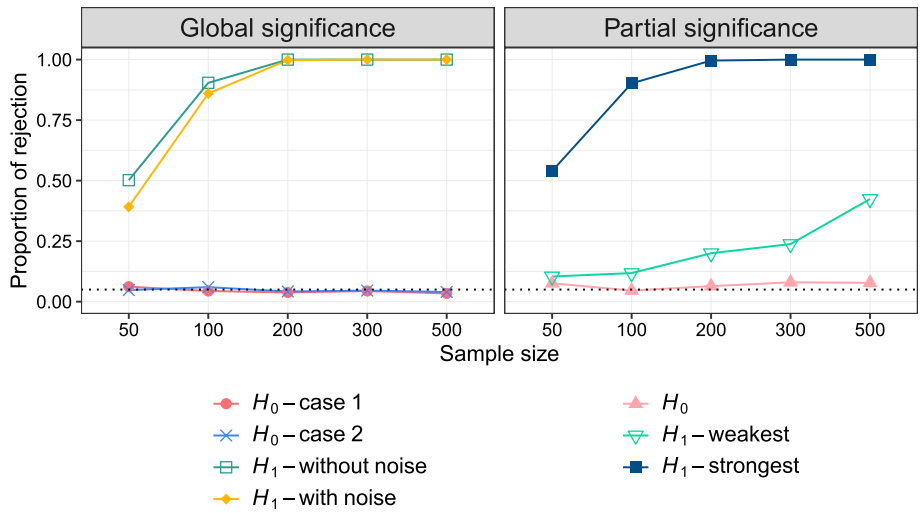


**Fig. 3** Significance test results. The left and right plots present the results for global and partial significance tests, respectively. The proportion of rejection corresponds to the type-1 error for $H_0$ scenarios, and power for $H_1$ scenarios. The dotted line represents the significance level of $\alpha = 0.05$

Serum levels of the thyroid hormones and TSH are used to evaluate subjects' thyroid function status and to identify subjects with a thyroid dysfunction. Therefore, establishing reference intervals for these hormones is critical in the diagnosis of thyroid dysfunction. However, reference ranges are affected by age and sex [37–41]. Furthermore, there is a relationship between TSH and thyroid hormone, and the effects of age and sex on this relationship have not been well described [27, 42]. Serum levels of these hormones are also affected by the subject's diagnosis, i.e. hormone levels would be within the reference ranges for normal subjects and out of range for subjects with thyroid dysfunction. The conditional mean of these hormones based on the covariates is studied in the literature, but to our knowledge, no study has yet explicitly investigated the effect of covariates on the conditional covariance matrix of these hormones. Hence, our contribution is to study the effect of age, sex and diagnosis on the covariance matrix of the thyroid hormones and TSH.

In this study, we investigate the thyroid disease data set from the UCI machine learning repository [43]. This data set originally included 9172 subjects and 30 variables including age, sex, hormone levels and diagnosis. Following the exclusion criteria applied in [42] and [40], we exclude pregnant women, subjects who have euthyroid sick syndrome (ESS), goitre, hypopituitarism or tumour, subjects who use antithyroid medication, thyroxine or lithium, who receive I131 treatment, or who have had thyroid surgery. The subjects have different diagnoses including hypothyroidism and hyperthyroidism, as well as normal subjects. Since the sample size of hyperthyroidism subjects is small, we exclude them from the analysis. We also exclude the very young and very old subjects, since there are only a few subjects on the extremes. The remaining data set consists of 324 hypothyroidism and 2951 normal subjects ($n = 3275$) between 20 and 80 years of age (2021 females/1254 males). We want to estimate the covariance matrix of four thyroid-related hormones-TSH, T3, TT4 (total T4) and FTI (free thyroxine index/free T4)-based on covariates and investigate how

the relationship between these hormones varies with the covariates. We apply the proposed method with the covariates age, sex and diagnosis to estimate the covariance matrix of the four hormones. We first perform the significance test with 500 permutations to evaluate the global effect of the three covariates. The estimated $p$-value with (5) is 0 and we reject the null hypothesis (6), which indicates that the conditional covariance matrices vary significantly with the set of covariates. Next, we apply the proposed method and obtain the covariance matrix estimates. We analyze the correlations between hormones as a function of covariates, and as shown in Fig. 4, age seems not to have much effect on the estimated correlations. We also compute the variable importance measures, and age (0.001) is found to be the least important variable where diagnosis (1.000) is the most important variable, followed by sex (0.011). Therefore, we apply the significance test to evaluate the effect of age on covariance matrices while controlling for sex and diagnosis. Using 500 permutations, the estimated $p$-value with (5) is 0.42 and we fail to reject the null hypothesis (3), indicating that we have insufficient evidence to prove that age has an effect on the estimated covariance matrices while sex and diagnosis are in the model. Although the mean levels of TSH and thyroid hormones differ with age [37–39, 41], the correlation between these hormones may not be affected by aging. Similarly, we apply the significance test for diagnosis and sex while controlling for the remaining two covariates, and the estimated $p$-values for both tests are 0, which indicates that both diagnosis and sex, taken individually, have an effect on the covariance matrix of the four hormones. We compare the estimated correlations using the proposed method to the sample correlations computed using the whole sample, which are represented with the black dashed lines in Fig. 4. For example, the sample correlation between TSH and T3 over all samples is −0.28 which is not close to the estimated correlation of either hypothyroidism or normal subjects. Furthermore, the estimated variances of the four hormones as a function of age, sex and diagnosis are presented in Supplementary Figure 9 of the Additional file 7. We can see that the variances also differ with covariates. For a mean regression analysis for any of these hormones, assuming a constant variance could yield misleading results.

The findings of this analysis suggest that there may be sex and diagnosis specific differences in the regulation of thyroid function, which could have important implications for the diagnosis and treatment of thyroid disorders in men and women. Clinicians can use this information to better understand the relationship between TSH and thyroid hormones in their patients, and to tailor their diagnostic and treatment approaches accordingly. It is known that the mean levels of TSH and thyroid hormones are different for hypothyroidism subjects compared to normal subjects. However, in Fig. 4, we also observe that there is a difference in correlation between hypothyroidism and normal subject classes. Moreover, we see that there is a difference between genders for hypothyroidism subjects for TSH and thyroid hormone correlations, Cor(TSH, T3), Cor(TSH, TT4), Cor(TSH, FTI).

## Concluding remarks

In this study, we propose a nonparametric covariance regression method, using a random forest framework, for estimating the covariance matrix of a multivariate response given a set of covariates. Random forest trees are built with a new splitting rule designed

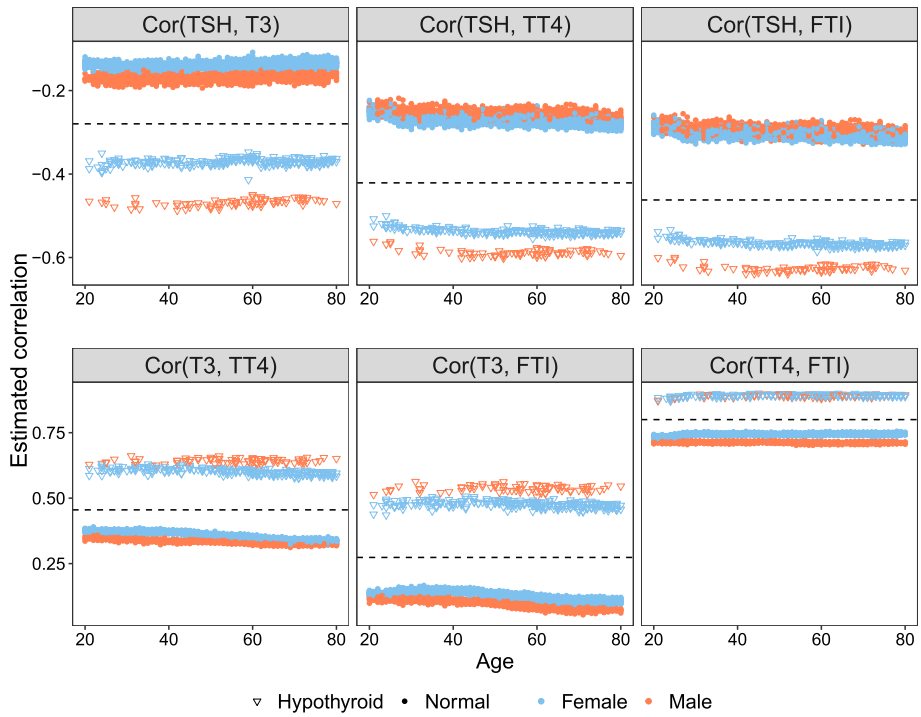


**Fig. 4** Estimated correlations between the four hormones as a function of age, sex and diagnosis. Dashed lines represent the sample correlations computed using the whole sample

to maximize the distance between the sample covariance matrix estimates of the child nodes. For a new observation, the random forest provides the set of nearest neighbour out-of-bag (OOB) observations which is used to estimate the conditional covariance matrix for that observation. We perform a simulation study to test the performance of the proposed method and compare it to the original Gaussian-based covariance regression model `covreg`. The average computational times of both methods for the simulations are presented in Supplementary Table 1 of the Additional file 8. We can see from the table that the proposed method is significantly faster than `covreg`. For the real data analysis, the computational time was 200.14 s. It should also be noted that `covreg` accounts for the uncertainty quantification in estimation of parameters which inevitably results in higher computational times compared to non-Bayesian methods. Furthermore, we propose a significance test to evaluate the effect of a subset of covariates while the other covariates are in the model. We investigate two particular cases: the global effect of covariates and the effect of a single covariate. We also propose a way to compute variable importance measures.

In this paper, we use the Euclidean distance between the upper triangular part of the two covariance matrices as splitting criterion. This is to avoid double counting the off-diagonal elements since covariance matrices are symmetric. However, several alternative splitting criteria are possible using other measures for computing distance between covariance matrices. We can use alternative distance metrics such as Frobenius norm, log-Euclidean, Kullback-Leibler divergence, Fisher Information metric, Bhattacharyya distance [44–46]. Another possibility is to use test statistics as splitting criteria. There is a large literature on testing the equality of covariance matrices. Here

are a few examples [47–51] and an R package [52] that implements them. Finally, we could use a weighted Euclidean distance between covariance matrices as $d(\mathbf{D},\mathbf{E}) = \sqrt{\sum_{i=1}^{q}\sum_{j=i}^{q} w_{ij}(\mathbf{D}_{ij} - \mathbf{E}_{ij})^2}$. This allows to finely control the weight we wish to give to each element of the matrix. This way, the splitting criterion could be based only on the the variance terms or on the covariance terms, for example. Another possibility is that for the final covariance matrix estimation for a new observation, we can use sparse or robust covariance matrix estimations [53, 54] using the nearest neighbour observations. Similarly, it is theoretically possible to use the sparse or robust covariance matrix estimations instead of the sample covariance matrix for the tree building process. However, the computational time could be a limiting factor. The proposed method can be applied to larger $\mathbf{X}$ dimensions. The computational time increases linearly with `mtry` which is the number of covariates to randomly split at each node. It can also be adapted to larger $\mathbf{Y}$ dimensions, but the computational time could be a limitation for very large $\mathbf{Y}$ dimensions. Computing the sample covariance matrix has a time complexity $\mathcal{O}(nq^2)$ for $q$ response variables and we compute covariance matrix for each node split in each tree of the forest which necessitates many covariance matrix computations.

In [21], we proposed a method, Random Forest with Canonical Correlation Analysis (RFCCA), which estimates the conditional canonical correlation between two multivariate data sets given the subject-related covariates. This method conditionally estimates a single parameter, the canonical correlation, that summarizes the strength of the dependency between two sets of variables. In this paper, we conditionally estimate the whole covariance matrix for one set of variables. Both methods use a splitting criterion that aims at maximizing the heterogeneity of the target parameter to build a forest of trees to obtain a set of local observations that is used to compute the final estimate. Hence, the general methodology in both papers is similar but the goals are different.

## Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05377-y.

**Additional file 1.** Results of a simulation study comparing different ways of estimating the final covariance matrix

**Additional file 2.** Results of a nodesize tuning experiment

**Additional file 3**. Details of the global significance test

**Additional file 4.** VIMP computation and figure presenting the performance of the estimated VIMP measures

**Additional file 5.** Details of the DGPs

**Additional file 6.** Figures presenting the difference in MAE and accuracy evaluation with Stein's loss

**Additional file 7.** Figure presenting the estimated variances of four thyroid-related hormones

**Additional file 8.** Table presenting the computational times

**Acknowledgements**
Not applicable.

**Author contributions**
CA, DL and AL developed the method, designed the simulation study and applied real data analysis. CA performed simulations, prepared figures, drafted manuscript and implemented the R package. DL and AL revised the manuscript. All authors read and approved the final manuscript.

**Funding**
This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and by Fondation HEC Montréal.

**Availability of data and materials**
The thyroid disease data set is available at https://archive.ics.uci.edu/ml/datasets/thyroid+disease. `CovRegRF` is implemented in a freely available R package on CRAN, available at https://cran.r-project.org/package=CovRegRF.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interest**
The authors declare that they have no competing interests.

Received: 8 December 2022 Accepted: 2 June 2023
Published online: 17 June 2023

**References**

1. Seiler C, Holmes S. Multivariate heteroscedasticity models for functional brain connectivity. Front Neurosci. 2017;11.
2. Le Goallec A, Patel CJ. Age-dependent co-dependency structure of biomarkers in the general population of the United States. Aging. 2019;11(5):1404–26.
3. Levy R, Borenstein E. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. Proc Natl Acad Sci. 2013;110(31):12804–9.
4. McGregor K, Labbe A, Greenwood CMT. MDiNE: a model to estimate differential co-occurrence networks in microbiome studies. Bioinformatics. 2020;36(6):1840–7.
5. Tu D, Mahony B, Moore TM, Bertolero MA, Alexander-Bloch AF, Gur R, et al. CoCoA: conditional correlation models with association size. Biostatistics. 2022.
6. Jiang L, Qiao K, Li C. Distance-based functional criticality in the human brain: Intelligence and emotional intelligence. BMC Bioinformatics. 2021;22(1):1–17.
7. Yin J, Geng Z, Li R, Wang H. Nonparametric covariance model. Stat Sin. 2010;20:469.
8. Hoff PD, Niu X. A covariance regression model. Stat Sin. 2012;22(2):729–53.
9. Niu X, Hoff PD. Joint mean and covariance modeling of multiple health outcome measures. Ann Appl Stat. 2019;13(1):321–39.
10. Fox EB, Dunson DB. Bayesian nonparametric covariance regression. J Mach Learn Res. 2015;16(1):2501–42.
11. Franks AM. Reducing subspace models for large-scale covariance regression. Biometrics. 2021.
12. Zou T, Lan W, Wang H, Tsai CL. Covariance regression analysis. J Am Stat Assoc. 2017;112(517):266–81.
13. Zhao Y, Wang B, Mostofsky SH, Caffo BS, Luo X. Covariate Assisted Principal regression for covariance matrix outcomes. Biostatistics. 2021;22(3):629–45.
14. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
15. Hothorn T, Lausen B, Benner A, Radespiel-Tröger M. Bagging survival trees. Stat Med. 2004;23(1):77–91.
16. Lin Y, Jeon Y. Random forests and adaptive nearest neighbors. J Am Stat Assoc. 2006;101(474):578–90.
17. Moradian H, Larocque D, Bellavance F. L1 splitting rules in survival forests. Lifetime Data Anal. 2017;23(4):671.
18. Moradian H, Larocque D, Bellavance F. Survival forests for data with dependent censoring. Stat Methods Med Res. 2019;28(2):445–61.
19. Roy MH, Larocque D. Prediction intervals with random forests. Stat Methods Med Res. 2020;29(1):205–29.
20. Tabib S, Larocque D. Non-parametric individual treatment effect estimation for survival data with random forests. Bioinformatics. 2020;36(2):629–36.
21. Alakuş C, Larocque D, Jacquemont S, Barlaam F, Martin CO, Agbogba K, et al. Conditional canonical correlation estimation based on covariates with random forests. Bioinformatics. 2021;37(17):2714–21.
22. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. Boca Raton: CRC Press; 1984.
23. Athey S, Tibshirani J, Wager S. Generalized random forests. Ann Stat. 2019;47(2):1148–78.
24. Lu B, Hardin J. A unified framework for random forest prediction error estimation. J Mach Learn Res. 2021;22(8):1–41.
25. Alakuş C, Larocque D, Labbe A. The R Journal: RFpredInterval: an R package for prediction intervals with random forests and boosted forests. R J. 2022;14(1):300–20.
26. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc. 2018;113(523):1228–42.
27. Lee K, Bargagli-Stoffi FJ, Dominici F. Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. arXiv preprint arXiv:2009.09036. 2020.
28. Spanbauer C, Sparapani R. Nonparametric machine learning for precision medicine with longitudinal clinical trials and Bayesian additive regression trees with mixed models. Stat Med. 2021;40(11):2665–91.
29. Bargagli-Stoffi FJ, De Beckker K, Maldonado JE, De Witte K. Assessing sensitivity of machine learning predictions. A novel toolbox with an application to financial literacy. arXiv preprint arXiv:2102.04382. 2021.
30. Bargagli-Stoffi FJ, Witte KD, Gnecco G. Heterogeneous causal effects with imperfect compliance: A Bayesian machine learning approach. Ann Appl Stat. 2022;16(3):1986–2009.
31. Meid AD, Gerharz A, Groll A. Machine learning for tumor growth inhibition: Interpretable predictive models for transparency and reproducibility. CPT Pharmacometrics Syst Pharmacol. 2022;11(3):257.
32. Ishwaran H, Tang F, Lu M, Kogalur UB. randomForestSRC: Multivariate splitting rule vignette; 2021.

33. Ishwaran H, Kogalur UB. Fast unified random forests for survival, regression, and classification (RF-SRC); 2022. R package version 3.1.0.
34. Niu X, Hoff P. covreg: A simultaneous regression model for the mean and covariance; 2014. R package version 1.0.
35. Yen PM. Physiological and molecular basis of thyroid hormone action. Physiol Rev. 2001;81(3):1097–142.
36. Shahid MA, Ashraf MA, Sharma S. Physiology, thyroid hormone. Treasure Island, FL: StatPearls Publishing;2022.
37. Kapelari K, Kirchlechner C, Högler W, Schweitzer K, Virgolini I, Moncayo R. Pediatric reference intervals for thyroid hormone levels from birth to adulthood: a retrospective study. BMC Endocr Disord. 2008;8(1):15.
38. Aggarwal N, Razvi S. Thyroid and aging or the aging thyroid? An evidence-based analysis of the literature. J Thyroid Res. 2013;2013.
39. Biondi B. The normal TSH reference range: what has changed in the last decade? J Clin Endocrinol Metab. 2013;98(9):3584–7.
40. Strich D, Karavani G, Edri S, Chay C, Gillis D. FT3 is higher in males than in females and decreases over the lifespan. Endocr Pract. 2017;23(7):803–7.
41. Park SY, Kim HI, Oh HK, Kim TH, Jang HW, Chung JH, et al. Age-and gender-specific reference intervals of TSH and free T4 in an iodine-replete area: data from Korean National Health and Nutrition Examination Survey IV (2013–2015). PLoS ONE. 2018;13(2): e0190738.
42. Hadlow NC, Rothacker KM, Wardrop R, Brown SJ, Lim EM, Walsh JP. The relationship between TSH and free T4 in a large population is complex and nonlinear and differs by age and sex. J Clin Endocrinol Metab. 2013;98(7):2936–43.
43. Dua D, Graff C. UCI machine learning repository; 2017.
44. Dryden IL, Koloydenko A, Zhou D. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. Ann Appl Stat. 2009;3(3):1102–23.
45. Costa SIR, Santos SA, Strapasson JE. Fisher information distance: a geometrical reading. Discrete Appl Math. 2015;197:59–69
46. Bhattacharyya A. On a measure of divergence between two multinomial populations. Sankhyā Indian J Stat (1933-1960). 1946;7(4):401–406.
47. Nagao H. On some test criteria for covariance matrix. Ann Stat. 1973;1(4):700–9.
48. R Schott J. Some tests for the equality of covariance matrices. J Stat Plann Inference. 2001;94(1):25–36.
49. Ledoit O, Wolf M. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. Ann Stat. 2002;30(4):1081–102.
50. Schott JR. A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. Comput Stat Data Anal. 2007;51(12):6535–42.
51. Srivastava MS, Yanagihara H, Kubokawa T. Tests for covariance matrices in high dimension with less sample size. J Multivar Anal. 2014;130:289–309.
52. Barnard B, Young D. Covariance matrix Tests; 2018. R package version 0.1.4.
53. Rousseeuw PJ, Driessen KV. A fast algorithm for the minimum covariance determinant estimator. Technometrics. 1999;41(3):212–23.
54. Bien J, Tibshirani RJ. Sparse estimation of a covariance matrix. Biometrika. 2011;98(4):807–20.