Computational Statistics (2024) 39:271–288
https://doi.org/10.1007/s00180-022-01268-7

ORIGINAL PAPER

A robust knockoff filter for sparse regression analysis of microbiome compositional data

Gianna Serafina Monti[1] · **Peter Filzmoser**[2]

Received: 29 April 2022 / Accepted: 31 July 2022 / Published online: 18 August 2022

Abstract
Microbiome data analysis often relies on the identification of a subset of potential biomarkers associated with a clinical outcome of interest. Robust ZeroSum regression, an elastic-net penalized compositional regression built on the least trimmed squares estimator, is a variable selection procedure capable to cope with the high dimensionality of these data, their compositional nature, and, at the same time, it guarantees robustness against the presence of outliers. The necessity of discovering "true" effects and to improve clinical research quality and reproducibility has motivated us to propose a two-step robust compositional knockoff filter procedure, which allows selecting the set of relevant biomarkers, among the many measured features having a nonzero effect on the response, controlling the expected fraction of false positives. We demonstrate the effectiveness of our proposal in an extensive simulation study, and illustrate its usefulness in an application to intestinal microbiome analysis.

Keywords False discovery rate (FDR) · High-dimensional regression · Knockoffs · Variable selection · Robustness

1 Introduction

Understanding the microbiome and how it is related to several aspects of the human health, including a wide range of diseases, is an intensive research area (The Human Microbiome Project Consortium 2012; Li 2015). The rapid advancement of human microbiome research resulted in the development of high-throughput sequencing technologies which enable to collect huge amounts of data. Due to great variation in the library size, the raw data have traditionally been normalized to allow for a

✉ Gianna Serafina Monti
gianna.monti@unimib.it

[1] Department of Economics, Management and Statistics, University of Milano Bicocca, Milan, Italy

[2] Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria

comparison among different samples. What is relevant for the analysis is the taxonomic relative abundance. This is precisely the concept of compositional data analysis, namely that (only) relative information counts for the analysis (Aitchison 1986). In fact, nowadays it is widely recognized that a proper analysis of microbiome data requires an appropriate compositional data analysis methodology (Gloor et al. 2017; Weiss et al. 2017; Nearing et al. 2022).

Different approaches have been proposed for a compositional analysis of these high-dimensional data. One interesting methodology is the linear log-contrast model for regression analysis (Lin et al. 2014; Shi et al. 2016), which accounts for the fact that the number of microbial taxa is usually bigger than the number of observations. This model is a natural extension of the log-contrast model, as it has been introduced in the seminal work of Aitchison and Bacon-Shone (1984), to address the issues derived from the compositional nature of the microbiome data, such as the collinearity and the non-Gaussian distribution of the compositional covariates.

A critical challenge to improve clinical research quality and reproducibility is to reliably select those microbial taxa, among a massive number of measured features, that are truly associated with a clinical outcome of interest. At the same time, however, there is the requirement to control the number of false positives, i.e., variables which have been selected by the method, but which actually do not have any significant effect on the outcome. Using false discovery rate (FDR)-controlling methods allows scientists to decide on the maximum threshold of the expected proportion of errors among the rejections they are willing to accept among all the discoveries, i.e., the significant results. A wider adoption of the FDR-controlling strategies has been recommended as a natural way for improving the power of association studies for complex phenomena of interest (Storey and Tibshirani 2003; Brzyski et al. 2017). Different methods for controlling the FDR have been presented in the literature: methods for marginal FDR control that examine each relative abundance taxa at a time followed by multiple comparison procedures (Benjamini and Hochberg 1995; Storey 2002), or the knockoff filter (Barber and Candés 2015; Candés et al. 2018; Barber and Candés 2019). The latter method gained popularity recently, and it is designed to control the expected fraction of false positives in a set of selected biomarkers.

The main idea behind the original fixed-X knockoff procedure (Barber and Candés 2015) was to construct a set of 'knockoff copy' variables which are not associated with the response, conditionally on the original variables, but with a correlation structure that mimics the one of the original variables. Knockoff variables act as controls for the original variables in the variable selection process. The knockoff filter procedure achieves an exact finite-sample FDR control in the homoscedastic Gaussian linear model when the number of observations is at least twice as big as the number of variables. Since this is not the case for microbiome data, the number of candidate variables is first reduced in a screening procedure.

However, these methods are not specifically designed for compositional data, thus they do not honor their nature, and can lead to inappropriate conclusions. In particular, the marginal method is often highly conservative, controlling the probability of any false positives, at the price of considerably reduced power in detecting true positives given the high dimension of the design matrix.

Candés et al. (2018) extended the idea of knockoff in the case of $p > n$ and treating the covariates as random, namely the model-X knockoff, which is based on the assumption that the distribution of the original features is completely specified in order to allow for the construction of the knockoff copies. The most used sampling scheme for knockoff generation is the sequential conditional independent pairs algorithm, whose implementations were only available for Gaussian distributions and discrete Markov chains (Candés et al. 2018; Sesia et al. 2019). However, for a compositional design matrix, the assumption of a Gaussian distribution is violated, and constructing exact or approximate knockoff features that do not follow a Gaussian distribution is nontrivial and still an open problem (Bates et al. 2021).

To address this problem, Srinivasan et al. (2021) proposed an FDR-controlled variable selection method, named compositional knockoff filter (CKF), specifically designed for the analysis of microbiome compositional data. However, the presence of anomalies in the data, such as observations that deviate from the majority, can undermine the CKF ability to control the FDR. This motivates us to propose a two-step robust compositional knockoff filter (RCKF) with the aim to control the finite-sample FDR, while maintaining robustness against outliers in the data.

This contribution is organized as follows: Sect. 2 details our proposal, the robust compositional knockoff filter (RCKF). Section 3 reports simulation studies to compare the RCKF with its non robust competitor CKF. A real data example on intestinal microbiome analysis is presented in Sect. 4. The final Sect. 5 concludes.

2 Robust compositional knockoff filter

In this section we introduce the robust compositional knockoff filter (RCKF), a robust version of the compositional knockoff filter (CKF) proposed by Srinivasan et al. (2021), to perform FDR-controlled variable selection for microbiome compositional data. The goal is to select a final subset of the originally measured biomarkers which are truly associated with the clinical outcome of interest by a procedure which is robust against vertical outliers and leverage points in the data. It is based on the recycled fixed-X knockoff procedure (Barber and Candés 2015, 2019), which requires that the number of observations used for the filtering procedure is at least twice as big as the number of variables. It consists of a two-step procedure: a robust compositional screening step, followed by a robust selection step. The main idea of the fixed-X knockoff is to construct for each feature X_j, screened in the first step, a synthetic fake copy $\tilde{X}_j$, which can play the role of a control covariate. Knockoff copies $\tilde{X}_1, \ldots, \tilde{X}_p$ mimic the correlation structure of the original features $X_1, \ldots, X_p$ and are conditionally independent of Y if they were null.

Before continuing describing the new methodology, we have to provide more background on regression modeling of compositions. Thus, in Sect. 2.1 we briefly review the log-contrast regression model in its original version (Aitchison and Bacon-Shone 1984) and its extension to high dimensions (Lin et al. 2014). In Sects. 2.2 and 2.3 we describe the two-step RCKF procedure, consisting of the robust compositional screening procedure and the subsequent robust controlled variable selection.

2.1 Log-contrast regression model

Let $\mathbf{Y} \in \mathbb{R}^n$ be the response vector and $\mathbf{X} \in \mathbb{R}^{n \times p}$ the compositional data matrix, where each row $\mathbf{x}_i$ of $\mathbf{X}$ lies in the simplex

$$\mathcal{S}^p = \left\{ \mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T,\ x_{ij} > 0,\ \sum_{j=1}^{p} x_{ij} = \kappa \right\},$$

where κ is a constant, usually taken to be one. Let $\mathbf{Z}^p \in \mathbb{R}^{n \times (p-1)}$ be the matrix of a log-ratio transformation of $\mathbf{X}$, where $z_{ij}^p = \log(x_{ij}/x_{ip})$, for $i = 1, \dots, n$ and $j = 1, \dots, p-1$, and p is the chosen reference component (Aitchison 1986).

To overcome the rank deficiency of the compositional design matrix $\mathbf{Z}^p$, Aitchison and Bacon-Shone (1984) formulated the log-contrast model defined as $\mathbf{Y} = \mathbf{Z}^p \boldsymbol{\beta}_{\setminus p} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta}_{\setminus p} = (\beta_1, \beta_2, \dots, \beta_{p-1})^T$ is the vector with the $(p-1)$ regression coefficients, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ contains the error terms. Lin et al. (2014) extended the log-contrast model to the high-dimensional setting, and they reformulated it into a symmetric form, which allows to avoid the choice of a reference component:

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{s.t.} \ \sum_{j=1}^{p} \beta_j = 0\,, \tag{1}$$

where $\mathbf{Z} \in \mathbb{R}^{n \times p}$ is the log-composition matrix with $z_{ij} = \log x_{ij}$, for $i = 1, \dots, n$ and $j = 1, \dots, p$, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ is the vector of coefficients. Due to the zero-sum linear constraint on the regression coefficients, model (1) is known as Zero-Sum regression; it preserves the simplex structure, and it treats all the components equally. In the high-dimensional setting, where the number of explanatory variables p is much larger than the number of observations n, Lin et al. (2014) suggested to estimate the regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$, which are supposed to be sparse, by a penalized estimation procedure with linear constraints,

$$\widehat{\boldsymbol{\beta}}_{\text{ZS}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{z}_i^T \boldsymbol{\beta})^2 + \lambda ||\boldsymbol{\beta}||_1 \right\}, \quad \text{s.t.} \ \sum_{j=1}^{p} \beta_j = 0, \tag{2}$$

where y_i and $\mathbf{z}_i$ are the i-th observations of the response $\mathbf{Y}$ and the explanatory matrix $\mathbf{Z}$, respectively, $\lambda > 0$ is a tuning parameter to control sparsity, and $||.||_1$ is the ℓ_1 or lasso penalty.

The zero-sum constraint is essential to ensure some desirable properties of the $\widehat{\boldsymbol{\beta}}_{\text{ZS}}$ estimator: the scale invariance, namely the regression coefficients are independent of an arbitrary scaling of the basis counts from which a composition is obtained; the permutation invariance, i.e. the estimator is invariant under any arbitrary permutation of the p components; and the selection invariance, that is

the estimator remains unaffected by correctly excluding some or all of the zero components (Lin et al. 2014).

With the aim to detect outlying observations, whose presence can seriously affect the prediction accuracy of the estimated log-contrast model, especially for high-dimensional data, Monti and Filzmoser (2021) introduced the Robust ZeroSum estimator (RZS), which is a compositional lasso version of the sparse least trimmed squares (SLTS) estimator (Alfons et al. 2013). RZS first tries to identify a homogeneous subset of the data, consisting of the majority of the observations, which best corresponds to the model. Then a weight is assigned to every observation which depends on the size of its residual to the fitted model. RZS is defined as

$$\hat{\boldsymbol{\beta}}_{\mathrm{RZS}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left(\sum_{i=1}^{n} w_i (y_i - \mathbf{z}_i^T \boldsymbol{\beta})^2 + n_w \lambda ||\boldsymbol{\beta}||_1 \right), \text{ s.t. } \sum_{j=1}^{p} \beta_j = 0, \tag{3}$$

where $n_w = \sum_{i=1}^{n} w_i$ is the sum of the binary weights w_i computed to reduce the influence of outliers detected by the final optimal solution of the minimization problem. If $w_i = 1$, the i-th observation is considered a regular one, and if $w_i = 0$, the i-th observation is identified as an outlier.

In its original formulation, estimator (3) added an elastic-net penalty for the coefficients to the objective function, but, for the purpose of this contribution, we limit the attention only to the lasso penalty.

2.2 Robust compositional screening procedure

To avoid selection bias in the screening step we split the sample into two halves: the first half (n_0 samples) is used to screen variables, and the second half ($n_1 = n - n_0$) is used for variable selection. The benefit of data splitting in implementing the two-step scheme of variable screening and a subsequent variable selection has been discussed and demonstrated by several authors (Fan and Lv 2008; Zhang and Xia 2008; Zhu and Yang 2015).

Following this idea, we randomly split the original data $(\mathbf{Z}, \mathbf{Y})$ into $(\mathbf{Z}^{(0)}, \mathbf{Y}^{(0)})$ and $(\mathbf{Z}^{(1)}, \mathbf{Y}^{(1)})$ where $\mathbf{Z}^{(0)} \in \mathbb{R}^{n_0 \times p}$ and $\mathbf{Y}^{(0)} \in \mathbb{R}^{n_0}$, $\mathbf{Z}^{(1)} \in \mathbb{R}^{n_1 \times p}$ and $\mathbf{Y}^{(1)} \in \mathbb{R}^{n_1}$.

The subset $(\mathbf{Z}^{(0)}, \mathbf{Y}^{(0)})$ is used to perform the screening step in order to obtain a subset of features $\hat{S}_0 \in \{1, \ldots, p\}$ such that $|\hat{S}_0| \leq \frac{n_1}{2}$, where $|\hat{S}_0|$ denotes the cardinality of set $\hat{S}_0$, whereas the subset $(\mathbf{Z}^{(1)}, \mathbf{Y}^{(1)})$ is used to perform the selection step. It is desirable that in the screening step all relevant features are selected, and various procedures have been proposed with this goal. However, common methods such as the Pearson correlation (Fan and Lv 2008) or the distance correlation (Szekely et al. 2007) do not take into account the compositional nature of the features. To this aim, Srinivasan et al. (2021) proposed a compositional screening procedure (CSP) adapting best-subset selection to the log-contrast model, that is

$$\hat{\boldsymbol{\beta}}_{\mathrm{BSS}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{z}_i^T \boldsymbol{\beta})^2 \right\}, \quad \text{s.t. } ||\boldsymbol{\beta}||_0 \leq k \text{ and } \sum_{j=1}^{p} \beta_j = 0\,, \tag{4}$$

where k is the size of the cardinality of the final set of the selected features. The objective function (4) could be interpreted as an ℓ_0-constrained sparse least-squares estimation problem. Common choices for the screening set size are $k = c\lfloor \frac{n_0}{\log(n_0)} \rfloor$, for some $c > 0$ (Fan and Lv 2008; Li et al. 2012). The minimization problem (4), which is NP-hard, could be solved by using mixed integer optimization (Konno and Yamamoto 2009; Bertsimas et al. 2016).

As the presence of anomalies in the data could seriously affect the mentioned likelihood-based compositional screening procedure, we present a novel robust compositional screening procedure (RCSP) that simultaneously attains variable screening and robustness against outliers.

RCSP is based on an adaptation of the RZS algorithm to obtain a subset of features $\hat{S}_0 \in \{1, \dots, p\}$ such that $|\hat{S}_0| \leq \frac{n_1}{2}$, to allow for an application of the fixed-X knockoff scheme, which requires that the number of observations is at least twice as big as the number of variables. The features in $\hat{S}_0$ are obtained as the set of active predictors which correspond to a sparsity parameter λ_k, closest to the minimum λ in (3), associated to the minimum cross-validated mean squared error (MSE), such that the number of selected variables, $\hat{\beta}_{\text{RZS},j\in\hat{S}_0} \neq 0$ as solution of problem (3), is in the neighborhood of a fixed screening set size k. The choice of k could be considered a further tuning parameter in the model.

The reduced log-contrast model after RCSP is $y_i = \sum_{j\in\hat{S}_0} z_{ij}\beta_j^r + \epsilon_i$, s.t. $\sum_{j\in\hat{S}_0} \beta_j^r = 0$. A further normalization step of the screened features is necessary to ensure model identifiability, thus $z_{ij} = \log x_{ij}^\star = x_{ij} / \sum_{j\in\hat{S}_0} x_{ij}$, where for simplicity we use the same notation for (the elements of) the compositional design matrix.

2.3 Robust controlled variable selection

We would like to estimate how many of the RZS discoveries in the first step are, in fact, null, i.e. $\hat{\beta}_{\text{RZS},j\in\hat{S}_0} = 0$. To this aim we consider the recycled knockoff procedure as follows (see Barber and Candés 2019, for more details).

Let $\mathbf{Z}^{(1)}_{\hat{S}_0} \in \mathbb{R}^{n_1\times|\hat{S}_0|}$, denote the columns of $\mathbf{Z}^{(1)}$ corresponding to $\hat{S}_0$, the selected set from the computed solution of the robust compositional screening procedure. The knockoff matrix $\tilde{\mathbf{Z}}^{(1)}_{\hat{S}_0}$, which has a "negative control" rule in the variable selection procedure, is built from $\mathbf{Z}^{(1)}_{\hat{S}_0}$ using the fixed-X knockoff procedure (Barber and Candés 2015). A review of the knockoff construction under the fixed-X design is briefly reported in Appendix A. Basically, the variables in the knockoff matrix hold the same correlation structure as the original variables, except that they are constructed to be conditionally independent from the response $\mathbf{Y}$. Note that to apply fixed-X knockoff it is necessary that the number of observations used is at least twice as big as the number of variables, which has to be guaranteed by the first screening step, but no further assumptions on the distribution of $\mathbf{Z}^{(1)}$ are required.

To increase the selection power, we consider the data recycling mechanism (Barber and Candés 2019) to construct the knockoff matrix, that is we concatenate the

original compositional design matrix $\mathbf{Z}^{(0)}_{\hat{S}_0}$ of the first n_0 observations with the knockoff matrix $\tilde{\mathbf{Z}}^{(1)}_{\hat{S}_0}$ related to the remaining n_1 observations,

$$\tilde{\mathbf{Z}}_{\hat{S}_0} = \begin{bmatrix} \mathbf{Z}^{(0)}_{\hat{S}_0} \\ \tilde{\mathbf{Z}}^{(1)}_{\hat{S}_0} \end{bmatrix} \in \mathbb{R}^{n \times |\hat{S}_0|}.$$

Then the knockoff filter described below is applied to the whole dataset of n samples. This procedure will involve the compositional design matrix $\mathbf{Z}_{\hat{S}_0}$, the knockoff matrix $\tilde{\mathbf{Z}}_{\hat{S}_0}$, and the original response $\mathbf{Y}$. The term "data recycling" refers to the fact that $(\mathbf{Z}^{(0)}_{\hat{S}_0}, \mathbf{Y}^{(0)})$ has already been used in the screening step.

For the knockoff filter we need to work with the augmented design matrix $\mathbb{Z}_{\hat{S}_0} = [\mathbf{Z}_{\hat{S}_0}, \tilde{\mathbf{Z}}_{\hat{S}_0}] \in \mathbb{R}^{n \times 2|\hat{S}_0|}$. We denote the rows of this matrix by $\mathbb{z}_i$. To account for possible outliers in the response variable as well as for outliers in the predictor space in the augmented data we propose to apply a robust penalized regression procedure, involving $\mathbf{Z}_{\hat{S}_0}$, $\tilde{\mathbf{Z}}_{\hat{S}_0}$ and $\mathbf{Y}$, by means of the sparse least trimmed squares (SLTS) estimator (Alfons et al. 2013), which has been demonstrated to exhibit good performance with respect to model selection and prediction in presence of contaminated data.

Consider $\bar{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}^T, \tilde{\boldsymbol{\beta}}^T)^T$ as the solution of the robust lasso optimization problem,

$$\bar{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{2|\hat{S}_0|}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{h} (r^2(\boldsymbol{\beta}))_{(i:n)} + h\lambda ||\boldsymbol{\beta}||_1 \right\}, \tag{5}$$

where $r_i = y_i - \mathbb{z}_i^T \boldsymbol{\beta}$ are the regression residuals, $(r^2(\boldsymbol{\beta}))_{(1:n)} \leq \cdots \leq (r^2(\boldsymbol{\beta}))_{(n:n)}$ are the order statistics of the squared residuals, and $h \leq n$ is a truncation number. The solution $\bar{\boldsymbol{\beta}}$ appends the first $|\hat{S}_0|$ components, the coefficients for the original variables, to the last $|\hat{S}_0|$ coefficients for the knockoffs features. Note that in the augmented robust lasso problem (5) the zero-sum constraint on $\boldsymbol{\beta}$ is no longer needed as the associated microbiome matrix $\mathbb{X}_{\hat{S}_0} = \exp(\mathbb{Z}_{\hat{S}_0})$ is no more compositional due to the augmented design matrix $\mathbb{Z}_{\hat{S}_0}$.

The idea is then to compare in the lasso path in which sequence the j-th variable of $\mathbf{Z}_{\hat{S}_0}$ and the j-th variable of the knockoff matrix $\tilde{\mathbf{Z}}_{\hat{S}_0}$ enter the model. Denote for simplicity these j-th variables by Z_j and $\tilde{Z}_j$, respectively.

Let $\bar{\boldsymbol{\beta}}(\lambda) = (\hat{\boldsymbol{\beta}}(\lambda)^T, \tilde{\boldsymbol{\beta}}(\lambda)^T)^T$ be the set of robust lasso coefficients for each value of the tuning parameter λ provided by the lasso path. $\bar{\boldsymbol{\beta}}(\lambda)$ is used to construct a feature importance statistic W_j for each variable in order to test the null hypothesis $H_0 : \beta_j = 0, \forall j \in \hat{S}_0$. This statistic W_j records the first time that Z_j or its knockoff $\tilde{Z}_j$ enters the robust lasso path, i.e. the largest penalty parameter value λ such that $\hat{\beta}_j \neq 0$ or $\tilde{\beta}_j \neq 0$, that is

$$\begin{aligned} W_j = &\text{ (largest } \lambda \text{ such that } Z_j \text{ or } \tilde{Z}_j \text{ enters the robust lasso path)} \\ &\times \begin{cases} 1 & \text{if } Z_j \text{ enters before } \tilde{Z}_j \\ -1 & \text{if } \tilde{Z}_j \text{ enters before } Z_j \end{cases}. \end{aligned} \tag{6}$$

Each W_j measures the evidence against the null hypothesis, where large and positive values of W_j would advocate a strong evidence against the null for the j-th feature, in other words, as the coefficient β_j stays for a long time in the lasso path, then most likely this feature is strongly associated with the clinical outcome. On the contrary, a negative or zero value of W_j would suggest that the j-th feature is irrelevant in the model, and the sign of W_j is independent coin flip (Barber and Candés 2015). The final knockoff selection set is calculated as $\hat{S} = \{j : W_j \geq T\}$, where T is the knockoff threshold,

$$T = \min\left\{ t \in \mathcal{W} : \frac{|\{j : W_j \leq -t\}|}{1 \vee |\{j : W_j \geq t\}|} \leq q \right\}. \tag{7}$$

Here, $q \in [0, 1]$ is the nominal FDR threshold, a predetermined target error rate, $\mathcal{W} = \{|W_j| : j \in \hat{S}_0\} \setminus \{0\}$ are the unique nonzero values of $|W_j|$'s, and $a \vee b$ denotes the maximum of a and b. Another threshold is also suggested in Barber and Candés (2015), namely the knockoff+ threshold, $T = \min\{t \in \mathcal{W} : (1 + |\{j : W_j \leq -t\}|)/(1 \vee |\{j : W_j \geq t\}|) \leq q\}$. However, for the purpose of this work we consider the threshold expressed in (7), because the knockoff+ threshold leads to more conservative solutions.

We call this robust FDR-control variable selection procedure the robust compositional knockoff filter (RCKF), which could be summarized in the following **algorithm**:

RCKF Algorithm:

Input: compositional $\mathbf{X}$, or log-compositional matrix $\mathbf{Z} = \log \mathbf{X}$, response $\mathbf{Y}$, FDR threshold q, screening sample size n_0, and screening set size $|\hat{S}_0|$.

Output: knockoff selection set $\hat{S}$.

Procedure:

1. Randomly split the data $(\mathbf{Y}, \mathbf{Z})$ into disjoint parts $(\mathbf{Y}^{(0)}, \mathbf{Z}^{(0)})$ and $(\mathbf{Y}^{(1)}, \mathbf{Z}^{(1)})$.
2. **Screening step:**
   (a) Run the robust compositional screening procedure method on $(\mathbf{Y}^{(0)}, \mathbf{Z}^{(0)})$ to identify $\hat{S}_0$.
   (b) Apply the normalization procedure $x_{ij}^{\star} = x_{ij}/(\sum_{j \in \hat{S}_0} x_{ij})$ and calculate the design matrix $\mathbf{Z}_{\hat{S}_0} = \log \mathbf{X}^{\star}$ which will be used in the following selection step.
3. **Selection step:**
   (a) Generate the recycled knockoff matrix $\tilde{\mathbf{Z}}_{\hat{S}_0}$ to construct the augmented design matrix $\mathbb{Z}_{\hat{S}_0} = [\mathbf{Z}_{\hat{S}_0}, \tilde{\mathbf{Z}}_{\hat{S}_0}]$.

(b) Solve Equation (5) to calculate $\bar{\boldsymbol{\beta}}(\lambda)$ and then the feature importance statistics W_j from $\bar{\beta}_j(\lambda)$ according to (6).
(c) Generate the selection set $\hat{S} = \{j : W_j \geq T\}$, where T is the knockoff threshold (7).

3 Simulation study

We compare RCKF with its classical counterpart CKF, which served as a benchmark for the evaluation of our procedure's performance. We considered in our comparison also a hybrid solution: RCSP as described in Sect. 2.2, followed by a classical selection step, in which the knockoff statistic W_j is the lasso path statistic. A comparison with this version will provide information on the importance of robust estimation only in the first step, or in both steps of the procedure. The fully robust version will be denoted as RCKFrob, while the hybrid version is named as RCKFcl in the following.

For each simulation scenario we generate microbiome data from the logistic normal distribution (Aitchison and Shen 1980; Lin et al. 2014; Srinivasan et al. 2021). We first generated an $n \times p$ data matrix $\mathbf{W} = (w_{ij})$ from a multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where all components of $\boldsymbol{\mu}$ are equal to 1, and the elements of $\boldsymbol{\Sigma}$ are $\sigma_{jk} = 0.5^{|j-k|}$, $j, k = 1, \ldots, p$. The matrix of covariates $\mathbf{X} = (x_{ij})$ is obtained by the transformation $x_{ij} = \exp(w_{ij}) / \sum_{k=1}^{p} \exp(w_{ik})$. We set $n = 250$ and $p = 400$ as in Srinivasan et al. (2021). For the first selection step, $n_0 = 100$ observations were randomly considered, while $n_1 = 150$ were used in the subsequent screening step. The response was generated according to the linear model (1), where $\mathbf{Z} = \log(\mathbf{X})$ is the log-compositional design matrix. The variance of the error term is chosen as $\sigma^2 = 1$. We consider different sparsity levels $|S^\star| \in \{10, 15, 20, 25\}$ for the vector of coefficients $\boldsymbol{\beta} = (-3, 3, 2.5, -1, -1.5; 3, 3, -2, -2, -2; 1, -1, 3, -2, -1; -1, 1, 2, -1, -1; 3, 3, -3, -2, -1; 0, \ldots, 0)^T$. For example, when $|S^\star| = 15$, only the first 15 elements of $\boldsymbol{\beta}$ are used, and the remaining coefficients are set to zero.

We consider the following simulation settings: a scheme without contamination, as described above, and a contaminated scenario with two contamination levels. To introduce contamination, we add to the first $\gamma\%$ (with $\gamma = 0.1,$ or 0.2) of the observations of the response variable a random error generated from a normal distribution $N(10, 1)$, and we replace the first $\gamma\%$ of the observations of the block of informative variables by values coming from a p-dimensional Logistic-Normal distribution with mean vector $(20, \ldots, 20)$ and uncorrelated components. The performance of the three methods is assessed by the empirical FDR and by the empirical power,

$$\widehat{\text{FRD}} = \text{ave}_R\left[\frac{|\{j : \beta_j = 0 \text{ and } j \in \hat{S}\}|}{|\hat{S}| \vee 1}\right],$$

$$\widehat{\text{Power}} = \text{ave}_R\left[\frac{|\{j : \beta_j \neq 0 \text{ and } j \in \hat{S}\}|}{|\hat{S}^\star|}\right],$$

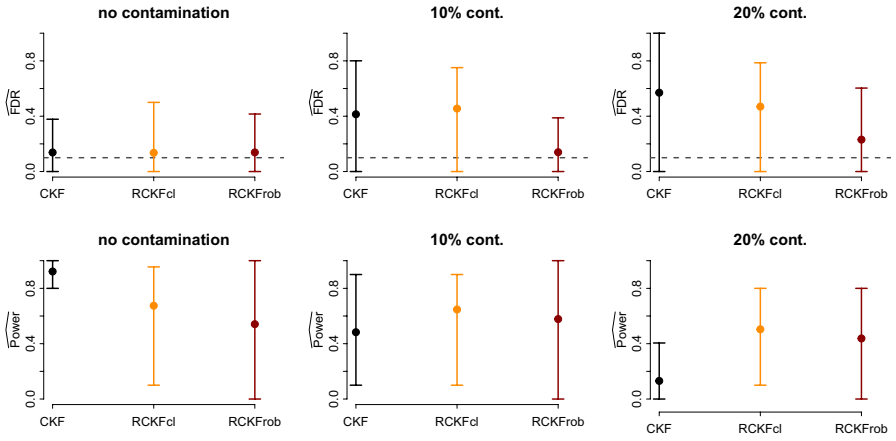

Fig. 1 Empirical FDR and power under nominal FDR of 0.1 based on 100 replicates. The dots give the average values, the error bars extend from the 5% to the 95% quantile. The dashed line represents the nominal FDR. $|S^\star| = 10$

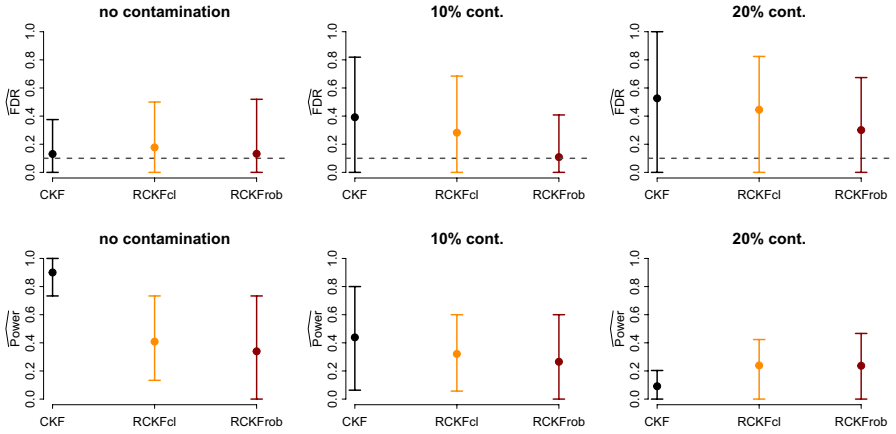

Fig. 2 Empirical FDR and power under nominal FDR of 0.1 based on 100 replicates. The dots give the average values, the error bars extend from the 5% to the 95% quantile. The dashed line represents the nominal FDR. $|S^\star| = 15$

where ave_R denotes the average over $R = 100$ simulation runs. In addition, we provide error bars to the results which cover the range from quantile 0.05 to 0.95 of the 100 replications. The results of the empirical FDR and empirical power under a nominal FDR $= 0.1$ are displayed in Figs. 1, 2, 3 and 4 for different sparsity levels.

In Fig. 1 the empirical FDR and power are reported in the most sparse model, i.e. $|S^\star| = 10$. The dots give the average values over the 100 simulation runs, the error bars extend from the 5% to the 95% quantile. The dashed line represents the

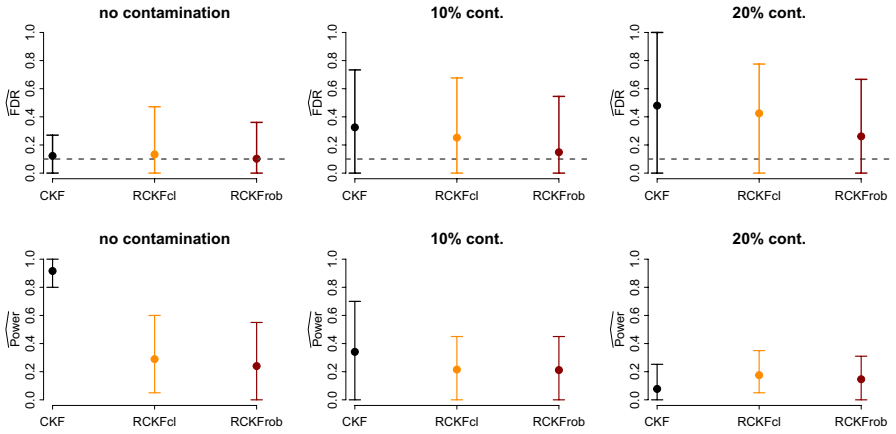

Fig. 3 Empirical FDR and power under nominal FDR of 0.1 based on 100 replicates. The dots give the average values, the error bars extend from the 5% to the 95% quantile. The dashed line represents the nominal FDR. $|S^\star| = 20$

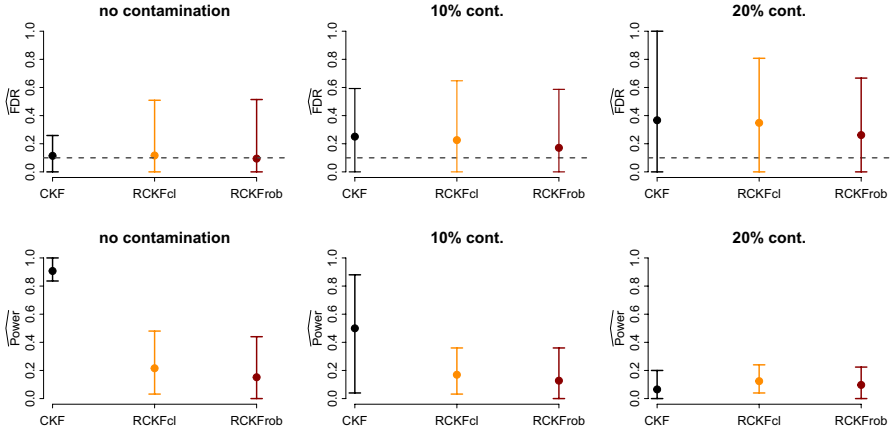

Fig. 4 Empirical FDR and power under nominal FDR of 0.1 based on 100 replicates. The dots give the average values, the error bars extend from the 5% to the 95% quantile. The dashed line represents the nominal FDR. $|S^\star| = 25$

nominal FDR of 0.1. In a non-contaminated scenario (left), CKF is on average very close to the nominal FDR, but also the robustified methods are close to that. The difference can be seen in the empirical power, where CKF is clearly superior compared to the robustified methods. In a contaminated scenario (with $\gamma = 0.1$), the non-robust CKF method suffers from a highly inflated average FDR exceeding 40%. RCKFrob is the only method that can control the nominal FDR level, while the empirical powers are quite comparable among all methods. In the more

extreme contaminated scheme (with $\gamma = 0.2$) again RCKFrob is the best method, albeit it shows an inflated average FDR level slightly above the nominal rate.

Figure 2 presents the results for $|S^\star| = 15$, and Fig. 3 those for $|S^\star| = 20$. Also in these less sparse settings, the overall picture is the same as seen before. Figure 4 highlights that as the model becomes dense, i.e., $|S^\star| = 25$, the differences among methods become less marked.

Overall, our simulation results indicate that the robust compositional knockoff filter with the robust lasso statistic (RCKFrob) has FDR control in all scenarios and is the best under the contaminated scenarios. When the models become less sparse, the empirical power suffers, especially in the uncontaminated setting, where CKF still yields very high empirical power. The CKF method works well when there are no anomalies in the data, but fails for the other scenarios when outliers exist by showing an extremely inflated FDR. In a contaminated scenario, RCKFcl performs better than CKF, though still clearly worse than RCKFrob. This empirically demonstrates the need to consider a robust second step in the compositional knockoff filter.

We conducted further simulations for each considered sparsity level to numerically evaluate the choice of the screening set size k, which can be viewed as a tuning parameter in the model as discussed in Sect. 2.2. For display purposes, we report in Figure 5 the empirical FDR and empirical power of the RCKFrob method under a nominal FDR of 0.1 based on 100 replicates, when $|S^\star| = 20$, for data with 10% of contamination, and by varying k in the grid $\{20, 25, \ldots, 45\}$. The results reveal that a screening set size equal to 20 is the best choice as it guarantees to achieve the nominal FDR with a higher power on average. However, the differences in $\widehat{\text{FDR}}$ and $\widehat{\text{Power}}$ among the different choices of k are not substantial, which suggests that the choice of k does not crucially affect the RCKF performance.

4 Application to microbiome data

The dataset considered here originates from a study presented in Altenbuchinger et al. (2017), where the association between the microbiome composition of allogenic stem cell transplants patients and urinary 3-indoxyl sulfate levels has been investigated. The authors made a pre-selection of 160 operational taxonomic units (OTUs) which are associated with the 3-indoxyl sulfate levels. In total, there are 37

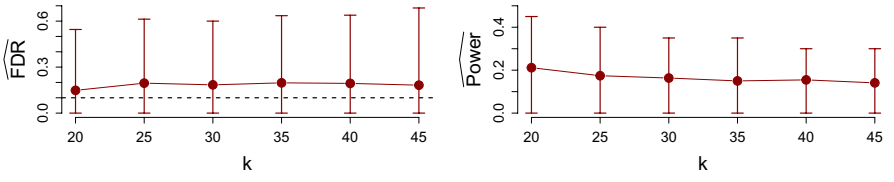

Fig. 5 Empirical FDR and power of RCKFrob method under nominal FDR of 0.1 based on 100 replicates. $|S^\star| = 20$, for data with 10% contamination, the number of screened variables of the first step varies in $\{20, 25, \ldots, 45\}$. The dots give the average values, the error bars extend from the 5% to the 95% quantile. The dashed line represents the nominal FDR

samples available, and thus we end up with a high-dimensional problem with low sample size. The OTUs contain many zeros, which we replaced by random uniform numbers independently generated in the interval [0.1, 0.5]. Different zero replacement techniques in microbiome compositional data analysis could be considered, see Lubbe et al. (2021) for a comparison. The response variable has been logarithmically transformed in order to obtain a more symmetric distribution.

For this experiment we have set the FDR to 0.25. Since the number of observations $n = 37$ is rather low, we have selected the number of observations for screening as $n_0 = 20$. As the results of the knockoff filter could strongly depend on the n_0 selected observations, we replicate the whole procedure 50 times, and then we count how often each variable has been selected by the classical CKF and the robust RCKF method. Afterwards we repeated the same experiment with contaminated data: we exchanged from the response variable the three smallest with the three biggest values.

The results are reported in Table 1. For every method we obtain 50 resulting variable sets (for the original and for the contaminated data). Among the 50 results we count how often the individual variables occur, see first column of the table. The upper part of the table is for the original data, the bottom part for the contaminated data. For example, CKF applied to the original data gives 69 unique variables (occurring at least once) in all 50 runs, 31 variables appearing at least twice, and 13 variables occur at least three times. The column "overlap" (numbers in italics) reports how many of these variables are in the overlap of the CKF and the RCKF results for the original data (top) and for the contminated data (bottom). Finally, the middle part of the table with the numbers in boldface shows the overlap for CKF original versus contaminated (left) and RCKF original versus contaminated (right). We can see that the overlap between CKF and RCKF is rather low, and quite

Table 1 Number of variables selected at least once (twice, and three times) by the classical and robust method, and number of common variables (overlap) between the classical and the robust method (italics), and the methods for original and contaminated data (boldface)

At least	CKF (orig.)	Overlap	RCKF (orig.)
Once	69	*14*	30
Twice	31	*4*	16
Three times	13	*0*	8

At least	Overlap	Overlap
Once	**39**	**12**
Twice	**7**	**5**
Three times	**1**	**3**

At least	CKF (cont.)	Overlap	RCKF (cont.)
Once	53	*17*	43
Twice	19	*2*	12
Three times	5	*0*	4

comparable whether we investigate the original or the contaminated data. Also when comparing the overlap for CKF for the original versus the contaminated data, we see a similar picture. However, this seems to be a bit different for the overlap for RCKF original versus contaminated: 12 variables in overlap if we just look at the unique variables, 5 variables in overlap among those which appear at least twice, and 3 variables in overlap when considering variables that appear at least three times (for which we have 8-original, and 4-contaminated). Thus this overlap seems to be much more stable for the robust method.

The next step is to compare regression models with those variables which have been selected at least three times (see the corresponding rows in Table 1). Since we deal with compositional covariates, we first transform them by an isometric log-ratio (ilr) transformation (Egozcue et al. 2003). For example, CKF applied to the

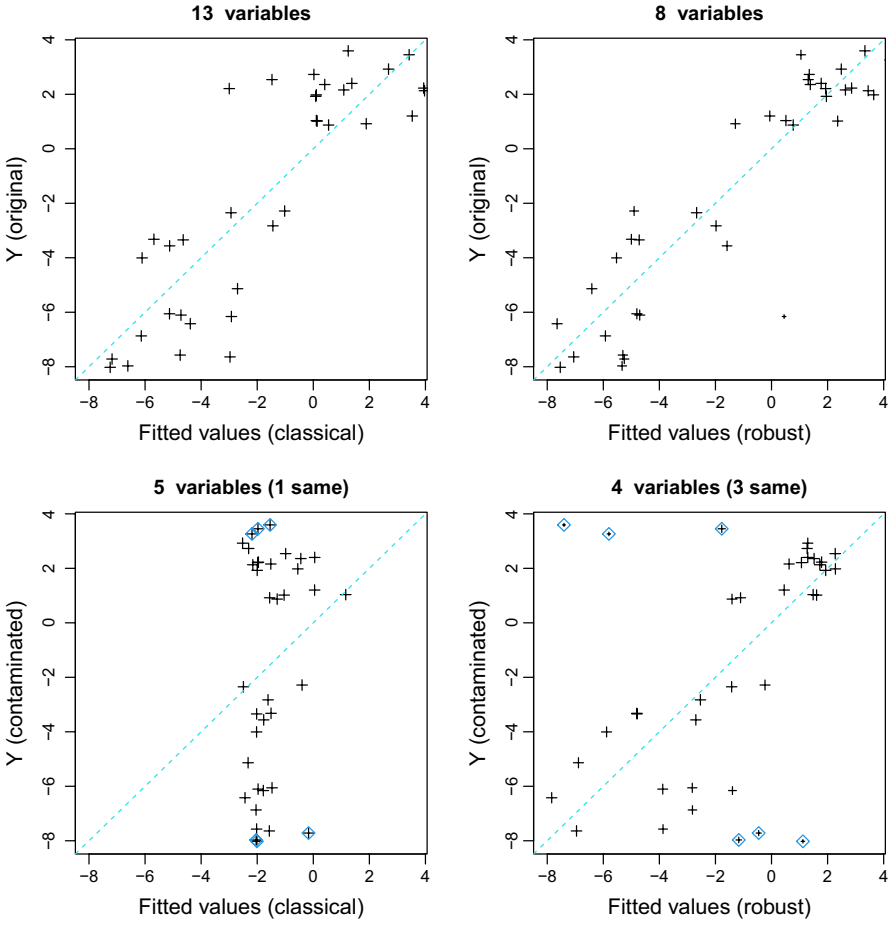

Fig. 6 Fitted values versus response for least-squares (left) and robust (right) regression models with the ilr-transformed selected variables from CFK (left) and RCFK (right) and the original (top) and contaminated (bottom) response

original data resulted in 13 variables which occurred at least three times. The corresponding ilr-transformed variables lead to a matrix of dimensionality 37×12, which is used to model the response by least-squares regression. Figure 6 (upper left) shows the resulting fitted values versus the response variable, with a very clear association. Thus, at least some of the 13 variables are related to the outcome variable. The plot on the upper right side shows the corresponding outcome with robust MM regression (Maronna et al. 2019) based on the ilr-transformed 8 selected variables by RCKF (occurring at least 3 times). Since MM regression also gives weights in [0, 1] for the observations as an output, indicating the outlyingness, we represent these weights as symbol sizes, where small symbols refer to small weight and thus to outliers. The plot shows just one deviating observation, but the remaining points reveal a quite good model for the whole range of the response variable. The bottom plots show the results from least-squares (left) and MM regression (right) when using the contaminated response variable; the points of the response which have been exchanged are marked in blue. The classical procedure seems to fail completely. Note that only 1 variable is in the intersection of the CKF selection for the original and the contaminated scenario. In contrast, the intersection for RCKF are 3 variables, and the lower right plot shows again a strong relationship between the fitted values and the contaminated response, where the robust model also downweights to a certain extent the exchanged points.

5 Conclusions

In microbiome analysis we face the methodological and computational challenge to correctly identify those abundant microbial taxa that are truly associated with an outcome of interest. To focus clinical research efforts it is crucial that the false positive identifications are properly kept under a certain limit. This problem has been addressed by the knockoff filter (Barber and Candés 2015; Candés et al. 2018; Barber and Candés 2019), which is also designed for high-dimensional data. However, this method has not been developed for compositional data, and its naive use in the microbiome context can lead to inconsistent results. Another challenge is the presence of outliers in the data, which can seriously affect all the research results.

To this aim, we have proposed a robust compositional knockoff filter for controlling the false discovery rate when performing variable selection with possibly contaminated microbiome data. For this method, the observations are randomly split into two groups, the first group serves to identify the set of possible relevant variables via a penalized robust linear log-contrast model (Monti and Filzmoser 2021), while the second is used for inference purposes by means of a robust version of the fixed-X knockoff filter procedure (Barber and Candés 2015) applied to the first screened set of features.

We have shown in numerical simulations that the RCKF ensures finite-sample FDR control under contaminated data. In such a setting, the non-robust compositional knockoff filter (CKF) of Srinivasan et al. (2021) produces a high number of false positives. Also in the uncontaminated case, and in settings with different

sparsity levels, the RCKF achieves an FDR comparable to the CKF. Although FDR control is of major concern in this context, we admit that the empirical power of the RCKF is clearly lower than for CKF in the uncontaminated case.

The application to a real microbiome dataset has shown that both, the CKF and the RCKF yield variable subsets that are strongly associated to the response. When we introduced artificial contamination, the obtained variables were more stable for the robust method than for the non-robust one. Most importantly, the subset obtained from CKF was only very weakly associated to the response, whereas the RCKF leads again to strong association, and to the ability to identify data outliers. The latter is achieved by the robust regression in the second step of the procedure. Thus, in contrast to the RCKF, outliers can seriously distort the selection abilities of the non-robust CKF.

For all these reasons we believe that the proposed RCKF based on a fixed-X knockoff machine is an attractive and feasible variable selection algorithm which guarantees a FDR control. This can bring great benefits in the analysis of high-throughput biological experiments, leading to reproducible and reliable results.

Appendix

We briefly review the knockoff construction under the fixed-X design, adopting the general setting of Barber and Candés (2015) to the notation used in this paper. The key idea is to generate a set of artificial covariates that have the same structure as the original ones but are known to be null, that is $\beta_j = 0$. For each feature $Z_j^{(1)}$, a knockoff copy $\tilde{Z}_j^{(1)}$ is constructed to satisfy

$$\left[\mathbf{Z}^{(1)}\ \tilde{\mathbf{Z}}^{(1)}\right]^\top \left[\mathbf{Z}^{(1)}\ \tilde{\mathbf{Z}}^{(1)}\right] = \begin{bmatrix} \mathbf{Z}^{(1)\top}\mathbf{Z}^{(1)} & \mathbf{Z}^{(1)\top}\tilde{\mathbf{Z}}^{(1)} \\ \tilde{\mathbf{Z}}^{(1)\top}\mathbf{Z}^{(1)} & \tilde{\mathbf{Z}}^{(1)\top}\tilde{\mathbf{Z}}^{(1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathrm{diag}\{\mathbf{s}\} \\ \boldsymbol{\Sigma} - \mathrm{diag}\{\mathbf{s}\} & \boldsymbol{\Sigma} \end{bmatrix} = \mathbf{G}, \tag{8}$$

where $\boldsymbol{\Sigma} = \mathbf{Z}^{(1)\top}\mathbf{Z}^{(1)}$ is the Gram matrix of the original $Z_j^{(1)}$ features, and $\mathbf{s} \geq \mathbf{0}$. In this way $\tilde{\mathbf{Z}}^{(1)}$ has the same covariance structure of the original design matrix $\mathbf{Z}^{(1)}$ and the cross-correlations are also preserved in the sense that $Z_j^{(1)\top} Z_k^{(1)} = \tilde{Z}_j^{(1)\top} Z_k^{(1)}$ for all $j \neq k$. A necessary and sufficient condition for $\tilde{\mathbf{Z}}^{(1)}$ to exist is that $\mathbf{G}$ is positive semidefinite. Barber and Candés (2015) suggested to choose $\mathbf{s} \in \mathbb{R}_+^{|S_0|}$ satisfying $\mathrm{diag}\{\mathbf{s}\} \preceq 2\boldsymbol{\Sigma}$, and then a valid knockoff matrix is

$$\tilde{\mathbf{Z}}^{(1)} = \mathbf{Z}^{(1)}(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\mathrm{diag}\{\mathbf{s}\}) + \tilde{\mathbf{U}}\mathbf{C}, \tag{9}$$

where $\tilde{\mathbf{U}}$ is an orthonormal matrix whose column space is orthogonal to $\mathbf{Z}^{(1)}$, and $\mathbf{C}^\top\mathbf{C} = 2\mathrm{diag}\{\mathbf{s}\} - \mathrm{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\mathrm{diag}\{\mathbf{s}\}$ is a Cholesky decomposition. It can be shown that setting $\tilde{\mathbf{Z}}^{(1)}$ as in (9) yields the correlation structure given in (8).

Acknowledgements We greatly acknowledge the DEMS Data Science Lab for supporting this work by providing computational resources. This research is supported by the FWF (Austrian Science Fund) and the GACR (Czech Science Fundation) project number I 5799-N.

Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Aitchison J (1986) The statistical analysis of compositional data. Chapman & Hall, London

Aitchison J, Bacon-Shone J (1984) Log contrast models for experiments with mixtures. Biometrika 71(2):323–330. https://doi.org/10.2307/2336249

Aitchison J, Shen SM (1980) Logistic-normal distributions: some properties and uses. Biometrika 67(2):261–272. https://doi.org/10.2307/2335470

Alfons A, Croux C, Gelper S (2013) Sparse least trimmed squares regression for analyzing high-dimensional large data sets. Ann Appl Stat 7(1):226–248. https://doi.org/10.1214/12-AOAS575

Altenbuchinger M, Rehberg T, Zacharias HU, Stämmler F, Dettmer K, Weber D, Hiergeist A, Gessner A, Holler E, Oefner PJ, Spang R (2017) Reference point insensitive molecular data analysis. Bioinformatics 33(2):219–226. https://doi.org/10.1093/bioinformatics/btw598

Barber RF, Candés EJ (2015) Controlling the false discovery rate via knockoffs. Ann Stat 43(5):2055–2085. https://doi.org/10.1214/15-AOS1337

Barber RF, Candés EJ (2019) A knockoff filter for high-dimensional selective inference. Ann Stat 47(5):2504–2537. https://doi.org/10.1214/18-AOS1755

Bates S, Candés E, Janson L, Wang W (2021) Metropolized knockoff sampling. J Am Stat Assoc 116(535):1413–1427. https://doi.org/10.1080/01621459.2020.1729163

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bertsimas D, King A, Mazumder R (2016) Best subset selection via a modern optimization lens. Ann Stat 44(2):813–852. https://doi.org/10.1214/15-AOS1388

Brzyski D, Peterson CB, Sobczyk P, Candés EJ, Bogdan M, Sabatti C (2017) Controlling the rate of GWAS false discoveries. Genetics 205(1):61–75. https://doi.org/10.1534/genetics.116.193987

Candés E, Fan Y, Janson L, Lv J (2018) Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. J R Stat Soc Ser B Stat Methodol 80(3):551–577. https://doi.org/10.1111/rssb.12265

Egozcue J, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. Math Geol 35:279–300. https://doi.org/10.1023/A:1023818214614

Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B Stat Methodol 70(5):849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ (2017) Microbiome datasets are compositional: and this is not optional. Front Microbiol 8:2224. https://doi.org/10.3389/fmicb.2017.02224

Konno H, Yamamoto R (2009) Choosing the best set of variables in regression analysis using integer programming. J Glob Optim 44(2):273–282. https://doi.org/10.1007/s10898-008-9323-9

Li H (2015) Microbiome, metagenomics, and high-dimensional compositional data analysis. Annu Rev Stat Appl 2:73–94. https://doi.org/10.1146/annurev-statistics-010814-020351
Li R, Zhong W, Zhu L (2012) Feature screening via distance correlation learning. J Am Stat Assoc 107(499):1129–1139. https://doi.org/10.1080/01621459.2012.695654
Lin W, Shi P, Feng R, Li H (2014) Variable selection in regression with compositional covariates. Biometrika 101(4):785–797. https://doi.org/10.1093/biomet/asu031
Lubbe S, Filzmoser P, Templ M (2021) Comparison of zero replacement strategies for compositional data with large numbers of zeros. Chemom Intell Lab Syst 210:104248. https://doi.org/10.1016/j.chemolab.2021.104248
Maronna RA, Martin RD, Yohai VJ, Salibián-Barrera M (2019) Robust statistics: theory and methods (with R). Wiley, Hoboken
Monti GS, Filzmoser P (2021) Sparse least trimmed squares regression with compositional covariates for high dimensional data. Bioinformatics 37(21):3805–3814. https://doi.org/10.1093/bioinformatics/btab572
Nearing JT, Douglas GM, Hayes MG, MacDonald J, Desai DK, Allward N, Jones CMA, Wright RJ, Dhanani AS, Comeau AM, Langille MGI (2022) Microbiome differential abundance methods produce different results across 38 datasets. Nat Commun 13(1):1–6. https://doi.org/10.1038/s41467-022-28034-z
Sesia M, Sabatti C, Candés EJ (2019) Gene hunting with hidden Markov model knockoffs. Biometrika 106(1):1–18. https://doi.org/10.1093/biomet/asy033
Shi P, Zhang A, Li H (2016) Regression analysis for microbiome compositional data. Ann Stat 10(2):1019–1040. https://doi.org/10.1214/16-AOAS928
Srinivasan A, Xue L, Zhan X (2021) Compositional knockoff filter for high-dimensional regression analysis of microbiome data. Biometrics 77(3):984–995. https://doi.org/10.1111/biom.13336
Storey JD (2002) A direct approach to false discovery rates. J R Stat Soc Ser B Stat Methodol 64(3):479–498. https://doi.org/10.1111/1467-9868.00346
Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci 100(16):9440–9445. https://doi.org/10.1073/pnas.1530509100
Szekely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. Ann Stat 35(6):2769–2794. https://doi.org/10.1214/009053607000000505
The Human Microbiome Project Consortium (2012) A framework for human microbiome research. Nature 486:215–221. https://doi.org/10.1038/nature11209
Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, Hyde ER, Knight R (2017) Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5(1):1–18. https://doi.org/10.1186/s40168-017-0237-y
Zhang W, Xia Y (2008) Discussion on "Sure independence screening for ultrahigh dimensional feature space''. J R Stat Soc Ser B Stat Methodol 70(2):849–911
Zhu X, Yang Y (2015) Variable selection after screening: with or without data splitting? Comput Stat 30(1):191–203. https://doi.org/10.1007/s00180-014-0528-8