Regression Models for Compositional Data: General Log-Contrast Formulations, Proximal Optimization, and Microbiome Data Applications

Patrick L. Combettes[1] · Christian L. Müller[2,3,4]

Abstract

Compositional data sets are ubiquitous in science, including geology, ecology, and microbiology. In microbiome research, compositional data primarily arise from high-throughput sequence-based profiling experiments. These data comprise microbial compositions in their natural habitat and are often paired with covariate measurements that characterize physicochemical habitat properties or the physiology of the host. Inferring parsimonious statistical associations between microbial compositions and habitat- or host-specific covariate data is an important step in exploratory data analysis. A standard statistical model linking compositional covariates to continuous outcomes is the linear log-contrast model. This model describes the response as a linear combination of log-ratios of the original compositions and has been extended to the high-dimensional setting via regularization. In this contribution, we propose a general convex optimization model for linear log-contrast regression which includes many previous proposals as special cases. We introduce a proximal algorithm that solves the resulting constrained optimization problem exactly with rigorous convergence guarantees. We illustrate the versatility of our approach by investigating the performance of several model instances on soil and gut microbiome data analysis tasks.

✉ Christian L. Müller
cmueller@flatironinstitute.org

Patrick L. Combettes
plc@math.ncsu.edu

[1] Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, USA

[2] Center for Computational Mathematics, Flatiron Institute, New York, NY 10010, USA

[3] Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany

[4] Department of Statistics, Ludwig-Maximilians-Universität München, München, Germany

🙢 Springer

1 Introduction

Compositional data sets are ubiquitous in many areas of science, spanning such disparate fields as geology and ecology. In microbiology, compositional data arise from high-throughput sequence-based microbiome profiling techniques, such as targeted amplicon sequencing (TAS) and metagenomic profiling. These methods generate large-scale genomic survey data of microbial community compositions in their natural habitat, ranging from marine ecosystems to host-associated environments. Elaborate bioinformatics processing tools [5, 6, 13, 17, 28] typically summarize TAS-based sequencing reads into sparse compositional counts of operational taxonomic units (OTUs). The quantification of the relative abundances of OTUs in the environment is often accompanied by measurements of other covariates, including physicochemical properties of the underlying habitats, variables related to the health status of the host, or those coming from other high-throughput protocols, such as metabolomics or flow cytometry.

An important step in initial exploratory analysis of such data sets is to infer parsimonious and robust statistical relationships between the microbial compositions and habitat- or host-specific measurements. Standard linear regression modeling cannot be applied in this context because the microbial count data carry only relative or compositional information. One of the most popular approaches to regression modeling with compositional covariates is the log-contrast regression model, originally proposed in [2] in the context of experiments with mixtures. The linear log-contrast model expresses the continuous outcome of interest as a linear combination of the log-transformed compositions subject to a zero-sum constraint on the regression vector. This leads to the intuitive interpretation of the response as a linear combination of log-ratios of the original compositions. In a series of papers, the linear log-contrast model has been generalized to the high-dimensional setting via regularization. The sparse linear log-contrast model, introduced in [20], considers variable selection via ℓ^1 regularization and has been extended (i) to multiple linear constraints for sub-compositional coherence across predefined groups of predictors [30]; (ii) to sub-composition selection via tree-structured sparsity-inducing penalties [33]; (iii) to longitudinal data modeling via a constraint group lasso penalty [32]; and (iv) to outlier detection via a mean shift modeling approach [23]. A common theme of these statistical approaches to log-contrast modeling is the formulation of the estimators as the solution of a convex optimization problem, and the theoretical analysis of the statistical properties of these estimators under suitable assumptions on the data.

In the present paper, we take a complementary approach and focus on the structure of the optimization problems underlying log-contrast modeling. We propose an general optimization model for linear log-contrast regression which includes previous proposals as special cases and allows for a number of novel formulations that have not yet been explored. A particular feature of our model is the joint estimation of regression vectors and associated scales for log-contrast models, similar to the scaled Lasso approach in high-dimensional linear regression [31]. This is achieved by leveraging recent results on the connection between

perspective functions and statistical models [8–10]. We introduce a Douglas–Rachford splitting algorithm that produces an exact solution to the resulting constrained optimization problems with rigorous guarantees on the convergence of the iterates. By contrast, most existing approaches to solve such problems proceed by first approximating it and then employing coordinate descent methods with less demanding convergence guarantees. We illustrate the versatility of our modeling approach by applying novel log-contrast model instances to environmental and gut microbiome data analysis tasks.

2 Linear Log-Contrast Models

We first introduce the statistical log-contrast data formation model under consideration. We then review several prominent estimators for regularized log-contrast regression models.

2.1 Statistical Log-Contrast Data Formation Model

Let Z be a known $(n \times p)$-dimensional compositional design matrix with rows $(z_i)_{1 \leqslant i \leqslant n}$ in the simplex $\left\{ (\zeta_1, \ldots, \zeta_p) \in]0, 1]^p \mid \sum_{k=1}^{p} \zeta_k = 1 \right\}$. In the microbiome context, each row represents a composition of p OTUs or components at a higher taxonomic rank. We apply a log transform $(x_i)_{1 \leqslant i \leqslant n} = (\log z_i)_{1 \leqslant i \leqslant n}$ resulting in the design matrix $X \in \mathbb{R}^{n \times p}$. In this context, we introduce the following log-contrast data formation model.

Model 1 The vector $y = (\eta_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$ of observations is

$$y = X\overline{b} + \overline{o} + Se, \quad \text{with} \quad C^\top \overline{b} = 0, \tag{1}$$

where $X \in \mathbb{R}^{n \times p}$ is the aforementioned design matrix with rows $(x_i)_{1 \leqslant i \leqslant n}$, $\overline{b} \in \mathbb{R}^p$ is the unknown regression vector (location), $\overline{o} \in \mathbb{R}^n$ is the unknown mean shift vector containing outliers, $e \in \mathbb{R}^n$ is a vector of realizations of i.i.d. zero mean random variables, $S \in [0, +\infty[^{n \times n}$ is a diagonal matrix the diagonal of which are the (unknown) standard deviations, and $C \in \mathbb{R}^{p \times K}$ is a matrix expressing K linear constraints on the regression vector.

The linear log-contrast data formation model is similar to the standard (heteroscedastic) linear model with the important difference that there are linear equality constraints on the regression vector. This stems from the fact that the entries in $X \in \mathbb{R}^{n \times p}$ are not independent due to the compositional nature. In the original model [2], the constraint matrix $C \in \mathbb{R}^{p \times K}$ is the p-dimensional all-ones vector $\mathbf{1}_p$, resulting in a zero-sum constraint on the regression vector. To gain some intuition about the implications of this constraint, consider a two-dimensional example with given estimates $b = (\beta_1, \beta_2)$, and denote by $\xi_{i,1}$ and $\xi_{i,2}$ the first and second column entries of X. The linear equality constraint enforces $\beta_2 = -\beta_1$, and thus each observation can be expressed as

$$\eta_i = \beta_1 \xi_{i,1} - \beta_1 \xi_{i,2} . \tag{2}$$

Due to the construction of the design matrix as the log transformation of the compositions, this model is equivalent to

$$\eta_i = \beta_1 \log \zeta_{i,1} - \beta_1 \log \zeta_{i,2} = \beta_1 \log \frac{\zeta_{i,1}}{\zeta_{i,2}} , \tag{3}$$

which expresses the response as a linear function of the log-ratios of the original compositional components. This example also shows that the regression coefficients in the log-contrast model bear a different interpretation than in the standard linear model. Combined log-ratio coefficients relate the response to log-fold changes of the corresponding component ratios.

2.2 Statistical Estimators for Log-Contrast Models

2.2.1 Sparse Log-Contrast Regression

In the low-dimensional setting, the standard log-contrast model with zero-sum constraints can be estimated by solving a least-squares problem subject to a linear constraint, or alternatively, via standard linear regression applied to isometrically log-ratio transformed compositions [14]. In the high-dimensional setting, we need structural assumptions on the regression vector for consistent estimation. To this end, the sparse log-contrast model was introduced in [20]. It is based on the optimization problem

$$\begin{array}{cl} \underset{\substack{b \in \mathbb{R}^p \\ \sum_{k=1}^{p} \beta_k = 0}}{\text{minimize}} & \frac{1}{2n} \|Xb - y\|_2^2 + \lambda \|b\|_1, \end{array} \tag{4}$$

where $\| \cdot \|_1$ is the ℓ^1 norm and $\lambda \in [0, +\infty[$ is a tuning parameter that balances model fit and sparsity of the solution. The estimator enjoys several desirable properties, including scale invariance, permutation invariance, and selection invariance. The latter property is intimately related to the principle of sub-compositional coherence [1] and means that the estimator is unchanged if one knew in advance the sparsity pattern of the solution and applied the procedure to the sub-compositions formed by the nonzero components. In [20], model consistency guarantees are derived for the estimator and the underlying optimization problem is approached via penalization. The proposed iterative algorithm alternates between estimating the Lagrange multipliers and solving a convex subproblem with a coordinate descent strategy. Model selection for the regularization parameter λ is performed with a generalized information criterion.

2.2.2 Sparse Log-Contrast Regression with Side Information

In many situations, it is desirable to incorporate side information about the covariates into log-contrast modeling. For instance, for microbial compositions, each component can be associated with taxonomic or phylogenetic information, thus relating the p components through a rooted taxonomic or phylogenetic tree $\mathcal{T}_p$. One way to use this hierarchical tree information is to perform log-contrast regression at a higher taxonomic level, effectively reducing the dimensionality of the regression problem. Let $\mathcal{T}_p$ be a tree with $1 \leqslant i_h \leqslant h$ levels and p leaves and assume that, at given level i_h, the p compositions split into K groups with sizes $(p_k)_{1 \leqslant k \leqslant K}$. Sub-compositional coherence across the groups can be expressed by the linear constraints $C^\top b = 0$, where C is an orthogonal $(p \times K)$-dimensional binary matrix. The kth column comprises p_k ones at the coordinates of the components that belong to the kth group. Sparse log-contrast regression with group-level compositional coherence can thus be achieved by solving the optimization problem

$$
\begin{array}{c}
\underset{\substack{b \in \mathbb{R}^p \\ C^\top b = 0}}{\text{minimize}} \quad \frac{1}{2n}\|Xb - y\|_2^2 + \lambda\|b\|_1,
\end{array}
\tag{5}
$$

where $\lambda \in [0, +\infty[$ is a tuning parameter. In [30], model consistency guarantees are derived for this estimator as well as a debiasing procedure for the final estimates. This is done by extending results from [16] to the log-contrast setting. In [20], the underlying optimization problem is approached via an augmented Lagrangian approach, while model selection is achieved by scaling a theoretically derived λ_0 with a data-driven heuristic estimate of the standard deviation σ [31], resulting in $\lambda = \lambda_0\sigma$.

An alternative way of incorporating tree information has been proposed in [33]. There, the tree structure is encoded in a parameterized matrix $J_\alpha \in \mathbb{R}^{m-1 \times p}$, where m is the number of vertices in the tree. An estimator based on the minimization problem

$$
\begin{array}{c}
\underset{\substack{b \in \mathbb{R}^p \\ \sum_{k=1}^p \beta_k = 0}}{\text{minimize}} \quad \frac{1}{2n}\|Xb - y\|_2^2 + \lambda\|J_\alpha b\|_1
\end{array}
\tag{6}
$$

is proposed, where $\lambda \in [0, +\infty[$ is a tuning parameter. The structure of J_α promotes tree-guided sparse sub-composition selection and comprises a weighting parameter $\alpha \in [0, 1]$. The authors of [33] are unable to solve the optimization in (6) exactly and resort to a heuristic that abandons the linear constraints and solves a generalized Lasso problem instead. The two tuning parameters λ and α are selected via an information criterion.

2.2.3 Robust Log-Contrast Regression

The previous estimators assume the response to be outlier-free with respect to the statistical model under consideration. One way to relax this assumption and

to guard against outliers in the response is to use a robust data fitting term. In [23], the robust log-contrast regression is introduced via mean shift modeling; see, e.g., [3, 29]. One specific instance of this framework considers the estimation problem

$$\underset{b \in \mathbb{R}^p, \, o \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2n}\|Xb - y - o\|_2^2 + \lambda_1\|b\|_1 + \lambda_2\|o\|_1, \quad \text{where} \quad C^\top b = 0, \tag{7}$$

and where nonzero elements in the mean shift vector $o \in \mathbb{R}^n$ capture outlier data, and λ_1 and λ_2 are tuning parameters. In [25], the objective function in (7) is approximated in the form of (5) with a single tuning parameter. As shown in [3] for partial linear models and in [29] for outlier detection, an equivalent form of (7) is to use the Huber function [15] as robust data fitting function and the ℓ^1 norm as regularizer. The Huber function is defined as

$$h_\rho : \mathbb{R} \to \mathbb{R} : u \mapsto \begin{cases} \rho|u| - \dfrac{\rho^2}{2}, & \text{if } |u| > \rho; \\ \dfrac{|u|^2}{2}, & \text{if } |u| \leqslant \rho, \end{cases} \tag{8}$$

where $\rho \in \,]1, +\infty[$ is a fixed parameter with default value $\rho = 1.345$ that determines the transition from the quadratic to the linear part. The model in (7) can be written as

$$\underset{\substack{b \in \mathbb{R}^p \\ C^\top b = 0}}{\text{minimize}} \quad \frac{1}{2n} \sum_{i=1}^{n} h_\rho(x_i b - \eta_i) + \lambda_1\|b\|_1. \tag{9}$$

After model estimation, each data point in the linear region of the Huber function is considered an outlier. The latter two models thus allow for joint sparse selection of predictors and outliers in a convex framework.

3 Optimization of General Log-Contrast Models

We introduce an optimization model for general log-contrast regression that includes all previous examples as special cases. We assume that the data follow the data formation model outlined in Model 1. Our model belongs to the class of perspective M-estimation models [10] and allows for joint estimation of regression parameters and corresponding scales while preserving the overall convexity of the model. We then present a proximal algorithm that can solve instances of the optimization model with theoretical guarantees on the convergence of the iterates. Finally, we propose two model selection schemes for practical regularization parameter selection that leverage the joint scale estimation capability of our optimization model.

3.1 Convex Optimization Model

Let us first introduce some notation (see [4, 27] for details). We denote by $\Gamma_0(\mathbb{R}^n)$ the class of lower semicontinuous convex functions $\varphi : \mathbb{R}^n \rightarrow]-\infty, +\infty]$ such that $\operatorname{dom} \varphi = \{x \in \mathbb{R}^n \mid \varphi(x) < +\infty\} \neq \emptyset$. Given $\varphi \in \Gamma_0(\mathbb{R}^n)$ and $x \in \mathbb{R}^n$, the unique minimizer of $\varphi + \|x - \cdot\|_2^2/2$ is denoted by $\operatorname{prox}_\varphi x$. In other words

$$\operatorname{prox}_\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^n : x \mapsto \underset{y \in \mathbb{R}^n}{\operatorname{argmin}} \left(\varphi(y) + \frac{1}{2}\|x - y\|^2 \right). \tag{10}$$

Now let D be a convex subset of $\mathbb{R}^n$. Then ι_D is the indicator function of D (it takes values 0 on D and $+\infty$ on its complement), $\operatorname{ri} D$ is the relative interior of D (its interior relative to its affine hull), and, if D is nonempty and closed, $\operatorname{proj}_D = \operatorname{prox}_{\iota_D}$ is the projection operator onto D.

The following general log-contrast optimization model enables the joint estimation of the regression vector $\bar{b} = (\bar{\beta}_k)_{1 \leqslant k \leqslant p} \in \mathbb{R}^p$ and of the scale vector $\bar{s} = (\bar{\sigma}_i)_{1 \leqslant i \leqslant N} \in \mathbb{R}^N$ in Model 1 within a convex optimization setting.

Problem 1 Consider the setting of Model 1. Let N and M be strictly positive integers, let D be a vector subspace of $\mathbb{R}^N$, let $(n_i)_{1 \leqslant i \leqslant N}$ be strictly positive integers such that $\sum_{i=1}^N n_i = n$, let $(m_i)_{1 \leqslant i \leqslant M}$ be strictly positive integers, and set $m = \sum_{i=1}^M m_i$. For every $i \in \{1, \ldots, N\}$, let $\varphi_i \in \Gamma_0(\mathbb{R}^{n_i})$, let

$$\tilde{\varphi}_i : \mathbb{R} \times \mathbb{R}^{n_i} \rightarrow]-\infty, +\infty]$$

$$(\sigma_i, u_i) \mapsto \begin{cases} \sigma_i \varphi_i(u_i/\sigma_i), & \text{if } \sigma_i > 0; \\ \sup_{u \in \operatorname{dom} \varphi_i} \left(\varphi_i(u + u_i) - \varphi_i(u) \right), & \text{if } \sigma_i = 0; \\ +\infty, & \text{if } \sigma_i < 0 \end{cases} \tag{11}$$

be the perspective of φ_i, let $X_i \in \mathbb{R}^{n_i \times p}$, and let $y_i \in \mathbb{R}^{n_i}$ be such that

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}. \tag{12}$$

Finally, set

$$E = \{b \in \mathbb{R}^p \mid C^\top b = 0\} \tag{13}$$

and, for every $i \in \{1, \ldots, M\}$, let $\psi_i \in \Gamma_0(\mathbb{R}^{m_i})$ and $L_i \in \mathbb{R}^{m_i \times p}$. The objective is to

$$\underset{s \in D, \, b \in E}{\operatorname{minimize}} \quad \sum_{i=1}^N \tilde{\varphi}_i(\sigma_i, X_i b - y_i) + \sum_{i=1}^M \psi_i(L_i b). \tag{14}$$

Remark 1 Problem 1 comprises four main components which are associated with different aspects of the general log-contrast regression model.

– The perspective functions $(\tilde{\varphi}_i)_{1 \leqslant i \leqslant N}$ play the role of the loss function in statistical estimation and couple the estimation of the scale vector s and the regression vector b. Because the functions $(\varphi_i)_{1 \leqslant i \leqslant N}$ are convex, the overall minimization problem (14) remains a convex one in (s, b).

– Problem 1 allows for the partitioning of the design matrix X and response y into N blocks with individual scale parameters $(\sigma_i)_{1 \leqslant i \leqslant N}$. This is beneficial when data from multiple measurement sources are available for the prediction of the response or when heteroscedasticity in the design matrix is expected for different groups of measurements. Introducing multiple scales has also numerical advantages. Indeed, as discussed in [10], certain proximity operators of perspective functions are easier to compute in separable form.

– The vector subspaces D and E (see (13)) enforce linear constraints on the scale vector $s = (\sigma_i)_{1 \leqslant i \leqslant N}$ and the regression vector b, respectively.

– Additional properties of the regression vector, such as (structured) sparsity, are promoted through the use of the penalization functions $(\psi_i)_{1 \leqslant i \leqslant M}$ and the matrices $(L_i)_{1 \leqslant i \leqslant M}$. The penalization functions typically contain a free parameter λ the setting of which requires a model selection strategy.

Perspective functions are discussed in [4, 8–10, 27]. The construction (11) guarantees that $(\forall i \in \{1, \ldots, N\}) \, \tilde{\varphi}_i \in \Gamma_0(\mathbb{R}^{n_i})$. We provide below two examples of perspective functions that will be used in the numerical investigations of Sect. 4.

Example 1 Consider the function $\varphi = \|\cdot\|_2^2 + 1/2$ defined on the standard Euclidean space $\mathbb{R}^P$. Then (11) yields (see Fig. 1)

$$
\tilde{\varphi} \,:\, \mathbb{R} \times \mathbb{R}^P \to \,]-\infty, +\infty]
$$

$$
(\sigma, u) \;\mapsto\; \begin{cases} \dfrac{\sigma}{2} + \dfrac{\|u\|_2^2}{\sigma}, & \text{if } \sigma > 0; \\ 0, & \text{if } \sigma = 0 \text{ and } u = 0; \\ +\infty, & \text{otherwise.} \end{cases} \tag{15}
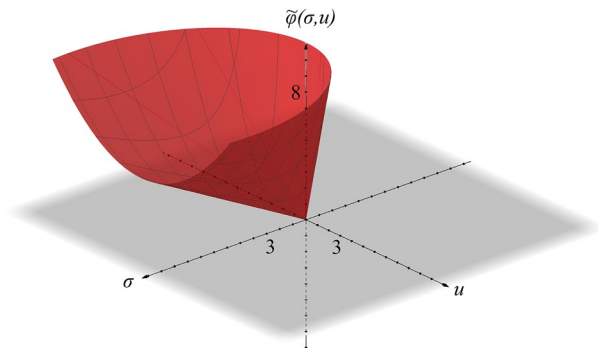$$

Fig. 1 Perspective of $\varphi = |\cdot|^2 + 1/2$

Now fix $(\sigma, u) \in \mathbb{R} \times \mathbb{R}^P$ and $\gamma \in]0, +\infty[$. If $4\gamma\sigma + \|u\|_2^2 > 2\gamma^2$, let t be the unique solution in $]0, +\infty[$ to the equation

$$\gamma t^3 + 2(2\sigma + 3\gamma)t - 8\|u\|_2 = 0, \tag{16}$$

and set $p = tu/\|u\|_2$ if $u \neq 0$, and $p = 0$ if $u = 0$. Then [10, Example 2.4] yields

$$\operatorname{prox}_{\gamma\tilde{\varphi}}(\sigma, u) = \begin{cases} \left(\sigma + \dfrac{\gamma}{2}\left(\dfrac{t^2}{2} - 1\right), u - \gamma p\right), & \text{if } 4\gamma\sigma + \|u\|_2^2 > 2\gamma^2; \\ (0, 0), & \text{if } 4\gamma\sigma + \|u\|_2^2 \leqslant 2\gamma^2. \end{cases} \tag{17}$$

A prominent estimator where the perspective function (15) is used as a loss function in conjunction with the ℓ^1 norm as penalization function is the scaled Lasso estimator for high-dimensional sparse linear regression [31].

Example 2 Set $\varphi = h_1 + 1/2$, where h_1 is the Huber function of (8). Then (11) yields (see Fig. 2)

$$\tilde{\varphi} : \mathbb{R} \times \mathbb{R} \to]-\infty, +\infty]$$

$$(\sigma, u) \mapsto \begin{cases} \dfrac{(1 - \rho^2)\sigma}{2} + \rho|u|, & \text{if } |u| > \sigma\rho \text{ and } \sigma > 0; \\ \dfrac{\sigma}{2} + \dfrac{|u|^2}{2\sigma}, & \text{if } |u| \leqslant \sigma\rho \text{ and } \sigma > 0; \\ \rho|u|, & \text{if } \sigma = 0; \\ +\infty, & \text{if } \sigma < 0. \end{cases} \tag{18}$$

Now fix $(\sigma, u) \in \mathbb{R} \times \mathbb{R}$ and $\gamma \in]0, +\infty[$. Then [10, Example 2.5] asserts that $\operatorname{prox}_{\gamma\tilde{\varphi}}(\sigma, u)$ is computed as follows.

(i) Suppose that $|u| \leqslant \gamma\rho$ and $|u|^2 \leqslant \gamma(\gamma - 2\sigma)$. Then $\operatorname{prox}_{\gamma\tilde{\varphi}}(\sigma, u) = (0, 0)$.
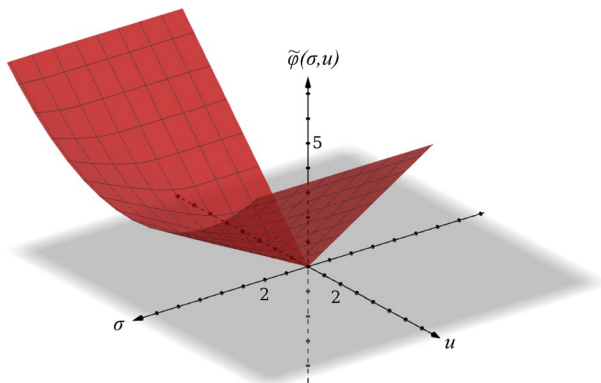

Fig. 2 Perspective of $\varphi = h_1 + 1/2$, where h_1 is the Huber function

(ii)   Suppose that $\sigma \leqslant \gamma(1 - \rho^2)/2$ and $|u| > \gamma\rho$. Then

$$\text{prox}_{\gamma\tilde{\varphi}}(\sigma, u) = \left(0, \left(1 - \frac{\gamma\rho}{|u|}\right)u\right).$$ (19)

(iii)   Suppose that $\sigma > \gamma(1 - \rho^2)/2$ and $|u| \geqslant \rho\sigma + \gamma\rho(1 + \rho^2)/2$. Then

$$\text{prox}_{\gamma\tilde{\varphi}}(\sigma, u) = \left(\sigma + \frac{\gamma}{2}(\rho^2 - 1), \left(1 - \frac{\gamma\rho}{|u|}\right)u\right).$$ (20)

(iv)   Suppose that $|u|^2 > \gamma(\gamma - 2\sigma)$ and $|u| < \rho\sigma + \gamma\rho(1 + \rho^2)/2$. If $u \neq 0$, let t be the unique solution in $]0, +\infty[$ to the equation

$$\gamma t^3 + (2\sigma + \gamma)t - 2|u| = 0.$$ (21)

Then

$$\text{prox}_{\gamma\tilde{\varphi}}(\sigma, u) = \begin{cases} \left(\sigma + \gamma(t^2 - 1)/2, u - \gamma t\text{sign}(u)\right), & \text{if } 2\gamma\sigma + |u|^2 > \gamma^2; \\ (0, 0), & \text{if } 2\gamma\sigma + |u|^2 \leqslant \gamma^2. \end{cases}$$ (22)

Using the perspective function (18) as a loss function and the ℓ^1 norm as a penalization function recovers a robust version of the scaled Lasso approach [10, 26].

3.2 Algorithm

Our algorithmic solution method to solve Problem 1 relies on an application of the Douglas–Rachford splitting algorithm in a higher-dimensional space. To describe our methodology let us first note that, since (14) involves non differentiable functions and hard constraints, it cannot be handled via methods which employ gradients. Rather, we must proceed with nonsmooth first order methods, i.e., methods which activate the functions present in the model via their proximity operators defined in (10). Let us consider the basic problem of minimizing the sum of two lower semicontinuous convex functions F and G in a Euclidean space $\mathcal{H}$, i.e.,

$$\begin{array}{c} \text{minimize } F(u) + G(u). \\ u \in \mathcal{H} \end{array}$$ (23)

Let us assume that this problem has a least one solution. A key property of the proximity operator prox_F is that its set of fixed points is the set of minimizers of F [4, Proposition 12.29]. A naive approach to solve (23) would therefore be to construct iteratively a fixed point of prox_{F+G}. However, this is not viable because prox_{F+G} is typically intractable. On the other hand, in many instances, the operators prox_F and prox_G are computable explicitly, which suggest that we design a splitting algorithm, i.e., one in which F and G are activated separately. The most popular splitting algorithm to solve (23) is the Douglas–Rachford algorithm [4, 7, 11, 12, 19, 21]. This algorithm exploits the following remarkable fact: given an arbitrary $\gamma \in]0, +\infty[$, if a point $v \in \mathcal{H}$ satisfies the fixed point property

$$\mathrm{prox}_{\gamma F}\left(2\mathrm{prox}_{\gamma G}v - v\right) = \mathrm{prox}_{\gamma G}v, \tag{24}$$

then the point $u = \mathrm{prox}_{\gamma G}v$ solves (23). This leads to the following result (see [4, Sect. 28.3]).

Theorem 2 (Douglas–Rachford algorithm) *Let $\mathcal{H}$ be a Euclidean space, let $\gamma \in\ ]0, +\infty[$, let $\varepsilon \in\ ]0, 1[$, let $v_0 \in \mathcal{H}$, and let $F \in \Gamma_0(\mathcal{H})$ and $G \in \Gamma_0(\mathcal{H})$ be such that $(\mathrm{ri\ dom}\,F) \cap (\mathrm{ri\ dom}\,G) \neq \varnothing$. Let $(\mu_k)_{k\in\mathbb{N}}$ be a sequence in $[\varepsilon, 2 - \varepsilon]$ and iterate*

$$\text{for } k = 0, 1, \ldots$$
$$\left\lfloor \begin{aligned} u_k &= \mathrm{prox}_{\gamma G}v_k \\ w_k &= \mathrm{prox}_{\gamma F}\left(2u_k - v_k\right) \\ v_{k+1} &= v_k + \mu_k(w_k - u_k). \end{aligned} \right. \tag{25}$$

Then $(u_k)_{k\in\mathbb{N}}$ converges to a solution to (23).

Our method for solving Problem 1 (Algorithm 3 below) is an implementation of (25) in a suitably constructed product space. The details of this construction are provided in Appendix A. To present the algorithm, it is convenient to introduce the matrices

$$A = \begin{bmatrix} X_1 \\ \vdots \\ X_N \\ L_1 \\ \vdots \\ L_M \end{bmatrix}, \quad Q = A^\top(\mathrm{Id} + AA^\top)^{-1}, \quad \text{and} \quad W = \mathrm{Id} - C(C^\top C)^{-1}C^\top, \tag{26}$$

together with the function

$$\mathbf{g} : \mathbb{R}^N \times \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_N} \times \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_M} \to\ ]-\infty, +\infty]$$
$$\left(s, u_1, \ldots, u_N, v_1, \ldots, v_M\right) \mapsto \sum_{i=1}^{N} \tilde{\varphi}_i(\sigma_i, u_i - y_i) + \sum_{i=1}^{M} \psi_i(v_i), \tag{27}$$

and to define, for every iteration index $k \in \mathbb{N}$, the vectors

$$\left\{ \begin{aligned} s_k &= (\sigma_{1,k}, \ldots, \sigma_{N,k}) \in \mathbb{R}^N \\ h_{s,k} &= (\eta_{1,k}, \ldots, \eta_{N,k}) \in \mathbb{R}^N \\ h_{b,k} &= (h_{1,k}, \ldots, h_{N,k}, h_{N+1,k}, \ldots, h_{N+M,k}) \\ &\quad \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_N} \times \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_M} \\ z_{b,k} &= (z_{1,k}, \ldots, z_{N,k}, z_{N+1,k}, \ldots, z_{N+M,k}) \\ &\quad \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_N} \times \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_M} \\ d_{s,k} &= (\delta_{1,k}, \ldots, \delta_{N,k}) \in \mathbb{R}^N \\ d_{b,k} &= (d_{1,k}, \ldots, d_{N,k}, d_{N+1,k}, \ldots, d_{N+M,k}) \\ &\quad \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_N} \times \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_M}. \end{aligned} \right. \tag{28}$$

Algorithm 3 Let $\gamma \in]0, +\infty[$, $\varepsilon \in]0, 1[$, $x_{s,0} \in \mathbb{R}^N$, $x_{b,0} \in \mathbb{R}^p$, $h_{s,0} \in \mathbb{R}^N$, and $h_{b,0} \in \mathbb{R}^{n+m}$. Iterate

$$
\begin{aligned}
&\text{for } k = 0, 1, \ldots \\
&\left|
\begin{aligned}
&\mu_k \in [\varepsilon, 2 - \varepsilon] \\
&q_{s,k} = x_{s,k} - h_{s,k} \\
&q_{b,k} = Ax_{b,k} - h_{b,k} \\
&s_k = x_{s,k} - q_{s,k}/2 \\
&b_k = x_{b,k} - Qq_{b,k} \\
&c_{s,k} = \text{proj}_D(2s_k - x_{s,k}) \\
&c_{b,k} = W(2b_k - x_{b,k}) \\
&x_{s,k+1} = x_{s,k} + \mu_k(c_{s,k} - s_k) \\
&x_{b,k+1} = x_{b,k} + \mu_k(c_{b,k} - b_k) \\
&\text{for } i = 1, \ldots, N \\
&\left|
\begin{aligned}
&z_{i,k} = X_i b_k \\
&(\delta_{i,k}, d_{i,k}) = (0, y_i) + \text{prox}_{\gamma\tilde{\varphi}_i}(2\sigma_{i,k} - \eta_{i,k}, 2z_{i,k} - h_{i,k} - y_i)
\end{aligned}
\right. \\
&\text{for } i = 1, \ldots, M \\
&\left|
\begin{aligned}
&z_{N+i,k} = L_i b_k \\
&d_{N+i,k} = \text{prox}_{\gamma\psi_i}(2z_{N+i,k} - h_{N+i,k})
\end{aligned}
\right. \\
&h_{s,k+1} = h_{s,k} + \mu_k(d_{s,k} - s_k) \\
&h_{b,k+1} = h_{b,k} + \mu_k(d_{b,k} - z_{b,k}).
\end{aligned}
\right.
\end{aligned}
\tag{29}
$$

Proposition 1 *Consider the setting of Problem* 1. *Suppose that*

$$
\lim_{\substack{s \in D, \ b \in E \\ \|s\|_2 + \|b\|_2 \to +\infty}} \mathbf{g}(s, Ab) = +\infty
\tag{30}
$$

and that

$$
\big(D \times A(E)\big) \cap \text{ri dom } \mathbf{g} \neq \varnothing.
\tag{31}
$$

Then Problem 1 *has at least one solution. Now let* $(s_k)_{k\in\mathbb{N}}$ *and* $(b_k)_{k\in\mathbb{N}}$ *be sequences generated by Algorithm* 3. *Then* $(s_k)_{k\in\mathbb{N}}$ *converges to some* $s \in \mathbb{R}^N$ *and* $(b_k)_{k\in\mathbb{N}}$ *converges to some* $b \in \mathbb{R}^p$ *such that* (s, b) *solves Problem* 1.

Proof See Appendix A. ☐

In most practical situations, (30) and (31) are typically satisfied. For example the following describes a scenario that will be encountered in Sect. 4.

Proposition 2 *Consider the setting of Problem* 1 *and suppose that the following additional properties hold:*

(i) *For every* $i \in \{1, \ldots, N\}$, $\varphi_i = \theta_i + \alpha_i$, *where* $\theta_i : \mathbb{R}^{n_i} \to [0, +\infty[$ *is convex and* $\alpha_i \in]0, +\infty[$.

(ii) *For every* $i \in \{1, \ldots, M\}$, $\psi_i : \mathbb{R}^{m_i} \to [0, +\infty[$.

(iii) *For some $j \in \{1, \dots, M\}$, $\psi_j(L_j b) \to +\infty$ as $\|b\|_2 \to +\infty$ while $C^\top b = 0$.*

(iv) $D \cap]0, +\infty[^N \neq \varnothing.$

Then (30) and (31) are satisfied.

Proof See Appendix B. □

3.3 Model Selection

In the context of log-contrast regression, a number of different model selection strategies have been proposed, including stability selection [20, 22] and Generalized Information Criteria [32]. In [30], a scale-dependent tuning parameter has been derived where the optimal scale has been found via line search. Our joint scale and regression modeling approach makes this line search obsolete, thus yielding a parameter-free model selection scheme. More specifically, we consider two model selection schemes. Firstly, following [30], we consider

$$\lambda_0 = \sqrt{2} q_n(r/p), \tag{32}$$

where $q_n(t) = n^{-1/2} \Phi^{-1}(1 - t)$, Φ^{-1} is the quantile function for the standard normal distribution, and r is the solution to the equation $r = q_1^4(r/p) + 2q_1^2(r/p)$. In practice, this data-independent model selection scheme may lead to inclusion of spurious coefficients. To assess the robustness of the inferred solutions we combine this theoretically derived regularization with stability selection [22]. The original stability selection approach selects, for every subsample, a small set of predictors from the regularization path, e.g., the first q predictors that appear along the path or the q coefficients that are largest in absolute value across the entire path. We here propose to select, for every subsample, the nonzero coefficients present at regularization parameter λ_0. Note that λ_0 is sample-size dependent and hence needs to be adapted to the specific subsample size used in stability selection. As default values, we consider a subsample size of $\lceil n/2 \rceil$ and generate 100 subsamples. The key diagnostic in stability selection is the selection frequency profile for each coefficient. To select a stable set of coefficients, a threshold parameter $t_s \in [0.6, 0.9]$ is recommended [22], where all coefficients with selection frequency above t_s are included in the final model.

4 Applications to Compositional Microbiome Data

We apply several instances of the general log-contrast model outlined in Problem 1 in the context of microbiome data analysis tasks. We set $M = 1$, $m_1 = m$, $L_1 = \mathrm{Id}$, and employ as a regularization function the ℓ^1 norm $\psi_1 = \lambda \|\cdot\|_1$. We use the functions in Examples 1 and 2 as instances of the perspective loss functions $\tilde{\varphi}_1$. We refer to these instances as Least Squares and Huber log-contrast model, respectively. Thus, in case of the Least Squares model, (14) becomes

$$\underset{\sigma \in \mathbb{R},\, b \in E}{\text{minimize}} \quad \widetilde{\|\cdot\|_2^2}(\sigma, Xb - y) + \lambda \|b\|_1, \tag{33}$$

while in the case of the Huber model it becomes

$$\begin{aligned}
&\underset{s \in D, \, b \in E}{\text{minimize}} \quad \sum_{i=1}^{n} \widetilde{h}_\rho \big(\sigma_i, x_i b - \eta_i \big) + \lambda \|b\|_1, \\
&\text{where} \quad D = \big\{ (\sigma, \dots, \sigma) \in \mathbb{R}^n \,\big|\, \sigma \in \mathbb{R} \big\}.
\end{aligned} \tag{34}$$

Note that the projection of a vector $s \in \mathbb{R}^n$ onto D, as required in Algorithm 3, is given by

$$\text{proj}_D s = \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i, \dots, \frac{1}{n} \sum_{i=1}^{n} \sigma_i \right). \tag{35}$$

Dependent on the application, we use different zero-sum constraints on b as specified by the matrix C. To solve the various instances of Problem 1, we use Algorithm 3 and set the parameter $\mu_k = 1.9$ and $\gamma = 1$. We consider that the algorithm has converged when $\|b_k - b_{k+1}\|_2 < 10^{-6}$. All computational experiments are fully reproducible with the code available at https://github.com/muellsen/PCM/tree/master/examples/LogContrastModels.

4.1 Body Mass Index Prediction from Gut Microbiome Data

We first consider a cross-sectional study that examines the relationship between diet and gut microbiome composition, where additional demographic covariates, including body mass index (BMI) are available, referred to as COMBO data set [34]. After pre-processing and filtering, the data set comprises the log-transformed relative abundances of $p = 87$ taxa at the genus level across $n = 96$ healthy subjects. Following previous analyses [20, 30], we investigate the relationship between BMI and the microbial compositions in a log-contrast regression framework. We use $C = \mathbf{1}_p$ to model the standard zero-sum constraint. In addition to the compositional covariates, two covariate measurements, fat and calorie intake, are also taken into account via joint unpenalized least squares. We investigate the influence of different loss functions, Least Squares and Huber, as well as the sub-compositional constraints on the quality of the estimation, the size of the support set, and the predictive power. Further numerical results can be found in Appendix C.

To highlight the ability of the algorithm to jointly estimate regression and scale we solve the two problems over the regularization path with 40 λ values on a log-linear grid in $[0.0069, \dots, 0.6989]$. We also consider the theoretically derived regularization parameter $\lambda_0 = 0.1997$ from (32). Figure 3a and b show the solution path of the regression vector b for the sparse Least Squares log-contrast model and the Huber model, respectively. Figure 3c displays the corresponding joint scale estimates σ for the Least Squares and the Huber models. The estimated regression coefficients at λ_0 are highlighted in Fig. 3d. Both models agree on a set of six genera, including Clostridium as strongest negative and Acidaminococcus as the strongest positive predictors. This implies that the log-ratio of Acidaminococcus to Clostridium is positively associated with BMI. Other genera include Alistipes, Megamonas, and Coprobacillus with negative coefficients, and Dorea with positive coefficient. In [20, 30], the genera Alistipes, Clostridium,

Acidaminococcus, and Allisonella have been identified as key predictors. The solutions of the perspective log-contrast models corroborates these finding for Clostridium and Acidaminococcus, and to a less extent to Alistipes, whereas the genus Allisonella has only a small strictly positive coefficient in both log-contrast models (Fig. 3d).

Next, we consider the stability selection scheme introduced in Sect. 3.3 with default parameters and threshold $t_s = 0.7$. Figure 4a shows the stability-based frequency profile for the sparse Least Squares and Huber log-contrast models. For both models, only Clostridium and Acidaminococcus are selected. Stability selection thus leads to a simple explanatory log-ratio model formed by the ratio of the relative abundances of Acidaminococcus to Clostridium. However, when considering the final model prediction results, as shown in Fig. 4b for the Huber model, this model can only explain normal to overweight participants (BMI 20–30) because 34 out of 96 participants are considered outliers in the Huber model. The overall refitted R^2 is 0.19 under the Huber model but increases to 0.43 for the 62 inlier participants.

Next, we investigate the influence of sub-compositional constraints on the stability selection frequency for the two estimation procedures. We follow the analysis of
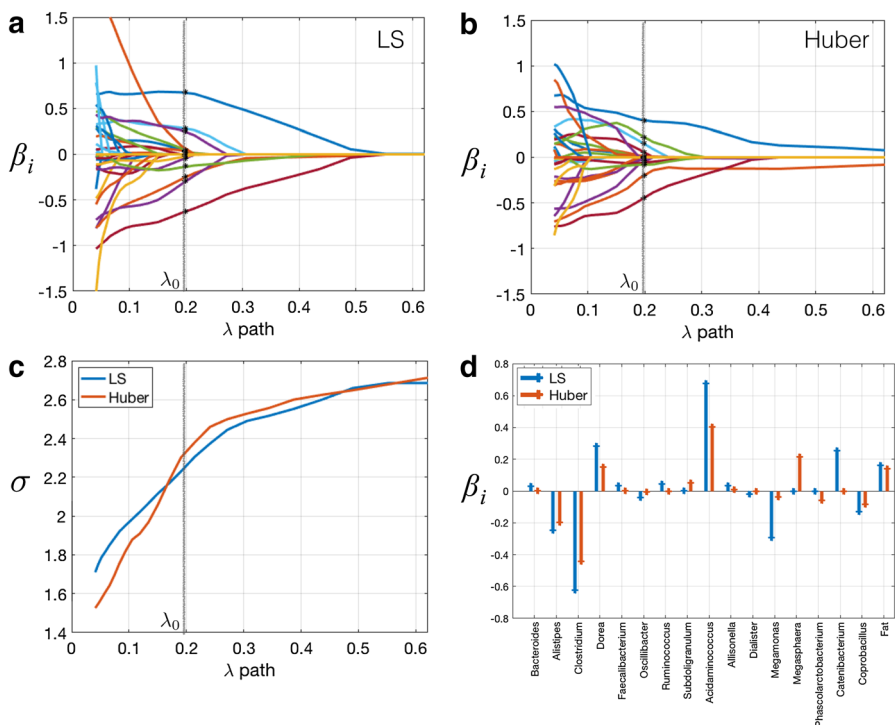
Fig. 3 **a** Solution path of regression vector b in the sparse Least Squares log-contrast model on the full COMBO data. The grey line marks the theoretical λ_0 from (32). **b** Solution path of regression vector b in sparse Huber log-contrast model on the full COMBO data. **c** Solution path of the scale estimates σ for both log-contrast models on the full COMBO data. **d** Comparison of the regression estimates of both models at regularization parameter λ_0 on the full data set. Both models agree on the two strongest predictors, the genera Clostridium and Acidaminococcus (Color figure online)

[30] and consider a subset of 45 genera that have the highest relative abundances in the data set. These 45 genera belong to $K = 4$ distinct phyla: Actinobacteria (two genera), Bacteroides (eight genera), Firmicutes (32 genera), and Proteobacteria (three genera). The constraint matrix C is hence an orthogonal (45×4)-dimensional binary matrix. Figure 5a and b show stability selection profile for both the Least Squares and the Huber model with and without compositional constraints, respectively. Figure 5c shows the difference in the selection frequency profiles. Although several genera, including Collinsella, Paraprevotella, Parabacteroides, Faecalibacterium, Oscillibacter, and Parasutterela display significant frequency differences, the two genera Clostridium and Acidaminococcus, both belonging to the Firmicutes phylum, demonstrate again the highest stability both with and without sub-compositional constraints.

4.2 Relationship Between Soil Microbiome and pH Concentration

We next consider a dataset put forward in [18] comprising $n = 88$ soil samples from North and South America. Both amplicon sequencing data and environmental covariates, including pH concentrations, are available and have been re-analyzed via a balance tree approach in [24]. The amplicon data contains $p = 116$ OTUs, and we consider $C = \mathbf{1}_p$. We perform stability selection with default parameters as outlined in Sect. 3.3. We refer to Appendix D for results regarding variable selection with the theoretical λ_0 value. The selection frequency of the different regression coefficients is shown Fig. 6a. At stability threshold $t_s = 0.7$, seven taxa are selected in the Least Squares models, and six taxa in the Huber model, respectively. After re-estimation of the two perspective log-contrast models on the selected subset, two taxa of order Ellin6513, one taxon of family Koribacteraceae, and one taxon of genus Rhodoplanes have negative coefficients whereas two taxa belonging to the genus Balneimonas as well as one Rubrobacter taxon and one taxon of order RB41 have positive coefficients

Fig. 4 **a** Stability selection profile for all taxa selected with a frequency > 0.1 in the Least Squares model (blue) or the Huber log-contrast model (red), respectively. The green solid line marks the stability threshold $t_s = 0.7$, selecting the genera Clostridium and Acidaminococcus. **b** Prediction of BMI from the log-contrast of the two genera in the Huber log-contrast model vs. measurements for 62 inliers (blue) and 34 outliers (red) (overall $R^2 = 0.19$) (Color figure online)

Fig. 5 **a** Stability selection profiles for the subset of 45 taxa selected with a frequency > 0.1 in either the sparse Least Squares (blue) or the Huber (red) log-contrast model with sub-compositional constraints. **b** Same as (**a**) but without sub-compositional constraints. **c** Stability selection frequency differences between the two approaches. Several genera show significant differences. The colors signify the different phyla that the genera belong to and the non-compositional covariates fat and diet intake (Color figure online)

(Fig. 6b). The seven taxa identified in the Least Squares model thus allow for a compact representation with four log-ratios of compositions. The Huber model with six identified taxa requires only three log-ratios. The five coefficients that are selected in both models agree in coefficient sign but show small variations in coefficient values. The Huber model ($R^2 = 0.86$) deems 33 data points to be outliers in the final estimate (Fig. 6c). For completeness, we include the mean absolute deviation (MAD) between model estimates and data in Fig. 6d. The selected taxa cover a wide range of average pH levels (as provided in [24]), ranging from 4.9 to 6.75, implying that the learned model may indeed generalize to other soil types not present in the current data set.

5 Discussion and Conclusion

Finding linear relationships between continuous variables and compositional predictors is a common task in many areas of science. We have proposed a general estimation model for high-dimensional log-contrast regression which includes many
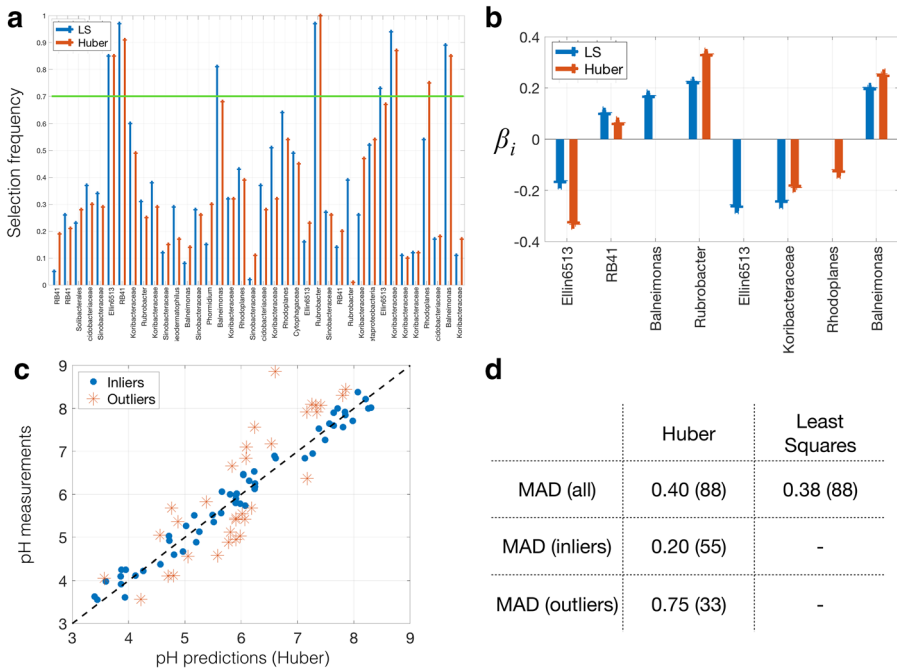
Fig. 6 **a** Stability selection profile for all taxa (denoted by the lowest taxonomic rank available) selected with a frequency > 0.1 in either the sparse Least Squares (blue) log-contrast model or the Huber model (red). The green solid line marks the stability selection threshold $t_s = 0.7$. **b** Refitted values of all selected log-contrast regression coefficients for Least Squares (seven coefficients in blue) and the Huber model (six coefficients in red). **c** Prediction of pH measurements from the Huber model for inliers (blue) and outliers (red) ($R^2 = 0.86$). **d** Table summarizing the mean absolute deviation (MAD) of the two model estimates on the data. Numbers in parentheses represent the number of inlier and outlier data points for the Huber model (Color figure online)

previous proposals as special cases [20, 23, 30, 33]. Our model belongs to the class of perspective M-estimation models [10] which allows for scale estimation in the data fitting term while preserving the overall convexity of the underlying model. This is made possible due to recent advances in the theory of perspective functions [8–10].

Several data fitting and penalty functions are available in the present framework. For instance, the robust Huber model is a convenient choice when outliers are suspected in the continuous outcome vector, or equivalently, when only a subset of the outcome data is expected to follow a linear log-contrast relationship with the compositional predictors [10, 23]. Combined with a sparsity-inducing penalty, the model allows for joint scale estimation, outlier detection, and variable selection in a single framework. Alternative choices for data fitting and regularization models are available in [10]. Our framework also enables sub-compositional coherence across groups of variables, e.g., bacterial phyla in microbiome data, via general linear constraints.

We have introduced a Douglas–Rachford algorithm that can solve the corresponding constrained nonsmooth convex optimization problems with rigorous guarantees on the convergence of the iterates. Furthermore, we have illustrated the viability of our approach on two microbiome data analysis tasks: body mass index (BMI) prediction

from gut microbiome data in the COMBO study and pH prediction from soil microbial abundance data. Our joint regression and scale estimation enabled the use of a universal single tuning parameter λ_0 [30] to control the sparsity of the estimates. We have combined this approach with stability-based model selection [22] to determine sparse stable sets of log-contrast predictors. For the gut microbiome BMI analysis, the robust Huber log-contrast model identified two genera whose log-ratio predicts BMI well for normal to overweight participants while simultaneously identifying outliers with respect to the log-contrast model. In the soil microbiome data set, we derived parsimonious pH prediction models. The Least Squares model requires four log-ratios of microbial compositions and achieves an overall $R^2 = 0.88$. The Huber model requires only three log-ratios of microbial compositions with an overall $R^2 = 0.86$.

Going forward, we believe that the general log-contrast model and the associated optimization and model selection techniques presented here will provide a valuable off-the-shelf tool for log-contrast regression analysis when compositional data such as microbial relative abundance data are used as predictors in exploratory data analysis. Future efforts will include the integration of the presented models in modern computational microbiome analysis software workflows.

Appendix A: Proof of Proposition 1

Define g as in (27) and set

$$
\begin{cases}
\mathsf{f} = \iota_{E \times D} \\
\mathsf{L} : \mathbb{R}^N \times \mathbb{R}^p \to \mathbb{R}^N \times \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_N} \times \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_M} \\
(s, b) \quad \mapsto (s, Ab) = \big(s, X_1 b, \ldots, X_N b, L_1 b, \ldots, L_M b\big).
\end{cases}
\tag{36}
$$

Then $\mathsf{f} \in \Gamma_0(\mathbb{R}^{N+p})$ as the indicator of the vector subspace $D \times E$, and

$$
(\forall (s, b) \in \mathbb{R}^{N+p}) \quad \mathrm{prox}_{\gamma \mathsf{f}}(s, b) = \big(\mathrm{proj}_D s, \mathrm{proj}_E b\big) = \big(\mathrm{proj}_D s, Wb\big),
\tag{37}
$$

where the last identity follows from [4, Proposition 29.17(iii)]. On the other hand, it follows from [8, Proposition 2.3(ii)] and [4, Proposition 8.6] that $\mathbf{g} \in \Gamma_0(\mathbb{R}^{N+n+m})$. Furthermore, we derive from [4, Propositions 24.11 and 24.8(ii)] that

$$(\forall(s, u_1, \ldots, u_N, v_1, \ldots, v_M) \in \mathbb{R}^N \times \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_N})$$

$$\text{prox}_{\gamma \mathbf{g}}(s, u_1, \ldots, u_N, v_1, \ldots, v_M) = \big((0, y_1) + \text{prox}_{\gamma \tilde{\varphi}_1}(\sigma_1, u_1 - y_1), \ldots \quad (38)$$

$$\ldots, (0, y_N) + \text{prox}_{\gamma \tilde{\varphi}_N}(\sigma_N, u_N - y_N), \text{prox}_{\gamma \psi_1} v_1, \ldots, \text{prox}_{\gamma \psi_M} v_M\big).$$

In addition, (31) implies that

$$\begin{aligned} \mathsf{L}(\text{dom}\, \mathbf{f}) \cap \text{dom}\, \mathbf{g} &= \big(\mathsf{L}(D \times E)\big) \cap \text{dom}\, \mathbf{g} \\ &= \big(D \times A(E)\big) \cap \text{dom}\, \mathbf{g} \\ &\neq \varnothing. \end{aligned} \quad (39)$$

Consequently, $\text{dom}\,(\mathbf{f} + \mathbf{g} \circ \mathsf{L}) \neq \varnothing$. Thus,

$$\mathbf{f} + \mathbf{g} \circ \mathsf{L} \in \Gamma_0(\mathbb{R}^{N+p}) \quad (40)$$

while, using the variable $\mathbf{w} = (s, b) \in \mathbb{R}^{N+p}$, (30) and (36) imply that

$$\lim_{\mathbf{w} \in \mathbb{R}^{N+p}, \|\mathbf{w}\|_2 \to +\infty} \mathbf{f}(\mathbf{w}) + \mathbf{g}(\mathsf{L}\mathbf{w}) = +\infty. \quad (41)$$

It therefore follows from [4, Proposition 11.15(i)] that

$$\text{Argmin}(\mathbf{f} + \mathbf{g} \circ \mathsf{L}) \neq \varnothing. \quad (42)$$

Since (14) is equivalent to

$$\begin{array}{c} \underset{\mathbf{w} \in \mathbb{R}^{N+p}}{\text{minimize}} \quad \mathbf{f}(\mathbf{w}) + \mathbf{g}(\mathsf{L}\mathbf{w}), \end{array} \quad (43)$$

we infer from (42) that Problem 1 admits at least one solution. Note that (43) can be rewritten as

$$\begin{array}{c} \underset{\substack{\mathbf{w} \in \mathbb{R}^{N+p} \\ \mathbf{z} \in \mathbb{R}^{N+n+m}}}{\text{minimize}} \quad \mathbf{f}(\mathbf{w}) + \mathbf{g}(\mathbf{z}). \\ \mathsf{L}\mathbf{w} = \mathbf{z} \end{array} \quad (44)$$

Now set $\boldsymbol{u} = (\mathbf{w}, \mathbf{z}) \in \mathcal{H} = \mathbb{R}^{2N+m+n+p}$ and

$$\begin{cases} \boldsymbol{F} : \mathcal{H} \to]-\infty, +\infty] : (\mathbf{w}, \mathbf{z}) \mapsto \mathbf{f}(\mathbf{w}) + \mathbf{g}(\mathbf{z}) \\ \boldsymbol{G} = \iota_V, \quad \text{where} \quad V = \big\{(\mathbf{x}, \mathbf{h}) \in \mathcal{H} \mid \mathsf{L}\mathbf{x} = \mathbf{h}\big\}. \end{cases} \quad (45)$$

Then $\boldsymbol{F} \in \Gamma_0(\mathcal{H}), \boldsymbol{G} \in \Gamma_0(\mathcal{H})$, and (44) is equivalent to

$$\begin{array}{c} \underset{\boldsymbol{u} \in \mathcal{H}}{\text{minimize}} \quad \boldsymbol{F}(\boldsymbol{u}) + \boldsymbol{G}(\boldsymbol{u}). \end{array} \quad (46)$$

Moreover, we deduce from (31) that

$$L(\text{dom f}) \cap \text{ri dom g} = \big(D \times A(E)\big) \cap \text{ri dom g} \neq \varnothing. \tag{47}$$

Consequently, using standard relative interior calculus [27, Sect. 6], (45) yields

$$
\begin{aligned}
\text{ri}\,(\text{dom}\,\boldsymbol{G}) \cap \text{ri}\,(\text{dom}\,\boldsymbol{F}) &= \boldsymbol{V} \cap \text{ri}\,\big(\text{dom f} \times \text{dom g}\big) \\
&= \boldsymbol{V} \cap \big(\text{ri dom f} \times \text{ri dom g}\big) \\
&= \boldsymbol{V} \cap (\text{dom f} \times \text{ri dom g}) \\
&= \big\{(\mathsf{x}, \mathsf{Lx}) \,\big|\, \mathsf{x} \in \mathbb{R}^{N+p}\big\} \cap (\text{dom f} \times \text{ri dom g}) \\
&\neq \varnothing.
\end{aligned} \tag{48}
$$

Therefore, given $\gamma \in\,]0, +\infty[$, $\varepsilon \in\,]0, 1[$, $v_0 \in \mathcal{H}$, and a sequence $(\mu_k)_{k\in\mathbb{N}}$ in $[\varepsilon, 2 - \varepsilon]$, the nonsmooth convex minimization problem (46) can be solved using the Douglas–Rachford algorithm (25) which, by Theorem 2, produces a sequence $(u_k)_{k\in\mathbb{N}}$ that converges to a solution to (46). Next, it follows from [4, Proposition 24.11 and Example 29.19(i)] that

$$
\begin{cases}
\text{prox}_{\gamma F} : (\mathsf{w}, \mathsf{z}) \mapsto \big(\text{prox}_{\gamma f}\mathsf{w}, \text{prox}_{\gamma g}\mathsf{z}\big) \\
\text{prox}_{\gamma G} : (\mathsf{x}, \mathsf{h}) \mapsto (\mathsf{w}, \mathsf{Lw}), \text{ where } \mathsf{w} = \mathsf{x} - \mathsf{L}^\top\big(\text{Id} + \mathsf{LL}^\top\big)^{-1}(\mathsf{Lx} - \mathsf{h}).
\end{cases} \tag{49}
$$

Now define

$$
\mathsf{R} = \mathsf{L}^\top(\text{Id} + \mathsf{LL}^\top)^{-1} \quad \text{and} \quad (\forall k \in \mathbb{N}) \quad
\begin{cases}
\boldsymbol{u}_k = (\mathsf{w}_k, \mathsf{z}_k) \\
\boldsymbol{v}_k = (\mathsf{x}_k, \mathsf{h}_k) \\
\boldsymbol{w}_k = (\mathsf{c}_k, \mathsf{d}_k).
\end{cases} \tag{50}
$$

Then we derive from (49) that, given $\mathsf{x}_0 \in \mathbb{R}^{N+p}$ and $\mathsf{h}_0 \in \mathbb{R}^{N+n+m}$, (25) becomes

for $k = 0, 1, \dots$
$$
\begin{cases}
\mathsf{q}_k = \mathsf{Lx}_k - \mathsf{h}_k \\
\mathsf{w}_k = \mathsf{x}_k - \mathsf{Rq}_k \\
\mathsf{z}_k = \mathsf{Lw}_k \\
\mathsf{c}_k = \text{prox}_{\gamma f}(2\mathsf{w}_k - \mathsf{x}_k) \\
\mathsf{d}_k = \text{prox}_{\gamma g}(2\mathsf{z}_k - \mathsf{h}_k) \\
\mathsf{x}_{k+1} = \mathsf{x}_k + \mu_k(\mathsf{c}_k - \mathsf{w}_k) \\
\mathsf{h}_{k+1} = \mathsf{h}_k + \mu_k(\mathsf{d}_k - \mathsf{z}_k).
\end{cases} \tag{51}
$$

Let us partition the vectors appearing in (51) according to their scale and regression components as

$$(\forall k \in \mathbb{N}) \quad \begin{cases} \mathsf{x}_k = (x_{s,k}, x_{b,k}) \in \mathbb{R}^N \times \mathbb{R}^p \\ \mathsf{h}_k = (h_{s,k}, h_{b,k}) \in \mathbb{R}^N \times \mathbb{R}^{n+m} \\ \mathsf{q}_k = (q_{s,k}, q_{b,k}) \in \mathbb{R}^N \times \mathbb{R}^{n+m} \\ \mathsf{w}_k = (s_k, b_k) \in \mathbb{R}^N \times \mathbb{R}^p \\ \mathsf{z}_k = (s_k, z_{b,k}) \in \mathbb{R}^N \times \mathbb{R}^{n+m} \\ \mathsf{c}_k = (c_{s,k}, c_{b,k}) \in \mathbb{R}^N \times \mathbb{R}^p \\ \mathsf{d}_k = (d_{s,k}, d_{b,k}) \in \mathbb{R}^N \times \mathbb{R}^{n+m}. \end{cases} \tag{52}$$

In terms of these new variables, using the matrix Q of (26), (36) and (50) yield

$$(\forall k \in \mathbb{N}) \quad \mathsf{R}\mathsf{q}_k = \left(q_{s,k}/2, Q q_{b,k} \right), \tag{53}$$

and it follows from (26), (28), (37), (38), and (53) that (51) is precisely (29). Altogether, since $(\boldsymbol{u}_k)_{k \in \mathbb{N}} = (\mathsf{w}_k, \mathsf{z}_k)_{k \in \mathbb{N}}$ converges to a solution to (46), $(\mathsf{w}_k)_{k \in \mathbb{N}} = (s_k, b_k)_{k \in \mathbb{N}}$ converges to a solution to Problem 1.

Appendix B: Proof of Proposition 2

- If $s \notin [0, +\infty[^N$, then (11) yields $(\forall b \in \mathbb{R}^p)$ $g(s, Ab) = +\infty$. On the other hand if, for some $i \in \{1, \dots, N\}$, $\sigma_i \in]0, +\infty[$ then we deduce from (i) that $(\forall b \in \mathbb{R}^p)$ $\tilde{\varphi}_i(\sigma_i, X_i b - y_i) = \sigma_i \theta_i((X_i b - y_i)/\sigma_i) + \alpha_i \sigma_i \geqslant \alpha_i \sigma_i \to +\infty$ as $\sigma_i \to +\infty$. Hence, (ii) entails that $(\forall b \in \mathbb{R}^p)$ $g(s, Ab) \to +\infty$ as $\|s\|_2 \to +\infty$ while $s \in [0, +\infty[^N$. On the other hand, it follows from (iii) that $(\forall s \in \mathbb{R}^N)(\forall b \in E)$ $g(s, Ab) \geqslant \psi_j(L_j b) \to +\infty$ as $\|b\|_2 \to +\infty$. Altogether, (30) holds.
- It follows from (i) and (11) that $(\forall i \in \{1, \dots, N\})$ ri dom $\tilde{\varphi}_i =]0, +\infty[\times \mathbb{R}^{n_i}$. Furthermore, (ii) yields $(\forall i \in \{1, \dots, M\})$ ri dom $\psi_i = \mathbb{R}^{m_i}$. Therefore ri dom $g =]0, +\infty[^N \times \mathbb{R}^n \times \mathbb{R}^m$. Since trivially $A(E) \subset \mathbb{R}^{n+m}$, (31) reduces to (iv).

Appendix C: Numerical Algorithm Comparison on BMI Data

We compare the numerical accuracy and the run time of the Algorithm 3 with the coordinate descent algorithm proposed in [30] for the special case of the constrained Lasso problem with joint scale estimation, defined in (33). We use a subset of the BMI dataset with $n = 96$ samples and $p = 45$ OTUs. The numerical example is reproduced with the following MATLAB script available at https://github.com/muellsen/PCM/tree/master/examples/LogContrastModels.

```
>>runtimeBMI
% Comparison of the runtime of exact proximal scheme
% and the coordinate descent (with interval search) method
% (Shi et al. 2016) on the BMI example without covariates

% Load all data in MATLAB format (located in
% /PerspectiveFunctions/misc/ConstrLasso/)
load allData.mat;
...
% Solve the model using the product space Douglas−Rachford
method
t1=now;
[betaPCMMat, sigmaPCMMat,funPCMMat,outPCM] = ...
pcmC2(x_cent, y_cent, pcmopts);
t2=now;
timePCM = (t2−t1)∗(60∗60∗24)
...
% Solve the model using the method of Shi et al. (2016)
t1=now;
[betaConLMat, sigmaConLMat,outConL] = ...
concomlasso(x_cent, y_cent, concomopts);
t2=now;
timeConcom = (t2−t1)∗(60∗60∗24);
...
```

On a MacBook Pro (2018) with 2.9 GHz Intel Core i9 processor and 32 GB 2400 MHz DDR4, the run time of the general Douglas–Rachford scheme (with fixed $\gamma = 0.5$) for solving the problem across the entire regularization path is 12 s at solution accuracy $\epsilon = 1e - 8$. Further run time improvement could be achieved by setting γ in a regularization-dependent fashion. The coordinate descent scheme requires 10.5 s. Since no convergence guarantees comparable to ours exist for the coordinate descent scheme, we also compare the solution quality across the regularization path. The results are summarized in Fig. 7. We observe that, for this example, both methods agree on the solution within six digits of accuracy, both for the regression vectors ($\Delta\beta$) and the scales ($\Delta\sigma$) across the entire regularization path.
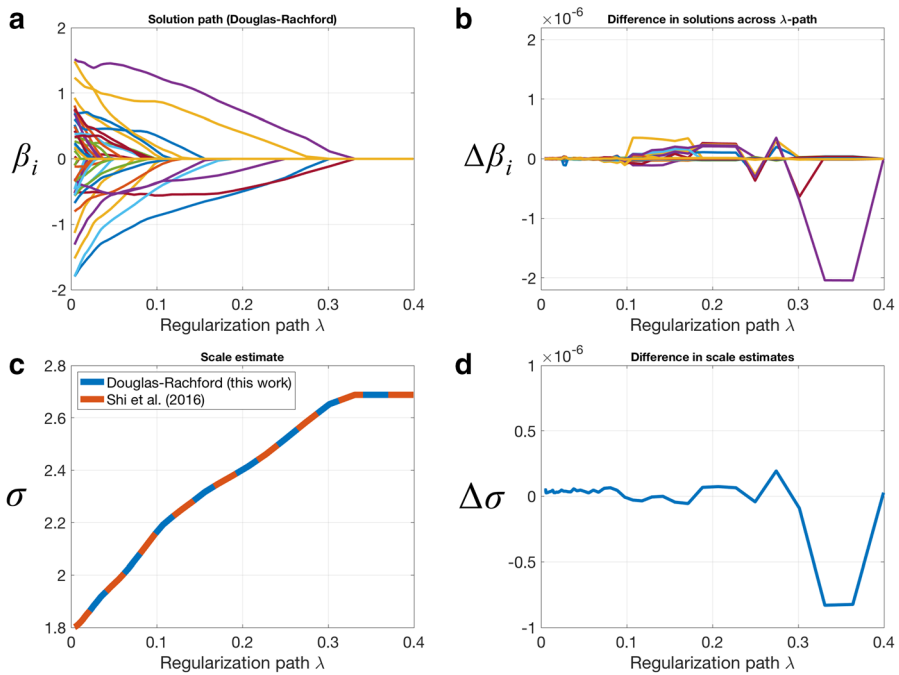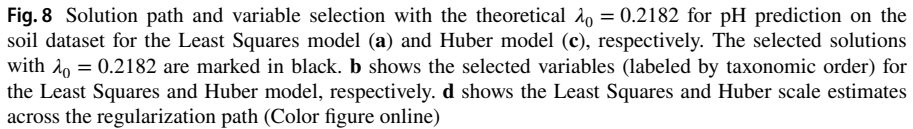
Fig. 7 Numerical comparison of Algorithm 3 and the coordinate descent (CD) scheme of [30] for the constrained Lasso problem with joint scale estimation, defined in (33). **a** Shows the solution path for the regression vector, found by Algorithm 3. **b** Shows the estimation difference between Algorithm 3 and CD. **c** shows the scale estimates for both methods, and **d** the difference between the respective estimates. The run times for both methods are 12 s and 10.5 s, respectively (Color figure online)

Appendix D: Solution Path and Variable Selection with Theoretical λ_0 Regularization for the Soil Dataset

We here complement the analysis of the soil dataset [18], analyzed in Sect. 4.2, comprising $n = 88$ soil samples and $p = 116$ OTUs. We show the full solution path and variable selection with the theoretical $\lambda_0 = 0.2182$ in Fig. 8, both for the Least Squares (Fig. 8a) and the Huber model (Fig. 8c). Using the theoretical λ_0, both models select 27 and 25 variables, respectively. The joint set of 30 variables is shown in Fig. 8b. For completeness, we also report the scale estimates for both models (Fig. 8d).

Fig. 8 Solution path and variable selection with the theoretical $\lambda_0 = 0.2182$ for pH prediction on the soil dataset for the Least Squares model (**a**) and Huber model (**c**), respectively. The selected solutions with $\lambda_0 = 0.2182$ are marked in black. **b** shows the selected variables (labeled by taxonomic order) for the Least Squares and Huber model, respectively. **d** shows the Least Squares and Huber scale estimates across the regularization path (Color figure online)

References

1. Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London
2. Aitchison J, Bacon-Shone J (1984) Log contrast models for experiments with mixtures. Biometrika 71:323–330
3. Antoniadis A (2007) Wavelet methods in statistics: some recent developments and their applications. Stat Surv 1:16–55. https://doi.org/10.1214/07-SS014
4. Bauschke HH, Combettes PL (2017) Convex analysis and monotone operator theory in Hilbert spaces, 2nd edn. Springer, New York
5. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: high-resolution sample inference from illumina amplicon data. Nat Methods 13:581–583
6. Caporaso JG et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336. https://doi.org/10.1038/nmeth0510-335
7. Combettes PL (2004) Solving monotone inclusions via compositions of nonexpansive averaged operators. Optimization 53:475–504
8. Combettes PL (2018) Perspective functions: properties, constructions, and examples. Set-Valued Var. Anal. 26:247–264
9. Combettes PL, Müller CL (2018) Perspective functions: proximal calculus and applications in high-dimensional statistics. J Math Anal Appl 457:1283–1306
10. Combettes PL, Müller CL (2020) Perspective maximum likelihood-type estimation via proximal decomposition. Electron J Stat 14:207–238

11. Combettes PL, Pesquet J-C (2007) A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery. IEEE J Select Topics Signal Process 1:564–574

12. Eckstein J, Bertsekas DP (1992) On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math Progr 55:293–318

13. Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 10:996–998

14. Hron K, Filzmoser P, Thompson K (2012) Linear regression with compositional explanatory variables. J Appl Stat 39:1115–1128

15. Huber P (1964) Robust estimation of a location parameter. Ann Math Stat 35:73–101

16. Javanmard A, Montanari A (2014) Confidence intervals and hypothesis testing for high-dimensional regression. J Mach Learn Res 15:2869–2909

17. Lagkouvardos I, Fischer S, Kumar N, Clavel T (2017) Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. PeerJ 5:e2836

18. Lauber CL, Hamady M, Knight R, Fierer N (2009) Pyrosequencing-based assessment of a soil pH as a predictor of soil bacterial community structure at the continental scale. Appl Environ Microbiol 75:5111–5120

19. Lenoir A, Mahey P (2017) A survey on operator splitting and decomposition of convex programs. RAIRO-Oper Res 51:17–41

20. Lin W, Shi P, Feng R, Li H (2014) Variable selection in regression with compositional covariates. Biometrika 101:785–797

21. Lions P-L, Mercier B (1979) Splitting algorithms for the sum of two nonlinear operators. SIAM J Numer Anal 16:964–979

22. Meinshausen N, Bühlmann P (2010) Stability selection. J R Stat Soc Ser B Stat Methodol 72:417–473

23. Mishra A, Müller CL (2019) Robust regression with compositional covariates. arxiv. https://arxiv.org/abs/1909.04990

24. Morton JT et al (2017) Balance trees reveal microbial niche differentiation. mSystems 2:e00162

25. Nguyen NH, Tran TD (2013) Robust lasso with missing and grossly corrupted observations. IEEE Trans Inform Theory 59:2036–2058

26. Owen AB (2007) A robust hybrid of lasso and ridge regression. Contemp Math 443:59–71

27. Rockafellar RT (1970) Convex analysis. Princeton University Press, Princeton

28. Schloss PD et al (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541

29. She Y, Owen AB (2011) Outlier detection using nonconvex penalized regression. J Am Stat Assoc 106:626–639

30. Shi P, Zhang A, Li H (2016) Regression analysis for microbiome compositional data. Ann Appl Stat 10:1019–1040

31. Sun T, Zhang CH, Campus B (2012) Scaled sparse linear regression. Biometrika 99:879–898

32. Sun Z, Xu W, Cong X, Chen K (2018) Log-contrast regression with functional compositional predictors: linking preterm infant's gut microbiome trajectories in early postnatal period to neurobehavioral outcome. http://arxiv.org/abs/1808.02403

33. Wang T, Zhao H (2017) Structured subcomposition selection in regression and its application to microbiome data analysis. Ann Appl Stat 11:771–791

34. Wu GD et al (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science 334:105–108