The Annals of Applied Statistics
2018, Vol. 12, No. 1, 540–566
https://doi.org/10.1214/17-AOAS1102

KERNEL-PENALIZED REGRESSION FOR ANALYSIS OF MICROBIOME DATA

BY TIMOTHY W. RANDOLPH[*,1,2], SEN ZHAO[†,3], WADE COPELAND[*], MEREDITH HULLAR[*,1] AND ALI SHOJAIE[†,2,3]

Fred Hutchinson Cancer Research Center and University of Washington†*

The analysis of human microbiome data is often based on dimension-reduced graphical displays and clusterings derived from vectors of microbial abundances in each sample. Common to these ordination methods is the use of biologically motivated definitions of similarity. Principal coordinate analysis, in particular, is often performed using ecologically defined distances, allowing analyses to incorporate context-dependent, non-Euclidean structure. In this paper, we go beyond dimension-reduced ordination methods and describe a framework of high-dimensional regression models that extends these distance-based methods. In particular, we use kernel-based methods to show how to incorporate a variety of extrinsic information, such as phylogeny, into penalized regression models that estimate taxon-specific associations with a phenotype or clinical outcome. Further, we show how this regression framework can be used to address the compositional nature of multivariate predictors comprised of relative abundances; that is, vectors whose entries sum to a constant. We illustrate this approach with several simulations using data from two recent studies on gut and vaginal microbiomes. We conclude with an application to our own data, where we also incorporate a significance test for the estimated coefficients that represent associations between microbial abundance and a percent fat.

1. Introduction. A common tool in the analysis of data from microbiome studies is a scatterplot of dimension-reduced microbial abundance vectors. This is a display of the samples' beta diversity which, in ecology, refers to differences among various habitats. When applied to human studies, beta diversity describes the variation in microbial community structure across sampling units (e.g., human subjects): a beta diversity plot displays the n sampling units with respect to the principal coordinates of their microbial abundance vectors, each consisting of measures on the p taxa (bacterial types) observed in the study; see, for example, Claesson et al. (2012), Goodrich et al. (2014), Koren et al. (2013), Kuczynski et al. (2010). This principal coordinates analysis (PCoA; or multidimensional scaling,

Received January 2017; revised May 2017.
[1]Supported by National Institutes of Health Grants P01-CA168530, U01-CA162077.
[2]Supported by National Institutes of Health Grant R01-GM114029.
[3]Supported by National Science Foundation Grants DMS-1161565 and DMS-1561814.
Key words and phrases. Compositional data, distance-based analysis, kernel methods, microbial community data, penalized regression.

MDS) begins with an $n \times n$ matrix of pairwise dissimilarities between vectors of taxon abundances. The choice of dissimilarity measure may greatly influence the biological interpretation [Lozupone et al. (2007), Fukuyama et al. (2012)]. Euclidean distance is rarely used.

A common assay of microbial content is based on counting sequences observed from the 16S rRNA gene, a marker used to identify bacterial species or other taxonomic categories. We will generically refer to "taxa" rather than specifying a category, such as genus or species. A single taxon may be placed within the context of a phylogenetic tree in order to provide evolutionary relationships among taxa. Dissimilarity measures that account for these phylogenetic relationships are assumed to enhance statistical analyses—for instance, to improve the power of statistical tests—because they incorporate the degree of divergence between sequences [Chen et al. (2012)] and do not ignore "the correlation between evolutionary and ecological similarity" [Hamady and Knight (2009)]. The UniFrac distance [Lozupone and Knight (2005)], in particular, is based on the premise that taxa which share a large fraction of the phylogenetic tree should be viewed as more similar than those sharing a small fraction of the tree. In the unweighted version of UniFrac, each taxon is quantified merely by its presence or absence; the distance between a pair of samples is based on the number of branches in the tree shared by both. Figure 1(a) is a beta diversity plot of $n = 100$ human microbial abundance vectors with $p = 149$ taxa based on data from Yatsunenko et al. (2012). Each sample is represented by 2-dimensional coordinates with respect to the unweighted UniFrac distance, and the size of each point is proportional to log(age) of the subject.

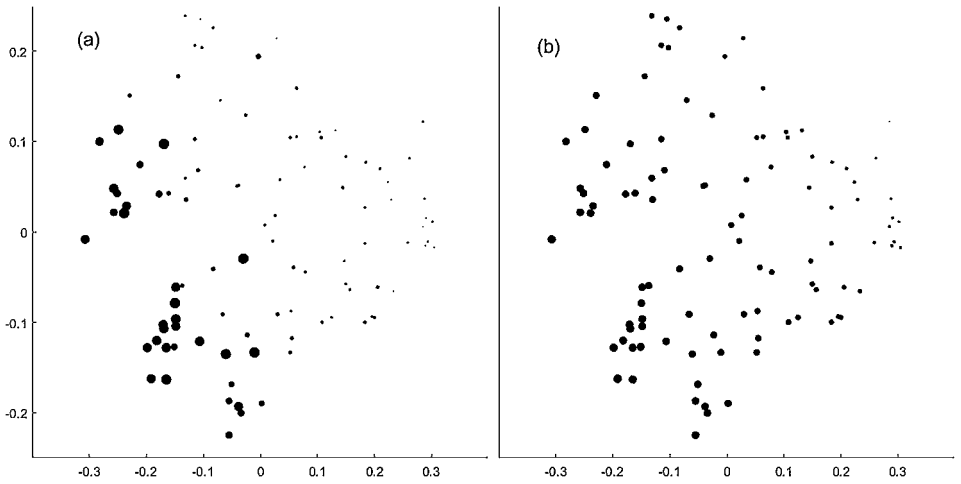


FIG. 1. *PCoA plots of data from Yatsunenko et al. (2012). (a): PCoA plot with respect to unweighted UniFrac distance where dot size is proportional to* log(*age*) *of the subject. (b): PCoA plot with respect to unweighted UniFrac distance, dot size is proportional to* y_{True} *from the model in equation (3.2) with* $\varepsilon = 0$.

Dissimilarity measures in microbiome studies are many and varied, with a rich collection that, like UniFrac, exploit the phylogentic structure: Chen et al. (2012) generalize UniFrac by reweighting rare and abundant lineages; double principal coordinate analysis (DPCoA) [Pavoine, Dufour and Chessel (2004)], as shown by Purdom (2011), generalizes PCA by incorporating the covariance that would arise if the data was created by a process modeled by the tree; the *edge PCA* method of Matsen and Evans (2013) incorporates taxon abundance information at all nodes in a phylogenetic tree, rather than just the leaves of the tree, and Evans and Matsen (2012) formalize the mathematical interpretation of UniFrac as just one example within a large family of Wasserstein (or earth mover's) metrics. A wide variety of non-phylogenetic dissimilarities are also in common use, such as Bray–Curtis [The Human Microbiome Project Consortium (2012)] and Jenson–Shannon [Koren et al. (2013)], among others.

While PCoA plots provide valuable graphical insight into the relationships among microbial profiles and an outcome or phenotype, they do not *quantify* this association. More importantly, the (sets of) taxa associated with the outcome—and the magnitude or statistical significance of such associations—are not ascertained from a PCoA plot; once a matrix of (dis)similarities between samples is formed, it is not clear how to identify individual taxa that are associated with an outcome. Specifically, given a PCoA plot as in Figure 1(a), with structure imposed by the chosen dissimilarity matrix (e.g., unweighted UniFrac) and with associations implied by a class label or continuous outcome (e.g., age), how does one estimate which taxa or subcommunities are associated with this outcome? We address this question by formulating multivariate regression models that are constrained by the structure of the (dis)similarity matrix. This is made possible by exploiting an equivalence between a taxon-based (primal space) and sample-based (dual space) formulation of our penalized regression models. While exploiting such an equivalence is straightforward in the special case of ridge regression (with purely Euclidean structure), it becomes complicated when more general distance measures are used. To this end, we show how a little-used regularization scheme by Franklin (1978) provides a dual-space regression coefficient estimate that naturally connects to primal-space coefficients. Because a dissimilarity matrix can be used to construct a similarity matrix (as commonly done in classical MDS [Mardia, Kent and Bibby (1980)], we work with kernels, rather than distances, and allow for general kernels, including those constructed from a nonlinear feature map.

In addition to complications stemming from more general distances, the analysis of microbiome data is also complicated by the compositional nature of the data itself. More specifically, taxon measures typically represent *relative*, rather than absolute, abundances. The p-variate relative abundance vectors are thus *compositional* in that they are constrained to a simplex within $\mathbb{R}^p$; such data do not reside in a Euclidean vector space [Aitchison (2003a)]. Consequently, spurious correlations arise and standard multiple regression models fail. Our proposed KPR

framework, however, addresses this: the centered log (CLR) transform of the relative abundance vectors first removes the vectors from the simplex, then the estimation process is constrained using a penalization term defined by Aitchison's variation matrix. This approach takes a different perspective from the recent proposal of Li (2015) which forces the estimated coefficient vector to reside in the simplex. Given that the CLR transforms the compositional vectors to Euclidean space and that the units of the Aitchison variation matrix are the same as the CLR transformed data [Egozcue and Pawlowsky-Glahn (2011)], our constraint seems more suitable for the geometry of the problem.

In summary, we describe a family of high-dimensional regression problems in Section 2, which are designed to incorporate the assumptions that are tacitly implied by various exploratory and graphically-focused PCoA plots common in microbiome studies. We show how phylogenetic and other structure can be incorporated via kernel penalized regression in either the primal (p-dimensional) feature space or the dual (n-dimensional) samples space; see Sections 2.2 and 2.3. Finally, our proposed framework leads to an approach, described in Section 2.4, for addressing well-known problems that arise from applying standard (Euclidean-based) statistical models to compositional data. Section 3 illustrates the proposed framework with simulations based on publicly available data, while Section 4 presents an application to our recent microbiome study of premenopausal women. In this analysis, we obtain estimates of associations between microbial species and percent fat measured in premenopausal women, and also provide inference for these estimates by applying a recent significance test [Zhao and Shojaie (2016)] in our kernel-penalized regression (KPR) framework.

2. Kernel penalized regression for microbiome data. We describe a family of multiple regression problems aimed at incorporating assumptions that are implicit in principal coordinate analysis (PCoA) plots common in microbiome studies. We begin in Section 2.1 by establishing notation and concepts from existing dimension-reduction (ordination) methods with the goal of extending them to non-truncated (penalized) regression models. Section 2.2 extends PCoA and principal component regression (PCR) to penalized regression models in the primal space in a manner that incorporates structures implicit in recent microbiome analyses. Section 2.3 extends kernel ridge regression to general (non-L^2) structure and the use of two kernels. This extension exploits a *dual-space* regularization scheme of Franklin [Franklin (1978)]. Section 2.4 describes how our proposed framework can be applied to formulate a penalized regression model that accounts for the compositional nature of relative abundance data.

We denote by y_i, $i = 1, \ldots, n$, a real-valued quantified trait and by $x_i = [x_{i1}, \ldots, x_{ip}]'$ a p-dimensional vector of microbial abundance values measured for each of n subjects. Denote by X the $n \times p$ sample-by-taxon matrix whose ith row is x_i'. We assume throughout that the columns of X are mean centered. For now, we assume that the abundance values are appropriately normalized/transformed

and postpone the treatment of compositional data to Section 2.4. The transpose of a matrix A is denoted by A' and the Frobenius norm is denoted as $\|A\|_F$. $I \equiv I_p$ denotes the identity matrix on $\mathbb{R}^p$ and the Euclidean norm of a vector $x \in \mathbb{R}^p$ is denoted $\|x\|_{\mathbb{R}^p}$ or simply $\|x\|$.

2.1. *Background for PCoA and principal component regression.* Consider first the Euclidean PCoA, which is obtained from the eigenvectors of the *kernel* matrix $K_I := XX'$ of inner products $K_{ij} = \langle x_i, x_j \rangle$ between samples. Let $\mathcal{J}$ be the centering matrix, $\mathcal{J} = I - \frac{1}{n}\mathbf{1}\mathbf{1}'$, where $\mathbf{1}$ is the $n \times 1$ vector of ones. Then it can be seen that $XX' = -\frac{1}{2}\mathcal{J}\Delta^E\mathcal{J}$, where Δ^E is the $n \times n$ matrix of *squared* Euclidean distances between samples: $\Delta^E_{i,j} = \|x_i - x_j\|^2_{\mathbb{R}^p}$. The relationship between a kernel and a distance matrix Δ is more general. In particular, if Δ is any $n \times n$ symmetric matrix of squared dissimilarities between vectors in $\mathbb{R}^p$ then $H = -\frac{1}{2}\mathcal{J}\Delta\mathcal{J}$ serves as a kernel matrix summarizing similarities; see, for example, Gower (1966), Pekalska, Paclik and Duin (2002). A particular case involves a $p \times p$ symmetric, positive definite matrix Q that defines an inner product $\langle x_i, x_j \rangle_Q = x_i' Q x_j$ on $\mathbb{R}^p$. If Δ^Q denotes the matrix of squared distances, $\Delta^Q_{i,j} = \|x_i - x_j\|^2_Q = \langle x_i - x_j, x_i - x_j \rangle_Q$, defined with respect to this inner product, then $XQX' = -\frac{1}{2}\mathcal{J}\Delta^Q\mathcal{J}$ is also a similarity kernel for the n samples. We will denote this kernel by $K_Q = XQX'$. Similarly, one may start with a matrix Δ^U of squared distances defined by a tree-based UniFrac dissimilarity [Lozupone and Knight (2005)], and define a similarity kernel by $H = -\frac{1}{2}\mathcal{J}\Delta^U\mathcal{J}$.

In graphical displays, two or three coordinates are typically used to explore the relationship between samples. Let $K = US^2U'$ be the eigen-decomposition of any similarity kernel, K, where U is the matrix whose columns are eigenvectors and $S^2 = \operatorname{diag}\{\sigma_j^2\}$ is the diagonal matrix of eigenvalues. The two-dimensional PCoA plot is then the collection of points $\{\eta_{i1}, \eta_{i2}\}_{i=1}^n := \{(\sigma_1 U_{i1}, \sigma_2 U_{i2})\}_{i=1}^n$; that is, a plot of the points represented by the first two columns of the matrix US. These points are often colored according to a grouping label or continuous values, $\{y_i\}_{i=1}^n$, to graphically explore the existence of an association between the outcome y and the sample profiles summarized by the first few columns of US. So, a PCoA plot may be viewed as a graphical depiction of a two-component regression model of association:

$$y_i = \gamma_1\eta_{i1} + \gamma_2\eta_{i2} + \varepsilon, \qquad i = 1, \ldots, n, \tag{2.1}$$

where η_1 and η_2 are the first two PCoA axes. Ordinary principal component regression corresponds to the case that η_1 and η_2 come from the Euclidean kernel $K_I = XX'$. On the other hand, the configuration of points in Figure 1(b) correspond to the first two eigenvectors of the kernel defined by an unweighted UniFrac distance matrix Δ^U, and the size of individual points correspond to the values of y from equation (2.1) with $\varepsilon = 0$.

Let $A_{(k)}$ denote the first k columns of a matrix A, or its first k rows and columns if A is diagonal. Then, using the singular value decomposition (SVD) $X = USV'$, if we express the dimension-reduced approximation of X as $\breve{X} := U_{(2)}S_{(2)}V'_{(2)}$, then equation (2.1) can be written as

$$\begin{aligned} y &= \gamma_1\eta_1 + \gamma_2\eta_2 + \varepsilon \\ &= U_{(2)}S_{(2)}\gamma + \varepsilon \\ &= \breve{X}V_{(2)}\gamma + \varepsilon, \end{aligned} \tag{2.2}$$

where $\gamma = [\gamma_1\ \gamma_2]'$. Here, $\breve{X}V_{(2)} = U_{(2)}S_{(2)}$ and $\mathrm{Range}(\breve{X}') = \mathrm{Range}(V_{(2)})$. Consequently, the model $y = \breve{X}V_{(2)}\gamma + \varepsilon$ can be written as $y = \breve{X}\beta + \varepsilon$, where β is some vector of the form $\beta = \breve{X}'\gamma$. So inherent in a Euclidean PCoA plot is an implicit coefficient vector, β, which models a linear association between y and $\breve{X}$. Using the SVD of X in (2.2), the PCR estimate of $\beta \in \mathbb{R}^p$ is expressed as

$$\hat{\beta}_{\mathrm{PCR}} = \left(\breve{X}'\breve{X}\right)^{\dagger}\breve{X}'y = V_{(2)}S_{(2)}^{-1}U'_{(2)}y = \sum_{k=1}^{2}\frac{1}{\sigma_k}u_k'y v_k, \tag{2.3}$$

where † denotes the Moore–Penrose inverse.

2.2. *Penalized regression and DPCoA.* An alternative to a Euclidean PCR is the ordinary ridge regression [Hoerl and Kennard (1970)],

$$\hat{\beta}_{\mathrm{ridge}} = \left(X'X + \lambda I\right)^{-1}X'y = \sum_{k=1}^{n}\left(\frac{\sigma_k^2}{\sigma_k^2 + \lambda^2}\right)\frac{1}{\sigma_k}u_k'y v_k, \tag{2.4}$$

in which the terms are reweighted, instead of being truncated as in $\hat{\beta}_{\mathrm{PCR}}$. The estimate in (2.4) is the solution of the penalized least squares regression problem, $\hat{\beta}_{\mathrm{ridge}} = \arg\min_{\beta}\{\|y - X\beta\|^2 + \lambda\|\beta\|^2\}$, where here and throughout λ is a tuning parameter that controls the amount of shrinkage or size of β in the penalty term. Here, the penalty is simply the Euclidean (or ℓ^2) norm on $\mathbb{R}^p$, but a wide range of penalty terms have been proposed to replace or extend this particular form of regularization; see Bühlmann, Kalisch and Meier (2014) for a review of the most established methods. These methods, such as the lasso, elastic net or SCAD do not incorporate any extrinsic information, but a variety of other penalization methods have been proposed which aim to do this. For instance, Tanaseichuk, Borneman and Jiang (2014) uses a tree-guided penalty [Kim and Xing (2010)] to incorporate such structure into a penalized logistic regression framework to encourage similar coefficients among taxa according to their relationships in the phylogenetic tree. Tibshirani and Taylor (2011) study the solution path for computing a "generalized lasso" estimate in which an ℓ^2 penalty is replaced with an ℓ^1 penalty applied to a linear transformation of the features, $\lambda\|L\beta\|_1$. Within the context of genetic networks, Li and Li (2008) accounted for network structure by augmenting the ℓ^1

penalty with a second penalty of the form $\lambda_2\|\beta\|_{\mathcal{L}}^2 = \beta'\mathcal{L}\beta$, where $\mathcal{L}$ denotes the graph Laplacian matrix corresponding to predefined connections between genes in a pathway.

For now, we consider a positive definite $p \times p$ matrix Q with a Cholesky decomposition $Q = LL'$, and a penalty term of the form $\|L^{-1}\beta\|^2 = \|\beta\|_{Q^{-1}}^2 = \beta' Q^{-1}\beta$. The generalized ridge (or Tikhonov regularization [Golub and van Loan (2012)]) estimate with respect to Q is then defined as

$$\begin{aligned}\hat{\beta}_Q &= \arg\min_{\beta}\{\|y - X\beta\|^2 + \lambda\|\beta\|_{Q^{-1}}^2\} = \left(X'X + \lambda Q^{-1}\right)^{-1} X'y \\ &= \sum_{k=1}^{n}\left(\frac{\sigma_k^2}{\sigma_k^2 + \lambda\mu_k^2}\right)\frac{1}{\sigma_k} u_k' y v_k.\end{aligned} \tag{2.5}$$

This estimate takes the same form as (2.4) but now the vectors u_k and v_k arise from the SVD of $XL = USV'$. Regarding the last equal sign, note that if A denotes any matrix with p columns, the structure of an estimate $\hat{\beta}_A$ from a penalty term of the form $\|A\beta\|_2^2$ is determined by the joint eigenstructure of the pair (X, A) via the generalized singular value decomposition.[4] In particular, the basis expansion of $\hat{\beta}_Q$ in (2.5) is given in terms of the generalized singular vectors of (X, L^{-1}). Although the ridge estimate (with $Q = I_p$) is biased, an informed choice of penalty term can reduce the bias [Randolph, Harezlak and Feng (2012)].

Now consider the context of phylogentic information and let δ represent the matrix of squared patristic distances between pairs of taxa, that is, the sum of branch lengths between each pair of taxa on the leaves of a phylogenetic tree. Set $Q = -\frac{1}{2}\mathcal{J}\delta\mathcal{J}$, a matrix of similarities between taxa. Double principal coordinate analysis (DPCoA) was proposed by Pavoine, Dufour and Chessel (2004) to provide an alternative to ordinary PCoA that incorporates structure among samples as well as structure implied by the taxa's distribution among subcommunities, as summarized by Q. Purdom (2011) clarified the original multistep DPCoA procedure and showed how it can be more simply understood as a generalized PCA (gPCA) in which one obtains the new coordinates from the eigenvectors of $K_Q = XQX'$. Note that when $Q = I_p$, DPCoA reduces to PCA/MDS. As emphasized in Purdom (2011), the use of a nonidentity Q matrix incorporates structure from known relationships between the p taxa by exploiting a matrix representation of phylogenetic relationships, thus providing a model for covariance structure.

If we let $Q = LL'$ be a Cholesky decomposition of Q and set $Z := XL$, then the kernel $K_Q = XQX'$ has an eigendecomposition of the form US^2U' with respect to the SVD of $Z = USV'$. This leads to a two-dimensional regression estimate that

[4]We refer here to the generalized singular value decomposition (GSVD) of Van Loan (1976), a simultaneous diagonlization of two matrices. A different SVD generalization [Greenacre (1984)] imposes constraints on left and right singular vectors of a matrix.

takes the same the form as $\hat{\beta}_{\mathrm{PCR}}$ in (2.3). Indeed, we can recover a primal space estimate in terms of singular vectors as

$$\hat{\beta}_{\mathrm{DPCR}} := V_{(2)} S_{(2)}^{-1} U_{(2)}' y = \sum_{k=1}^{2} \frac{1}{\sigma_k} u_k' y v_k. \tag{2.6}$$

That is, implicit in a DPCoA plot is a coefficient vector $\hat{\beta}_{\mathrm{DPCR}}$ which models a two-dimensional linear association between y and $Z = XL$ in the same way that $\hat{\beta}_{\mathrm{PCR}}$ represents a two-dimensional linear association between y and X. Further, $XQX' = (XL)(XL)'$ and so U, S and V in $\hat{\beta}_{\mathrm{DPCR}}$ of (2.6) are the same as those in the penalized (nontruncated) estimate, $\hat{\beta}_Q$, in (2.5). When $Q = I$, these two estimates reduce to $\hat{\beta}_{\mathrm{PCR}}$ and $\hat{\beta}_{\mathrm{ridge}}$, respectively.

2.3. *Kernel-based regression with two kernels.* In addition to similarities among taxa, as in Q, it is often of interest to incorporate similarities among samples as derived, for instance, from UniFrac distances: $H = -\frac{1}{2}\mathcal{J}\Delta^U\mathcal{J}$. The symmetric positive definite $n \times n$ kernel H defines a new inner product on $\mathbb{R}^n$ given by $\langle u, w\rangle_H = u'Hw$, with the corresponding norm $\|u\|_H^2 = \langle u, u\rangle_H$. If we consider both a general kernel, H, and a DPCoA kernel $K_Q = XQX'$, the generalized ridge estimate $\hat{\beta}_Q$ in (2.5) can be extended to

$$\begin{aligned} \hat{\beta}_{Q,H} &:= \arg\min_{\beta \in \mathbb{R}^p} \{\|y - X\beta\|_H^2 + \lambda \|\beta\|_{Q^{-1}}^2\} \\ &= (X'HX + \lambda Q^{-1})^{-1} X'Hy. \end{aligned} \tag{2.7}$$

In this section, we show that the estimate in (2.7) is directly defined based on the generalized eigenvectors of the two kernels K_Q and H. Before proceeding to the general case, let us examine the special case of ridge regression for which $H = I_n$ and $Q = I_p$. A ridge estimate can be obtained by solving an equivalent optimization problem in the dual space $\mathbb{R}^n$, known as *kernel ridge regression* [Schölkopf and Smola (2002)]. Specifically, taking $K_I = XX'$, the ridge estimate in (2.4) can be obtained as $\hat{\beta}_{\mathrm{ridge}} = X'\hat{\gamma}_{\mathrm{kernel\ ridge}}$, where

$$\begin{aligned} \hat{\gamma}_{\mathrm{kernel\ ridge}} &= (K_I + \lambda I)^{-1} y = (K_I^2 + \lambda K_I)^{-1} K_I y \\ &= \arg\min_{\gamma \in \mathbb{R}^n} \{\|y - K_I\gamma\|^2 + \lambda \|\gamma\|_{K_I}^2\}. \end{aligned} \tag{2.8}$$

In the case of ridge, the connection between the dual- and primal-space estimates, $\hat{\gamma}_{\mathrm{kernel\ ridge}}$ and $\hat{\beta}_{\mathrm{ridge}}$, relies on the form $K_I = XX'$. Unfortunately, it is less clear how to extend this connection to a general kernel (e.g., UniFrac or polynomial). One way to incorporate a more general kernel K and a second kernel H in (2.8) is to define the penalty in terms of H as

$$\hat{\gamma}_* = (K^2 + \lambda H^{-1})^{-1} Ky = \arg\min_{\gamma} \{\|y - K\gamma\|^2 + \lambda \|\gamma\|_{H^{-1}}^2\}, \tag{2.9}$$

which is exactly Tikhonov regularization, but in the dual space; compare equation (2.5). However, $\hat{\gamma}_* \in \mathbb{R}^n$ has no obvious connection to a penalized estimate of $\beta \in \mathbb{R}^p$ and cannot be used to obtain a penalized regression estimate in the primal space, even if $K = K_I = XX'$.

To bridge this gap, we instead apply the Franklin regularization scheme [Franklin (1978)], a little-used alternative to Tikhonov regularization. More specifically, for *any* kernels K and H, we *define* the dual estimate

$$\hat{\gamma}_{H,K} := \big(K + \lambda H^{-1}\big)^{-1} y = \underset{\gamma \in \mathbb{R}^n}{\arg\min} \big\{ \|y - K\gamma\|_{K^{-1}}^2 + \lambda \|\gamma\|_{H^{-1}}^2 \big\}, \tag{2.10}$$

where the justification for the second equality is given in the supplementary material [Randolph et al. (2018)].

Comparing (2.9) and (2.10), one sees that the analytic form of (2.10) involves just K rather than $K^2 = K'K$. As shown in Proposition 2.2, this subtle difference is a key for relating a dual-space estimate $\hat{\gamma}_{H,K_Q}$ and its primal-space counterpart, $\hat{\beta}_{Q,H} = QX'\hat{\gamma}_{H,K_Q}$. Before presenting the main result of this section, we provide several equivalent forms of $\hat{\gamma}_{H,K}$

$$\begin{aligned}
\hat{\gamma}_{H,K} &= \big(K + \lambda H^{-1}\big)^{-1} y \\
&= \underset{\gamma \in \mathbb{R}^n}{\arg\min} \big\{ \|y - K\gamma\|_{K^{-1}}^2 + \lambda \|\gamma\|_{H^{-1}}^2 \big\} \\
&= \underset{\gamma \in \mathbb{R}^n}{\arg\min} \big\{ \|y - K\gamma\|_{H}^2 + \lambda \|\gamma\|_{K}^2 \big\} \\
&= (HK + \lambda I)^{-1} H y \\
&= H(KH + \lambda I)^{-1} y.
\end{aligned} \tag{2.11}$$

In Proposition 2.2, we also refer to the special case corresponding to the DPCoA ordination. As before, let $Z = XL$ so that $K_Q = XQX' = XLL'X' = ZZ'$. Taking $H = I$, the dual-space estimate in (2.10) is $\hat{\gamma}_{I,K_Q} = (K_Q + \lambda I)^{-1} y$, and so the corresponding primal-space estimate is $\hat{\beta} \equiv Z'\hat{\gamma}_{I,K_Q}$. Since this estimate arises from the DPCoA kernel, we make the following definition.

DEFINITION 2.1. A primal space DPCoA estimate is of the form $\hat{\beta}_{\mathrm{DPCoA}} = Z'\hat{\gamma}_{I,K_Q} = L'X'(XQX' + \lambda I)^{-1} y$.

The next proposition collects several properties that emphasize the roles of H and K in our penalized regression framework. In particular, we show that the primal space estimate $\hat{\beta}_{Q,H}$ can be recovered in terms of two kernels, H and K_Q.

PROPOSITION 2.2. *Let H and K be any two kernels constructed using the rows of X in the regression model $y = X\beta + \varepsilon$. Then*:

1. *$\hat{\gamma}_{H,K}$ is a linear combination of the eigenvectors of the matrix product HK.*
2. *For any kernel H and DPCoA kernel $K_Q = XQX'$, then the primal- and dual-space estimates in (2.7) and (2.10), respectively, are related as: $\hat{\beta}_{Q,H} = QX'\hat{\gamma}_{H,K_Q}$.*
3. *For $H = I$ and $Q = LL'$, the generalized ridge and DPCoA estimates are related as $\hat{\beta}_Q = QX'(K_Q + \lambda I_n)^{-1}y = L\hat{\beta}_{\mathrm{DPCoA}}$.*

The proof, given in the supplementary material [Randolph et al. (2018)], makes use of some linear algebraic identities which show, in particular, that

$$\hat{\beta}_{Q,H} = QX'\big(XQX' + \lambda H^{-1}\big)^{-1}y = QX'\hat{\gamma}_{H,K_Q}. \tag{2.12}$$

REMARKS. (A) *Types of similarity kernels.* In general, a sufficient condition for a matrix K to be a similarity kernel is that it is induced by a feature map $\phi\colon \mathbb{R}^p \to \mathcal{K}$. More specifically, the i, j entry of K is defined as the inner product of the observations $x_i \in \mathbb{R}^p$ with respect to their transformed versions $K_{ij} = \langle \phi(x_i), \phi(x_j)\rangle$ in the new inner product space, $(\mathcal{K}, \langle\cdot,\cdot\rangle)$. Examples include $K_I = XX'$ or $K_Q = XQX'$, where $\mathcal{K}$ is $\mathbb{R}^p$ with inner product $\langle\cdot,\cdot\rangle_Q$ (as in DPCoA). It is this quadratic form that we require for K_Q in Proposition 2.2(2)–(3); see Freytag et al. (2013) for genomic applications of this form. On the other hand, H can be any symmetric positive semidefinite matrix. Here, we are more interested in biologically-motivated kernels, such as UniFrac or DPCoA, than mathematically-derived ones, such as those constructed from polynomials or radial basis functions [Schölkopf and Smola (2002)].

(B) *Co-informative kernels and the HSIC.* Any kernels K and H may be used in (2.10) and (2.11), but to be useful in this framework, we assume that they are "co-informative" in the sense that they exhibit a shared eigenstructure; for instance, both should be informative for classifying samples. This concept is illustrated in the simulation of Section 3.3 and Figure 4. The co-informativeness can be made precise using the Hilbert–Schmidt information criteria (HSIC) [Gretton et al. (2005)] or its relatives—the distance covariance [Székely and Rizzo (2009)] and the RV statistic [Robert and Escoufier (1976)]. Josse and Holmes (2016) provide a nice review of these and related kernel-based tests. The HSIC provides a test for the statistical dependence of two data sets, X_1 $(n \times p)$ and X_2 $(n \times q)$, and is based on the eigenspectrum of covariance operators defined by kernels created from X_1 and X_2. For two kernels K and H, the empirical HSIC is simply $\mathrm{trace}(HK)$. The HSIC is thus of particular interest in item (1) of Proposition 2.2, which shows how two co-informative kernels may be used to obtain a penalized estimate $\hat{\beta}_{Q,H}$.

(C) *Linear mixed models and KPR.* As an alternative to the regularization framework presented here, one may consider a kernel as a generalized covariance among either the p variables (using Q) or n subjects (using H) [Purdom (2011), Schaid (2010)]. This alternative representation can be made precise using the linear mixed

model (LMM) framework [Ruppert, Wand and Carroll (2003)]. Specifically, recall from equations (2.7) and (2.11) that

$$\begin{aligned}\hat{\beta}_{Q,H} &= \arg\min_{\beta\in\mathbb{R}^p}\{\|y - X\beta\|_H^2 + \lambda\|\beta\|_{Q^{-1}}^2\}\\ &= QX'\big(K_Q + \lambda H^{-1}\big)^{-1}y\\ &= QX'\arg\min_{\gamma\in\mathbb{R}^n}\{\|y - K_Q\gamma\|_H^2 + \lambda\|\gamma\|_{K_Q}^2\} = QX'\hat{\gamma}_{H,K_Q}.\end{aligned}$$

These regression estimates are compatible with $\beta \sim N(0, \sigma_b^2 Q)$, $\varepsilon \sim N(0, \sigma_e^2 H^{-1})$ and $\operatorname{var}(y) = (\tau K_Q + \lambda H^{-1})^{-1}$. And the estimate $\hat{\gamma}_{H,K_Q}$ is compatible with $\gamma \sim N(0, \sigma_a^2 K_Q^{-1})$ and $\varepsilon \sim N(0, \sigma_e^2 H^{-1})$. With regard to the latter, a genetic similarity between subjects (e.g., kinship) is often used for grouping subjects and several authors have proposed this form of kernel for testing the (global) genetic association with a trait or phenotype, y; see, for example, Schifano et al. (2012). In particular, these methods use the LMM framework to motivate and define a "kernel association test." The variance score statistic for testing the null hypothesis of no association between y and X ($H_0 : \beta = 0$) is, using our notation above, $\mathcal{T} := \|y\|_{H^{1/2}K_Q H^{1/2}}^2$. The kernel association testing framework has been applied to microbiome data using a single kernel at a time derived from UniFrac [Zhao et al. (2015)], but this is a test for whether $\beta \neq 0$ and, unlike our KPR framework, provides no insight about which taxa, as represented by coordinates of β, are associated with y.

2.4. *Regression with compositional data.* Data from 16S rRNA gene sequencing methods are random counts of the molecules in each sample. The number of sequence reads assigned to a taxon contains no information about the *actual* number of molecules in the sample; the total number of reads observed in two samples can vary by several orders of magnitude. Hence, only *relative* amounts can be investigated. Common approaches for normalizing these data include converting them to proportions (relative percent) or subsampling the sequences to create equal library sizes for each sample (rarefying). These data are "compositional" in the sense that the microbial abundances represent a proportion of a constant total. It is known, however, that compositional measures can result in spurious correlations among taxa [Pearson (1896), Aitchison (2003a), Friedman and Alm (2012)], an effect that can be quite extreme when there are a few dominant taxa.

Compositional data reside on the *simplex* $\mathbb{S}^{p-1}$ of unit-sum vectors in $\mathbb{R}^p$ and so standard multivariate methods do not apply [Aitchison (2003b), Egozcue and Pawlowsky-Glahn (2011), Li (2015), Lovell et al. (2015)]. In particular, because these data do not naturally reside in a Euclidean vector space, standard regression models based on Euclidean covariance measures are inappropriate. However, ordinary least-squares and ridge regression estimates are of the form

$\hat{\beta} = (X'X + \lambda I)^{-1}X'y$ (with $\lambda = 0$ and $\lambda > 0$, resp.). Thus, these estimates depend on the empirical covariance structure, $X'X$, among taxa, which may include spurious correlations. Similarly, Li (2015) points out that a naïve application of lasso regression is not expected to perform well due to the compositional nature of the covariates. He addresses this issue by applying a lasso regression model to the log-ratio abundances and imposing an additional constant-sum constraint on the coefficient vector, β.

We next show that the generality of KPR for handling non-Euclidean structures can be used to address the compositional nature of microbiome data. In particular, we propose an approach that uses the centered log-ratio transformation of the compositional vectors and an estimate of covariance among the log taxa counts that is obtained via Aitchison's variance matrix [Aitchison (2003b), Egozcue and Pawlowsky-Glahn (2011)].

Let X be the $n \times p$ sample-by-taxon matrix whose rows are relative percent (compositional) vectors $\{x_i\}_{i=1}^n \subset \mathbb{S}^{p-1}$. The columns of X will be denoted by x^k, corresponding to $k = 1, \ldots, p$ taxa. Let $g(z) = (\prod_{k=1}^p z^k)^{1/p}$ be the geometric mean of a row vector, z, and denote the centered log-ratio (CLR) transform of x_i by $\tilde{x}_i = \operatorname{clr}(x_i) := [\log \frac{x_i^1}{g(x_i)}, \ldots, \log \frac{x_i^p}{g(x_i)}]$. In what follows, we denote the matrix of CLR vectors by $\tilde{X}$, and use the normalized variation matrix T, of X, as defined by Aitchison (1982): $T_{k,\ell} = \operatorname{var}(\frac{1}{\sqrt{2}} \log \frac{x^k}{x^\ell})$. T is a symmetric dissimilarity matrix with zeros on the diagonal and entries that have squared Aitchison distance units: the Aitchison norm of a vector $x \in \mathbb{S}^{p-1}$ is defined as $\|x\|_a^2 = \frac{1}{2p} \sum_{k,\ell} (\log \frac{x^k}{x^\ell})^2$. In fact, $\|x\|_a^2 = \|\operatorname{clr}(x)\|^2$. One can show that T is related to the covariance matrix, C, of the log of the true unobserved taxa counts via $T = v\mathbf{1}' + \mathbf{1}v' - 2C$ [Li (2015)]. Consequently, $C = -\frac{1}{2}\mathcal{J}T\mathcal{J}$, and we can use C in place of Q in equation (2.5) to obtain

$$\tilde{\beta}_C = \arg\min_{\beta} \{\|y - \tilde{X}\beta\|_{\mathbb{R}^n}^2 + \lambda \|\beta\|_{C^{-1}}^2\}. \tag{2.13}$$

As a comparison, we observe that Li (2015) proposed a constrained regression

$$\mathrm{E}(y_i) = \beta_1 \log x_i^1 + \cdots + \beta_p \log x_i^p \quad \text{subject to} \quad \sum_{j=1}^p \beta_j = 0, \tag{2.14}$$

augmented with a lasso penalty to obtain an estimate of the form

$$\arg\min_{\beta} \left\{ \frac{1}{2n} \left\| y - \sum_j \log(x^j)\beta_j \right\|_{\mathbb{R}^n}^2 + \lambda \sum_j |\beta_j| \right\} \quad \text{subject to} \quad \sum_{j=1}^p \beta_j = 0.$$

The zero-sum constraint on β was emphasized for interpretability advantages over the standard lasso estimate. Temporarily denoting $\beta_p = -\sum_{j=1}^{p-1} \beta_j$, we see that

(2.14) is equivalent to

$$\begin{aligned}\mathrm{E}(y_i) &= \beta_1 \log \frac{x_i^1}{x_i^p} + \beta_2 \log \frac{x_i^2}{x_i^p} + \cdots + \beta_{p-1} \log \frac{x_i^{p-1}}{x_i^p} \\ &= \beta_1 \log x_i^1 + \beta_2 \log x_i^2 + \cdots + \beta_{p-1} \log x_i^{p-1} - \sum_{j=1}^{p-1} \beta_j \cdot \log x_i^p.\end{aligned}$$

Since $\sum_{j=1}^{p} \beta_j = 0$, this can be rewritten as

$$\begin{aligned}\mathrm{E}(y_i) &= \beta_1 \log x_i^1 + \cdots + \beta_p \log x_i^p - \left(\sum_{j=1}^{p} \beta_j\right) \log g(x_i) \\ &= \beta_1 \log \frac{x_i^1}{g(x_i)} + \cdots + \beta_p \log \frac{x_i^p}{g(x_i)} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j = 0.\end{aligned}$$

Therefore, Li's proposal of regression on log-ratio abundances is equivalent to regression on the CLR-transformed data $\tilde{X}$ provided a zero-sum constraint is imposed on β. In contrast, however, our formulation does not explicitly impose a constant-sum constraint. In fact, this constraint is not needed because the CLR transform removes the analysis from the simplex to allow an analysis in Euclidean vector space algebra [Egozcue and Pawlowsky-Glahn (2011)]. Our model instead incorporates the appropriate covariance structure for the CLR transformation, C.

As a final observation, a positive-definite C in (2.13), or more generally Q in (2.5), can be decomposed as a sum $Q = I + \tilde{Q}$ of the identity plus a positive semi-definite singular matrix $\tilde{Q}$. The identity term constrains $\sum_{j=1}^{p} \beta_j^2$ to be small while, overall, $\tilde{Q}$ encourages extrinsic structure (e.g., smoothness). One may also control the size of $\sum_{j=1}^{p} \beta_j^2$ by adding or subtracting values in the diagonal entries of Q. This idea is similar to that of "Grace-ridge" in Zhao and Shojaie (2016) where, in addition to the penalty induced by Q, the authors propose to further impose a ridge-type penalty in the objective. We apply the significance testing framework of Zhao and Shojaie (2016) in Section 4.

3. Numerical experiments. To illustrate the proposed framework, we perform several data-driven simulations using publicly available microbiome data. We consider three scenarios from the literature that exploit extrinsic structure from a phylogenetic tree, including DPCoA, UniFrac and edge PCA. To achieve realistic simulations, we simulate "true" signals of the type implied by each of these methods in order to create benchmarks for performance evaluation. Our emphasis is on formalizing the role that such structure plays in penalized regression when modeling associations between the multivariate data, X, and a response variable, y. Since y is directly simulated from X in these settings, the compositional nature

of the data discussed in Section 2.4 does not affect the simulation results. We will return to this topic when analyzing the relative abundance data in Section 4.

The numerical experiments in this section are motivated by the relationship between the PCoA plots and PCR described in Section 2.1 and Figure 1(b). This connection can be generalized to a number of other commonly-used graphical representations in the microbiome literature. For instance, any two-dimensional DPCoA plot involves an implicit coefficient vector, β, of associations between y and X.

Throughout this section, we compare the performance of KPR with ridge regression and lasso. Ridge regression provides a direct extension of ordinary least squares and thus is a natural benchmark for comparing various KPR estimates. Lasso, which gives sparse estimates, is used as a benchmark in settings where the true β is sparsely nonzero. The choice of competing methods is limited by our emphasis on *estimating* β, rather than predicting the outcome y. Indeed, most kernel methods focus on prediction which renders them inappropriate for comparison.

In all simulation experiments, the tuning parameters for KPR, ridge and lasso are chosen using 10-fold cross-validation. Specifically, to compare the prediction performance of KPR, ridge and lasso, we choose the tuning parameters that minimize squared test error in held-out cross validation samples (CV min). On the other hand, the task of estimation usually requires more smoothing than prediction [Cai and Hall (2006)]. Therefore, when examining the estimation performances of KPR, ridge and lasso, we use the largest tuning parameters such that the squared test errors are within one standard error of the minimum squared test error (CV 1se), as suggested in Hastie, Tibshirani and Friedman (2009). For comparison, we also consider the tuning parameters corresponding to the minimum squared test error for ridge and lasso.

3.1. *Regression and DPCoA.* In our first example, we compare the estimation and prediction performances of KPR, ridge and lasso using the data depicted in Figure 1. The rows of X represent relative abundances of $p = 149$ taxa from $n = 100$ subjects in a study by Yatsunenko et al. (2012). The outcome y is log-transformed age of each subject. For KPR, we use $K_Q = XQX'$ and $H = I$, where $Q = -\frac{1}{2}\mathcal{J}\delta\mathcal{J}$ is a matrix of similarities between taxa obtained from the matrix of squared patristic distances, δ. Motivated by DPCoA plots, we assume the underlying "true" response y_{True} is generated from the first two eigenvectors of K_Q. Let L be the Cholesky factor of Q, that is, $Q = LL'$, and let $XL = U^L S^L (V^L)'$. Recall that $A_{(k)}$ denotes the first k columns of matrix A, or its first k rows and columns if A is diagonal. Motivated by (2.6), we let

$$\beta_{\text{True}} = s\big(V^L_{(2)}\big(S^L_{(2)}\big)^{-1}\big(U^L_{(2)}\big)'y, \tau\big), \tag{3.1}$$

where, $s(\cdot, \tau)$ is the hard-thresholding operator, that is, $s(x, \tau) = x \cdot 1(|x| > \tau)$. The threshold $\tau \geq 0$ is set to achieve various levels of sparsity: $\|\beta_{\text{True}}\|_0 \in$

$\{\lfloor 0.2p \rfloor, \lfloor 0.6p \rfloor, p\}$. After generating β_{True}, we simulate

$$y_{\text{True}} = U^L_{(2)} S^L_{(2)} (V^L_{(2)})' \beta_{\text{True}}.$$

The simulation is repeated 500 times, each with a different $\varepsilon \sim N_n(0, \sigma_\varepsilon^2 I_n)$ in $y_{\text{obs}} = y_{\text{True}} + \varepsilon$. Further, σ_ε^2 is set to achieve $R^2 = \text{var}(y_{\text{True}})/(\text{var}(y_{\text{True}}) + \sigma_\varepsilon^2) \in \{0.1, 0.2, \ldots, 0.9\}$. In each repetition, we estimate $\hat{\beta}_{\text{DPCoA}}$ from y_{obs} according to Definition 2.1. To make the simulation more realistic, we add error to the matrix Q used to simulate β_{True} and y_{True}, that is, we use Q_{obs}, obtained by adding random Gaussian noise to Q, to estimate $\hat{\beta}_{\text{DPCoA}}$. The eigenvalues of Q_{obs} are adjusted to be equal to the eigenvalues of Q. The amount of Gaussian noise added to the entries of Q_{obs} is empirically determined to achieve $\|Q - Q_{\text{obs}}\|_F / \|Q\|_F \in \{0, 0.25, 0.5\}$. As a comparison, we estimate $\hat{\beta}_{\text{Ridge}}$ and $\hat{\beta}_{\text{Lasso}}$ using only X and y_{obs}, without incorporating Q. From the estimated coefficients, we compute $\hat{y}_{\text{DPCoA}} = X L_{\text{obs}} \hat{\beta}_{\text{DPCoA}}$, $\hat{y}_{\text{Ridge}} = X\hat{\beta}_{\text{Ridge}}$ and $\hat{y}_{\text{Lasso}} = X\hat{\beta}_{\text{Lasso}}$. The performance metrics are the prediction sum of squared error (PSSE) from y_{True} and estimation sum squared error (ESSE) from β_{True}.

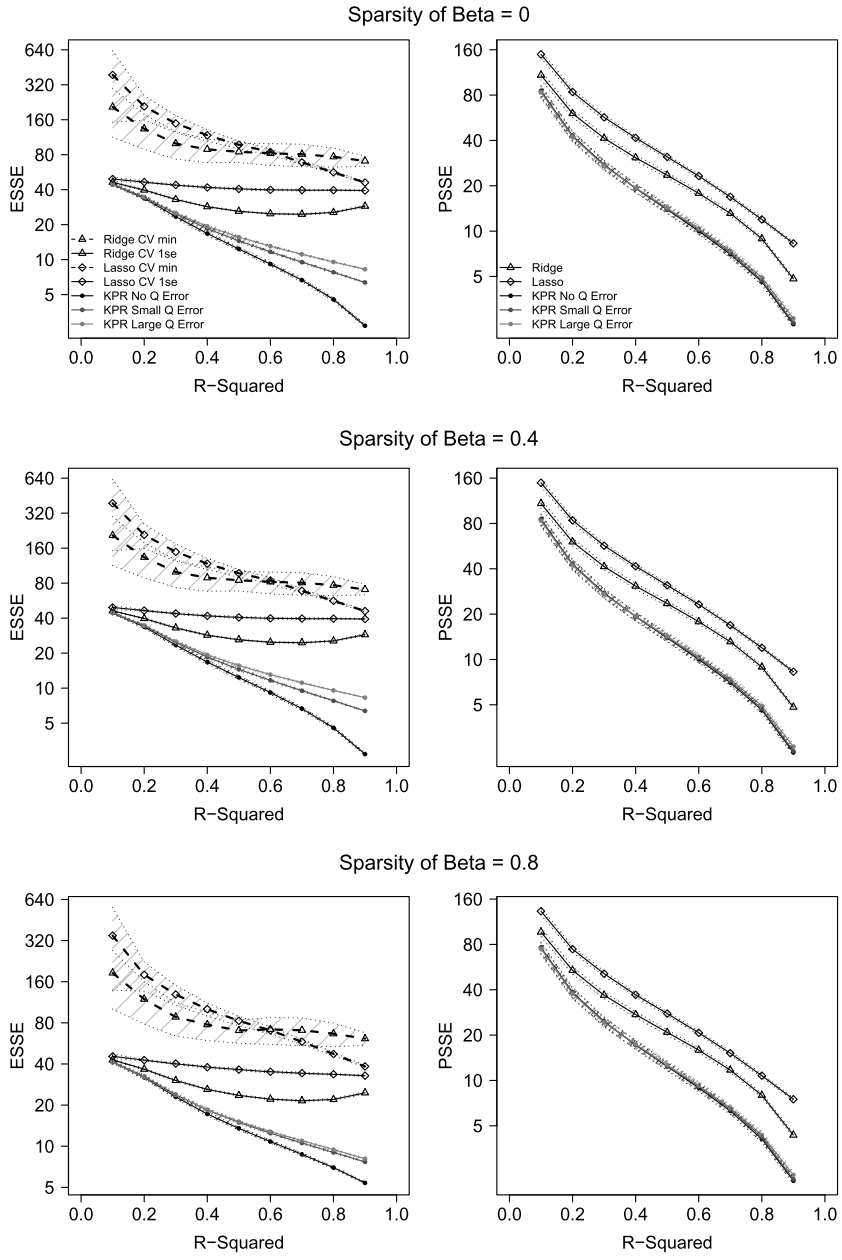
Figure 2 shows the estimation and prediction performance of KPR, ridge and lasso. KPR significantly outperforms both ridge regression and lasso for both prediction and estimation in all settings. As expected, the performance of ridge and lasso for *estimation* improve when using a larger tuning parameter. On the other hand, neither misspecification of Q nor sparsity of β_{True} seems to substantially impact the relative performance of the three methods. This may be due to the fact that KPR estimates the correct target β_{True}, even with misspecified Q, whereas ridge regression and lasso estimate the wrong target.

3.2. *Regression and PCoA with respect to a UniFrac kernel.* In the case of PCoA with respect to a UniFrac matrix Δ^U of squared dissimilarities, the graphical displays are based on the eigendecomposition of $H = -\frac{1}{2}\mathcal{J}\Delta^U\mathcal{J}$. That is, for $H = U^H(S^H)^2(U^H)' \approx U^H_{(2)}(S^H_{(2)})^2(U^H_{(2)})'$, the n samples are represented in two dimensions by the columns of $U^H_{(2)}S^H_{(2)}$; this results in points $\{\eta_{i1}, \eta_{i2}\}_{i=1}^n := \{(\sigma_1 U^H_{i1}, \sigma_2 U^H_{i2})\}_{i=1}^n$, as plotted in Figure 1. When the points are colored according to a response variable, $\{y_i\}_{i=1}^n$, the implied regression model is

$$\begin{aligned} y &= \gamma_1\eta_1 + \gamma_2\eta_2 + \varepsilon \\ &= U^H_{(2)} S^H_{(2)} \gamma + \varepsilon. \end{aligned} \tag{3.2}$$

However, in contrast to PCR in equation (2.2), where $US = XV$, it is not obvious how to connect γ directly to the p-coordinates corresponding to the p columns of X. Here, we exploit the joint eigenstructure of kernels K_I and H by proceeding as in (2.11) to obtain the estimate $\hat{\beta}_H = X'\hat{\gamma}$ as in (2.12), with $Q = I$.

In this example, we use the same data as in Section 3.1. For KPR, we use $K = XX'$ and obtain $H = -\frac{1}{2}\mathcal{J}\Delta^U\mathcal{J}$ using the UniFrac distance matrix. We simulate


Sparsity of Beta = 0
ESSE
PSSE
R−Squared
Ridge CV min
Ridge CV 1se
Lasso CV min
Lasso CV 1se
KPR No Q Error
KPR Small Q Error
KPR Large Q Error
Ridge
Lasso
Sparsity of Beta = 0.4
Sparsity of Beta = 0.8

γ_{True} and y_{True} from the first two eigenvectors of H, as in (3.2):

$$\begin{aligned} \gamma_{\text{True}} &= \big((U^H_{(2)})'(U^H_{(2)})\big)^{-1} S^H_{(2)} (U^H_{(2)})' y \\ y_{\text{True}} &= U^H_{(2)} S^H_{(2)} \gamma_{\text{True}}. \end{aligned} \tag{3.3}$$

This bivariate ordinary least squares regression is illustrated in Figure 1(b).

The simulation is repeated 500 times, each with a different $\varepsilon \sim N_n(0, \sigma_\varepsilon^2 I_n)$ to produce various values of $R^2 \in \{0.1, 0.2, \ldots, 0.9\}$. We compute $\hat{y}_{\text{KPR}} = K\hat{\gamma}_{\text{KPR}}$, where $\hat{\gamma}_{\text{KPR}}$ is estimated using (2.10). Similar to the last example, we do not assume we always observe the H matrix that is used to generate γ_{True} and y_{True}; rather, we use a noisy version, H_{obs}, of H in KPR with $\|H - H_{\text{obs}}\|_F / \|H\|_F \in \{0, 0.25, 0.5\}$.

Although we estimate β here as in (2.7) (with $Q = I$), there is no obvious way to simulate a β_{True} using UniFrac and so we do not compare the methods based on their estimation performances, and only consider prediction. For all three methods, we find the tuning parameters that minimize the cross-validated H_{obs}-weighted squared test error. While the use of H in tuning ridge and lasso penalties deviates from the common practice, it results in improved performances, given the important role of H in this simulation. The H matrix also defines the valid distance in this example. Thus, to evaluate the prediction performances of various methods, we use the H-weighted prediction sum of squared error (HPSSE), $\|\hat{y} - y_{\text{True}}\|^2_H$.

Figure 3 shows that KPR consistently outperforms ridge regression and lasso in prediction, even with a reasonable amount of misspecification of H. This may be due to the fact that, with the incorporation of the H matrix, KPR estimates the correct target whereas ridge and lasso do not.

3.3. *Regression and PCoA using an edge-matrix kernel.* In this section, simulations are based on data from a study of bacterial vaginosis (BV) by Srinivasan et al. (2012) in which 16S rRNA gene samples were collected using vaginal swabs from $n = 220$ women with and without BV. Here, the outcome y represents pH measured from vaginal fluid of each subject and we consider the association of

FIG. 2. *Estimation sum squared error* (*ESSE*: *left panels*) *and prediction sum squared errors* (*PSSE*: *right panels*) *of KPR, ridge regression and lasso and their* 95% *confidence bands. Standard errors for ESSE and PSSE are estimated based on* 500 *simulation runs, and are roughly* 0.5%–2% (*ESSE*) *and* 2%–5% (*PSSE*) *for KPR. We consider three sparsity settings for* β_{True}, *based on* (3.1): $\|\beta_{\text{True}}\|_0 = p$ *in top panels,* $\|\beta_{\text{True}}\|_0 = \lfloor 0.6p \rfloor$ *in center panels and* $\|\beta_{\text{True}}\|_0 = \lfloor 0.2p \rfloor$ *in bottom panels. For ridge and lasso, tuning parameters that produce the smallest cross-validated squared test error* (*CV min*), *and the largest tuning parameters such that the cross-validated squared test errors are within one standard error of the minimum cross-validated squared test error* (*CV* 1*se*) *are considered. For KPR, we consider* $\|Q - Q_{\text{obs}}\|_F / \|Q\|_F = 0$ (*no Q error*), 0.25 (*small Q error*) *and* 0.5 (*large Q error*).

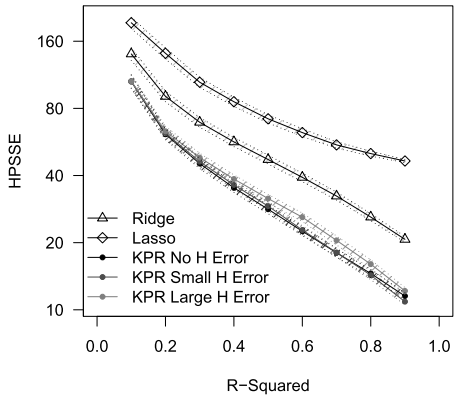


FIG. 3. *H-weighted prediction sum of squared error (HPSSE) of KPR, ridge and lasso, with* 95% *confidence bands. Standard errors for HPSSE are estimated based on* 500 *simulation runs, and are roughly* 1%–4% *for KPR. For KPR, we consider* $\|H - H_{\text{obs}}\|_F / \|H\|_F = 0$ (*no H Error*), 0.25 (*small H Error*) *and* 0.5 (*large H Error*).

y with genus-level taxa. In this example, we use the $p = 62$ genera that exhibit nonzero sequence counts in at least 20% of the subjects. So here, X represents 220×62 abundances in a sample-by-genus matrix, and we use a kernel $K = XX'$. Additionally, however, we define a second kernel $H = EE'$ based on the "edge mass difference matrix," E, originally introduced by Matsen and Evans (2013). If the full phylogenetic tree has q edges, each sample can be represented by a vector indexed by all q edges, the eth coordinate of which quantifies the difference between the fraction of sequence reads on either side of the edge; that is, the fraction of reads observed on the root side of the tree minus the fraction of reads on the nonroot side. We refer to Matsen and Evans (2013) for details and a discussion of "edge PCA," which refers to PCA applied to the $n \times q$ matrix E. Note, in particular, that abundances from *every* taxon level in the tree contribute to a similarity between subjects as opposed to abundances at a single taxon level, which is used in UniFrac or DPCoA.

In summary, X represents $p = 62$ genus-level abundances while E is based on all $q = 1770$ edges in the original phylogenetic tree. Figure 4(a) shows a PCA plot of the 220 subjects in which their similarity is defined using the edge kernel $H = EE'$; the color of each dot represents the subject's pH. Figure 4(b) is a heatmap of the kernel H used to create Figure 4(a). The columns and rows of H represent similarities between samples based on the edge mass difference matrix, ordered by subjects' pH measurements. Similarly, Figure 4(c) is a (Euclidean) PCA plot based on similarities defined using the genus-level abundance kernel, $K = XX'$. Figure 4(d) is a heatmap of the kernel K used to create Figure 4(c), and subjects are again ordered by pH. These figures illustrate how two different measures of similarity (two separate kernels) may be co-informative in the sense that they both

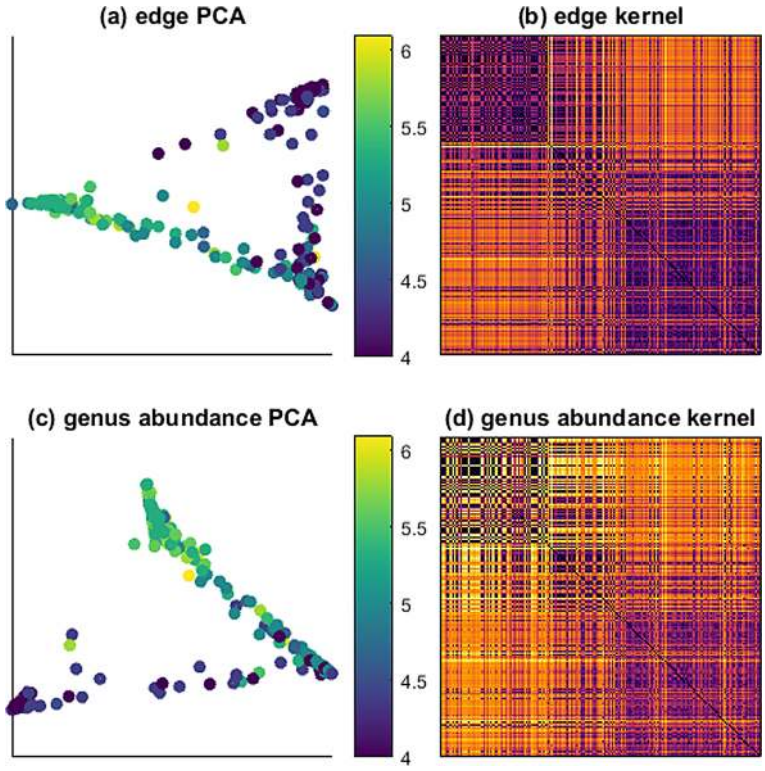


FIG. 4. *Analysis of bacterial vaginosis data from Srinivasan et al.* (2012). (a): *Representation of the samples in the space of the first two PCs of the edge-matrix kernel* $H = EE'$. *The color corresponds to the pH of the sample*; (b): *Heatmap of edge-matrix kernel used to generate the plot in* (a); (c): *Two-dimensional PCA plot based on the genus-level relative abundances, colored according to pH*; (d): *Heatmap of the genus-abundance kernel* $K = XX'$ *used to create the plot in* (c). *In* (b) *and* (d), *subjects are ordered by the pH values*.

provide information about grouping of subjects' microbiota in relation to their pH. It is thus natural to expect that incorporating information from both H and K within the KPR framework may result in improved estimates of association between $y = \text{pH}$ and the microbial abundances.

For the simulation, we define a "true" association between pH and the genus-level taxa in X using the 2-dimensional PCR model in equations (2.2) and (2.3). Specifically, we use the apparent association between $y = \text{pH}$ and genus-level abundances in Figure 4(c) to construct a "true" coefficient vector β_{True} as follows. Using the SVD of $X = USV'$, and proceeding as in (3.3), define

$$\gamma_{\text{True}} = \left[(U_{(2)}S_{(2)})'(U_{(2)}S_{(2)})\right]^{-1}(U_{(2)}S_{(2)})'y,$$
$$y_{\text{True}} = U_{(2)}S_{(2)}\gamma_{\text{True}}.$$

We then project y_{True} onto the space spanned by the first two singular vectors of X to define a true coefficient vector as

$$\beta_{\text{True}} = V_{(2)} S_{(2)}^{-1} U_{(2)}' y_{\text{True}}.$$

We now consider how the contribution of $H = EE'$ can aid in both the prediction of y_{True} and the estimation of β_{True} even though, by construction, neither are informed by E.

Taking $H = EE'$ in a KPR model of the form (2.10), we compare the resulting estimate of β with ridge and lasso estimates. The simulation is repeated 500 times, each with a different $\varepsilon \sim N_n(0, \sigma_\varepsilon^2 I_n)$ to produce various values of $R^2 \in \{0.1, 0.2, \ldots, 0.9\}$. The performance metrics are the estimation sum squared error (ESSE) the H-weighted prediction sum squared error (HPSSE) as in the previous section. In this numerical example, we do not assume we always observe the true H matrix; rather, we use a noisy version, H_{obs}, of H in KPR with $\|H - H_{\text{obs}}\|_F / \|H\|_F \in \{0, 0.25, 0.5\}$. For all three methods, tuning parameter values are chosen to minimize the sum of squared test error weighted by H_{obs}. As in the simulation for DPCoA, we also allow for using the largest tuning parameters such that the squared test error weighted by H is within one standard error of the minimum squared test error.

Figure 5 shows that KPR significantly outperforms ridge and lasso in both prediction and estimation. Even though H is not used to simulate the true association, using the edge kernel in KPR enhances the performance of both estimation and prediction, as long as H is not severely misspecified. Once again, the performance of ridge and lasso estimates improve when using a larger tuning parameter (CV 1se).

4. Application to an observational study. We apply our kernel-penalized regression framework to data from 16S rRNA gene collected in a study of premenopausal women [Hullar et al. (2015)]. This study investigated aspects of gut microbial communities in stool samples from premenopausal women using 454 pyrosequencing of the 16S rRNA gene. The abundances of 127 species were zero for more than 90% of the subjects and were removed from our analysis. The data set we consider consists of $p = 128$ species sampled from $n = 102$ women.

To make the measurements comparable between subjects, the species abundances were scaled by the total number of sequences measured in each sample. This scaling produces compositional data (the relative abundances in each sample sum to 1) which introduces analytical complications. In particular, regression analysis using compositional covariates must somehow account for their unit sum constraint [Kurtz et al. (2015), Li (2015)]. For this reason, we apply the CLR transformation to the relative abundance values and use this transformed data $\tilde{X}$ as the matrix of predictors in the KPR model. Additionally, using Aitchison's variation matrix [Aitchison (1982)], T, we obtain the covariance matrix, C, as described

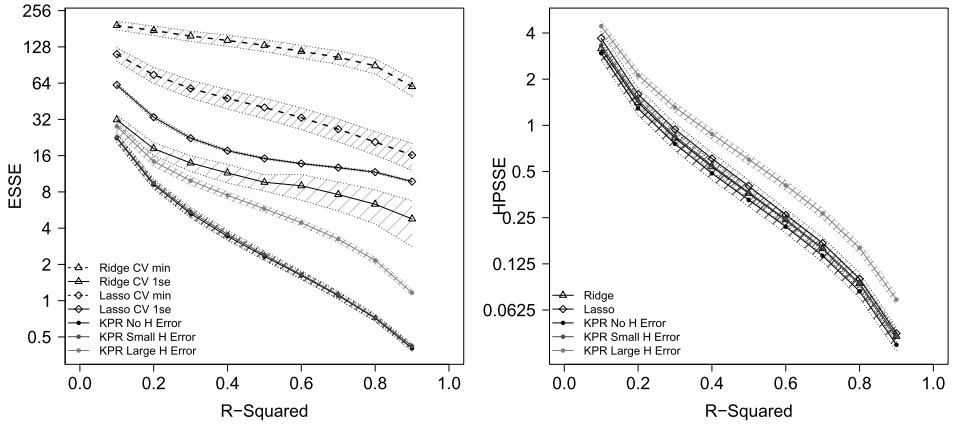


FIG. 5. *In silico evaluation of using tree-based edge information in regression models. Estimation sum squared error* (*ESSE*) *and* H*-weighted prediction sum squared error* (*HPSSE*) *of KPR, ridge regression and lasso, with the* 95% *confidence bands. Standard errors for ESSE and HPSSE are estimated based on* 500 *simulation runs, and are roughly* 2%–5% (*ESSE*) *and* 3%–5% (*hPSSE*) *for KPR. For KPR, we consider* $\|H - H_{\text{obs}}\|_F / \|H\|_F = 0$ (*no* H *error*), 0.25 (*small* H *error*) *and* 0.5 (*large* H *error*).

prior to equation (2.13). As C provides more accurate information on the covariance among the true abundances than does the empirical covariance matrix from relative abundances, X, or their CLR transform, $\tilde{X}$, we use C in place of Q in (2.5).

In this example, we examine the effect of using the CLR transformed data $\tilde{X}$ and covariance C as in (2.13) and fit penalized regression models with the goal of estimating $\tilde{\beta}_C$ in (2.13) for the purpose of identifying specific species that may be associated with percent fat in the cohort described above. To this end, we apply a recently developed significance testing procedure to three high-dimensional models in order to identify species exhibiting evidence of association with subjects' adiposity. This significance test for graph-constrained estimation, called Grace [Zhao and Shojaie (2016)], provides a means to assign significance to estimates from penalized regression models that incorporate structure of the type provided by Q in (2.5) [or C in (2.13)]. The method asymptotically controls the type-I error rate regardless of the choice of Q. The special case with $Q = I$ provides a significance test for ordinary ridge regression. In each application of the Grace test, tuning parameters are selected based on the smallest squared test error using 10-fold cross validation. Following Zhao and Shojaie (2016), the assumed sparsity parameter is set to be $\xi = 0.05$. The tuning parameter for the initial estimator is set to be $\lambda_{\text{init}} = 4\hat{\sigma}_\varepsilon \sqrt{3 \log p/n}$, where $\hat{\sigma}_\varepsilon$ is the estimated standard deviation of the random error ε, using the scaled lasso [Sun and Zhang (2012)]. To assess significance for the sparse models using lasso, we apply the recently proposed significance test for lasso regressions based on low-dimensional projection estimator

TABLE 1

Species found to be associated with percent fat (in increasing order of p-values) at different significant levels using: KPR with centered log-ratio transformed abundances (CLR); ridge and lasso regression with centered log-ratio transformed abundances; and ridge and lasso regression with untransformed relative abundances (rel%)

	$p < 0.01$	$p < 0.005$	FDR < 0.1
KPR + CLR	*Bacteroides*, *Anaerovorax*, *Acidaminococcus*, *Blautia*, *Dethiosulfatibacter*, *Asaccharobacter*, *Turicibacter*, *Lebetimonas*, *Streptobacillus*, *Anoxynatronum*	*Bacteroides*, *Turicibacter*, *Acidaminococcus*, *Dethiosulfatibacter*	(none)
Ridge + CLR	(none)	(none)	(none)
Ridge + rel%	*Catonella*, *Dethiosulfatibacter*	(none)	(none)
Lasso + CLR	*Roseburia*	(none)	(none)
Lasso + rel%	*Dethiosulfatibacter*, *Micropruina*	*Dethiosulfatibacter*	(none)

(LDPE) [Zhang and Zhang (2014), Van de Geer et al. (2014)], which provides an asymptotically valid test for lasso-penalized regression estimates.

We report on five regression estimation methods for which the significance of regression coefficients can be evaluated using existing high-dimensional testing methods. Two are obtained using the relative abundances, X, with respect to: (i) an ordinary ridge penalty and (ii) a lasso penalty. Three are obtained using the CLR transformed abundances, $\tilde{X}$, with respect to: (iii) an ordinary ridge penalty, (iv) a lasso penalty and (v) the KPR estimate in (2.13). None of these methods results in any species associated with the outcome of percent fat when controlled for false discovery rate (FDR) at 0.1 using the Benjamini–Yekutieli procedure [Benjamini and Yekutieli (2001)]. However, when using a cut-off of $p = 0.01$, the KPR estimate (2.13) results in ten species. With a cut-off of $p = 0.005$, KPR results in four species. Ordinary ridge regressions using the CLR-transformed vectors find no associations at a cut-off of $p = 0.01$, whereas using the relative abundances, ridge finds two species at the $p = 0.01$ cut-off and none at $p = 0.005$. Lasso regression with the CLR-transformed vectors identifies one specie at the $p = 0.01$ cut-off and none at $p = 0.005$ cut-off. When using the relative abundances, lasso identifies two species as significant at the $p = 0.01$ cut-off and one at the $p = 0.005$ cutoff. See Table 1 for the list of identified species.

5. Discussion. We have formulated a family of regression models that naturally extends the dimension-reduced graphical explorations common to microbiome studies. In this sense, we have simply refocused the role of the eigenstructures used in ordination methods toward exploiting this structure in penalized regression models. The large family of models developed here provides a supervised

statistical learning counterpart to the unsupervised methods of principal coordinate analysis (PCoA).

A primary motivation for PCoA graphical displays is the ability to incorporate biologically-inclined measures of (dis)similarity. The popular use of UniFrac, for instance, is motivated by the desire to impose phylogeny into the analysis. These dissimilarities have also been used for rigorous statistical testing in the context of Anderson's nonparametric MANOVA [Anderson (2006)] or the closely-related kernel machine regression score test [Chen et al. (2012), Pan (2011), Zhao et al. (2015)] for global association of a multivariate predictor with an outcome. However, the use of UniFrac and other non-Euclidean distances make it difficult to identify specific associations between the microbial abundance profiles and a phenotype; indeed, none of these analyses proceed to estimate the individual associations. In addition to ordination displays and global tests for associations, a variety of machine learning approaches have emphasized on models that *predict* a response. In contrast, we focus on *estimating* the coefficient vector, which is a key aspect of any approach used to draw scientific conclusions based on the association of microbial communities with an outcome or phenotype.

An interesting feature of the proposed kernel-penalized regression framework is its ability to sidestep some of the problems inherent in compositional data analysis. Indeed, as emphasized by Li (2015) regression analysis with compositional covariates must somehow acknowledge their unit-sum constraint and spurious correlations. Our approach, which differs somewhat from that of Li (2015), may also be viewed as a penalized version of the low-dimensional linear model for compositions by Tolosana-Delgado and Van Den Boogart (2011), who use the isometric log-ratio (ILR) coordinates. We note that ILR coordinates arise from the SVD of mean-centered CLR-transformed data, $\tilde{X}$ [see Egozcue and Pawlowsky-Glahn (2011)], which is also used in our model. However, to estimate $\beta \in \mathbb{R}^p$, we used instead a regularization framework; our penalty in Section 2.4 arises from Aithison's total variation matrix whose singular values are the total variances of ILR components. Moreover, the proposed framework also allows us to use existing inference frameworks for high-dimensional regression, and in particular the Grace test [Zhao and Shojaie (2016)], to assess the significance of estimated regression coefficients.

SUPPLEMENTARY MATERIAL

Supplement to "Kernel-penalized regression for analysis of microbiome data" (DOI: 10.1214/17-AOAS1102SUPP; .pdf). Mathematical justification and remarks for equation (2.10) and Proposition 2.2.

REFERENCES

AITCHISON, J. (1982). The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. B* **44** 139–177. With discussion. MR0676206

AITCHISON, J. (2003a). A concise guide to compositional data analysis. In 2*nd Compositional Data Analysis Workshop*.

AITCHISON, J. (2003b). *The Statistical Analysis of Compositional Data*. The Blackburn Press, Caldwell, NJ.

ANDERSON, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* **62** 245–253, 320. MR2226579

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

BÜHLMANN, P., KALISCH, M. and MEIER, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annu. Rev. Statist. Appl.* **1** 255–278.

CAI, T. T. and HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179.

CHEN, J., BITTINGER, K., CHARLSON, E. S., HOFFMANN, C., LEWIS, J., WU, G. D., COLLMAN, R. G., BUSHMAN, F. D. and LI, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* **28** 2106–2113.

CLAESSON, M. J., JEFFERY, I. B., CONDE, S., POWER, S. E., O'CONNOR, E. M., CUSACK, S., HARRIS, H. M., COAKLEY, M., LAKSHMINARAYANAN, B., O'SULLIVAN, O., FITZGERALD, G. F., DEANE, J., O'CONNOR, M., HARNEDY, N., O'CONNOR, K., O'MAHONY, D., VAN SINDEREN, D., WALLACE, M., BRENNAN, L., STANTON, C., MARCHESI, J. R., FITZGERALD, A. P., SHANAHAN, F., HILL, C., ROSS, R. P. and O'TOOLE, P. W. (2012). Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488** 178–184.

EGOZCUE, J. J. and PAWLOWSKY-GLAHN, V. (2011). Basic concepts and procedures. In *Compositional Data Analysis*: *Theory and Applications* (V. Pawlowsky-Glahn and A. Buccianti, eds.) 12–28. Wiley, Chichester.

EVANS, S. N. and MATSEN, F. A. (2012). The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 569–592.

FRANKLIN, J. N. (1978). Minimum principles for ill-posed problems. *SIAM J. Math. Anal.* **9** 638–650.

FREYTAG, S., MANITZ, J., SCHLATHER, M., KNEIB, T., AMOS, C. I., RISCH, A., CHANG-CLAUDE, J., HEINRICH, J. and BICKEBÖLLER, H. (2013). A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. *Hum. Hered.* **76** 64–75.

FRIEDMAN, J. and ALM, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8** e1002687.

FUKUYAMA, J., MCMURDIE, P. J., DETHLEFSEN, L., RELMAN, D. A. and HOLMES, S. (2012). Comparisons of distance methods for combining covariates and abundances in microbiome studies. In *Pacific Symposium on Biocomputing* 2012 213–224.

GOLUB, G. H. and VAN LOAN, C. F. (2012). *Matrix Computations*. Johns Hopkins Univ. Press, Baltimore, MD.

GOODRICH, J. K., RIENZI, S. C. D., POOLE, A. C., KOREN, O., WALTERS, W. A., CAPORASO, J. G., KNIGHT, R. and LEY, R. E. (2014). Conducting a microbiome study. *Cell* **158** 250–262.

GOWER, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53** 325–338.

GREENACRE, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, Cambridge, MA.

GRETTON, A., HERBRICH, R., SMOLA, A., BOUSQUET, O. and SCHÖLKOPF, B. (2005). Kernel methods for measuring independence. *J. Mach. Learn. Res.* **6** 2075–2129.

HAMADY, M. and KNIGHT, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res.* **19** 1141–1152.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*. Springer, New York.

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

HULLAR, M. A., LANCASTER, S. M., LI, F., TSENG, E., BEER, K., ATKINSON, C., WÄHÄLÄ, K., COPELAND, W. K., RANDOLPH, T. W., NEWTON, K. M. and LAMPE, J. W. (2015). Enterolignan-producing phenotypes are associated with increased gut microbial diversity and altered composition in premenopausal women in the United States. *Cancer Epidemiol. Biomark. Prev.* **24** 546–554.

JOSSE, J. and HOLMES, S. (2016). Measuring multivariate association and beyond. *Stat. Surv.* **10** 132–167.

KIM, S. and XING, E. P. (2010). Tree-guided group lasso for multi-task regression with structured sparsity. In *Proceedings of the* 27*th International Conference on Machine Learning* (*ICML* 2010) (J. Fürnkranz and T. Joachims, eds.) 543–550.

KOREN, O., KNIGHTS, D., GONZALEZ, A., WALDRON, L., SEGATA, N., KNIGHT, R., HUTTENHOWER, C. and LEY, R. E. (2013). A guide to enterotypes across the human body: Meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput. Biol.* **9** e1002863.

KUCZYNSKI, J., LIU, Z., LOZUPONE, C., MCDONALD, D., FIERER, N. and KNIGHT, R. (2010). Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Methods* **7** 813–819.

KURTZ, Z. D., MUELLER, C. L., MIRALDI, E. R., LITTMAN, D. R., BLASER, M. J. and BONNEAU, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11** e1004226.

LI, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu. Rev. Statist. Appl.* **2** 73–94.

LI, C. and LI, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24** 1175–1182.

LOVELL, D., PAWLOWSKY-GLAHN, V., EGOZCUE, J. J., MARGUERAT, S. and BÄHLER, J. (2015). Proportionality: A valid alternative to correlation for relative data. *PLoS Comput. Biol.* **11** e1004075.

LOZUPONE, C. and KNIGHT, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71** 8228–8235.

LOZUPONE, C. A., HAMADY, M., KELLEY, S. T. and KNIGHT, R. (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.* **73** 1576–1585.

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1980). *Multivariate Analysis*. Academic Press, San Diego, CA.

MATSEN, F. A. and EVANS, S. N. (2013). Edge principal components and squash clustering: Using the special structure of phylogenetic placement data for sample comparison. *PLoS ONE* **8** e56859.

PAN, W. (2011). Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genet. Epidemiol.* **35** 211–216.

PAVOINE, S., DUFOUR, A.-B. and CHESSEL, D. (2004). From dissimilarities among species to dissimilarities among communities: A double principal coordinate analysis. *J. Theoret. Biol.* **228** 523–537.

PEARSON, K. (1896). Mathematical contributions to the theory of evolution—On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. R. Soc. Lond.* **60** 489–498.

PEKALSKA, E., PACLIK, P. and DUIN, R. P. (2002). A generalized kernel approach to dissimilarity-based classification. *J. Mach. Learn. Res.* **2** 175–211.

PURDOM, E. (2011). Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree. *Ann. Appl. Stat.* **5** 2326–2358.

RANDOLPH, T. W., HAREZLAK, J. and FENG, Z. (2012). Structured penalties for functional linear models—Partially empirical eigenvectors for regression. *Electron. J. Stat.* **6** 323–353.

RANDOLPH, W. T., ZHAO, S., COPELAND, W., HULLAR, M. and SHOJAIE, A. (2018). Supplement to "Kernel-penalized regression for analysis of microbiome data." DOI:10.1214/17-AOAS1102SUPP.

ROBERT, P. and ESCOUFIER, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **25** 257–265. MR0440801

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge Univ. Press, New York.

SCHAID, D. J. (2010). Genomic similarity and kernel methods I: Advancements by building on mathematical and statistical foundations. *Hum. Hered.* **70** 109–131.

SCHIFANO, E. D., EPSTEIN, M. P., BIELAK, L. F., JHUN, M. A., KARDIA, S. L. R., PEYSER, P. A. and LIN, X. (2012). SNP set association analysis for familial data. *Genet. Epidemiol.* **36** 797–810.

SCHÖLKOPF, B. and SMOLA, A. J. (2002). *Learning with Kernels*: *Support Vector Machines*, *Regularization*, *Optimization*, *and Beyond*. MIT Press, Cambridge, MA.

SRINIVASAN, S., HOFFMAN, N. G., MORGAN, M. T., MATSEN, F. A., FIEDLER, T. L., HALL, R. W., ROSS, F. J., MCCOY, C. O., BUMGARNER, R., MARRAZZO, J. M. et al. (2012). Bacterial communities in women with bacterial vaginosis: High resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PLoS ONE* **7** e37818.

SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898.

SZÉKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265.

TANASEICHUK, O., BORNEMAN, J. and JIANG, T. (2014). Phylogeny-based classification of microbial communities. *Bioinformatics* **30** 449–456.

THE HUMAN MICROBIOME PROJECT CONSORTIUM (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486** 207–214.

TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. MR2850205

TOLOSANA-DELGADO, V. and VAN DEN BOOGART, K. G. (2011). Linear models with compositions in R. In *Compositional Data Analysis*: *Theory and Applications* (V. Pawlowsky-Glahn and A. Buccianti, eds.) 12–28. Wiley, Chichester.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202.

VAN LOAN, C. F. (1976). Generalizing the singular value decomposition. *SIAM J. Numer. Anal.* **13** 76–83.

YATSUNENKO, T., REY, F. E., MANARY, M. J., TREHAN, I., DOMINGUEZ-BELLO, M. G., CONTRERAS, M., MAGRIS, M., HIDALGO, G., BALDASSANO, R. N., ANOKHIN, A. P., HEATH, A. C., WARNER, B., REEDER, J., KUCZYNSKI, J., CAPORASO, J. G., LOZUPONE, C. A., LAUBER, C., CLEMENTE, J. C., KNIGHTS, D., KNIGHT, R. and GORDON, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature* **486** 222–227.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940

ZHAO, S. and SHOJAIE, A. (2016). A significance test for graph-constrained estimation. *Biometrics* **72** 484–493. MR3515775

ZHAO, N., CHEN, J., CARROLL, I. M., RINGEL-KULKA, T., EPSTEIN, M. P., ZHOU, H., ZHOU, J. J., RINGEL, Y., LI, H. and WU, M. C. (2015). Testing in microbiome-profiling studies

with MiRKAT, the microbiome regression-based kernel association test. *Am. J. Hum. Genet.* **96** 797–807.

T. W. RANDOLPH
W. COPELAND
M. HULLAR
FRED HUTCHINSON CANCER RESEARCH CENTER
1100 FAIRVIEW AVE N
SEATTLE, WASHINGTON 98109
USA
E-MAIL: trandolp@fredhutch.org

S. ZHAO
A. SHOJAIE
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF WASHINGTON
BOX 357232
SEATTLE, WASHINGTON 98195
USA