



Multitable Methods for Microbiome Data Integration

Kris Sankaran^{1*} and Susan P. Holmes²

¹ Mila, Université de Montréal, Montréal, QC, Canada, ² Department of Statistics, Stanford University, Stanford, CA, United States

The simultaneous study of multiple measurement types is a frequently encountered problem in practical data analysis. It is especially common in microbiome research, where several sources of data—for example, 16s-rRNA, metagenomic, metabolomic, or transcriptomic data—can be collected on the same physical samples. There has been a proliferation of proposals for analyzing such multitable microbiome data, as is often the case when new data sources become more readily available, facilitating inquiry into new types of scientific questions. However, stepping back from the rush for new methods for multitable analysis in the microbiome literature, it is worthwhile to recognize the broader landscape of multitable methods, as they have been relevant in problem domains ranging across economics, robotics, genomics, chemometrics, and neuroscience. In different contexts, these techniques are called data integration, multi-omic, and multitask methods, for example. Of course, there is no unique optimal algorithm to use across domains—different instances of the multitable problem possess specific structure or variation that are worth incorporating in methodology. Our purpose here is not to develop new algorithms, but rather to 1) distill relevant themes across different analysis approaches and 2) provide concrete workflows for approaching analysis, as a function of ultimate analysis goals and data characteristics (heterogeneity, dimensionality, sparsity). Towards the second goal, we have made code for all analysis and figures available online at https://github.com/krisrs1128/multitable_review.

Keywords: microbiome, data integration, multiomics, dimensionality reduction, heterogeneity

Most methods in statistics expect data to be available as a single table. To a researcher confronted with multiple sources of data, it might therefore seem most natural to either analyze each source separately, one at a time, or else combine all data into a single, unified table. However, neither of these approaches is entirely satisfactory. First, many scientific problems can only be answered by collecting several complementary measurement types. Indeed, the situation is analogous to using many types of sensors to study a single system from many perspectives. Further, while in certain supervised problems, it is enough to predict a single measurement of interest, with other sources collected primarily to provide better features, there are often additional relational components to the analysis: how do different types of measurements co-vary with one another? Here, it is of interest to provide a representation of the data that facilitates comparisons across tables, rather than just comparing each table with a single response of interest. This richer scientific question motivates the development of methods distinct from those used to analyze a single measurement type at a time.

For more concrete motivation, we consider data from the WELL-China study, which is focused on the relationships between various indicators of wellness (Min et al., 2019). In this study,

OPEN ACCESS

Edited by:

Lingling An,
University of Arizona,
United States

Reviewed by:

Kui Zhang,
Michigan Technological University,
United States
Jing Ma,
Fred Hutchinson Cancer
Research Center, United States

*Correspondence:

Kris Sankaran
kris.sankaran@umontreal.ca

Specialty section:

This article was submitted to
Statistical Genetics and Methodology,
a section of the journal
Frontiers in Genetics

Received: 14 October 2018

Accepted: 17 June 2019

Published: 28 August 2019

Citation:

Sankaran K and Holmes SP (2019)
Multitable Methods for Microbiome
Data Integration.
Front. Genet. 10:627.
doi: 10.3389/fgene.2019.00627

1,969 individuals¹ underwent clinical examinations, filled out wellness surveys (covering topics such as exercise, sleep, diet, and mental health, for example), and provided stool samples, used for 16s-rRNA sequencing and metabolomic analysis. To date, 16s-rRNA sequencing data are available for 221 of these participants. Evidently, various interesting relational questions can be investigated using this data source.

For the purpose of illustration, we focus on one relatively narrow question that can be addressed using these data: How is the distribution of lean and fat mass across the body related to patterns of microbial abundance? The measurement types most relevant in this analysis are DEXA scans and 16s-rRNA sequencing abundances. DEXA scans use relative X-ray absorption to gauge the amount of lean and fat body mass within a region of the body being scanned. We have access to these lean and fat body mass measurements at several body sites—arms, legs, trunk, etc.—along with related body type variables, like height, age, and android and gynoid fat measurements. In total, there are 36 of these variables. 16s-rRNA sequencing is a technology for gauging the abundance of different bacterial species in the gut by counting the alignments of reads to the 16s-rRNA gene, a component of all bacterial genomes with enough variation to allow discrimination between different individual species. We have counts associated with 2,565 species across 181 genera, though the vast majority are present in low abundances.

This question of the relationship between lean and fat mass distribution (informally, “body type”) and the microbiome is motivated by findings that certain taxonomic groups are over- or underrepresented as a function of an individual’s body mass index (BMI) (Ley et al., 2005; Ley et al., 2006; Turnbaugh et al., 2009; Ley, 2010). Further, since the distribution of fat is often more related to underlying biological mechanisms than overall body mass (Matsuzawa, 2008), and since this distribution is mediated by specific metabolic pathways, there is reason to suspect that a joint analysis of DEXA and 16s-rRNA microbial abundance data might yield a more complete view of the relationship between the microbiome and body type.

We use this motivating dataset in the examples that follow. Additional numerical examples, for methods only discussed abstractly in this review, are available in the github repository associated with this paper.

CLASSICAL MULTIVARIATE METHODS

Methods from classical multivariate statistics are a mainstay of single-table microbiome data analysis, so it is natural to revisit them before surveying extensions to the multitable setting. Here, we explore a few of the classically studied multitable methods that fit nicely into the modern microbiome data analysis toolbox. We first describe a naive approach based on Principal Components Analysis (PCA)—naive because it lifts a single-table method to the multiple table setting without any special considerations—before studying approaches that directly characterize covariation across several tables: Canonical

Correlation Analysis (CCA), Multiple Factor Analysis (MFA), and Principal Component Analysis with Instrumental Variables (PCA-IV).

The earliest multitable method (CCA) was published in 1936, motivated by the problem of relating prices of groups of commodities (Hotelling, 1936). There are two notable aspects of data analysis in this classical paradigm that no longer hold in modern statistics,

- Even when many samples could be collected, there were typically only a few features for each sample, and it was straightforward to study all of them simultaneously. It is now possible to automatically collect a large number of features for each observation (or subject).
- Before electronic computers had been invented, it was important that all statistical quantities be easy to calculate, typically necessitating analytical formulas for parameter estimates. This is no longer an important limitation due to modern computation.

These changes have driven the development of high-dimensional methods and facilitated the adoption of iterative, more computationally intensive approaches.

Nonetheless, it is worth reviewing these original approaches, both to understand the context for many modern techniques and to have an easy starting point for practical data analysis. Indeed, these more established methods tend to be the most readily available through statistical computing packages and can provide a benchmark with which to compare more elaborate, modern methods.

PCA

The simplest approach to dealing with multiple tables is to combine them into one and apply a single-table method, for example, PCA. That is, write

$$X = [X^{(1)} | \dots | X^{(L)}] \in \mathbb{R}^{n \times p},$$

where $p = \sum_{l=1}^L p_l$, and compute the SVD $X = UDV^T$. The K -principal component directions are the first K columns v_1, \dots, v_K , while the associated scores are reweighted rows $d_1 u_1, \dots, d_K u_K$. We call this method concatenated PCA.

While this does not account for the multitable structure of the data, it does accomplish two goals:

- Through the principal component scores, it provides a visualization of the relationships between samples, based on all features.
- Through the principal component directions, it gives a way of relating features within and across the multiple tables.

However, two drawbacks of this approach are worth noting:

- It does not provide a summary of the relationship between the sets of variables defining the tables—it can only relate pairs of variables.

¹ Though sampling is still ongoing.

- If some tables have many more variables than others, they can dominate the resulting ordination.

These limitations are addressed by CCA and MFA, discussed in sections CCA and MFA, respectively.

We provide one geometric and one statistical motivation for PCA. The geometric motivation is that, if each row x_i of X is viewed as a point in p -dimensional space, then the principal component directions provide the best K -dimensional approximation to the data. The second interpretation is that PCA finds a low-dimensional representation of the x_i such that the resulting points have maximal variance. Qualitatively, this is a desirable property, because it means that the simpler representation preserves most of the variation present in the original data.

PCA is a very widely used technique, and some standard references include Mardia et al. (1980), Friedman et al. (2001), and Pagés (2014). Nonetheless, it is not ideal in the multitable setting.

Example

Figure 1 illustrates this approach on body composition and bacterial abundance data from the WELL-China study. Note that we have subsetted to only women, since men and women have very different body compositions, and we have slightly more data for women. Further, the 16s-rRNA data have been variance stabilized according to the methodology proposed in Anders and Huber (2010) and filtered to only those species that have count ≥ 5 in at least 7% of samples.

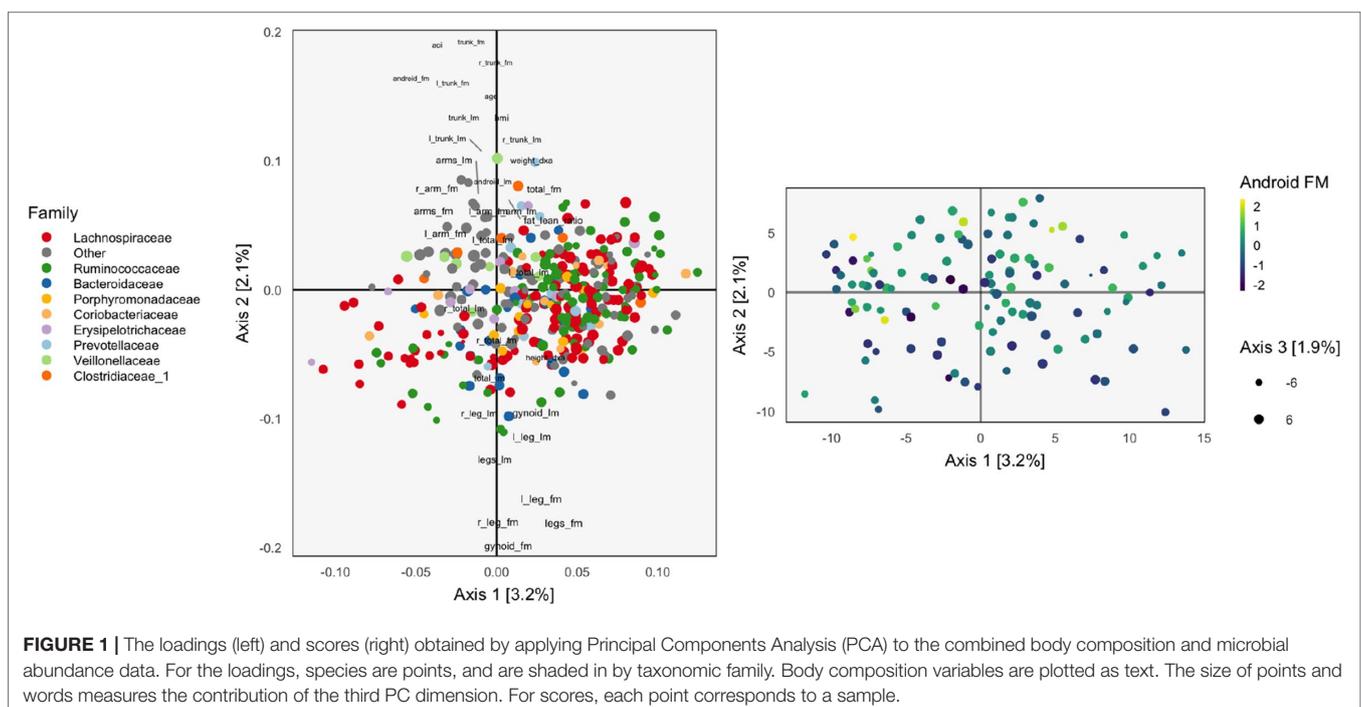
The left panel of **Figure 1** displays the loadings associated with this concatenated PCA approach, where body composition

(36 columns) and 16s-rRNA abundances (372 columns) were combined into one dataset (408 columns). Columns associated with bacterial species are displayed as points, shaded by taxonomic family, while columns associated with body composition variables are labeled with text. Note that the fraction of variance explained by each axis is on the order of a few percent—this is to be expected, considering that the baseline proportion would be $\frac{1}{408} \approx 0.25\%$ in the orthogonal case.

Most body composition variables lie close to the vertical axis, in a direction approximately orthogonal to the main direction of variation among species. Columns that are highly correlated—e.g., right (R) and left (L) leg fat mass (FM)—have loadings nearly equal to one another. Among species, the most notable pattern is the concentration of Ruminococcaceae on the right.

To identify relationships between species and body composition variables, it would be of interest to isolate those species with large contributions along the axis defined by linking the center of the variables and the origin. Relatively few such species stand out, though note that there is nothing in this algorithm's objective that would seek covariation across tables directly, so the fact that such associations seem weak with respect to the top two principal components does not mean such relationships do not exist.

We can study individual samples with respect to these loadings, by plotting their projections onto the top two principal components. This is the content of the right panel of **Figure 1**, which displays samples in the same positions, but shaded by android (i.e., abdominal) fat mass. This shading



confirms the observations from the loadings directly using observed data. Indeed, the increasing android fat mass among samples in the top of the scores in that panel exactly corresponds to the fact that related variables lie at the top in the left panel.

In this approach, the loadings provide a description of the relationship between variables across datasets. Further, scores summarize variation in samples across multiple datasets. Hence, this heuristic is a natural first step in analyzing multiple table data. However, considering the difficulty in directly interpreting the covariation across datasets, as well as the method's failure to use any sense of covariation in the dimensionality reductions strategy, suggests that this method should not be the last step of an analysis workflow. Nevertheless, we now have a baseline with which to compare the more elaborate methods of subsequent sections.

CCA

CCA is a close relative of PCA, designed to compare sets of features across tables. Like PCA, it provides low-dimensional representations of observations, but it also allows comparisons at the table level. Suppose for now that there are only two tables of interest, $X \in \mathbb{R}^{n \times p_1}$ and $Y \in \mathbb{R}^{n \times p_2}$. Let $\hat{\Sigma}_{XX}, \hat{\Sigma}_{YY}$, and $\hat{\Sigma}_{XY}$ be the associated covariance estimates. Take the SVD, $\hat{\Sigma}_{XX}^{-\frac{1}{2}} \hat{\Sigma}_{XY} \hat{\Sigma}_{YY}^{-\frac{1}{2}} = \tilde{U} D \tilde{V}^T$. The canonical correlation directions associated with the two tables are $u_k \sum_{XX}^{-\frac{1}{2}} \tilde{u}_k \in \mathbb{R}^{p_1}$ and $v_k = \sum_{YY}^{-\frac{1}{2}} \tilde{v}_k \in \mathbb{R}^{p_2}$. These directions give two sets of low-dimensional representations for each sample, one for each table: $z_k^{(1)} = X u_k \in \mathbb{R}^n$ and $z_k^{(2)} = Y v_k \in \mathbb{R}^n$. If the two tables are closely related, then the $z_k^{(1)}$ and $z_k^{(2)}$ will be very correlated. The singular values d_k are called the canonical correlation coefficients. Like the eigenvalues in PCA, they characterize the amount of covariation across tables that can be captured by each additional pair of directions.

As with PCA, there are many ways to view this procedure—here we discuss geometric, statistical, and probabilistic interpretations. Unlike the geometric interpretation of PCA, the geometric interpretation for CCA identifies point locations with features, not samples. Specifically, the columns of X and Y are thought of as points in \mathbb{R}^n . Consider two subspaces spanning the columns of X and Y , respectively. These subspaces correspond to the linear combinations of features within each table. Place two ellipses on the respective subspaces, centered at the origin and with size and shape depending on the within-table covariances $\hat{\Sigma}_{XX}$ and $\hat{\Sigma}_{YY}$. The first canonical correlation directions are the pair of points, one lying on each ellipse, such that the angle from the origin to those two points is smallest. In this sense, it finds a pair of variance-constrained linear combinations of features within the two tables such that the two combinations appear “close” to one another. The second pair of canonical correlation directions identify a pair of points with a similar interpretation, except they are required to be orthogonal to the first pair, with respect to the inner product induced by the covariances in each table.

For a statistical interpretation, the idea of CCA is to find the low-dimensional representations of the two tables with maximal

covariance—this is analogous to the maximum variance interpretation. Formally, rows of the two tables are imagined to be i.i.d. draws from \mathbb{P}^{XY} , which has marginals \mathbb{P}^X and \mathbb{P}^Y . Consider arbitrary linear combinations $z_i^{(1)}(u) = u^T x_i$ and $z_i^{(2)}(v) = v^T y_i$ of samples from the two tables. The first pair of CCA directions u_i^* and v_i^* are chosen to optimize

$$\begin{aligned} & \text{maximize}_{u \in \mathbb{R}^{p_1}, v \in \mathbb{R}^{p_2}} \text{Cov}_{\mathbb{P}^{XY}} [z_i^{(1)}(u), z_i^{(2)}(v)] \\ & \text{subject to } \text{Var}_{\mathbb{P}^X} (z_i^{(1)}(u)) = 1 \\ & \text{Var}_{\mathbb{P}^Y} (z_i^{(2)}(v)) = 1 \end{aligned} \tag{1}$$

To produce subsequent directions, the same optimization is performed, but with the additional constraint that the directions must be orthogonal to all the previous directions identified for that table. Of course, in actual applications, we estimate these covariances and variances empirically.

This perspective makes it easy to derive the algorithm given at the start of this section. The empirical version of the optimization problem (1) is

$$\begin{aligned} & \text{maximize}_{u \in \mathbb{R}^{p_1}, v \in \mathbb{R}^{p_2}} u^T \hat{\Sigma}_{XY} v \\ & \text{subject to } u^T \hat{\Sigma}_{XX} u = 1 \\ & v^T \hat{\Sigma}_{YY} v = 1. \end{aligned} \tag{2}$$

Consider the transformed data, $\tilde{u} = \hat{\Sigma}_{XX}^{-\frac{1}{2}} u$ and $\tilde{v} = \hat{\Sigma}_{YY}^{-\frac{1}{2}} v$. The optimization can be now be expressed as

$$\begin{aligned} & \text{maximize}_{\tilde{u} \in \mathbb{R}^{p_1}, \tilde{v} \in \mathbb{R}^{p_2}} \tilde{u}^T \hat{\Sigma}_{XX}^{-\frac{1}{2}} \hat{\Sigma}_{XY} \hat{\Sigma}_{YY}^{-\frac{1}{2}} \tilde{v} \\ & \text{such that } \|\tilde{u}\|_2 = 1 \\ & \|\tilde{v}\|_2 = 1. \end{aligned} \tag{3}$$

The optimal \tilde{u}_1 and \tilde{v}_1 for this problem are well known—they are exactly the first left and right eigenvectors of $\hat{\Sigma}_{XX}^{-\frac{1}{2}} \hat{\Sigma}_{XY} \hat{\Sigma}_{YY}^{-\frac{1}{2}} = \tilde{U} D \tilde{V}^T$, respectively.

A probabilistic interpretation of this procedure views it as estimating the factors in an implicit latent variable model. In particular, (Bach and Jordan, 2005) supposes that x_i and y_i are drawn i.i.d. from the model,

$$\begin{aligned} \xi_i & := (\xi_i^S, \xi_i^X, \xi_i^Y) \sim \mathcal{N}(0, Id) \\ x_i | \xi_i & \sim \mathcal{N}(\mu_x + W_X \xi_i^S + B_X \xi_i^X, I_d) \\ y_i | \xi_i & \sim \mathcal{N}(\mu_y + W_Y \xi_i^S + B_Y \xi_i^Y, I_d) \end{aligned}$$

That is, each sample is associated with a d -dimensional latent variable ξ_i , drawn from a spherical normal prior. A few of the coordinates of these latent variables, ξ_i^S , contribute to shared structure, through W_X and W_Y . The remaining coordinates model table-specific structure, through B_X and B_Y . It can be shown that the posterior expectations of the latent ξ_i^S given the observed tables must lie on the subspace defined by the CCA directions.

Example

We next apply CCA to the WELL-China body composition and microbiome data, with particular interest in how the results compare with those of section Example. We provide analogous loadings and scores plots in **Figure 2**. However, note that the data are not quite the same between the two analysis—we have filtered down to species passing a filter, which reduces the number of species to 66, from 2,565. This very aggressive filtering is necessary because CCA requires estimation of covariances matrices, and Σ_{XX} , Σ_{XY} , and Σ_{YY} , which is impossible for $p > n$ and highly unstable when p is a large fraction of n . Besides this stronger filtering, all preprocessing steps remain the same as in section Example.

The left panel of **Figure 2** provides the analog of CCA loadings. To be precise, let $X \in \mathbb{R}^{102 \times 36}$ be the matrix of body composition measurements and $Y \in \mathbb{R}^{102 \times 66}$ be the variance-stabilized microbial abundances. As before, write $u_k \in \mathbb{R}^{36}$, $v_k \in \mathbb{R}^{66}$ for the k^{th} canonical correlation directions. Text labels from column j of the body composition variables are displayed at position $(u_{j1}, u_{j2})_{j=1}^{36}$ and shaded points for the j^{th} species at position $(v_{j1}, v_{j2})_{j=1}^{66}$.

As in the concatenated PCA, we find that the groups of variables occupy separate spaces. Our interpretation is that sequences further to the left are correlated with the body variables further to the left, which are all in some way variants of body mass. Note that age is negatively correlated with total fat mass, which is why it appears on the opposite end. Among the abundant species that remain, there is limited clustering according to taxonomic group, though the Bacteroidaceae and Ruminococcus do appear restricted to the bottom right and left, respectively.

In the right panel of **Figure 2**, we plot the corresponding scores. Note that in CCA, there are two sets of scores for each k , the Xu_k and Yv_k . Indeed, the CCA objective finds directions that maximize the correlation between these scores. We use a different color legend

for the two panels, each of which represents one set of scores. The legend for scores from species abundances are colored by family, while those for the body composition associates samples with android fat mass. The pairs of scores for each individual sample are drawn with small links. Since most links are relatively short, linear combinations of the two tables could be found that optimized the objective—indeed, the top two canonical correlations are 0.968 and 0.957. However, some caution is necessary here, and a more honest evaluation would be based on scores obtained by projecting new samples onto the original CCA directions. This is especially important in this nearly high-dimensional setting, where covariance estimation may be unreliable.

Aside from the fact that samples appear as pairs, interpretation proceeds as in a PCA scores plot, as in **Figure 1**. The association between these variables and the sample positions is not as strong as when performing PCA on the combined table. This is to be expected, however, as PCA maximizes variance without any thought to covariance, and the body composition table alone has a large portion of its variance related to android fat mass.

Co-Inertia Analysis

Co-inertia Analysis (CoIA) emerged in ecology to facilitate analysis of variation in species abundance as a function of environmental conditions (Dolédéc and Chessel, 1994). It can be viewed as a slight modification of CCA. Again, we seek sets of orthonormal directions $(u_k)_{k=1}^K$ and $(v_k)_{k=1}^K$ such that the associated projections Xu_k and Yv_k explain most of the covariation between the tables. Unlike CCA, CoIA finds its first directions by maximizing the covariance—not the correlation—between scores,

$$\text{maximize}_{u \in \mathbb{R}^{p^1}, v \in \mathbb{R}^{p^2}} u^T X^T Y v$$

$$\text{such that } \|u\| = 1$$

$$\|v\| = 1,$$

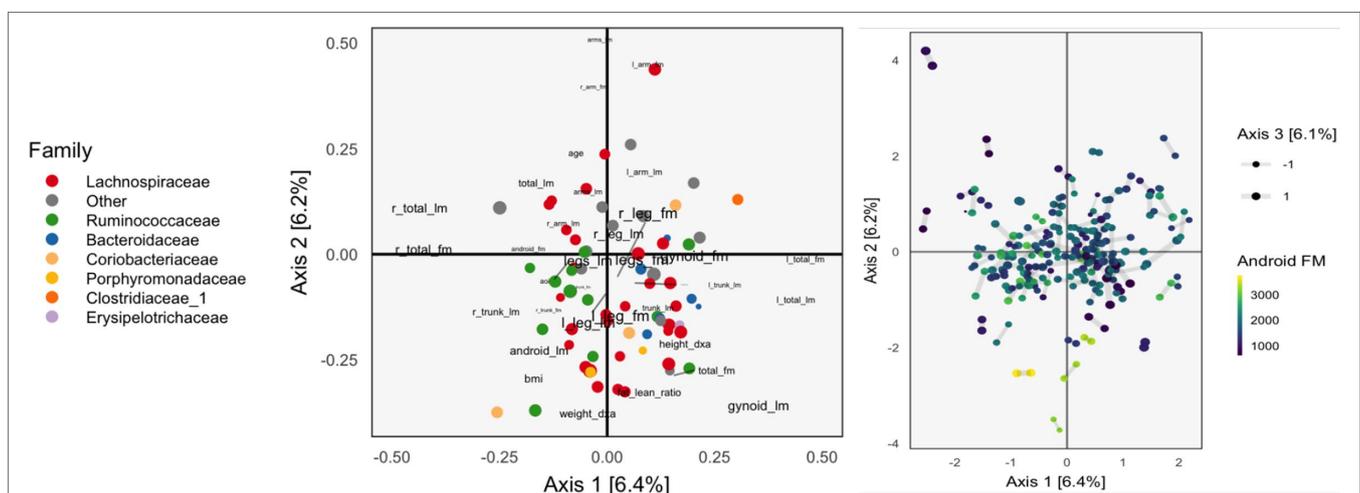


FIGURE 2 | The Canonical Correlation Analysis (CCA) analog of the PCA biplot in **Figure 1**, obtained by applying CCA to the combined body composition and microbial abundance data. Since each sample is associated with a pair of scores, one from each table, we use a different symbol to represent the scores: two points joined by an edge, where each point gives the score from one of the tables. Aside from this exception, the PCA biplot interpretation still applies. The higher the CCA objective, the shorter the links between pairs. The first two CCA dimensions suggest smooth variation across samples, according to amount of android fat mass.

with subsequent directions found by the same optimization, after adding the constraint that they are orthogonal to the previously derived directions.

The only difference with the objective in equation (2) is that norm constraint is imposed on u and v directly, rather than their transformations $\frac{1}{\lambda_X}u$ and $\frac{1}{\lambda_Y}v$. It is in this sense that the CCA objective maximizes the correlation between scores, while CoIA maximizes the covariance.

The solution $(u_k)_{k=1}^K$ and $(v_k)_{k=1}^K$ can be obtained as the first K left and right eigenvectors from the SVD of $X^T Y$, as opposed to the first K generalized eigenvectors, as in CCA. The proof of this fact is almost identical to the derivation in section CCA, for CCA.

Example

We apply CoIA to the same data as used in section Example, as CoIA also needs to estimate the covariance between tables, which is difficult when the number of species is large. We find that the associated scores are quite different from those found using CCA. Compare **Figure 3**, which shades samples by android fat mass with **Figure 2** for CCA. The scores for CoIA are not so closely aligned across tables, but they exhibit a clearer gradient across android fat mass. We find that the scores are not nearly as closely aligned as they are for CCA, but that they are more strongly associated with variation in android fat mass, as in the concatenated PCA result of **Figure 1**. It is not clear whether this phenomenon—the CoIA scores being more similar to those from PCA than CCA—holds in general, or what it is about the change in inner products between CoIA and CCA that is responsible for this difference.

MFA

MFA gives an alternative approach to producing scores and relating features across multiple tables (Pagés, 2014). It can be understood as a refined version of the concatenated PCA described in section PCA that reweights tables in a way that prevents any one table from dominating the resulting ordination. Specifically, MFA is a concatenated PCA on the matrix

$$X := \left[\frac{1}{\lambda_1(X^{(1)})} X^{(1)} \mid \dots \mid \frac{1}{\lambda_1(X^{(L)})} X^{(L)} \right],$$

which reweights each table $X^{(k)}$ by its largest eigenvalue, $\lambda(X^{(k)})$. This procedure is the multitable analog of the common practice of standardizing variables before performing PCA.

The resulting MFA directions and scores can be interpreted in the same way as those from PCA—the MFA directions still specify the relationship between measured features, and the position of each sample's projection describes the relative value of each feature for that sample. Moreover, MFA gives a way of comparing entire tables to each other, called a “canonical analysis” (Pagés et al., 2004). A K -dimensional representation of the l^{th} group is given by

$$\left[\mathcal{L}(z_1, X^{(l)}), \dots, \mathcal{L}(z_K, X^{(l)}) \right],$$

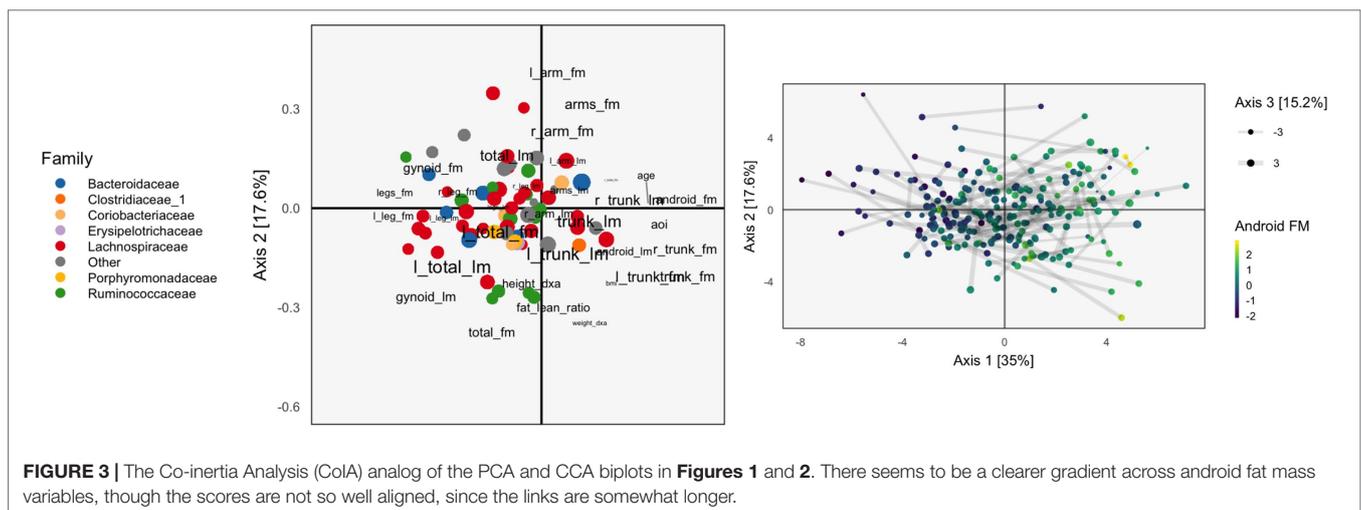
where $z_k = d_k u_k \in \mathbb{R}^n$ is the k^{th} column of principal component scores and

$$\mathcal{L}(z_k, X^{(l)}) = \frac{\lambda_k(X)}{\lambda_1(X^{(l)})} \text{tr}(X^{(l)} X^{(l)T} z_k z_k^T) = \frac{\lambda_k(X)}{\lambda_1(X^{(l)})} \|X^{(l)T} z_k\|_2^2$$

is a measure of aggregate similarity between the coordinates in the l^{th} table and the k^{th} column of scores. According to this definition, if the samples, as represented by the l^{th} table, have high correlation with the k^{th} dimension of scores, then the canonical analysis displays positions the l^{th} table far in the k^{th} direction. Plotting these table-level coordinates helps resolve which tables measure similar underlying variation.

PCA-IV

PCA-IV adapts the dimensionality reduction ideas of PCA to the multivariate regression setting (Rao, 1964). It can also be



viewed as a version of PCA that chooses a dimension reduction of X based on its ability to predict Y . In this sense, it anticipates methods like Partial Least Squares, Canonical Correspondence Analysis, the Curds & Whey procedure, and the Graph-Fused Lasso, which are described in sections Partial Least Squares, CCpnA, Curds & Whey, and Graph-Fused Lasso.

Formally, suppose we are predicting $y_i \in \mathbb{R}^{p_1}$ from $x_i \in \mathbb{R}^{p_2}$. Since p_2 may be large, it might be useful to work with a lower-dimensional representation $z_i = V^T x_i \in \mathbb{R}^K$, which is potentially more interpretable but still as (or more) predictive of y_i . As in PCA, we require that V be orthonormal.

The criterion that PCA-IV uses to identify the loadings V and scores Z mirrors the maximum variance criterion for PCA. Instead of choosing V to maximize the variance of the z_i , we choose it to minimize the residual covariance of y_i given z_i . That is, suppose that y_i and x_i are jointly normal with mean 0 and covariance

$$\text{Var}_{\mathbb{P}} \begin{pmatrix} y_i \\ x_i \end{pmatrix} = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix}.$$

If $z_i = V^T x_i$, then the joint covariance of y_i and z_i is

$$\text{Var}_{\mathbb{P}} \begin{pmatrix} y_i \\ z_i \end{pmatrix} = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX}V \\ V^T \Sigma_{XY} & V^T \Sigma_{XX}V \end{pmatrix},$$

so the residual covariance of y_i given z_i is

$$\Sigma_{YY} - \Sigma_{YX}V(V^T \Sigma_{XX}V)^{-1}V^T \Sigma_{XY}. \tag{4}$$

Rao (Rao, 1964) uses the trace to measure the “size” of this matrix. The true population covariances are unknown to us, so we replace them by their empirical estimates. The formal optimization for PCA-IV then becomes

$$\underset{V \in \mathbb{R}^{p_2 \times K} \text{ orthonormal}}{\text{minimize}} \quad \text{tr}(\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX}V(V^T \hat{\Sigma}_{XX}V)^{-1}V^T \hat{\Sigma}_{XY}) \tag{5}$$

The optimal V are the top K generalized eigenvectors of $\hat{\Sigma}_{XY} \hat{\Sigma}_{YX}$ with respect to $\hat{\Sigma}_{XX}$, that is, the orthonormal set of (v_k) satisfying

$$\hat{\Sigma}_{XY} \hat{\Sigma}_{YX} V = (\lambda_1 \hat{\Sigma}_{XX} v_1 \mid \dots \mid \lambda_K \hat{\Sigma}_{XX} v_K) = \hat{\Sigma}_{XX} V \Lambda,$$

where $\Lambda = \text{diag}(\lambda_k) \in \mathbb{R}^{K \times K}$. A derivation for why this choice is optimal is provided in section *Derivation Details for PCA-IV*.

For a geometric interpretation of PCA-IV, view each column y_j in Y and x_j in X as a point in \mathbb{R}^n . Assuming X and Y are full rank, the collections (y_j) and (x_j) span p_1 - and p_2 -dimensional subspaces. A set of independent regressions of y_j on X projects each individual y_j onto the span of the (x_j) , and the squared residuals are the distance to this subspace. The PCA-IV procedure is an attempt to find a further K -dimensional subspace within the span of the (x_j) such that the residuals of the regressions from y_j

onto this further subspace is not much worse. This is displayed in **Figure 4**.

Example

Continuing our WELL-China case study, we now illustrate results from PCA-IV. The idea of scores and loadings in this context requires some clarification. By PCA-IV scores, we mean the coordinates of projections z_i of samples onto the subspace defined by V , and by loadings, we mean the correlation between columns² of X and Y with the PCA-IV axes defining V .

The scores and loadings are given in **Figure 5**. Interpretation of the species loadings is simple, since species seem well separated by taxa. Interpretation of the body composition variables is less clear—pairs of variables that would be expected to be near to one another are not, in many cases. Indeed, leg fat mass (leg_fm) and left leg fat mass (l_leg_fm) should have a small angle between one another, but they do not. It is possible that by approximating the covariation across tables, the quality of within-table approximations deteriorates.

We find that the scores, displayed in figures, are similar to those that found by the concatenated PCA of section PCA. One possible explanation for this behavior is that the PCA-IV-generalized SVD of X is similar to an ordinary PCA of X , and that in the concatenated PCA of $(Y X)$, the fact that X has many more columns than Y means that the result is similar to a PCA on X alone.

Partial Triadic Analysis

Partial Triadic Analysis (PTA) gives an approach for working with multitable data when each table has the same dimension, $p_1 = p_2$ (Kroonenberg, 2008; Thioulouse, 2011). Specifically, it gives a way of analyzing data of the form $(X_{..l})_{l=1}^L$, where each $X_{..l} \in \mathbb{R}^{n \times p}$. This is called a data cube because it can also be written as a three-dimensional array $X \in \mathbb{R}^{n \times p \times L}$. We denote the j^{th} feature measured on the i^{th} sample in the l^{th} table by x_{ijl} , and the slices over fixed i, j , and l by $X_{i..}$, $X_{.j.}$, and $X_{..l}$. This type of data arises frequently in longitudinal data analysis, where the same features are collected for the same samples over a series of L times. However, the actual ordering of the L tables is not ever used by this method: if we scrambled the time ordering for L tables, the algorithm’s result would not change.

The main idea in PTA is to divide the analysis into two steps:

- Combine the L tables into a single compromise or consensus table.
- Apply any standard single-table method, e.g., PCA, on the compromise table.

A naive approach to constructing the compromise table would be to average each entry across the L tables. Instead, PTA upweights tables that are more similar to the average table, as these are considered more representative. Formally, the compromise is defined as $X_c = \sum_{l=1}^L \alpha_l X_{..l} = X \alpha \in \mathbb{R}^{n \times p}$, where α (constrained to norm one) is chosen to maximize $\sum_{l=1}^L \alpha_l \langle \bar{X}, X_{..l} \rangle$,

²Geometrically, the angle between original columns and the subspace, in the sense of **Figure 4**.

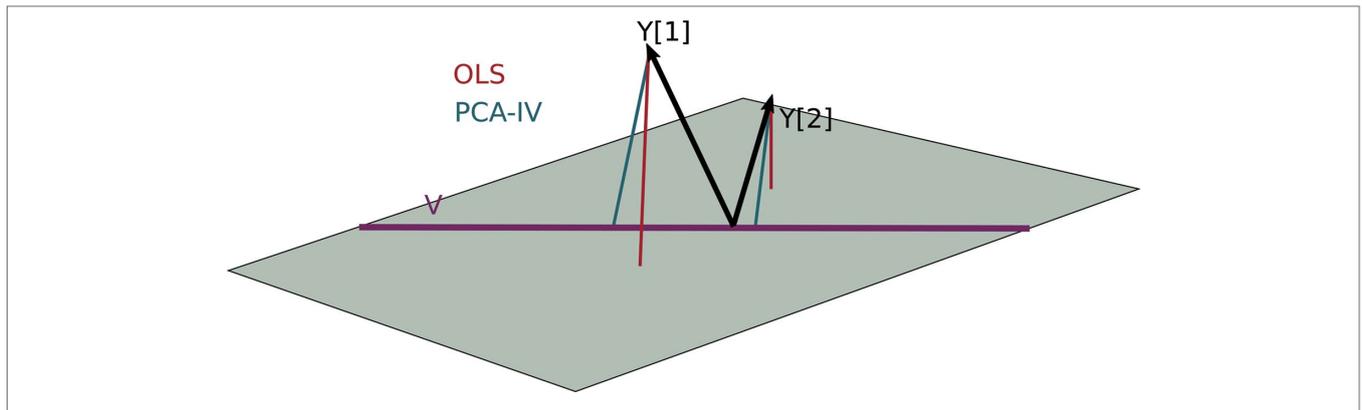


FIGURE 4 | A geometric view of Principal Component Analysis with Instrumental Variables (PCA-IV). The columns of the response Y are views as n -dimensional vectors. The gray plane is the span of X . Multivariate OLS simply projects the columns of Y onto the plane, while PCA-IV searches for a further subspace V on which to project all responses.

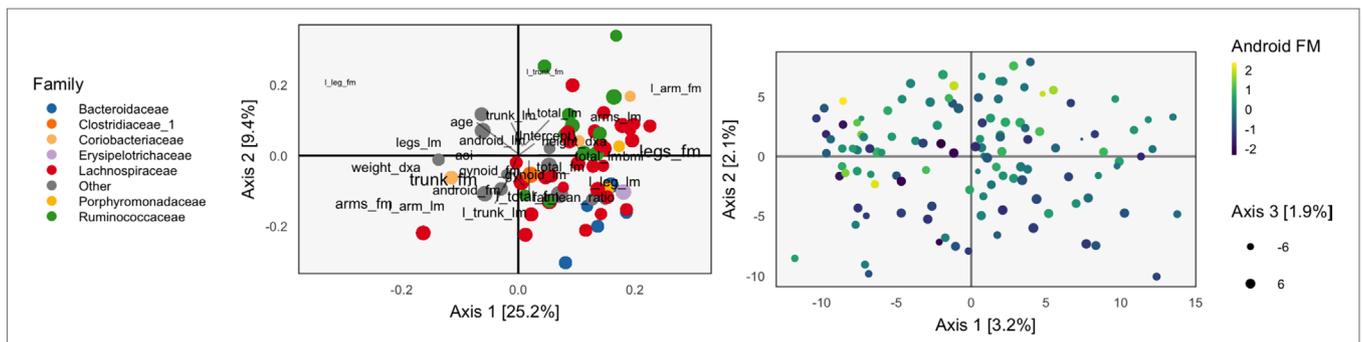


FIGURE 5 | The PCA-IV biplot can be interpreted like biplots from previous methods, for example, **Figure 1**. Some of the relationships between variables seem less intuitive than those observed previously.

a weighted average of inner-products³ between each of the L tables and the naive-average table, $\bar{X} = \frac{1}{L} \sum_{l=1}^L X_{..l}$.

The optimal α can be derived using Lagrange multipliers (see *Derivation of PTA α*) and leads to the compromise table,

$$X_c = \sum_{l=1}^L \frac{\langle \bar{X}, X_{..l} \rangle}{\sqrt{\sum_{r=1}^L \langle \bar{X}, X_{..r} \rangle^2}} X_{..l}$$

We can try to interpret the compromise matrix geometrically. Suppose the $X_{..l}$ define an orthonormal basis, so that $\langle X^l, X^{l'} \rangle = \mathbb{1}(l=l')$. Then, we can write the compromise table as

$$X_c = \sqrt{L} \sum_{l=1}^L \langle \bar{X}, X_{..l} \rangle X_{..l} = \sqrt{L} \bar{X},$$

a scaled version of the mean.

³We are using $\langle A, B \rangle = \text{tr}(A^T B)$.

If, however, the tables are not orthonormal, then we place more weight on directions that are correlated. For example, if $X^{(1)} = X^{(2)}$, but the rest of the tables are orthogonal to each other and to these first two tables, then the compromise double counts the direction $X^{(1)}$. Therefore, compared to the naive average \bar{X} , X_c upweights more highly represented tables.

Statico and Costatis

In the multivariate ecology literature, it is common to have a pair of data cubes, giving species abundances and environmental variables over time, respectively. We write these as $Y \in \mathbb{R}^{n \times p_1 \times L}$ and $X \in \mathbb{R}^{n \times p_2 \times L}$. Costatis and Statico are two approaches for analyzing such data (Thioulouse, 2011). They are easiest to understand as divide-and-conquer approaches, where the general problem of analyzing a pair of data cubes is divided into two steps, one designed for analyzing individual cubes, and another for studying covariation across tables. In Statico, the covariation problem is dealt with first, then followed by a data cube analysis, while in Costatis, that order is reversed.

Specifically, in Statico, an empirical cross-covariance matrix is constructed at each time point, $Z^l = \frac{1}{n_l} Y_{..l}^T X_{..l}$. For example, this is the correlation between the environmental variables and species counts at a specific time point l . The L matrices Z^l are then

input into a PTA, yielding a compromise table Z_c that can then be studied with PCA.

Alternatively, in Costatis, a compromise table is constructed for each of the data cubes Y and X , using PTA. Call these Y_c and X_c . These are now simply two matrices, each with n rows, and they can be analyzed by any two-table dimensionality reduction method, for example, CoIA.

Hence, we see that the only difference between these methods is the order in which CoIA and PTA are applied. Indeed, this is reflected in the names of the methods: Statis is an abbreviation for a PTA, and Statico performs a CoIA before a Statis while Costatis does the reverse.

MODERN MULTIVARIATE METHODS

Compared to classical approaches, modern multivariate methods are typically designed for more high-dimensional, heterogeneous settings. The two methods reviewed in this section are examples of this trend: Partial Least Squares (PLS) is well-suited for finding predictors in the presence of high-dimensional response matrices, while Canonical Correspondence Analysis (CCpNA) was designed to facilitate joint analysis of heterogeneous continuous and count data necessary. Unlike traditional statistical methods, neither approach is explicitly model-based, and both are iterative, requiring more extensive computation than earlier techniques.

Partial Least Squares

PLS sequentially derives a set of mutually orthogonal features $(z_k)_{k=1}^K$ that characterizes the relationship between two tables, Y and X (Wold, 1985). To obtain the first PLS direction, z_1 , compute the first left singular vector u_1 of the cross-covariance matrix between the two tables, $\hat{\Sigma}_{YX} = \frac{1}{n} Y^T X$. Then, for each of the p_2 columns of X , compute the univariate (i.e., partial) regression coefficient $\hat{\phi}_j = \frac{1}{\|x_{\cdot j}\|_2} x_{\cdot j}^T u_1$, for $j = 1, \dots, p_1$. The first PLS direction is defined

as $z_1 = \sum_{j=1}^{p_1} \hat{\phi}_j x_{\cdot j}$ a weighted average of $x_{\cdot j}$ according to their partial correlation with u_1 . To generate subsequent directions z_k , orthogonalize both Y and X with respect to the current directions z_1, \dots, z_{k-1} , and repeat the process.

This procedure is appealing because, like PCA, it reduces a potentially high-dimensional matrix X with many correlated columns into a smaller set of orthogonal directions. Moreover, it achieves this reduction in a way that accounts for correlation with columns in Y : columns of X that are uncorrelated with Y will have no contribution to the PLS directions, even if they account for a large proportion of variation in X .

We have stated the procedure in the form it was originally proposed, but this algorithmic description is difficult to understand geometrically or probabilistically. However, interpretational aids have since been developed. Frank and Friedman (1993) and Stone and Brooks (1990) studied the case where $p_1 = 1$, so y is a single column vector. By assuming that the rows of y and X are drawn i.i.d. from distribution \mathbb{P}^{YX} , with marginals \mathbb{P}^Y and \mathbb{P}^X ,

they found that the k^{th} PLS direction z_k is the z that solves the optimization

$$\begin{aligned} & \underset{z}{\text{maximize}} \text{Corr}_{\mathbb{P}^{YX}} [x_i^T z_k, y_i] \text{Var}_{\mathbb{P}^X} (z^T x_i) \\ & \text{such that } z^T X^T X z_j = 0 \text{ for all } j \leq k-1 \\ & \|z\|_2 = 1. \end{aligned} \quad (6)$$

If the covariance term is omitted, the optimization is identical to the maximum variance problem that gives the principal component directions based on X . This formulation makes precise the idea that PLS is a version of principal components that accounts for correlation with Y .

An alternative interpretation, due to (Gustafsson, 2001), is that PLS fits a particular latent variable model. Suppose $\xi_i = (\xi_i^s, \xi_i^X)$ are drawn i.i.d. from a $K_1 + K_2 = K$ dimensional spherical normal. PLS assumes the observed tables Y and X have rows drawn i.i.d. from

$$\begin{aligned} y_i | \xi_i & \sim \mathcal{N}(\mu_Y + W_Y \xi_i^s, \sigma^2 I_{p_1}) \\ x_i | \xi_i & \sim \mathcal{N}(\mu_X + W_X \xi_i^s + B_X \xi_i^X, \sigma^2 I_{p_2}). \end{aligned}$$

That is, each table is the sum of two components, one that is a table-specific linear combination of a shared latent variable, and another that is an arbitrary linear combination of a table-specific latent variable. The shared feature ξ^s is the object of interest, and is what PLS implicitly estimates.

Sparse Partial Least Squares

PLS suffers from two of the same problems as PCA:

- It can be unstable in high-dimensional settings, since it requires estimation of covariances, and isn't well defined when $p > n$.
- PLS directions are linear combinations of all features in x_p , which can be difficult to interpret when there are many features.

Different regularized, sparse modifications of PCA have been proposed to remedy these issues in the PCA context (Jolliffe et al., 2003; Zou et al., 2006; Witten et al., 2009). For PLS, similar analysis leads to sparse PLS (Lê Cao et al., 2008; Chun and Kele, 2010), and we briefly review this method here.

Directly regularizing the multiresponse version of the PLS optimization (6) leads to the problem

$$\begin{aligned} & \underset{z_k}{\text{maximize}} \sum_{j=1}^{p_1} \text{Cov}_{\mathbb{P}^{YX}} [x_i^T z_k, y_{ij}] \\ & \text{such that } z^T X^T X z_j = 0 \text{ for all } j \leq k-1 \\ & \|z_k\|_2 = 1 \\ & \|z_k\|_1 \leq \lambda, \end{aligned}$$

which can be applied to real data by replacing the objective with its sample version, $z_k^T M z_k$, where $M = X^T Y Y^T X$. This version

of the problem falls into the Penalized Matrix Decomposition framework of Witten et al. (2009), reviewed in the section penalized matrix decomposition.

However, Chun and Kele (2010) argue that this formulation does not lead to “sparse enough” solutions. Instead, they adapt the SPCA approach of Zou et al. (2006) to PLS. The resulting objective identifies two sets of directions, a set (a_k) that maximizes the PLS-defining covariance and another, (z_k), that approximates the first set by a sparser alternative. Formally,

$$\begin{aligned} \underset{z_k, a_k}{\text{maximize}} & -\kappa \|a_k\|_M^2 + (1-\kappa) \|z_k - a_k\|_M^2 \\ \text{such that} & \|a_k\|_2^2 = 1 \\ & \|z_k\|_1 \leq \lambda_1 \\ & \|z_k\|_2 \leq \lambda_2, \end{aligned} \quad (7)$$

where we have defined $\|x\|_M = \sqrt{x^T M x}$ and κ , λ_1 , and λ_2 are tuning parameters. The first term in the objective is the PLS-defining covariance, the second ensures that the solutions z_k and a_k are similar, and the norm constraints induce sparsity and stability on z_k . Note that while this objective is not convex, for fixed a_k , it is an elastic-net regression, while for fixed z_k , it is a type of eigenvalue problem.

Example

Next we apply the sparse partial least squares (SPLS) implementation of Chung et al. (2012) to the WELL-China body composition data. We use the body composition variables as the response Y and the microbiome community composition as X . In this direction, a well-fitting model would allow the microbiome community measurements X to serve as a proxy for the variables in Y , in case those data were not easily accessible. To an extent, however, this choice of directionality is arbitrary—regressing abundances on body composition variables would also be sensible—and reflects the basic limitations of using an asymmetric method to study a symmetric problem.

We subset to female subjects and filter species, keeping only those species with a count of at least 5 in at least 7% of samples. This leaves 372 species over 119 participants. All species abundances are variance-stabilized using the approach of Anders and Huber (2010). We cross-validate with five folds, searching through a grid over $K \in \{4, \dots, 8\}$ and $\lambda_1 \in \{0, 0.05, \dots, 0.7\}$. This grid is used to prevent the model from regularizing to the point that there is no information to visualize. For example, if we set $K = 1$, every row of **Figure 6** would look identical. The predictive accuracy is poor, which is unsurprising considering the spike at 0 in the abundances histogram—the held out error is ≈ 1.29 , after having scaled and centered the body composition variables.

Figure 6 displays fitted coefficients relating body composition variables with species abundances. By fitted coefficients, we mean we display $\hat{B} = ZQ^T$, where Z are the SPLS directions and a multiresponse linear regression model is used. Specifically, $Y = XB + E = XZQ^T + E$ where X is a matrix with rows x_i , Y is a matrix with columns y_j , and Z is a matrix with columns z_k .

Positive associations tend to occur across all responses simultaneously, while negative associations can be unique to either lean or fat mass. Most taxonomic families seem to have slightly more negative than positive associations, with the possible exception of Porphyromonodaceae.

To interpret these coefficients in the raw data, we can visualize individual species with strong associations to body composition. Specifically, we study associations with the android and gynoid fat mass variables. In the left panel of **Figure 7**, we display the abundances X for species against android fat mass, respectively. The species are chosen according to whether the two-dimensional coefficient across android and gynoid fat mass has large norm⁴. The main associations that are visible are those between the body composition and species presence or absence. That is, there don't seem to be any cases where a body composition feature varies smoothly as a species becomes more or less abundant. Instead, SPLS has identified species whose samples have lower or higher android or gynoid fat mass, depending on whether that species is present or absent.

CCpna

CCpna is a method, originally developed in ecology, useful for joint analysis of count and continuous data. The canonical application has a site-by-species count matrix $Y \in \mathbb{R}^{n \times p_1}$ and an environmental features matrix $X \in \mathbb{R}^{n \times p_2}$, for example, historical rainfall and temperature measurements. In the WELL context, Y would be the samples by community abundance matrix, while X would contain the body composition measurements.

The scientific goal might be to identify species that are more abundant in sites with more rainfall or higher temperature. If these environmental variables were uncorrelated, it would be enough to fit a separate regression to each. This, however, is rarely the case, motivating the development for CCpna.

Translating to the language of the WELL-study, individual samples can be thought of sites, and the supplemental data—that is, the body composition variables—are analogous to environmental variables.

CCpna produces low-dimensional representations of both the rows and columns of Y (the samples and species), along with latent subspaces on which these representations are defined. Algorithmically, CCpna first constructs the following matrices, where 1_r denotes a column vector of r ones,

1. An overall frequency matrix,

$$F = \frac{1}{n_{\cdot}^Y} Y,$$

where n_{\cdot}^Y is the sum of all counts in matrix Y .

2. A diagonal matrix of row (site) proportions,

$$D_r = \text{diag}(F_{p_1}^1) \in \mathbb{R}^{n \times n}.$$

⁴Specifically, $\left\| \begin{pmatrix} \beta_{\text{android}} \\ \beta_{\text{gynoid}} \end{pmatrix} \right\|_2 > 0.065$.

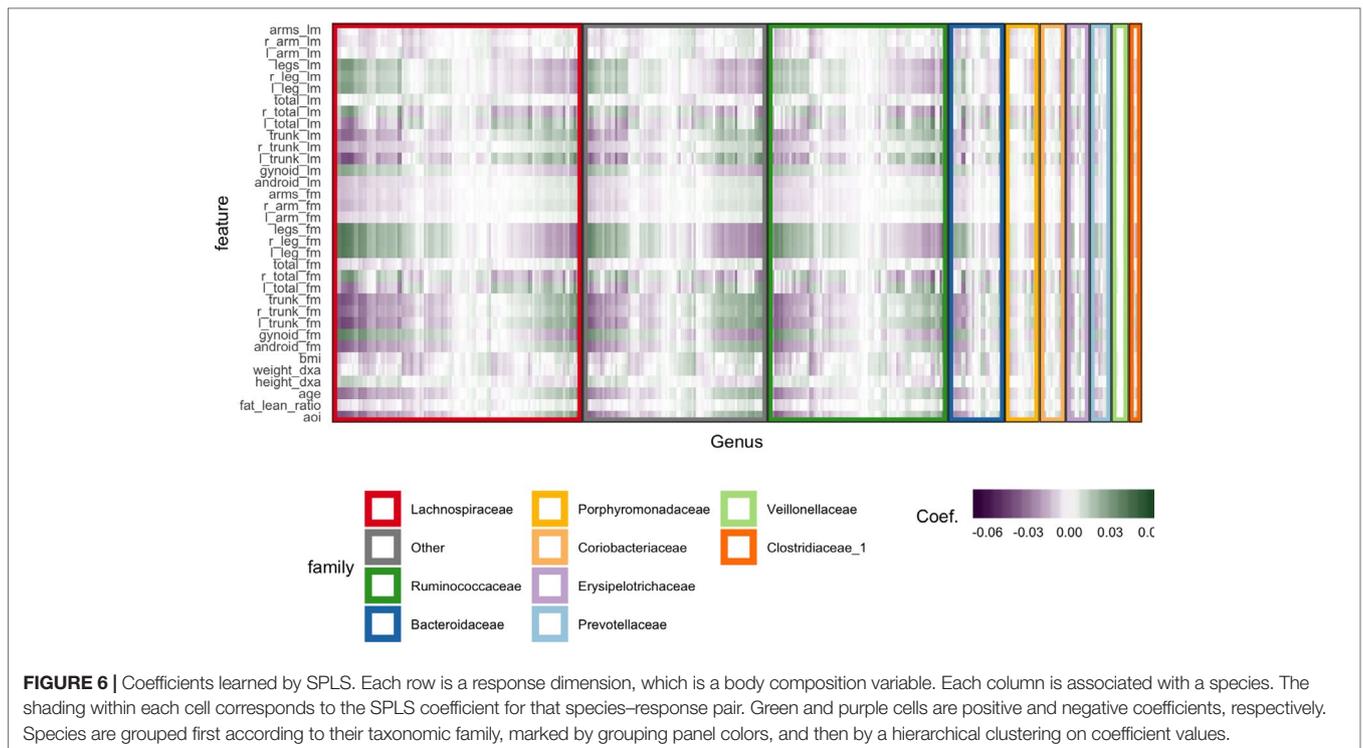


FIGURE 6 | Coefficients learned by SPLS. Each row is a response dimension, which is a body composition variable. Each column is associated with a species. The shading within each cell corresponds to the SPLS coefficient for that species–response pair. Green and purple cells are positive and negative coefficients, respectively. Species are grouped first according to their taxonomic family, marked by grouping panel colors, and then by a hierarchical clustering on coefficient values.

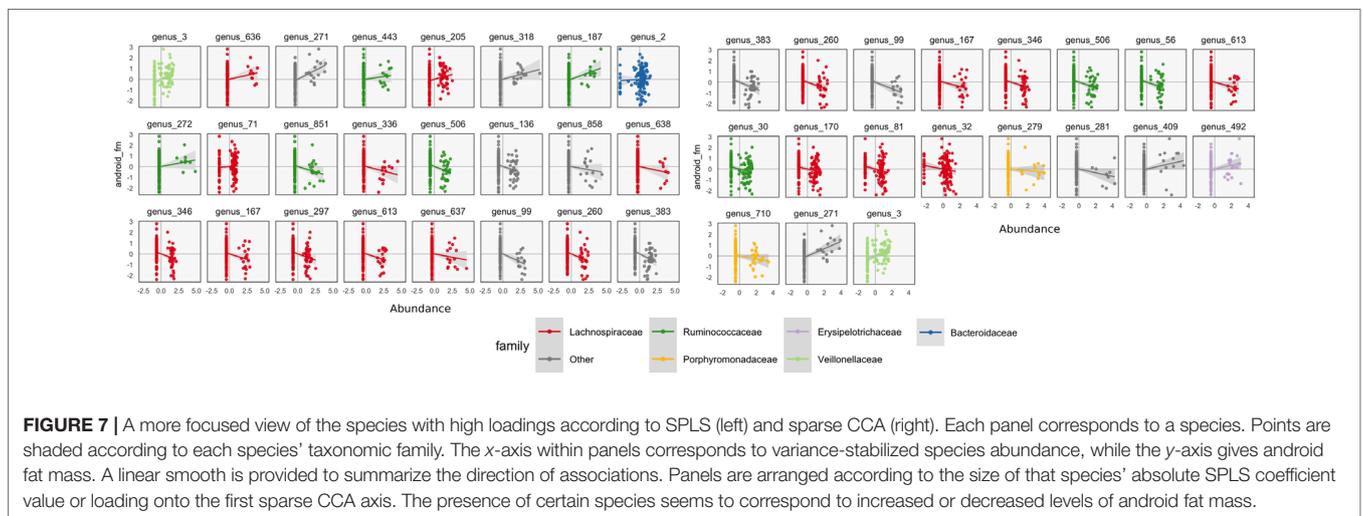


FIGURE 7 | A more focused view of the species with high loadings according to SPLS (left) and sparse CCA (right). Each panel corresponds to a species. Points are shaded according to each species' taxonomic family. The x-axis within panels corresponds to variance-stabilized species abundance, while the y-axis gives android fat mass. A linear smooth is provided to summarize the direction of associations. Panels are arranged according to the size of that species' absolute SPLS coefficient value or loading onto the first sparse CCA axis. The presence of certain species seems to correspond to increased or decreased levels of android fat mass.

3. A diagonal matrix of column (species) proportions,

$$D_c = \text{diag}(F^T \mathbf{1}_n) \in \mathbb{R}^{p_1 \times p_1}.$$

4. A projection onto the columns of the supplemental matrix X , reweighting samples according to their species counts,

$$P_X = D_r^{-1/2} X (X^T D_r X)^{-1} X^T D_r^{-1/2} \in \mathbb{R}^{n \times n},$$

$$D_r^{-1/2} = (F - F \mathbf{1}_{p_1} \mathbf{1}_{p_1}^T F) D_c^{-1/2} P_X = USV^T,$$

and define row and column scores Z and Q by

$$Z = D_r^{-1/2} US$$

$$Q = D_c^{-1/2} V^T S.$$

With this notation, compute an SVD,

There are several ways to interpret this procedure. CCpNA was originally proposed as the solution to a fixed-point

iteration called reciprocal averaging (Ter Braak, 1986). Later, Greenacre (1984) and Greenacre and Hastie (1987), provided a geometric view and Zhu et al. (2005) gave an exact probabilistic interpretation.

The intuition for the reciprocal averaging procedure is simple: the scores for different samples should be a weighted average of the species scores, with larger weights for the species that are more common at those sites. Similarly, species scores can be defined according to a weighted average of sample scores. That is,

$$z_i \propto \frac{1}{f_i} \sum_{j=1}^{p_1} f_{ij} q_{ij}$$

$$q_i \propto \frac{1}{f_j} \sum_{i=1}^n f_{ij} z_{ij},$$

or, in matrix form,

$$Z \propto \text{diag}(F1p_1)^{-1} FQ^T$$

$$Q \propto \text{diag}(F^T 1_n)^{-1} Z.$$

This formulation suggests an algorithm for finding Z and Q —arbitrarily initialize one and iterate these calculations until convergence.

As is, this is not yet the setup that yields CCpnA—it does not use information in the supplemental table X . To recover CCpnA, a projection step needs to be inserted before the calculation of row scores,

1. Arbitrarily initialize Z .
2. While not converged,
 - a. Solve $Q' \propto \text{diag}(F^T 1_n)^{-1} F^T Z$.
 - b. Project $Q = P_X Q'$.
 - c. Solve $Z \propto \text{diag}(Z1_{p_1})^{-1} FQ^T$.

The fixed point of this iteration is the previously described CCpnA solution.

A second interpretation is due to Zhu et al. (2005). Suppose first that we are only interested in a one-dimensional score for rows and columns. Let α be a latent gradient, for example, between warm-dry and cold-wet sites, or low and high android-fat mass samples. For each of the p_1 species, define a normal density over the supplemental variables, $f_j(x_i) = \mathcal{N}(x_i | \mu_j, \Sigma_j)$. The mode of this density represents the preferred environment for species j . Next, project these densities onto the gradient, giving a univariate $f_j^\alpha(z_i) = \mathcal{N}(z_i | \alpha^T \mu_j, \alpha^T \Sigma_j \alpha)$ for each species. The z_i represent the scores for species i along the gradient α .

The generative model views species–sample pairs one at a time. For each pair involving sample i and species j , draw a score according to $f_j^\alpha(z_i)$. Hence, each site i draws species according to a p_1 -class linear discriminant (LDA) model.

To use this idea to compute scores, we need to estimate the gradient α , which is also of interest in its own right. This is done by supposing equal covariances across species, $\Sigma_j = \Sigma$ for all j ,

and finding the $\hat{\alpha}$ maximizing the between vs. total variance across species,

$$\frac{\alpha^T \Sigma_B \alpha}{\alpha^T \Sigma \alpha},$$

where

$$\Sigma_B = \sum_{j=1}^{p_1} f_{\cdot j} (\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^T$$

is a between-species covariance matrix. Estimating $\hat{\alpha}$ in this way and writing $z_i = \hat{\alpha}^T x_i$ gives the original site scores from CCpnA.

We have omitted a detailed numerical example of this method in this review, but note that codes for applying this method are available in the github repository associated with this review.

Penalized Matrix Decomposition

In high-dimensional settings, sparsity is a desirable property, for both qualitative interpretability and statistical stability. A regression model using only a few features is easier to understand than one involving a linear combination of all possible features. Further, regularized models typically outperform their unregularized counterparts in terms of both predictive accuracy and inferential power (Buhlmann and Van De Geer, 2011). In fact, it is impossible to fit an unregularized linear regression when the number of features is greater than the number of samples.

The Penalized Matrix Decomposition (PMD) is a general approach to adapting the regularization machinery developed around regression to the multivariate analysis setting (Witten et al., 2009). The CCA and MultiCCA instances of PMD have been particularly well-studied (Witten et al., 2009; Witten et al., 2013).

The general setup is as follows. Suppose we want a one-dimensional representation of the samples (rows) in $X \in \mathbb{R}^{n \times p}$. Recall that the first k -eigenvectors recovered by PCA span a subspace that minimizes the ℓ^2 -distance from the original data to their projections onto that subspace. In particular, when $k = 1$, the associated PCA coordinates $u \in \mathbb{R}^n$ and eigenvector v are the optimal values in the problem

$$\begin{aligned} & \underset{u \in \mathbb{R}^n, v \in \mathbb{R}^p, d \in \mathbb{R}}{\text{minimize}} \quad \|X - duv^T\|_2^2 \\ & \text{subject to} \quad \|u\|_2 = \|v\|_2 = 1. \end{aligned}$$

The PMD generalizes this formulation of rank-one PCA to enforce additional structure on u and v . The PMD solutions u and v are defined as the optimizers of

$$\begin{aligned} & \underset{u \in \mathbb{R}^n, v \in \mathbb{R}^p, d \in \mathbb{R}}{\text{minimize}} \quad \|X - duv^T\|_2^2 \\ & \text{subject to} \quad \|u\|_2 = \|v\|_2 = 1. \quad (8) \\ & \text{Pen}_u(u) \leq \mu_1 \\ & \text{Pen}_v(v) \leq \mu_2 \end{aligned}$$

where Pen_u and Pen_v are arbitrary constraints u on and v .

To choose the regularization parameters μ_1 and μ_2 , Witten et al. (2009) applied cross-validation to the reconstruction errors after holding out random entries in X . To obtain a sequence of scores $(u_k)_{k=1}^K$ and $(v_k)_{k=1}^K$ for $K > 1$, define u_k and v_k as the optimizers of the problem (equation 8) on the residual: $X^k := X^{k-1} - d_{k-1}u_{k-1}v_{k-1}^T$ where $d_k = u_k^T X^k v_k$ and $X^1 = X$.

This view can be specialized to develop regularized versions of a number of multivariate analysis problems. We consider applications to the CCA and MultiCCA problems. Recalling that $\|A\|_F^2 = \text{tr}(A^T A)$ along with the linearity and the cyclic properties of the trace, the objective in equation (8) can be rewritten, using \equiv to mean equality up to terms constant in u and v ,

$$\begin{aligned} \|X - duv^T\|_F^2 &= \text{tr}\left((X - duv^T)^T(X - duv^T)\right) \\ &\equiv -2d \text{tr}(X^T uv^T) + d^2 \text{tr}(uv^T uv^T) \\ &\equiv -2dv^T X^T u + d^2, \end{aligned}$$

where for the last equivalence we used that $v^T v = u^T u = 1$.

From this expression, and by partially minimizing out $d = v^T X^T u$, we see that the PMD solutions u and v in equation (8) can be found as the optimizers of

$$\begin{aligned} &\underset{u \in \mathbb{R}^n, v \in \mathbb{R}^p}{\text{maximize}} u^T X^T v \\ &\text{subject to } \|u\|_2 = \|v\|_2 = 1 \\ &\text{Pen}_u(u) \leq \mu_1 \\ &\text{Pen}_v(v) \leq \mu_2 \end{aligned}$$

Notice that, as long as the penalties are convex in u and v , the optimization is biconvex, so a local maximum can be found by alternately maximizing over u and v .

From this form, we can derive a sparsity-inducing version of CCA. Recall the maximal-covariance interpretation of CCA,

$$\begin{aligned} &\underset{u \in \mathbb{R}^{p_1}, v \in \mathbb{R}^{p_2}}{\text{maximize}} u^T \hat{\Sigma}_{XY} v \\ &\text{subject to } u^T \hat{\Sigma}_{XX} u = v^T \hat{\Sigma}_{YY} v = 1 \end{aligned}$$

Witten et al. (2009) argue for diagonalized CCA, in which the variance constraints are replaced by unit norm constraints, and sparsity-inducing ℓ^1 constraints are added,

$$\begin{aligned} &\underset{u \in \mathbb{R}^{p_1}, v \in \mathbb{R}^{p_2}}{\text{maximize}} u^T \hat{\Sigma}_{XY} v \\ &\text{subject to } \|u\|_2 = \|v\|_2 = 1 \\ &\|u\|_1 \leq \mu_1 \\ &\|v\|_1 \leq \mu_2 \end{aligned}$$

which is exactly of the form of equation (9) where $X = \hat{\Sigma}_{XY}$.

Multiple CCA can also be described in this framework, by replacing the objective with the sum over all pairwise covariances,

$\sum_{l,l'=1}^L c_1^{(l)T} X^{(l)T} X^{(l')} c_1^{(l')}$, and introducing constraints for each of the $c_1^{(l)}$.

Example

We apply the PMD formulation of sparse CCA to the WELL-China data. As before, we k -over- A filter the microbiome data, requiring species to have counts of at least 5 in at least 7% of samples. Further, we first variance-stabilize, center, and scale these species abundances. For the regularization parameters, we set $\mu_1 = 0.7$ for the body composition data and $\mu_2 = 0.3$ for the species count data. The reasoning behind the relative values of these two tuning parameters is that sparsity in species loadings is more important than sparsity across body composition variables, because the microbiome data are more high-dimensional. The choice of the tuning parameters' overall magnitude was guided by the overall number of factors that we wanted to retain.

We only compute the first three PMD directions, and the associated correlations between scores are $(d_1, d_2, d_3) = (0.700, 0.435, 0.632)$. Note that the correlation can increase in subsequent directions, since directions are computed iteratively and cannot be defined and sorted all at once.

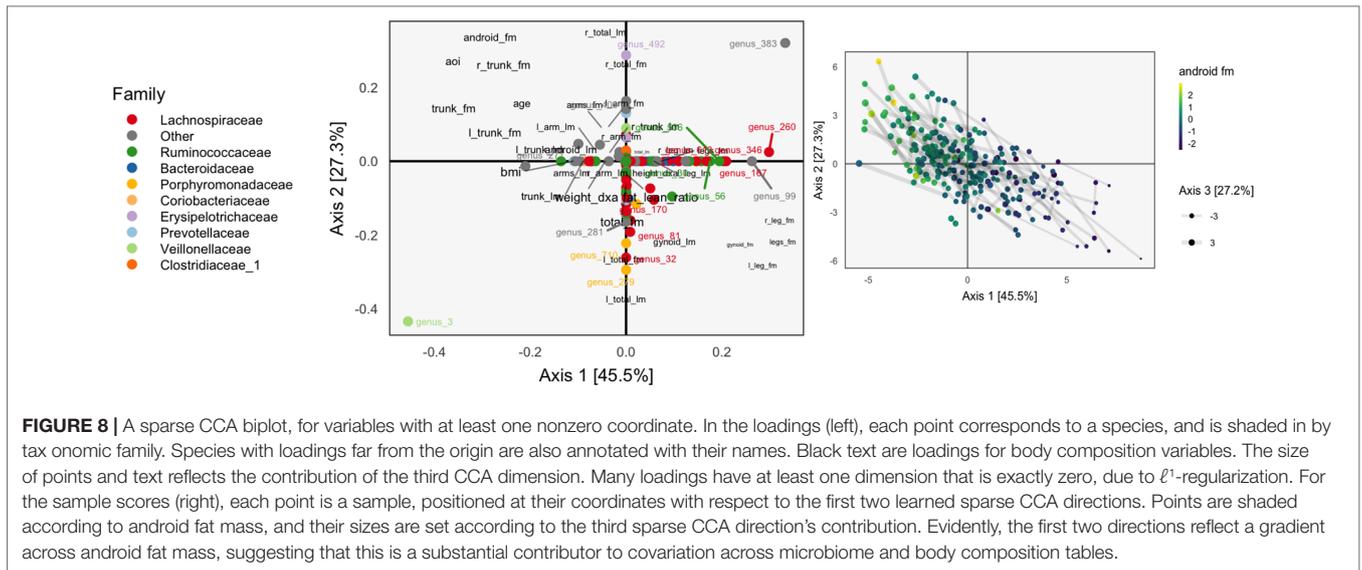
The learned loadings and scores are displayed in **Figure 8**. The x -axis in the loadings differentiates between high android and gynoid fat mass. The y -axes in the loadings reflect a gradient between overall right and left body mass. The size of points corresponds to the third PMD direction, and it seems to highlight high BMI, ratio of fat to lean mass, and overall weight. We interpret species based on their positions relative to these body composition variables, as in an ordinary biplot. For example, genus 492, located in the center-top, seems to be more common among people with higher android and lower gynoid fat mass.

The associated scores are displayed in the right panel, shaded according to android fat mass. The gradient between android and gynoid fat mass suggested by the loadings is clearly visible from this display. The length of links reflects the correlation between sets of scores. They are somewhat longer in the sparse CCA compared to the ordinary CCA on a subset of species, but this is likely a consequence of regularization and overfitting on the part of ordinary CCA.

We can follow up these displays by focusing on species that seemed related to the CCA axes. In the right panel of **Figure 7**, we isolate species with loadings a distance of at least 0.15 from the origin. These are the same ones that are labeled by text in **Figure 8**. We can see associations between abundance and android fat mass, as suggested by the loadings. Generally, there is a difference between android fat mass among people with and without particular species—there is no smooth function between the quantity of a species android fat mass, even in these cases where an association exists. Further, no individual taxonomic group seems to dominate the set of associated species.

Multitable Mixed-Membership

In section CCA, a latent variable interpretation of CCA was provided as an alternative to the standard covariance maximization perspective. Since likelihood-based methods are easily adapted to different data types, it is natural to consider versions of CCA designed



for non-Gaussian data, using section CCA as a starting point. We are particularly interested in data with the same structure as the WELL-China body composition and microbiome data, namely, two table data where one table is continuous with Gaussian marginals and correlated columns and the other is a high-dimensional collection of counts, where many entries are exactly zero.

As before, define a set of shared scores $\xi_i^s \in \mathbb{R}^K$, and two sets of within-table scores $\xi_i^X \in \mathbb{R}^{L_1}$ and $\xi_i^Y \in \mathbb{R}^{L_2}$. As before, we model the body composition variables using essentially a Gaussian factor analysis model, $y_i | \xi_i^X, \xi_i^Y \sim \mathcal{N}(B^Y \xi_i^s + W^Y \xi_i^Y, \sigma^2 I_{p_2})$ with a spherical Gaussian prior ξ_i^X, ξ_i^Y on. For the counts matrix, we might consider a few different approaches:

- *Bayesian Exponential Family PCA* (Mohamed et al., 2009): By requiring low-rank structure on the natural parameters of an exponential family model, we could naturally model high-dimensional count data, using a Poisson or multinomial likelihood, for example.
- *Nonnegative Matrix Factorization* (Lee and Seung, 2001): A variant of the exponential family approach is to model the counts matrix as a Poisson likelihood over a low-rank product of Gamma random matrices.
- *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003): We can model the observed samples as Dirichlet mixtures of a few underlying “topics,” which are themselves drawn from a Dirichlet prior.

Here, we focus on the LDA approach, though we suspect that the other two approaches are potentially interesting as well. Formally, this model supposes that counts are drawn according to

$$x_i | (\theta_k) \sim \text{Mult} \left(x_i | N_i \sum_{k=1}^K \theta_{ik} \beta_k \right)$$

$$\theta_i \sim \text{Dir}(\alpha)$$

$$\beta_k \sim \text{Dir}(\gamma),$$

where $N_i = \sum_{j=1}^{p_1} x_{ij}$ is the total count in sample i . This has the flavor of a factor analysis where $(\theta_{ik})_{k=1}^K$ are scores for the i^{th} sample and (β_k) are K underlying topics.

The only complexity with using an LDA model of X together with a Gaussian factor analysis on Y is that the shared scores ξ_i^s typically have different priors—a Dirichlet for LDA and a spherical Gaussian for factor analysis. In any formulation of probabilistic CCA that uses both models, this must be reconciled. One approach is to continue to place Dirichlet priors on all the scores, ξ_i^s, ξ_i^X , and ξ_i^Y . While the model for the Gaussian data is no longer exactly traditional factor analysis, it has a similar interpretation. Alternatively, we could use a spherical Gaussian prior on all scores and then recover probability vectors by

$$\text{applying the softmax function, } [\mathcal{S}(v)]_k = \frac{\exp(v_k)}{\sum_{k'} \exp(v_{k'})},$$

$$x_i | \xi_i^s, \xi_i^X \sim \text{Mult} \left(x_i | N_i, \mathcal{S} \left(B^X \xi_i^s + W^X \xi_i^X \right) \right)$$

$$\xi_i^s \sim \mathcal{N}(\xi_i^s | 0, \tau^2).$$

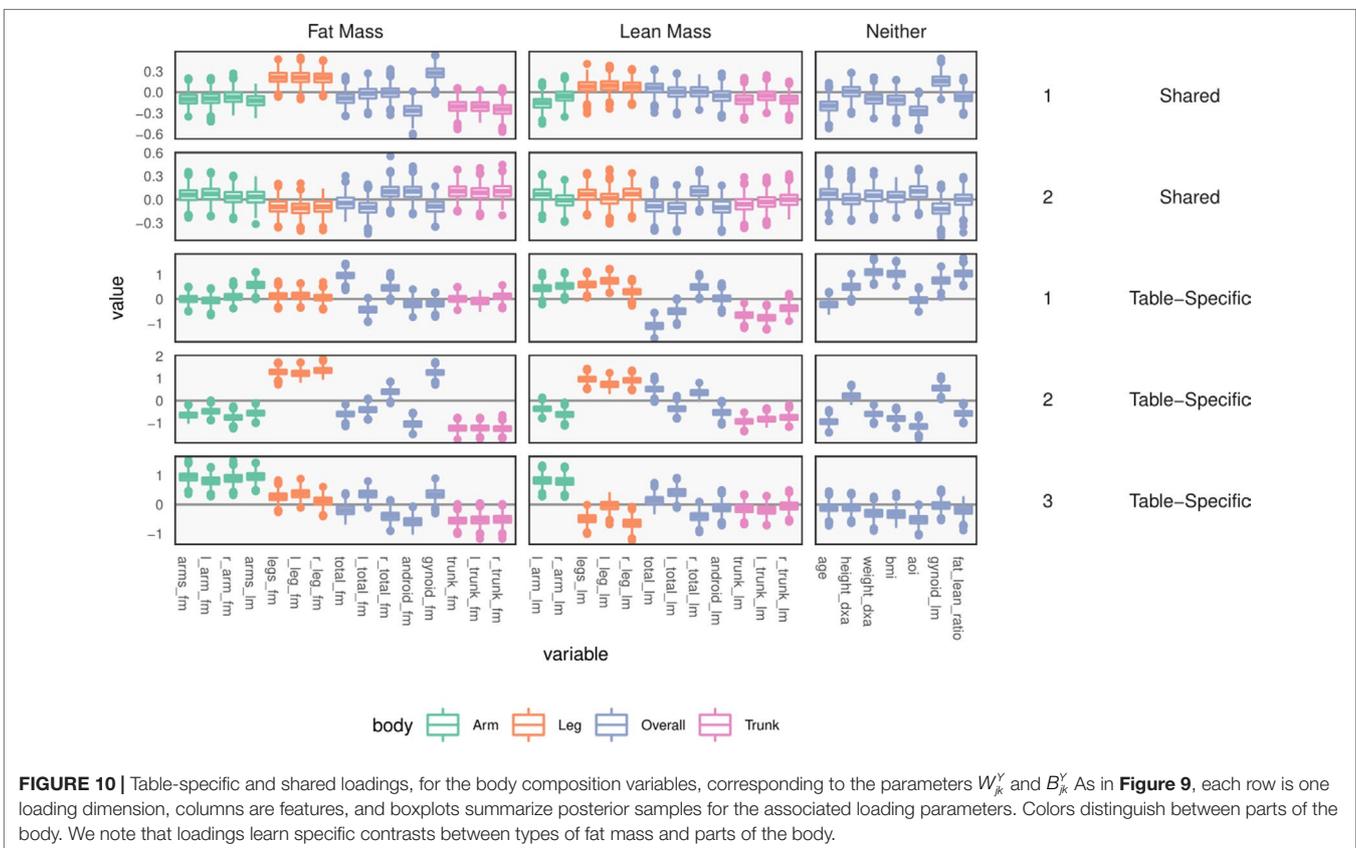
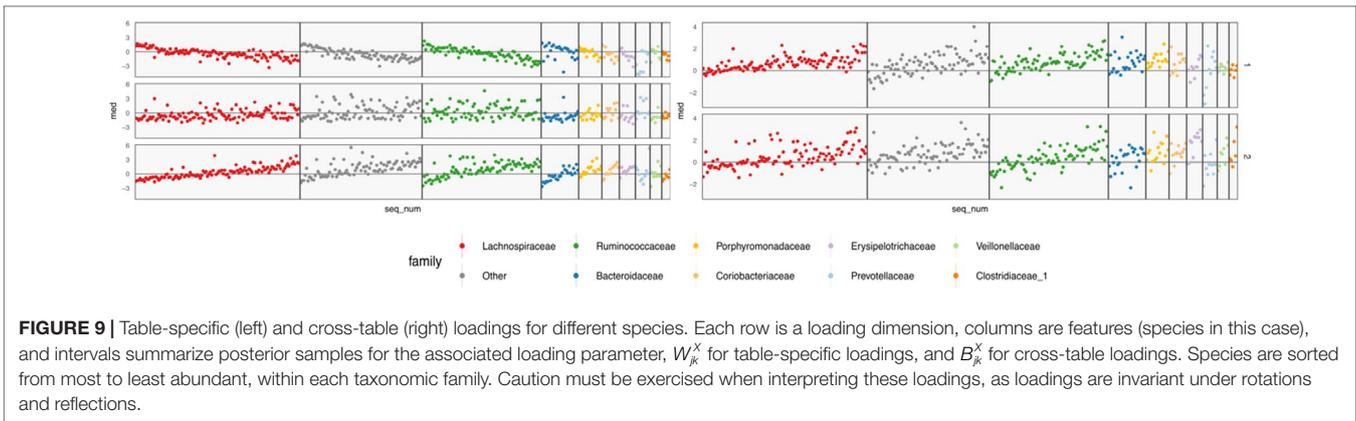
It is this second model that we use in our experiments below.

Example

We illustrate this multitable mixed-membership approach on the WELL-China data. We choose $K = 2$ for the number of shared topics and $L_1 = L_2 = 3$ for the number of unshared topics per table. We initialize scores and loadings using results from the PMD formulation of sparse CCA. While the use of shared ξ_i^s and unshared (ξ_i^X, ξ_i^Y) scores gives more flexibility in modeling, it also leads to additional complexity in interpretation—there are both more scores and more loadings that need to be visualized.

Consider the loadings W^X and W^Y , provided in the left panel of Figure 9 and bottom three rows of Figure 10. Note that there is no notion of variance explained by different axes in this case.

The loadings W^X of Figure 9 summarize table-specific variation in bacterial abundances. Invariance under rotation and reflection



complicates interpretation of these estimates. If we flip the sign of all the loadings axes, then the more abundant species have larger loadings, so the direction of different trends is irrelevant. The main distinction between the first and second loadings is the rate of decay in frequencies, especially among Lachnospiraceae and Ruminococcaceae. For example, topic 1 seems to include species from these taxonomic families that are not very abundant. The main characteristic of the third loading is that it has higher values for Porphyromonadaceae, so samples with high weight on this loading have decreased levels of these taxa.

Next, consider within-table body composition loadings, given in the bottom three rows of **Figure 10**, which suggests that the first and

third axes of W^Y capture variation between overall and android vs. gynoid fat mass. The first axis has high loadings for weight, BMI, and total fat mass, and the third contrasts areas with high android and high gynoid fat mass. The second axis distinguishes between right and left total lean and fat mass variation, while the third axis captures difference between mass in the trunk versus arms and legs.

These summaries could have been obtained by analyzing each table separately. Covariation between the two tables is captured by the shared scores ξ_i^s and loadings B^X , B^Y . The shared body composition loadings are given in the top two rows of **Figure 10**. These loadings again differentiate between android and gynoid fat mass, learning contrasts between body mass in arms and legs,

for example, though the effects are less pronounced than in the table-specific loadings.

The shared bacterial abundance loadings are given in the right panel of **Figure 9**. The most notable observation is that the first axes places more weight on rarer species, while the second places proportionally more weight on abundant species. Further, the two axes seem to have very different behaviors with respect to Prevotellaceae and Veillonellaceae.

In general, we find the results from the LDA-CCA approach less satisfying than those of the sparse CCA of section Penalized Matrix Decomposition. It seems that inference of a probabilistic model with shared and unshared parameters is more difficult than optimization of a single set of shared parameters. It may be possible to improve this approach through the following strategies:

- Applying LDA-CCA only to those species that are not sent entirely to zero by sparse CCA.
- Placing a sparsity-inducing prior on the scores $B^X, B^Y, W^X,$ and $W^Y,$ respectively, in the spirit of Archambeau and Bach (2009).

Curds & Whey

The Curds & Whey (C&W) procedure is a “soft” version of reduced-rank regression, differentially shrinking the ordinary least squares (OLS) fits with respect to the response canonical correlation directions (Breiman and Friedman, 1997). This is in contrast to reduced-rank regression, whose projection onto the first K response canonical correlation directions is a hard-thresholding analog. Hence, C&W is to reduced-rank regression what ridge regression is to principal component regression.

More precisely, the C&W algorithm fits a table Y according to

$$\hat{Y} = P_X Y V \Lambda^{-1}, \quad (10)$$

where again $V \in \mathbb{R}^{p_1 \times p_1}$ are the CCA directions associated with the response Y and P_X is the projection operator onto the column space of X . Λ is defined to be a diagonal matrix that determines the degree of shrinkage for the different canonical directions.

The main difficulty in C&W is the choice of Λ , and Breiman and Friedman (1997) suggest several possibilities. One choice is derived from a generalized cross-validation point of view, and results in shrinkage towards the response canonical correlation directions, without assuming the form of equation (10) *a priori*. This derivation is provided in section *Derivation of Curds & Whey Shrinkage*.

Graph-Fused Lasso

An approach to multiresponse regression, introduced by Chen et al. (2010), incorporates prior knowledge about the relationship between responses. Specifically, they use the correlation network between responses to induce structured regularization on the regression parameters.

Let $Y \in \mathbb{R}^{n \times p_1}$ and $X \in \mathbb{R}^{n \times p_2}$ and assume a correlation network between the p_2 tasks. This is denoted by $G = (V, E)$, where $V = \{1, \dots, p_2\}$. Each edge e is associated with a weight, $r(e)$, giving the correlation between the pair of responses.

The graph-fused lasso estimates a coefficient matrix $B \in \mathbb{R}^{p_2 \times p_1}$ whose columns $\beta^{(r)}$ are the regression coefficients across tasks, but which have been pooled together, with the strength of the pooling depending on the separately computed strength of the relationship between tasks. Formally, $\hat{\beta}$ is defined as the solution to the optimization,

$$\underset{B \in \mathbb{R}^{p_2 \times p_1}}{\text{minimize}} \frac{1}{2} \|Y - XB\|_F^2 + \lambda \|B\|_1 + \gamma \sum_{e \in E} \sum_{j=1}^{p_2} |r_e| |\beta_j^{(e^+)} - \text{sign}(r_e) \beta_j^{(e^-)}|, \quad (11)$$

where $\|B\|_1$ is the sum of the absolute values of all entries of B , β_j is the j^{th} row of B , and e^- and e^+ denote the nodes at either end of the edge e . The last regularization term in the objective is called the graph fused-lasso penalty, and it is this element that encourages pooling of information across regression problems.

Example

We apply the graph-fused lasso to the body composition problem and compare it to a naive version of the lasso that does not share any information across responses. We consider predicting the body composition variables, many of which are strongly correlated with one another, using variance-stabilized bacterial abundances.

We filter away species that do not appear in at least 7% of samples, as in the original PCA approach. We set the smoothing parameter to $\mu = 0.01$, while the ℓ^1 and graph-regularization parameters are set to $\lambda = 0.1$ and $\gamma = 0.01$, respectively, after they were heuristically found to provide interpretable levels of sparsity and smoothness in the fitted coefficients.

The graph-fused lasso requires a correlation graph between response variables. We estimate such a graph using the graphical lasso (Friedman et al., 2008), since there are only ~ 100 with which to estimate the 36-dimensional covariance matrix. The estimated correlation matrix is displayed in **Figure 11**.

The fitted coefficients from the graph-fused lasso are given in the top panel of **Figure 12**. The analogous display when the problem is decoupled into parallel lasso regressions is given in the bottom panel of the same figure.

Generally, both approaches highlight the same directions and size of association between individual species and the response variables, though those returned by the graph-fused lasso are smoother across responses. This smoothing may obscure true variation—for example, the stronger association between height_dxa and a few Ruminococcus species—that appears in the parallel-lasso approach. On the other hand, regularization reduces the number of one-off nonzero coefficients, which are likely just noise.

There appear to be real associations between Lachnospiraceae and Ruminococcaceae and the body composition measurements. The strongest negative association between species abundance and fat mass occurs among a few species of Ruminococcaceae. Most species that have any association tend to have the same direction and magnitude of association across all body composition variables, not just those restricted to one mass type. This seems to be the case even in the parallel-lasso context, where such structure has not been directly imposed.

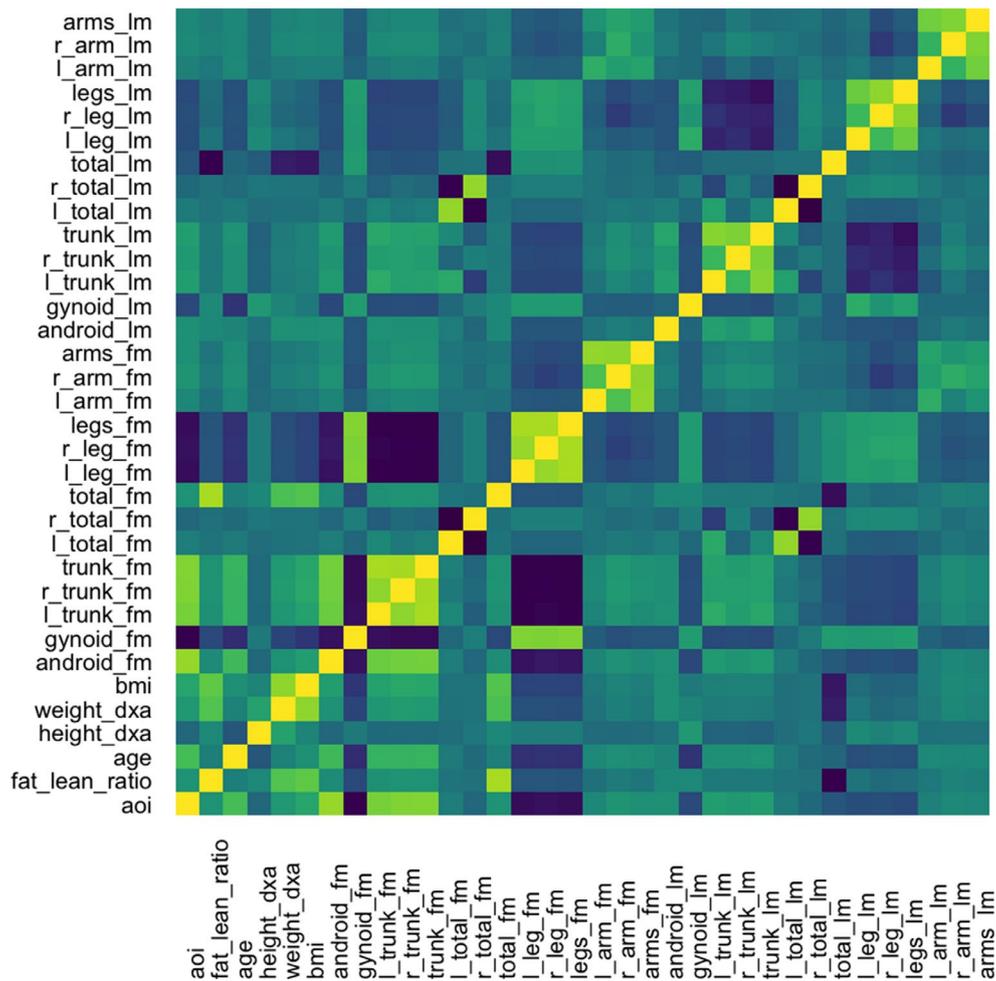


FIGURE 11 | Correlation matrix used as the input graph R for the graph-fused lasso, estimated itself according to the graphical lasso.

DISCUSSION

In this work, we have studied the problem of multitable data analysis, reviewing both the algorithmic foundations and practical applications of various methods. We have described approaches that are usually confined to particular literature areas and highlighted certain similarities in the process—for example, PCA-IV (section PCA-IV) and the graph-fused lasso (section Graph-Fused Lasso) were proposed in very different contexts, but have similar goals. By writing short, self-contained descriptions of various methods, we hope to contribute to an effort to distill ideas from the wide multitable data analysis literature to make them easily understandable to researchers interested in entering this field and useful for scientists hoping to apply these methods. A “cheat-sheet” summarizing some of the key properties of these methods is given in **Table 1**, and relevant packages can be found in **Table 2**.

In developing our WELL-China case study, we have both 1) described the types of interpretations facilitated by different approaches and 2) provided accessible implementations that can be incorporated into practical scientific workflows. Though our focus

on a single application has allowed side-by-side comparisons of methods, we do not want to leave the reader with the impression that these methods are tied in any way to this particular biological analysis task. Indeed, the value of mathematical abstractions is that they can be applied to situations outside the imaginations of the original method designers. For example, consider these potential use cases:

- *Microbiome and metabolites*: If we replace the body composition table with the concentrations of different metabolites across samples, we can begin to make claims about covariation between microbiome community composition and host metabolic processes (Chong and Xia, 2017; Fukuyama et al., 2017).
- *Microbiome and metagenomics*: In addition to a species composition matrix, we might have data quantifying the presence of various genes. The methods in this review could be used to understand the relationship between community composition and functional capacity (Gill et al., 2006; Kurokawa et al., 2007).
- *Microbiome and perturbations*: If we had a matrix tracking the application of various perturbations to the host—the use of various medications, for example—we could use

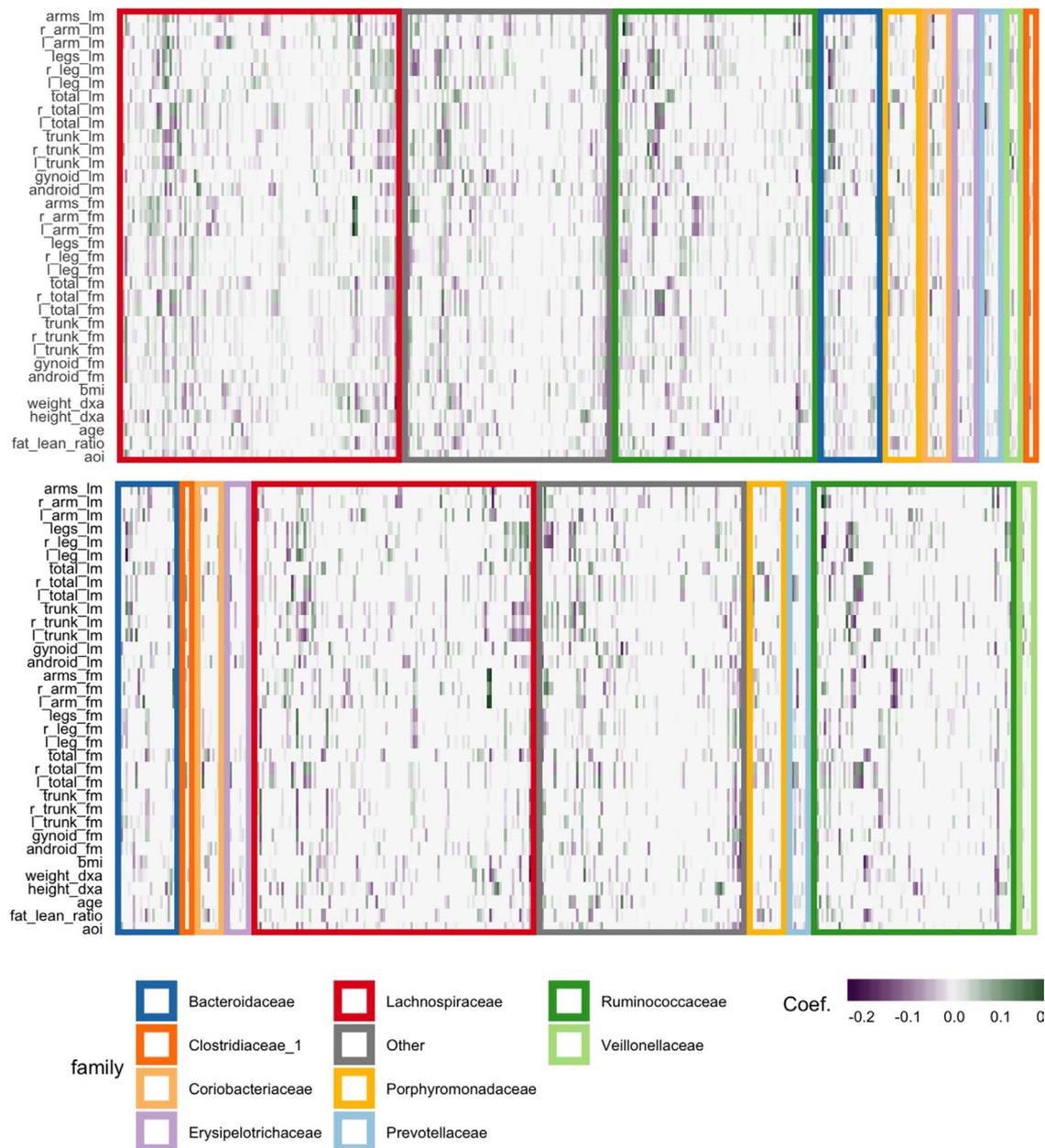


FIGURE 12 | Coefficients for the graph-fused (top) and decoupled (bottom) lasso fits highlight groups of species with similar profiles across response variables. Colored rectangles demarcate taxonomic families. Individual cells give the coefficient for a particular species (column) for a given response variable (row). Purple and green denote negative and positive coefficients, respectively. Note that coefficient graph-fused panels have been smoothed according to correlation network between variables, as given in **Figure 11**. Species with similar coefficients are placed near one another. Note that even in the decoupled case, where there is no sharing across response problems, the coefficients nonetheless seem to be similar within lean and fat mass response groups, respectively. However, they are not as smooth as in the graph-fused lasso. As there is some consistency within these groups of variables, the form of structured regularization imposed by the graph seems appropriate.

multitable methods to describe ways these (multidimensional) perturbations are related to microbiome community structure (Dethlefsen and Relman, 2011).

Our case study includes carefully thought-through visualizations of model results, a step that is crucial in scientific

study but often overlooked in methodological research, where model results are reduced to tables of performance metrics. Recognizing that a good deal of effort in statistical work goes into data preparation and visualization of model results, we have ensured that codes for all steps are available, so that our work is fully reproducible.

TABLE 1 | A high-level comparison of the multitable analysis methods discussed in this review. The purpose of this table is to give rules-of-thumb that can guide practical application, where choices invariably depend on the scale and structure of the data, the goals of the analysis, the expected number of future workflow applications, and availability of programming computation time.

Property	Algorithms	Consequence
Analytical solution	Concat. PCA, CCA, CoIA, MFA, PTA, Statico/Costatis	Methods with analytical solutions generally run much faster than those that require iterative updates, optimization, or Monte Carlo sampling. They tend to be restricted to more classical settings, however.
Require covariance estimate	Concat. PCA, CCA, CoIA, MFA, PTA, Statico/Costatis	Methods that require estimates of covariance matrices cannot be applied to data with more variables than samples, and become unstable in high-dimensional settings.
Sparsity	SPLS, Graph-Fused Lasso, Graph-Fused Lasso	Encouraging sparsity on scores or loadings can result in more interpretable, results for high-dimensional data sets. These methods provide automatic variable selection in the multitable analysis problem.
Tuning parameters	<i>Sparsity</i> : Graph-Fused Lasso, PMD, SPLS <i>Number of Factors</i> : PCA-IV, Red. Rank Regression, Mixed-Membership CCA <i>Prior Parameters</i> : Mixed-Membership CCA, Bayesian Multitask Regression	Methods with many tuning parameters are often more expressive than those without any, since it makes it possible to adapt to different degrees of model complexity. However, in the absence of automatic tuning strategies, these methods are typically more difficult to use effectively.
Probabilistic	Mixed-Membership CCA, Bayesian Multitask Regression	Probabilistic techniques provide estimates of uncertainty, along with representations of cross-table covariation. This comes at the cost of more involved computation and difficulty in assessing convergence.
Not Normal or Nonlinear	CCpNA, Mixed-Membership CCA, Bayesian Multitask Regression	When data are not normal (and are difficult to transform to normality) or there are sources of nonlinear covariation across tables, it can be beneficial to directly model this structure.
>2 Tables	Concat. PCA, CCA, MFA, PMD	Methods that allow more than two tables are applicable in a wider range of multitable problems. Note that these are a subset of the cross-table symmetric methods.
Cross-Table Symmetry	Concat. PCA, CCA, CoIA, Statico/Costatis, MFA, PMD	Cross-table symmetry refers to the idea that some methods don't need a supervised or multitask setup, where one table contains response variable and the other requires predictors. The results of these methods do not change when the two tables are swapped in the method input.

TABLE 2 | Pointers to R package that can be used to implement methods discussed in this survey. The vignettes in these packages go into more depth on the capabilities of these packages than do the short scripts used in our case study, available at https://github.com/krisrs1128/multitable_review.

Package	Methods	Documentation	Link
ade4	PCA, CCA, CoIA, Statico, Costatis, PCA-IV	Average	https://cran.r-project.org/web/packages/ade4/
FactoMineR	PCA, MFA	High	https://cran.r-project.org/web/packages/FactoMineR/
vegan	CCA, CCpNA	High	https://cran.r-project.org/web/packages/vegan/
spls	SPLS	High	https://cran.r-project.org/web/packages/spls/
PMA	PMD	High	https://cran.r-project.org/web/packages/PMA/
pls	PLS	High	https://cran.r-project.org/web/packages/pls/
Base R	PCA, CCA	High	https://cran.r-project.org/
GFLasso	Graph-Fused Lasso	Low	https://github.com/krisrs1128/gflasso
bayesMult	Bayesian Multitask Regression	Low	https://github.com/krisrs1128/bayesmult

We have found that multitable data analysis problems have motivated a wide range of analysis approaches. This is not surprising, considering the variety of contexts in which it arises, and it speaks to the richness of this methodological problem. As new data sources arise and as science evolves, we expect these ideas will inspire future generations of multitable research advances.

AUTHOR CONTRIBUTIONS

SH and KS conceived and designed the review, drafted the manuscript, and prepared all figures. KS implemented code for data analysis.

FUNDING

KS was supported by a Stanford University Weiland fellowship and the National Institutes of Health T32 grant 5T32GM096982-04. SH is supported by the National Institutes of Health TR01 grant AI112401.

ACKNOWLEDGMENTS

We thank the WELL-China study team for sharing the data appearing in this study and Yan Min for useful discussions.

An earlier version of this work first appeared in KS's PhD thesis (Sankaran, 2018).

REFERENCES

- Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11, R106. doi: 10.1186/gb-2010-11-10-r106
- Archambeau, C., and Bach, F. R. (2009). Sparse probabilistic projections. In *Advances in neural information processing systems*. 73–80.
- Ashish, N., Ambite, J. L., Muslea, M., and Turner, J. A. (2010). Neuroscience data integration through mediation: an (f) birn case study. *Front. Neuroinform.* 4, 118. doi: 10.3389/fninf.2010.00118
- Bach, F. R., and Jordan, M. I. (2005). *A probabilistic interpretation of canonical correlation analysis*. Berkeley: Technical Report 688 Department of Statistics, University of California.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022. doi: 10.1162/jmlr.2003.3.4-5.993
- Breiman, L., and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *J. R. Stat. Soc. Series B Stat. Methodol.* 59, 3–54. doi: 10.1111/1467-9868.00054
- Buhlmann, P., and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Berlin Heidelberg: Springer Science & Business Media. doi: 10.1007/978-3-642-20192-9
- Chalise, P., and Fridley, B. L. (2017). Integrative clustering of multi-level ‘omic’ data based on non-negative matrix factorization algorithm. *PLOS ONE* 12, e0176278. doi: 10.1371/journal.pone.0176278
- Chaudhary, K., Poirion, O. B., Lu, L., and Garmire, L. X., (2017). Deep learning based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.* 24 (6), 1248–1259. doi: 10.1101/114892
- Chen, X., Kim, S., Lin, Q., Carbonell, J. G., and Xing, E. P. (2010). Graph-structured multi-task regression and an efficient optimization method for general fused lasso. *arXiv preprint arXiv:1005.3579*. <https://arxiv.org/abs/1005.3579>
- Chong, J., and Xia, J. (2017). Computational approaches for integrative analysis of the metabolome and microbiome. *Metabolites* 7, 62. doi: 10.3390/metabo7040062
- Chun, H., and Kele, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Series B Stat. Methodol.* 72, 3–25. doi: 10.1111/j.1467-9868.2009.00723.x
- Chung, D., Chun, H., and Keles, S. (2012). Spls: Sparse partial least squares (spls) regression and classification. *R package, version 2*, 1–1.
- Dethlefsen, L., and Relman, D. A. (2011). Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci.* 108, 4554–4561. doi: 10.1073/pnas.1000087107
- Dolédéc, S., and Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species–environment relationships. *Freshw. Biol.* 31, 277–294. doi: 10.1111/j.1365-2427.1994.tb01741.x
- Frank, I. E., and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* 35, 109–135. doi: 10.1080/00401706.1993.10485033
- Franzosa, E. A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X. C., et al. (2015). Sequencing and beyond: integrating molecular ‘omics’ for microbial community profiling. *Nat. Rev. Microbiol.* 13, 360–372. doi: 10.1038/nrmicro3451
- Friedman, J., Hastie, T., and Tibshirani, R., (2001). *The elements of statistical learning* Vol. 1. Berlin: Springer series in statistics Springer. doi: 10.1007/978-0-387-21606-5_1
- Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432–441. doi: 10.1093/biostatistics/kxm045
- Fukuyama, J., Rumker, L., Sankaran, K., Jeganathan, P., Dethlefsen, L., Relman, D. A., et al. (2017). Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. *PLoS Comput. Biol.* 13, e1005706. doi: 10.1371/journal.pcbi.1005706
- Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., et al. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355–1359. doi: 10.1126/science.1124234
- Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkschlager, M., Gisel, A., et al. (2014). Data integration in the era of omics: Current and future challenges. *BMC Syst. Biol.* 8, I1. doi: 10.1186/1752-0509-8-S2-I1
- Greenacre, M. J. (1984). Theory and applications of correspondence analysis. *J. Am. Stat. Assoc.* 82 (398), 437–447.
- Greenacre, M., and Hastie, T. (1987). The geometric interpretation of correspondence analysis. *J. Am. Stat. Assoc.* 82, 437–447. doi: 10.1080/01621459.1987.10478446
- Gustafsson, M. G. (2001). A probabilistic derivation of the partial least-squares algorithm. *J. Chem. Inf. Comput. Sci.* 41, 288–294. doi: 10.1021/ci0003909
- Hannan, E. (1967). Canonical correlation and multiple equation systems in economics. *Econometrica*, 35(1), 123–138. doi: 10.2307/1909387
- Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* 28, 321–377. doi: 10.1093/biomet/28.3-4.321
- Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *J. Comput. Graph. Stat.* 12, 531–547. doi: 10.1198/1061860032148
- Kroonenberg, P. M. (2008). *Applied multiway data analysis*. Hoboken: John Wiley & Sons. doi: 10.1002/9780470238004
- Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., et al. (2007). Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* 14, 169–181. doi: 10.1093/dnares/dsm018
- Lê Cao, K.-A., Rossouw, D., Robert-Granic, C., and Besse, P. (2008). A sparse pls for variable selection when integrating omics data. *Stat. Appl. Genet. Mol. Biol.* 7, 1544–6115. doi: 10.2202/1544-6115.1390
- Lee, D. D., and Seung, H. S. (2001). “Algorithms for non-negative matrix factorization,” in *Advances in neural information processing systems*. Eds. T. K. Leen, T. G. Dietterich, and V. Tresp (Cambridge, MA: MIT Press), 556–562.
- Ley, R. E. (2010). Obesity and the human microbiome. *Curr. Opin. Gastroenterol.* 26, 5–11. doi: 10.1097/MOG.0b013e328333d751
- Ley, R. E., Backhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., and Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. U.S.A.* 102, 11070–11075. doi: 10.1073/pnas.0504978102
- Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Microbial ecology: human gut microbes associated with obesity. *Nature* 444, 1022–1023. doi: 10.1038/4441022a
- Mardia, K. V., Kent, J. T., and Bibby, J. M., (1980). *Multivariate analysis*. London: Academic Press
- Matsuzawa, Y. (2008). The role of fat topology in the risk of disease. *Int. J. Obes.* 32, S83. doi: 10.1038/ijo.2008.243
- McHardy, I. H., Goudarzi, M., Tong, M., Ruegger, P. M., Schwager, E., Weger, J. R., et al. (2013). Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome* 1, 17. doi: 10.1186/2049-2618-1-17
- Min, Y., Ma, X., Sankaran, K., Ru, Y., Chen, L., Baiocchi, M., et al. (2019). Sex-specific association between gut microbiome and fat distribution. *Nat. Commun.* 10, 2408. doi: 10.1038/s41467-019-10440-5
- Mohamed, S., Ghahramani, Z., and Heller, K. A. (2009). Bayesian exponential family pca proceedings of advances in neural information processing systems. *Adv. Neural. Inf. Process. Syst.* 1089–1096.
- Pagés, J. (2014). *Multiple Factor Analysis by example using R*. CRC Press. doi: 10.1201/b17700
- Pagés, J., and Tenenhaus, M. (2001). Multiple factor analysis combined with PLS path modelling. Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgements. *Chemom. Intell. Lab. Syst.* 58, 261–273. doi: 10.1016/S0169-7439(01)00165-4
- Pagés, J. (2004). Multiple factor analysis: main features and application to sensory data. *Rev. Colomb. Estad.* 27 (1), 1.
- Perez, P., and de Los Campos, G. (2014). Genome-wide regression & prediction with the bgrr statistical package. *Genetics*. 198 (2), 483–495. doi: 10.1534/genetics.114.164442
- Rahnavard, G., Franzosa, E. A., McIver, L. J., Schwager, E., Weingart, G., Moon, Y. S. et al., (2017). High-sensitivity pattern discovery in large multiomic datasets. <http://huttenhower.sph.harvard.edu/halla>
- Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhā*, 26 (4), 329–358. <https://www.jstor.org/stable/25049339>
- Sankaran, K. (2018). *Discovery and visualization of latent structure with applications to the microbiome*. Ph.D. thesis, Stanford University.
- Stone, M., and Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *J. R. Stat. Soc. Series B Stat. Methodol.* 52 (2), 237–269. doi: 10.1111/j.2517-6161.1990.tb01786.x

- Ter Braak, C. J. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67, 1167–1179. doi: 10.2307/1938672
- Thioulouse, J. (2011). Simultaneous analysis of a sequence of paired ecological tables: a comparison of several methods. *Ann. Appl. Stat.* 5 (4), 2300–2325. doi: 10.1214/10-AOAS372
- Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457, 480. doi: 10.1038/nature07540
- Vlassis, N., Motomura, Y., and Krose, B., (2000). “Supervised linear feature extraction for mobile robot localization,” in *Robotics and Automation, 2000. Proceedings. ICRA’00. IEEE International Conference on (IEEE)*, vol. 3, 2979–2984.
- Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*. 10 (3), 515–534. doi: 10.1093/biostatistics/kxp008
- Witten, D., Tibshirani, R., Gross, S., and Narasimhan, B. (2013). Package ‘pma’. *Genet. Mol. Biol.* 8, 28.
- Wold, H. (1985). “Partial least squares,” in *Encyclopedia of statistical sciences*. Vol. 6. New York: John Wiley, 581–591.
- Zhu, M., Hastie, T. J., and Walther, G. (2005). Constrained ordination analysis with flexible response functions. *Ecol. Modell.* 187, 524–536. doi: 10.1016/j.ecolmodel.2005.01.049
- Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Stat.* 15, 265–286. doi: 10.1198/106186006X113430

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Sankaran and Holmes. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

This appendix includes derivations and technical discussion of several methods surveyed in the main text: PCA-IV, PTA, and the C&W algorithm. While these methods can be understood and applied based on their computational description, these mathematical discussions provide motivation and context for their particular form.

DERIVATION DETAILS FOR PCA-IV

In this section, we provide the argument for why the generalized eigendecomposition $\Sigma_{XY} \Sigma_{YX} = \Sigma_{XX} V \Lambda V^T$ provides the optimal V used in PCA-IV.

First consider $k = 1$. For any \tilde{v} , the objective in equation (5) has the form

$$\begin{aligned} \text{tr} \left(\hat{\Sigma}_{YX} \tilde{v} \left(\tilde{v}^T \hat{\Sigma}_{XX} \tilde{v} \right)^{-1} \left(\hat{\Sigma}_{YX} \tilde{v} \right)^T \right) &= \frac{\tilde{v}^T \hat{\Sigma}_{XY} \hat{\Sigma}_{YX} \tilde{v}}{\tilde{v}^T \hat{\Sigma}_{XX} \tilde{v}} \\ &= \frac{\tilde{w}^T \Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XY}^{-\frac{1}{2}} \Sigma_{YX} \Sigma_{XX}^{-\frac{1}{2}} \tilde{w}}{\|\tilde{w}\|_2^2} \end{aligned} \quad (12)$$

where we change variables $\tilde{w} = \Sigma_{XX}^{-\frac{1}{2}} \tilde{v}$. But to maximize equation (12), just choose \tilde{w} to be the top eigenvector of $\Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XY} \Sigma_{YX} \Sigma_{XX}^{-\frac{1}{2}}$, which implies that \tilde{v} is the top generalized eigenvector of $\Sigma_{XY} \Sigma_{YX}$ with respect to Σ_{XX} . Indeed, in this case,

$$\begin{aligned} \Sigma_{XY} \Sigma_{YX} \tilde{v} &= \Sigma_{XY} \Sigma_{YX} \Sigma_{XX}^{-\frac{1}{2}} \tilde{w} \\ &= \Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XY}^{-\frac{1}{2}} \Sigma_{YX} \Sigma_{XX}^{-\frac{1}{2}} \tilde{w} \\ &= \Sigma_{XX}^{-\frac{1}{2}} \lambda_1 \tilde{w} \\ &= \lambda_1 \Sigma_{XX} \tilde{v}. \end{aligned}$$

Hence, in the case $K = 1$, the criterion is maximized by the top generalized eigenvector. For larger K , recall that the problem of maximizing $\frac{v^T A v}{\|v\|^2}$ over v subject to being orthogonal to the first $K - 1$ eigenvectors of A is solved by the K^{th} eigenvector of A , and applying this fact in step 12 of the argument above gives the result for general K .

DERIVATION OF PTA α

The Lagrangian of the optimization defined by PTA is

$$\mathcal{L}(\alpha, \lambda) = \sum_{i=1}^L \alpha_i \langle \bar{X}, X_{..i} \rangle + \lambda (\|\alpha\|_2^2 - 1),$$

Which, when differentiated with respect to α , yields $\alpha_i = -\frac{1}{2\lambda} \langle \bar{X}, X_{..i} \rangle$ for all i . The constraint that $\|\alpha\|_2^2 = 1$ implies that $\frac{1}{4\lambda^2} \sum_{i=1}^L \langle \bar{X}, X_{..i} \rangle^2 = 1$, which gives $\lambda = \frac{1}{2} \sqrt{\sum_{i=1}^L \langle \bar{X}, X_{..i} \rangle^2}$, so $\alpha_i = \frac{\langle \bar{X}, X_{..i} \rangle}{\sqrt{\sum_{i=1}^L \langle \bar{X}, X_{..i} \rangle^2}}$.

DERIVATION OF CURDS & WHEY SHRINKAGE

Consider prediction across many related response variables. One way to pool information across responses is to define new fitted values from a linear combination of independent OLS fits. That is, to predict a response $y_i \in \mathbb{R}^{p_i}$, we set $\hat{y}_i^{\text{cw}} = B \hat{y}_i^{\text{ols}}$ for some square matrix $B \in \mathbb{R}^{p_i \times p_i}$. But how to choose B ?

One reasonable idea is to choose a B that has the best performance in a generalized cross-validation (GCV). The GCV approximation is that the h_{ii} can be approximated by their average across all diagonal elements of H : $h_{ii} \approx h := \frac{1}{n} \text{tr}(H)$ for all i . In this spirit, define $g = \frac{1}{1-h}$ and approximate

$$\hat{y}_{-i} \approx (1-g)y_i + g\hat{y}_i$$

Then, the leave-one-out CV error can be simplified to

$$\sum_{i=1}^n \|y_i - B\hat{y}_{-i}\|_2^2 = \sum_{i=1}^n \|y_i - B((1-g)y_i + g\hat{y}_{-i})\|_2^2,$$

and differentiating with respect to B , we find that the optimal \hat{B}^{cw} in this GCV framework must satisfy

$$\sum_{i=1}^n (y_i - B((1-g)y_i + g\hat{y}_{-i}))((1-g)y_i + g\hat{y}_{-i})^T,$$

or equivalently

$$\sum_{i=1}^n y_i ((1-g)y_i + g\hat{y}_{-i})^T = \sum_{i=1}^n B((1-g)y_i + g\hat{y}_{-i})((1-g)y_i + g\hat{y}_{-i})^T,$$

which in matrix form is

$$(1-g)Y^T Y + g\hat{Y}^T Y = B((1-g)Y + g\hat{Y})^T ((1-g)Y + g\hat{Y}), \quad (13)$$

where $\hat{Y} \in \mathbb{R}^{n \times p_i}$ has i^{th} row \hat{y}_{-i} .

Next, we can represent these cross-products in a way that is suggestive of CCA,

$$\begin{aligned}
 Y^T Y &= n \hat{\Sigma}_{YY} \\
 \hat{Y}^T Y &= Y^T H Y = Y^T X (X^T X)^{-1} X^T Y = n \hat{\Sigma}_{YX} \hat{\Sigma}_{XX}^{-1} \hat{\Sigma}_{XY} \\
 \hat{Y}^T \hat{Y} &= Y^T P_X^2 Y = Y^T P_X Y = n \hat{\Sigma}_{YX} \hat{\Sigma}_{XX}^{-1} \hat{\Sigma}_{XY},
 \end{aligned}$$

Substituting this into equation (13) and ignoring the scaling n yields

$$\begin{aligned}
 (1-g) \hat{\Sigma}_{YY} + g \hat{\Sigma}_{YX} \hat{\Sigma}_{XX}^{-1} \hat{\Sigma}_{XY} &= \\
 B \left[(-g) \hat{\Sigma}_{YY} + (2g-g^2) \hat{\Sigma}_{YX} \hat{\Sigma}_{XX}^{-1} \hat{\Sigma}_{XY} \right].
 \end{aligned}$$

Postmultiplying by $\hat{\Sigma}_{YY}^{-1}$ gives

$$(1-g) I_{p_1} + g \hat{Q}^T = B [(1-g) I_{p_1} + (2g-g^2) \hat{Q}^T], \quad (14)$$

where

$$\hat{Q} := \hat{\Sigma}_{YX}^{-1} \hat{\Sigma}_{YX} \hat{\Sigma}_{XX}^{-1} \hat{\Sigma}_{XY} \in \mathbb{R}^{p_1 \times p_1}$$

Now, we claim that we can decompose $\hat{Q} = VD^2V^{-1}$, where $V \in \mathbb{R}^{p_1 \times p_1}$ is the full matrix of CCA response directions and D is diagonal with the canonical correlations. Indeed, the usual CCA response directions V can be recovered by setting $V = \hat{\Sigma}_{YY}^{-\frac{1}{2}} \tilde{V}$, where \tilde{V} comes from the SVD of $A := \frac{-\frac{1}{2}}{XX} \Sigma_{XY} \Sigma_{XX}^{-\frac{1}{2}} = \tilde{U} D \tilde{V}^T$. Hence

$$\begin{aligned}
 Q &= \Sigma_{YY}^{-\frac{1}{2}} A^T A \Sigma_{YY}^{\frac{1}{2}} \\
 &= \Sigma_{YY}^{-\frac{1}{2}} \tilde{V}^T D^2 \tilde{V} \Sigma_{YY}^{\frac{1}{2}} \\
 &= VD^2V^{-1},
 \end{aligned}$$

where we are able to write $V^{-1} = \tilde{V}^T \Sigma_{YY}^{-\frac{1}{2}}$ because \tilde{V} is the full (untruncated) matrix of eigenvectors, so $\tilde{V} \tilde{V}^T = I$ in addition to the usual $\tilde{V}^T \tilde{V} = I$, which holds even for the truncated SVD.

Therefore, equation (14) can be expressed as

$$V^{-T} [(1-g) I_{p_1} + g D^2] V^T = B V^{-T} [(1-g) I_{p_1} + (2g-g^2) D^2] V^T$$

and the B satisfying the normal equations has the form

$$\hat{B}^{cw} = V^{-T} \Lambda V^T,$$

where Λ is a diagonal matrix with entries

$$\lambda_{jj} = \frac{1-g + d_{jj}^2}{1-g + (2g-g^2) d_{jj}^2}$$

Notice that when n is large, $\frac{1}{n} \text{tr} P_X$ will be small, leading to a smaller $g \approx 0$ and less shrinkage. Recall that \hat{B}^{cw} is used to pool across OLS fits, $\hat{y}_i^{cw} = \hat{B}^{cw} \hat{y}_i^{ols}$. That is,

$$\hat{Y}^{cw} = \hat{Y}^{ols} B^T = \hat{Y}^{ols} V \Lambda V^{-1}$$

which we can also view as $\hat{Y}^{cw} V = (\hat{Y}^{ols} V) \Lambda$. This means that the C&W coordinates along the canonical directions V are set as the OLS fits \hat{Y}^{ols} along the canonical directions V , with weights defined by Λ . The actual \hat{Y}^{cw} are recovered by transforming back to the original coordinate system. A similar way to view the C&W fits is to note $\hat{Y}^{cw} V = P_X (YV) \Lambda$, which is the original data Y according to the canonical directions, then projects the shrunk data onto the subspace defined by the columns of X . In any case, we see that C&W pools across regression problems through a soft shrinkage weighted along canonical response directions.