Multitable Data Analysis for the Microbiome

Kris Sankaran and Susan P. Holmes

December 12, 2017

Abstract

Many modern microbiome studies require the analysis of data collected across multiple measurement sources – for example, 16s sequencing, metagenomic, metabolomic, or transcriptomic data. We abstract away the essential scientific questions underlying these studies and provide a self-contained review of methods that have emerged to address them. In addition to summarizing methods abstractly, we describe an application to a multitable problem encountered at the intersection of the microbiome and epidemiology literatures. Code for all methods and figures is publicly available at `https://github.com/krisrs1128/well_microbiome_expers`. We hope this work can both (1) help practitioners understand and easily apply multitable methods in their day-to-day work and (2) provide a concrete case study to motivate the design and comparison of multitable techniques among methods researchers.

Contents

1 Introduction

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 sequencing, metagenomic, metabolomic, or transcriptomic data – can be collected on the same physical samples [McHardy et al., 2013, Franzosa et al., 2015]. 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 [Fukuyama et al., 2017, Rahnavard et al., Chaudhary et al., 2017, Chalise and Fridley, 2017].

However, stepping back from the rush of 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 from economics [Hannan, 1967] to robotics [Vlassis et al., 2000] to computational biology [Gomez-Cabrero et al., 2014]. 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 the 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, ...) of the data. Towards the second goal, we have made code for all analysis and figures available online at `https://github.com/krisrs1128/well_microbiome_expers`.

First, though, why can't multiple sources of data simply be combined into a single, unified table for subsequent analysis? One answer is that 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 primarily collected to provide better features, there are often an additional relational components to the analysis: how do different types of measurements covary 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, consider data from the WELL-China study, which is focused on the relationships between various indicators of wellness [Center]. In this study, 1969 individuals[1] 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 sequencing and metabolomic analysis. To date, 16s sequencing data is 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 this 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 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 sequencing is a technology for gauging the abundance of differnet bacterial species in the gut by counting the alignments of reads to the 16s gene, a component of all bacterial genomes with enough variation to allow discimination between different individual species. We have counts associated with 2565 species across 181 genuses, 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 BMI [Ley et al., 2006, Turnbaugh et al., 2009, Ley et al., 2005, Ley, 2010]. Fur-

[1] Though sampling is still ongoing.

ther, 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 microbial abundance data might yield a more complete view of the relationship between the microbiome and body type.

2 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, where the motivating data analysis problem was to relate prices of groups of commodities [Hotelling, 1936]. There are two notable aspects of data analysis in this classical paradigm which 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 straightforwards to study all of them simultaneously. It is now possible to automatically collect a large number of features for each sample.

- 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 in an environment 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, as well as 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.

2.1 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

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

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 $v_1, \ldots, v_K$, while the associated scores are reweighted rows $d_1u_1, \ldots, d_Ku_K$. 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.
- 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 2.2 and 2.4, 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, see Figure 1. Formally, recall that VV^Tx_i is the projection of x_i onto the subspace spanned by the columns of V. PCA identifies the orthogonal matrix $V \in \mathbb{R}^{p\times K}$ such that

$$\sum_{i=1}^{n} \|x_i - VV^Tx_i\|_2^2$$

is minimized. The principal component scores are then the coordinates of the projected points with respect to this subspace.

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, see Figure 2. In this figure, histograms of the scores associated with four linear combinations of the original body composition measurements are displayed side by side, to

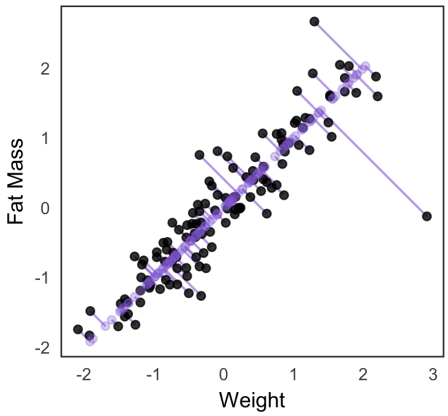

Figure 1: A geometric motivation for PCA. The first principal component spans the one-dimensional subspace V minimizing the squared error between points and their projections, $\sum_{i=1}^{n} \| \left(I - VV^T\right) x_i \|_2^2$. The black points represent the scaled values for two of the body composition variables – weight and total fat mass – while the purple points are the projections onto the first principal component. Contrast this with Figures 27 and 28, where variables other than and fat mass are taken into account.

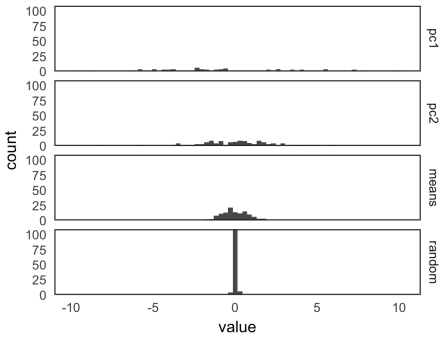

Figure 2: A statistical motivation for PCA. Each histogram is computed from the linear combinations $\left(c^{j^T} x_i\right)$, for c^j equal to the first and second principal components, the vector $\frac{1}{n}\mathbf{1}$, and a normalized random multivariate Gaussian. At least among these four vectors, the scores obtained from the first PC have the largest variance. It is a theorem that its variance is in fact maximal among all linear combinations with unit-length vectors.

emphasize the fact that the linear combination $x^T v_1$ associated with the first principal component v_1 has the largest variance among a few natural choices.

Formally, suppose that the $x_i \in \mathbb{R}^p$ are drawn independently from some distribution $\mathbb{P}$, so that the variance is $\mathrm{Cov}_{\mathbb{P}}[x_i] = \Sigma$. Consider an arbitrary linear combination of x_i's p coordinates: $z_i := c^T x_i$ for some $c \in \mathbb{R}^p$. The first PCA direction gives the unit-length c such that the variance of this coordinate, $\mathrm{Var}_{\mathbb{P}}(z_i) = c^T \Sigma c$, is maximal. The second direction gives the linear combination that maximizes variance, subject to being orthogonal to the first, and so forth.

While our description of the method of concatenating multiple tables into a single one has focused on PCA, note that other single table methods could be applied instead. For example, suppose there are multiple data types – say count, categorical, and continuous – across tables. Then, it is possible to define a new distance between samples as a mixture of distances based on several tables. For example, Jaccard, χ^2, and Euclidean distances can be applied to binary, count, and real valued tables. The combined distance can then be input into any distance-based single-table procedure, like multidimensional scaling or hierarchical clustering. The primary downside of this approach is that the resulting distance only allows a comparison between samples, but not across features.

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

2.1.1 Example

Figure 3 and 4 illustrate 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 data have been variance stabilized according the methodology proposed in [Anders and Huber, 2010a] and filtered to only those species that have count ≥ 5 in at least 7% of samples.

Figure 3 displays the loadings associated with this concatenated PCA approach, where body composition (36 columns) and 16s 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.

Most body composition variables lie on the top right, 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. Age appears in generally negatively correlated with the other variables. 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

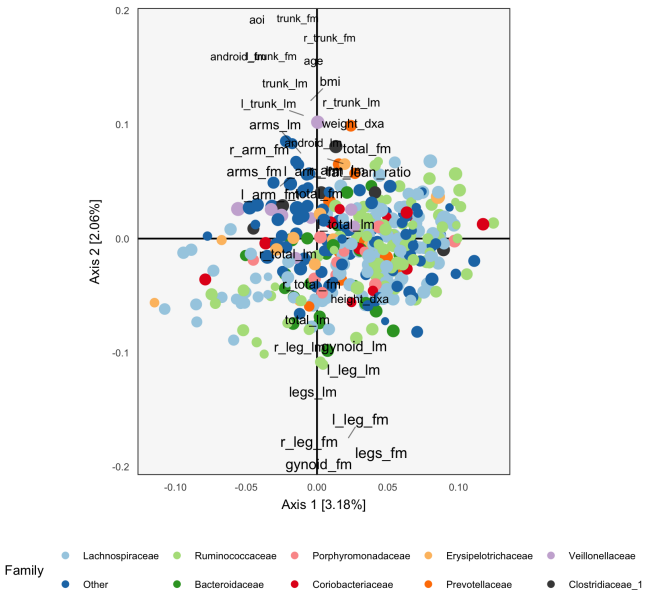

Figure 3: The loadings obtained by applying PCA to the combined body composition and microbial abundance data. Species are points, and are shaded in by taxonomic family. Body composition variables are plotted as text. The size of points and words measures the contribution of the third PC dimension.

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 Figures 4 and 5. These figures display samples in the same positions, but shaded by android (i.e., abdominal) fat mass and the difference between variance-stabilized Bacteroidaceae and Ruminococcaceae[2] counts, respectively. This shading confirms the observations from the loadings directly using observed data. Indeed, the increasing android fat mass among samples in the top of Figure 4 exactly corresponds to the fact that related variables lie at the top in Figure 3. The largest Bacteroidaceae vs. Ruminococcaceae differentiation is visible in Figure 5, which is consistent with the loadings.

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

[2]We pass the difference through tanh to squash very large differences.

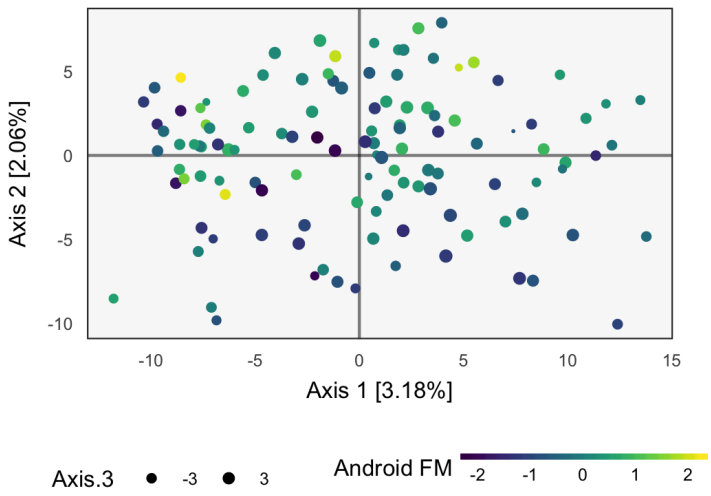

Figure 4: The scores resulting from PCA applied to the combined body composition and microbial abundance data. The first two principal components explain about 4% of the variance. Points are species, shaded in by taxonomic family. Text are columns from the body composition data. Points and text with small angle between one another are more correlated.

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.

2.2 CCA

CCA is a close relative of PCA, designed to compare sets of features across tables. Like PCA, it provides low-dimensional representations of samples, 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 $u_k = \Sigma_{XX}^{-\frac{1}{2}} \tilde{u}_k \in \mathbb{R}^{p_1}$ and $v_k = \Sigma_{YY}^{-\frac{1}{2}} \tilde{v}_k \in \mathbb{R}^{p_2}$. 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.

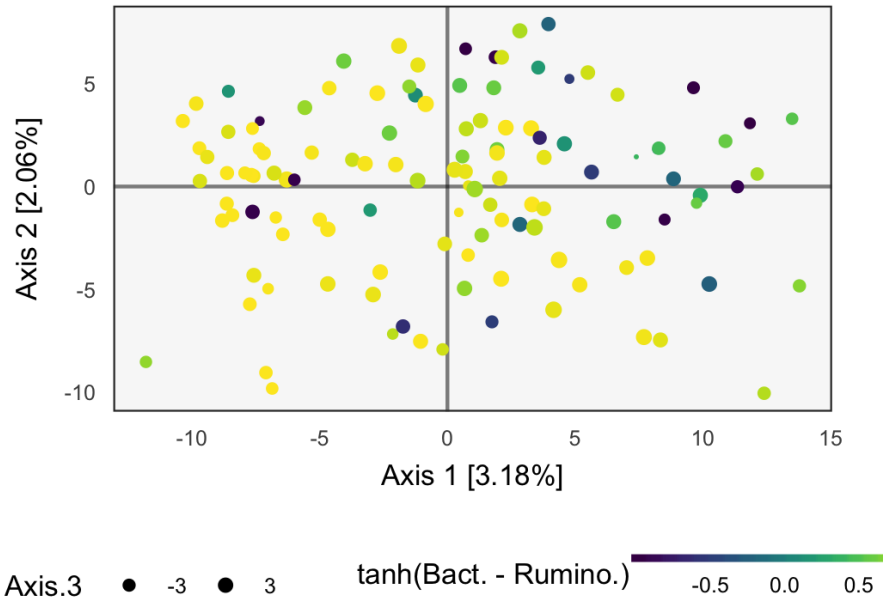

Figure 5: The same scores as Figure 4, but shaded now by the transformed ratio of Bacteroidaceae over Ruminococcaceae abundances. The slight increase in Ruminococcaceae from left to right is consistent with the loadings observed in Figure 3.

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_1^* and v_1^* are chosen to optimize

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

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

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

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 now be expressed as

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

The optimal $\tilde{u}_1$ and $\tilde{v}_1$ for this problem are well known – they're 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. The argument is standard[3], but we include it for completeness.

Let ξ and ν be potential maximizers of length one. We can find w_u and w_v such that $\xi = \tilde{U}w_u, \nu = \tilde{V}w_v$, since $\tilde{U}$ and $\tilde{V}$ are both orthonormal bases. Since they are length one, $1 = \|\xi\|_2^2 = w_u^T\tilde{U}^T\tilde{U}w_u = \|w_u\|_2^2$ and similarly $\|w_v\|_2^2 = 1$. The objective 3 can be bounded by

$$\begin{aligned} \xi^T\hat{\Sigma}_{XX}^{-\frac{1}{2}}\hat{\Sigma}_{XY}\hat{\Sigma}_{YY}^{-\frac{1}{2}}\nu &= w_u^T\tilde{U}^T\tilde{U}D\tilde{V}^T\tilde{V}w_v \\ &= w_u^T D w_v \\ &= \sum_{k=1}^{p_1\wedge p_2} d_k w_{uk} w_{vk} \\ &\leq d_1 \sum_{k=1}^{p_1\wedge p_2} w_{uk} w_{vk} \\ &\leq d_1 \|w_u\|\|w_v\| = d_1, \end{aligned}$$

and this maximum is attained when w_u and w_v both put all their on the first coordinate, that is $\xi = \tilde{u}_1$ and $\nu = \tilde{v}_1$. For subsequent directions, we repeat the argument but require that w_u and w_v have zero on the first columns of $\tilde{U}$ and $\tilde{V}$.

To recover the solutions to the original problem, we reverse the original variable transformation, yielding optimal $u_1 = \Sigma_{XX}^{-\frac{1}{2}}\tilde{u}_1$ and $v_1 = \Sigma_{YY}^{-\frac{1}{2}}\tilde{v}_1$.

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,

$$\begin{aligned} \xi_i := (\xi_i^s, \xi_i^x, \xi_i^y) &\sim \mathcal{N}(0, I_d) \\ x_i|\xi_i &\sim \mathcal{N}(\mu_x + W_X\xi_i^s + B_X\xi_i^x, I_d) \\ y_i|\xi_i &\sim \mathcal{N}(\mu_Y + W_Y\xi_i^s + B_Y\xi_i^y, I_d) \end{aligned}$$

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

[3]See [Mardia et al., 1980], for example.

the posterior expectations of the latent ξ_i^s (assumed $\in \mathbb{R}^K$) given the observed tables must lie on the subspace defined by the CCA directions. More precisely,

$$\mathbb{E}\left[\xi_i \mid x_i^{(1)}\right] \in \text{span}\left(z_1^{(1)}, \ldots, z_K^{(1)}\right),$$

and

$$\mathbb{E}\left[\xi_i \mid x_i^{(2)}\right] \in \text{span}\left(z_1^{(2)}, \ldots, z_K^{(2)}\right).$$

Finally, we observe that the logic for CCA generalizes to an arbitrary number L of tables, by summing all pairwise covariances. That is, the instead of finding directions $c_k^{(1)}$ and $c_k^{(2)}$ maximizing $\text{Cov}_{\mathbb{P}^{(1)},\mathbb{P}^{(2)}}\left[c_k^{(1)T}x_i^{(1)}, c_k^{(2)T}x_i^{(2)}\right]$ subject to normalization and orthogonality constraints, we can seek directions $c_k^{(1)}, \ldots, c_k^{(L)}$ that maximize the sum of cross-covariances $\sum_{l,l'=1}^{L} \text{Cov}_{\mathbb{P}^{(l)},\mathbb{P}^{(l')}}\left[c_k^{(l)T}x_i^{(l)}, c_k^{(l')}x_i^{(l')}\right]$.

2.2.1 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 2.1.1. To this end, we provide analogous loadings and scores plots in Figures 6 through 9. 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 2565. This very aggressive filtering is necessary because CCA requires estimation of covariances matrices 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 2.1.1.

Figure 6 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. Then, Figure 6 displays text labels for column j of the body composition variables at location $(u_{j1}, u_{j2})_{j=1}^{36}$ 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 separates 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 Bacteroideceae and Ruminoccocus do appear restricted to the bottom and top right, respectively.

In Figure 7 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 maximizes the correlation between these scores. This figure displays both sets of scores, colored differently according to the table to which

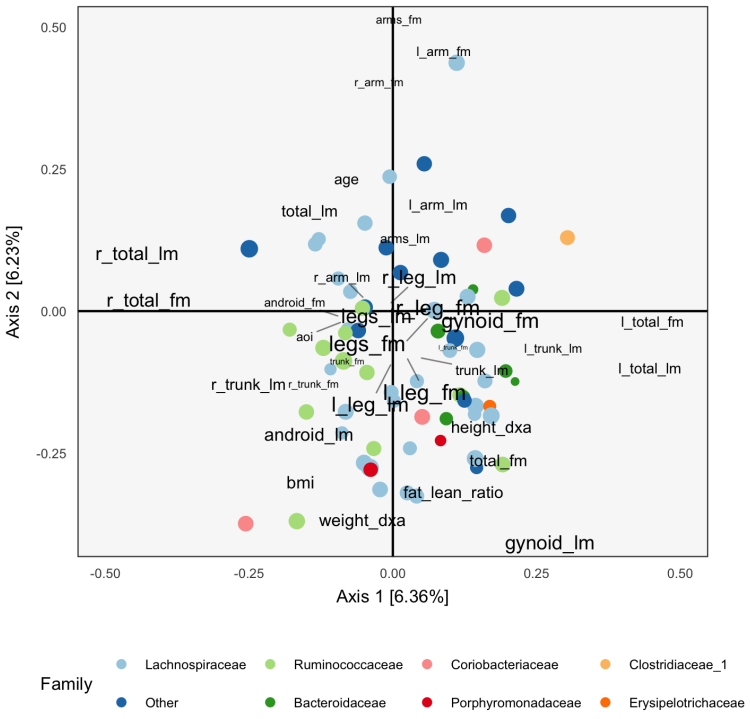

Figure 6: The CCA analog of the PCA loadings plot in Figure 3. loadings obtained by applying CCA to the combined body composition and microbial abundance data. Each point is a species, and text are body composition variables. Species are shaded in according to taxonomic family, and the size of points corresponds to the third CCA dimension. Text and point that are close to one another have similar patterns of abundance across samples.

they are associated. 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 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 4. For example, Figures 8 and 9 display the samples shaded in by android fat mass and Bacteroidaceae vs. Ruminococcaceae difference, respectively, as in Figures 4 and 9. 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 on covariance, and the body composition table alone has a large portion of its variance related to android fat mass.

2.3 Co-Inertia Analysis

Co-Inertia Analysis (CoIA) emerged in ecology to facilitate analysis of variation in species abundance as a function of environmental conditions [Dolédec 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,

$$\begin{aligned} \underset{u \in \mathbb{R}^{p_1}, v \in \mathbb{R}^{p_2}}{\text{maximize}} \; & u^T X^T Y v \\ \text{such that } & \|u\| = 1 \\ & \|v\| = 1, \end{aligned}$$

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 $\Sigma_{XX}^{\frac{1}{2}} u$ and $\Sigma_{YY}^{\frac{1}{2}} v$. It is in this sense that the CCA objective maximizes the correlation between scores, while CoIA maximizes the covariance.

The solutions $(u_k)_{k=1}^K$ 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 2.2, for CCA.

2.3.1 Example

We apply CoIA to the same data as used in Section 2.2.1, as CoIA also needs to estimate the covariance between tables, which is difficult when the number

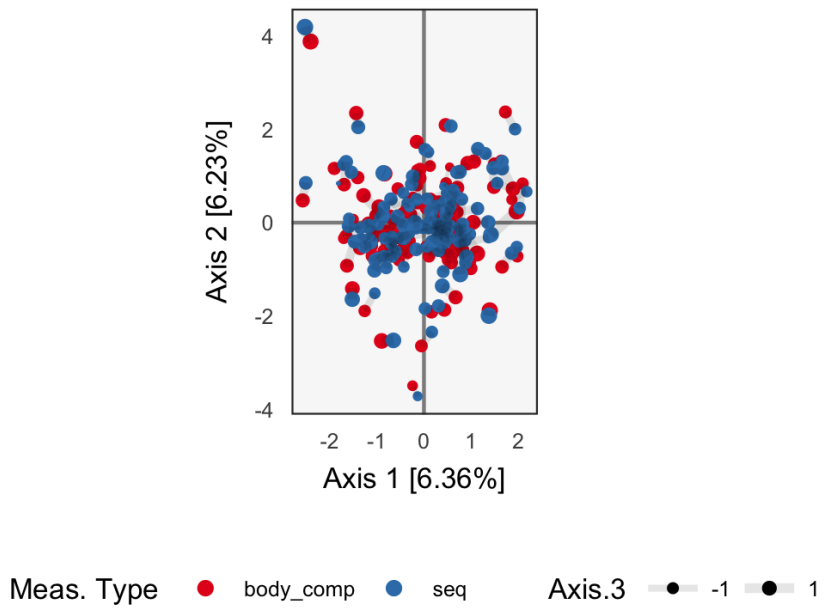

Figure 7: The scores resulting from CCA applied to the combined body composition and microbial abundance data. Each measurement type provides a different set of scores for the observed samples, and the similarity between the sets of scores is reflected in the CCA objective. Each sample appears as two separate circles, since there is one score per measurement type. A light line links the two appearances of each sample. Colors distinguish the two table sources. The higher the CCA objective, the shorter the links between pairs.

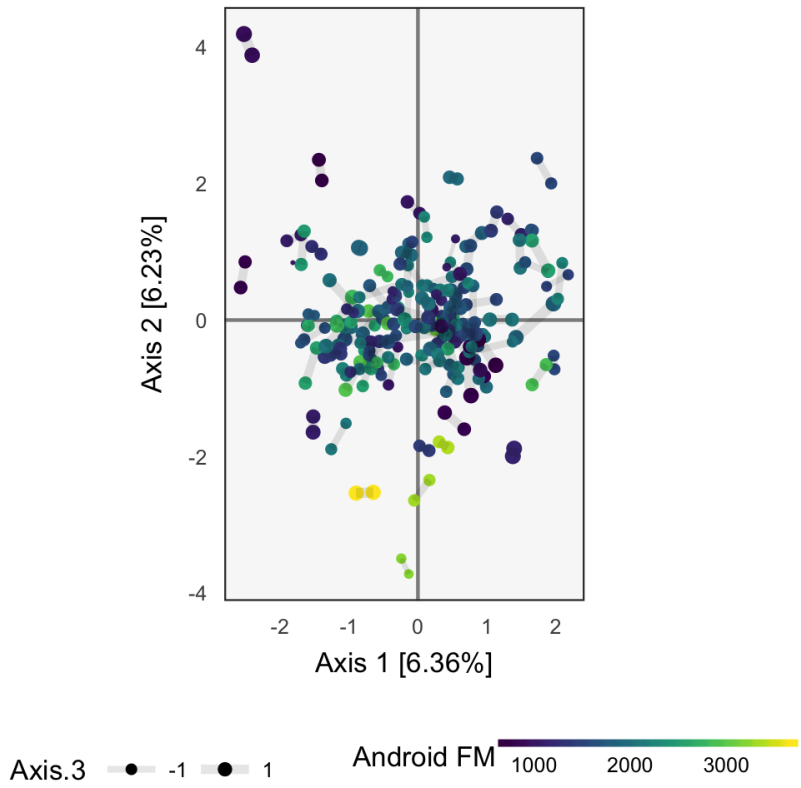

Figure 8: The scores resulting from CCA applied to the combined body composition and microbial abundance data, shaded in by android fat mass. The positions of points are identical to those in Figure 7, and other than the shading, the plots are read in the same way. The first two CCA dimensions suggest smooth variation across samples, according to amount of android fat mass.

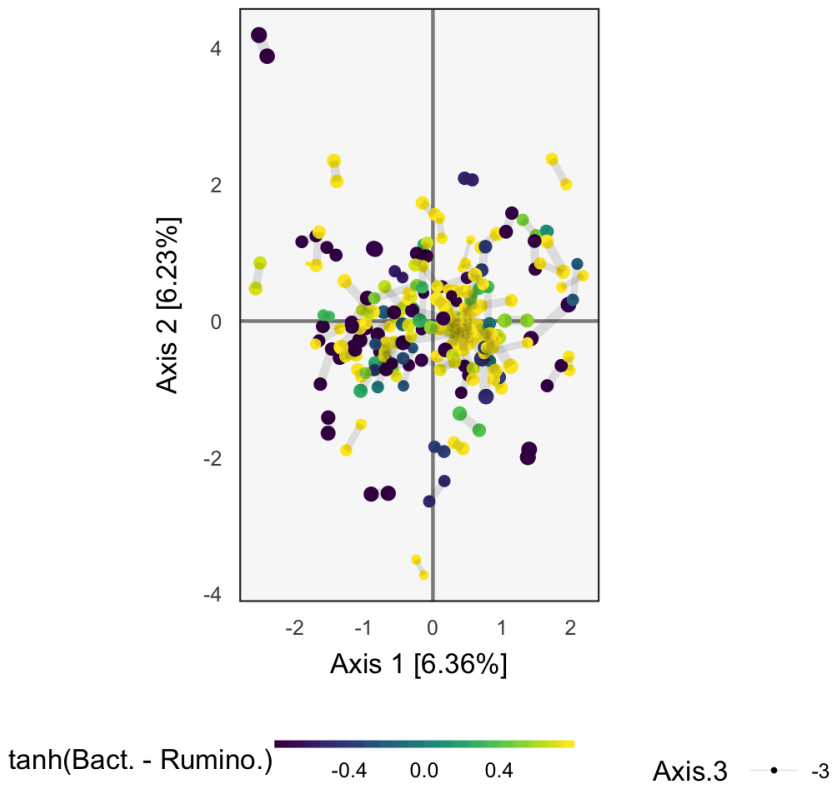

Figure 9: The same scores as Figure 8, but shaded now by the difference in Bacteroidaceae to Ruminococcaceae abundances.

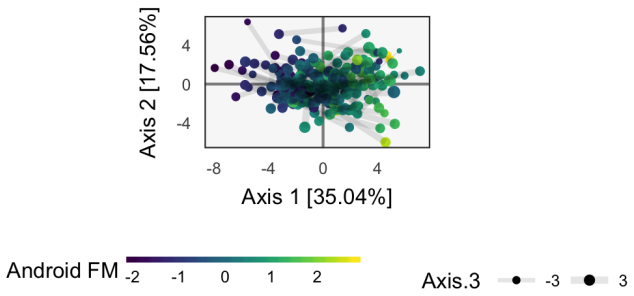

Figure 10: The normalized scores from each table, displayed simultaneously, as obtained by CoIA. This is the analog of Figure 8 from CCA, though the aspect ratio has been adjusted according to the relative amount of variance explained by the first two CoIA axes.

of species is large. The loadings are displayed in Supplementary Figure 29, and they are similar to those in Figure 6. However, the associated scores are quite different than those found using CCA. Compare Figure 10, which shades samples by android fat mass, or Supplemental Figure 30, which shades them by Bacteroides vs. Ruminococcaceae differences, with Figures 8 and 9 for CCA. The scores for CoIA are not so closely aligned across tables, but they exhibit a clearer gradient across supplemental variables. We find that the scores are not nearly as closely aligned as they are for CCA (see Figure 8), but that they are more strongly associated with variation in android fat mass, as in the concatenated PCA result of Figure 4. It is not clear whether this phenomena – 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.

2.4 MFA

MFA gives an alternative approach to producing scores and relating features across multiple tables [Pagès, 2014]. It can be understood as a refined version of the concatenated PCA described in Section 2.1 that res tables in a way that prevents any one table from dominating the resulting ordination Specifically, MFA is a concatenated PCA on the matrix

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

which reweights each table by its largest eigenvalue. 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" [Pages et al., 2004]. A K-dimensional representation of the l^{th} group is given by

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

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

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

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.

2.5 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, the Graph-Fused Lasso and Bayesian multitask learning, which are described in Sections 3.1, 3.3, 3.7, 3.8, and 3.9 respectively.

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 $z_i = V^T x_i \in \mathbb{R}^K$, that 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

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

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

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

so the residual covariance of y_i given z_i is

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

[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

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

or, equivalently,

$$\underset{V \in \mathbb{R}^{p_2 \times K} \text{ orthonormal}}{\text{maximize}} \operatorname{tr}\left(\hat{\Sigma}_{YX} V \left(V^T \hat{\Sigma}_{XX} V\right)^{-1} V^T \hat{\Sigma}_{XY}\right), \tag{6}$$

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

$$\hat{\Sigma}_{XY}\hat{\Sigma}_{YX} v_k = \lambda_k \hat{\Sigma}_{XX} v_k, \text{for } k = 1, \ldots, K$$

or more concisely,

$$\hat{\Sigma}_{XY}\hat{\Sigma}_{YX} V = \left(\lambda_1 \hat{\Sigma}_{XX} v_1 | \ldots | \lambda_k \hat{\Sigma}_{XX} v_k\right) = \hat{\Sigma}_{XX} V \Lambda,$$

where $\Lambda = \operatorname{diag}(\lambda_k) \in \mathbb{R}^{K \times K}$. A derivation for why this choice is optimal is provided in Supplemental Section 5.2.

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 X onto the y_j projects the y_j onto the span of (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 11.

Indeed, write the usual estimates for the covariance matrices of interest,

$$\begin{aligned}
\hat{\Sigma}_{YY} &= \frac{1}{n} Y^T Y \\
\hat{\Sigma}_{YX} &= \frac{1}{n} Y^T X \\
\hat{\Sigma}_{XX} &= \frac{1}{n} X^T X
\end{aligned}$$

and observe that the residual covariance of equation 4 can be expressed

$$\begin{aligned}
&\frac{1}{n}\left[Y^T Y - Y^T X V \left(V^T X^T X V\right)^{-1} V^T X^T Y\right] \\
=&\frac{1}{n}\left[Y^T Y - Y^T Z \left(Z^T Z\right)^{-1} Z^T Y\right] \\
=&Y^T (I - P_Z) Y,
\end{aligned}$$

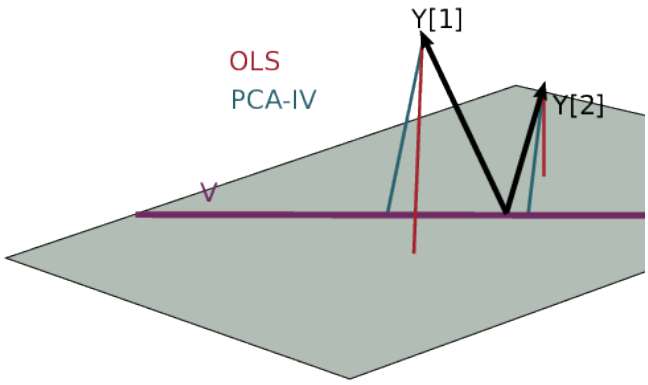

Figure 11: A geometric view of PCA-IV. The columns of the response Y are views as n-dimensional vectors. The grey plane is the span of X. Multivariate OLS simply projects the columns of Y onto the plane, while PCA-IV searches for a further subspace V on which to project all responses.

where $P_Z = Z\left(Z^TZ\right)^{-1}Z^T$ is the projection operator onto the columns of Z. Minimizing the trace of this matrix is equivalent to minimizing

$$\begin{aligned}\operatorname{tr}\left(Y^T\left(I-P_Z\right)Y\right) &= \operatorname{tr}\left(Y^T\left(I-P_Z\right)^T\left(I-P_Z\right)Y\right)\\ &= \left\|\left(I-P_Z\right)Y\right\|_F^2\\ &= \sum_{j=1}^{p_1}\left\|\left(I-P_j\right)y_{\cdot j}\right\|_2^2,\end{aligned}$$

which is exactly the sum of squared residuals from the columns of Y onto the span of the PCA-IV subspace, justifying the earlier geometric picture.

2.5.1 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 columns[4] of X and Y with the PCA-IV axes defining V.

We find that the scores, displayed in Supplementary Figures 11 are similar to those that found by the concatenated PCA of Section 2.1. 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 $\begin{pmatrix}Y & X\end{pmatrix}$, the fact that X has many more columns than Y means that the result is similar to a PCA on X alone.

[4]Geometrically, the angle between original columns and the subspace, in the sense of Figure 11.

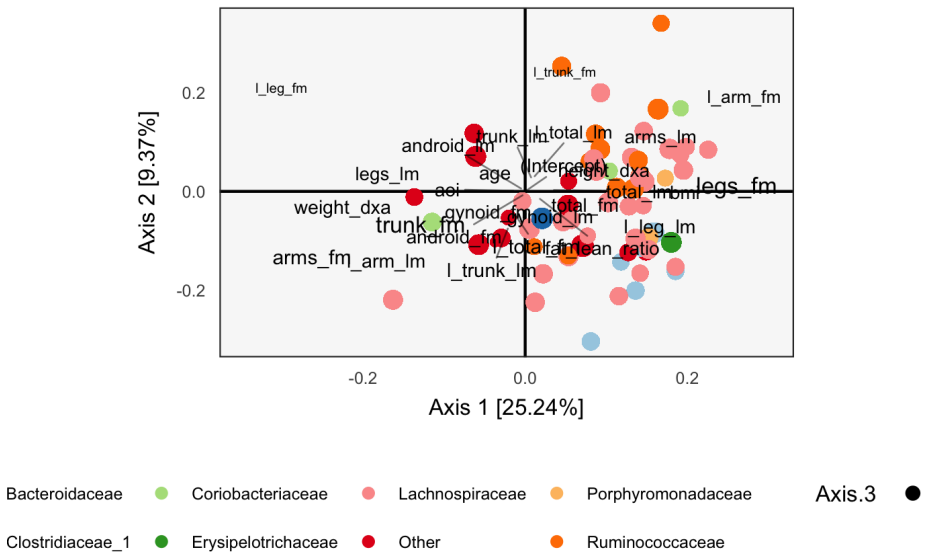

Figure 12: The loadings for PCA-IV can be interpreted like loadings from previous methods, for example Figure 3. Some of the relationships between variables seem less intuitive than those observed previously.

The loadings are given in Figure 12. 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.

2.6 Partial Triadic Analysis

Partial Triadic Analysis (PTA) gives an approach to working with multitable data when each table has the same dimension, $p_1 = p_2$ [Thioulouse, 2011]. Specifically, it gives a way of analyzing data of the form $(X_{\cdot\cdot l})_{l=1}^{L}$, where each $X_{\cdot\cdot l} \in \mathbb{R}^{n \times p}$. 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 $X_{i\cdot\cdot}$, $X_{\cdot j\cdot}$ and $X_{\cdot\cdot l}$. 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 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 $X_c = \sum_{l=1}^{L} \alpha_l X_{\cdot\cdot l} = X\alpha \in \mathbb{R}^{n\times p}$, where α (constrained to norm one) is chosen to maximize $\sum_{l=1}^{L} \alpha_l \left\langle \overline{X}, X_{\cdot\cdot l} \right\rangle$, a weighted average of inner-products[5] between each of the L tables and the naive-average table, $\overline{X} = \frac{1}{L}\sum_{l=1}^{L} X_{\cdot\cdot l}$.

The optimal α can be derived using Lagrange multipliers (see Supplemental Section 5.3), and leads to the compromise table,

$$X_c = \sum_{l=1}^{L} \frac{\left\langle \overline{X}, X_{\cdot\cdot l} \right\rangle}{\sqrt{\sum_{l'=1}^{L} \left\langle \overline{X}, X_{\cdot\cdot l'} \right\rangle^2}} X_{\cdot\cdot l}.$$

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

$$X_c = \sqrt{L} \sum_{l=1}^{L} \left\langle \overline{X}, X_{\cdot\cdot l} \right\rangle X_{\cdot\cdot l} = \sqrt{L}\overline{X},$$

a scaled version of the mean.

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 $\overline{X}$, X_c upweights more highly represented tables.

2.7 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.

[5] We are using $\langle A, B\rangle = \mathrm{tr}\left(A^T B\right)$.

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

2.8 Reduced-rank regression

Reduced-rank regression is an approach to multiresponse regression which pools information across responses [Izenman, 1975, Mukherjee and Zhu, 2011]. Compared to performing separate regressions for each response, this pooling can lead to meaningful performance improvements. Further, reduced-rank estimates can often be more interpretable.

Suppose we have collected p_1 responses and p_2 features across n samples, $y_i \in \mathbb{R}^{p_1}$ and $x_i \in \mathbb{R}^{p_2}$, respectively. Our goal is to use this training data to predict the response y^* given a new sample x^*. Arrange these data into two matrices, $Y \in \mathbb{R}^{n \times p_1}$ and $X \in \mathbb{R}^{n \times p_2}$.

The simplest approach to this problem is to fit a coefficient matrix $B \in \mathbb{R}^{p_2 \times p_1}$ relating the p_2 features to the p_1 response coordinates by minimizing $\|Y - XB\|_F^2$. A slight modification supposes that the responses might be correlated, and instead optimizes a whitened version of the problem[6], $\| (Y - XB) \hat{\Sigma}_{YY}^{-\frac{1}{2}} \|_F^2$

The optimal B for these two approaches are $\left(X^T X\right)^{-1} X^T Y$ and $\left(X^T \hat{\Sigma}_{YY} X\right)^{-1} X^T \hat{\Sigma}_{YY} Y$ respectively. These simply concatenate coefficients from p_1 independent linear regressions, one per response dimension.

This is not a very satisfactory solution, because we imagine there is information to share across the different response dimensions: we should be able to improve performance compared to parallel univariate regressions. Towards this goal, consider that while there may be p_1 responses, their effective dimension may be relatively low. Reduced-rank regression formalizes this with an explicit constraint on the rank of B, defining an estimate $\hat{B}^{\text{rr}}$ by the optimal value of

$$\begin{aligned} &\underset{B \in \mathbb{R}^{p_2 \times p_1}}{\text{minimize}} \| (Y - XB) \Sigma_{YY}^{-\frac{1}{2}} \|_F^2 \\ &\text{such that } \operatorname{rank}(B) \leq K, \end{aligned} \tag{7}$$

for some $K < p_1 \wedge p_2$.

[6] This can also be viewed as using a Mahalanobis distance.

The optimal value is given by $\hat{B}^{\text{ols}} V_K V_K^-$, where the columns of V_K are the top K response CCA directions and V_K^- denotes the pseudoinverse of V_k. The derivation is provided in Supplementary Section 5.4.

Therefore $\hat{Y}^{\text{rr}} = X\hat{B}^{\text{rr}} = P_X Y V_k V_k^-$, which means that the reduced-rank fits can be obtained by first projecting the columns of Y onto the top K response canonical directions, and then projecting these pooled Y onto the span of X. If the Y had not been pooled, then the projection onto the span of the X's is exactly the independent linear regressions. Hence, we have a clear geometric picture of the effect of the reduced-rank constraint.

3 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 high-dimensional response matrices, while Canonical Correspondence Analysis (CCpnA) was facilitates 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.

3.1 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 $\hat{\varphi}_j$ from the model $u_{1i} = \alpha_{0j} + \varphi_j x_{ij}$, for $i = 1, \ldots, p_1$. The first PLS direction is defined as $z_1 = X\hat{\varphi}_1$. To generate subsequent directions, orthogonalize both Y and X with respect to the current directions, 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, statistical interpretations 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

$$\begin{aligned} &\underset{z}{\text{maximize}} \text{Corr}_{\mathbb{P}^{YX}} \left[x_i^T z_k, y_i\right] \text{Var}_{\mathbb{P}^X} \left(z^T x_i\right) \\ &\text{such that } z^T X^T X z_j = 0 \text{ for all } j \leq k-1 \\ &\|z\|_2 = 1. \end{aligned} \tag{8}$$

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. For $p_1 > 1$, this objective can be generalized to the total covariance across response dimensions, $\sum_{j=1}^{p_1} \text{Cov}_{\mathbb{P}^{YX}} \left[x_i^T z_k, y_{ij}\right]$ [Chun and Keleş, 2010].

An alternative interptetation, due to [Gustafsson, 2001], is that PLS fits a particular latent variable model. Suppose $\xi_i = \left(\xi_i^s, \xi_i^X\right)$ 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

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

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 implictly estimates.

3.2 Sparse PLS

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_i, 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 Keleş, 2010], and we briefly review this method here.

Directly regularizing the multiresponse version of the PLS optimization 8 leads to the problem

$$\begin{aligned} &\underset{z_k}{\text{maximize}} \sum_{j=1}^{p1} \text{Cov}_{\mathbb{P}^{YX}} \left[x_i^T z_k, y_{ij}\right] \\ &\text{such that } z^T x^T x z_j = 0 \text{ for all } j \leq k-1 \\ &\|z_k\|_2 = 1 \\ &\|z_k\|_1 \leq \lambda, \end{aligned}$$

which can be applied to real data by replacing the objective with its sample version, $z_k^T M z_k$, where $M = X^T Y Y^T X$. In this sample version, the problem falls into the Penalized Matrix Decomposition framework of [Witten et al., 2009], reviewed in Section 3.5.

However, Chun and Keleş [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 maximize the PLS-defining covariance and another, (z_k), that approximates the first set by a sparser alternative. Formally, consider

$$\begin{aligned} \underset{z_k, a_k}{\text{minimize}} & -\kappa \|a_k\|_M^2 + (1-\kappa)\,\|z_k - a_k\|_M^2 \\ \text{such that } & \|a_k\|_2^2 = 1 \\ & \|z_k\|_1 \le \lambda_1 \\ & \|z_k\|_2 \le \lambda_2, \end{aligned} \tag{9}$$

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. κ trades off the importance of the two components of the objective. 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.

To develop some intuition for this optimization problem, we review the motivation behind the analogous sparse PCA objective. Suppose we have the SVD, $X = UDV^T$. It is a simple observation that the k^{th} principal component is proportional to the solution of a ridge regression onto $y_k = u_k d_k$,

$$\hat{v}_k \propto \arg\min_{v_k} \|y_k - X v_k\|_2^2 + \lambda \|v_k\|_2.$$

Indeed, according to the usual ridge regression formula, the minimizer is

$$\begin{aligned} \left(X^T X + \lambda I\right)^{-1} X^T y_k &= \left(V^T D^2 V + \lambda I\right)^{-1} V D U^T u_k d_k \\ &= V \left(D^2 + \lambda I\right)^{-1} D e_k d_k \\ &= \frac{d_k^2}{d_k^2 + \lambda} v_k. \end{aligned}$$

A less obvious fact is that still this connection between PCA continues to hold even without direct access to the u_k and d_k. Indeed, it turns out that in the problem

$$\begin{aligned} \text{minimize}_{A,V} & \sum_{i=1}^{n} \|x_i - A V^T x_i\|_2^2 + \lambda \sum_{k=1}^{K} \|v_k\|_2 \\ \text{subject to } & A^T A = I_K, \end{aligned} \tag{10}$$

the optimal $\hat{V}$ corresponds exactly to the top K right singular vectors of X. The idea of Zou et al. [2006] is to induce sparsity on v_k by constraining their

ℓ^1 norm, resulting in sparse principle component directions. The objective 9 is the analog of this problem when setting $z_k = v_k$, $\kappa = \frac{1}{2}$ and $M = X^T X$.

Reformulating the optimization 10 suggests a simple algorithm for solving it. First, notice that if we extend $\tilde{A} = \begin{pmatrix} A & A^\perp \end{pmatrix}$ so that it is orthonormal, then since norms are preserved under orthonormal rotations,

$$\begin{aligned}
\sum_{i=1}^{n} \left\| x_i - AV^T x_i \right\|_2^2 &= \|X - XVA\|_F^2 \\
&= \left\| \begin{pmatrix} XA - XV \\ XA^\perp \end{pmatrix} \right\|_F^2 \\
&= \left\| \begin{pmatrix} XA - XV \\ XA^\perp \end{pmatrix} \tilde{A} \right\|_F^2 \\
&= \sum_{k=1}^{K} \|Xa_k - Xv_k\|_2^2 + C
\end{aligned}$$

where C is constant in A, and hence the sparse PCA objective can be written as

$$\begin{aligned}
\text{minimize}_{A,V} & \sum_{k=1}^{K} \|Xa_k - Xv_k\|_2^2 + \sum_{k=1}^{K} \lambda_2 \|v_k\|_2 \\
\text{such that } & \|a_k\|_2 = 1 \\
& \|v_k\|_1 \leq \lambda_1 \\
& A^T A = I_K.
\end{aligned}$$

For a fixed A, this is an elastic-net regression. For fixed V, considering that $\sum_{k=1}^{K} \|Xa_k - Xv_k\|_2^2 = \|XA - XV\|_F^2$, this is a Procrustes rotation problem. Hence, a local minimum can be found by alternating these two steps until convergence.

3.2.1 Example

Next we apply the 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. We subset to female subjects and filter species according to a K-over-A filter with $K = 7\%$ of samples and $A = 5$. This leaves 372 species over 119 participants. All species abundances are variance-stabilized using the approach of Anders and Huber [2010b]. We cross-validate with 5 folds, searching through a grid over $K \in \{4, \ldots, 8\}$ and $\lambda_1 \in \{0, 0.05, \ldots, 0.7\}$. 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 13 would look identical. The predictive accuracy is poor, which is unsurprising considering the spike at 0 in

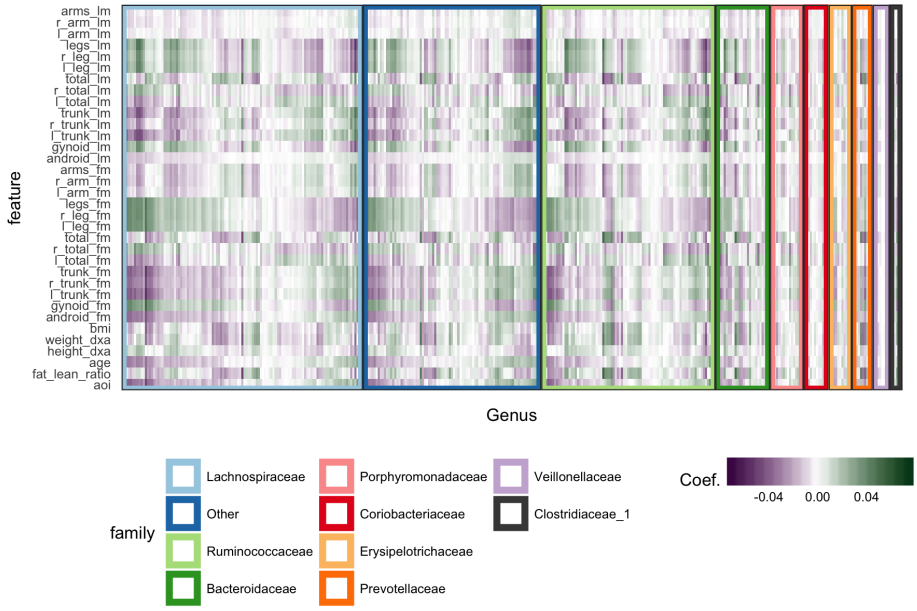

Figure 13: Coefficients learned by SPLS. Each row is a response dimension, which is a body composition variable. Each column is associated with a species. The shading within each cell corresponds to the SPLS coefficient for that species-response pair. Green and purple cells are positive and negative 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.

the abundances histogram – the held out error is ≈ 1.29, after having scaled and centered the body composition variables.

Figure 13 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, $Y = XB + E = XZQ^T + E$. 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 Figures 14 and 15, we display the abundances X for species against android and gynoid fat mass, respectively. The species are chosen according to whether the two-

dimensional coefficient across android and gynoid fat mass has large norm[7]. 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. This suggests that a logistic regression version of SPLS [Chung and Keles, 2010], applied to the presence-absence transformed version of this data may fit the data just as well, with the advantage of being somewhat more interpretable.

3.3 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. 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.

CCpnA produces low-dimensional representations of both the rows and columns of Y (the sites 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,

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

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

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

$$D_r = \operatorname{diag}\left(F 1_{p_1}\right) \in \mathbb{R}^{n \times n}.$$

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

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

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

$$P_X = D_r^{\frac{1}{2}} X \left(X^T D_r X\right)^{-1} X^T D_r^{\frac{1}{2}} \in \mathbb{R}^{n \times n}.$$

[7] Specifically, $\left\| \begin{pmatrix} \beta_{android} \\ \beta_{gynoid} \end{pmatrix} \right\|_2 > 0.065$.

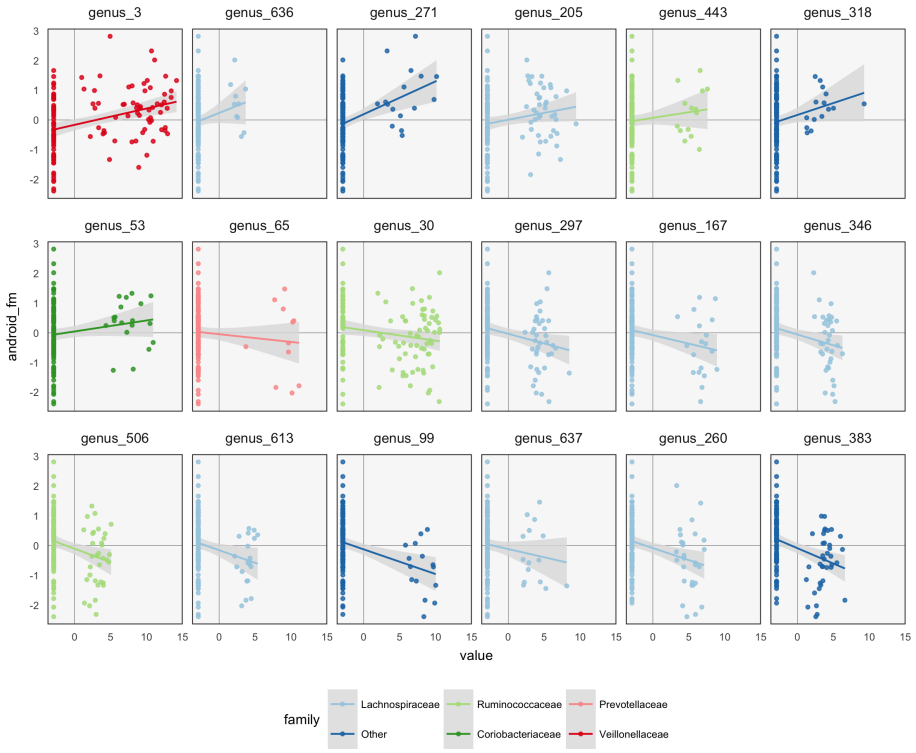

Figure 14: A subset of the species whose coefficients for total lean and fat mass are high in magnitude. Each panel represents one species, and each point is a samples $\times$ species combination. The x-axis gives the abundance of that species across samples, and the y-axis gives android fat mass of the associated sample. Colors indicate taxonomic family membership. Panels are sorted from most positive to most negative association. Within each panel, a linear smooth is given, independent of the estimated coefficient.

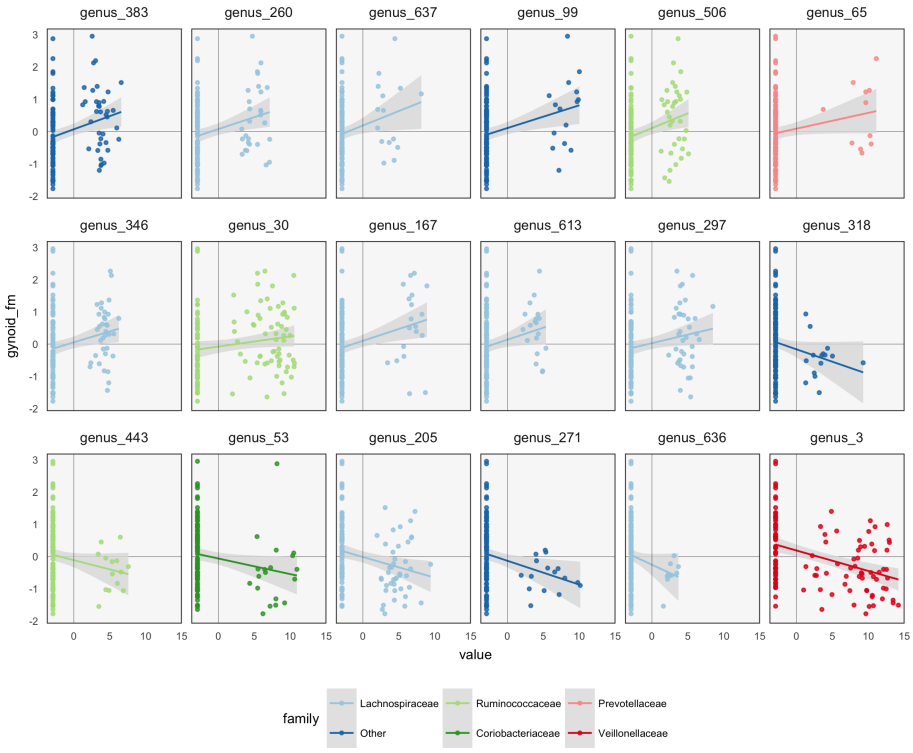

Figure 15: The analog of Figure 14, but with gynoid, rather than android, fat mass along the y-axis.

With this notation, compute an SVD,

$$D_r^{-\frac{1}{2}} \left(F - F1_p 1_p^T F\right) D_c^{-\frac{1}{2}} P_X = UDV^T,$$

and define row and column scores Z and Q by

$$Z = D_r^{-\frac{1}{2}} UD$$
$$Q = D_c^{-\frac{1}{2}} V^T D.$$

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 and Hastie [1987], Greenacre [1984], 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 sites 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 site scores. That is,

$$z_i \propto \frac{1}{f_{i\cdot}} \sum_{j=1}^{p_1} f_{ij} q_{ij}$$
$$q_j \propto \frac{1}{f_{\cdot j}} \sum_{i=1}^{n} f_{ij} z_{ij},$$

or, in matrix form,

$$Z \propto \operatorname{diag}\left(F1_p\right)^{-1} FQ^T$$
$$Q \propto \operatorname{diag}\left(F^T 1_n\right)^{-1} Z.$$

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[8] – it doesn't use information in the environmental 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 $Q' \propto \operatorname{diag}\left(F^T 1_n\right)^{-1} F^T Z$.
   (b) Project $Q = P_X Q'$.
   (c) Solve $Z \propto \operatorname{diag}\left(Z1_p\right)^{-1} FQ^T$.

[8] It in fact gives the solution to the Correspondence Analysis problem (the similarity is the reason for the name Canonical *Correspondence Analysis*).

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 environmental gradient, for example, between warm-dry and cold-wet sites. For each of the p_1 species, define a normal density over the environmental 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 environmental gradient, giving a univariate $f_j^\alpha(z_i) = \mathcal{N}\left(z_i|\alpha^T\mu_j, \alpha^T\Sigma_j\alpha\right)$ for each species. The z_i represent the scores for species i along the environmental gradient α.

The generative model views species-site pairs one at a time. For each pair involving site 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 LDA model.

To use this idea to compute scores, we need to estimate the environmental 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

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

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.

3.4 Kernel CCA

Since CCA is based on correlation, it can only summarize linear relationships across tables – Kernel CCA (KCCA) is a modification that is sensitive to non-linear associations [Akaho, 2006, Bach and Jordan, 2003, Lanckriet et al., 2004]. It does this by implicitly lifting the original data into a richer feature space, with the hope that nonlinear associations in the original data become linear associations in the richer space – this is often called the "kernel trick" [Schölkopf, 2001]. Informaly, KCCA is the algorithm that emerges after applying the kernel trick to CCA.

More precisely, let $\varphi^Y : \mathbb{R}^{p_1} \to H^Y$ and $\varphi^X : \mathbb{R}^{p_2} \to H^X$ be mappings from the features directly measured by X and Y into richer spaces H^Y and H^X. For example, the φ might map vectors into an expansion of all polynomial products of original feature values, up to some fixed degree, this mapping is called the polynomial kernel. As in CCA, let $\mathbb{P}^{YX}$ denote the sampling distributions associated with the two tables, and write x_i and y_i for generic draws from these distribution.

In the same way that the first CCA direction maximizes the covariance between linear combinations $u^T x_i$ and $v^T y_i$, KCCA maximizes the correlation between the more general inner products, $z_i(u) = \langle u, \varphi(x_i) \rangle$ and $z_i(v) = \langle v, \varphi(y_i) \rangle$,

$$\begin{aligned} &\underset{u \in H^X, v \in H^Y}{\operatorname{argmax}} \operatorname{Cov}_{\mathbb{P}^{YX}} \left[z_i(u), z_i(v) \right] \\ &\text{subject to } \operatorname{Var}_{\mathbb{P}^X} \left(z_i(u) \right) = \operatorname{Var}_{\mathbb{P}^Y} \left(z_i(v) \right) = 1. \end{aligned} \tag{11}$$

As is, the problem is not well-posed, and it is necessarily to regularize. The regularized Lagrangian associated with the optimization in equation 11 is

$$\begin{aligned} &\operatorname{Cov}_{\mathbb{P}^{YX}} \left[z_i(u), z_i(v) \right] - \frac{\rho^X}{2} \operatorname{Var}_{\mathbb{P}^X} \left(z_i(u) \right) - \frac{\rho^Y}{2} \operatorname{Var}_{\mathbb{P}^Y} \left(z_i(v) \right) + \\ &\frac{\lambda^X}{2} \operatorname{Pen}(u) + \frac{\lambda^Y}{2} \operatorname{Pen}(v), \end{aligned} \tag{12}$$

where $\operatorname{Pen}(x)$ is some regularizer, the ℓ^1 or ℓ^2 norm, for example.

The optimal u and v must lie in the spans of $\left(\varphi^X(x_i) \right)_{i=1}^n$ and $\left(\varphi^Y(y_i) \right)_{i=1}^n$, respectively, since directions orthogonal to these subspaces cannot improve the correlation in the objective. Therefore,

$$\begin{aligned} u &= \Phi^X \alpha^X \\ v &= \Phi^Y \alpha^Y, \end{aligned}$$

for some α^X, α^Y, where $\Phi^X \in \mathbb{R}^{n \times \dim\left(H^X\right)}$ has i^{th} row $\varphi^Y(x_i)$ and Φ^Y is defined similarly.

Substituting this into the Lagrangian in equation 12, it becomes clear that only the cross-products ${\Phi^X}^T \Phi^X$ and ${\Phi^Y}^T \Phi^Y$ appear. Since these inner products can be written as kernel matrices – call them K^X and K^Y – the optimization can be fully expressed in terms of kernels, without reference to the original X or Y. It can then be shown that the optimal α^X and α^Y are the solutions to the generalized eigenvalue problem,

$$\begin{pmatrix} 0 & K^X K^Y \\ K^Y K^X & 0 \end{pmatrix} \begin{pmatrix} \alpha^X \\ \alpha^Y \end{pmatrix} = \rho \begin{pmatrix} \left(K^X + \lambda^X I_{p_1} \right)^2 & 0 \\ 0 & \left(K^Y + \lambda^Y I_{p_2} \right)^2 \end{pmatrix} \begin{pmatrix} \alpha^X \\ \alpha^Y \end{pmatrix}.$$

A geometric interpretation of Kernel CCA is given in Kuss and Graepel [2003], which translates the Euclidean picture associated with CCA to the more general RKHS setting. Often, however, KCCA results can be difficult to use in exploratory analysis, because only sample scores are provided. Eigenvectors are never computed in the spaces H^X and H^Y, so it's not possible to make a biplot. Consequently, KCCA scores are typically interpreted using supplementary characteristics of the samples.

3.5 Penalized Matrix Decomposition

In high-dimensional settings, sparsity is a desirable property, both for 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, and in fact, it is impossible to fit a 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 reguarlization 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, 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

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

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

$$\begin{aligned} &\underset{u \in \mathbb{R}^n, v \in \mathbb{R}^p, d \in \mathbb{R}}{\text{minimize}} \|X - duv^T\|_2^2 \\ &\text{subject to } \|u\|_2^2 = \|v\|_2^2 = 1 \\ &\qquad \text{Pen}_u(u) \leq \mu_1 \\ &\qquad \text{Pen}_v(v) \leq \mu_2, \end{aligned} \tag{13}$$

where Pen_u and Pen_v are arbitrary constraints on u 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 and eigenvectors $(u_k)_{k=1}^K$ and $(v_k)_{k=1}^K$ for $K > 1$, define u_k and v_k as the optimizers of the problem 13 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 $X^1 = X$. The effect of regularization is illustrated in Figure 16.

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 $\|A\|_F^2 = \text{tr}\left(A^T A\right)$ along with the linearity and the cyclic properties of the trace, the objective in 13 can be rewritten, using

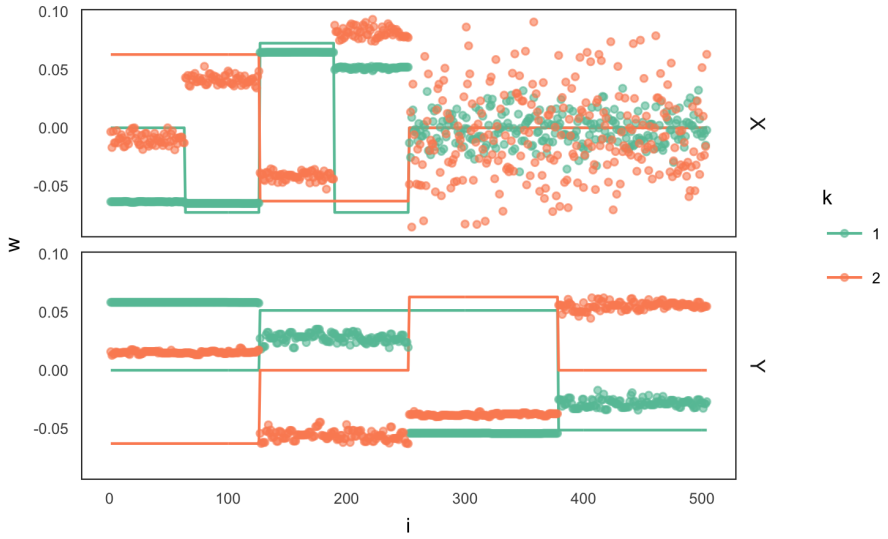

Figure 16: An example of the PMD CCA solution in a high-dimensional source separation problem, based on the simulation experiment in [Witten et al., 2009]. The solid lines indicate true "sources" underlying observed data tables $X \in \mathbb{R}^{20\times504}$ and $Y \in \mathbb{R}^{20\times504}$. Specifically, X and Y are defined as random linear combinations of two sources each, which are then corrupted with noise. The recovered sources from CCA between X and Y using a fused-lasso penalty are plotted as points. The fused-lasso regularization parameter was chosen by cross-validation. The long flat lines in the recovered solutions reflect this penalty. Note that PMD is able to recover the positions at which changepoints in the source signal occur, but that sources are nonidentifiable without further constraints. Further, in the table X, a region with truly zero signal is estimated very poorly, for reasons that are unclear. A scatterplot view of this same data is provided in Figure 33.

$\equiv$ to mean equality up to terms constant in u and v,

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

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 13 can be found as the optimizers of

$$\begin{aligned}&\underset{u\in\mathbb{R}^n, v\in\mathbb{R}^p}{\text{maximize}}\ u^T X^T v \\ &\text{subject to } \|u\|_2^2 = \|v\|_2^2 = 1 \\ &\qquad\qquad \operatorname{Pen}_u(u) \le \mu_1 \\ &\qquad\qquad \operatorname{Pen}_v(v) \le \mu_2\end{aligned} \tag{14}$$

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] argues for diagonalized CCA, in which the variance constraints are replaced by unit norm constraints, and sparsity-inducing ℓ^1-constraints are added,

$$\begin{aligned}&\underset{u\in\mathbb{R}^{p_1}, v\in\mathbb{R}^{p_2}}{\text{maximize}}\ u^T\hat{\Sigma}_{XY}v \\ &\text{subject to } \|u\|_2^2 = \|v\|_2^2 = 1 \\ &\qquad\qquad \|u\|_1 \le \mu_1 \\ &\qquad\qquad \|v\|_1 \le \mu_2\end{aligned}$$

which is exactly of the form of equation 14 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 $c_1^{(l)}$.

3.5.1 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. Our reasoning is that sparsity within species loadings is more important than sparsity across body composition variables, because the microbiome data is more high-dimensional. 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 Figures 17 and 18, respectively. 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, species 271 seems to be more common among people with higher android and lower gynoid fat mass.

The associated scores are displayed in Figure 18, shaded in 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 folow-up these displays by focusing on species that seemed related to the CCA axes. In Figure 19, 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 17. 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 generally 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. Instead, a few isolated members of a few different taxonomic groups seem to be associated with android fat mass.

3.6 Multitable Mixed-Membership

In Section 2.2, 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 2.2 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.

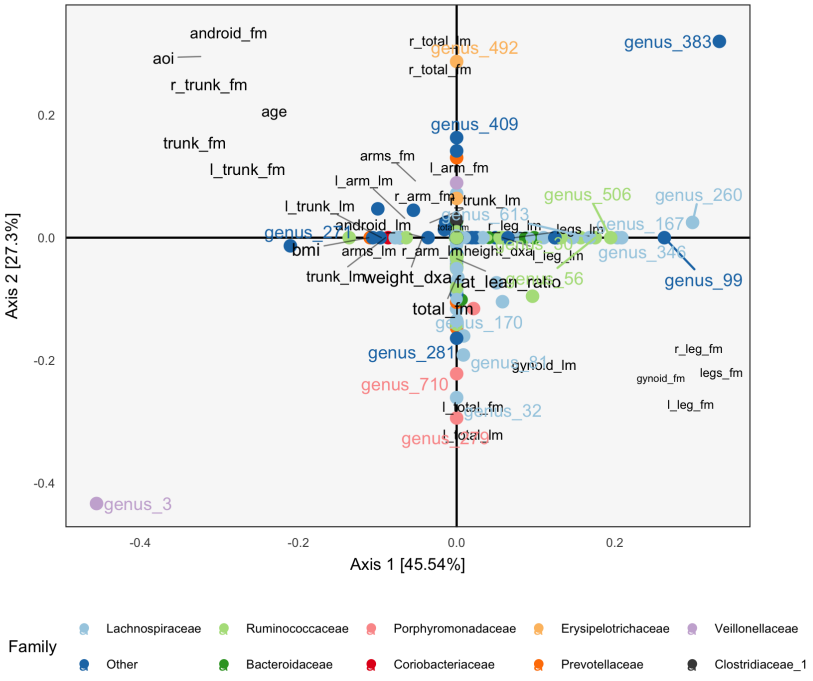

Figure 17: Body composition and species loadings produced by sparse CCA, for variables with at least one nonzero coordinate. Each point corresponds to a species loading, and is shaded in by taxonomic family. Species with loadings far from the origin are also annotated with their names. Black text are loadings for body composition variables. The size of points and text reflects the contribution of the third CCA dimension. Many loadings have at least one dimension that is exactly zero, due to ℓ^1-regularization. There appear to be a gradient with android and gynoid weight, running from left to right, and this is confirmed in Figures 18 and 19.

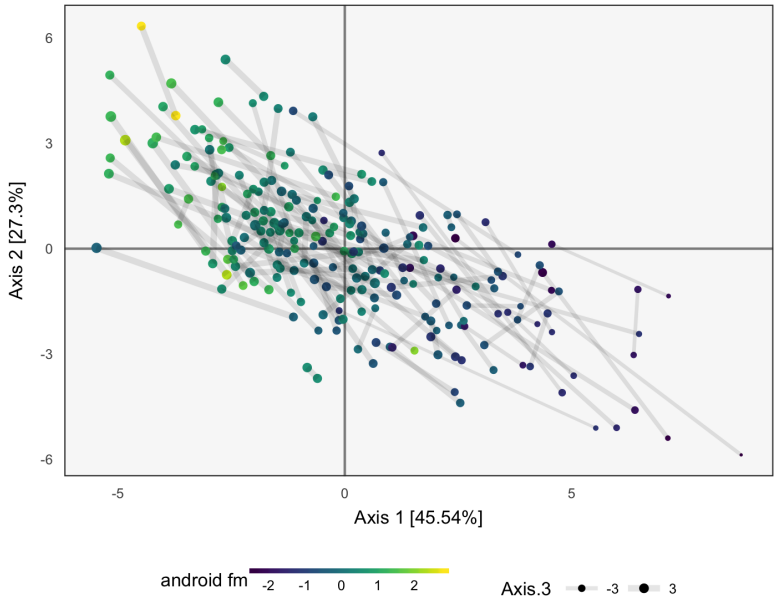

Figure 18: Sample scores provided by sparse CCA. Each point is a sample, positioned at their coordinates with respect to the first two learned sparse CCA directions. Points are shaded in according to android fat mass, and their sizes are set according to the third sparse CCA direction's contribution. Evidently, the first two directions reflect a gradient across android fat mass, suggesting that this is a substantial contributor to covariation across microbiome and body composition tables.

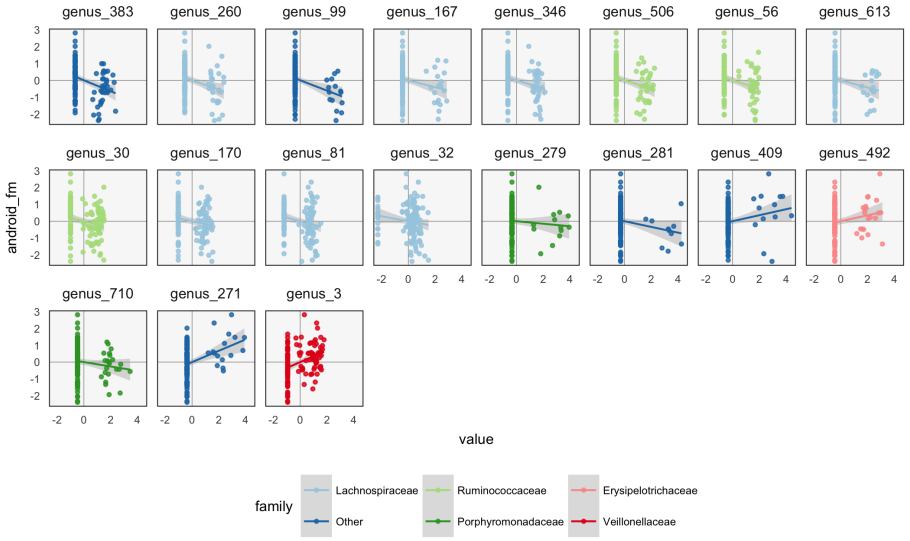

Figure 19: A more focused view of the species with high loadings in Figure 17. Each panel corresponds to a species. Points are shaded in according to each species' taxonomic family. The x-axis within panels corresponds to variance-stabilized species abundance, while the y-axis gives android fat mass. A linear smooth is provided to summarize the direction of associations. Panels are arranged according to the size of that species' loading onto the first sparse CCA axis. The presence of certain species seems to correspond to increased or decreased levels of android fat mass.

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 a essentially a Gaussian factor analysis model, $y_i|\xi_i^X, \xi_i^Y \sim \mathcal{N}\left(B^y\xi_i^s + W^y\xi_i^y, \sigma^2 I_{p_2}\right)$ with a spherical Gaussian prior on ξ_i^X, ξ_i^Y. 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 the other two approaches are potentially interesting as well. Formally, this model supposes that counts are drawn according to

$$\begin{aligned} x_i|\left(\theta_k\right) &\sim \text{Mult}\left(x_i|N_i, \sum_{k=1}^K \theta_{ik}\beta_k\right) \\ \theta_i &\sim \text{Dir}\left(\alpha\right) \\ \beta_k &\sim \text{Dir}\left(\gamma\right), \end{aligned}$$

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

$$\begin{aligned} x_i|\xi_i^s, \xi_i^x &\sim \text{Mult}\left(x_i|N_i, \mathcal{S}\left(B^X\xi_i^s + W^X\xi_i^x\right)\right) \\ \xi_i^s &\sim \mathcal{N}\left(\xi_i^s|0, \tau^2\right). \end{aligned}$$

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

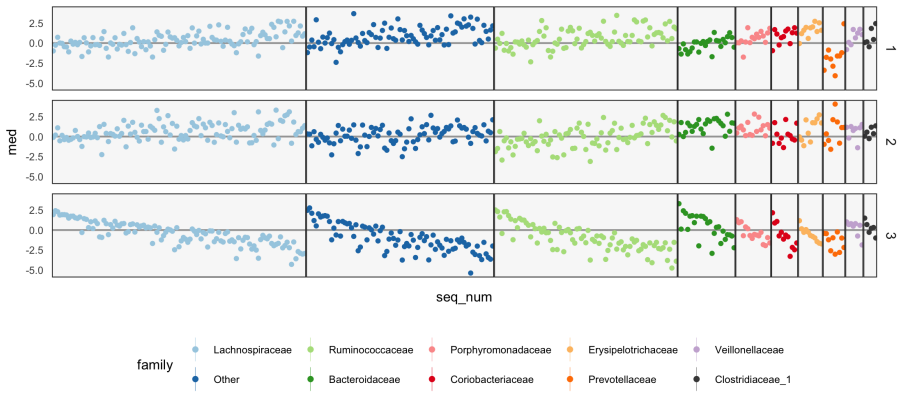

Figure 20: Table-specific loadings for different species. As in Figure 21, each row is a loading dimension, columns are features (species in this case), and intervals summarize posterior samples for the associated loading parameter, W^X_{jk}. Species are sorted from most to least abundant, within each taxonomic family. Caution must be exercised when interpreting these loadings, as loadings are invariant under rotations and reflections.

3.6.1 Example

We illustrate this multitable mixed-membership approach on the WELL-China data. We choose $K = 3$ 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^s_i) and unshared (ξ^x_i, ξ^y_i) 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 table specific loadings W^X and W^Y, provided in Figures 20 and 21. Note that there is no notion of variance explained by different axes, and we use an aspect ratio of 1 throughout.

Figure 20 summarizes table-specific variation in bacterial abundances. Invariance under rotation and reflection complicates interpretation of these loadings. If we flip the sign of all the loadings axes, then the more abundant species have larger loadings. The main difference between the first and second loadings is the rate of decay in frequencies, especially among Lachnospiraceae and and Ruminococcaceae. For example, topic 1 seems to species from these taxonomic families that are not very abundant. The second topic also seems to have generally lower loadings (hence higher abundances) for the Bacteroidaceae, and higher loadings (lower abundances) for Prevotella, compared to other topics. 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 this taxa.

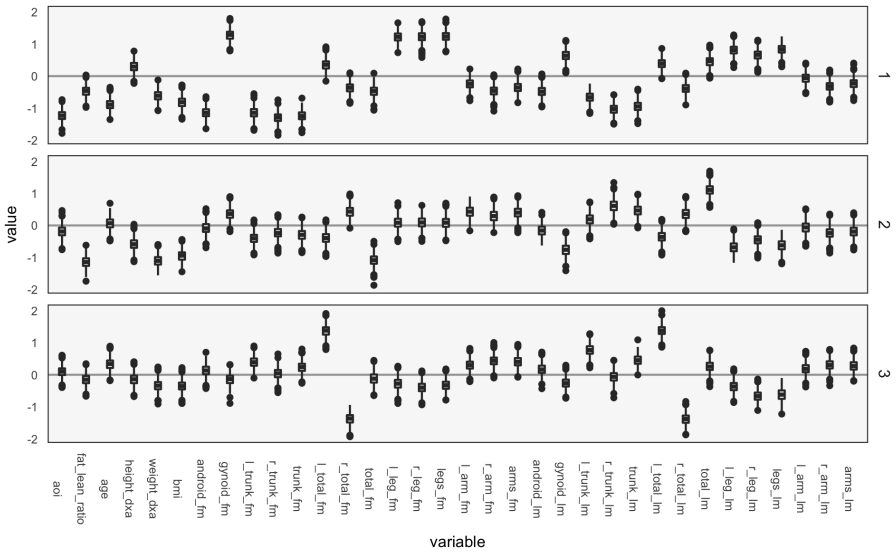

Figure 21: Table-specific loadings for the body composition variables. Each row is one loading dimension, columns are features, and boxplots summarize posterior samples for the associated loading parameter, W^Y_{jk}.

Figure 21 suggests that the first and third axes of W^Y captures variation between overall and android vs. gynoid fat mass, which can be used to interpret the scores in Supplementary Figure 34. The first axes 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 axes captures difference between mass in the trunk versus arms and legs.

These summaries could have been obtained by analyzing each table separately. More interest lies in covariation between the two tables, captured by the shared scores ξ^s_i and loadings B^X, B^Y. The shared body composition loadings are given in Figure 23. The first row of loadings again differentiates android and gynoid and fat mass variables. The second axis slightly differentiates left and right total fat mass, though the effect is less pronounced than in the table-specific loadings.

The bacterial abundance loadings are given in Figures 20 and 22. 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 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 3.5. It seems that inference of a probabilistic model with shared and unshared parameters is more difficult than

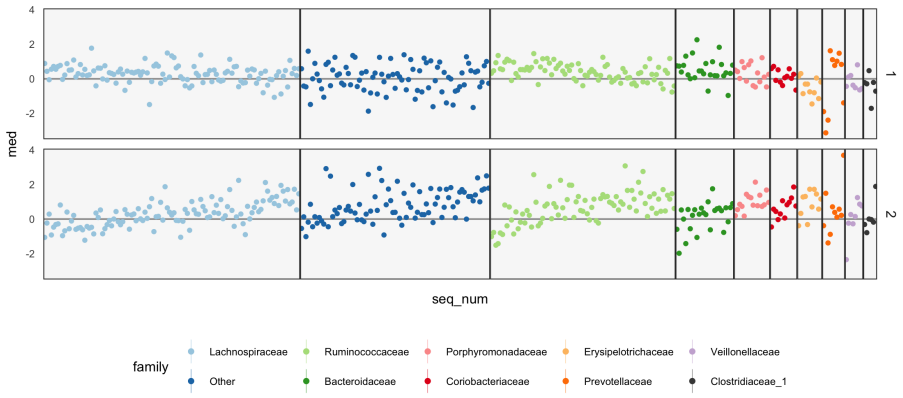

Figure 22: Cross-table loadings for species, as in Figure 20, but for B^X_{jk}.

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 the scores B^X, B^Y, W^X, and W^Y, respectively, in the spirit of [Archambeau and Bach, 2009].

3.7 Curds & Whey

The Curds & Whey (C&W) procedure is a "soft" version of reduced-rank regression, differentially shrinking the 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

$$\hat{Y} = P_X Y V \Lambda V^{-1}, \tag{15}$$

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 15 a priori. This derivation is provided in Supplementary Section 5.5.

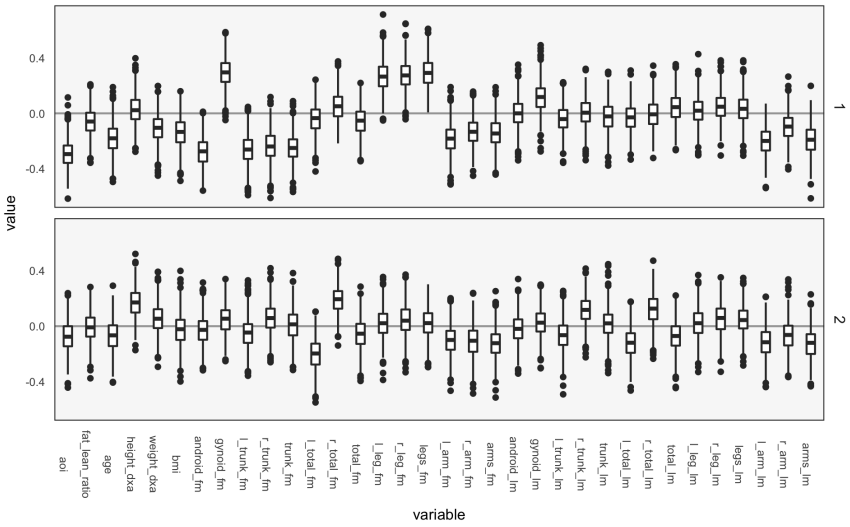

Figure 23: Cross-table loadings for the body composition variables, read exactly as in Figure 21, but for for the parameters B^Y_{jk}.

3.8 Graph-Fused Lasso

Chen et al. [2010] describe an approach to multiresponse regression that 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 $V = \{1, \ldots, p_1\}$. Each edge e is associated with a weight, $r(e)$, giving the correlation between the pair of responses.

The graph-fused lasso idea is to estimate 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, $\hat{B}^{gf}$ is defined as the solution to the optimization,

$$\underset{B \in \mathbb{R}^{p_2 \times p_1}}{\text{minimize}} \frac{1}{2}\|Y - XB\|_F^2 + \lambda\|B\|_1 + \gamma \sum_{e \in E} \sum_{j=1}^{p_2} |r_e| \left| \beta_j^{(e^+)} - \text{sign}(r_e)\, \beta_j^{(e^-)} \right|, \tag{16}$$

where $\|B\|_1$ is the sum of the absolute values of all entries of B, β_j is the j^{th} column 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.

To write the graph fused-lasso penalty in matrix form, define $H \in \mathbb{R}^{p_1 \times |E|}$ by, for each edge (column of H), placing $|r_e|$ at the row corresponding to one endpoint of that edge and $-\operatorname{sign}(r_e)|r_e|$ at the other. This yields,

$$\|BH\|_1 = \sum_{e \in E} \sum_{j=1}^{p_2} |r_e| \left| \beta_j^{(e^+)} - \beta_j^{(e^-)} \right|,$$

and in particular we can write the objective in equation 16 as

$$\frac{1}{2}\|Y - XB\|_F^2 + \|BC\|_1, \tag{17}$$

where $C = (\lambda I_{p_1}, \gamma H)$.

We next describe estimation of $\hat{B}^{\text{gf}}$. There are many ways to solve ℓ^1-regularization problems – for example, quadratic programming, iterative soft-thresholding, coordinate descent, proximal smoothing, and ADMM – but Chen et al. [2010] advocates a proximal smoothing approach. That is, they apply gradient descent on a smooth surrogate loss function – this is similar in spirit to optimizing a Huber loss instead of an ℓ^1-penalty. Rather than directly smoothing the loss in equation 17, they first reformulate it as

$$\frac{1}{2}\|Y - XB\|_F^2 + \max_{\|A\|_\infty \leq 1} \langle A, BC \rangle,$$

using the duality between the ℓ^1 and ℓ^∞ norms. It is for this objective that a family of smooth surrogates is introduced,

$$f_\mu(B) := \max_{\|A\|_\infty \leq 1} \left[\langle A, BC \rangle - \frac{\mu}{2}\|A\|_F^2 \right],$$

and the new objective is to minimize

$$\frac{1}{2}\|Y - XB\|_F^2 + f_\mu(B).$$

When μ is 0, we recover the objective of equation 17. When $\mu > 0$, the problem is smooth, and its gradient can be found in closed form. Towards this, let $g(X) = \frac{1}{2}\|X\|_F^2$ with domain restricted to $\|X\|_\infty \leq 1$. Then g has Fenchel conjugate $g^*(Y) = \max_{\|X\|_\infty \leq 1} \langle Y, X \rangle - \frac{1}{2}\|X\|_F^2$. In particular,

$$\begin{aligned} f_\mu(B) &= \mu \max_{\|A\|_\infty \leq 1} \left[\left\langle A, \frac{1}{\mu} BC \right\rangle - \frac{1}{2}\|A\|_F^2 \right] \\ &= \mu g^*(BC) \end{aligned}$$

Using the fact that the derivative of a Fenchel conjugate function is given by the argmax of the optimization that defines it, we find $\nabla g^*(Y) = P_{\ell^\infty}(Y)$ the projection of Y onto the ℓ^∞ ball[9]. Together with the chain rule, this gives

$$\nabla f_\mu(B) = P_{\ell^\infty}\left(\frac{1}{\mu} BC\right) C^T,$$

[9]This is achieved by taking all entries larger than one and setting them to 1, and setting all entries smaller than -1 to -1).

and so the gradient of the objective 17 has the form

$$X^T (Y - XB) + P_{\ell\infty}\left(\frac{1}{\mu}BC\right)C^T,$$

which can be input to any number of gradient-based routines. Chen et al. [2010] uses Nesterov's accelerated method, a version of gradient descent with a momentum term, and this is what we use in our examples below.

3.8.1 Example

We apply the graph-fused lasso to the body composition problem and compare it to a naive version of lasso that doesn't 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 resulting graph has two major components, corresponding to the lean and fat mass variables, respectively, though arm and leg mass do deviate somewhat from trunk and total mass. The full correlation matrix is displayed in Figure 24.

The fitted coefficients from the graph-fused lasso are given in Figure 25. For reference, the analogous display when the problem is decoupled into parallel lasso regressions, is given in Figure 26. A version of this parallel-lasso figure across a range of λ regularization values is given in Supplementary Figure 36.

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 Ruminoccocus 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 Ruminoccocaceae. 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.

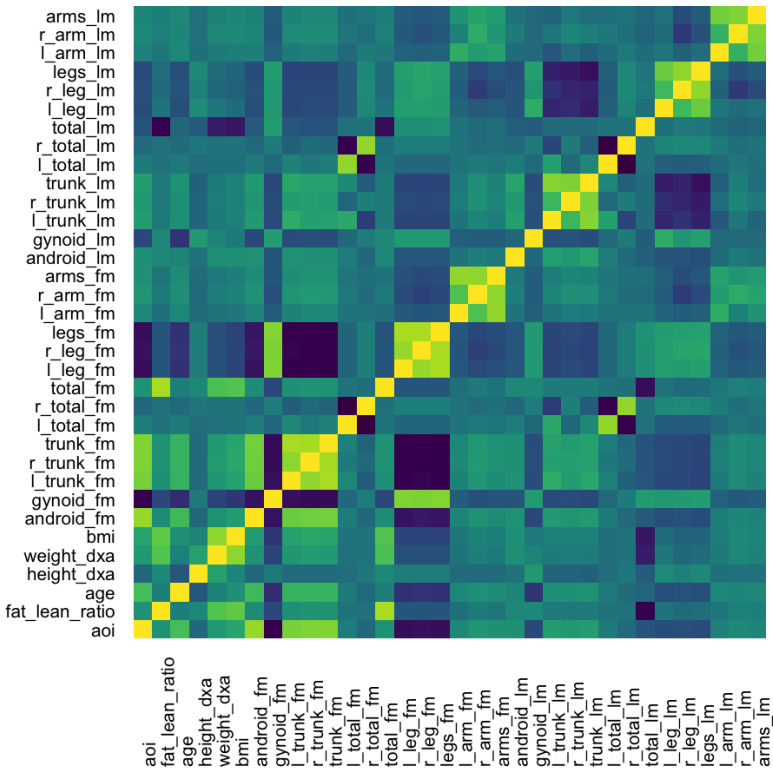

Figure 24: Correlation matrix used as the input graph R for the graph-fused lasso, estimated itself according to the graphical lasso. The two main blocks correspond to lean and fat mass variables, respectively.

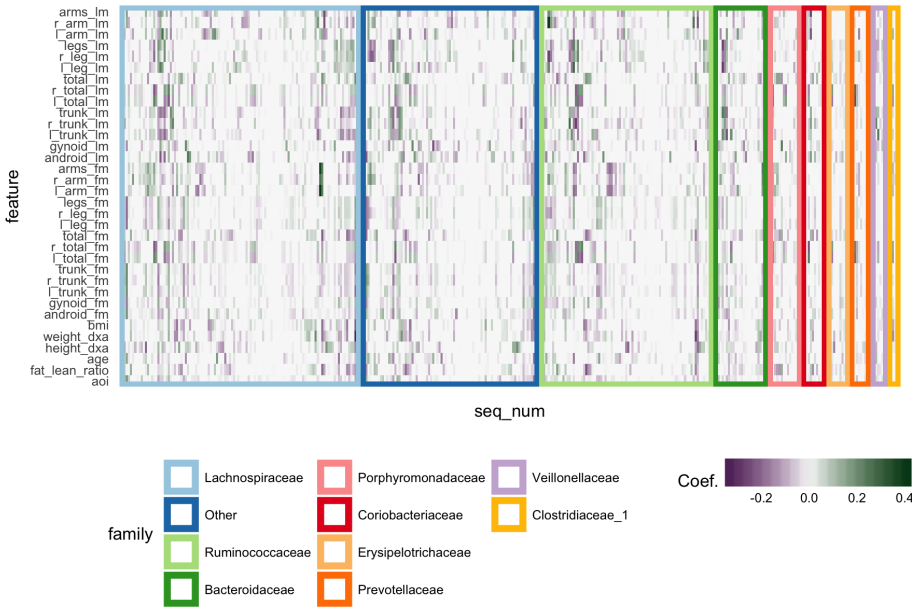

Figure 25: Coefficients for the graph-fused lasso highlight groups of species with similar profiles across response variables. Colored rectangles demarcate taxonomic families. Individual cells give the coefficient for a particular species (column) for a given response variable (row). Purple and green denote negative and positive coefficients, respectively. Note that coefficients have been smoothed according to correlation network between variables, as given in Figure 24. Further, species with similar coefficients are placed near one another.

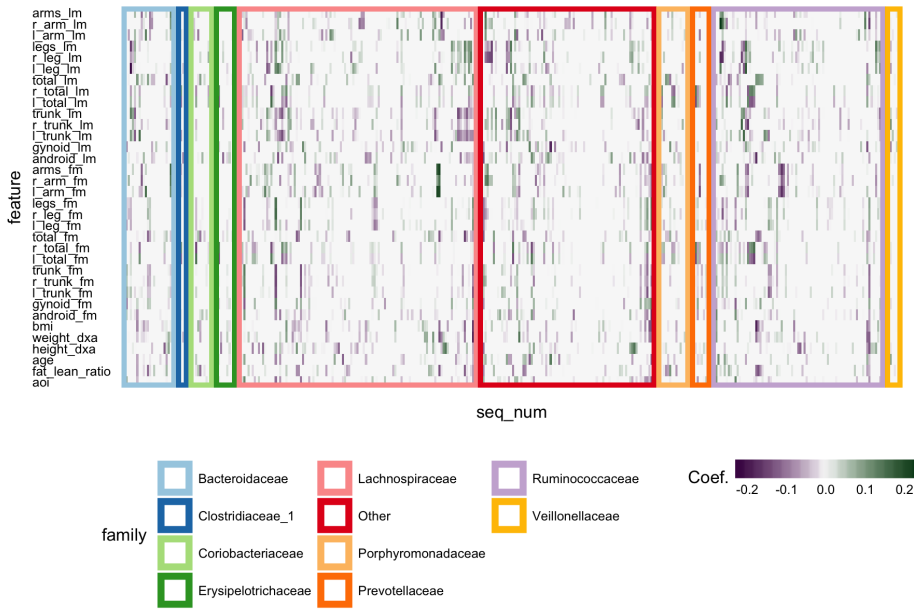

Figure 26: The analog of Figure 25 when there is no sharing across response problems. Note that coefficients nonetheless seem to be similar within lean and fat mass response groups, respectively, but are not as smooth as in the graph-fused lasso. As there is some consistency within these groups of variables, the form of structured regularization imposed by the graph seems appropriate.

3.9 Bayesian multitask learning

Zhang et al. [2005] formulates a general probabilistic approach to multitask learning. Their framework encapsulates both regression and classification, though for simplicity we specialize to regression only. The main idea is to pool the coefficients across otherwise separate regression problems via careful prior specification.

The data come from p_1 regression problems, which we call tasks. Within the r^{th} task, we have features $x_i^{(r)} \in \mathbb{R}^{p_2}$ and responses $y_i^{(r)} \in \mathbb{R}$ for the i^{th} sample. Note that the features $x_i^{(r)}$ are allowed to vary across the regression problems. We use the shorthand $D = \left\{ \left(x_i^{(r)}, y_i^{(r)} \right)_{i=1}^{n_r} \right\}_{r=1}^{p_1}$.

The responses are assumed drawn i.i.d. from a linear model with problem-specific coefficient $\beta^{(r)}$,

$$y_i^{(r)} \sim \mathcal{N}\left(y_i^{(r)} | x_i^{(r)T} \beta^{(r)}, \sigma^2 \right).$$

To tie problems together, the $\beta^{(r)}$ are modeled using K-dimensional latent factors,

$$\beta^{(r)} \sim \mathcal{N}\left(\beta^{(r)} | S w^{(r)}, \Psi \right), \tag{18}$$

for some latent source matrix $S \in \mathbb{R}^{p_2 \times K}$ and mixing weights $w^{(r)}$. Different models come from choosing different priors on $w^{(r)}$; for now suppose the $w^{(r)}$ are drawn jointly from

$$w^{(1)}, \ldots, w^{(r)} \sim p_\Phi\left(\left(w^{(r)} \right)_{r=1}^{p_1} \right).$$

The nonrandom parameters in this model are $\theta := \{\sigma^2, S, \Psi, \Phi\}$. The weights and coefficients $w^{(r)}$ and $\beta^{(r)}$ are random parameters, and for inference they will be treated as latent data.

Note that the form of the $\beta^{(r)}$ in equation 18 can be reparameterized as

$$\beta^{(r)} = S w^{(r)} + \eta^{(r)},$$

where $\eta^{(r)} \sim \mathcal{N}(0, \Psi)$. This clarifies what the latent factor model form of $\beta^{(r)}$ does: it decomposes the slope into a part shared across all models, via $S w^{(r)}$, and a part specific to individual regression problems, the $\eta^{(r)}$. Different priors on $\left(w^{(r)} \right)$ allow the modeler to trade-off the degree of sharing across regression problems: setting $w^{(r)} \equiv 0$ decomposes the model into independent regression problems, while enforcing that they have large variance will ensure that they are more important than the problem-specific component $\eta^{(r)}$.

The authors propose special cases corresponding to particular choices for the prior $p_\Phi\left(\left(w^{(r)} \right)_{r=1}^{p_1} \right)$.

- $w^{(r)} \equiv 0$ decomposes the problem into independent regressions.

- $w^{(r)} = 1$ turns the model into a "contaminated-signal" model of the form $\beta^{(r)} \sim \mathcal{N}(\mu, \Psi)$.
- Drawing $w^{(r)} \sim \text{Mult}(1, \Phi)$ clusters the overall regression problem into subproblems, each sharing the same coefficient.
- Drawing $w^{(r)} \sim \text{Lap}(0, \Phi)$ encourages sparse weights. Though not technically a part of the specification above, it is possible to put a similar prior on S to induce sparsity on the shared sources.
- Drawing $w^{(r)} \sim DP(\alpha, G_0)$ has a similar effect as the clustered-regressions model, though now the number of clusters need not be explicitly specified in advance[10].
- Drawing the $w^{(r)}$ jointly according to a dynamical system, for example $w^{(r)} = \Phi w^{(r-1)} + \epsilon_r$, can be used to model the evolution of the regression coefficients. In general, prior information on how the coefficients should be related can be incorporated by correlating the $w^{(r)}$ across regression problems.

To perform inference, Zhang et al. [2005] employ a variational strategy. Inference of the true posterior jointly over $\left(z^{(r)}\right)_{r=1}^{p_1} := \left(\beta^{(r)}, w^{(r)}\right)_{r=1}^{p_1}$ is generally intractable[11]. However, we can find a point q^* in a variational family $(q_\gamma(z))_{\gamma \in \Gamma}$ such that the expected complete data likelihood is large, and this q^* can be serve as a proxy for the true posterior.

The loglikelihood assuming that the latent $z^{(r)}$ are known is called the complete-data loglikelihood. In our multitask setting, this has the form

$$\begin{aligned}\ell_c(\theta) &= \sum_{r=1}^{p_1}\left[\sum_{i=1}^{n_r}\log p\left(y_i^{(r)} \mid \beta^{(r)}; X^{(r)}, \sigma^2\right)\right] + \log p\left(\beta^{(r)} \mid w^{(r)}; \Lambda, \Psi\right) + \log p_\Phi\left(w^{(r)}\right) \\ &= \sum_{r=1}^{p_1}\sum_{i=1}^{n}\log \mathcal{N}\left(y_i^{(r)} \mid x_i^{(r)T}\beta^{(r)}; \sigma^2\right) + \sum_{r=1}^{p_1}\log \mathcal{N}\left(\beta^{(r)} \mid Sw^{(r)}; \Psi\right) + \log p_\Phi\left(w^{(r)}\right),\end{aligned}$$

where we have assumed that, in the prior, the $w^{(r)}$ are independent across problems – this covers most cases above, though variations are possible in case this assumption is not plausible.

Since the $\beta^{(r)}$ and $w^{(r)}$ are unknown, $\ell_c(\theta)$ cannot actually be evaluated. However, for different choices of q_γ, it may be possible to evaluate $\mathbb{E}_{q_\gamma}[\ell_c(\theta)]$. Further, it can be shown that this expected complete data loglikelihood is maximized when the $z^{(r)}$ are drawn from the posterior $p_\theta((z^r) \mid (x_i), y_i)$. Since the posterior is intractable, we instead search for the element q_{γ^*} that is closest to

[10] The concentration parameter α does modulate the general number of clusters identified, however.

[11] An exception is when $w^{(r)}$ are given standard multivariate Gaussian priors, in which case Gaussian-Gaussian conjugacy can be used.

the posterior in KL-divergence sense,

$$q_{\gamma^*} := \arg\min_{\gamma \in \Gamma} KL\left(q_\gamma\left[\left(z^{(r)}\right)\right] || p_\theta\left[\left(z^{(r)}\right) | (x_i), (y_i)\right]\right).$$

This is a convenient measure to use, because it can be rewritten as

$$\begin{aligned}
& KL\left(q_\gamma\left[\left(z^{(r)}\right)\right] || p_\theta\left[(z^r) | (x_i), (y_i)\right]\right) \\
=& \mathbb{E}_{q_\gamma}\left[\log \frac{q_\gamma\left[\left(z^{(r)}\right)\right]}{p_\theta\left[\left(z^{(r)}\right) | (x_i), (y_i)\right]}\right] \\
=& \mathbb{E}_{q_\gamma}\left[\log \frac{q_\gamma\left[\left(z^{(r)}\right]\right)}{p_\theta\left[\left(z^{(r)}\right), (x_i), (y_i)\right]}\right] + \log p_\theta\left[(x_i), (y_i)\right] \\
=& \mathbb{E}_{q_\gamma}\left[\log q_\gamma\left[\left(z^{(r)}\right)\right]\right] - \mathbb{E}_{q_\gamma}\left[\ell_c(\theta)\right] + \log p_\theta\left[(x_i), (y_i)\right],
\end{aligned}$$

so we can equivalently choose the q_{γ^*} to maximize

$$\mathcal{F}(\gamma, \theta) := -\mathbb{E}_{q_\gamma}\left[\log q_\gamma\left(z^{(r)}\right)\right] + \mathbb{E}_{q_\gamma}\left[\ell_c(\theta)\right],$$

which can be interpreted as a regularized complete data loglikelihood.

Note that this expression only involves the expected complete data loglikelihood, which is usually simple to compute. In particular, in many cases it is easy to find a γ to maximize this quantity, when θ is kept fixed.

Further, for a fixed choice of θ, it may be possible to choose γ to maximize this quantity, and vice versa. This is the variational strategy adopted by [Zhang et al., 2005] – initialize θ^0, then choose γ^1 to minimize $\mathcal{F}\left(\gamma, \theta^0\right)$, then find a θ^1 to maximize $\mathcal{F}\left(\gamma^{(1)}, \theta\right)$, and so forth. The final posterior is approximated by q_{γ^t} for some large t.

4 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 usually confined to particular literature areas and highlighted certain similarities in the process – for example, PCA-IV (Section 2.5) and Bayesian multitask regression (Section 3.9) 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 **??**.

In developing our WELL-China case study, we have both (1) described the types of interpretations facilitated by different approaches and (2) provided

Property	Algorithms
Require covariance estimate	Concat. PCA, CCA, CoIA, MFA, PTA, Statico / Costatis
Analytical Solution	Concat. PCA, CCA, CoIA, MFA, PTA, Statico / Costatis
SPLS, Graph-Fused Lasso, Graph-Fused Lasso	Encouraging sparsity on scores or loadings can result in mo
Tuning Parameters	Sparsity: Graph-Fused Lasso, PMD, SPLS Number of Factors: PCA-IV, Red. Rank Regression, Mixe Prior Parameters: Mixed-Membership CCA, Bayesian Mul Kernel: KCCA
Probabilistic	Mixed-Membership CCA, Bayesian Multitask Regression
Not Normal or Nonlinear	KCCA, CCpNA, Mixed-Membership CCA, Bayesian Multi
>2 Tables	Concat. PCA, CCA, MFA, PMD, KCCA
Cross-Table Symmetry	Concat. PCA, CCA, CoIA, Statico / Costatis, MFA, PMD

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 invariabily 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.

Package	Methods	Documentation	Link
ade4	PCA, CCA, CoIA, Statico / Costatis, PCA-IV	Average	`https://cran.r-project`
FactoMineR	PCA, MFA	High	`https://cran.r-project`
vegan	CCA, CCpnA	High	`https://cran.r-project`
GFLasso	Graph-Fused Lasso	Low	`https://github.com/kri`
bayesMult	Bayesian Multitask Regression	Low	`https://github.com/kri`
spls	SPLS	High	`https://cran.r-project`
kernlab	KCCA	High	`https://cran.r-project`
PMA	PMD	High	`https://cran.r-project`
Base R	PCA, CCA	High	`https://cran.r-project`
pls	PLS	High	`https://cran.r-project`

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/well_microbiome_expers`.

accessible implementations that can be incorporated into practical scientific workflows. Packages providing the implementations of the multitable methods we have reviewed in this work are linked in Table 2. 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 code for all steps are available, so that our work is fully reproducible.

We have found that multitable data analysis problem 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.

References

Shotaro Akaho. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*, 2006.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010a. doi: 10.1186/gb-2010-11-10-r106. URL `https://doi.org/10.1186/gb-2010-11-10-r106`.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010b.

Cédric Archambeau and Francis R Bach. Sparse probabilistic projections. pages 73–80, 2009.

Francis R Bach and Michael I Jordan. Kernel independent component analysis. *The Journal of Machine Learning Research*, 3:1–48, 2003.

Francis R Bach and Michael I Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Leo Breiman and Jerome H Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):3–54, 1997.

Stanford Prevention Research Center. Well-china: New wellness solutions. URL `https://prevention.stanford.edu/content/dam/sm/prevention/documents/about/WELL-CHINA.pdf`.

Prabhakar Chalise and Brooke L. Fridley. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PLOS ONE*, 12(5):e0176278, may 2017. doi: 10.1371/journal.pone.0176278. URL `https://doi.org/10.1371/journal.pone.0176278`.

Kumardeep Chaudhary, Olivier B. Poirion, Liangqun Lu, and Lana X. Garmire. Deep learning based multi-omics integration robustly predicts survival in liver cancer. mar 2017. doi: 10.1101/114892. URL `https://doi.org/10.1101/114892`.

Xi Chen, Seyoung Kim, Qihang Lin, Jaime G Carbonell, and Eric P Xing. Graph-structured multi-task regression and an efficient optimization method for general fused lasso. *arXiv preprint arXiv:1005.3579*, 2010.

Hyonho Chun and Sündüz Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.

Dongjun Chung and Sunduz Keles. Sparse partial least squares classification for high dimensional data. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

Dongjun Chung, Hyonho Chun, and Sunduz Keles. Spls: Sparse partial least squares (spls) regression and classification. *R package, version*, 2:1–1, 2012.

Sylvain Dolédec and Daniel Chessel. Co-inertia analysis: an alternative method for studying species–environment relationships. *Freshwater biology*, 31(3): 277–294, 1994.

Ildiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

Eric A. Franzosa, Tiffany Hsu, Alexandra Sirota-Madi, Afrah Shafquat, Galeb Abu-Ali, Xochitl C. Morgan, and Curtis Huttenhower. Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nature Reviews Microbiology*, 13(6):360–372, apr 2015. doi: 10.1038/nrmicro3451. URL `https://doi.org/10.1038/nrmicro3451`.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Julia Fukuyama, Laurie Rumker, Kris Sankaran, Pratheepa Jeganathan, Les Dethlefsen, David A. Relman, and Susan P. Holmes. Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. *PLOS Computational Biology*, 13(8):e1005706, aug 2017. doi: 10.1371/journal.pcbi.1005706. URL `https://doi.org/10.1371/journal.pcbi.1005706`.

David Gomez-Cabrero, Imad Abugessaisa, Dieter Maier, Andrew Teschendorff, Matthias Merkenschlager, Andreas Gisel, Esteban Ballestar, Erik Bongcam-Rudloff, Ana Conesa, and Jesper Tegnér. Data integration in the era of omics: current and future challenges. *BMC systems biology*, 8(2):I1, 2014.

Michael Greenacre and Trevor Hastie. The geometric interpretation of correspondence analysis. *Journal of the American statistical association*, 82(398): 437–447, 1987.

Michael J Greenacre. *Theory and applications of correspondence analysis*. 1984.

Mats G Gustafsson. A probabilistic derivation of the partial least-squares algorithm. *Journal of chemical information and computer sciences*, 41(2):288–294, 2001.

EJ Hannan. Canonical correlation and multiple equation systems in economics. *Econometrica: Journal of the Econometric Society*, pages 123–138, 1967.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4): 321–377, 1936.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3):531–547, 2003.

Malte Kuss and Thore Graepel. The geometry of kernel canonical correlation analysis. Technical report, 2003.

Gert RG Lanckriet, Tijl De Bie, Nello Cristianini, Michael I Jordan, and William Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

Kim-Anh Lê Cao, Debra Rossouw, Christele Robert-Granié, and Philippe Besse. A sparse pls for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7(1), 2008.

Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

Ruth E Ley. Obesity and the human microbiome. *Current opinion in gastroenterology*, 26(1):5–11, 2010.

Ruth E Ley, Fredrik Bäckhed, Peter Turnbaugh, Catherine A Lozupone, Robin D Knight, and Jeffrey I Gordon. Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31):11070–11075, 2005.

Ruth E Ley, Peter J Turnbaugh, Samuel Klein, and Jeffrey I Gordon. Microbial ecology: human gut microbes associated with obesity. *Nature*, 444(7122): 1022–1023, 2006.

Kantilal Varichand Mardia, John T Kent, and John M Bibby. Multivariate analysis. 1980.

Y Matsuzawa. The role of fat topology in the risk of disease. *International journal of obesity*, 32(S7):S83, 2008.

Ian H McHardy, Maryam Goudarzi, Maomeng Tong, Paul M Ruegger, Emma Schwager, John R Weger, Thomas G Graeber, Justin L Sonnenburg, Steve Horvath, Curtis Huttenhower, Dermot PB McGovern, Albert J Fornace, James Borneman, and Jonathan Braun. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome*, 1(1):17, 2013. doi: 10.1186/2049-2618-1-17. URL `https://doi.org/10.1186/2049-2618-1-17`.

Shakir Mohamed, Zoubin Ghahramani, and Katherine A Heller. Bayesian exponential family pca. In *Advances in neural information processing systems*, pages 1089–1096, 2009.

Ashin Mukherjee and Ji Zhu. Reduced rank ridge regression and its kernel extensions. *Statistical analysis and data mining*, 4(6):612–622, 2011.

Jérôme Pagès. *Multiple Factor Analysis by example using R*. CRC Press, 2014.

Jérôme Pages et al. Multiple factor analysis: main features and application to sensory data. *Revista Colombiana de Estadística*, 27(1):1–26, 2004.

Gholamali Rahnavard, Eric A. Franzosa, Lauren J. McIver, Emma Schwager, George Weingart, Yo Sup Moon, Xochitl C. Morgan, Levi Waldron, and Curtis Huttenhower. High-sensitivity pattern discovery in large multi'omic datasets. URL `https://huttenhower.sph.harvard.edu/halla`.

C Radhakrishna Rao. The use and interpretation of principal component analysis in applied research. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 329–358, 1964.

Bernhard Schölkopf. The kernel trick for distances. In *Advances in neural information processing systems*, pages 301–307, 2001.

Mervyn Stone and Rodney J Brooks. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 237–269, 1990.

Cajo JF Ter Braak. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67(5):1167–1179, 1986.

Jean Thioulouse. Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods. *The Annals of Applied Statistics*, pages 2300–2325, 2011.

Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. A core gut microbiome in obese and lean twins. *nature*, 457(7228):480, 2009.

Nikos Vlassis, Yoichi Motomura, and Ben Krose. Supervised linear feature extraction for mobile robot localization. In *Robotics and Automation, 2000. Proceedings. ICRA'00. IEEE International Conference on*, volume 3, pages 2979–2984. IEEE, 2000.

Daniela Witten, Rob Tibshirani, Sam Gross, Balasubramanian Narasimhan, and Maintainer Daniela Witten. Package pma. *Genetics and Molecular Biology*, 8(1):28, 2013.

Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008commit, 2009.

Herman Wold. Partial least squares. *Encyclopedia of statistical sciences*, 1985.

Jian Zhang, Zoubin Ghahramani, and Yiming Yang. Learning multiple related tasks using latent independent component analysis. In *Advances in neural information processing systems*, pages 1585–1592, 2005.

Mu Zhu, Trevor J Hastie, and Guenther Walther. Constrained ordination analysis with flexible response functions. *Ecological Modelling*, 187(4):524–536, 2005.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

5 Supplementary Material

5.1 Additional figures

5.2 Derivation details for PCA-IV

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

Statistics	Ecology	Machine Learning
Samples	Sites	Samples
Variables	Species	Features
Scores	Scores	Embeddings
Biplot	Ordination	—
Variance explained	Inertia	Training Error
Entropy	Diversity	Entropy

Table 3: A brief translation of terms used in multitable data analysis, across three of the fields where many methods have been proposed.

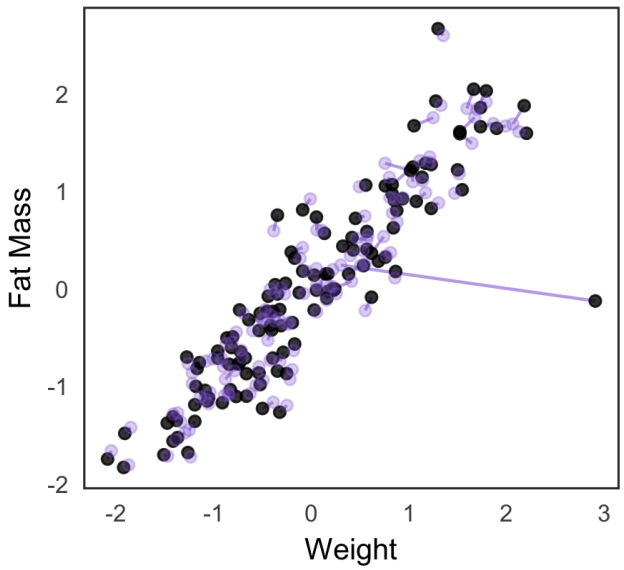

Figure 27: This is the analog of Figure 1 in the case that PCA is run on all body composition variables, rather than just lean and fat mass. That is, we project the original values for these two features onto the top two PCs obtained from a PCA on all body composition variables. Since the PCA is working in a large space, the projected points are generally not too far from their original positions. However, note that one outlier on the far right is projected into the bulk of points in the center – the variation coming from this one point is too specific to be preserved by PCA.

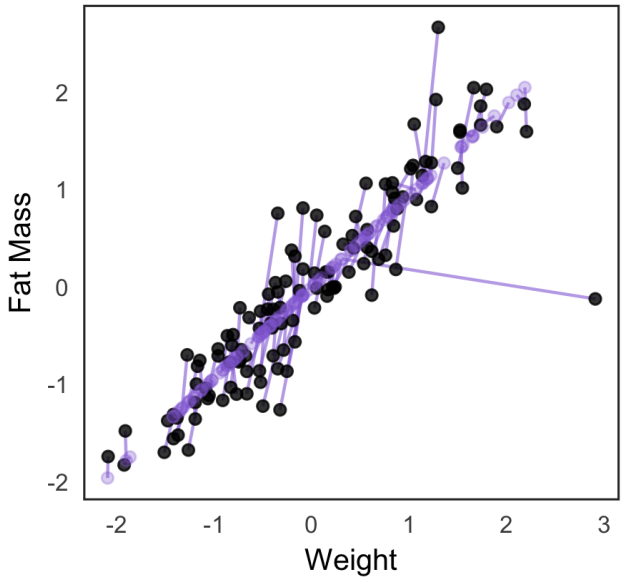

Figure 28: Here we perform the same procedure as in Figure 27, except instead of projecting onto the top two PCs, we project onto only the top PC. The main point is that, while in two dimensions (Figure 1), the behavior of the projection is easy to understand in terms of orthogonal errors, the corresponding orthogonal projection in higher dimensions is more complex.

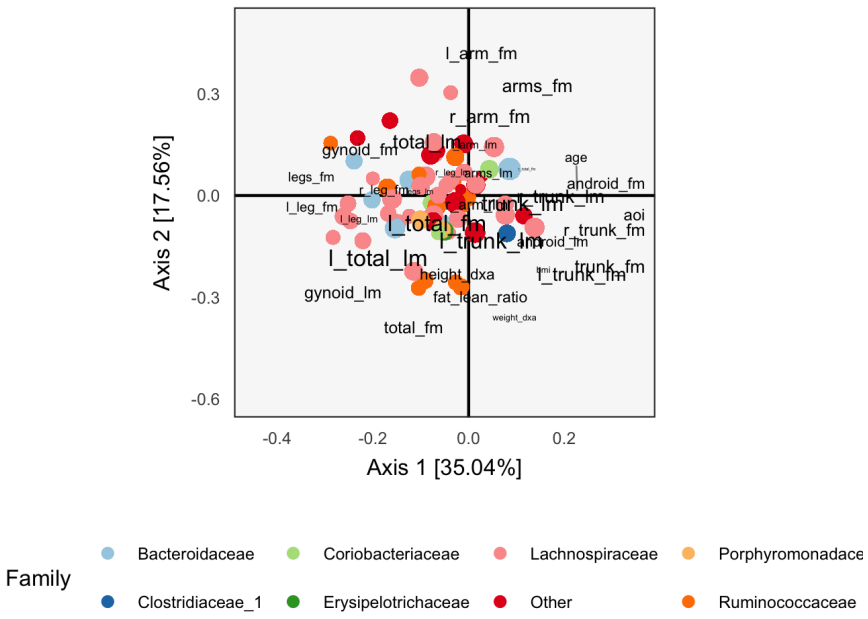

Figure 29: These are the loadings obtained from CoIA, which are analogous to those obtained from concatenated PCA (Figure 3) and CCA (Figure 6).

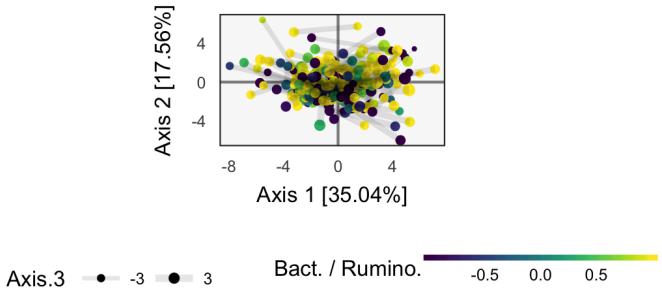

Figure 30: These the same CoIA scores as in Figure 10, but shaded instead by Ruminococcaceae / Lachnospiraceae ratio, as in Figures 5 and 9.

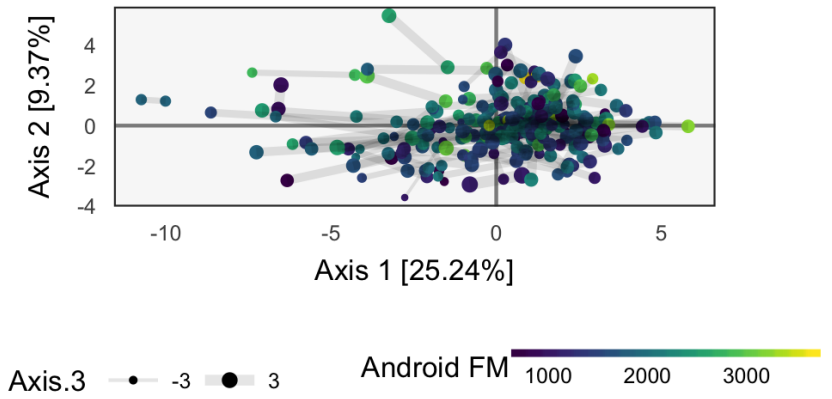

Figure 31: We can display the scores obtained by PCA-IV. The results are similar to those from combined-PCA, which is not surprising, since they are related to the PCA based only on microbial abundances.

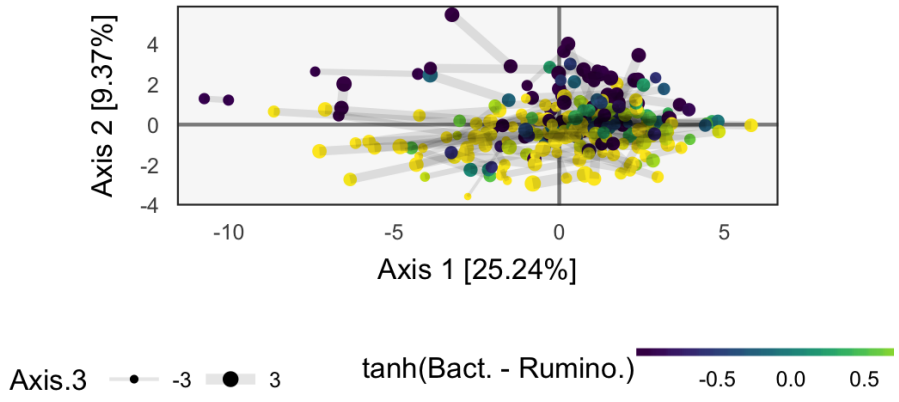

Figure 32: This provides the same scores as Figure 31, but shaded by Ruminococcaceae vs. Lachnospiraceae ratio.

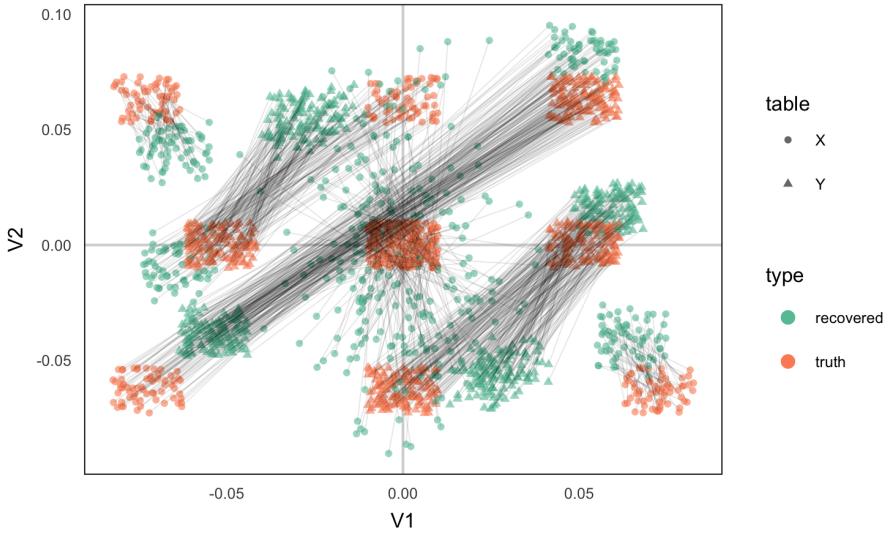

Figure 33: An alternative view of Figure 16, where the two underlying sources are plotted in two dimensions rather than one. Basic unidentifiabilities in the model are visible as long swaps between true and recovered points.

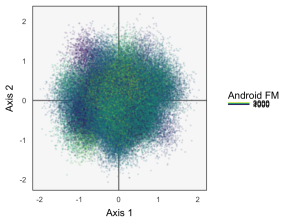

Figure 34: Posterior samples of scores ξ_i^Y associated with the body composition variables. These are much more spread out than either the ξ_i^s or ξ_i^Y, likely because each scores is based on just the 36 body composition variables, rather than counts for all ~ 400 species.

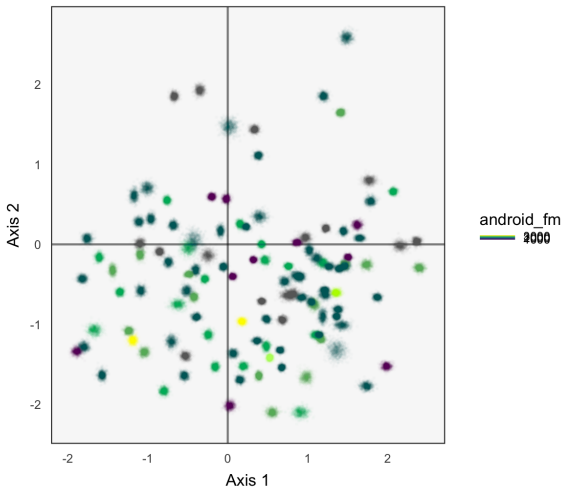

Figure 35: Posterior samples of scores ξ_i^s shared across both bacterial abundance and body composition variables. Points are positioned according to their value in the first two dimensions shared body composition loading dimensions, $B^Y_{\cdot,1:2}$. Clouds of points correspond to participants, and they are shaded by their android fat mass. This method does not appear to pick up on any associations between body composition and microbiome structure.

First consider $k = 1$. For any $\tilde{v}$, the objective in equation 6 has the form

$$\begin{aligned}\operatorname{tr}\left(\hat{\Sigma}_{YX}\tilde{v}\left(\tilde{v}\hat{\Sigma}_{XX}\tilde{v}\right)^{-1}\left(\hat{\Sigma}_{YX}\tilde{v}\right)^{T}\right) &= \frac{\tilde{v}^T\Sigma_{XY}\Sigma_{YX}\tilde{v}}{\tilde{v}^T\Sigma_{XX}\tilde{v}} \\ &= \frac{\tilde{w}^T\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YX}\Sigma_{XX}^{-\frac{1}{2}}\tilde{w}}{\|\tilde{w}\|_2^2}, \end{aligned} \tag{19}$$

where we change variables $\tilde{w} = \Sigma_{XX}^{\frac{1}{2}}\tilde{v}$. But to maximize equation 19, 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,

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

Hence, in the case $K = 1$, the criterion is maximized by the top generalized eigenvector. For larger K, recall that the problem of maximizing $\frac{v^TAv}{\|v\|^2}$ over v

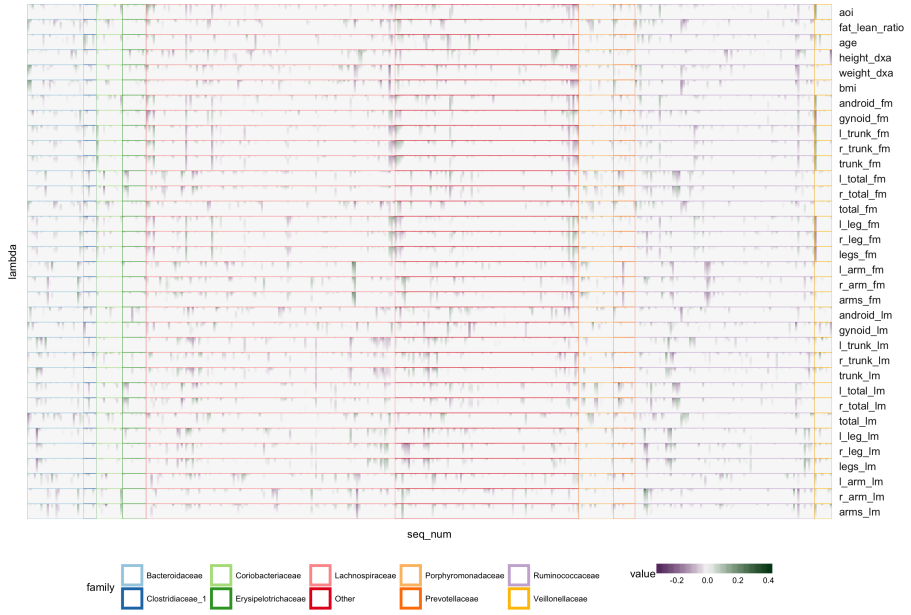

Figure 36: Inspecting coefficients from multitask lasso across λ regularization parameters highlights species with different directions and strengths of association with body composition variables. Different panel columns indicate different taxonomic families, while rows correspond to different response variables. Individual tiles give the coefficients of an individual species (column) against a response at a particular λ (row). Within panels, the λs are sorted from least to most regularization. Species have been clustered according to their lasso coefficients, revealing groups of species with similar coefficient profiles across groups of responses.

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 19 of the argument above gives the result for general K.

5.3 Derivation of PTA α

The Lagrangian of the optimization defined by PTA is

$$\mathcal{L}(\alpha, \lambda) = \sum_{l=1}^{L} \alpha_l \left\langle \overline{X}, X_{\cdot\cdot l} \right\rangle + \lambda \left(\|\alpha\|_2^2 - 1 \right),$$

which when differentiated with respect to α yields $\alpha_l = -\frac{1}{2\lambda} \left\langle \overline{X}, X_{\cdot\cdot l} \right\rangle$ for all l. The constraint that $\|\alpha\|_2^2 = 1$ implies that $\frac{1}{4\lambda^2} \sum_{l'=1}^{L} \left\langle \overline{X}, X_{\cdot\cdot l'} \right\rangle^2 = 1$, which gives $\lambda = \frac{1}{2} \sqrt{\sum_{l'=1}^{L} \left\langle \overline{X}, X_{\cdot\cdot l'} \right\rangle^2}$, so $\alpha_l = \frac{\left\langle \overline{X}, X_{\cdot\cdot l} \right\rangle}{\sqrt{\sum_{l'=1}^{L} \left\langle \overline{X}, X_{\cdot\cdot l'} \right\rangle^2}}$.

5.4 Derivation of Reduced Rank Solution

Consider the data and parameters in the whitened space, $Y^* = Y\hat{\Sigma}_{YY}^{-\frac{1}{2}}$ and $B^* = B\hat{\Sigma}_{YY}^{-\frac{1}{2}}$, and rewrite the objective 7 as

$$\|Y^* - XB^*\|_F^2 = \|Y^* - \hat{Y}^{*\text{ols}}\|_F^2 + \|\hat{Y}^{*\text{ols}} - XB^*\|_F^2,$$

where we used the fact that the residuals are orthogonal to the column space of X to remove the cross term. The first term does not involve B^*, so we can focus on minimizing the second. Consider the SVD $\hat{Y}^{*\text{ols}} = \dot{U}\dot{D}\dot{V}^T$. We know that the matrix A of rank K that minimizes $\|\hat{Y}^{*\text{ols}} - A\|_F^2$ is $A = \dot{U}_K \dot{D}_K \dot{V}_K^T = Y^{*\text{ols}} \dot{V}_k \dot{V}_k^T$, the truncated SVD of $\hat{Y}^{*\text{ols}}$, or alternatively its projection onto the top K right eigenvectors.

In particular, any matrix B that satisfies

$$XB^* = \hat{Y}^{*\text{ols}} \dot{V}_k \dot{V}_k^T = X\hat{B}^{*\text{ols}} \dot{V}_k \dot{V}_k^T$$

solves the reduced rank regression problem, so we can choose

$$\hat{B}^{\text{rr}} = \hat{B}^{*\text{ols}} \dot{V}_k \dot{V}_k^T,$$

which involves $\hat{B}^{*\text{ols}}$, the OLS fit of Y^* on X, and V_k, the top K right eigenvectors of the resulting fitted vector $\hat{Y}^{*\text{ols}}$.

There is a connection between this fit and the response canonical directions of $\hat{Y}$. In particular, consider the eigendecomposition that follows from the earlier SVD,

$$\begin{aligned} \dot{V}\dot{D}\dot{V}^T &= \hat{Y}^{*\text{ols}T} \hat{Y}^{*\text{ols}} \\ &= \left(P_X Y \Sigma_{YY}^{-\frac{1}{2}} \right)^T \left(P_X Y \Sigma_{YY}^{-\frac{1}{2}} \right) \\ &= \Sigma_{YY}^{-\frac{1}{2}} \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-\frac{1}{2}}. \end{aligned} \tag{20}$$

Recall that the response canonical directions V are derived by taking the SVD of $\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YY}^{-\frac{1}{2}} = \tilde{U}\tilde{D}\tilde{V}^T$ and setting $V = \Sigma_{YY}^{-\frac{1}{2}}\tilde{V}$. But comparing this to the form of Equation 20, we find that $\dot{V} = \tilde{V}$, the eigenvectors from which the CCA response directions are derived are equal to the eigenvectors of the cross-products of the OLS fits in the whitened space. Hence,

$$\begin{aligned}
\hat{B}^{\text{rr}} &= \hat{B}^{*\text{ols}}\tilde{V}_K\tilde{V}_K^T \\
&= \left(X^TX\right)^{-1}X^TY^*\tilde{V}_K\tilde{V}_K^T \\
&= \left(X^TX\right)^{-1}X^TY\Sigma_{YY}^{-\frac{1}{2}}\tilde{V}_K\tilde{V}_K^T \\
&= \left(X^TX\right)^{-1}X^TYV_K\Sigma_{YY}^{\frac{1}{2}}V_K^T \\
&= \hat{B}^{\text{ols}}V_KV_K^-,
\end{aligned}$$

Therefore, the reduced-rank coefficients are just the projection of the original OLS coefficients onto the subspace spanned by the top K response canonical directions.

5.5 Derivation of Curds & Whey Shrinkage

Consider prediction across many related response variables. One way to pool information across respones is to define new fitted values from a linear combination of independent OLS fits. That is, to predict a response $y_i \in \mathbb{R}^{p_1}$, we set $\hat{y}_i^{\text{cw}} = B\hat{y}_i^{\text{ols}}$ for some square matrix $B \in \mathbb{R}^{p_1 \times p_1}$. 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: $h_{ii} \approx h := \frac{1}{n}\operatorname{tr}(H)$ 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

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

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

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

or equivalently

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

which in matrix form is

$$(1-g)\,Y^TY + g\hat{Y}^TY = B\left((1-g)\,Y + g\hat{Y}\right)^T\left((1-g)\,Y + g\hat{Y}\right), \qquad (21)$$

where $\hat{Y} \in \mathbb{R}^{n\times p_1}$ has i^{th} row $\hat{y}_{-i}$.

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

$$\begin{aligned}
Y^TY &= n\hat{\Sigma}_{YY} \\
\hat{Y}^TY &= Y^THY = Y^TX\left(X^TX\right)^{-1}X^TY = n\hat{\Sigma}_{YX}\hat{\Sigma}_{XX}^{-1}\hat{\Sigma}_{XY} \\
\hat{Y}^T\hat{Y} &= Y^TP_X^2Y = Y^TP_XY = n\hat{\Sigma}_{YX}\hat{\Sigma}_{XX}^{-1}\hat{\Sigma}_{XY},
\end{aligned}$$

Substituting this into equation 21 and ignoring the scaling n yields

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

Postmultiplying by $\hat{\Sigma}_{YY}$ gives

$$(1-g)\,I_{p_1} + g\hat{Q}^T = B\left[(1-g)\,I_{p_1} + \left(2g-g^2\right)\hat{Q}^T\right], \qquad (22)$$

where,

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

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 $A := \Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{XX}^{-\frac{1}{2}} = \tilde{U}D\tilde{V}^T$. Hence,

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

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 22 can be expressed as

$$V^{-T}\left[(1-g)\,I_{p_1} + gD^2\right]V^T = BV^{-T}\left[(1-g)\,I_{p_1} + \left(2g-g^2\right)D^2\right]V^T$$

and the B satisfying the normal equations has the form

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

where Λ is a diagonal matrix with entries

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

Notice that when n is large, $\frac{1}{n} \operatorname{tr} P_X$ will be small, leading to a smaller $g \approx 0$ less shrinkage.

Recall that $\hat{B}^{\text{cw}}$ is used to pool across OLS fits, $\hat{y}_i^{\text{cw}} = \hat{B}^{\text{cw}} \hat{y}_i^{\text{ols}}$. That is,

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

which we can also view as $\hat{Y}^{\text{cw}} V = \left(\hat{Y}^{\text{ols}} V \right) \Lambda$. This means that the C&W coordinates along the canonical directions V are set as the OLS fits $\hat{Y}^{\text{ols}}$ along the canonical directions V, with weights defined by Λ. The actual $\hat{Y}^{\text{cw}}$ are recovered by transforming back to the original coordinate system. A similar way to view the C&W fits is to note $\hat{Y}^{\text{cw}} V = P_X (YV) \Lambda$, which 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.