The New England Journal of Statistics in Data Science Volume 0, 1–21 (2023)
DOI: https://doi.org/10.51387/23-NEJSDS31

New Perspectives on Centering

Jack PROTHERO*, Jan HANNIG, and J.S. MARRON

Abstract

Data matrix centering is an ever-present yet under-examined aspect of data analysis. Functional data analysis (FDA) often operates with a default of centering such that the vectors in one dimension have mean zero. We find that centering along the other dimension identifies a novel useful mode of variation beyond those familiar in FDA. We explore ambiguities in both matrix orientation and nomenclature. Differences between centerings and their potential interaction can be easily misunderstood. We propose a unified framework and new terminology for centering operations. We clearly demonstrate the intuition behind and consequences of each centering choice with informative graphics. We also propose a new direction energy hypothesis test as part of a series of diagnostics for determining which choice of centering is best for a data set. We explore the application of these diagnostics in several FDA settings.

keywords and phrases: Data Matrix, Object Centering, Trait Centering, Functional Data Analysis.

1. INTRODUCTION

Many data processing pipelines involve transformations such as *centering*: the subtraction of the mean of a set of values resulting in the transformed data having 0 mean. Despite the pervasiveness of such transformations, there are surprising misunderstandings concerning their meaning and implications. In this paper we present a survey of the effects of different forms of centering on data and the consequences of those effects within widely-used data analysis methods. We first seek to disambiguate the terminology used to discuss centering colloquially by putting forth a carefully-considered nomenclature framework. With a unified lexical understanding we discuss the geometric effects of each centering in all relevant vector spaces. We find overall that new and more complete data insights are available via a new mode of variation derived from non-standard centering. The case studies and hypothesis tests presented in this paper provide a blueprint for how to determine which centering to ultimately opt for in new analyses.

While the issues explored in this paper are relevant for a wide range of settings, they play a central role in Functional Data Analysis (FDA) where each item in a data set is a function observed at finitely many points. The notion of a mean function and the exact spaces in which centering takes places must be carefully considered for these sorts of data. Several excellent foundational references on FDA include [7], [9], and [5]. For a more brief introduction, we refer to the review paper by [13].

We consider the special case where each function in the data set was observed at the exact same points. In this special case, we may organize the data into a $d \times n$ matrix with one of rows or columns considered as curves. In this paper, we follow the convention in [10] of columns representing those curves. We refer to each d-dimensional column vector as a *data object* (i.e. experimental unit, data point, observation, case). This terminology appropriately reflects the full generality of the kinds of data collected and stored in matrix form in modern settings. We refer to each n-dimensional row vector as a *trait* (i.e. feature, variable). While this term is non-standard, it avoids potential ambiguity in using the more popular term "feature." In some areas of data science, "feature vector" refers to what are called data objects here. Because vectors along both dimensions of data matrices are critical to our discussion, we need an appropriately distinct name for each. As one of the main goals of this paper is disambiguating the terminology surrounding centering, we prefer the term "trait vector" to represent the vectors in the dimension opposite the data objects.

Some researchers and software packages opt for the transpose of our convention: using columns as traits and rows as data objects. In fact, the legacy of structural limitations of data analysis software reverberates through our choices of matrix orientation to this day. Many tools placed stricter limits on the number of columns a data table could have, mirroring mathematical preferences for "long and skinny" matrices. Most fields during this time period analyzed data with many more objects than traits, so data was typically entered and stored such that objects were rows and traits were columns. Bioinformatics and related fields were in the opposite position and often collected data on a very large number of traits from a relatively small group of objects. This led to data matrices being stored with the opposite orientation: data objects as columns and traits as rows. We follow the bioinformatics convention here. An agreement

[*] Corresponding author.

as to whether rows or columns are data objects is important to facilitate discussion of data analysis between fields. As we'll see shortly, ambiguity in matrix orientation choice has an acutely confounding effect when discussing centering choices.

A time-honored, broadly-used tool in FDA is principal component analysis (PCA). PCA decomposes the data into *modes of variation* about the mean of the data objects. These modes can be calculated from an eigenanalysis of the covariance matrix of the data. To construct the covariance matrix one must first center the data matrix such that the data objects have a mean vector of 0. While this choice of centering is very natural, it is unclear whether it should be called "column centering" or "row centering" regardless of matrix orientation convention. In our convention one might first consider this a vector operation and call it "column centering". However, the operation is equivalent to finding the mean value of each trait row vector and subtracting it from each of that trait row vector's entries. From this perspective the operation could be called "row centering" as the entries of each row have mean zero after the operation. In the other matrix orientation convention, the same could be said of "column centering." We propose new terminology specifically aimed at avoiding this sort of ambiguity. As this translation of the data objects in $\mathbb{R}^d$ (*data object space*) such that they are centered at the origin is an important effect of this centering operation, we will refer to this operation as *object centering* a data matrix. Referring to the intended target of the centering (object vs trait) as opposed to the matrix dimension (column vs row) clarifies the intended meaning while also unifying terminology regardless of choice of matrix orientation.

Including object centering the matrix, there are in fact four total centerings available besides leaving the matrix uncentered.

- *Trait centering* is the dual operation to object centering. From a vector point of view, the trait vectors are translated in $\mathbb{R}^n$ (*trait space*) such that their mean vector is at the origin. As a result, the entries of each individual data object have a mean of 0.
- *Grand mean centering* finds the mean of all entries of the data matrix (the *grand mean*) and subtracts that value from each entry.
- *Double centering* is the result of performing object centering followed by trait centering (or vice versa, the operations commute) on the matrix. The resulting matrix has all the properties of both object-centered matrices and trait-centered matrices.

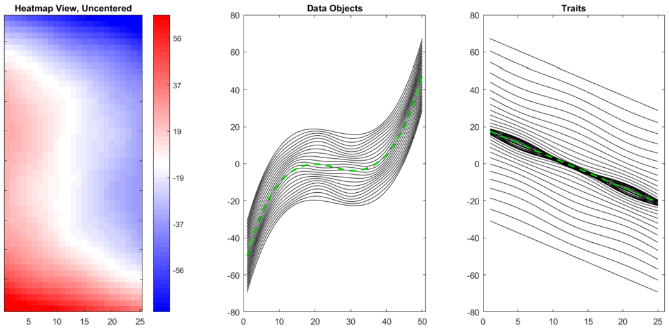


Figure 1: Heatmap (left) and functional data (center, right) views of synthetic data example. Heatmap shows a clear undulating pattern. Data object functions (center panel), corresponding to columns of the heatmap, are distorted cubic functions. Trait functions (right panel), corresponding to rows of the heatmap, are distorted linear functions. Green dashed lines are mean curves.

[15] examine the effects of each centering on the quality of low-rank matrix approximations. Here our purpose is more focused on the interpretability and insights from the data gained or lost by using these different forms of centering in statistical analyses. Notably, that manuscript opts for the ambiguous convention of referring to different centerings according to matrix dimension ("row" and "column") rather than according to the goal of the centering operation. We submit that our nomenclature allows for clearer explorations of these kinds of topics.

Figures 1–4 visually explore the centerings listed above through different points of view of a common synthetic data set. The synthetic data matrix is 50×25; we display its contents in Figure 1. The left panel shows a heatmap view of the data matrix. In a heatmap view, the numerical value of each entry is encoded as a color, with hue indicating a positive (red) or negative (blue) entry and saturation indicating magnitude. The heatmap reveals strong patterns across both the traits and the data objects. We alternatively display these patterns with functional data views of both the data objects and the traits in the center and right panels respectively. The data objects (heatmap columns, center panel curves) are a bundle of distorted and vertically shifted cubic functions with a cubic function mean (center panel green dashed line). The traits (heatmap rows, right panel curves) are a bundle of distorted linear functions with a linear function mean (right panel green dashed line). Curve height in the center & right panels corresponds with pixel color in the heatmap.

In subsequent views of this data we will perform three of the four centerings on the original matrix and examine the changes in the visual patterns of the data matrix. Each view will substantially and uniquely alter which aspects of the data are prominent and which are hidden. We omit grand mean centering as it amounts to a simple modification of the heatmap colors and a vertical shift of the curves in the functional data plots.

We first perform object centering, with results shown in Figure 2. The center panel now shows a series of vertical shifts and amplitude scalings of a sine wave as the cubic structure was removed with the object mean (green dashed

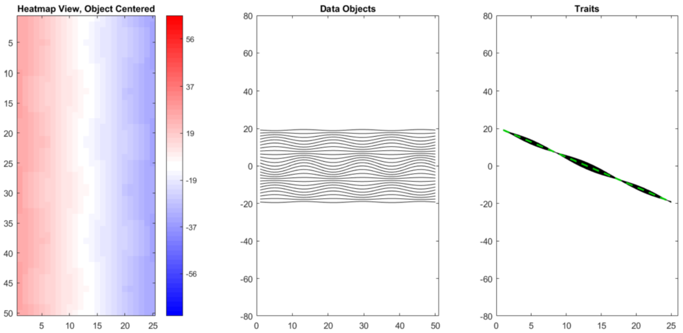


Figure 2: Heatmap (left) and functional data (center, right) view of object-centered synthetic data example. Cubic effect in data objects (center), corresponding to columns of the heatmap, is removed and linear effect now dominates. Data objects are now shifted and scaled low-amplitude sine waves. Traits (right), corresponding to rows of the heatmap, all coalesce around the linear mean curve.

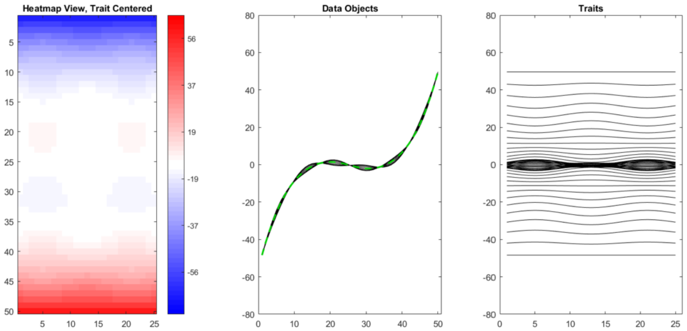


Figure 3: Heatmap (left) and functional data view (center, right) of trait-centered synthetic data example. Linear effect in traits (right), corresponding to rows of the heatmap, is removed to reveal another smaller-scale wave effect. Data objects (center), corresponding to columns of the heatmap, all coalesce around the cubic mean curve.

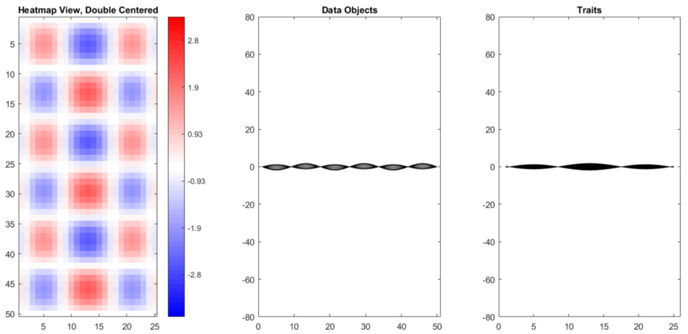


Figure 4: Heatmap (right) and functional data (center, left) view of double-centered synthetic data example. Both the data object (center) and trait (right) curve bundles are clear sine waves. Heatmap shows a clear planar wave pattern due to the outer product of the sine waves along both dimensions (data objects, traits). Scale of the heatmap color bar is changed to emphasize this subtle but meaningful effect.

line in center panel of Figure 1). The right panel shows a strong vertical shift of each of the trait curves, bringing them together around their linear function mean. The heatmap in the left panel is now dominated by the linear effect in each row. We have kept the heatmap color saturation scale and vertical axes in the center & right panel identical to those in Figure 1 for effective comparison.

Next we examine the effects of trait centering. As this is the dual operation to object centering, Figure 3 displays effects dual to those from Figure 2. The right panel now shows a series of vertically shifted waves as the sloped linear structure was removed with the trait mean. The center panel shows a strong vertical shift of each of the data object curves, bringing them together around their cubic function mean. The heatmap in the left panel is now dominated by the cubic effect in each column, with some columns containing small, higher-frequency oscillations reflecting the sine wave distortion that was hard to see in Figure 1. Once again, this figure uses the same scalings as those in Figure 1.

Finally, we perform double centering, which will remove both the cubic function mean among the data objects as well as the linear function mean among the traits. In Figure 4, the residual curves along both dimensions are pure sine waves, and the resulting heatmap in the left panel shows a very clear planar wave pattern. This underlying mode of variation was obscured in Figure 1 by the mean effects along either dimension. In this figure the curve plots use the same vertical axes as Figures 1–3 but the color saturation scale of the heatmap is adjusted because the pattern would appear too faint otherwise.

The fundamental point is that each form of centering leads to a substantially different interpretation of the prominent traits of the data matrix. The visual impression of the raw data differs greatly from each of the centered versions. The distinct and interesting pattern that remains in the double-centered data is largely hidden in views containing either the object mean or the trait mean. An important premise of this paper is that paying more attention to this phenomenon can lead to improved insights from exploratory analysis. In particular, we propose a new, insightful mode of variation based on the trait mean for FDA decompositions.

Notably, while our analyses throughout this paper employ FDA perspectives for visualization and interpretation, they all ultimately involve data that can also be considered in a matrix without missing values. This framing provides $\mathbb{R}^d$ and $\mathbb{R}^n$ as natural choices for data object space and trait space respectively. Potential extensions of the ideas presented in this paper to more general types of data such as irregularly sampled functional data are discussed in Section 5.

The rest of the paper proceeds as follows. In Section 2, we will analyze mortality and genomic data sets under multiple centering regimes to demonstrate the value of exploring non-standard centerings. In both cases we find enhanced visual interpretability after additional centering operations. In Section 3 we mathematically investigate the geometry of different forms of centering in the dual data object and trait spaces. Combining insights from both of these analyses, in Section 4 we develop a novel statistical test which determines whether a significant mean effect is lurking as a substantial portion of a mode of variation. Section 5 contains some brief discussion of results. Finally, in Appendix A, we examine how these lessons on centering can be applied in a multi-block data integration context.

2. FDA CASE STUDIES

In FDA, the data objects are typically vectors representing digitized curves. As with other kinds of data objects we're interested in how the objects vary in the space they occupy. We can use traditional tools like PCA and singular value decomposition (SVD) to discover informative modes of variation in the data, and then use the functional interpretation of the data in question to produce insightful visualization of those modes of variation. In the following subsections we present functional data analyses of two data sets: a collection of mortality rates in Spain during the 20th century and a cohort of base-pair level RNAseq observations. In both cases we'll examine the effects that different centering choices have on the visual interpretation of the analysis.

2.1 Spanish Mortality

[10] consider a data matrix containing mortality rates (proportion of the population of a given age that died in a year) of Spanish males from 1908 to 2002 using information downloaded from the Human Mortality Database [14]. We are interested in how mortality rates, as a function of age from 0 (birth) to 98, changed over this time span. Hence, we will treat each year as a data object (column) and the mortality rates of each age as a trait (row). We first conduct a classical FDA based on object centering. We then compare those results to a naive uncentered SVD and a double centered FDA.

Figure 5 displays the data curves for the Spanish mortality data. Each curve represents a year of data, and the points along each curve encode the mortality rates for each age in that particular year. The curve colors represent chronology, with earlier years displayed in cooler colors and later years displayed in warmer colors. Each entry was adjusted by a $\log_{10}$ transformation because mortality rates tend to vary across several orders of magnitude.

Prominent details include higher mortality rates for newborns and the elderly as well as overall improvement in mortality rate over the course of the 20th century. Both the varied contributions to mortality by age group and the broad decline in mortality across Europe during the 20th century are well-documented in [12], [1], and elsewhere. We observe systematic spikes every 10 years, reflecting strong decadal rounding in death records for older men in the earlier half of the century.

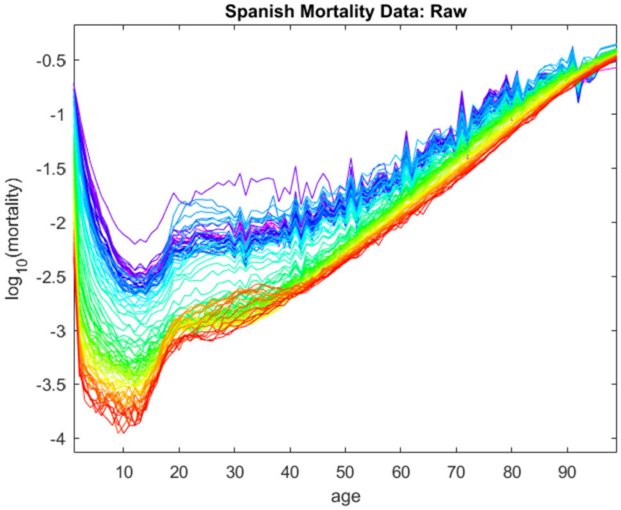


Figure 5: Data curve view of the $\log_{10}$ Spanish mortality data.

We conduct a conventional FDA to find interesting modes of variation in mortality rates over the course of the 20th century. Insights into these modes of variation come from considering both loadings and scores of a PCA. Figure 6 shows the loadings vectors as curves scaled by the scores. The top panel shows the object mean curve as a function of age, and subsequent panels show additional modes of variation about that mean. The second panel (first mode of variation) shows an overall decrease in mortality rate over time which benefitted younger individuals more strongly. The year 1918 is visually distinct at the top of the plot due to the global flu pandemic that year. The third panel (second mode of variation) shows a contrast in mortality rate trends between 18–49 year olds and the rest of the population. This reflects three bursts in mortality for this age group, including the flu pandemic, the Spanish Civil War, and automobile fatalities. The causes of the rise in Spanish male mortality in the late 20th century are discussed further in [4]. Throughout the first and third modes of variation there are remnants of “age-rounding” due to imprecise records. This manifests visually as a repeating pattern over time of length 10.

While Figure 6 explores the modes of variation of the data via loadings, Figure 7 explores relationships between the data objects by looking at scatter plots of projections of the data onto score vectors. We generate one-dimensional views of each score vector with score value on the horizontal axis and chronology on the vertical axis overlaid with a smooth histogram. We also plot two-dimensional views showing projections onto the two-dimensional planes generated by each

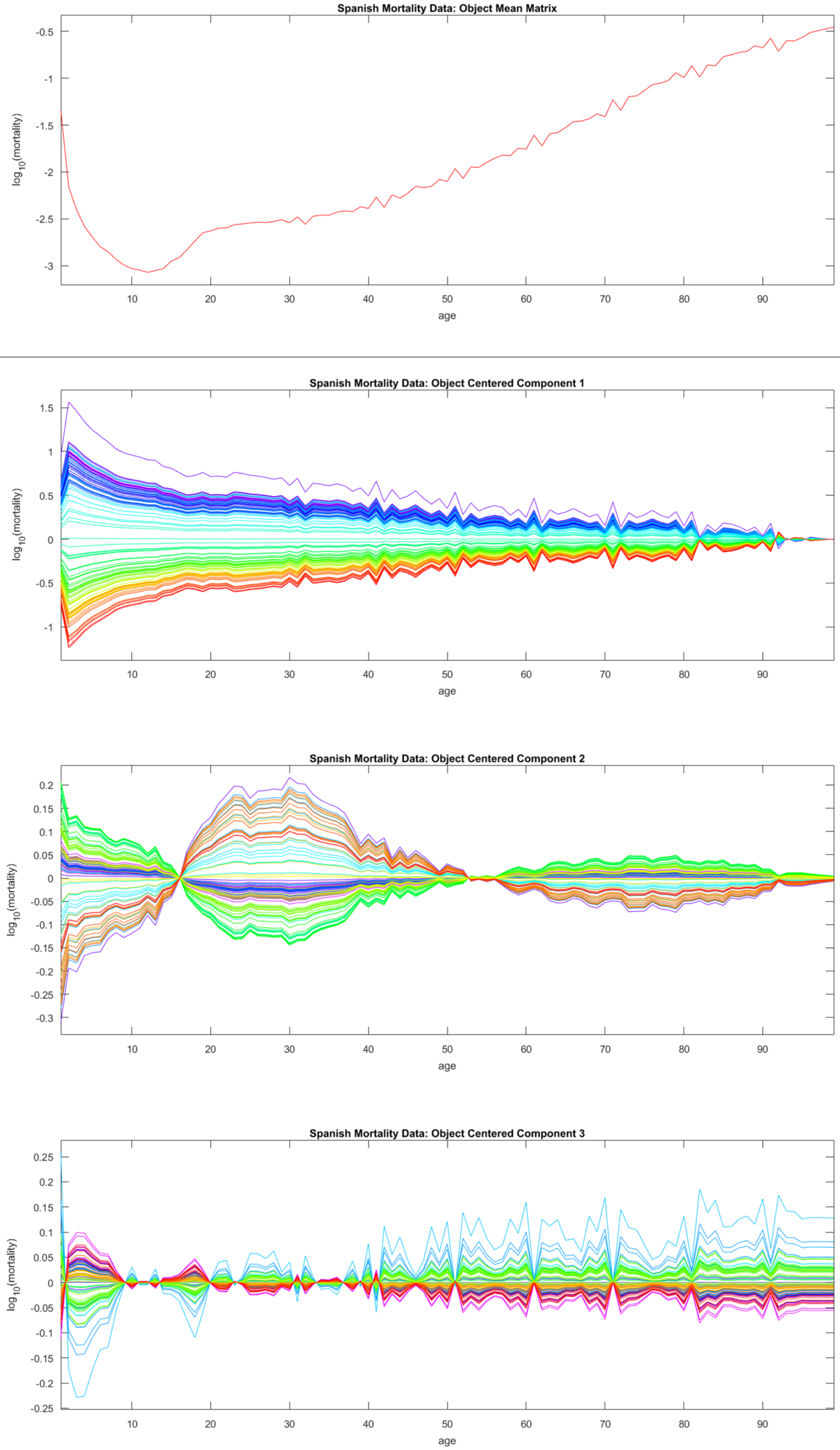


Figure 6: Data curve view of the mean and first three principal modes of variation for the Spanish mortality data. Component 1 (second panel) shows overall improvement over time and component 2 (third panel) shows differences between young/middle-age adults and children/the elderly. Note that for each year adding the corresponding curves from each plot together results in an approximation of the original data curves in Figure 5.

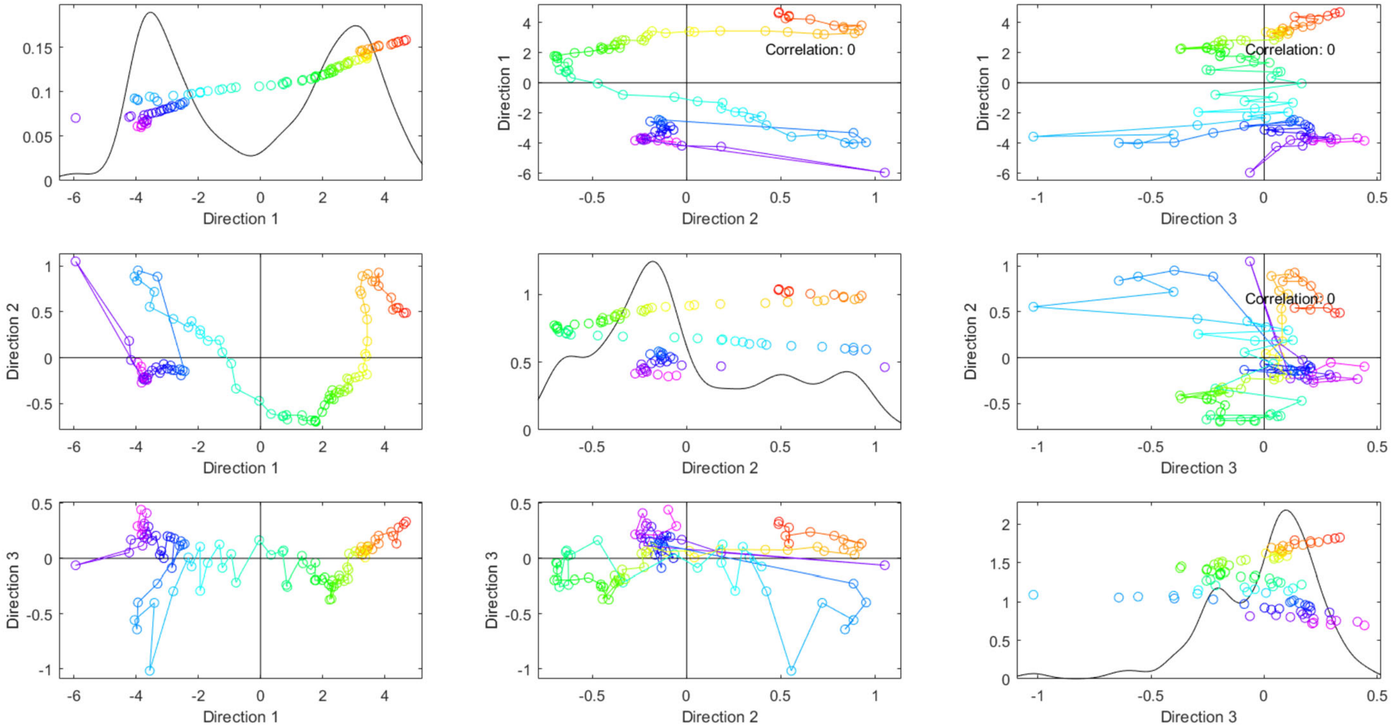


Figure 7: Scatter plot view of Spanish mortality data. Most explainable and interpretable trends appear in the two-way plot of components 1 and 2. Due to the object centering performed for conventional FDA, all 2D plots display 0 correlation.

pair of score vectors. These are all organized into a matrix of plots with 1D views on the diagonal and corresponding 2D views in respective off-diagonal slots. The year-based coloring in each plot is consistent with other views. In the 2D scatter plots, we connect the dots in chronological order. The 2D plot between components 1 and 2 shows many of the trends discussed previously. We can track overall improvement over time with obstacles to that improvement arising in the early 20th century (small cluster of blue points in the bottom right) and late 20th century (cluster of orange points in the top right). Notably the correlation in each 2D scatterplot is zero. As we will show in greater detail in Section 3, this is a consequence of the object centering operation that takes place as the first step of a conventional FDA.

The above PCA can be viewed either as an eigenanalysis of a covariance matrix or as an SVD of the data matrix after it has been object-centered. The right-singular vectors from the SVD are the score vectors and the left-singular vectors are the loadings vectors in our matrix orientation convention. The SVD formulation suggests potential use of other centerings. We could examine left and right singular vectors for an uncentered version of the matrix or a differently-centered version of the matrix. In such cases, the interpretation of the decomposition into modes of variation will typically change substantially.

For instance, Figure 8 shows the modes of variation for the uncentered version of the data matrix. The first component contains information about both the general mortality pattern across ages and the overall improvement over time. The second component has a new contrast between young children and everyone else, and the third component combines many of the patterns separating young adults from older adults with an additional infant effect. Finally, the fourth component reveals a new contrast between younger middle-aged men (ages 25–40) and the rest of the population. The first component contains much of the information taken out by the object mean in the conventional FDA in Figure 6, but it also contains much of what is found in that analysis's first mode of variation.

Next, the double-centered FDA is studied in Figure 9. The rank 2 double mean (first panel) contains both the differences across ages found in the object mean and the constant component of the overall improvement over time found in the trait mean. This visualization of the double mean matrix provides further meaning and context to the first component of the uncentered FDA in Figure 8. We can now see that component is a slightly perturbed and lower-rank version of the double mean matrix. Subsequent panels of Figure 9 then each show one additional effect, and each panel's effect roughly corresponds to the respective panel from Figure 8. The second panel shows stronger improvement over time for younger people, which was also shown in the second panel of the previous figure, but for a more lopsided age group. The third shows differences between the

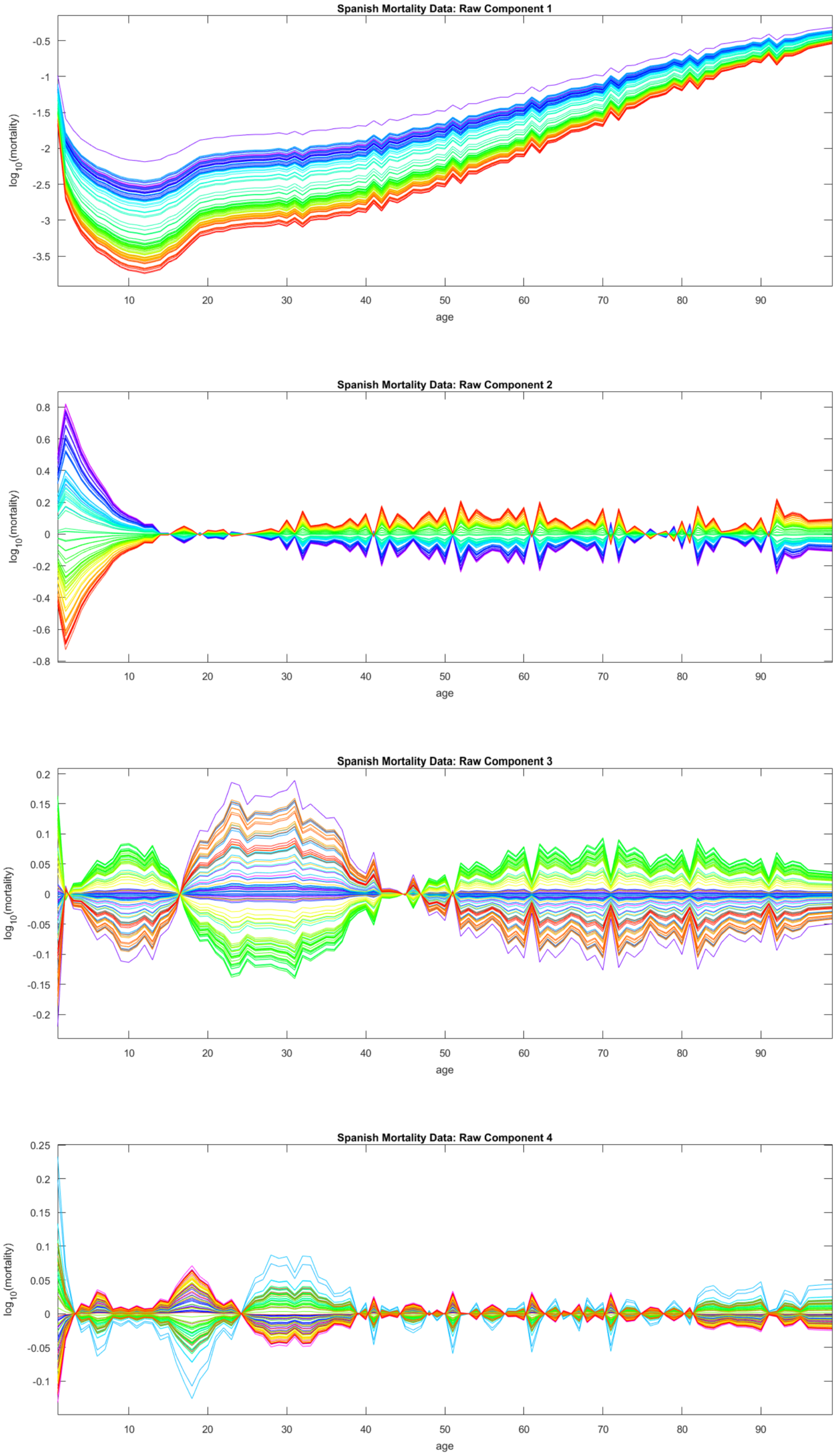


Figure 8: Data curve view of uncentered Spanish mortality data. Different centering dramatically changes the visual analytic impression.

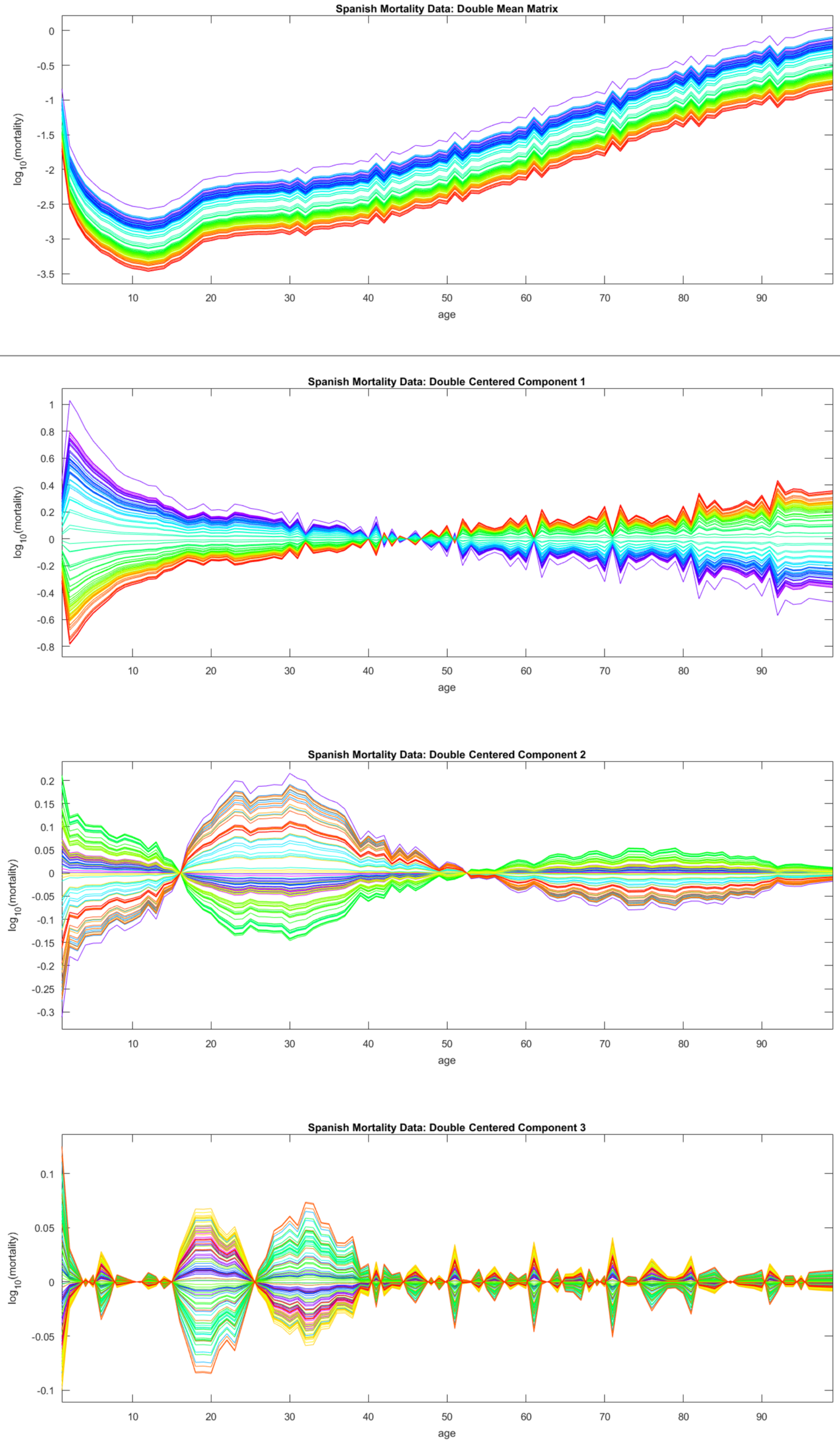


Figure 9: Data curve view of double-centered Spanish mortality data. Both the overall improvement over time and differences across ages are contained within the mean, leaving more specific effects for each subsequent mode of variation. The double mean matrix is the sum of the object and trait mean matrices, and is typically rank 2.

Table 1. Phenomena in FDA components after different forms of centering. Missing phenomena are indicated by empty cells.

↓ Phenomenon, Centering →	None	Object	Double
Difference across ages	1	Mean	Object Mean
Overall mortality reduction	1	1	Trait Mean
Stronger reduction for young	2	1	1
Contrast: 18–49 and others	3	2	2
Contrast: 18–25 and 25–40	4		3
Infant Effects	3,4	3	3
Age Rounding	1,2,3	Mean,1,3	Object Mean,1,3

18–49 year-olds and the rest of the population, which again lines up well with the effect shown in the third panel of the previous figure. The fourth panel shows a difference between older and younger individuals within the 18–49 age range, representing a clearer picture of the contrast hinted at in the fourth panel of Figure 8. Each component is cleanly interpretable and untethered from interference due to mean effects. The one aspect of the data still spread throughout components is the age-rounding effect for older individuals, though this happens regardless of the centering chosen.

The distribution of interpretable effects varies uniquely with each form of centering. To summarize the differences, Table 1 displays for each centering (columns), which phenomena (rows) are contained in which component (numbers). For instance, the first component of the object-centered analysis contains information about both overall mortality improvement and the stronger improvement in mortality rate for younger people, whereas those two phenomena are split up in the double-centered analysis. The former is contained in the mean and the latter is contained in the first component.

The table shows that choice of centering determines in which component different phenomena appear. Different analysts may well have different preferences. We prefer double centering for this data set because it provides the cleanest separation of phenomena into individual modes of variation. Object centering fails to find the additional contrast among the younger adults found with no centering and double centering. While most of the effects are present in the uncentered FDA, double centering allows for clearer attribution of each phenomenon to a specific effect, centering or otherwise. The contrast among younger men is also more prominent and more straightforward in the third mode of variation of the double-centered FDA as compared to the fourth mode in the uncentered FDA.

Often the two most meaningful decompositions into modes of variation will derive from object-centered and double-centered data. In Section 4, we present a statistical test to help determine whether object centering or double centering may be more appropriate for a given data set. As will be seen in Section 3, these two centerings result in mutually uncorrelated score vectors, and double centering additionally results in mutually uncorrelated loadings vectors.

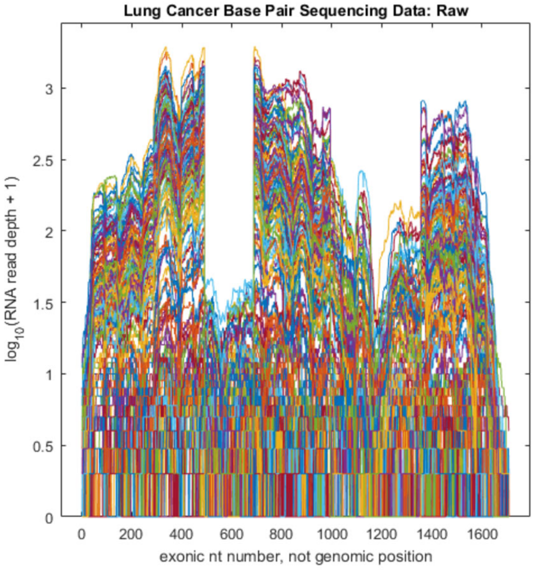


Figure 10: Data curve view of lung cancer RNAseq data. Important relationships in the data are hard to discern. The large steps at the bottom are an artifact of the shifted log transformation.

2.2 Lung Cancer Data

The default form of centering (usually object centering) can sometimes be the best choice depending on the goals of the analysis. One such situation is clustering in the context of RNAseq lung cancer gene expression data from [8]. Here our data matrix contains 180 observations from lung cancer patients of 1709 base pairs along the gene CDKN2A. Figure 10 displays the data as a curve bundle. The horizontal axis represents base pair location on the chromosome and for each location the vertical axis displays the $\log_{10}$ of the counts of RNA reads plus 1. Because these reads overlap, traits near one another appear to be strongly correlated.

To look for relationships within the data, we perform a traditional FDA. In particular, we project the data onto the subspace defined by the first few modes of variation as shown in the left half of Figure 11. The four panels on the left are 1D and 2D scores plots laid out in a similar format to Figure 7. The first two modes of variation suggest three distinct clusters. We study those clusters via *brushing*: manually coloring data based on visual information in the left side of Figure 11 and then transferring those colors to the curve bundle plot in the right panel of Figure 11.

The brushed clusters have a clear, obvious visual interpretation in the curve view of the data. The red individuals have low expression levels across the entire gene (these are classically called *unexpressed*), while the blue & gold individuals are similar but differ in an important way within the range of base pairs between 1000 and 1400. This event is called *alternate splicing* and is very important in cancer research. Focusing on such differences has led to new

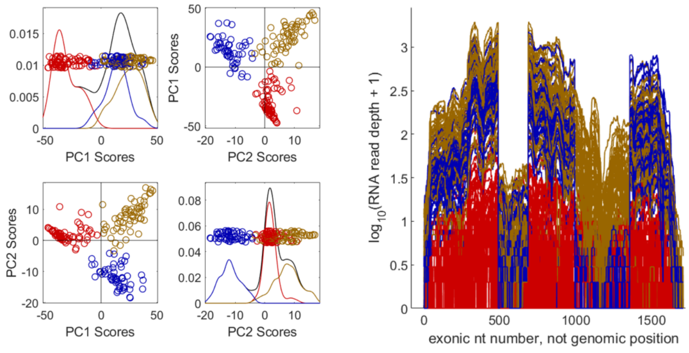


Figure 11: (Left) Brushed scores view of traditional FDA of lung cancer base pair RNA expression data. We have three prominent clusters among the first two modes of variation. (Right) Curve view colored with clusters. Red cluster has low expression everywhere, blue & gold clusters differ between base pairs 1000 and 1400, suggesting alternate splicing as discussed in [8].

discoveries by [8]. This data has a clear correspondence between clusters and modes of variation. In particular, the first mode separates the red observations from the others, while the second mode separates blue from gold with red in the middle.

Given this straightforward and interpretable analysis from traditional FDA, what happens when we double-center the matrix instead? Figure 12 displays a matrix of 1D and 2D scores plots for the trait mean component and first two orthogonal modes of variation. Note that the three clusters are less visually distinct in these views and the correspondence between modes and clusters is less clear. The separation of the red observations is spread over the trait mean and first orthogonal component, and the separation between blue & gold is spread across all three directions. Choosing new clusters by brushing this figure would also be much more challenging as no single two-dimensional view shows three clearly distinguished point clouds like those seen in Figure 11.

In this case, introducing an additional form of centering reduced the interpretability of the results without adding any additional insights. The three-dimensional subspace from the double-centered FDA does no better of a job delineating clusters than the two-dimensional subspace from the typical FDA, and the separation of each group is spread across multiple modes of variation in the double-centered FDA. Opting for double centering over object centering can either enhance interpretability, as in the mortality data, or obscure it, as in the lung cancer data.

3. FORMALISM

3.1 Consequences of Different Forms of Centering

We investigate the effects of grand mean, object, trait, and double centering on a small example data matrix **X**

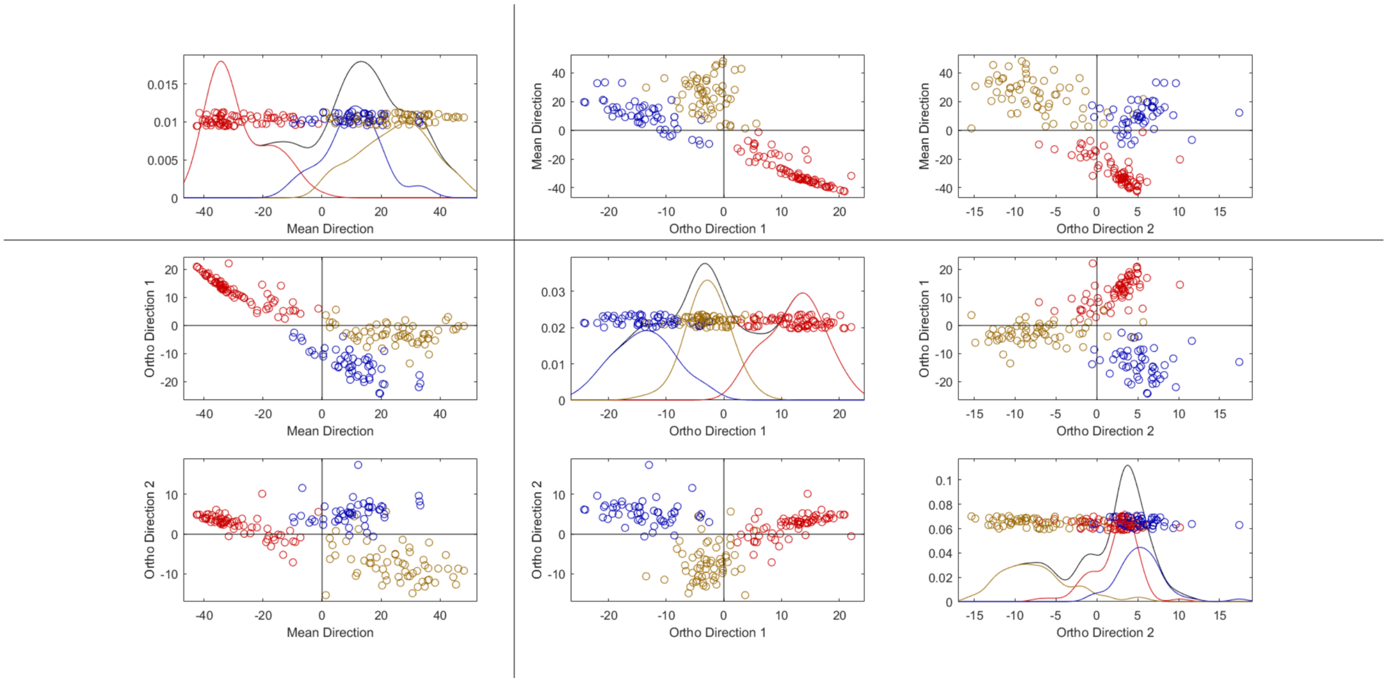


Figure 12: Scores view of double-centered FDA of lung cancer base pair sequencing data. Clusters are made less distinct by involving the trait mean in the visualization.

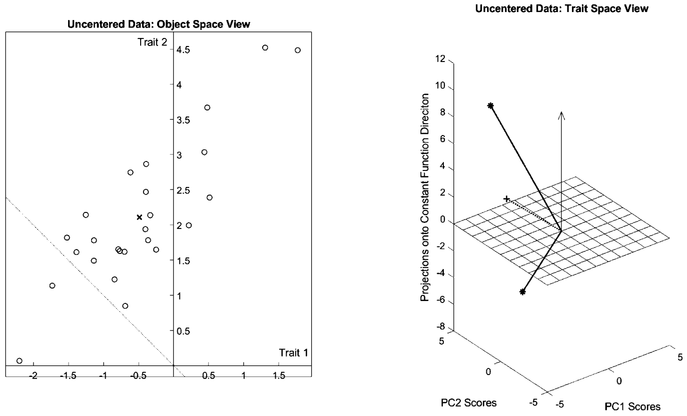


Figure 13: Uncentered data matrix $\mathbf{X}$ shown in both object space ($\mathbb{R}^2$, left) and the three-dimensional subspace of trait space generated by the constant function direction and the data ($\mathbb{R}^{25}$, right). Notably, the asterisks and plus sign in the right panel do not lie in the mesh plane orthogonal to the constant function direction (vertical axis).

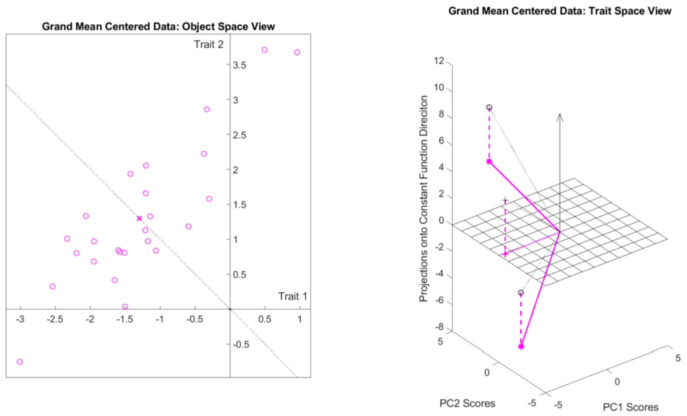


Figure 14: Data matrix $\mathbf{X}_G$, centered version of $\mathbf{X}$ such that all the entries have mean 0, shown in both object space ($\mathbb{R}^2$, left) and trait space ($\mathbb{R}^{25}$, right). Shows grand mean centering is a translation of point clouds in both spaces.

with 2 traits (rows) and 25 data objects (columns). We can then think of $\mathbb{R}^2$ as the *object space*, and $\mathbb{R}^{25}$ as the *trait space*. Figure 13 shows the entries of $\mathbf{X}$ as they exist in both object space and trait space. The left panel shows the 25 ordered pairs (circles) as a scatter plot in $\mathbb{R}^2$. Visualization in $\mathbb{R}^{25}$ is more challenging. For studying centering, the *constant function direction*, i.e. vector of 1's, and the subspace it generates are pivotal. Therefore, the right panel shows the two 25-dimensional trait vectors (asterisks) projected into the three-dimensional subspace of $\mathbb{R}^{25}$ generated by the constant function direction (z-coordinate), and the two orthogonal trait space principal components (x and y coordinates). Note that the subspace orthogonal to the constant function direction contains every vector whose entries have mean 0. The mesh plane represents the projection of that subspace of $\mathbb{R}^{25}$ into the chosen three-dimensional subspace. Also note that in both spaces, the mean vectors of the points ($\times$ in object space, $+$ in trait space) are nonzero. In the right panel, the two data points and their mean are shown as vectors from the origin. In addition to a different symbol, the mean vector is distinguished with a dashed line type.

The following subsections each discuss the results of a different centering on this data. Each subsection has an accompanying figure that visually demonstrates the impacts of each type of centering on the example data matrix in both spaces. Each accompanying figure is formatted similarly to Figure 13: the left panel will show object space, and the right panel will show a projection onto the same subspace of trait space. In each subsection we will also discuss how each centering can be interpreted in a third space: $\mathbb{R}^{d\times n}$, the space of $d \times n$ matrices endowed with the Frobenius norm.

3.1.1 Grand Mean Centering

We begin our geometric exploration of centering with grand mean centering: the form of centering that finds the scalar grand mean value of all the entries of the matrix and subtracts that value from each entry.

We calculate the *grand mean matrix* $\mathbf{M}_G = \mathbf{1}_{d\times n}\mu_G$, where μ_G is the average of all entries of $\mathbf{X}$. The grand-mean-centered version of $\mathbf{X}$ is then denoted $\mathbf{X}_G = \mathbf{X} - \mathbf{M}_G$. While this centering is not often performed on its own in data analysis, it serves as an appropriate first step for analyzing the geometric implications of each subsequent centering. Figure 14 shows the results of this centering in both object space and trait space, where the point clouds retain their shapes but have been translated to different locations. In both spaces, the data are translated parallel to their corresponding constant function direction such that each mean ($\times$ in $\mathbb{R}^2$ and $+$ in $\mathbb{R}^{25}$) lies in the subspace orthogonal to their constant function direction.

While we have described the geometric implications of grand mean centering in both object space and trait space, grand mean centering alone is typically not useful in data analysis. The interpretation and consequences of grand mean centering are better studied in $\mathbb{R}^{d\times n}$. In this space $\mathbf{M}_G$ lies in the constant function direction. Therefore when we subtract $\mathbf{M}_G$ from $\mathbf{X}$, the resulting matrix $\mathbf{X}_G$ is orthogonal to the constant function in the space of matrices. This property then enforces a further orthogonal relationship between the object mean and trait mean matrices, denoted $\mathbf{M}_O$ and $\mathbf{M}_T$ respectively. We calculate $\mathbf{M}_O = \boldsymbol{\mu}_d \mathbf{1}_n^T$ and $\mathbf{M}_T = \mathbf{1}_d \boldsymbol{\mu}_n^T$, where $\boldsymbol{\mu}_d$ is the d-dimensional vector whose entries are the mean of each trait of $\mathbf{X}$ and $\boldsymbol{\mu}_n$ is the n-dimensional vector whose entries are the mean of each data object of $\mathbf{X}$. Each of $\mathbf{M}_O$ and $\mathbf{M}_T$ are rank 1, and $\mathbf{M}_O$ has identical columns while $\mathbf{M}_T$ has identical rows. If these mean matrices are calculated with respect to $\mathbf{X}_G$ rather than $\mathbf{X}$,

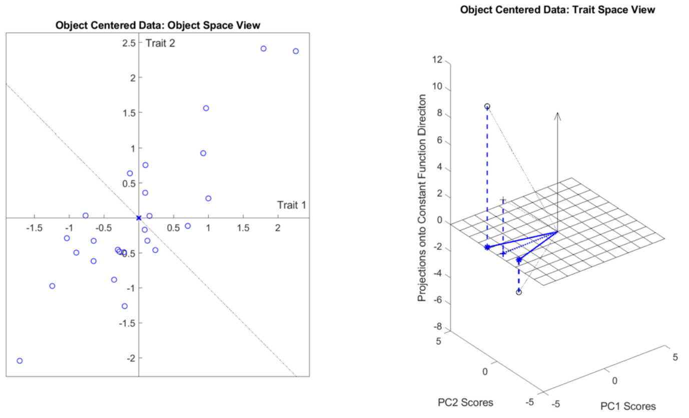


Figure 15: Data matrix $\mathbf{X}_O$, centered version of $\mathbf{X}$ such that the data objects have mean vector $\mathbf{0}$, shown in both object space ($\mathbb{R}^2$, left) and trait space ($\mathbb{R}^{25}$, right). The trait vectors are projected onto the subspace orthogonal to the constant function direction.

both of these matrices have entries that sum to 0. In this case each column of $\mathbf{M}_O$ sums to 0 and each row of $\mathbf{M}_T$ sums to 0. This means that with respect to the Frobenius inner product in $\mathbb{R}^{d\times n}$, $\mathbf{M}_O$ and $\mathbf{M}_T$ are orthogonal after grand mean centering.

3.1.2 Object Centering

Now we explore the centering which is performed on data matrices as a part of typical FDA. We calculate $\mathbf{X}_O$, the object-centered version of $\mathbf{X}$ such that the data objects have a mean of the d-dimensional $\mathbf{0}$ vector. Figure 15 shows the results of this form of centering on $\mathbf{X}$. As shown in the left panel, the points in object space now have a mean vector at exactly the origin. The points have all been translated from their locations in Figure 13 by the same amount and in the same direction. The two trait vectors undergo a different transformation in $\mathbb{R}^{25}$. Each vector, as well as their vector mean, is projected into the 24-dimensional subspace which is orthogonal to the constant function direction. This 24-dimensional subspace is again represented by the mesh plane.

To facilitate a PCA-like decomposition into modes of variation of $\mathbf{X}_O$, consider an SVD of $\mathbf{X}_O = \mathbf{U}_O\mathbf{D}_O\mathbf{V}_O^T$. In our matrix orientation convention, $\mathbf{U}_O$ is associated with loadings and $\mathbf{V}_O$ is associated with scores. The trait vectors of $\mathbf{X}_O$ lie in the subspace orthogonal to the constant function direction in $\mathbb{R}^{25}$. Therefore, the orthonormal basis for their span, i.e. the columns of $\mathbf{V}_O$, will also be composed of vectors orthogonal to the constant function direction. These entries represent the scores of each observation along each direction, and centering this way guarantees that each set of scores has mean 0.

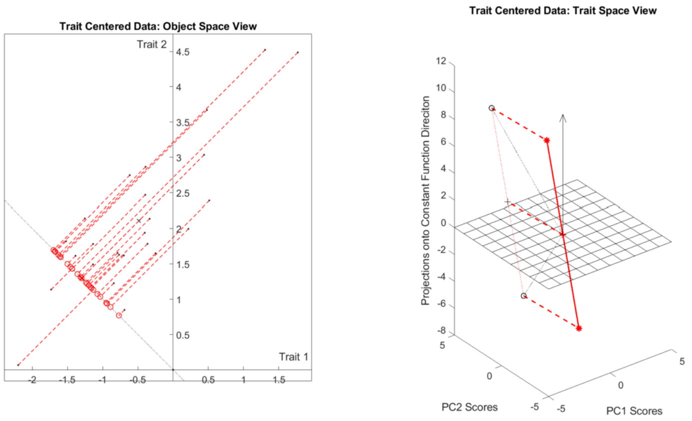


Figure 16: Data matrix $\mathbf{X}_T$, centered version of $\mathbf{X}$ such that the traits have mean vector $\mathbf{0}$, shown in both object space ($\mathbb{R}^2$, left) and trait space ($\mathbb{R}^{25}$, right). As a consequence, the objects are projected onto the subspace orthogonal to the respective constant function direction. This projection demonstrates that $\mathbf{X}_T$ is of lower rank than $\mathbf{X}$.

3.1.3 Trait Centering

The second centering is the dual of the centering used in PCA. We calculate $\mathbf{X}_T$: the centered version of $\mathbf{X}$ such that the traits have a mean of the n-dimensional $\mathbf{0}$ vector. Figure 16 shows the results of this centering on the data matrix $\mathbf{X}$ from Figure 13. The left panel shows that the points in object space have been projected onto the subspace orthogonal to the $\mathbb{R}^2$ constant function direction, while the right panel shows that the points in trait space have been translated such that their mean is at exactly the origin. This result is of course the dual of the previous form of centering. Note that the resulting matrix $\mathbf{X}_T$ is now rank 1 instead of rank 2.

To similarly find the modes of variation of $\mathbf{X}_T$, consider an SVD of $\mathbf{X}_T = \mathbf{U}_T\mathbf{D}_T\mathbf{V}_T^T$. The data object vectors of $\mathbf{X}_T$ lie in the subspace orthogonal to the constant function direction in $\mathbb{R}^2$. Therefore, the orthonormal basis for their span, i.e. the columns of $\mathbf{U}_T$, will also be composed of vectors in this subspace. These entries represent the unweighted loadings of each trait within each mode of variation of the data objects, and centering this way guarantees that each set of loadings has mean 0.

3.1.4 Double Centering

The final mode of centering combines the operations of both previous forms into a single transformation. We calculate $\mathbf{X}_D$, the double-centered version of $\mathbf{X}$ where the traits have a mean of the n-dimensional $\mathbf{0}$ vector and the data objects have a mean of the d-dimensional $\mathbf{0}$ vector. Figure 17 shows the results of double centering the matrix $\mathbf{X}$ from Figure 13. In both panels, the original points have been translated so that their mean lies at the origin and they

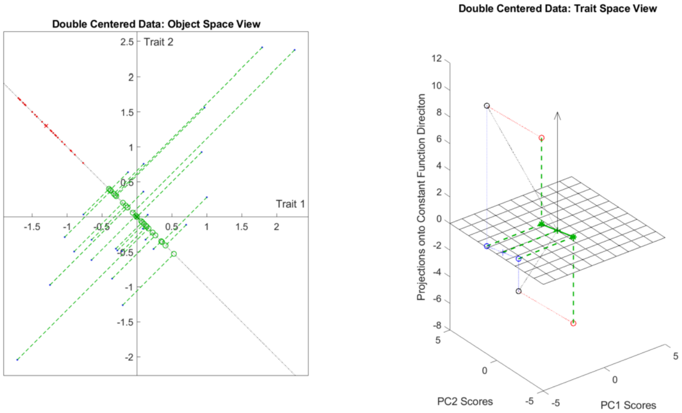


Figure 17: Data matrix $\mathbf{X}_D$, centered version of $\mathbf{X}$ such that the data objects and traits both have mean vector $\mathbf{0}$, shown in both object space ($\mathbb{R}^2$, left) and trait space ($\mathbb{R}^{25}$, right). As a consequence, both the data objects and traits are projected onto respective subspaces orthogonal to the constant function direction.

are projected onto the subspace orthogonal to the corresponding constant function direction. The double-centered ordered pair objects in the left panel and corresponding double-centered trait vectors are all shown in green.

Note that these two operations, projection and translation, are commutative. Projecting the first-translated points results in the same transformed data as translating the first-projected points. We can see this commutation in both panels of Figure 17. In the left panel we can arrive at the green points either by projecting the previously-translated blue points onto the line orthogonal to the constant function direction or by translating the previously-projected red points such that their vector mean now lies at the origin. In the dual situation in the right panel, we can arrive at the green points either by projecting the previously-translated red points onto the mesh plane indicating the subspace orthogonal to the constant function or by translating the previously-projected blue points such that their vector mean now lies at the origin.

This matrix is also of lower rank than $\mathbf{X}$.

To similarly find the modes of variation of $\mathbf{X}_D$, consider an SVD of $\mathbf{X}_D = \mathbf{U}_D\mathbf{D}_D\mathbf{V}_D^T$. Both the data objects and trait vectors lie in subspaces orthogonal to their respective constant function direction, so both the sets of loadings and the sets of scores for this data will have mean 0.

3.2 Discussion

There is a strong connection between mutual orthogonality of vectors and correlation in corresponding scatterplots that is driven by the means of the entries of each vector. For two observed unit vectors $\mathbf{x}$ and $\mathbf{y}$, because $\sum x_i^2 = \sum y_i^2 = 1$, the correlation of their entries is $Corr(\mathbf{x},\mathbf{y}) = \sum(x_iy_i) - \sum(x_i)\sum(y_i)$. Since FDA scores and loadings vectors are mutually orthogonal in their respective spaces, we will always have $\sum(x_iy_i) = 0$. Therefore a sufficient condition for the entries of two scores and/or loadings vectors to be *uncorrelated* is for the entries of one vector to have mean zero.

Table 2. Summary of which centerings produce which outcome for sets of scores and loadings vectors in FDA.

↓ Effect, Centering Type →	None	Object	Trait	Double
Orthogonal Score Vectors	✓	✓	✓	✓
Uncorrelated Score Vectors		✓		✓
Orthogonal Loadings Vectors	✓	✓	✓	✓
Uncorrelated Loadings Vectors			✓	✓

Table 2 summarizes how this condition enforces uncorrelatedness in scores and loadings vectors found via FDA of differently-centered matrices. Whether FDA of a single matrix is treated as an eigenanalysis of a covariance matrix or as an SVD of a (possibly centered) data matrix, the scores and loadings vectors will always be mutually orthogonal. This fact combined with the projection operations involved in different centerings can produce mutually uncorrelated scores and/or loadings vectors. This uncorrelatedness is most prominent and important when forming scatter plots like those shown in Figure 7.

As a remark, some of the centerings resulted in loss of rank in our synthetic data matrix. Recall that the matrix was 2 × 25, and the matrix became rank 1 after trait centering and double centering. The centerings that involved translation in $\mathbb{R}^{25}$, and therefore projection in $\mathbb{R}^2$, were the ones that reduced the rank of the matrix. In general, the centering that involves projection in the lower-dimensional space out of the trait vector and object vector spaces will result in loss of rank.

4. QUANTIFYING DOUBLE CENTERING IN FUNCTIONAL DATA ANALYSIS

As discussed in Sections 2 and 3, object centering is the standard default for FDA. This is recommended because interesting structure is often found in variation about that mean vector so its dominating effect is removed and treated separately. As seen in the transition between the left panels of Figures 13 and 15, subtraction of the object mean results in a translation of the data objects in object space $\left(\mathbb{R}^d\right)$ so their mean vector becomes the origin. The FDA modes of variation among the now-translated point cloud are then readily calculable via SVD.

However, in cases like the Spanish Mortality data studied in Section 2, an additional dominating effect due to the trait mean can remain within the point cloud even after removal of the object mean. In object space, the trait mean manifests through projection onto the constant function direction. Each entry of the trait vector mean is the signed magnitude of the projection of a corresponding data object

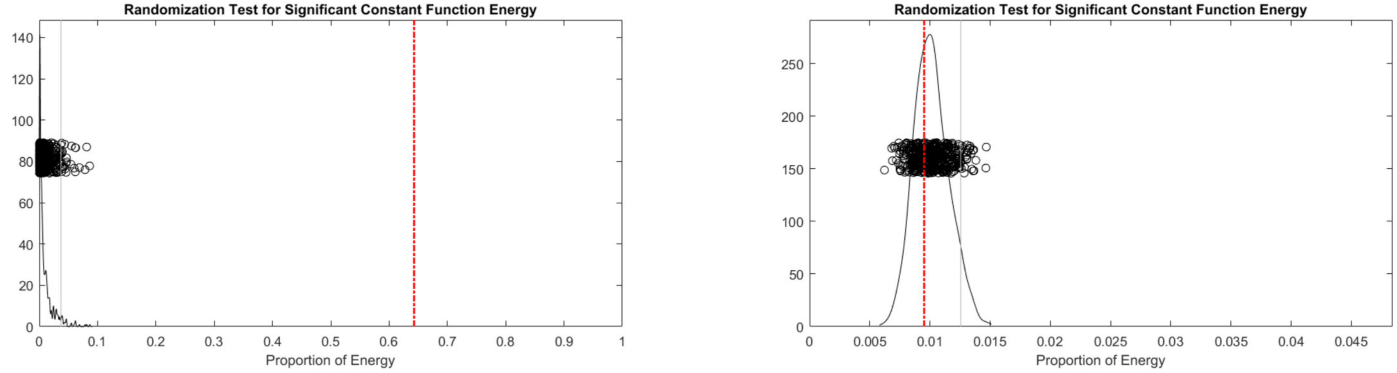


Figure 18: Left Panel: Direction Energy Hypothesis Test on Spanish Mortality data. Energy proportion in constant function is much larger than what would be expected due to random chance. Right Panel: Direction energy hypothesis test on $n = 100$ synthetic 100-dimensional Gaussian observations. Energy proportion in constant function direction is not distinct from empirical null distribution.

vector onto the constant function direction. If a substantial proportion of the object-centered point cloud energy lies along the constant function direction, the trait mean effect may be concealing more interesting structure. In this section we develop a *direction-energy* hypothesis test to determine when the proportion of energy in the constant function direction becomes "substantial" enough to warrant potential separate consideration of the trait mean mode of variation and the remaining (double centered) modes.

Consider an object-centered data matrix $\mathbf{X}_O$. Define the *total energy* of the data matrix as its squared Frobenius norm, $E_{total} = ||\mathbf{X}_O||_2^2$. Define the *energy in a direction* for the data matrix as the squared Frobenius norm of the projection of $\mathbf{X}_O$ onto the 1-dimensional subspace spanned by the unit vector $\mathbf{v}$ in $\mathbb{R}^d$, $E_{\mathbf{v}} = ||\mathbf{v}\mathbf{v}^\top \mathbf{X}_O||_2^2$. For any given unit vector $\mathbf{v}$, primary interest is in the proportion $p_{\mathbf{v}} = \frac{E_{\mathbf{v}}}{E_{total}}$ of the total energy attributable to that direction. To evaluate the potential significance of the proportion of energy in the constant function direction $p_{\mathbf{1}_d}$, we generate an empirical null distribution of energy proportions in D randomly chosen directions. If $p_{\mathbf{1}_d}$ lies above a high percentile of that empirical null distribution, we say there is evidence of statistically significant energy in the constant function direction for the data matrix. Note that the number of sampled random directions D should scale with the dimension of object space d to ensure a representative sample of the possible energies for the empirical null distribution.

Figure 18 visually displays the results of the hypothesis test described above for two data sets. The left panel shows the test as administered to the mortality data, and the right panel shows the test administered to a synthetically generated matrix of 100 observations from a 100-dimensional standard normal distribution. We plot energy proportion on the horizontal axis with the vertical axis representing density of the randomly generated energy proportions The black dots in each panel represent $D = 500$ energy proportions from the randomly chosen directions. The black curve in each panel is a smooth histogram representing the empirical null distribution of random direction energy proportions. The red dot-dash line in each panel represents the energy proportion in the constant function direction. For the Spanish mortality data in the left panel, nearly 65% of the energy is congregated around the constant function direction, as is apparent from Figures 6, 8, and 9. This proportion is much higher than that in any random direction, whose energy proportions are shown with the black circles. This result indicates that substantial gains in interpretability are possible for this data set via opting to double-center the data matrix before performing FDA, as demonstrated in Figure 9 and Table 1. Contrastingly, in the right panel, the energy proportion in the constant function direction is not remarkable in any way in the spherically-symmetrical synthetic data displayed.

We can further study the effects of removal of the trait mean by plotting the energies in each FDA component before and after double centering. Figure 19 shows such a breakdown for the Spanish mortality data (left panel) and breast cancer RNAseq data studied in [2] (right panel). The solid blue lines show how much energy is accounted for in each object-centered FDA component; components are shown in the order they're found from bottom to top in the figure. The red dashed lines show how much energy is accounted for in each double-centered FDA component; and they're displayed in a similar fashion to the blue lines. All energy proportions are in terms of the total energy in the object-centered data matrix, so the constant function direction energy is included in the total for the double-centered FDA. Consequentially, there is less energy to be allocated for the double-centered components. The blue lines in the left panel correspond with the components shown in Figure 6 while the red dashed lines in the left panel correspond with the components shown in Figure 9.

In the object-centered FDA of the mortality data, the first component accounted for more than 95% of the energy

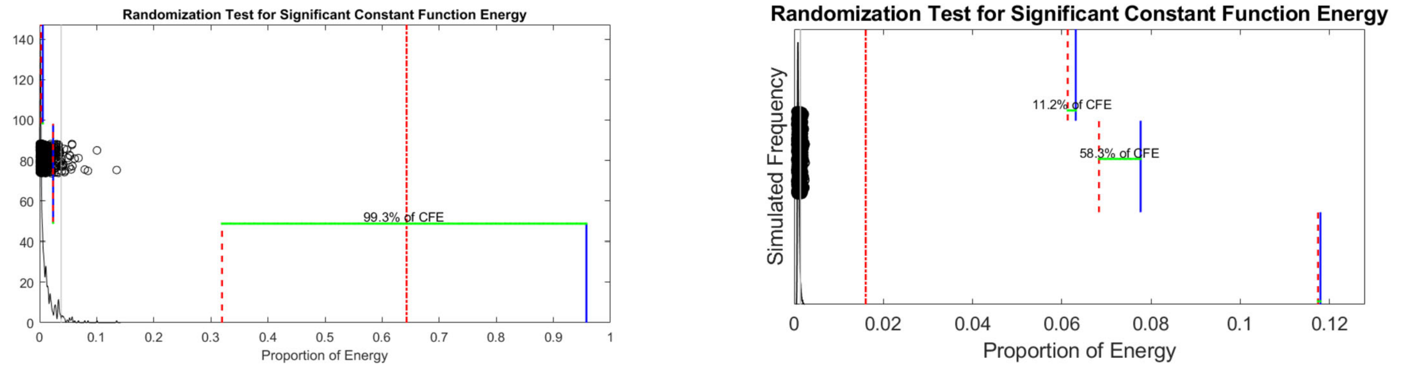


Figure 19: Direction energy hypothesis tests on Spanish Mortality (left) and Breast Cancer RNAseq (right) with energy breakdowns by FDA component. Left Panel: Constant function direction almost entirely takes energy from the first component. Right Panel: Constant function direction contains significant information, but its spread across many components and its total energy share is small relative to the first several components.

in the data, but after double centering that energy is split between the constant function direction and the first orthogonal component. In fact the drop in energy share of the first component between object-centered & double-centered FDA accounts for 99.3% of the energy share of the constant function direction. This corresponds with the interpretation of this operation in Section 2, where the first object-centered FDA component contained information about both overall improvement and greater improvement for young people, while the first double-centered FDA component is only about greater improvement for young people as the overall improvement is sequestered to the constant function mode.

The effect of the constant function direction is much less pronounced in the RNAseq data. While it appears to include a statistically significant amount of the overall matrix energy, its energy proportion is still trumped by those of several orthogonal principal components, including the three shown in the right panel. This is likely because a procedure of a similar flavor to removal of the trait mean has already been performed on this data. The columns of this data are normalized such that each has an identical upper quartile. While this operation doesn't entirely remove the effect of the trait mean, it still removes much of the variation in the data objects not explained by the traits.

5. DISCUSSION

In this manuscript we presented a unified framework for describing, understanding, and implementing centering as part of data matrix analysis. We put forth disambiguating terminology for describing data matrix dimensions and centering operations. We highlighted double centering in FDA as a way to incorporate the constant function direction mode of variation in data analyses. Correspondingly, we proposed a hypothesis test for determining whether the constant function direction mode of variation is significant for a given data set.

Functional data made up of curves with differing numbers of observations or observations at differing points in time don't align neatly with the framework presented above. To consider the impacts of different forms of centering for more complex data, suitable choices for data object space and trait space must first be identified. One common approach for this form of functional data is to choose a basis in the space of square-integrable functions over $\mathbb{R}$ and approximate each observation as a linear combination of functions in that basis. In this case the data object space is populated by the vectors of coefficients of that linear combination, and the trait space is a subspace of the infinite-dimensional function space from which the basis was chosen. Consequently, the ideas presented in this paper should generalized to the settings where the functions may be observed at irregular and/or differing points, and the constant function will still represent an important direction in the trait space. In section 2.4 of the FDA review paper by [13], they describe an example of *functional principal components analysis* (FPCA) on observations of helper T cell counts in 25 patients at differing time points. Figure 4 of the review paper displays plots of the first four functional modes of variation after the FPCA. The first and by far the most explanatory mode for this data in fact appears to be a slightly perturbed version of the constant function. Projection into the space orthogonal to the constant function appears to be a potentially fruitful further direction of research.

APPENDIX A. CENTERING IN DATA INTEGRATION

In data integration tasks, two or more data matrices that share common row and/or column dimensions are analyzed in tandem. The goals of such tasks are to reveal what information is shared between the data matrices and to combine the information from each data block to arrive at a more complete picture of the population and/or traits in question. Data matrix centering can play an even more complex role in these situations than in single-matrix FDA as both data objects and traits play critical roles in the analysis. Depending on the chosen methodology, different centering choices may affect the outputs in surprising ways.

We explore the integrative analysis of two data blocks using partial least squares (PLS) under different centering regimes. We choose PLS as a simple and direct method that takes into account potential scaling differences between data blocks.

A.1 Partial Least Squares

PLS as a data integration procedure derives shared information from the *cross-covariance matrix* $\mathbf{\Sigma}_{1,2}$ between the two data blocks. Here, the cross-covariance matrix refers to the submatrix of the grand covariance matrix of all traits in either data block associated with the covariances between traits across blocks:

$$\mathbf{cov}\left(\begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\right) = \begin{bmatrix}\mathbf{\Sigma}_1 & \mathbf{\Sigma}_{1,2}\\ \mathbf{\Sigma}_{1,2}^T & \mathbf{\Sigma}_2\end{bmatrix}$$

We can also find the cross-covariance matrix by multiplying object-centered versions of the two data blocks together: $\mathbf{\Sigma}_{1,2} = \frac{1}{n}\mathbf{X}_{1O}\mathbf{X}_{2O}^T$. As PLS operates on covariance matrices, it chooses pairs of score vectors for each data block with maximal covariance between them.

Importantly, as discussed in [11], different variations of PLS lead to different centering-based consequences. As is the case in many data integration methods, one piece of information, either scores or loadings, is calculated first while the other is found subsequently with a projection operation involving the first piece of information and the original data blocks. Whichever set of vectors is found first will be predictably affected by centering choices during preprocessing of the data blocks, but the subsequently found set of vectors are typically not even mutually orthogonal due to the projection.

One approach is to directly take a singular value decomposition of the cross-covariance matrix; the resulting left and right singular vectors then constitute the estimated loadings vectors for $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively. This results in loadings vectors that are uncorrelated only when the data blocks are double centered. As per Table 2, trait centering and double centering are the two choices that result in loadings vectors with uncorrelated entries. We do not consider trait centering as a possible choice since object centering is required to correctly form the cross-covariance matrix in the first place.

Another approach is to sequentially and algorithmically calculate each score vector, then its corresponding loadings vector, then remove the one-dimensional subspace approximation defined by those vectors before searching for subsequent scores and loadings vectors. As this procedure calculates score vectors first, we can guarantee that the calculated score vectors will be uncorrelated due to object-centering the data blocks. We opt for this approach to mirror other data integration methods that first locate score vectors, including canonical correlations analysis (CCA) from [6] and angle-based joint and individual variation explained (AJIVE) from [3].

A.2 Synthetic Data Example

To demonstrate the additional complexities involved in centering choice for data integration, we first use the synthetic two-block data set shown in Figure 20. The first block, $\mathbf{X}_1$, is 300 × 200, and the second block $\mathbf{X}_2$, is 500 × 200. Note that each block has the same number of data objects (columns) but different numbers of traits (rows). For example, one could represent demographic data and the other could represent various biomarker observations about a cohort of patients. Each data block is formed by adding a rank-two signal matrix to a full-rank Gaussian noise matrix. The underlying components of each signal matrix lie in the same common subspace of trait space, representing shared information between the blocks. However, the overlapping subspaces are obscured by object and trait mean effects in each matrix.

Figure 20 uses heatmaps to display the construction of the synthetic data example we will use to demonstrate the value of exploring double centering in data integration contexts. The left panels show the observed data matrices, which are formed by additively combining the other matrices in each respective row. The heatmaps in the second column display a shared, underlying rank-two signal in both $\mathbf{X}_1$ and $\mathbf{X}_2$. By construction, this underlying rank-two joint signal is double centered. The heatmaps in the third column show the mean effects added to each matrix. The matrix added to $\mathbf{X}_1$ is rank 1 and represents an object mean matrix as each column is identical. The matrix added to $\mathbf{X}_2$ is rank 2 and represents a double mean matrix with both object mean and trait mean components. Finally, the heatmaps in the fourth column display the i.i.d. Gaussian noise added to the observations. The color scale is kept constant across all heatmaps in Figure 20 to appropriately convey differences in effect size between the shared signal and mean effects.

The object mean vector added to columns of $\mathbf{X}_1$ increased the values in the top 100 rows and decreased the values in the bottom 100 rows. An object mean vector was added to $\mathbf{X}_2$, but its visual impression is swamped by that of the trait mean effect. The trait mean vector has entries that gradually increase from the first observation's entry to the last. This creates the color gradient visual effect seen in the third panel of the second row.

We perform PLS on this two block data set after object centering and after double centering. Figure 21 displays the first two PLS components of each data block found using the object-centered versions of the matrices. The top panels show the $\mathbf{X}_1$ components and the bottom panels show the

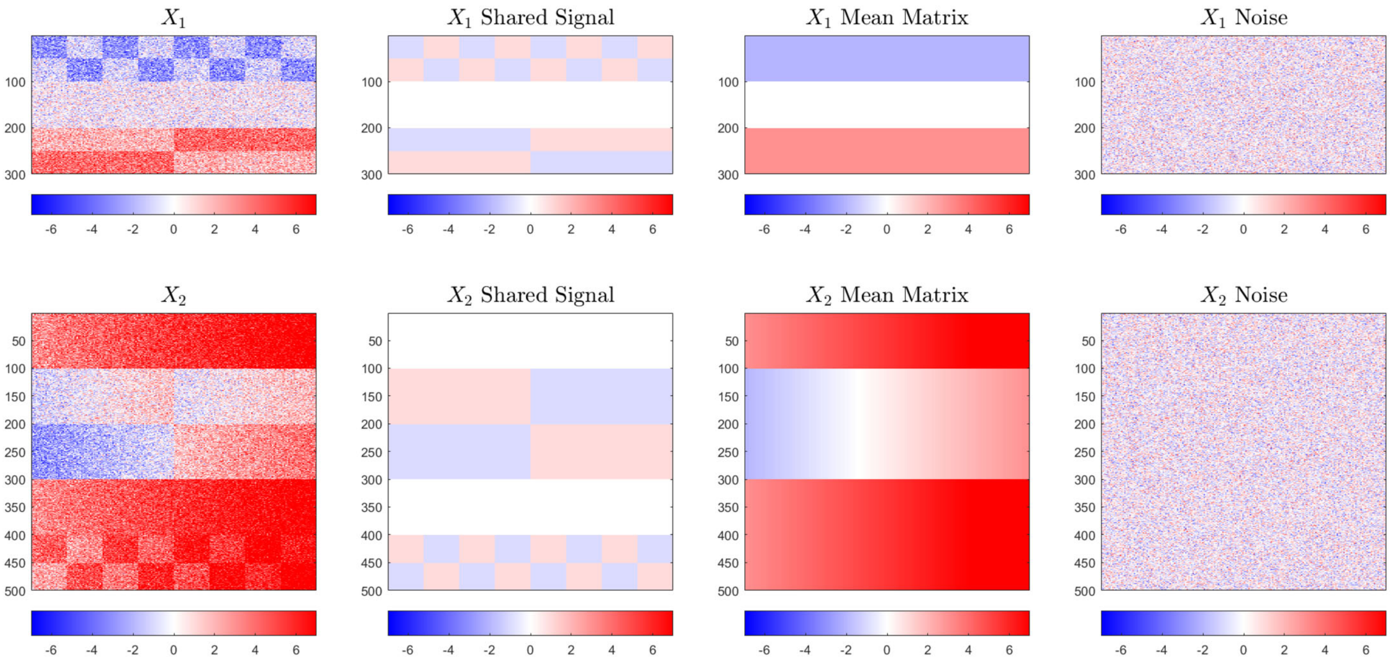


Figure 20: Three stages of synthetic data example construction. Underlying rank-two signal (left), underlying signal with added mean effect perturbation (middle), noise perturbation (right).

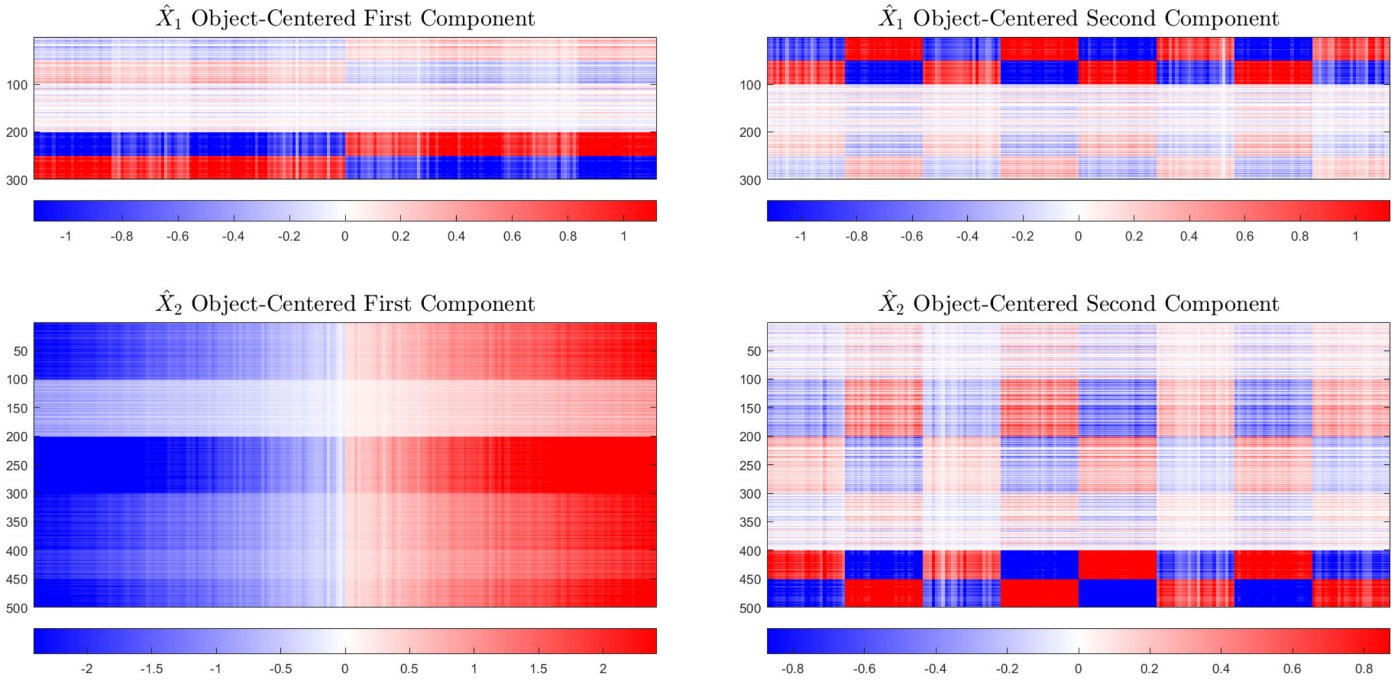
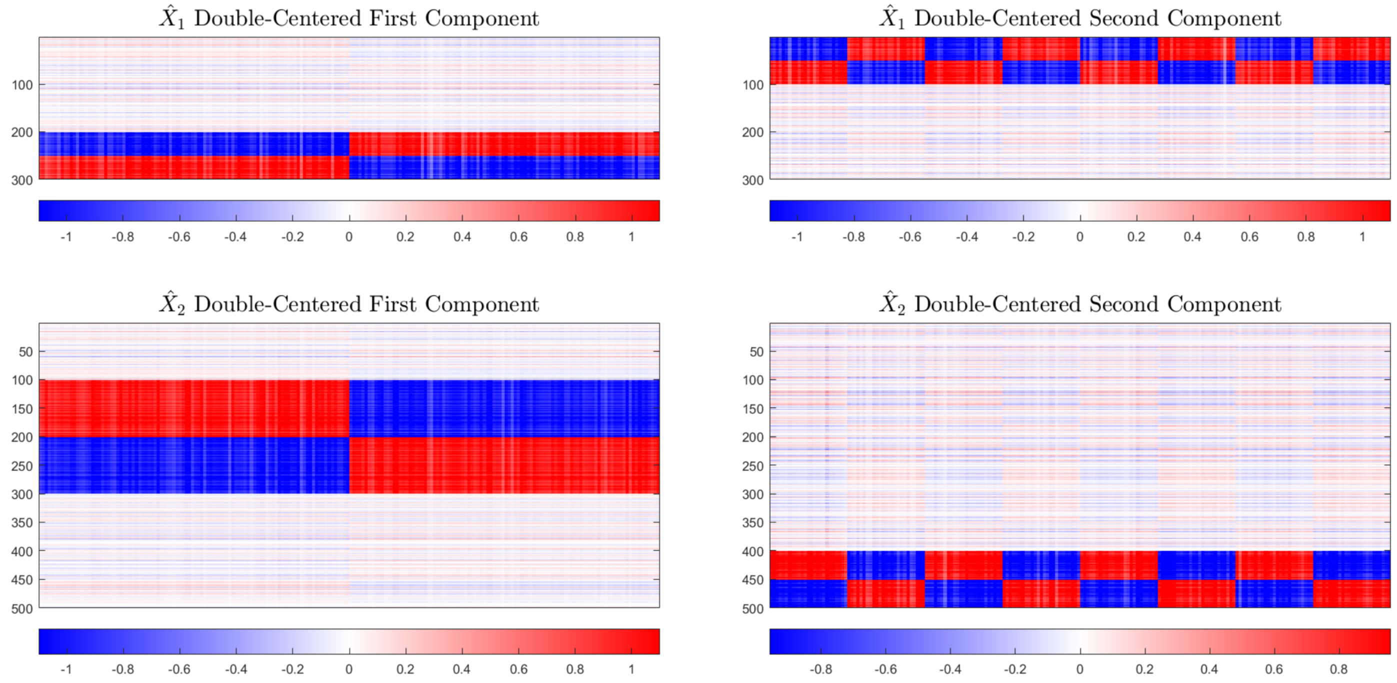


Figure 21: First two PLS components of each object centered synthetic data block. Recovery and parsing of distinct underlying signal pieces is reasonable for $\mathbf{X}_1$ but the trait mean effect dominates the first component of $\mathbf{X}_2$.

$\mathbf{X}_2$ components. Each estimated $\mathbf{X}_1$ component roughly corresponds with one of the rank one underlying signal components shown in the left panels of Figure 20, as expected. This is because $\mathbf{X}_1$ only had an object mean added to its shared signal. However, the first $\mathbf{X}_2$ component is completely dominated by the large linear trend in the trait mean rather than one of the underlying shared effects. This is a consequence of PLS choosing score vectors to maximize covari-



$\hat{X}_1$ Double-Centered First Component

$\hat{X}_1$ Double-Centered Second Component

$\hat{X}_2$ Double-Centered First Component

$\hat{X}_2$ Double-Centered Second Component

Figure 22: First two PLS components of each double centered synthetic data block. Recovery and parsing of distinct underlying signal pieces is strong for both blocks.

ance rather than correlation. The trait mean effect is much larger in magnitude than the underlying shared structure, so the best way to maximize covariance is to choose a score vector close to the trait mean effect for $\mathbf{X}_2$.

Figure 22 displays the first two PLS components of each data block found using the double centered versions of the matrices. The top panels show the $\mathbf{X}_1$ components and the bottom two panels show the $\mathbf{X}_2$ components. Now that the strong trait mean effect in $\mathbf{X}_2$ has been removed, the recovery of the underlying shared signal is greatly improved in both blocks. The first component in both blocks is distinctly the long-checkered pattern and the second component in both blocks is distinctly the short-checkered pattern.

In this synthetically constructed example, a strong trait mean effect dominated the calculated components from data integration. Removal of the trait means of each block in addition to the typical and necessary removal of the object means of each block drastically improved interpretability of results.

A.3 Spanish Mortality: Males versus Females

We return to the Spanish mortality data from Section 2 to further explore the implications of additional centering in data integration tasks. Here we combine the observations of male mortality rates from 1908 to 2002 with corresponding measurements of female mortality rates over the same time period. We perform PLS on these two data blocks to locate the shared information between them. We opt for the algorithmic approach outlined in Section A.1 to ensure score vectors are orthogonal. We will compare the analysis after object centering and double centering. In both of these centering regimes, the score vectors will be mutually uncorrelated (See Table 2).

Figure 23 shows the results of PLS on the two data blocks after each has been object centered. We display the loadings vectors scaled by the scores of each observation in a similar fashion to Figures 6, 8, and 9. The object mean and first three joint modes of variation for males are shown on the left, and the corresponding modes for females are shown on the right. While each mode of variation manifests differently in each data block, the same broad trends are identifiable for each gender. The first mode shows overall improvement and more dramatic improvement for younger people, and the second mode shows a contrast between younger adults and the rest of the population. This contrast highlights ages 18–50 for males and ages 15–45 for females. The third mode is much harder to interpret as there is no obvious commonality between the patterns for each gender outside of the appearance of age rounding. Overall these modes correspond with those found via the PCA analysis of male mortality in Section 2. Since PCA finds modes of maximal variation and PLS tries to find directions with maximal covariance between blocks, this correspondence between PCA and PLS modes of variation is not surprising.

Figure 24 shows the results of PLS on the two data blocks after each has been double centered, organized in a similar fashion to Figure 23. Again the first two modes of variation match expectations. The first mode shows stronger improve-

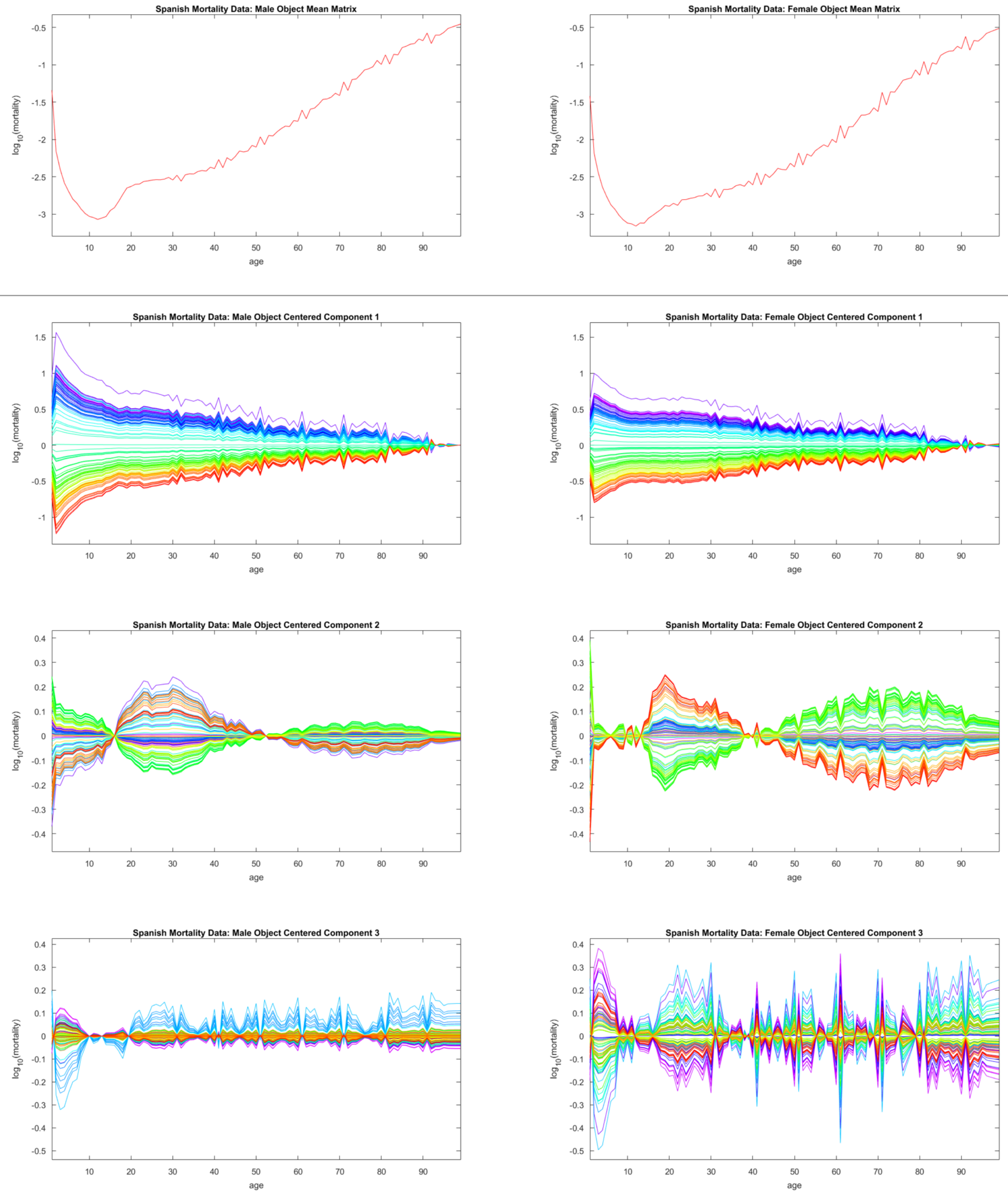


Figure 23: Object-centered PLS of male and female mortality rates. First and second mode show expected trends, but third mode is challenging to interpret.

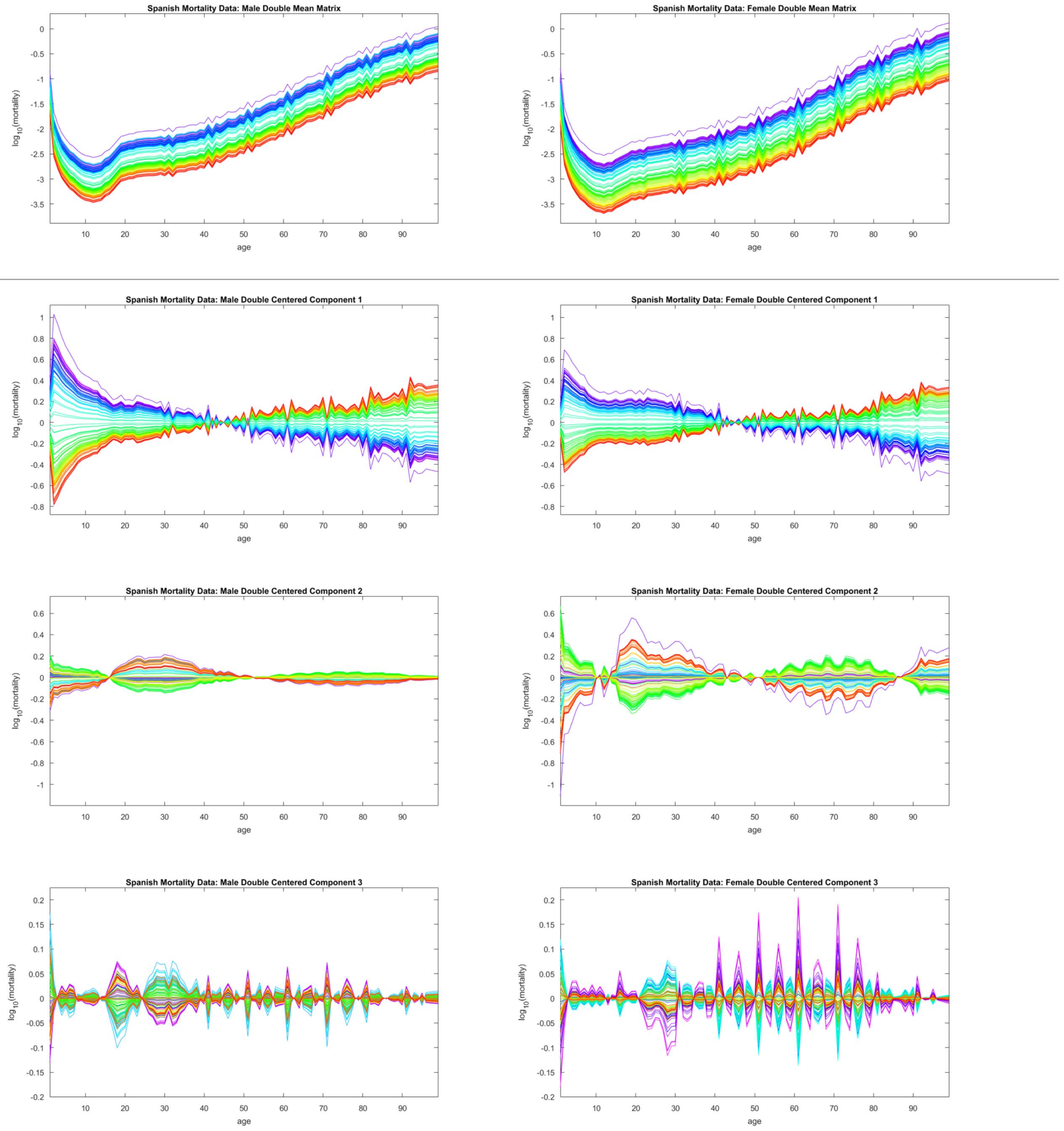


Figure 24: Double-centered PLS of male and female mortality rates. First and second mode show expected trends, and third mode highlights record-keeping anomalies in each gender.

ment for younger individuals as the overall improvement has been removed with the trait mean, and the second mode again shows a contrast between younger adults and the rest of the population. The third mode now more clearly pertains to age rounding for both males and females. In addition to large spikes every ten years, we also see smaller spikes every five years, further reflecting a bias towards rounder numbers on death certificates of older individuals.

As in previous analyses of this kind of data, we feel the choice to double-center each data block enhances the interpretability of the results.

ACKNOWLEDGEMENTS

The authors would like to thank Charles Perou, Katherine Hoadley, Chris Fan, and Andres Alonso for the collection and maintenance of the data used in this manuscript.

FUNDING

Research partially supported by the National Science Foundation under Grant Nos. IIS-1633074, DMS-1916115, DMS-2113404, and DMS-2210337.

Accepted 20 March 2023

REFERENCES

[1] Canudas-Romo, V., Glei, D., Gómez-Redondo, R., Coelho, E. and Boe, C. (2008). Mortality changes in the Iberian Peninsula in the last decades of the twentieth century. *Population* **63** 319–343.

[2] Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., Bowlby, R., Shen, H., Hayat, S., Fieldhouse, R., Lester, S. C., Tse, G. M. K., Factor, R. E., Collins, L. C., Allison, K. H., Chen, Y.-Y., Jensen, K., Johnson, N. B., Oesterreich, S., Mills, G. B., Cherniack, A. D., Robertson, G., Benz, C., Sander, C., Laird, P. W., Hoadley, K. A., King, T. A., Network, T. R. and Perou, C. M. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163**(2) 506–519. https://doi.org/10.1016/j.cell.2015.09.033

[3] Feng, Q., Jiang, M., Hannig, J. and Marron, J. S. (2018). Angle-based joint and individual variation explained. *J. Multivariate Anal.* **166** 241–265. https://doi.org/10.1016/j.jmva.2018.03.008. MR3799646

[4] Gómez-Redondo, R. and Boe, C. (2005). Decomposition analysis of Spanish life expectancy at birth: Evolution and changes in the components by sex and age. *Demographic Research* **S4**(20) 521–546. https://www.demographic-research.org/special/4/20/s4-20.pdf. https://doi.org/10.4054/DemRes.2005.13.20

[5] Horváth, L. and Kokoszka, P. (2012) *Inference for Functional Data with Applications. Springer Series in Statistics*. Springer New York. https://books.google.com/books?id=OVezLB___ZpYC. https://doi.org/10.1007/978-1-4614-3655-3. MR2920735

[6] Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* **28**(3-4) 321–377. https://academic.oup.com/biomet/article-pdf/28/3-4/321/586830/28-3-4-321.pdf. https://doi.org/10.1093/biomet/28.3-4.321

[7] Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley Series in Probability and Statistics*. Wiley. https://books.google.com/books?id=YjsbAAAACAAJ. https://doi.org/10.1002/9781118762547. MR3379106

[8] Kimes, P. K., Cabanski, C. R., Wilkerson, M. D., Zhao, N., Johnson, A. R., Perou, C. M., Makowski, L., Maher, C. A., Liu, Y., Marron, J. S. and Hayes, D. N. (2014). SigFuge: single gene clustering of RNA-seq reveals differential isoform usage among cancer samples. *Nucleic Acids Res.* **42**(14) 113.

[9] Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis. Chapman & Hall/CRC Texts in Statistical Science*. CRC Press. https://books.google.com/books?id=aHE3DwAAQBAJ.

[10] Marron, J. S. and Alonso, A. M. (2014). Overview of object oriented data analysis. *Biom. J.* **56**(5) 732–753. https://doi.org/10.1002/bimj.201300072. MR3258083

[11] Rosipal, R. and Krämer, N. (2005). Overview and Recent Advances in Partial Least Squares. **3940** 34–51. https://doi.org/10.1007/11752790_2

[12] Schofield, R., Reher, D. and Bideau, A., eds. (1991). *The Decline of Mortality in Europe*. Oxford University Press. https://EconPapers.repec.org/RePEc:oxp:obooks:9780198283287.

[13] Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016). Functional Data Analysis. *Annual Review of Statistics and Its Application* **3**(1) 257–295. https://doi.org/10.1146/annurev-statistics-041715-033624

[14] Wilmoth, J. R. and Shkolnikov, V. *Human mortality database*. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany).

[15] Zhang, L., Marron, J. S., Shen, H. and Zhu, Z. (2007). Singular Value Decomposition and Its Visualization. *Journal of Computational and Graphical Statistics* **16**(4) 833–854. https://doi.org/10.1198/106186007X256080

Jack Prothero. University of North Carolina Chapel Hill, USA.
E-mail address: jbprothero@gmail.com

Jan Hannig. University of North Carolina Chapel Hill, USA.
E-mail address: jan.hannig@unc.edu

J.S. Marron. University of North Carolina Chapel Hill, USA.
E-mail address: marron@unc.edu