Comparison of dimension reduction methods for the identification of heart-healthy dietary patterns

Natalie C. Gasca, Robyn L. McClelland

Observational Studies, Volume 9, Issue 2, 2023, pp. 123-156 (Article)

Published by University of Pennsylvania Press

DOI: https://doi.org/10.1353/obs.2023.0020

For additional information about this article
https://muse.jhu.edu/article/883480
Comparison of dimension reduction methods for the identification of heart-healthy dietary patterns

Natalie C. Gasca  
ncgasca.uw@gmail.com  
Department of Biostatistics  
University of Washington  
Seattle, WA, 98195, USA

Robyn L. McClelland  
rmccllell@u.washington.edu  
Department of Biostatistics  
University of Washington  
Seattle, WA, 98195, USA

Abstract

Most nutritional epidemiology studies investigating diet-disease trends use unsupervised dimension reduction methods, like principal component regression (PCR) and sparse PCR (SPCR), to create dietary patterns. Supervised methods, such as partial least squares (PLS), sparse PLS (SPLS), and Lasso, offer the possibility of more concisely summarizing the foods most related to a disease. In this study we evaluate these five methods for interpretable reduction of food frequency questionnaire (FFQ) data when analyzing a univariate continuous cardiac-related outcome via a simulation study and data application. We also demonstrate that to control for covariates, various scientific premises require different adjustment approaches when using PLS. To emulate food groups, we generated blocks of normally distributed predictors with varying intra-block covariances; only nine of 24 predictors contributed to the normal response. When block covariances were informed by FFQ data, the only methods that performed variable selection were Lasso and SPLS, which selected two and four irrelevant variables, respectively. SPLS had the lowest prediction error, and both PLS-based methods constructed four patterns, while PCR and SPCR created 24 patterns. These methods were applied to 120 FFQ variables and baseline body mass index (BMI) from the Multi-Ethnic Study of Atherosclerosis, which includes 6814 participants aged 45-84, and we adjusted for age, gender, race/ethnicity, exercise, and total energy intake. From 120 variables, PCR created 17 BMI-related patterns and PLS selected one pattern; SPLS only used five variables to create two patterns. All methods exhibited similar predictive performance. Specifically, SPLS’s first pattern highlighted hamburger and diet soda intake (positive associations with BMI), reflecting a fast food diet. By selecting fewer patterns and foods, SPLS can create interpretable dietary patterns while maintaining predictive ability.

Keywords: Cardiovascular, Dimension reduction, Heart, Nutritional epidemiology, Partial least squares, Supervised methods, Variable selection

©2023 Natalie C. Gasca and Robyn L. McClelland.
1. Introduction

Early investigations of the link between diet and cardiovascular disease (CVD), which has been identified as the leading cause of mortality in American men and women (Heron and Anderson, 2016), focused on individual nutrients or foods. As research transitioned to studying the concurrent effect of foods, results were often discouraging with frequent unintuitive trends due to strong correlations among some food measurements (Hu, 2002). Attention shifted to dimension reduction methods in order to obtain understandable findings from the most relevant aspects of diet.

Figure 1: Classification of select dimension reduction methods used to create diet patterns.

In nutritional epidemiology, a common approach to constructing composite dietary scores is to use previous (a-priori) scientific knowledge to assign “heart-healthiness” to a limited number of foods and nutrients (as categorized in Figure 1, adapted from Schulze and Hoffmann (2006)). A few examples include the Dietary Approaches to Stop Hypertension diet (Svetkey et al., 1999), the Healthy Eating Index (Krebs-Smith et al., 2018), and the Mediterranean Diet (Trichopoulou et al., 2003). However, many of these scores rely on study-specific quantiles to distinguish levels of intake, making it difficult to generalize results outside of the study population.

Another popular way to construct dietary patterns is to apply an unsupervised method on all the food measurements collected in a study. One of the most widespread methods is principal component analysis (PCA), which creates dietary patterns that summarize how people in a study eat overall (Pearson, 1901). When using these patterns as the covariates in a regression model, this method is called principal component regression (PCR). Similarly, cluster (Driver and Kroeber, 1932) and factor analysis (Carroll, 1985) are examples of unsupervised methods. While these methods describe different aspects of food consumption, their patterns are not inherently predictive of CVD as they are not informed by disease data.

To better characterize links between diet and heart disease, we propose utilizing supervised methods, which incorporate disease data directly into the construction of diet summaries. These techniques have the potential to concisely reduce nutritional information in a way that is tailored to heart disease. Among these techniques, reduced-rank regression (RRR) and canonical-correlation analysis have been the most frequently applied in nutritional epidemiology (Schulze and Hoffmann, 2006). These methods are applicable when considering a multivariate response (e.g., multiple biomarkers at once), but we focus on a univariate response in this study. One of the most widespread supervised methods across
other fields is the Lasso (least absolute shrinkage and selection operator) (Tibshirani, 1996), which is an appealing candidate due to its ability to select variables that are useful for prediction. It accomplishes this property by shrinking its regression coefficients towards zero. It was not until recently that this supervised method was applied by Zhang et al. (2018) to explore how well certain foods predicted CVD risk factors.

Another supervised method is partial least squares (PLS) (Wold, 1966), which can create dietary patterns that are correlated with a continuous response while also capturing variation in diet composition. Similar to PCA, these patterns are linear combinations of all food measurements. Although PLS is a standard technique in chemometrics, econometrics and genomics (Mehmood and Ahmed, 2016), it has only recently begun to gain attention in nutritional epidemiology, particularly in the context of comparing its performance to RRR and PCA when analyzing multivariate responses (see Hoffmann et al., 2004; DiBello et al., 2008; Yang et al., 2017).

An appealing extension of PLS is Chun and Keles’s sparse PLS (SPLS), which incorporates variable selection when constructing diet patterns, as we do not presume that every food is inherently related to heart disease. There has been a similar extension to PCA called sparse PCA (SPCA) (Zou et al., 2006), which we include in this study to provide a fairer comparison between methods. However there have been other types of sparsity integrated into PLS and PCA (e.g., Jolliffe et al. (2003); Lê Cao et al. (2008)), so we review their differences to select the most appropriate one for our scientific setting of interest.

When constructing dietary patterns to evaluate heart-healthiness, it is important to adjust for potential confounders and precision variables. Though adjustment is straightforward for standard regression analyses, it is particularly tricky when using (S)PLS on only one set of covariates. While there have been a few proposed adjustment approaches for PLS methods, their properties and modeling implications have not been thoroughly studied. We will demonstrate that there are different ways to adjust for covariates when using PLS, depending on the scientific premises and goals.

Our illustrative example comes from the Multi-Ethnic Study of Atherosclerosis (MESA) (Bild et al., 2002), an observational cohort study which investigates the prevalence and progression of subclinical CVD. At the baseline exam, 6531 participants reported their consumption of 120 types of foods and beverages during a typical week of the previous year in a food frequency questionnaire (FFQ). We want to extract the most biologically-relevant information to characterize a heart-healthy diet that could be protective against CVD. More specifically, we aim to construct “interpretable” dietary patterns that—in descending order of importance—select all of the relevant foods, discard irrelevant foods, capture a few dimensions of diet, and maintain predictive ability of a CVD-related outcome.

This paper first reviews and draws connections between the dimension reduction methods which will be assessed: two PCR-based methods, two PLS-based methods, and Lasso. Techniques and implications of covariate adjustment for each method are then presented in Section 2.2. Simulation studies are conducted in Section 3 to evaluate how the methods perform when blocks of variables (imitating food groups) exhibit varying amounts of covariance within each block. We compare both the predictive ability and scientific interpretability of the constructed patterns. These methods are applied to MESA data in Section 4 to create dietary patterns related to body mass index (BMI), as excess body weight is an important pathway between nutrition and cardiovascular disease (Powell-Wiley et al., 2021). Since
we utilize data from an observational study, we advise that these results are not causal in nature, but rather associations between foods and BMI. Code for this section is available at ncgasca.github.io. In Section 5, we discuss the study findings that SPLS offers advantages compared to PCA-based methods and Lasso, as it can identify more interpretable patterns by using fewer predictors and components, avoid arbitrarily dropping highly correlated variables, and maintain predictive ability. Additional results and proofs can be found in the Appendices.

2. Methodology

The notation convention we use throughout is upper-case bold letters for matrices (\(X\)), lower-case bold letters for vectors (\(y\)), italic letters for scalars (\(c\)), and \(\|w\|_p = \left(\sum_i |w_i|^p\right)^{1/p}\) for the L\(_p\) norm of \(w\). In our setting, we would like to predict a continuous univariate response \(y\) (like BMI) using a function of the original, potentially highly-correlated predictors in matrix \(X_{n \times p}\) (such as diet), where the number of observations \(n\) is greater than the number of predictors \(p\). We assume that each predictor in \(X\) is centered, without loss of generality. In order to create predictive reduced-dimension summaries of \(X\), we consider a class of methods that construct \(K\) mutually orthogonal linear combinations of the covariates, denoted by matrix \(T_{n \times K} = [t_1 \cdots t_K]\), where each column is a linear combination. These combinations are called components or patterns (also known as scores or latent variables, depending on the field). We call this class “component-based methods”, and all of the techniques we compare apart from Lasso fall under this category.

2.1 Review of select dimension reduction methods

Component-based methods construct weighted sums of the predictors in descending order of “importance”. For example, the first PLS component explains the largest amount of covariance between \(X\) and \(y\) that can be constructed by a weighted sum of the predictors, with higher weights assigned to predictors highly correlated with \(y\). After subtracting out the first component’s information from \(X\) and/or \(y\) (to preserve the orthogonality between components), the second PLS component describes the most covariance between the residual predictors and response. This process is repeated for \(K\) components, chosen by the researcher, with \(K \leq \text{rank}(X^TX) \leq p\). If the maximal number of components are constructed, no dimension reduction has occurred, so the PLS solution (and similarly the PCA solution) simplifies to the least squares (LS) estimate from linear regression (Frank and Friedman, 1993). Indeed, Helland (1988) showed that the full set of components form an orthogonal basis for \(X\), so the same information in \(X^TX\) is contained in \(TT^T\).

Sparse methods (i.e., methods intending to reduce the number of non-zero weights or regression coefficients) are useful when not all of the predictors are associated with the response. Indeed, if many irrelevant variables are kept in a component-based model, they could attenuate the estimated coefficients of the relevant variables (Chun and Keleş, 2010). To improve interpretability, sparse techniques have incorporated variable selection by penalizing a quantity related to the regression coefficients \(\hat{\beta}\)—typically the weights in component-based methods.
Dimension reduction to identify heart-healthy patterns

**Principal Component Analysis and Regression** Formally, the PCA model is \( X = TP^T = TKP_K^T + E_K \), where the latter notation emphasizes that some information is left out when using fewer dimensions. As the components are linear combinations of \( X \), we can rewrite this as \( T = XP \). The key is to find appropriate weights \( P \) (called *loadings* in PCA literature) that maximize the variability of \( T \). One approach is to perform a singular value decomposition (SVD) of \( X \) (i.e., \( X = UDV^T \)) and set \( P = V \). Another more popular way to obtain \( P_{p \times K} = [p_1 \cdots p_K] \) is to utilize the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm (Wold, 1966), which can sequentially calculate the most important weights. Here, the optimization problem for the \( k \)th weight is

\[
\hat{p}_k = \text{argmax} \left\{ \langle p^T X^T X p \rangle \right\} \quad \text{s.t.} \quad p^T p = 1 \text{ and } p^T X^T X p_j = 0 \text{ for } j < k.
\]

The first constraint ensures the identifiability of the weights and the second ensures that components are orthogonal to all previously constructed components. This optimization problem is equivalent to maximizing \( \text{Var}(t) = t^T t / n \) with the same constraints.

If one is interested in assessing the relationship between \( X \) and an outcome \( y \), principal component regression simply utilizes a chosen number of PCA components as the new covariates (as \( T = XP \)). The estimated response becomes \( \hat{y} = T(T^T T)^{-1} T^T y \), resulting in PCR regression coefficients \( \hat{\beta} = P(T^T T)^{-1} T^T y \). An advantage of PCR over LS is that it can reduce the dimension of the problem by choosing fewer components than predictors while still preserving most of the variability of \( X \). In this manner, high-dimensional problems \( (n \leq p) \) can become tractable. Even in low-dimensional \( (n > p) \) settings, the PCR estimator is more stable than that of LS when there is multicollinearity (high correlation among some \( X \)). If there is interest in testing associations between \( T \) and \( y \) (a slightly different aim), an advantage of any unsupervised method over a supervised one is that the resulting inference will be valid, since outcome data will not be used multiple times.

**Sparse Principal Component Analysis** Researchers sought to improve the interpretability of PCA components, so Jolliffe et al. (2003) proposed Simplified Component Technique-LASSO (SCoTLASS), adding an \( L_1 \) penalty to the PCA weight constraints. By shrinking the weights towards 0, variables that describe a small amount of variability in each component step can potentially be eliminated. However, SCoTLASS is often not sparse enough (e.g., many non-0 weights), there is no guarantee that the final solution is the global optimum (rather than a local optimum), and it is computationally intensive.

To address these issues, Zou et al. (2006) proposed a Sparse PCA method that generalizes a ridge regression formulation of PCA (to increase speed) and then adds an \( L_1 \) penalty to the PCA weights (to improve interpretability). Their Sparse PCA optimizes

\[
\begin{align*}
\text{argmin}_{C, P} \sum_{k=1}^{K} & \left\{ \|Xc_k - Xp_k\|^2_2 + \lambda\|p_k\|^2_2 + \lambda_k\|p_k\|_1 \right\} \\
\text{s.t.} & \quad C^T C = I_{K \times K}.
\end{align*}
\]

The estimated sparse weights \( \hat{P} \) are obtained by iteratively solving for \( C \) and \( P \) until convergence. Then, the sparse weights are utilized to construct SPCA components, which are used as new predictors for sparse PCR (SPCR). While SPCA allows for separate tuning parameters \( (\lambda_k) \) for each component’s weight, we simplify this to a single parameter for a fairer comparison with the other sparse methods and reduced computation time.
Though both sparse PCA methods improve interpretation because each component is formed by fewer variables (i.e., component-wise sparsity), they do not guarantee variable selection (i.e., overall sparsity) unless hardly any components are selected with each being composed of only a few variables. We posit that overall sparsity is a more beneficial aim than component-wise sparsity if it is likely that many predictors are irrelevant, such as from a food frequency questionnaire.

**Partial Least Squares** PLS is similar to PCA in that it decomposes $X$ into components, but these are constructed to explain the most (squared) covariance between $X$ and $y$. Formally, the univariate PLS model is

$$X = TP^T + \mathcal{E}_x \quad \text{and} \quad y = Tq^T + \epsilon_y,$$

where $\mathcal{E}_x, \epsilon_y \perp T$.

There is debate as to how to interpret these models from a statistical standpoint (Wold et al., 2009), as some view $\mathcal{E}_x$ as the residual $X$ information after reducing its dimension. However, the “latent variable” perspective posits that $T$ represents the true phenomena of interest, in which case $\mathcal{E}_x$ represents purely random noise.

Regardless of interpretation, the components are a set of mutually orthogonal linear combinations of $X$, which can be expressed as $T = XR = X(W(P^T W)^{-1})$, and consequently $\beta = R (T^T T)^{-1} T^T y = R q^T$. Again, one must find the appropriate weights to obtain $T$—either Gram-Schmidt-orthogonalized $W$ from the NIPALS algorithm or oblique projection $R$ from the Statistically Inspired Modification of PLS algorithm (De Jong, 1993). For univariate $y$, the sequential algorithms are equivalent, with the $k$th weight solving

$$r_k = \arg\max_r \{ r^T X^T y y^T X r \} \quad \text{s.t.} \quad r^T r = 1 \text{ and } r^T X^T X r_j = 0 \text{ for } j < k. $$

This optimization problem is equivalent to maximizing $\text{Cov}^2 (y, t)$, or more specifically $\text{Cor}^2 (y, t) \text{ Var}(t)$, with the same constraints (Rosipal and Krämer, 2006).

The PLS estimator with $K$ components has a few possible interpretations, as

$$\hat{\beta}^{(K)} = R_{p \times K} (T_{n \times K}^T T_{n \times K})^{-1} T_{n \times K}^T y_{n \times 1} = \sum_{k=1}^{K} s_k d_k = \arg\min_{\beta \in \mathcal{K}_K} \| y - X \beta \|_2^2.$$

The first form demonstrates that the PLS regression coefficient (on the $X$ scale) is the projection of the component-scale coefficient $q^T$. The second notation indicates that the PLS estimator is equivalent to the conjugate gradient estimate after $K$ steps (Phatak and de Hoog, 2002). Specifically, $s_k$ is an optimized step size and $d_k$ is the conjugated residual covariance. Finally, these estimates are restricted to a particular subspace, which provides a convenient way to connect the objective of PLS with that of typical regression models; the $K$-th order Krylov subspace $\mathcal{K}_K(S, s)$ is defined as $\text{span}\{ s, S s, S^2 s, \ldots, S^{K-1} s \}$, where we define $s = X^T y$ and $S = X^T X$. Each of these forms are a useful way to consider how PLS creates a reduced-dimension (i.e., shrunken) summary of the LS estimator.

Although the PLS estimator’s magnitude increases with each component included, it is important to note that some coefficients may be attenuated from one step to another. Additionally, the degrees of freedom are not as easily defined when using PLS models; the information used is usually greater than $K$ but less than $p$ (Krämer and Sugiyama, 2011).
Sparse Partial Least Squares  To obtain more interpretable PLS components, Lé Cao et al. (2008) proposed a Sparse PLS method that adds an \( L_1 \) penalty to the PLS weight constraints. However, similar to SCoTLASS and SPCA, inducing component-wise sparsity does not guarantee variable selection. Additionally, by shrinking the PLS weights, they are no longer orthogonal. Consequently, this Sparse PLS estimator is not guaranteed to equal the LS estimator when the maximal number of components are constructed.

To remedy this, Chun and Keleš (2010) proposed a two-stage Sparse PLS procedure that directly induces variable selection while preserving orthogonal weights. At each component step, the variables that have the strongest covariance with the (residual) response are added to the index set \( A \). The subset of variables \( X_A \) is chosen based on a given sparsity percentage \( \eta \in [0, 1) \) and \( K \) components. Initializing \( \hat{\beta}^{(0)} = 0 \) and index set \( A_0 = \{\} \), the \( k \)th set is

\[
A_k = A_{k-1} \cup \left\{ j : \left| X_j^T (y - X_{A_{k-1}} \hat{\beta}_{A_{k-1}}^{(k-1)}) \right| > \eta \max_{1 \leq h \leq p} \left| X_h^T (y - X_{A_{k-1}} \hat{\beta}_{A_{k-1}}^{(k-1)}) \right| \right\},
\]

where \( \hat{\beta}_{A_{k-1}}^{(k-1)} \) is the coefficient from a \( (k - 1) \)-component PLS model using the covariates \( X_{A_{k-1}} \). The final selection of variables \( A = A_K \) includes those covariates that have the strongest relationship with the response or that best capture the variation among covariates when transformed into \( K \) independent linear combinations. This procedure then fits a \( K \)-component PLS model with the final selected variables. As such, this sparse PLS estimate is denoted by \( \hat{\beta}^{(K)}_A \) and simplifies to PLS if \( \eta = 0 \).

SPLS directly selects the variables that are most related to \( y \)—even when fitting more than one component—as long as the sparsity level (\( \eta \)) and signal-to-noise ratio are high. In comparison, the initial Sparse PLS method may add in variables with weak relationships to \( y \) as early as the second component (if the first component has included most of the information from the strongly related variables). Additionally, by fitting PLS on a reduced set of variables, Chun and Keleš’s weights are still orthogonal, so the SPLS estimator with the maximal number of components is equivalent to the LS estimator with the same set of reduced variables. However, this method no longer promotes component-wise sparsity, since all of the selected variables are utilized to construct each component.

Lasso  The Lasso is a penalized linear regression method that directly induces variable selection by shrinking \( \beta \) towards \( 0 \) using an \( L_1 \) penalty. Its estimate takes the form \( \hat{\beta}_{Lasso} = \min_\beta \{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \} \). If the selected sparsity level \( \lambda \geq 0 \) is large enough, smaller coefficients will be set to 0, resulting in variable selection. While the Lasso is optimized for predictive accuracy, Zou and Hastie (2005) pointed out that it struggles with highly-correlated covariates, as it will select only one correlated variable (regardless of etiology). Although the Lasso does not create components, we say that it constructs one pattern \( (X\hat{\beta}_{Lasso}) \) for comparability.

2.2 Adjustment for low-dimensional covariates

We now present adjustment approaches for these dimension-reduction methods and provide a connection between the least squares and PLS estimators. To better characterize the relative contributions of \( X \) and pre-specified low-dimensional adjustment variables \( L \), we
only consider joint models. Rather than fitting a model on \( L \) alone and then relating \( X \) to the subsequent residuals, we restrict to techniques that model \( X \) while accounting for \( L \) to obtain a final adjusted estimate.

Formally, we consider two models for the centered outcome. The first presumes that \( X \) is inherently related to \( y \): \( y = X\beta + L\gamma + \epsilon \). If this is the scientific premise, PLS would simply summarize \( X \) using fewer dimensions (i.e., covariate-level summary). In the context of nutritional epidemiology, this approach would help if we wanted to characterize diet using individual foods as the unit of interest.

The second model posits that components \( T \) are truly related to \( y \): \( y = Tq^T + L\gamma + \epsilon \). In this case, PLS would construct the latent variables of interest (i.e., component-level summary). For example, this would be our aim in nutritional epidemiology if we wanted to characterize diet using dietary patterns—emphasizing the interpretation of each composite pattern over individual foods. The distinction between these scientific premises influences how much of the residual outcome to use when fitting PLS, which affects the resulting coefficient estimates (as PLS is a supervised method).

We introduce a few more notation conventions. Let \( L = LX^T \) be the orthogonal projection of \( L \), and similarly define \( X \) and \( T \). We define “orthogonalizing” as the process of projecting a block of covariates off of the column space of the other set of covariates (e.g., orthogonalizing \( X \) means \( (I - L)X \)). For the upcoming remarks, we use the convention \texttt{model.fit[ outcome ~ covariates ]}. The symbol \( \rightarrow \) signifies that we construct the component \( T = XR \) utilizing the corresponding weights from the indicated PLS fit.

Adjustment for LS Assuming that \( X^TX \) and \( L^TL \) are invertible, we show in Appendix A that the coefficient estimates from a multiple linear regression are

\[
\begin{bmatrix}
\hat{\beta} \\
\hat{\gamma}
\end{bmatrix} = \begin{bmatrix}
X^T(I - L)X^{-1}X^T(I - L)y \\
L^T(I - X)L^{-1}L^T(I - X)y
\end{bmatrix} = \begin{bmatrix}
X^T\gamma + X^T(\epsilon - L\gamma) \\
L^T\gamma + L^T(\epsilon - X\beta)
\end{bmatrix}.
\]

These LS coefficient estimates can also be recovered via separate regression fits, as outlined in Remark 1 and proven in Appendix A. Of note, the orthogonalized fits allow for a convenient partition between the sets of covariates, whereas the residual fits highlight a typical regression interpretation given that the opposite adjusted coefficient is known.

\textbf{Remark 1} Given an outcome \( y \), adjustment via a multiple least squares regression can be accomplished by any of the following ways:

1. \textit{Concurrent fit:} \( \begin{bmatrix}
\hat{\beta} \\
\hat{\gamma}
\end{bmatrix} = \text{LS}[y \sim X + L] = \begin{bmatrix}
X^T(I - L)X^{-1}X^T(I - L)y \\
L^T(I - X)L^{-1}L^T(I - X)y
\end{bmatrix} \)

2. \textit{Orthogonalized fits:} \( \begin{bmatrix}
\hat{\beta} \\
\hat{\gamma}
\end{bmatrix} = \text{LS}[y \sim (I - L)X] = \begin{bmatrix}
X^T(I - L)X^{-1}X^T(I - L)y \\
L^T(I - X)L^{-1}L^T(I - X)y
\end{bmatrix} \)

3. \textit{Residual fits:} \( \begin{bmatrix}
\hat{\beta} \\
\hat{\gamma}
\end{bmatrix} = \text{LS}[y - L\hat{\gamma} \sim X] = (X^TX)^{-1}X^T(y - L\hat{\gamma}) \;
\begin{bmatrix}
\hat{\gamma} = \text{LS}[y - X\hat{\beta} \sim L] = (L^TL)^{-1}L^T(y - X\hat{\beta}).
\end{bmatrix} \)

\textbf{Covariate-summary adjustment for (S)PLS} If we opt to use PLS simply as a tool to create a lower-dimensional summary of \( X \), we would begin with a multiple linear regression using \( X \) and \( L \), as above. To construct a shrinkage summary of the adjusted
relationship between \( y \) and \( X \), one would use (S)PLS to model \( X \) with respect to \( y - L\hat{\gamma} \). One could also subsequently refit the \( \gamma \) estimate using the constructed (S)PLS components in order to explain as much of the outcome as possible in a joint model.

**Component-summary adjustment for (S)PLS** If one instead assumes that PLS components are the true scale, the goal would be to evaluate a model with \( T \) and \( L \). This implies that the (S)PLS components and weights should be constructed using \((I - L)y\) in order to obtain components that are adjusted for \( L \). The resulting estimators take the form

\[
\begin{align*}
\hat{q}^T &= [T^T(I - L)T]^{-1} T^T(I - L)y \\
\hat{\gamma} &= [L^T(I - T)L]^{-1} L^T(I - T)y
\end{align*}
\]

As before, these coefficient estimates can be recovered in various ways outlined in Remark 2 and proven in Appendix A. Since SPLS can be viewed as applying PLS to a reduced set of variables, we only describe adjustment in the context of PLS for clarity.

**Remark 2** Given an outcome \( y \), adjustment via a multiple least squares regression with PLS components \( T \) can be accomplished by any of the following ways:

1. **Concurrent fit**: PLS\([y \sim X] \rightarrow T \)
   \[
   \begin{align*}
   \hat{q}^T &= \text{LS}[y \sim T + L] = [T^T(I - L)T]^{-1} T^T(I - L)y \\
   \hat{\gamma} &= \text{LS}[y \sim (I - T)L] = [L^T(I - T)L]^{-1} L^T(I - T)y
   \end{align*}
   \]

2. **Orthogonalized fits**: \( \hat{q}^T = \begin{cases} \text{PLS}[y \sim (I - L)X] \rightarrow T \\ T^T(I - L)T]^{-1} T^T(I - L)y \end{cases} \)
   \[
   \hat{\gamma} = \text{LS}[y \sim (I - T)L] = [L^T(I - T)L]^{-1} L^T(I - T)y
   \]

3. **Residual fits**: PLS\([y \sim X] \rightarrow T \)
   \[
   \begin{align*}
   \hat{q}^T &= \text{LS}[y \sim L\hat{\gamma} \sim T] = (T^T T)^{-1} T^T (y - L\hat{\gamma}) \\
   \hat{\gamma} &= \text{LS}[y \sim T\hat{q}^T \sim L] = (L^T L)^{-1} L^T (y - T\hat{q}^T)
   \end{align*}
   \]

Although it is less ideal to require a separate PLS step for the first and third approaches, it is necessary to obtain PLS weights (and thus components) that are maximizing the correct value. Alternatively, the orthogonalized fits provide a computationally expedient way to obtain the adjusted coefficient estimates, regardless of number of components used. Compared to the covariate-summary version, the component-summary \( \hat{\gamma} \) explains more of the outcome (as \( T \) is inherently smaller in dimension than \( X \)). Also, any \( X \) variable that is correlated with \( L \) will have a smaller magnitude estimator when doing a component-summary, as it orthogonalizes \( X \). Jørgensen et al. (2004) proposed an iterative PLS adjustment scheme based on the concurrent fit framework. It reduces to the component-summary adjustment if one orthogonalizes \( X \) (suggested by Bazzoli and Lambert-Lacroix (2018)).

**Adjustment for (S)PCR** Unlike PLS, PCA is not a supervised method. Orthogonalizing \( X \) during the PCA component construction will result in a different estimate, so the equivalence between the concurrent and orthogonalized fits in Remark 2 does not hold for PCR. To adjust for \( L \), we create (S)PCA components as usual and then include them in a multiple linear regression model with the adjustment variables.

**Adjustment for Lasso** The Lasso is easily equipped to adjust for other variables by including them without any penalization. Its adjusted estimates are expressed as

\[
\begin{align*}
\hat{\beta}_{\text{Lasso}} & = \min_{\beta, \gamma} \{ \|y - X\beta - L\gamma\|^2_2 + \lambda\|\beta\|_1 \}
\end{align*}
\]
2.3 Selection of tuning parameters

These dimension reduction methods require the specification of a sparsity level and/or number of components. We used 10-fold cross-validation (CV) to assess a potential model’s predictive accuracy, with the root mean squared error of prediction (RMSE) as the criterion. To select an accurate yet parsimonious model (e.g., with fewer components or more sparsity) from various candidate models, we employed the one standard error (1-se) test, as described by Hastie et al. (2009) for the Lasso. After identifying the “reference” model with the smallest CV-RMSE, the most parsimonious model whose CV-RMSE is within one standard error of the reference model’s RMSE is selected as the final model. Since SPCR and SPLS have two tuning parameters, we extended the 1-se test to optimize the parameter search across a grid of candidate pairs, favoring a decrease in number of components over an increase in sparsity. This two-dimensional 1-se test is described further in Appendix B.

2.4 Computation

The software environment R version 3.6.1 (Team et al., 2013) was used throughout the investigation process. PCR and PLS were run with the \texttt{pls} package (Mevik et al., 2011), SPCR was run with the \texttt{elasticnet} package (Zou and Hastie, 2012), SPLS was implemented using functions in the \texttt{spls} package (Chung et al., 2013), and Lasso was run with the \texttt{glmnet} package (Friedman et al., 2009). To conduct the 1-se test for SPCR and SPLS, we developed code that called functions from the corresponding R packages referenced above (i.e., wrappers). For component-based methods, we followed the data pre-processing convention in \texttt{pls} of scaling each predictor by its standard deviation and then centering it, as well as centering the response $y$. For the Lasso, we only scaled each predictor by its standard deviation. When possible, we applied the same cross-validation folds for all methods, with the intent to establish fair and equal comparisons.

3. Simulation studies

To evaluate the performance of various methods in controlled scenarios, we generate (i.e., simulate) pseudo-data, thereby knowing which variables truly influence the outcome. This allows us to compare each method’s variable selection and prediction results with the underlying truth, enabling us to better assess each model’s performance than if we were to test these methods on real data (as it is tricky to discern between true signal and noise).

3.1 Simulation design

As our study is motivated by nutritional epidemiology, we generate predictors that mimic the features of FFQ data. In particular, it is natural to classify foods into food groups, so we create blocks of variables that are more similar to each other. Such “multiblock data” occur in many other scientific domains, since blocks can also represent different types of experiments or different sample collection sites from the same subject (see e.g., Ingerslev et al. (2015); Karaman et al. (2015)). We have designed the simulation study with blocks of differing size and structure, as commonly found in practice.

We generate a multivariate-normal predictor $X$ with 1000 observations and 24 variables, only nine of which contribute to the response $y$. The predictors are grouped into four blocks;
Block 1 has two variables (one is relevant), Block 2 has six variables (two are relevant), Block 3 has nine variables (six are relevant), and Block 4 has seven variables (none are relevant). Each predictor has mean zero and unit variance, with the variability representing different relative levels of food consumption in the context of nutritional epidemiology.

The first setting we consider (Case 1) is when all the variables are independent of each other, which is the simplest scenario since relevant and irrelevant variables are more easily distinguishable. As the intake of foods within a food group could be related in practice, we also vary the covariance between variables of the same block in further settings. We first test scenarios of “matching-covariance block structure”, where all blocks have the same level of intra-block covariance ranging from 0.3 to 1 (see Appendix C).

Next, we consider “mixed-covariance block structure”, where blocks can have different intra-block covariance, using the MESA diet data at baseline as a guide for reasonable correlations in food frequency questionnaires. Thus, the variance of $\mathbf{X}$ has a block diagonal structure, with independence between blocks for simplicity. In Case 2, the intra-block covariances are 0.05, 0.50, 0.30, and 0.10, respectively. In this setting, we would expect that it would be most difficult to distinguish the relevant variables from the irrelevant ones in Block 2. Further mixed-covariance block structures were also tested (see Appendix C).

We assign the nine relevant variables to have unit positive or negative associations with the response as follows: the Block 1 and three Block 3 relevant variables are positive while the remaining are negative. The generated response is univariate normal with unit variance and its mean being the sum of the relevant variables (multiplied by the true association). In the context of nutritional epidemiology, this variability represents the contribution of factors other than diet to any response. We also compare how the SPLS adjustment techniques perform in an expanded simulation setting (see Appendix C).

For each of 100 generated data sets, we split our observations into a training (n=700) and test (n=300) set, and we use 10-fold cross validation on the training data to select the number of components and/or sparsity level for each method. Due to computational burden, we only perform 5-fold CV for SPCR. Predictions are then made on the test data. To compare model interpretability, we report the average number of selected variables, number of selected patterns, and percentage of $y$’s variance explained ($R^2$). To evaluate predictive performance, we report the average RMSE from the test sets.

### 3.2 Simulation Results

Table 1(a) summarizes the performance of the five methods on data from Case 1, when variables are independent. Among the sparse methods, SPLS and Lasso select all relevant variables and up to one irrelevant variable, on average. While SPCR imposes sparsity in the patterns (selecting one or two variables per pattern), it keeps almost all of the predictors because of the numerous patterns selected in order to have competitive predictive ability. When it does discard variables (in 47 data sets), SPCR tends to discard at least one relevant variable. All methods except for SPCR have reasonable predictive ability, with SPLS having the lowest RMSE, on average. Among the component-based methods, PLS and SPLS select the fewest patterns (two). PLS’s first pattern highlights most of the relevant variables, while PCR describes mostly irrelevant variables at first. Case 1 is the setting in which the methods capture the most information about the response, with most $R^2$ values between 0.89–0.90.
Table 1: Simulation results across 100 data sets, with mean and (standard deviation).

(a) Case 1: Independent variables

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of Variables</th>
<th>Number of Patterns</th>
<th>RMSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCR</td>
<td>24.0 (0.0)</td>
<td>23.8 (0.5)</td>
<td>1.026 (0.040)</td>
<td>0.901 (0.007)</td>
</tr>
<tr>
<td>SPCR</td>
<td>23.1 (1.4)</td>
<td>23.0 (1.4)</td>
<td>1.140 (0.211)</td>
<td>0.875 (0.053)</td>
</tr>
<tr>
<td>PLS</td>
<td>24.0 (0.0)</td>
<td>2.0 (0.0)</td>
<td>1.027 (0.040)</td>
<td>0.901 (0.008)</td>
</tr>
<tr>
<td>SPLS</td>
<td>9.1 (0.6)</td>
<td>2.0 (0.4)</td>
<td>1.013 (0.038)</td>
<td>0.899 (0.008)</td>
</tr>
<tr>
<td>Lasso</td>
<td>9.8 (1.0)</td>
<td>1.0 (0.0)</td>
<td>1.043 (0.039)</td>
<td>0.893 (0.008)</td>
</tr>
</tbody>
</table>

(b) Case 2: Mixed-covariance block structure

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of Variables</th>
<th>Number of Patterns</th>
<th>RMSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCR</td>
<td>24.0 (0.0)</td>
<td>23.4 (0.8)</td>
<td>1.034 (0.041)</td>
<td>0.893 (0.008)</td>
</tr>
<tr>
<td>SPCR</td>
<td>24.0 (0.0)</td>
<td>23.8 (0.5)</td>
<td>1.032 (0.044)</td>
<td>0.893 (0.008)</td>
</tr>
<tr>
<td>PLS</td>
<td>24.0 (0.0)</td>
<td>3.5 (0.5)</td>
<td>1.036 (0.042)</td>
<td>0.893 (0.008)</td>
</tr>
<tr>
<td>SPLS</td>
<td>13.2 (0.7)</td>
<td>3.7 (0.7)</td>
<td>1.022 (0.041)</td>
<td>0.890 (0.008)</td>
</tr>
<tr>
<td>Lasso</td>
<td>10.7 (1.3)</td>
<td>1.0 (0.0)</td>
<td>1.038 (0.040)</td>
<td>0.885 (0.009)</td>
</tr>
</tbody>
</table>

RMSE=root mean squared error. †† All methods selected the nine relevant variables, except for SPCR in Case 1, which discarded one to two relevant variables in 23 data sets. †† Case 2 included intra-block covariances of 0.05, 0.50, 0.30, and 0.10.

As more covariance is introduced in the matching-covariance block structure, the relevant and irrelevant variables become harder to distinguish in the first three blocks (see Appendix C). With 0.3 intra-block covariance, variables in each block behave more similarly, inducing a noticeable association between the response and the five irrelevant variables in Blocks 1 and 2. SPLS incorrectly selects four of these irrelevant variables, while Lasso only selects 1.5 of them incorrectly on average. SPLS still maintains its predictive advantage (RMSE=1.019), with Lasso and PLS performing comparably (RMSE=1.043). It is not until 0.8 intra-block covariance that SPLS provides little improvement in sparsity, as it selects 21 variables compared to Lasso’s 12 selected variables.

Table 1(b) shows the method performance for Case 2, a mixed-covariance block structure, which is more realistic in practice. As Block 2 has the highest covariance, its four irrelevant variables behave more similarly to its relevant variables, inducing an association with $y$. SPLS selects these four irrelevant variables, Lasso only selects 1.7 irrelevant variables, and SPCR includes all variables (selecting only one or two variables per pattern). All methods have similar predictive accuracy (RMSEs between 1.032–1.038) except for SPLS, which results in the lowest RMSE (1.022). PLS and SPLS again create the fewest patterns (3.6 on average). Both PLS and PCR highlight Block 2 in their first component, while SPLS additionally emphasizes the other relevant variables. PCR’s second component picks Block 3 variables (the next most correlated block), whereas PLS’s second component emphasizes Block 2 variables alongside other relevant variables. In Case 2, all methods achieved an
Dimension reduction to identify heart-healthy patterns

$R^2$ value of roughly 0.89. In the other tested mixed-covariance block structures, we note that the supervised methods perform better when blocks with more relevant variables have lower intra-block covariances.

Across all the tested simulation scenarios, SPLS consistently predicts the best, on average. When predictors are independent (Case 1), SPLS outperforms Lasso in variable selection by discarding more irrelevant variables. Once there is moderate matching covariance within blocks, SPLS selects about 3–5 more irrelevant variables than Lasso. When the covariance structure changes by block, SPLS selects up to three more irrelevant variables than Lasso. Except for settings with high (0.8–1) matching covariance within blocks, SPLS always selects the relevant variables. SPCR imposes sparsity on the number of variables used in each component, rather than on the actual number of variables selected. Thus, SPCR at most drops one variable.

In the simplest settings, SPLS selects 2.0 components, and in the trickiest (high matching covariance) it selects 6.7 components, on average. In all settings except perfect correlation within blocks, SPLS selects slightly more components than PLS, as it is using far fewer variables to describe the response. Meanwhile, PCR selects 22–24 patterns for all cases except perfect correlation and the simplest mixed covariance by block structure. SPCR selects a similar number of patterns as PCR (typically 23–24). Most methods explained a similar amount of $y$ variability; PCR described the most and the Lasso usually described the least (but only differing by 0.01). Case 1 resulted in the highest $R^2$ values (0.89–0.90), and perfect correlation across all blocks resulted in the lowest values (0.82–0.83).

When comparing SPLS adjustment techniques (see Appendix C), the covariate-summary adjustment resulted in better variable selection, though the component-summary was generally similar. They were most sensitive to the variability level of the noise added to the generated response, followed by the presence or not of a confounder—which impacted PLS estimator accuracy. Both joint adjustment techniques had better PLS estimation accuracy than the two-stage residual adjustment technique (as measured by RMSE).

4. Data application

4.1 Scientific background and study design

In our motivating illustration, our goal is to identify understandable biologically-driven diet patterns related to cardiovascular disease. As excess body weight is an influential pathway between diet and CVD, we utilize body mass index as a continuous heart-health related response to identify relevant diet patterns. In the Multi-Ethnic Study of Atherosclerosis (Bild et al., 2002), 6499 participants (95% of the cohort) had complete information on demographics, diet, BMI, and intentional exercise. These participants were aged 45–84 (53% women), free of clinical CVD at baseline (2000-2002), and from six United States communities (New York City, Baltimore, Chicago, Los Angeles, the Twin Cities, and Winston Salem). They self-identified as 39% white, 27% Black, 22% Hispanic, and 12% Chinese.

The MESA Block-inspired FFQ was modified from the Insulin Resistance Atherosclerosis Study’s FFQ (Mayer-Davis et al., 1999), which was validated in their multi-ethnic cohort, and it includes both ethnic and regional foods. In MESA’s FFQ, participants reported the frequency and serving size of 120 foods and beverages that they consumed during a typical week of the previous year. There were nine frequency options, ranging from “rare
or never” and “once per month” to “multiple times per day”, and three serving size options (small, medium, or large compared to others of the same gender and age). The calculated servings per day of each food item were used as our predictors, after standardizing them. By adapting the American MyPlate guidelines (Committee et al., 2015a), we categorized the diet variables into eight food groups: fruits, vegetables, grains, proteins, mixed entrees, dairy, sweets/oils, and beverages. See Appendix D for the categorized list of 120 food items.

There are also many factors that could potentially confound the association between BMI and diet (such as age, gender, and race/ethnicity), so we included these in all models. Exercise is also important to consider as it is related to BMI independently of diet. In the MESA Typical Week Physical Activity Survey, adapted from the Cross-Cultural Activity Participation Study (Ainsworth et al., 1999), participants reported their time spent and frequency of various physical activities during the previous month. We used the log of the metabolic equivalent minutes per week of conditioning activities such as walking for exercise, dancing, and sports in our analysis.

Studying the effect of diet on a relative scale has been established by nutritional epidemiologists as a useful type of analysis (Willett et al., 1997; Freedman et al., 1997), in part because relative diet composition is more feasible to change than absolute levels of food consumption—which is influenced by other unalterable factors. This can be accomplished by adjusting for total energy (caloric) intake as a potential confounder, which we performed. Lastly, we opted for a component-summary adjustment when applying (S)PLS.

4.2 MESA Results

Average servings per day tended to be quite low, since many people did not consume some of the foods. The most popular food was coffee (1.13 servings), followed by diet soda (0.42 servings). There were 21 foods that more than 25% of the MESA participants consumed, such as apples and bananas, and 20 foods that less than 25% of the cohort consumed. BMI ranged from 15 to 62 (median of 27.5) for those with complete data. It has been suggested that a recommended normal BMI range for seniors is [23, 31), and 60% of the MESA cohort were in this category (Winter et al., 2014). When modeling with the adjustment variables alone, they explained 15% of the variation in BMI. Foods and beverages were only mildly correlated with BMI residuals, with correlations ranging from -0.07 to 0.11.

Table 2 summarizes the adjusted BMI-related dietary patterns in MESA created by the five methods. We describe these results in more depth subsequently. Foods are listed by the magnitude of their regression coefficients; foods positively associated with BMI are indicated by (+), and those negatively associated with BMI are indicated by (-).

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of Variables</th>
<th>Number of Patterns</th>
<th>CV-RMSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCR</td>
<td>120</td>
<td>17</td>
<td>4.92</td>
<td>0.177</td>
</tr>
<tr>
<td>SPCR</td>
<td>116</td>
<td>67</td>
<td>4.86</td>
<td>0.197</td>
</tr>
<tr>
<td>PLS</td>
<td>120</td>
<td>1</td>
<td>4.88</td>
<td>0.190</td>
</tr>
<tr>
<td>SPLS</td>
<td>5</td>
<td>2</td>
<td>4.91</td>
<td>0.179</td>
</tr>
<tr>
<td>Lasso</td>
<td>11</td>
<td>1</td>
<td>4.94</td>
<td>0.175</td>
</tr>
</tbody>
</table>

CV-RMSE = cross-validated root mean squared error.
Using the 120 food items, PCR constructed 17 patterns, which had a CV-RMSE of 4.92 and described 17.7% of the BMI variability. The foods with the largest magnitude associations were hamburgers (+), steak (+), fries (+), and diet soda (+). Sparse PCR achieved the best CV-RMSE and $R^2$ values, but it required 67 patterns to do so. As expected, SPCR created sparse patterns (each pattern being constructed from one or a handful of foods) which allows for easier scientific interpretation, but it only discarded four foods. The first three patterns described the consumption of Western/fast foods, stir fried Asian meals, and Hispanic meals, respectively. The most important foods overall according to SPCR were hamburgers (+), diet soda (+), light green lettuce (+), and fries (+).

PLS selected one diet pattern using all of the foods (CV-RMSE=4.88, $R^2=0.190$). The most relevant ones were hamburgers (+), diet soda (+), natural sweeteners added to coffee or tea (-), and light green lettuce (+). SPLS constructed two diet patterns with only five foods. With this reduction in covariates, it still had comparable predictive accuracy and $R^2$ to the other methods (CV-RMSE=4.91, $R^2=0.179$), even outperforming PCR. The first diet pattern emphasized hamburger and diet soda intake, typical fast food staples. The second diet pattern again highlighted hamburger consumption, but it also focused on natural sweeteners added to coffee and tea. The most relevant foods overall with SPLS were hamburgers (+), diet soda (+), light green lettuce (+), natural sweeteners (-), and wine (-).

Figure 2: SPLS’s first BMI-related diet pattern in MESA, ordered by strength of relationship with BMI.
as to be analogous to plotting PCA loadings. This novel graphic can provide an added perspective about how each food group contributes to the creation of (S)PLS components, lending insight into this heart-health related diet pattern.

The Lasso selected 11 foods, the most important being hamburgers (+), light green lettuce (+), diet soda (+), and sweet potatoes (-). Although the Lasso selected the same five foods that SPLS did along with six others, its model performance was the worst (CV-RMSE=4.94, $R^2=0.175$). Interestingly, most of the foods selected by Lasso but discarded by SPLS tended to be highly correlated with total energy intake.

Figure 3 displays the adjusted associations between foods and BMI for the five selected models. Inherently, PCR and PLS (solid and dashed lines) model all 120 foods, but the inclusion of potentially many irrelevant foods might be contributing to the attenuated coefficients. SPCR (filled dots) selects most of the foods, but it does give more magnitude to some of the same foods highlighted by SPLS (open dots) and Lasso (diagonal crosses). Lasso’s regression coefficients vary the most, with the hamburger coefficient being five times larger than any other method’s coefficients. Though these relationships are specific to BMI, they can also be helpful to consider the potential trends between diet and heart disease.

Figure 3: Adjusted associations between foods and BMI in MESA, ordered by food groups.

Notes: Dot symbols are used for sparse methods, whereas lines are used for non-sparse methods. Food labels are included for those foods that had a magnitude of at least 0.25 (using any method), placed near its largest coefficient when possible. For sparse methods, regression coefficients ($\hat{\beta}$) can be interpreted as the expected change in BMI when substituting one unit of a selected food’s intake with the intake of foods not selected by that model, keeping adjustment variable and other selected food levels constant. See Appendix D for the list of 120 food items.
5. Discussion

When the scientific goal is to construct patterns that reflect the underlying relationship between correlated predictors and a continuous outcome, PLS-based methods have useful properties to consider. If variable selection is not a priority, PLS is a good alternative to PCR since it requires fewer components to describe a similar amount of variability in the outcome, demonstrated across all simulation settings and the data application. If not all predictors are believed to be relevant, we suggest that variable selection is a more advantageous aim than component-wise sparsity. As was illustrated in our data application, SPLS can create more interpretable patterns that are reflective of a specific disease process than PCR, SPCR, or PLS by using fewer predictors and components without much loss of predictive ability. The Lasso also excelled at discarding covariates, but it can arbitrarily drop some variables if they are highly correlated.

We were surprised by the simulation performance of certain methods. SPCR in particular had trouble with variable selection, convergence, and computational burden. Though SPCR created the most parsimonious patterns (selecting 1-2 variables per pattern), it required almost all of the variables and patterns to achieve reasonable predictions. Next, PCR often selected 24 components, which is equivalent to a linear regression. These behaviors are likely due to the fact that predictive ability was the criterion for component selection, as prediction is part of the study goal. If we instead wanted to explain the most covariate variability, PCA-based methods would have selected far fewer patterns. There has been recent work (Jones et al., 2020) showing that the first PCA components tend to be more correlated with \( y \) than later components under certain conditions, if one accepts working with transformations of the predictors that cannot be interpreted on the original X scale. However, we have shown that for our scientific scope, PLS is more reliable at creating initial components that are correlated with the outcome than PCA. Last, we expected the Lasso to outperform component-based methods in out-of-sample predictive ability (RMSE column in Table 1), but most methods had similar predictive performance. Perhaps Lasso’s advantage in prediction would become apparent if the signal-to-noise ratio were smaller, when PLS methods can behave poorly (Stoica and Söderström, 1998), or if we constructed covariates with differing magnitudes of true associations, which Lasso can more easily model (Friedman and Popescu, 2003).

Fundamental to our study is the notion of “interpretability”. One way of constructing understandable patterns is capturing variations in diet composition. As shown in the data analysis, PCA-based models excel at this; SPCR’s first patterns highlighted the consumption of Western/fast foods, stir fried Asian meals, and Hispanic meals. However, to form patterns that can help us better understand the scientific mechanisms that relate a large set of correlated covariates to a disease or risk factor, other approaches should be considered.

We proposed that an interpretable model should (in descending order of priority) select all of the relevant variables, discard irrelevant variables, capture a few dimensions of the scientific relationship, and maintain predictive ability of the outcome. The Lasso accomplishes most of these goals, except that its variable selection is solely focused on how well a subset of variables can predict an outcome, making it less appealing for settings of moderately correlated food measurements. Additionally, many nutritional epidemiologists have favored reducing the dimension of diet variables into scores or dietary patterns when the goal is to
interpret complex diet-disease associations. These prompted us to investigate PLS-based methods, and we conclude that SPLS is a strong contender that can achieve these goals, particularly when the signal-to-noise ratio is sufficiently large. However, interpreting SPLS patterns still requires scientific expertise, as they are combinations of the predictors.

Given the advantages of PLS, we investigated the modeling implications of different proposed techniques when adjusting for low-dimensional pre-specified covariates. The scientific premise of whether $X$ or $T$ is the underlying predictor of interest dictates how to perform covariate adjustment. We also presented three ways to obtain adjusted coefficients from a multiple least squares regression model, and the orthogonalized fits outlined a computationally expedient regardless of whether PLS is used. These connections have not been defined previously for PLS or presented as clearly for LS.

Our data application illustrates how SPLS can summarize scientifically-interpretable relationships. Dietary pattern research has found that consuming more vegetables, fruits, whole grains, low-fat dairy, seafood, legumes, nuts, and alcohol (in moderation) is associated with lower risks of obesity and chronic diseases, as is consuming less red and processed meats, refined grains, and sugar-sweetened food and beverages Committee et al. (2015b). SPLS’s diet patterns mostly align with this health recommendation. The main discrepancies are that we found natural sweeteners added to coffee or tea to be negatively associated and light green lettuce to be positively associated with BMI. A possible explanation for the former is that people who add honey or sugar to their beverages are conscious of how much sweetener they are consuming, as opposed to someone drinking a pre-sweetened beverage, so they might consume less sweetener in that way. Next, the “light green lettuce” category only includes lettuces similar to iceberg, not romaine. Apart from iceberg lettuce being relatively low in nutrients, this trend could reflect a fast food diet, as iceberg lettuce salads tend to be served at pizza parlors or fast food restaurants as a “healthy” side. Given what we know about obesity, the SPLS trends are fairly conventional; it is reassuring that this method seems to select reasonable aspects of nutrition—in this case highlighting the association between “fast food” or a more “Western” diet with BMI in MESA.

To more completely characterize a heart-healthy diet, it would be ideal to consider multiple pathways through which diet may influence heart disease: obesity, n−3 fatty acids, blood pressure, and cholesterol, among others (De Caterina et al., 2006). As such, the presented BMI-related diet patterns are one aspect of the diet-CVD link, but future studies should investigate the other pathways. On a different note, we acknowledge that any dietary survey is prone to measurement error, which should be taken in consideration when interpreting these results (regardless of method used).

One limitation of supervised methods (e.g., PLS, SPLS, Lasso) versus unsupervised methods is that the resulting components should not be used to test associations with $y$; the inference will be overly-optimistic as the same outcome data is used for component development and association. Instead, one can stop at the component stage and interpret the scientific relevance of the pattern to generate hypotheses for future studies, as we have illustrated. Using components in a prediction setting is also a reasonable next step. If inference is the ultimate goal, researchers in practice have compared a disease outcome with PLS components that were derived from a different loosely-related response (see e.g., Yang et al. (2017); DiBello et al. (2008)). Another alternative is to derive components with a hold-out sample and then test associations in the remainder of the data, or even to
Dimension reduction to identify heart-healthy patterns

derive components using one study population and then perform inference in an independent population, if the selected predictors are available.

Regardless of method or aim, we emphasize that when using data from observational studies such as MESA, results can only be interpreted as associations between predictors and a disease, not as causal relationships. However, these studies still serve a valuable role in increasing our understanding of complex diseases. For instance, if similar results are found in future studies, this could lend support to the idea that only a few foods contribute most to obesity. On a practical note, this could motivate investigating whether the current 120-item FFQ could be condensed to asking about fast food consumption, reducing the burden on participants and data centers when collecting detailed diet intake information.

Overall, we recommend that future nutritional epidemiology studies consider utilizing sparse PLS methods if their scientific goal is to succinctly summarize aspects of exposure data that are most related to an outcome. These results can complement the previous nutritional literature, which is predominantly based on unsupervised methods, and provide further insight into how diet and diseases are related, as previously suggested (Schulze and Hoffmann, 2006; Jacobs Jr and Steffen, 2003). While we have focused on illustrating the usefulness of SPLS in the context of nutritional epidemiology, these dimension reduction methods also apply to many other disciplines whose aim is to reduce the dimension of their predictors—such as bioinformatics, environmental studies, -omics fields, and other settings with high-throughput data—as demonstrated by previous comparisons of component-based methods for different scientific aims (Du et al., 2018; Andersen et al., 2012).

Acknowledgments

This research was supported by funding for the first author from the National Science Foundation (NSF) Graduate Research Fellowship under Grant No. DGE-1256082 as well as the National Institutes of Health (NIH) National Heart, Lung, and Blood Institute (NHLBI) Diversity Supplement under Contract No. 75N92020D00001. The second author was also funded by NHLBI Contract No. 75N92020D00001.

The authors thank the other investigators, the staff, and the participants of the Multi-Ethnic Study of Atherosclerosis (MESA) for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org. This research was also supported by contracts HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168 and N01-HC-95169 from the NHLBI, and by grants UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420 from the National Center for Advancing Translational Sciences (NCATS).
Appendix A: Proofs of Remarks

Useful algebraic properties and assumptions: We begin by presenting several properties and assumptions that will be useful for the proofs of the remarks. When possible, we will present the weighted versions of the properties, with linear results following by setting the given weighting matrix \( V \) to \( I_n \) and using the linear response. For simplicity, we denote the (working) response by \( y \). Let \( L = L(L^T VL)^{-1}L^T \) be the hat matrix corresponding to \( L \), and similarly define hat matrices \( \mathcal{X} \) and \( \mathcal{T} \).

First we point out that in the linear case, \( (I - L) \) is symmetric and idempotent; in the weighted case \( (I - L)^T V = V - VL(L^T VL)^{-1}L^T V = V(I - L) \), and it is idempotent since \( (I - L)^2 = I - 2L + L = (I - L) \). It follows that \( L^T V(I - L) = 0 \).

We can use properties of matrix blockwise inversion and the assumption that both \( X^T VX \) and \( L^T VL \) are invertible to show:

\[
\begin{bmatrix}
X^T VX & X^T VL \\
L^T VX & L^T VL
\end{bmatrix}^{-1}
= \begin{bmatrix}
[X^T VX - X^T V L X]^{-1} & 0 \\
0 & [L^T VL - L^T V L X]^{-1}
\end{bmatrix} \cdot
\begin{bmatrix}
I & -X^T VL(L^T VL)^{-1} \\
-L^T VX(X^T VX)^{-1} & I
\end{bmatrix}
= \begin{bmatrix}
[X^T V(I - L) X]^{-1} & -[X^T V(I - L) X]^{-1} X^T VL(L^T VL)^{-1} \\
-[L^T V(I - \mathcal{X}) L]^{-1} L^T VX(X^T VX)^{-1} & [L^T V(I - \mathcal{X}) L]^{-1}
\end{bmatrix}.
\]

Thus, given a working outcome \( y \), we can express the coefficient estimates from a multiple weighted least squares (WLS) regression as follows:

\[
\begin{bmatrix}
\hat{\beta} \\
\hat{\gamma}
\end{bmatrix} = \begin{bmatrix}
X^T VX & X^T VL \\
L^T VX & L^T VL
\end{bmatrix}^{-1} \begin{bmatrix}
X^T Vy \\
L^T Vy
\end{bmatrix} = \begin{bmatrix}
[X^T V(I - L) X]^{-1} X^T V(I - L) y \\
[L^T V(I - \mathcal{X}) L]^{-1} L^T V(I - \mathcal{X}) y
\end{bmatrix}.
\]

A similar expression follows when using weighted PLS to estimate \( \hat{\beta} \) by replacing \( X \) with components \( T = XR \) whose PLS weights \( R \) are constructed to maximize \( R^T X^T V(I - L) y \).

Lastly, we will make use of the identity \( (I - A)^{-1} = I + (I - A)^{-1} A \) and two forms of the inverse of a difference of matrices presented (for sums) in Henderson and Searle (1981):

\[
(S - ABC)^{-1} = S^{-1} + S^{-1} A (I - BC S^{-1} A)^{-1} BC S^{-1} \quad (\dagger)
\]

\[
= S^{-1} + S^{-1} AB (I - CS^{-1} AB)^{-1} CS^{-1}. \quad (\dagger\dagger)
\]

Proof of Remarks 1 and 2: Assume the same given \( X \), \( L \), \( V \), and outcome \( y \) as above. We will first show that the multiple WLS estimates can be recovered via orthogonalized fits. Using the projection properties mentioned previously, see that modeling orthogonalized covariates \( (I - L)X \) on the full outcome \( y \) will result in \( \hat{\beta} \):

\[
\]

The same holds for \( \hat{\gamma} \) by substituting \( Xs \) and \( Ls \).
Next we show that these coefficients can be obtained using residual fits. Specifically, we prove that \([X^T V(I - \mathcal{L})]^{-1} X^T V(I - \mathcal{L}) y = [X^T V X]^{-1} X^T V(y - \mathcal{L})\). For purposes of space, let \(\Sigma := X^T V X\). We begin by proving the equality with all terms except \(y\). We indicate the lines that use the matrix inverse forms by their corresponding daggers.

\[
[X^T V(I - \mathcal{L})]^{-1} X^T V(I - \mathcal{L}) = [\Sigma - X^T V \mathcal{L} X]^{-1} X^T V(I - \mathcal{L})
\]

\[
= \Sigma^{-1} X^T V(I - \mathcal{L}) + \Sigma^{-1} X^T V \mathcal{L} (I - X \Sigma^{-1} X^T V \mathcal{L})^{-1} X \Sigma^{-1} X^T V(I - \mathcal{L})
\]

\[
= \Sigma^{-1} X^T V - \Sigma^{-1} X^T V \mathcal{L} [I - (I - \mathcal{L} \mathcal{L})^{-1} \mathcal{L} (I - \mathcal{L})] \Sigma^{-1} X^T V(I - \mathcal{L})
\]

\[
= \Sigma^{-1} X^T V - \Sigma^{-1} X^T V \mathcal{L} (I - \mathcal{L})^{-1} \mathcal{L} + \mathcal{L} \mathcal{L}
\]

\[
= \Sigma^{-1} X^T V - \Sigma^{-1} X^T V \mathcal{L} \left[ I + [I - \mathcal{L}]^{-1} \mathcal{L} \right] (I - \mathcal{L})
\]

\[
= \Sigma^{-1} X^T V - \Sigma^{-1} X^T V L (L^T V L)^{-1} \left[ I + L^T V[I - \mathcal{L}]^{-1} \mathcal{L} L(L^T V L)^{-1} \right] L^T V(I - \mathcal{L})
\]

\[
= \Sigma^{-1} X^T V - \Sigma^{-1} X^T V L [L^T V(I - \mathcal{L}) L]^{-1} L^T V(I - \mathcal{L})
\]

\[
= \Sigma^{-1} X^T V \left( I - L [L^T V(I - \mathcal{L}) L]^{-1} L^T V(I - \mathcal{L}) \right).
\]

Multiplying by \(y\) on the right side of the equations gives the \(\hat{\beta}\) equality. Substituting \(Xs\) and \(Ls\) results in the \(\hat{\gamma}\) equality, concluding the proof of Remark 1. \(\square\)

We will next show that the multiple WLS coefficient estimates with PLS components \(T\) as covariates can be obtained by various fits. Recall that we fit a weighted PLS (WPLS) model on the residual outcome \(y - L \hat{\gamma}_0 = (I - \mathcal{L}) y\) to obtain our components. The corresponding PLS weight \(\mathcal{R}\) therefore maximizes \(\mathcal{R}^T X^T V(I - \mathcal{L}) y\). The coefficient estimates from a multiple WLS with these constructed components \(T\) and \(L\) can be expressed as:

\[
\begin{bmatrix} \hat{\mathcal{q}}^T \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} T^T V T & T^T V L \end{bmatrix}^{-1} \begin{bmatrix} T^T V y \\ L^T V y \end{bmatrix} = \begin{bmatrix} T^T V(I - \mathcal{L}) T^{-1} T^T V(I - \mathcal{L}) y \\ L^T V(I - \mathcal{L} L)^{-1} L^T V(I - \mathcal{L} y) \end{bmatrix}.
\]

To demonstrate that residual fits can obtain these coefficients, simply substitute \(T\) for \(X\) in the proof of Remark 1. However, it is reasonable to question whether we can condense the residual fits to two lines, as in WLS. Specifically, does \(\hat{\mathcal{q}}^T = \text{WPLS}[y - L \hat{\gamma} \sim X]\)? Here, the PLS weight \(\mathcal{R}\) is maximizing \(\mathcal{R}^T X^T V(y - L \hat{\gamma}) \neq \mathcal{R}^T X^T V(y - L \hat{\gamma}_0)\). Thus, \(\mathcal{R} \neq \mathcal{R}\), so the constructed components will differ (unless \(X^T V L = 0\)).

We can recover the adjusted coefficients using only two orthogonalized fits. Consider modeling orthogonalized covariates \((I - \mathcal{L}) X\) on the full outcome \(y\). Its PLS weight \(\hat{\mathcal{R}}\) maximizes \(\hat{\mathcal{R}}^T X^T (I - \mathcal{L})^T V y = \hat{\mathcal{R}}^T X^T V(I - \mathcal{L}) y\), so \(\hat{\mathcal{R}}\) is equivalent to \(\mathcal{R}\). The resulting
PLS coefficient is $[\hat{R}^T X^T (I - L)^T V (I - L)\hat{R}]^{-1} \hat{R}^T X^T (I - L)^T V y$, which reduces to $[\hat{R}^T X^T V (I - L)\hat{R}]^{-1} \hat{R}^T X^T V (I - L)y$ after using properties of projections. As $\hat{R}$ is equivalent to $R$, this PLS coefficient is equal to $q^T$. After defining the PLS component $T = XR$, the same projection properties allow us to recover the correct $\hat{\gamma}$ by modeling orthogonalized covariates $(I - T)L$ on the full outcome $y$:


Therefore, we have proven that exact equality holds between the three regression procedures presented in Remark 2.

Appendix B: One standard error test for multiple tuning parameters

**Background:** To assess the predictive ability of candidate models, we use cross-validation (CV) and compare each model’s root mean squared error of prediction (RMSE), where the mean is taken across the CV folds. There is a trade-off between model accuracy and complexity, as there are typically more parsimonious models (with fewer components or more sparsity) that have similar predictive accuracy as the model with the minimum RMSE.

In the context of component-based methods, the one standard error (1-se) test chooses the model with the fewest components whose RMSE is within one standard error of the minimum RMSE. In the example in Figure 4, two components results in the lowest RMSE (red dotted line). After calculating the standard error of the squared error of prediction for the 2-component model across the CV folds, we search for the most parsimonious model whose predictive ability is within 1-se of the minimum RMSE (i.e., whose RMSE is no larger than the upper red bar). Here, the model with one component (blue dashed line) satisfies this criterion, so we select this simpler model.

**Extension:** Sparse component-based methods have two tuning parameters, so we extended the 1-se test to optimize the parameter search across a grid of candidate pairs. Figure 5
Dimension reduction to identify heart-healthy patterns

illustrates our approach in the context of SPLS, where its sparsity parameter is a fraction \( \{\eta \in [0, 1]\} \). In this example, the model with three components and 45% sparsity results in the lowest RMSE (black solid square with “min”). We then consider all models in the region of potential parsimony improvement (red rectangle), which emphasizes a decrease in the number of components for our scientific goal. There are three pairs of candidates (black dashed squares and “1-se” square) whose RMSEs are within 1-se of the minimum RMSE, so we use a weighted metric to determine which pair provides the best improvement in parsimony. Since we favor a decrease in components over an increase in sparsity, we select the model with two components and 45% sparsity (black solid square with “1-se”).

Figure 5: Illustration of the one standard error test for two tuning parameter (for SPLS).

Formally, the weighted metric is given by

\[
\text{improve} = \alpha_1 \ast (\Delta K) + \alpha_2 \ast (\Delta \lambda).
\]

\( \Delta K = K_{\text{min}} - K_{\text{candidate}} \) refers to the change in number of components, \( \Delta \lambda = \lambda_{\text{min}} - \lambda_{\text{candidate}} \) refers to the change in sparsity level, as are the weights assigned to each change in tuning parameter, and \( \text{improve} > 0 \) means that the candidate model provides an increase in parsimony over the model with minimum RMSE. The candidate model with the largest \( \text{improve} \) value is selected by the extended 1-se test.

Example with SPLS: In our setting, we would like fewer components, so we set \( \alpha_1 = +1 \), meaning that the greater the decrease in components, the larger \( \text{improve} \) will be. We would also like more sparsity, so we set \( \alpha_2 = -5 \); the greater the increase in sparsity, the larger \( \text{improve} \) will be. For SPLS, we use its sparsity fraction for the sparsity unit (e.g., \( \lambda = \eta \)). We choose these specific magnitudes because we prefer a decrease in components over an increase in sparsity. Specifically, \( \frac{\alpha_1}{\alpha_2} = 0.20 \) implies that we will select a model with a 1-unit decrease in components even if its sparsity level is smaller by up to 19% (absolute difference). So, if we were considering between a model with 4 components and 70% sparsity and a model with 3 components and 55% sparsity, we would select the latter model. However, we would not select a model with 3 components and 50% sparsity.

Example with SPCR: In Sparse PCR, its sparsity parameter takes a much larger (positive) range, based on the size of the weight \( p \). We strategically select 9 sparsity values to test for each data set based on the size of the largest element in \( p \). To make this test comparable across all data sets, we instead use the index of sparsity value as our \( \lambda \) (e.g., \( \lambda = \{1, 2, ..., 9\} \)). Again, we prefer a decrease in components over an increase in sparsity,
but now we change the magnitudes. We set $\alpha_1 = +2$ and $\alpha_2 = -1$ (with $|\frac{\alpha_1}{\alpha_2}| = 2$), meaning that we will select a model with a 1-unit decrease in components even if its sparsity index is 1 unit smaller.

Appendix C: Further simulation results

Simulations with base covariates only: We first present results from seven “matching-covariance block structure” settings. Here, all blocks are generated with the same level of intra-block covariance ($\rho$), specifically with $\rho = \{0,.1,.3,.5,.6,.8,1\}$. Figure 6 displays average results across 100 generated data sets, as well as 95% confidence intervals. Panel A displays the RMSE of the test sets, and we see that SPLS consistently has the lowest prediction error. This difference is not statistically significant, apart from the independent case ($\rho = 0$), when SPCR has a statistically worse prediction error compared to all the other methods. Panel B displays the number of true variables selected. While SPCR has some difficulty when $\rho = 0$, SPLS does not discard any true variables until $\rho = 0.8$, though it is most noticeable for SPLS and Lasso when $\rho = 1$. Panel C displays the number of false variables selected. SPLS and Lasso select less than one false variable until $\rho > 0.1$. Lasso is less sensitive to the block-covariance structure, selecting at most 2.6 false variables when $\rho = 0.8$, whereas SPLS selects 12 false variables. However, when the block-covariance is more realistic to food frequency questionnaire data ($\rho \leq 0.5$), SPLS tends to select less than a third of the false variables (i.e., 5). Panel D displays the number of components selected. Except for with perfect correlation ($\rho = 1$), PCR and SPCR select 23–24 components. PLS-based methods tend to select 2–7 components, with SPLS generally selecting half a component more than PLS. Panel E displays the percent of $y$ variance explained ($R^2$). Most methods exhibit a gradual decrease in $R^2$ as the block-covariances increase. Typically, PCR and PLS explain the most $y$ variance, followed by SPLS, SPCR, and Lasso.

We also tested “mixed-covariance block structures”, in which we permuted each block’s covariance from among $\rho = \{0.05,0.10,0.30,0.50\}$. Since we still imposed independence across blocks, each block behaves similarly to the corresponding covariance structure from the “matching” scenarios. In general, the supervised methods have a harder time distinguishing between true and false variables when the blocks with more relevant variables (Blocks 2 and 3) have higher intra-block covariances. In the easiest mixed settings, SPLS and Lasso select less than one false variable; in the hardest mixed setting, SPLS selects 5 false variables while Lasso selects 2. SPLS again has the lowest RMSE across all mixed settings, though PCR comes close when the blocks with more relevant variables have high intra-block covariances. Interestingly, in that setting when the blocks with more relevant variables have low intra-block covariances, PCR and SPCR select only 20 components, as compared to their usual 23–24 components.
Figure 6: Simulation results across 100 data sets for matching-covariance block structures.
Simulations with adjustment covariates: To compare how the adjustment techniques perform for Chun and Keleş’s SPLS procedure, we include an expanded simulation setting. We generate the same multiblock $X$ predictors as in Case 2. Then, we create $L$ such that three of its variables emulate precision variables and the last either another precision variable or a confounder. Specifically, the first two variables are binomial and Poisson distributed with roughly .40 correlation but independent of $X$. The second two are normally-distributed with .18 correlation, the last of which is either independent or correlated (magnitude of .30) with two $X$ variables. The mean model of the outcome is constructed from the true $X$ covariates (with $\beta = \pm 1$) and $L$ covariates (with $\gamma = \pm 5$). We vary the amount of Gaussian noise added to the outcome, ranging from sd=1 to sd=11.

Across 50 generated data sets, we split our observations into a training ($n=700$) and test ($n=300$) set, and we use 10-fold CV on the training data to select tuning parameters using the minimum RMSE criterion. On the test data, we evaluate variable selection (true variables of 9 and false variables of 15), predictive ability (RMSE of outcome; not shown), and PLS estimation error (RMSE of $\beta$). The last metric compares $\hat{\beta}$ with the known $\beta$; an ideal method will select all true variables without overly-shrinking their estimates and discard most false variables, assigning small magnitudes for the estimates of false selections.

Figure 7: Average simulation results on 50 multiblock data sets ($n = 1000$, $p = 28$, $|\gamma| = 5$) with $\text{sd}(y)=3$, across adjustment variable settings.

In the left panel of Figure 7, we see that all methods were able to select the true covariates when the outcome had smaller variability (sd=3). The covariate-summary adjustment was better equipped to discard false variables across adjustment correlation settings. However, the largest discrepancy was that the two-stage residual adjustment (included for comparison) resulted in the worst PLS estimation error by far. With more substantial outcome
variability (sd=7), the adjustment techniques dropped a fraction of a true variable while selecting roughly four false variables (see Figure 8). Again, covariate-summary adjustment tended to do the best in variable selection across correlation settings. This technique also performed the best in PLS estimation error in the presence of a confounder (as expected), but component-summary adjustment outperformed it when \( L \) was associated with two false \( X \) variables (also as expected, since it orthogonalizes \( X \)). The two-stage residual adjustment always had worse PLS estimation error than the component-summary adjustment.

Figure 8: Average simulation results on 50 multiblock data sets \((n = 1000, \ p = 28, |\gamma_\beta| = 5)\) with \( sd(y)=7 \), across adjustment variable settings.

Appendix D: Food frequency questionnaire items in the Multi-Ethnic Study of Atherosclerosis

Table 3 lists the 120 food frequency questionnaire items in the Multi-Ethnic Study of Atherosclerosis (MESA). These items are ordered by food group, which is the same order as presented in Figure 3. For certain items, examples of foods and beverages are provided in brackets.

Table 3: List of food frequency questionnaire items in MESA, ordered by food group

<table>
<thead>
<tr>
<th>Fruits</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Peaches, apricots, nectarines, or plums</td>
</tr>
<tr>
<td>2</td>
<td>Cantaloupe, mango, or papaya</td>
</tr>
<tr>
<td>3</td>
<td>Berries</td>
</tr>
<tr>
<td>4</td>
<td>Apples or pears</td>
</tr>
<tr>
<td>5</td>
<td>Bananas or plantains</td>
</tr>
<tr>
<td>6</td>
<td>Sweet citrus or kiwi [e.g., orange]</td>
</tr>
<tr>
<td>7</td>
<td>Dried fruit</td>
</tr>
<tr>
<td>8</td>
<td>All other fruit</td>
</tr>
<tr>
<td>9</td>
<td>Orange or grapefruit juice</td>
</tr>
<tr>
<td>10</td>
<td>Non-citrus fruit juice [e.g., apple]</td>
</tr>
<tr>
<td>11</td>
<td>Tomato</td>
</tr>
<tr>
<td>12</td>
<td>Avocado</td>
</tr>
<tr>
<td>13</td>
<td>Light green lettuce salad [e.g., iceberg lettuce]</td>
</tr>
<tr>
<td>14</td>
<td>Dark green lettuce salad or cooked leafy greens [e.g., romaine lettuce, spinach]</td>
</tr>
<tr>
<td>15</td>
<td>Carrots</td>
</tr>
<tr>
<td>16</td>
<td>Cruciferous vegetables [e.g., broccoli, cabbage]</td>
</tr>
<tr>
<td>17</td>
<td>Green beans or peas</td>
</tr>
<tr>
<td>18</td>
<td>Corn or hominy</td>
</tr>
<tr>
<td>19</td>
<td>Winter or acorn squash</td>
</tr>
<tr>
<td>20</td>
<td>All other vegetables</td>
</tr>
<tr>
<td>21</td>
<td>Unfried potatoes or turnips</td>
</tr>
<tr>
<td>22</td>
<td>Sweet potatoes or yams</td>
</tr>
<tr>
<td>23</td>
<td>Stir-fried vegetables without added protein</td>
</tr>
<tr>
<td>24</td>
<td>Pancakes, waffles, or French toast</td>
</tr>
<tr>
<td>25</td>
<td>Oatmeal</td>
</tr>
<tr>
<td>26</td>
<td>Hot cereal besides oatmeal [e.g., grits]</td>
</tr>
<tr>
<td>27</td>
<td>Cold cereal</td>
</tr>
<tr>
<td>28</td>
<td>White bread or rolls</td>
</tr>
<tr>
<td>29</td>
<td>Dark, whole grain bread or rolls</td>
</tr>
<tr>
<td>30</td>
<td>Bran muffins</td>
</tr>
<tr>
<td>31</td>
<td>All other bread</td>
</tr>
<tr>
<td>32</td>
<td>Crackers, pretzels, or popcorn</td>
</tr>
<tr>
<td>33</td>
<td>White rice</td>
</tr>
<tr>
<td>34</td>
<td>Brown or wild rice</td>
</tr>
<tr>
<td>35</td>
<td>Tortillas on the side</td>
</tr>
<tr>
<td>36</td>
<td>Eggs or omelettes</td>
</tr>
<tr>
<td>37</td>
<td>Sausage, bacon, or chorizo</td>
</tr>
<tr>
<td>38</td>
<td>Nuts (except peanuts) [e.g., almonds]</td>
</tr>
<tr>
<td>39</td>
<td>Seeds [e.g., sunflower]</td>
</tr>
<tr>
<td>40</td>
<td>Peanuts or peanut butter</td>
</tr>
<tr>
<td>41</td>
<td>Unfried beans [e.g., black beans]</td>
</tr>
<tr>
<td>42</td>
<td>Refried beans</td>
</tr>
<tr>
<td>43</td>
<td>Hamburger or meat loaf</td>
</tr>
<tr>
<td>44</td>
<td>Steaks, roasts, or ribs</td>
</tr>
<tr>
<td>45</td>
<td>Ham hocks, pig’s feet, or chicharrones</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>46</td>
<td>Ham or processed pork meat [e.g., hot dogs]</td>
</tr>
<tr>
<td>47</td>
<td>Non-fried chicken or turkey</td>
</tr>
<tr>
<td>48</td>
<td>Fried chicken</td>
</tr>
<tr>
<td>49</td>
<td>Organ meats</td>
</tr>
<tr>
<td>50</td>
<td>Meat-based gravy</td>
</tr>
<tr>
<td>51</td>
<td>Fried seafood</td>
</tr>
<tr>
<td>52</td>
<td>Unfried molluscs and crustaceans [e.g., shrimp, oysters]</td>
</tr>
<tr>
<td>53</td>
<td>Unfried tuna, salmon, or sardines</td>
</tr>
<tr>
<td>54</td>
<td>All other unfried fish</td>
</tr>
</tbody>
</table>

**Mixed entrees**
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>Cream-based soups [e.g., chowder]</td>
</tr>
<tr>
<td>56</td>
<td>Pea or bean-based soups [e.g., lentil soup]</td>
</tr>
<tr>
<td>57</td>
<td>Miso soup</td>
</tr>
<tr>
<td>58</td>
<td>Other soups</td>
</tr>
<tr>
<td>59</td>
<td>Fried rice</td>
</tr>
<tr>
<td>60</td>
<td>Asian noodle soups [e.g., ramen]</td>
</tr>
<tr>
<td>61</td>
<td>Asian dumplings and buns [e.g., dim sum]</td>
</tr>
<tr>
<td>62</td>
<td>Chow mein</td>
</tr>
<tr>
<td>63</td>
<td>Stir-fried meat or poultry with vegetables</td>
</tr>
<tr>
<td>64</td>
<td>Stir-fried seafood with vegetables</td>
</tr>
<tr>
<td>65</td>
<td>Stir-fried tofu with vegetables</td>
</tr>
<tr>
<td>66</td>
<td>Vegetarian burrito or quesadilla</td>
</tr>
<tr>
<td>67</td>
<td>Non-vegetarian burrito, quesadilla, or fajitas</td>
</tr>
<tr>
<td>68</td>
<td>Vegetarian Hispanic dish (except burrito or quesadilla)</td>
</tr>
<tr>
<td>69</td>
<td>Non-vegetarian tacos, nachos, enchiladas, or tamales</td>
</tr>
<tr>
<td>70</td>
<td>Hispanic spiced stew or ground meat [e.g., picadillo]</td>
</tr>
<tr>
<td>71</td>
<td>Hispanic chicken with rice</td>
</tr>
<tr>
<td>72</td>
<td>Non-vegetarian chili</td>
</tr>
<tr>
<td>73</td>
<td>Hispanic red chili with meat</td>
</tr>
<tr>
<td>74</td>
<td>Hispanic green chili with meat</td>
</tr>
<tr>
<td>75</td>
<td>Salsa</td>
</tr>
<tr>
<td>76</td>
<td>Vegetarian cream-based pasta [e.g., macaroni and cheese]</td>
</tr>
<tr>
<td>77</td>
<td>Non-vegetarian cream-based pasta [e.g., tuna noodle casserole]</td>
</tr>
<tr>
<td>78</td>
<td>Vegetarian tomato-based pasta</td>
</tr>
<tr>
<td>79</td>
<td>Non-vegetarian tomato-based pasta [e.g., spaghetti]</td>
</tr>
<tr>
<td>80</td>
<td>Pizza</td>
</tr>
<tr>
<td>81</td>
<td>Meat or poultry stew or empanadas</td>
</tr>
<tr>
<td>82</td>
<td>Seafood stew [e.g., seafood gumbo]</td>
</tr>
<tr>
<td>83</td>
<td>Chicken, tuna, or egg salad</td>
</tr>
<tr>
<td>84</td>
<td>Pasta, macaroni, or potato salad</td>
</tr>
</tbody>
</table>

**Dairy**
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>85</td>
<td>Cottage or ricotta cheese</td>
</tr>
<tr>
<td>86</td>
<td>All cheeses (except cottage, ricotta)</td>
</tr>
<tr>
<td>87</td>
<td>Plain yogurt</td>
</tr>
<tr>
<td>88</td>
<td>Flavored yogurt</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>89</td>
<td>Whole milk beverages</td>
</tr>
<tr>
<td>90</td>
<td>2% milk beverages</td>
</tr>
<tr>
<td>91</td>
<td>Skim or 1% milk beverages</td>
</tr>
<tr>
<td>92</td>
<td>Milk added to coffee or tea (except lattes)</td>
</tr>
<tr>
<td><strong>Sweets/oils</strong></td>
<td></td>
</tr>
<tr>
<td>93</td>
<td>Margarine or mayonnaise on bread or rolls</td>
</tr>
<tr>
<td>94</td>
<td>Butter on bread or rolls</td>
</tr>
<tr>
<td>95</td>
<td>Potato, corn, or tortilla chips</td>
</tr>
<tr>
<td>96</td>
<td>Fried potatoes</td>
</tr>
<tr>
<td>97</td>
<td>Butter or oil added to meal (except bread)</td>
</tr>
<tr>
<td>98</td>
<td>Sweet spreads on bread or cereal [e.g., jam]</td>
</tr>
<tr>
<td>99</td>
<td>Full-fat ice cream</td>
</tr>
<tr>
<td>100</td>
<td>Frozen desserts (except full-fat ice cream)</td>
</tr>
<tr>
<td>101</td>
<td>Tofu-based desserts</td>
</tr>
<tr>
<td>102</td>
<td>Non-chocolate sweet pastries [e.g., cookies]</td>
</tr>
<tr>
<td>103</td>
<td>Chocolate pastries or candy [e.g., brownies]</td>
</tr>
<tr>
<td>104</td>
<td>Pies</td>
</tr>
<tr>
<td>105</td>
<td>Pudding or custard</td>
</tr>
<tr>
<td>106</td>
<td>Other candy</td>
</tr>
<tr>
<td>107</td>
<td>Sweetened condensed milk</td>
</tr>
<tr>
<td>108</td>
<td>Cream additives to coffee or tea [e.g., half-and-half]</td>
</tr>
<tr>
<td>109</td>
<td>Natural sweeteners added to coffee or tea [e.g., honey]</td>
</tr>
<tr>
<td><strong>Beverages</strong></td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>Soy milk</td>
</tr>
<tr>
<td>111</td>
<td>Soda or sweetened beverage</td>
</tr>
<tr>
<td>112</td>
<td>Diet soda</td>
</tr>
<tr>
<td>113</td>
<td>Meal replacement drinks [e.g., Ensure]</td>
</tr>
<tr>
<td>114</td>
<td>Hot chocolate</td>
</tr>
<tr>
<td>115</td>
<td>Coffee (except lattes)</td>
</tr>
<tr>
<td>116</td>
<td>Herbal tea</td>
</tr>
<tr>
<td>117</td>
<td>Black or green tea</td>
</tr>
<tr>
<td>118</td>
<td>Wine</td>
</tr>
<tr>
<td>119</td>
<td>Beer</td>
</tr>
<tr>
<td>120</td>
<td>Liquor or mixed drinks</td>
</tr>
</tbody>
</table>

**References**


Dimension reduction to identify heart-healthy patterns


R Core Team et al. R: A language and environment for statistical computing, 2013.


