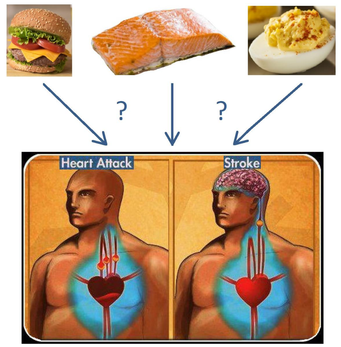
USING DIMENSION-REDUCTION METHODS TO IDENTIFY INTERPRETABLE DIETARY PATTERNS RELATED TO BODY MASS INDEX (BMI) IN THE MULTI-ETHNIC STUDY OF ATHEROSCLEROSIS (MESA)

Natalie Gasca (ncgasca@uw.edu) & Robyn McClelland, PhD *University of Washington*

Motivation

Figure 1. What is a heart-healthy diet?



Nutritional Epidemiology Problem

Difficult interpreting diet patterns from **unsupervised methods**, as they aren't driven by heart disease (CVD) risk

Study Proposal

Use **supervised** dimension-reduction methods, which do include disease data

Goal: BMI-Related Dietary Patterns

Why? Excess body weight is an important pathway between nutrition and heart disease

How? Using component-based methods on food frequency data

Who? MESA is a prospective cohort study that tracks > 6,400 healthy multi-ethnic older adults since 2000 as some progress from subclinical to clinical CVD

Methods

Component-Based Methods

construct mutually orthogonal components that are weighted sums of X. The difference is what each method's "weight" maximizes:

Principal Component Analysis (PCA)
Var(X)

Sparse PCA (SPCA)
Var(X) + elastic net penalty on weight

Partial Least Squares (PLS)
Cov(X,Y)

Sparse PLS (SPLS)
Cov (X,Y) + soft-threshold (L_1 norm) penalty on surrogate weight

Tuning Parameter Selection

10-fold cross validation to select the "simplest" reasonably-predictive model using the 1-standard error test.

For sparse methods, I adapted this test to search across a grid of candidate parameters.

Data Pre-Processing

Baseline BMI was pre-adjusted for age, gender, race/ethnicity, intentional exercise, and total caloric intake

Baseline diet data (120 food/beverage servings-per-day) were normalized

Analyzed in *R* :: `pls`, `elasticnet`, & `spls`

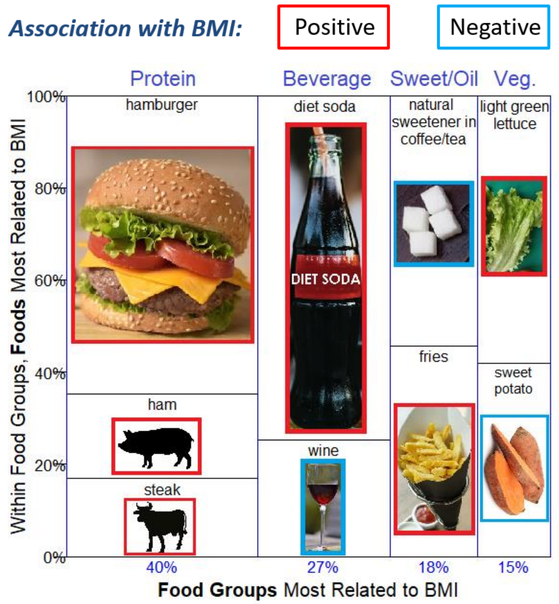
Results from MESA

Table 1. Summary of BMI-related diet patterns

Method	CV-RMSE	# Comp.	# Var.	% R^2
PCA	4.97	8	120	1.5
SPCA	4.95	27	104	3.5
PLS	4.94	1	120	3.9
SPLS	4.94	2	9	3.2

Note: CV-RMSE = Cross-Validated Root Mean Squared Error, Comp. = Components, and Var. = Variables.

Figure 2. BMI-related diet pattern from SPLS



Note: "Relatedness" = squared correlation with BMI.

Study Findings

- **SPCA ≤ PCA:** fewer foods and streamlined patterns, but more components and huge computational burden
- **PLS > PCA:** better predictive ability and fewer, more-relevant components
- **SPLS > PLS:** far fewer variables needed for similar number of components and similar predictive ability

Conclusion

- By using BMI data and variable selection, **SPLS created a more interpretable dietary pattern**
  - → In simulations, PLS-based methods picked the fewest patterns and SPLS discarded most noise variables
- **SPLS can provide useful scientific insight** for nutritional epidemiologists who want dietary patterns that are tailored to heart disease risk

References

PCA Pearson K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science* **2**, 559-572.

SPCA Zou H., Hastie T., and Tibshirani R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics* **15**(2), 265-286.

PLS Wold H. (1966). Nonlinear Estimation by Iterative Least Square Procedures. *Research Papers in Statistics*. Festschrift for J. Neyman, edited by F. N. David. Wiley. pg 411-444.

SPLS Chun H. and Keles S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society, Series B* **72**(1), 3-25.

Funding: National Science Foundation Graduate Research Fellowship under DGE-1256082

Acknowledgements: The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org.