Alternating Direction Methods for Dimensionality Reduction, Classification, and Feature Selection

Brendan Ames

Department of Mathematics
The University of Alabama

UA Applied Mathematics Seminar
October 10, 2014

Outline

Consider an alternating direction method for simultaneously performing classification and feature selection of high dimensional data.

Outline:

- Review of linear discriminant analysis.
- Penalized LDA in the high-dimensional setting.
- Our ADMM algorithm.
- Numerical results.

Joint with Mingyi Hong, Iowa State University, IMSE.

Classification in automobile manufacturing

Want to design automated rule for detecting flaws in manufacturing process.

Want to automatically identify bad cars before leaving the factory to be sold.

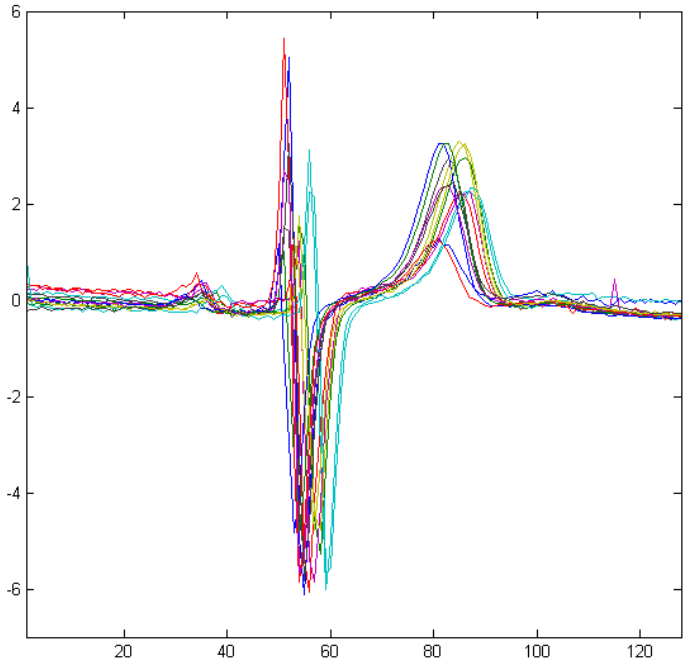
Can make many measurements during manufacturing (high dimension).

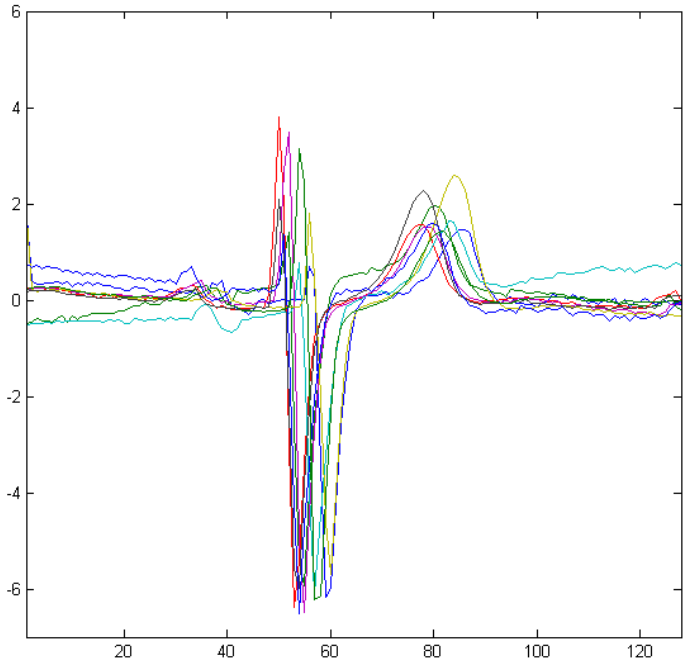
Want to use as few samples as possible, to avoid producing many defective cars (small sample size).

Also want to identify where process is defective (feature selection).

The EKGFiveDays data set

- Data set from UCR Time Series repository. www.cs.ucr.edu/∼eamonn/time_series_data/
- Training set consists of 23 heartbeat/EKG signals in $\mathbf{R}^{128}$.
- 14 from Class D (healthy), 9 from Class E (unhealthy).
- Want to design rule to distinguish between classes.





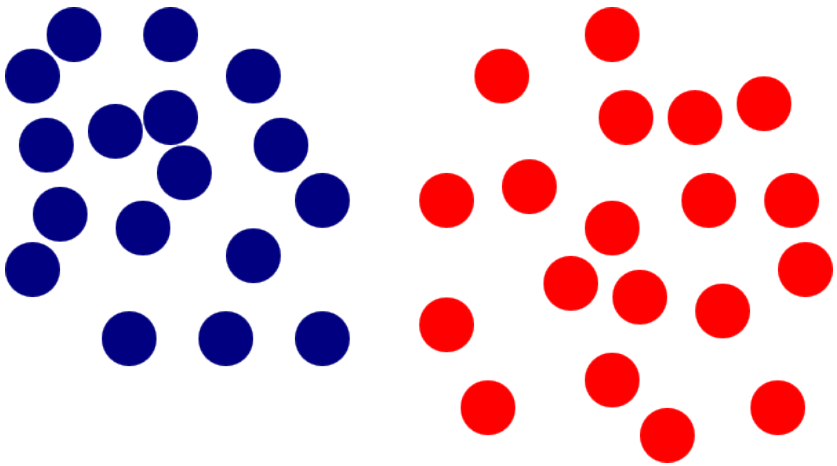
The Classification Problem

Given n observations $\boldsymbol{x}_i \in \mathbf{R}^p$.

$\boldsymbol{x}_i$ belongs to exactly one of k classes $C_1, C_2, \ldots, C_k$

Class labels are known.

Problem: design a decision rule to assign new observations to exactly one of the k classes.



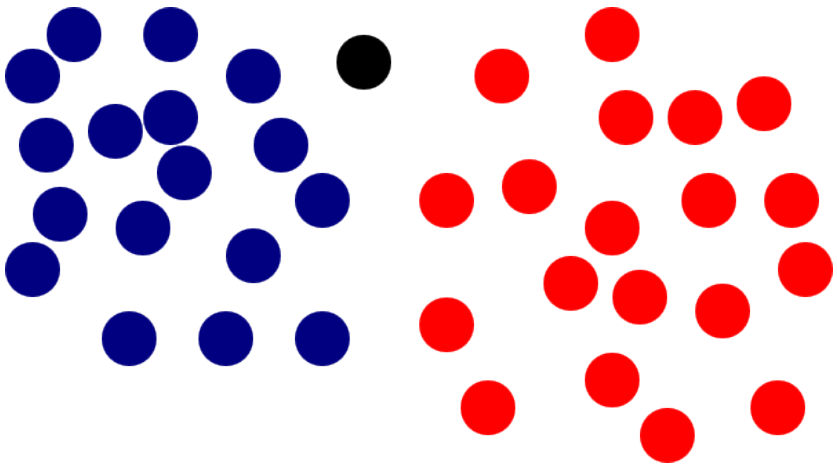
The Classification Problem

Given n observations $\boldsymbol{x}_i \in \mathbf{R}^p$.

$\boldsymbol{x}_i$ belongs to exactly one of k classes $C_1, C_2, \ldots, C_k$

Class labels are known.

Problem: design a decision rule to assign new observations to exactly one of the k classes.



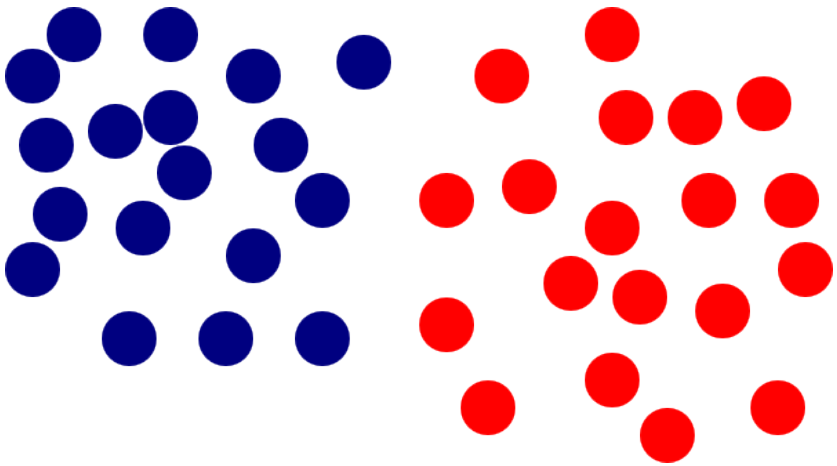
The Classification Problem

Given n observations $\boldsymbol{x}_i \in \mathbf{R}^p$.

$\boldsymbol{x}_i$ belongs to exactly one of k classes $C_1, C_2, \ldots, C_k$

Class labels are known.

Problem: design a decision rule to assign new observations to exactly one of the k classes.



Dimensionality Reduction

Measured data vectors are high-dimensional but we have some reason to expect the data to lie on/near a low-dimensional manifold.

Data are multiple measurements of some underlying source.

Want to identify set of degrees of freedom which reproduce most of the variability of the data set.

Why Reduce Dimension?

Gives compact low-dimensional encoding of high-dimensional data set.

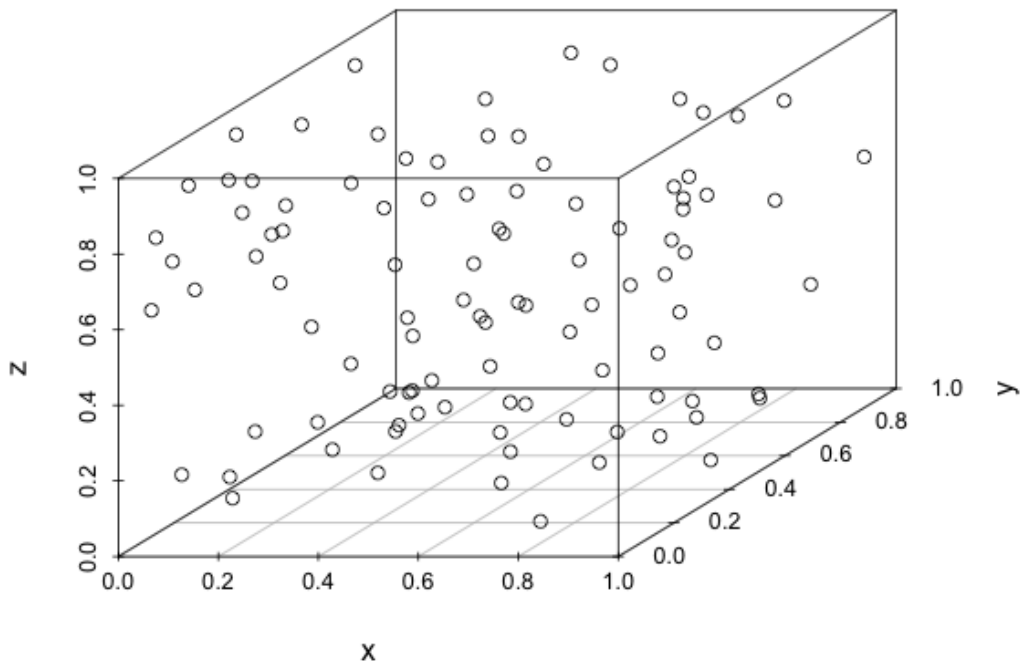
Especially useful for visualization.

Preprocess data for learning task.

e.g., clustering, classification, etc. might be easier in the lower dimensional encoding.

Why Reduce Dimension? (2)

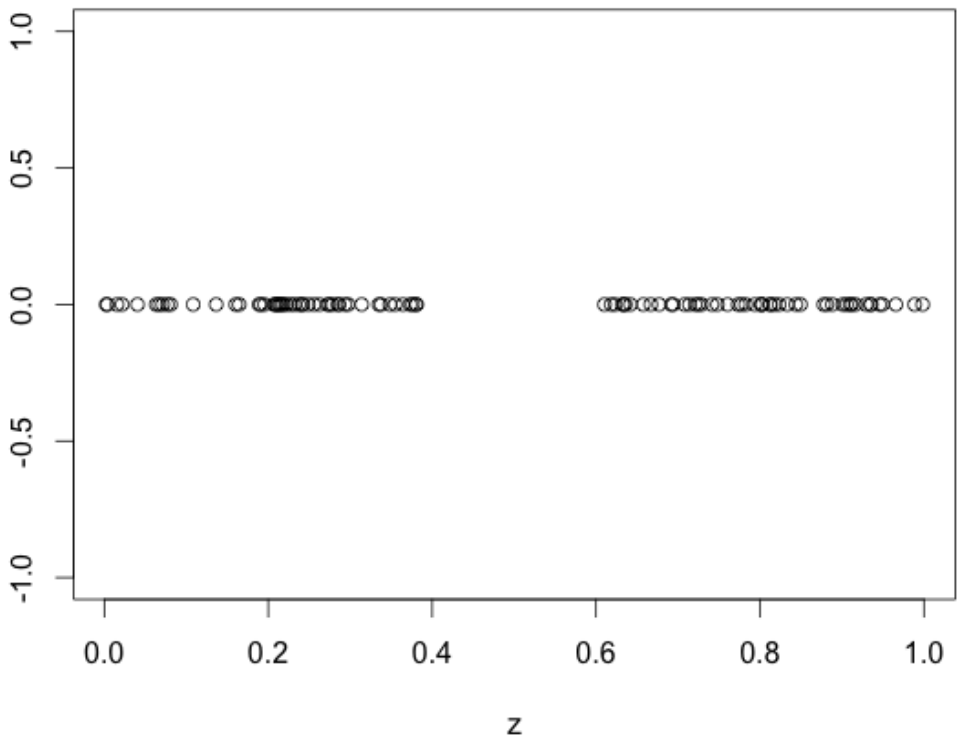
Example: Data clustering.



Difficult to identify clusters in 3D.

Why Reduce Dimension? (2)

Example: Data clustering.



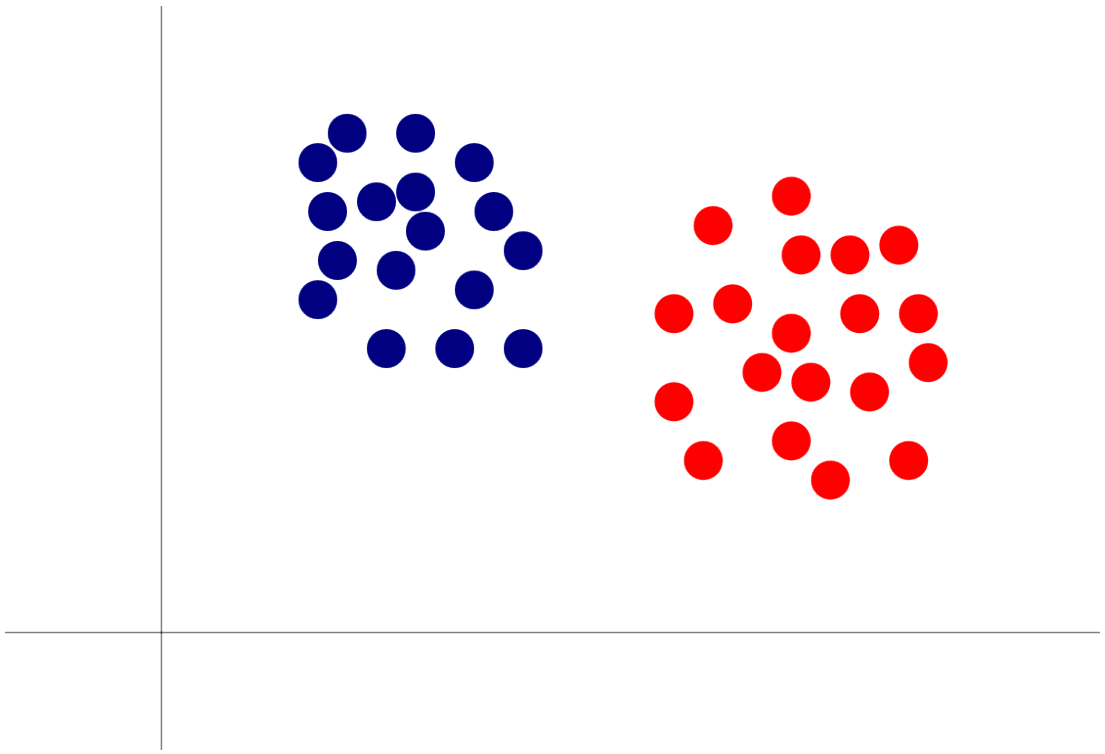
Difficult to identify clusters in 3D.

Easy after projecting onto z-axis.

Nearest Centroid Rule

Project data to lower dimensional subspace where classes are well-separated.

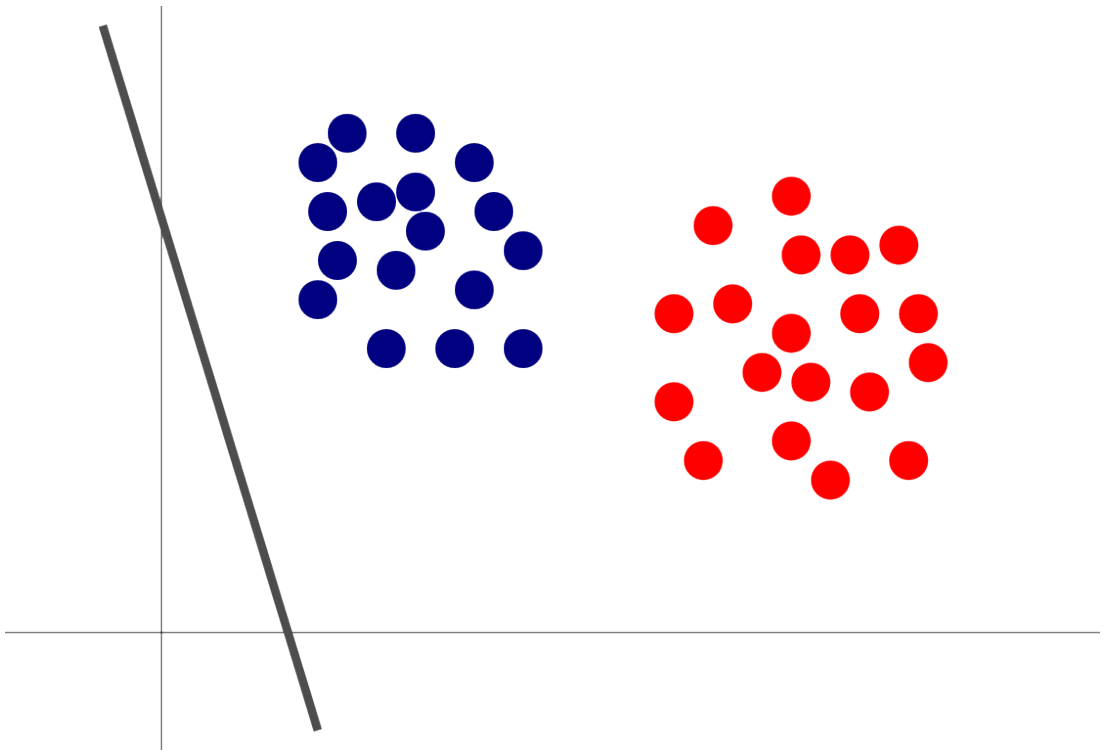
Nearest centroid rule: assign test data to class with nearest projected centroid.



Nearest Centroid Rule

Project data to lower dimensional subspace where classes are well-separated.

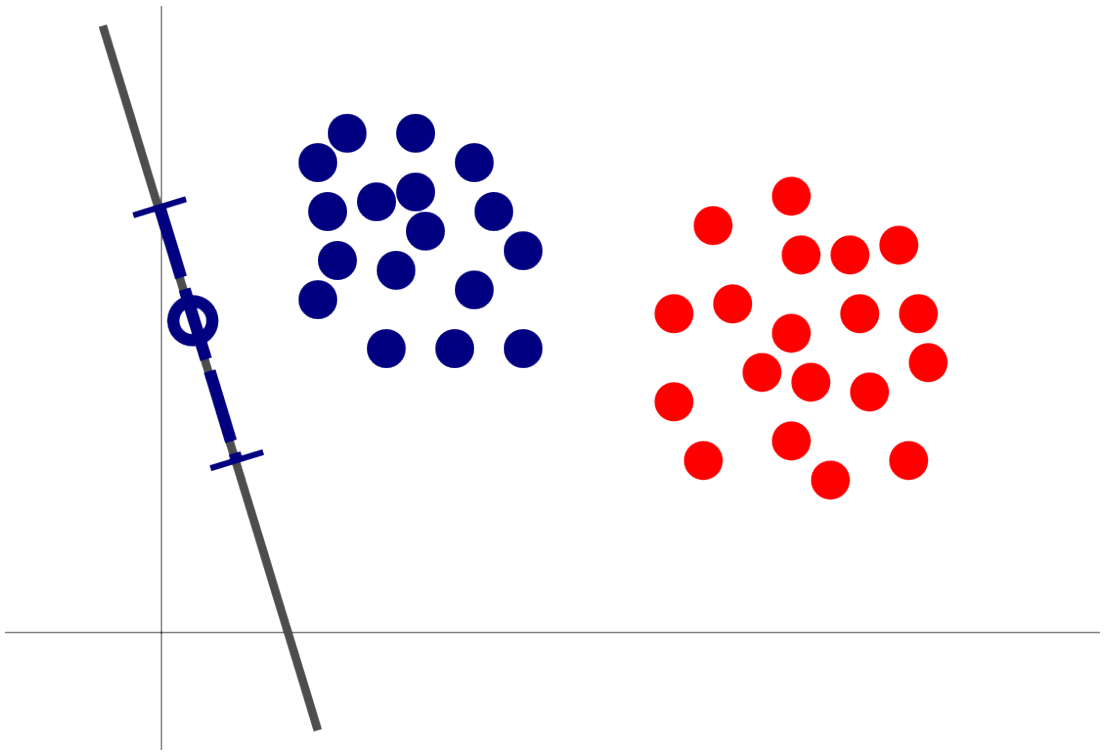
Nearest centroid rule: assign test data to class with nearest projected centroid.



Nearest Centroid Rule

Project data to lower dimensional subspace where classes are well-separated.

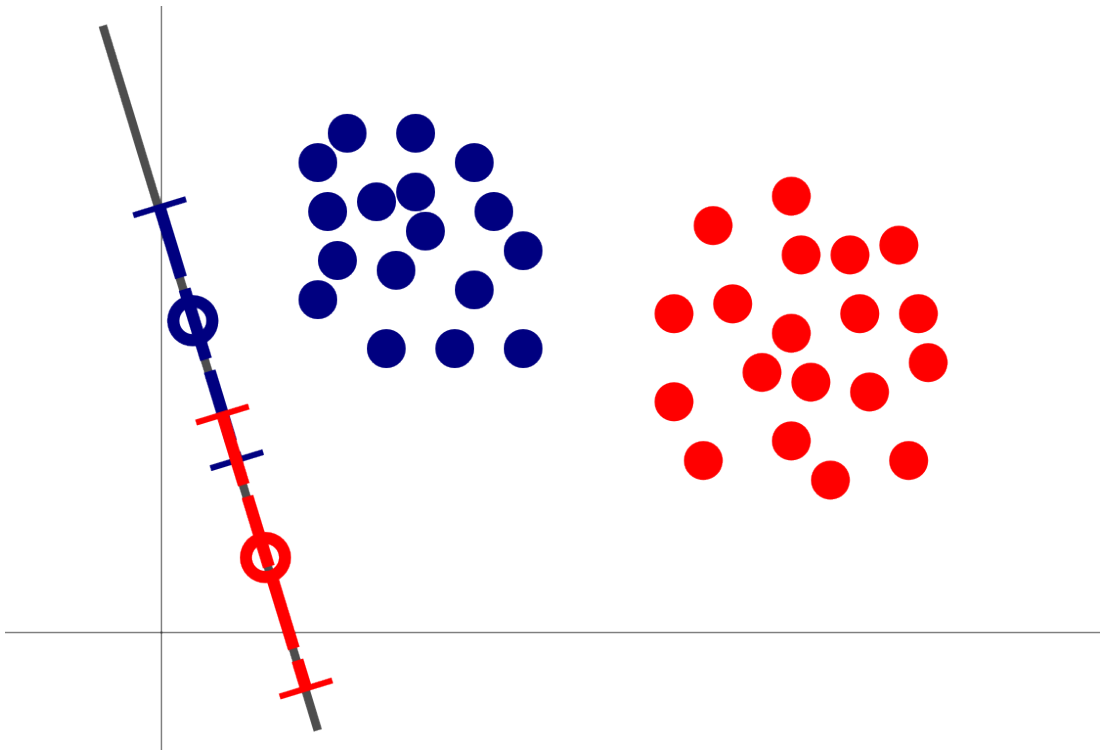
Nearest centroid rule: assign test data to class with nearest projected centroid.



Nearest Centroid Rule

Project data to lower dimensional subspace where classes are well-separated.

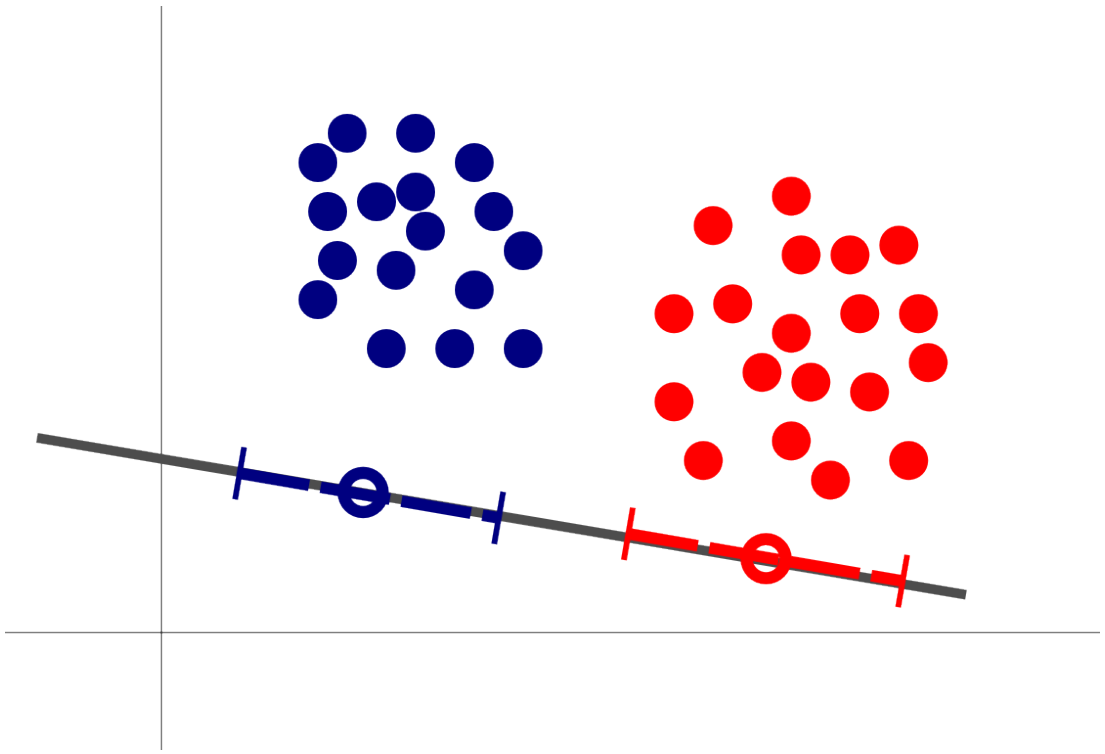
Nearest centroid rule: assign test data to class with nearest projected centroid.



Nearest Centroid Rule

Project data to lower dimensional subspace where classes are well-separated.

Nearest centroid rule: assign test data to class with nearest projected centroid.



Assumptions

Classes $C_1, C_2, \ldots, C_k$ have common covariance matrix $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$.

The k centroids lie in some affine subspace of dimension $k - 1$.

Classes linearly separable in the common subspace under these assumptions.

Data is normalized and centered so that each feature has sample mean 0 and sample variance 1.

Fisher Linear Discriminant Analysis

Want to find subspace $\boldsymbol{S} \subseteq \mathbf{R}^p$ such that centroids of classes are as far apart as possible when projected onto $\boldsymbol{S}$, while keeping within-class scatter relatively small.

Between and Within-Class Scatter Matrices

We can treat each class as a separate data set, with sample mean and scatter matrix

$$\boldsymbol{\mu}_i = \sum_{j \in C_i} \frac{\boldsymbol{x}_j}{|C_i|} \qquad \boldsymbol{S}_i = \sum_{j \in C_i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T$$

Total within-class scatter is defined as

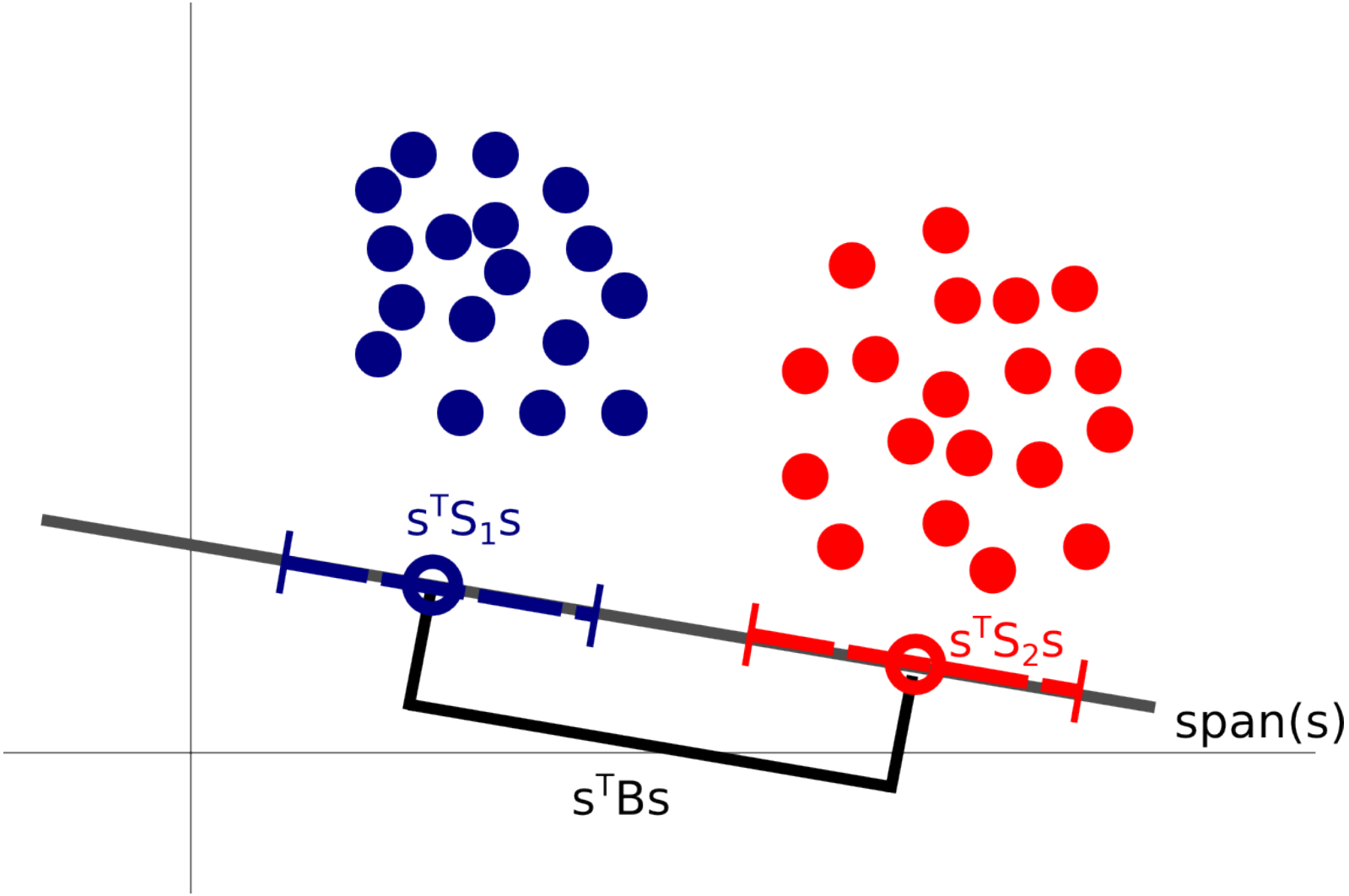
$$\boldsymbol{W} = \boldsymbol{S}_1 + \boldsymbol{S}_2 + \cdots + \boldsymbol{S}_k.$$

Total between-class scatter is defined as

$$\boldsymbol{B} = \boldsymbol{\mu}_1 \boldsymbol{\mu}_1^T + \boldsymbol{\mu}_2 \boldsymbol{\mu}_2^T + \cdots + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T$$

(assuming data has been centered).

Between and Within-Class Scatter Matrices (2)



$\boldsymbol{S}_i$ measures variance within class i.

$\boldsymbol{B}$ measures variance within the set of class means.

Fisher LDA

Want $\ell \leq k-1$ $\boldsymbol{W}$-conjugate discriminant vectors $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_\ell$ maximizing:

$$J(\boldsymbol{s}) = \frac{\boldsymbol{s}^T \boldsymbol{B} \boldsymbol{s}}{\boldsymbol{s}^T \boldsymbol{W} \boldsymbol{s}}.$$

Can show that

$$\boldsymbol{s}_i = \underset{\boldsymbol{s} \in \mathbf{R}^p}{\arg\max} \left\{ \boldsymbol{s}^T \boldsymbol{B} \boldsymbol{s} : \boldsymbol{s}^T \boldsymbol{W} \boldsymbol{s} \leq 1,\ \boldsymbol{s}^T \boldsymbol{W} \boldsymbol{s}_j = 0\ \forall j < i \right\}$$

If $\boldsymbol{W}$ is positive definite, can show that

$$\boldsymbol{s}_i = \boldsymbol{W}^{1/2} \boldsymbol{z}_i$$

where $\boldsymbol{z}_i$ is the eigenvector corresponding to the ith largest eigenvalue of $\boldsymbol{W}^{-1/2} \boldsymbol{B} \boldsymbol{W}^{-1/2}$.

The High-Dimension/Small Sample Size problem

Three significant issues when $p >> n$:

The High-Dimension/Small Sample Size problem

Three significant issues when $p >> n$:

1. $\boldsymbol{W}$ is singular when $n < p$.
   - Cannot solve by change of variables.
   - $\boldsymbol{s}^T\boldsymbol{B}\boldsymbol{s}/\boldsymbol{s}^T\boldsymbol{W}\boldsymbol{s}$ is unbounded if $\mathrm{Null}(\boldsymbol{W}) \not\subseteq \mathrm{Null}(\boldsymbol{B})$.

The High-Dimension/Small Sample Size problem

Three significant issues when $p >> n$:

1. $\boldsymbol{W}$ is singular when $n < p$.
   - Cannot solve by change of variables.
   - $\boldsymbol{s}^T\boldsymbol{B}\boldsymbol{s}/\boldsymbol{s}^T\boldsymbol{W}\boldsymbol{s}$ is unbounded if $\mathrm{Null}(\boldsymbol{W}) \not\subseteq \mathrm{Null}(\boldsymbol{B})$.
2. The discriminant vectors are dense, difficult to interpret.

The High-Dimension/Small Sample Size problem

Three significant issues when $p >> n$:

1. $\boldsymbol{W}$ is singular when $n < p$.
   - Cannot solve by change of variables.
   - $\boldsymbol{s}^T\boldsymbol{B}\boldsymbol{s}/\boldsymbol{s}^T\boldsymbol{W}\boldsymbol{s}$ is unbounded if $\mathrm{Null}(\boldsymbol{W}) \not\subseteq \mathrm{Null}(\boldsymbol{B})$.
2. The discriminant vectors are dense, difficult to interpret.
3. The sample covariance/scatter matrices $\boldsymbol{B}$, $\boldsymbol{W}$ are generally inconsistent approximations of the true population between and within-class covariance matrices.

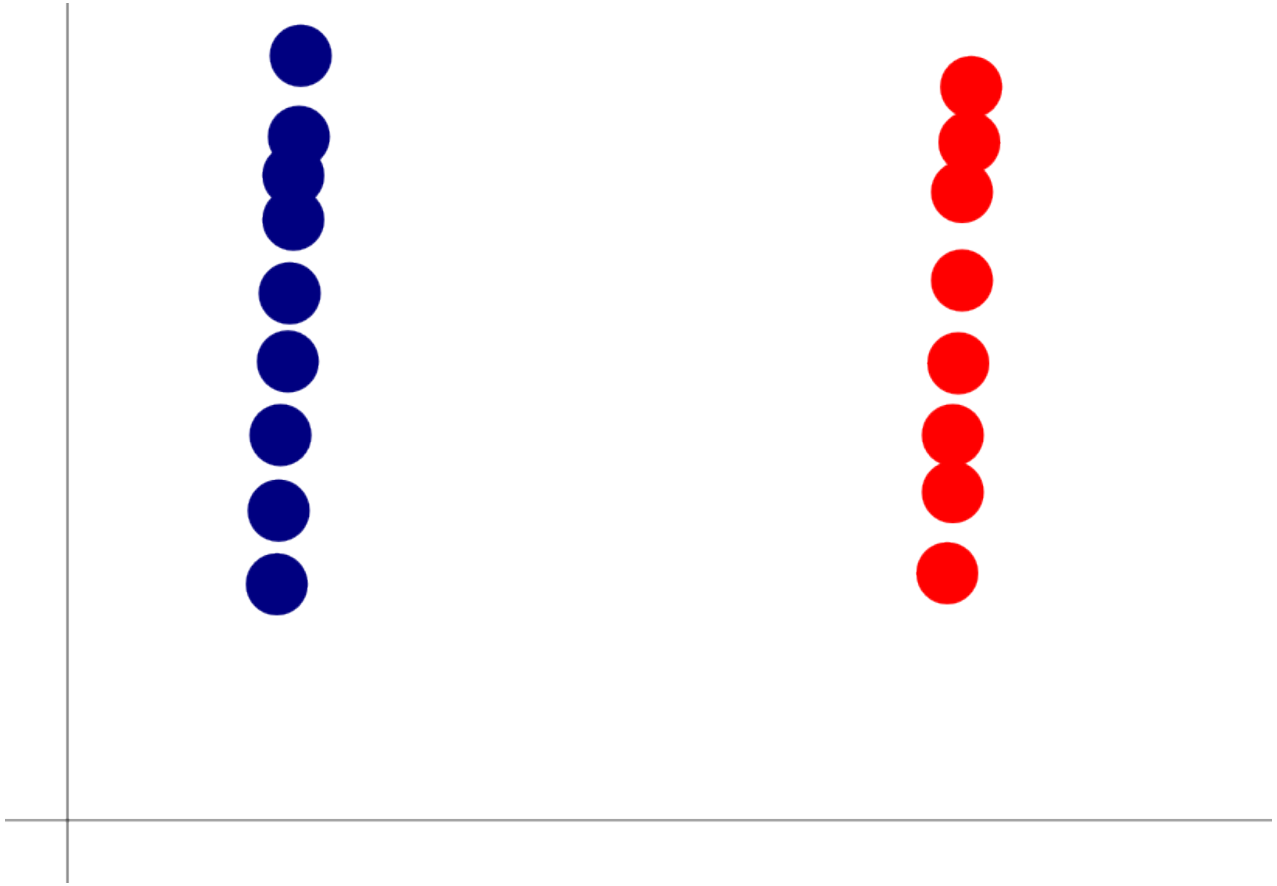
Our approach

Combine

1. Zero-variance discriminant analysis (to address singularity issues), with

2. ℓ_1-regularization for feature selection (to increase interpretability of discriminant vectors and consistency).

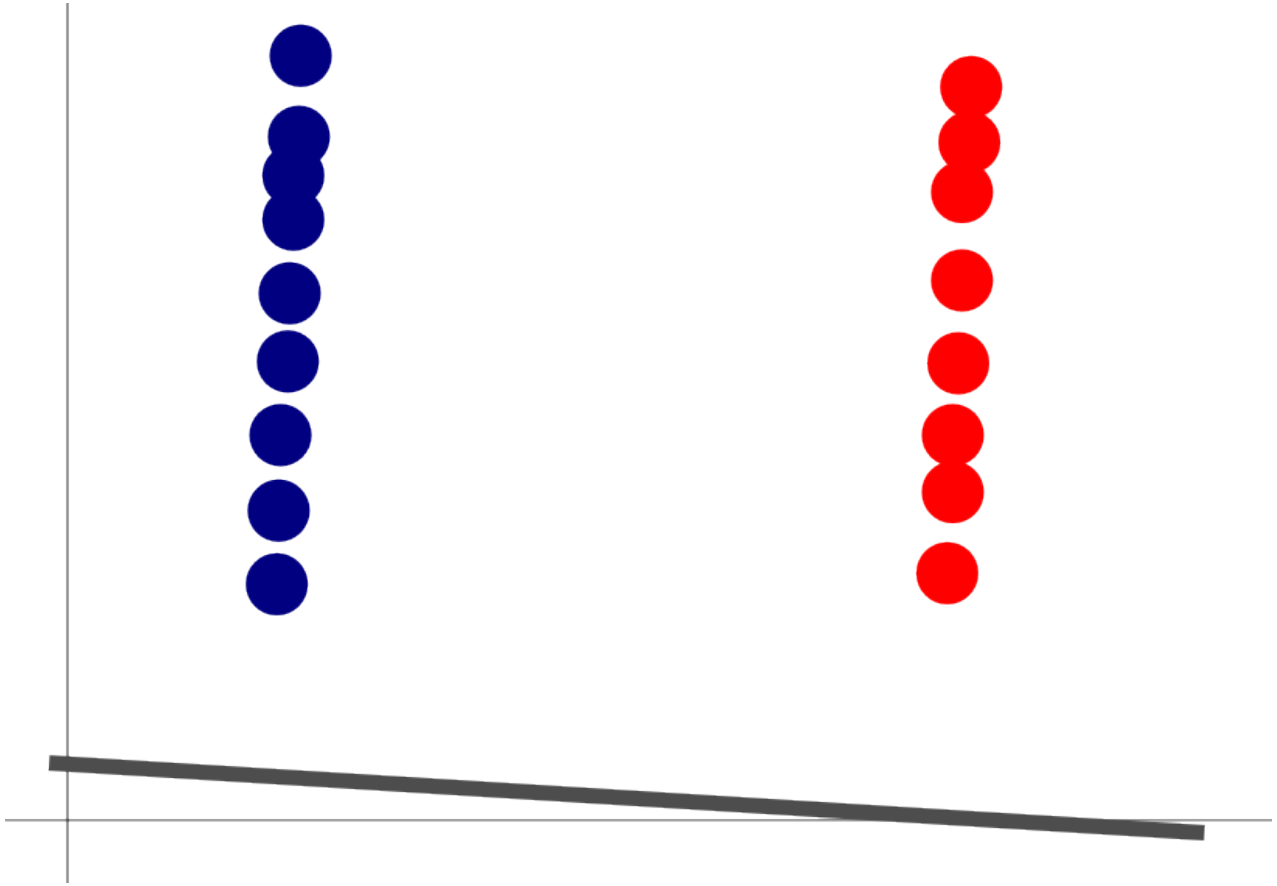
Zero-Variance Discriminant Analysis



Krzanowski et al. 1995: search for discriminant vectors in $\mathrm{Null}(\boldsymbol{W})$:

$$\boldsymbol{s}_i = \underset{\boldsymbol{s} \in \mathbf{R}^p}{\arg\max} \left\{ \boldsymbol{s}^T \boldsymbol{B} \boldsymbol{s} : \boldsymbol{W} \boldsymbol{s} = \mathbf{0},\ \boldsymbol{s}^T \boldsymbol{s} \leq 1,\ \boldsymbol{s}^T \boldsymbol{s}_j = 0\ \forall j < i \right\}$$

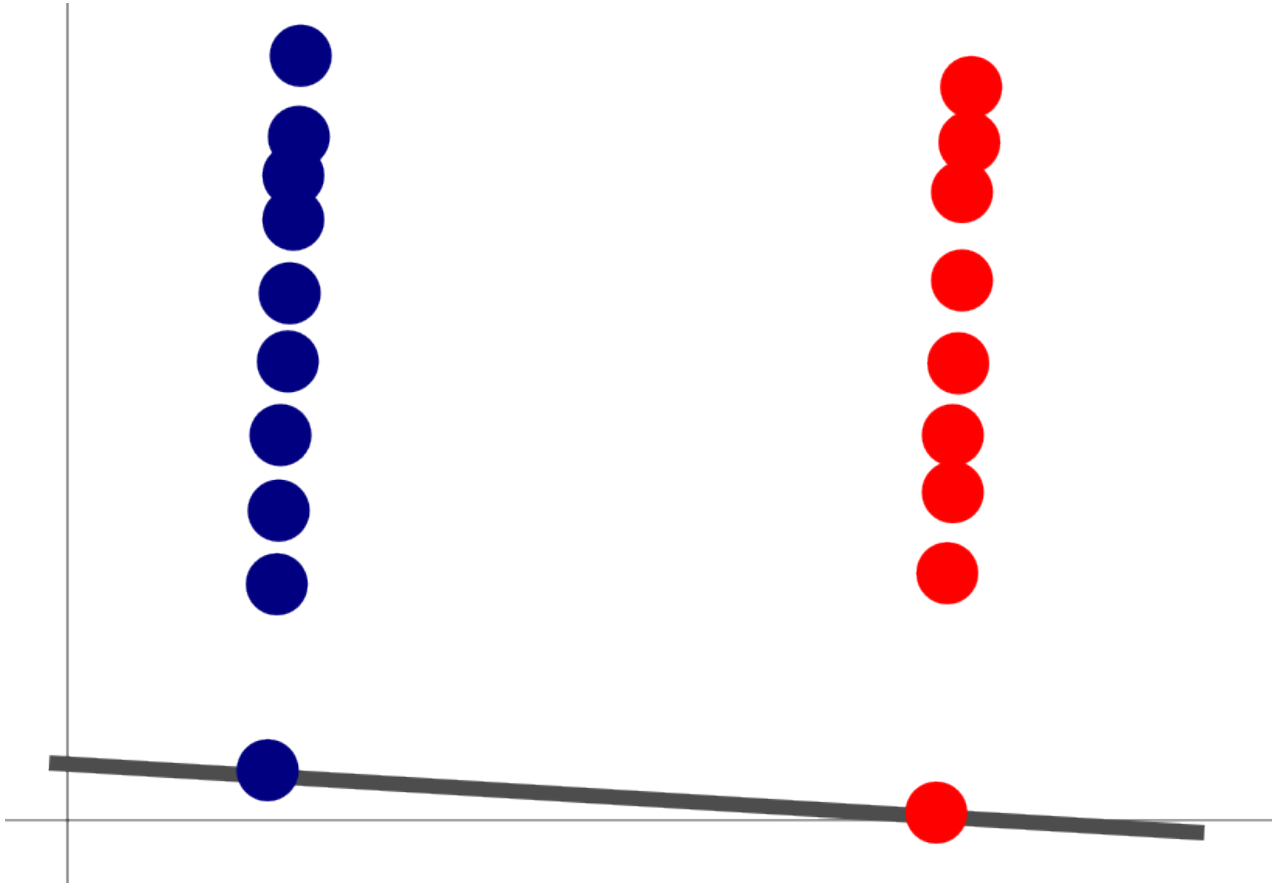
Zero-Variance Discriminant Analysis



Krzanowski et al. 1995: search for discriminant vectors in $\mathrm{Null}(\boldsymbol{W})$:

$$\boldsymbol{s}_i = \underset{\boldsymbol{s}\in\mathbf{R}^p}{\arg\max}\left\{\boldsymbol{s}^T\boldsymbol{B}\boldsymbol{s} : \boldsymbol{W}\boldsymbol{s} = \mathbf{0},\ \boldsymbol{s}^T\boldsymbol{s} \leq 1,\ \boldsymbol{s}^T\boldsymbol{s}_j = 0\ \forall\, j < i\right\}$$

Zero-Variance Discriminant Analysis



Krzanowski et al. 1995: search for discriminant vectors in $\mathrm{Null}(\boldsymbol{W})$:

$$\boldsymbol{s}_i = \underset{\boldsymbol{s}\in\mathbf{R}^p}{\arg\max}\left\{\boldsymbol{s}^T\boldsymbol{B}\boldsymbol{s} : \boldsymbol{W}\boldsymbol{s} = \mathbf{0},\ \boldsymbol{s}^T\boldsymbol{s} \leq 1,\ \boldsymbol{s}^T\boldsymbol{s}_j = 0\ \forall\, j < i\right\}$$

ZVD as Generalized Eigenproblem

Let columns of $\boldsymbol{N} \in \mathbf{R}^{p \times d}$ form an orthonormal basis of $\mathrm{Null}(\boldsymbol{W})$.

Perform change of variables $\boldsymbol{s} = \boldsymbol{N}\boldsymbol{x}$.

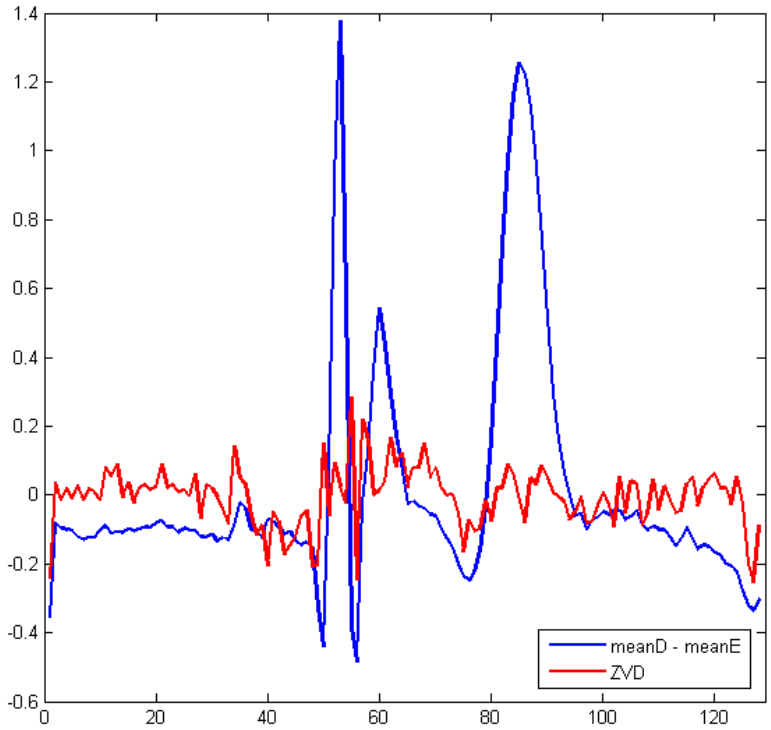
Problem reduces to finding nonzero eigenvectors of $\boldsymbol{N}^T\boldsymbol{B}\boldsymbol{N}$:

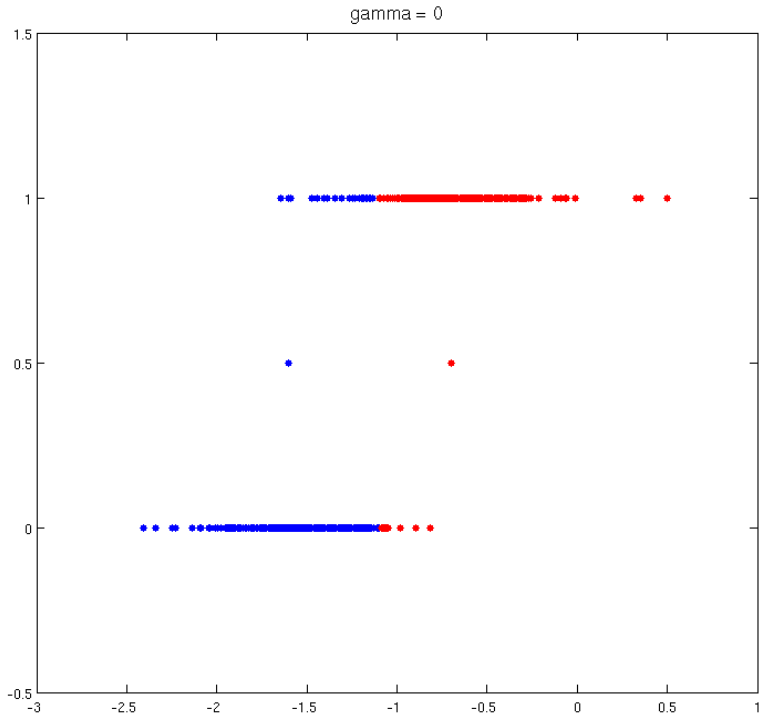
$$\boldsymbol{s}_i = \boldsymbol{N} \cdot \underset{\boldsymbol{x} \in \mathbf{R}^d}{\arg\max} \left\{ \boldsymbol{x}^T\boldsymbol{N}^T\boldsymbol{B}\boldsymbol{N}\boldsymbol{x} : \boldsymbol{x}^T\boldsymbol{x} \leq 1,\ \boldsymbol{x}^T\boldsymbol{N}^T\boldsymbol{s}_j = 0\ \forall\, j < i \right\}$$

Back to the EKG Data

Apply ZVD to the EKGFiveDays data:

- Test data: 428 signals from class D, 433 from class E.
- Assigned $\boldsymbol{x}$ to D if $\left|\boldsymbol{\mu}_D^T \boldsymbol{s}^* - \boldsymbol{x}^T \boldsymbol{s}^*\right| < \left|\boldsymbol{\mu}_E^T \boldsymbol{s}^* - \boldsymbol{x}^T \boldsymbol{s}^*.\right|$;





Penalized ZVD

We want to find sparse zero-variance discriminant vectors.

To do so, we repeatedly solve penalized generalized eigenproblems

$$\max_{\boldsymbol{s}\in\mathbf{R}^p}\left\{\frac{1}{2}\boldsymbol{s}^T\boldsymbol{B}\boldsymbol{s}-\gamma\rho(\boldsymbol{D}\boldsymbol{s}) : \boldsymbol{W}\boldsymbol{s}=\mathbf{0},\ \boldsymbol{s}^T\boldsymbol{s}\leq 1\right\},$$

where

- $\rho(\boldsymbol{x})=\sum\sigma_i|x_i|$ is weighted ℓ_1-norm,
- $\gamma>0$ is a regularization parameter tuning between the two objectives,
- $\boldsymbol{D}\in\mathcal{O}^p$ is a $p\times p$ orthogonal matrix.

A Change of Variables

Let $\boldsymbol{s} = \boldsymbol{N}\boldsymbol{x}$ where columns of $\boldsymbol{N} \in \mathbf{R}^{p \times d}$ form an orthonormal basis for $\mathrm{Null}(\boldsymbol{W})$:

$$\max_{\boldsymbol{x} \in \mathbf{R}^d} \left\{ \frac{1}{2} \boldsymbol{x}^T (\boldsymbol{N}^T \boldsymbol{B} \boldsymbol{N}) \boldsymbol{x} - \gamma \rho(\boldsymbol{D}\boldsymbol{N}\boldsymbol{x}) : \boldsymbol{x}^T \boldsymbol{x} \leq 1 \right\}$$

A Change of Variables

Let $\boldsymbol{s} = \boldsymbol{N}\boldsymbol{x}$ where columns of $\boldsymbol{N} \in \mathbf{R}^{p \times d}$ form an orthonormal basis for $\mathrm{Null}(\boldsymbol{W})$:

$$\max_{\boldsymbol{x} \in \mathbf{R}^d} \left\{ \frac{1}{2} \boldsymbol{x}^T (\boldsymbol{N}^T \boldsymbol{B} \boldsymbol{N}) \boldsymbol{x} - \gamma \rho(\boldsymbol{D}\boldsymbol{N}\boldsymbol{x}) : \boldsymbol{x}^T \boldsymbol{x} \leq 1 \right\}$$

Challenges:

- Nonconvex in general.
- ρ is not separable in $\boldsymbol{x}$, even if $\boldsymbol{D} = \boldsymbol{I}$.

A Change of Variables

Let $\boldsymbol{s} = \boldsymbol{N}\boldsymbol{x}$ where columns of $\boldsymbol{N} \in \mathbf{R}^{p \times d}$ form an orthonormal basis for $\mathrm{Null}(\boldsymbol{W})$:

$$\max_{\boldsymbol{x} \in \mathbf{R}^d} \left\{ \frac{1}{2} \boldsymbol{x}^T (\boldsymbol{N}^T \boldsymbol{B} \boldsymbol{N}) \boldsymbol{x} - \gamma \rho(\boldsymbol{D}\boldsymbol{N}\boldsymbol{x}) : \boldsymbol{x}^T \boldsymbol{x} \leq 1 \right\}$$

Challenges:

- Nonconvex in general.
- ρ is not separable in $\boldsymbol{x}$, even if $\boldsymbol{D} = \boldsymbol{I}$.

Solution: split variables as $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{N}\boldsymbol{x}$, solve equivalent problem

$$\max_{\boldsymbol{x} \in \mathbf{R}^d} \left\{ \frac{1}{2} \boldsymbol{x}^T (\boldsymbol{N}^T \boldsymbol{B} \boldsymbol{N}) \boldsymbol{x} - \gamma \rho(\boldsymbol{y}) : \boldsymbol{y}^T \boldsymbol{y} \leq 1, \ \boldsymbol{D}\boldsymbol{N}\boldsymbol{x} - \boldsymbol{y} = \mathbf{0} \right\}. \quad (\mathbf{SZVD})$$

Alternating Direction Method of Multipliers

Iterative method for solving

$$\min_{\boldsymbol{x},\boldsymbol{y}} \left\{ f(\boldsymbol{x}) + g(\boldsymbol{y}) : \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} = \boldsymbol{C} \right\}.$$

The general problem has augmented Lagrangian

$$L_\mu(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{z}) = f(\boldsymbol{x}) + g(\boldsymbol{y}) + \boldsymbol{z}^T(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} - \boldsymbol{C}) + \frac{\mu}{2}\|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} - \boldsymbol{C}\|^2$$

ADMM: each iteration minimize L_μ with respect to $\boldsymbol{x}$, then with respect to $\boldsymbol{y}$, then with respect to $\boldsymbol{z}$.

Converges at a linear rate under mild assumptions on $f, g, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$.

ADMM for Penalized ZVD

Let $\boldsymbol{A} := \boldsymbol{N}^T \boldsymbol{B} \boldsymbol{N}$.

We apply ADMM to the problem

$$\min_{\boldsymbol{x} \in \mathbf{R}^d, \boldsymbol{y} \in \mathbf{R}^p} \left\{ -\frac{1}{2} \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} + \gamma \rho(\boldsymbol{y}) : \boldsymbol{y}^T \boldsymbol{y} \leq 1, \ \boldsymbol{D} \boldsymbol{N} \boldsymbol{x} - \boldsymbol{y} = \boldsymbol{0} \right\}.$$

The augmented Lagrangian is

$$\begin{aligned} L_\mu(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \quad = \quad & -\frac{1}{2} \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} + \gamma \rho(\boldsymbol{y}) + \delta_{\{\boldsymbol{y} : \boldsymbol{y}^T \boldsymbol{y} \leq 1\}}(\boldsymbol{y}) \\ & + \boldsymbol{z}^T (\boldsymbol{D} \boldsymbol{N} \boldsymbol{x} - \boldsymbol{y}) + \frac{\mu}{2} \| \boldsymbol{D} \boldsymbol{N} \boldsymbol{x} - \boldsymbol{y} \|^2. \end{aligned}$$

Note: “mild” assumptions are not met by our problem (not convex).

Updating $\boldsymbol{x}$

Suppose we have iterate $(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k)$.

Take

$$\boldsymbol{x}^{k+1} = \arg\min_{\boldsymbol{x} \in \mathbf{R}^d} \frac{1}{2} \boldsymbol{x}^T (\mu \boldsymbol{I} - \boldsymbol{A}) \boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{N}^T \boldsymbol{D}^T (\boldsymbol{z}^k - \mu \boldsymbol{y}^k).$$

This is a convex quadratic for μ large enough.

Update $\boldsymbol{x}$ as the solution of the linear system

$$(\mu \boldsymbol{I} - \boldsymbol{A}) \boldsymbol{x}^{k+1} = \boldsymbol{N}^T \boldsymbol{D}^T (\mu \boldsymbol{y}^k - \boldsymbol{z}^k).$$

Updating $\boldsymbol{y}$

We next update $\boldsymbol{y}$ as

$$\boldsymbol{y}^{k+1} = \arg\min_{\boldsymbol{y}^T\boldsymbol{y}\leq 1} \gamma \sum_{i=1}^{p} \sigma_i |y_i| + \frac{\mu}{2}\|\boldsymbol{y} - (\boldsymbol{D}\boldsymbol{N}\boldsymbol{x}^{k+1} - \boldsymbol{z}^k/\mu)\|^2.$$

Update for $\boldsymbol{y}$ has closed form given by soft thresholding followed by an appropriate normalization:

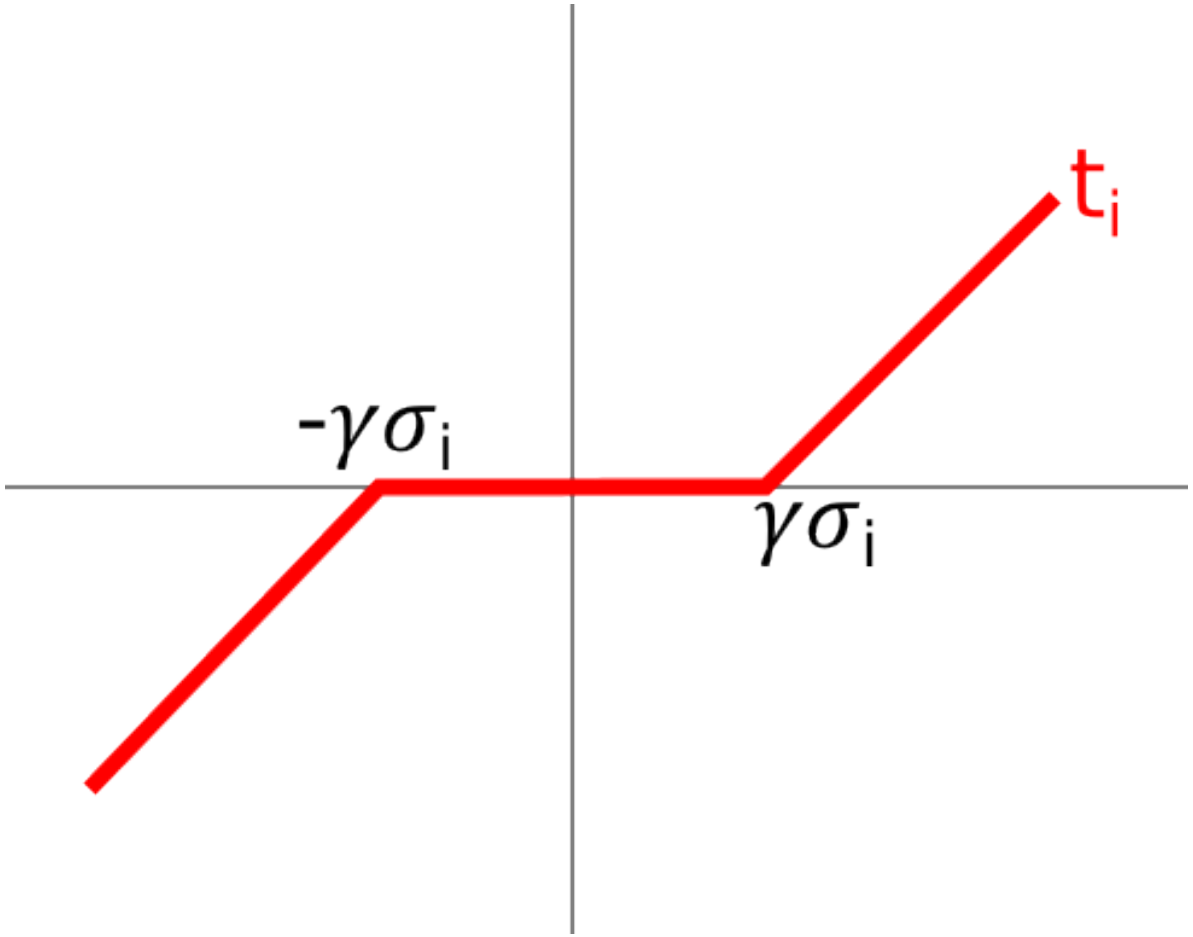
- Let $\boldsymbol{t} \in \mathbf{R}^p$ be given by

$$t_i = \mathrm{sign}(b_i) \cdot \max\{|b_i| - \gamma\sigma_i, 0\},$$

  where $\boldsymbol{b} = \boldsymbol{D}\boldsymbol{N}\boldsymbol{x}^{k+1} - \boldsymbol{z}^k/\mu$.
- Then take

$$\boldsymbol{y}^{k+1} = \frac{\boldsymbol{t}}{\mu + \max\{0, \|\boldsymbol{t}\| - \mu\}}.$$

The Soft-Thresholding Operator



$$t_i = \begin{cases} b_i - \gamma\sigma_i, & \text{if } b_i > \gamma\sigma_i \\ 0, & \text{if } -\gamma\sigma_i \leq b_i \leq \gamma\sigma_i \\ -b_i + \gamma\sigma_i, & \text{if } b_i < -\gamma\sigma_i. \end{cases}$$

Updating $\boldsymbol{z}$

We complete the iteration by updating $\boldsymbol{z}$ by approximate gradient step (dual ascent):

$$\boldsymbol{z}^{k+1} = \boldsymbol{z}^k + \mu(\boldsymbol{DNx}^{k+1} - \boldsymbol{y}^{k+1}).$$

Stop when surrogates for primal and dual feasibility gaps are sufficient small:

$$\|\boldsymbol{DNx}^k - \boldsymbol{y}^k\| < tol \qquad \mu\|\boldsymbol{y}^k - \boldsymbol{y}^{k-1}\| < tol.$$

Convergence

Theorem (Ames and Hong 2014):

- Suppose columns of $[\boldsymbol{DN}, \boldsymbol{C}]$ form an orthonormal basis for $\mathbf{R}^p$.
- Suppose the sequence of iterates generated by this ADMM algorithm satisfy

$$\boldsymbol{C}^T(\boldsymbol{z}^{k+1} - \boldsymbol{z}^k) = \mathbf{0}$$

  for all k, and μ large enough that L_μ is convex in each of $\boldsymbol{x}, \boldsymbol{y}$.
- Then $\{(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k)\}_{k=0}^{\infty}$ converges to a stationary point of **(SZVD)**.

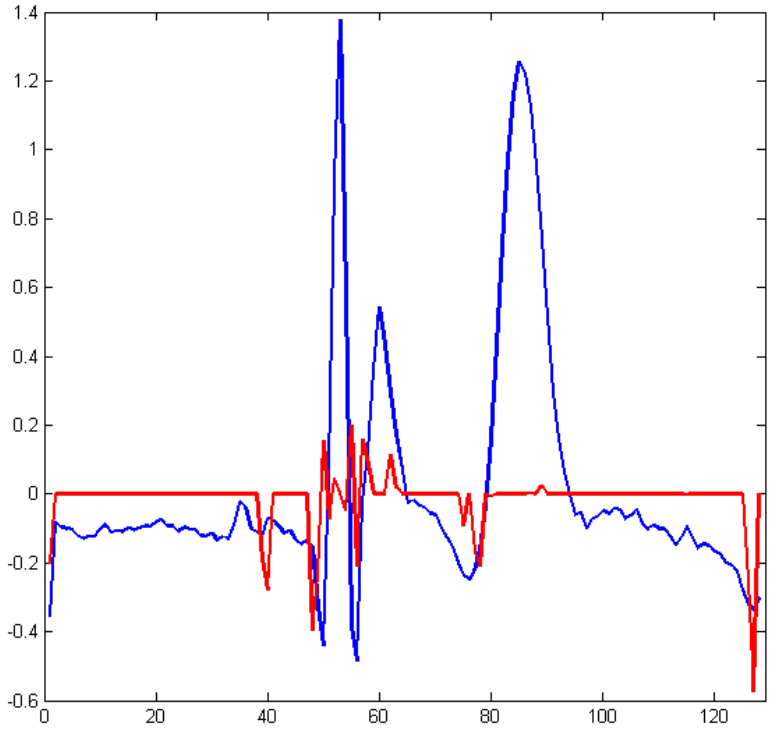
Holds if $[\boldsymbol{DN}, \boldsymbol{C}]$ forms standard basis, or $\boldsymbol{N} = \boldsymbol{I}$ (sparse PCA case).

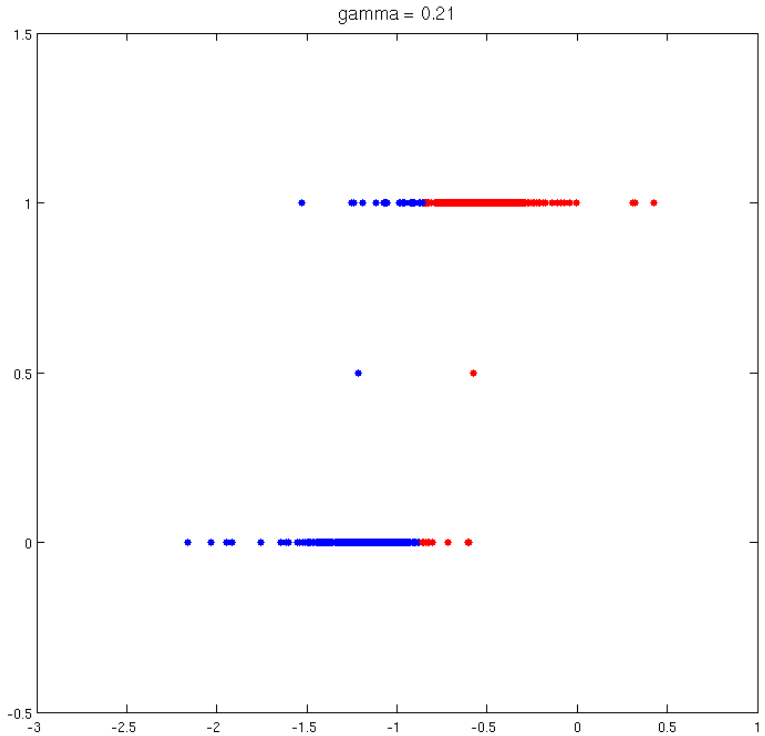
Applying SZVD to the EKG data

γ	$\|\boldsymbol{s}^*\|_0$	$\|\boldsymbol{s}^*\|_1$	$(\boldsymbol{\mu}_D - \boldsymbol{\mu}_E)^T \boldsymbol{s}^*$	misclassified
0	116	8.2659	0.89998	0.041812
0.02	104	7.6835	0.89704	0.044135
0.04	86	7.0615	0.88737	0.045296
0.06	71	6.5048	0.87219	0.051103
0.08	59	6.0422	0.85422	0.051103
0.10	49	5.6578	0.83432	0.04878
0.12	45	5.3861	0.81685	0.051103
0.14	40	5.064	0.79168	0.051103
0.16	33	4.486	0.73512	0.046458
0.18	24	3.9582	0.67231	0.038328
0.20	21	3.7551	0.64339	0.039489
0.21	21	3.7221	0.63938	0.038328

Applying SZVD to the EKG data (2)

The discriminant vector for the $\gamma = 0.21$ case:





Synthetic Data

For $k \in \{2, 4\}$, $r = 0.1, 0.5, 0.9$, and $p = 500$ sample $(25k, 25k, 250k)$ training, validation, and test observations as follows.

For each $i = 1, 2, \ldots, k$, sample $(25, 25, 250)$ observations from Normal distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where

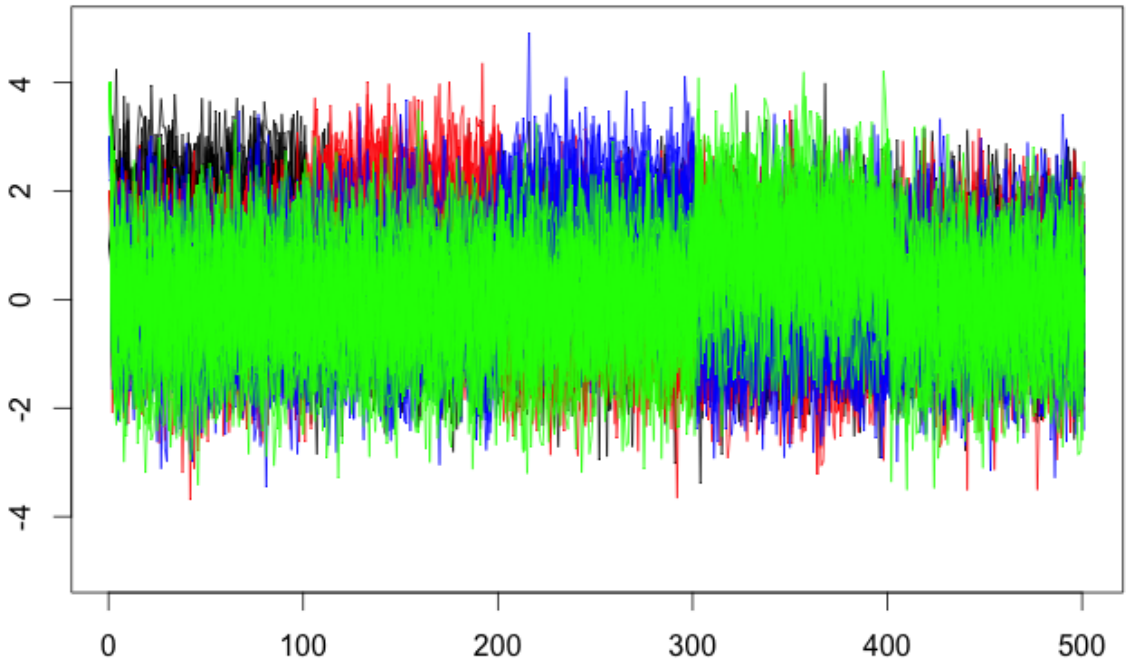
$$[\mu_i]_j = \begin{cases} 0.7, & \text{if } 100(i-1)+1 \leq j \leq 100i \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\Sigma_{k\ell} = \begin{cases} 1, & \text{if } k = \ell \\ r, & \text{otherwise.} \end{cases}$$

Synthetic Data (2)

Plotting training data:



Signals differ only in blocks of 100 features.

Synthetic Data (3)

Compare performance of **SZVD** with **ZVD**, penalized linear discriminant analysis **PL1/PFL** of Witten and Tibshirani 2011 and sparse discriminant analysis **SDA** of Clemmensen et al. 2011.

Tune any regularization parameters using validation: choose value maximizing

$$\frac{\#\text{misclassified}}{\#\text{validation obs}} + \frac{1}{2}\left(\frac{\#\text{feat}}{p(k-1)}\right).$$

Experiment repeated 20 times for each (k, r) pair.

Synthetic Data (3)

(k, r)		ZVD	SZVD	PL1	PFL	SDA
$k = 2$	*Err*	0 (0)	1.6 (1.9)	12.2 (17.6)	11 (26.7)	17.0 (14.7)
$r = 0.1$	*Feat*	490.9 (2.7)	104.2 (10.5)	113.2 (33.4)	178.1 (31.9)	57.8 (8.2)
	Time	0.10 (0.004)	1.1 (0.08)	0.02 (0.002)	0.03 (0.002)	2.04 (0.025)
$k = 2$	*Err*	0 (0)	0.1 (0.447)	86.55 (49.55)	58.6 (57.62)	56.9 (46.09)
$r = 0.5$	*Feat*	488.7 (3.8)	112.7 (10.3)	139.0 (44.0)	184.3 (44.5)	51.4 (4.7)
	Time	0.10 (0.003)	1.1 (0.06)	0.02 (0.001)	0.04 (0.003)	2.0 (0.01)
$k = 2$	*Err*	0 (0)	0 (0)	100.2 (68.5)	95.8 (71.7)	115.2 (58.0)
$r = 0.9$	*Feat*	485.5 (3.3)	143.8 (9.4)	171.4 (60.6)	160.3 (50.7)	49.2 (0.9)
	Time	0.10 (0.004)	1.1 (0.051)	0.03 (0.004)	0.04 (0.005)	2.01 (0.02)
$k = 4$	*Err*	0.6 (0.7)	23.3 (16.7)	115.2 (36.4)	116.6 (23.5)	53.1 (38.3)
$r = 0.1$	*Feat*	1474.0 (3.8)	305.0 (66.1)	402.6 (109.0)	428.4 (59.1)	313.7 (211.1)
	Time	0.19 (0.003)	5.8 (0.4)	0.13 (0.01)	0.21 (0.02)	14.3 (1.4)
$k = 4$	*Err*	0.1 (0.2)	30.9 (28.6)	318.4 (40.3)	307.6 (43.7)	39.5 (32.2)
$r = 0.5$	*Feat*	1475.1(3.8)	239.8 (73.7)	369.9 (61.1)	388.9 (81.5)	369.3 (194.9)
	Time	0.20 (0.01)	6.1 (0.3)	0.14 (0.01)	0.23 (0.02)	14.1 (1.4)
$k = 4$	*Err*	0 (0)	0 (0)	412.0 (55.9)	415.5 (60.5)	1.0 (4.5)
$r = 0.9$	*Feat*	1471.1(6.2)	336.0 (136.2)	386.2 (100.4)	369.7 (92.4)	371.3 (288.7)
	Time	0.20 (0.01)	6.0 (0.4)	0.15 (0.02)	0.23 (0.02)	14.0 (2.5)

Application to Generalized Eigenproblems

We can apply our ADMM heuristic to approximately solve **any** penalized eigenproblem

$$\max_{\boldsymbol{x} \in \mathbf{R}^d} \left\{ \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} - \gamma \rho(\boldsymbol{x}) : \boldsymbol{x}^T \boldsymbol{x} \leq 1 \right\}$$

If ρ is not a weighted ℓ_1-norm penalty, y may not have closed form updates.

Update y by solving a convex optimization problem.

For example, gives a simple heuristic for Sparse Principal Component Analysis (SPCA).

Principal Component Analysis

Given data set $\boldsymbol{X} = [\boldsymbol{x}_1 | \boldsymbol{x}_2 | \boldsymbol{x}_3 | \ldots | \boldsymbol{x}_n]$.

Principal component analysis (PCA): identify orthogonal coordinate system such that variance with respect to each coordinate is ordered decreasingly.

- Most variance is in direction of the first component $\boldsymbol{w}_1$, next most in the direction of the second component $\boldsymbol{w}_2$, and so on.

If data is centered, reduces to solving eigenproblems of the form

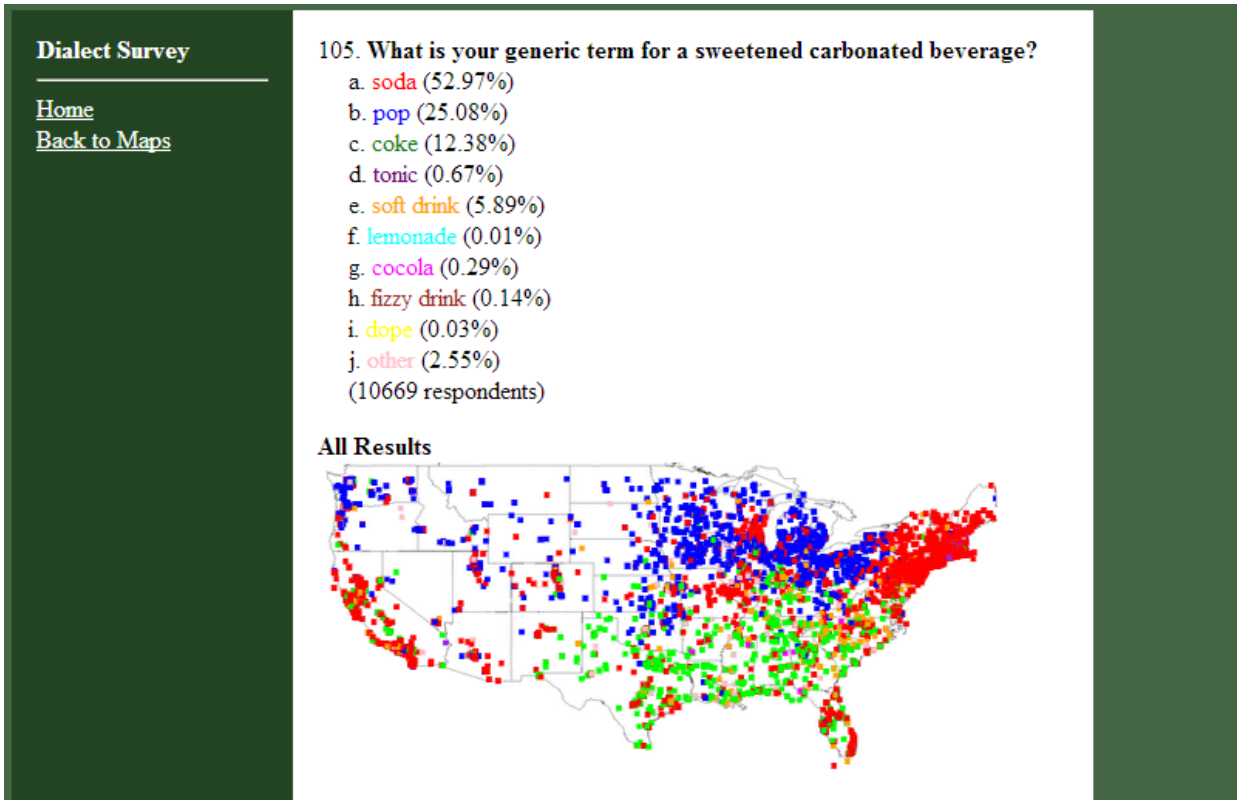
$$\boldsymbol{w}_i = \max_{\boldsymbol{w} \in \mathbf{R}^p} \left\{ \boldsymbol{w}^T \boldsymbol{X} \boldsymbol{X}^T \boldsymbol{w} : \boldsymbol{w}^T \boldsymbol{w} \leq 1,\ \boldsymbol{w}^T \boldsymbol{w}_j = 0\ \forall j < i \right\}.$$

The Harvard Dialect Survey

Researchers asked series of vocabulary and pronunciation questions to map regional dialects.

- Results available at dialect.redlog.net
- New survey can be taken at www.tekstlab.uio.no/cambridge_survey/

The Harvard Dialect Survey



Users responses encoded as a binary vector:

- if $\boldsymbol{x}$ answered (c) for Question 105, we have

$$x_{105c} = 1, \qquad x_{105a} = x_{105b} = \cdots = x_{105j} = 0$$

Visualizing the data

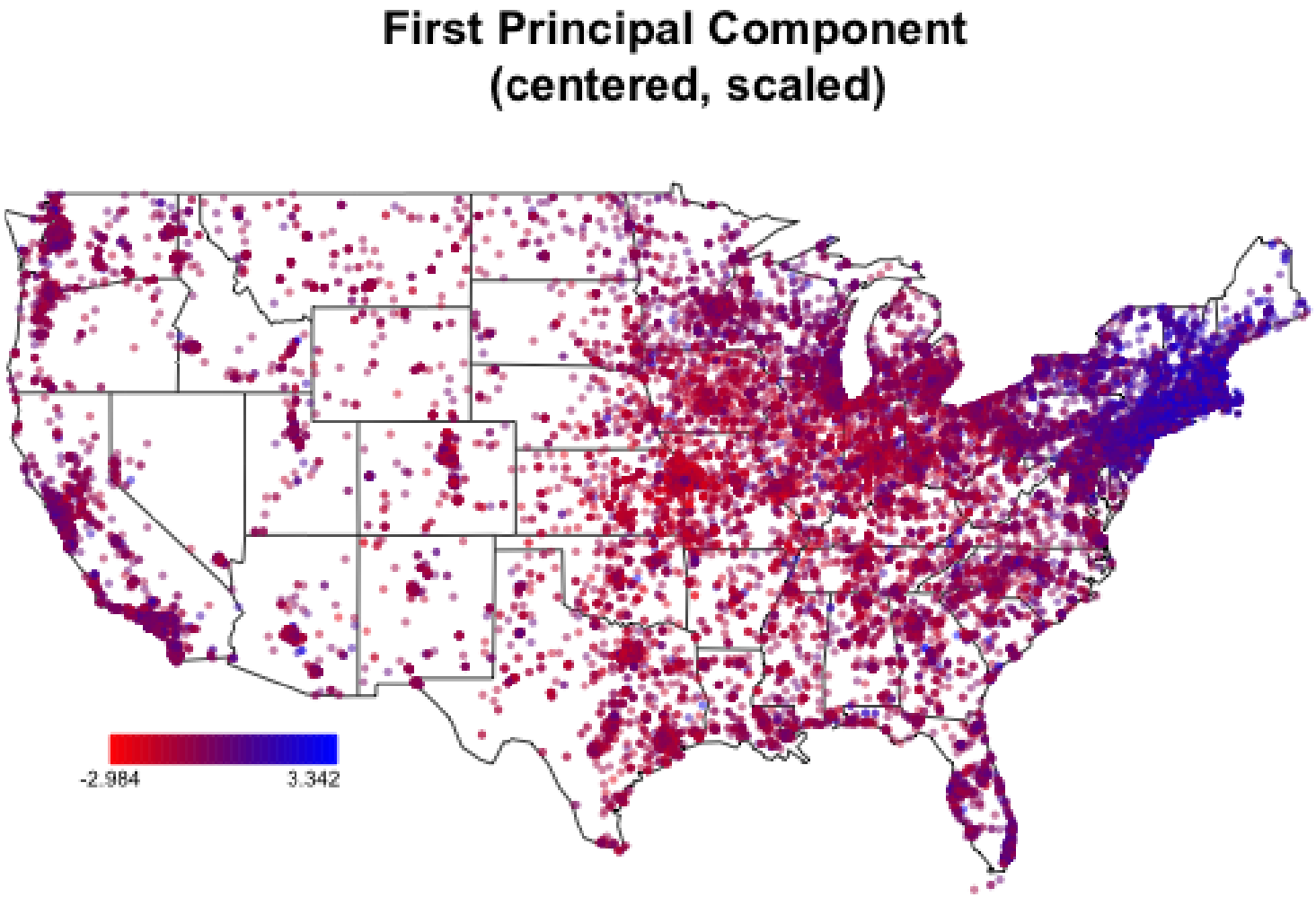
Data is encoded as 468 dimensional binary vectors.

How can we use this information to detect regional dialects?

Use dimensionality reduction via PCA to visualize the data.

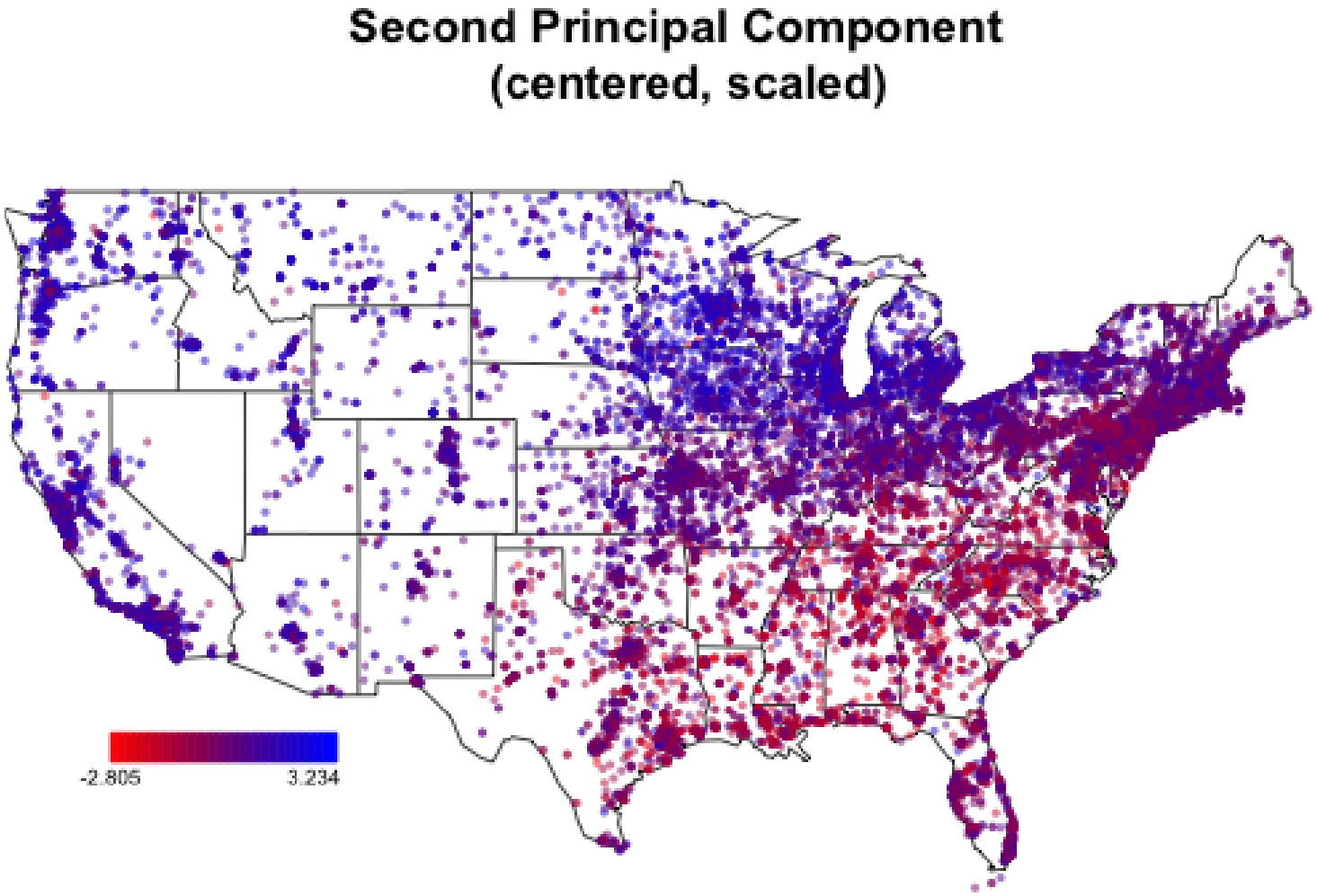
Projection onto the first principal component

Each dot represents a user $\boldsymbol{x}$.
Intensity represents value of $\boldsymbol{w}_1^T \boldsymbol{x}$



Projection onto the second principal component

Intensity represents value of $\boldsymbol{w}_2^T \boldsymbol{x}$



Interpretation

First PC $\boldsymbol{w}_1$ detects regional differences between the Northeast and Midwest/rest of country.

Largest magnitude entries of $\boldsymbol{w}_1$ tell us what questions are important to making this distinction:

- “sneakers” vs “tennis shoes” (73), “soda” vs “pop” (105), “sunshower” vs “no term for that” (80)

Second PC $\boldsymbol{w}_2$ detects differences between the Southeast and rest of country:

- “Kitty-corner” vs “catty-corner” (76), “drinking” vs “water fountain” (103), “you guys” vs ”y’all” (50)

Sparse Principal Component Analysis

In the high-dimension, small sample size setting, the sample covariance matrix $\boldsymbol{X}\boldsymbol{X}^T/n$ is an inconsistent approximation of the population covariance matrix.

Need to regularize to improve performance of the estimator.

Choose sparsity via the ℓ_1-norm as regularizer to induce more interpretable principal components:

$$\boldsymbol{w}_i = \max_{\boldsymbol{w} \in \mathbf{R}^p} \left\{ \boldsymbol{w}^T \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{w} - \gamma \|\boldsymbol{w}\|_1 : \boldsymbol{w}^T \boldsymbol{w} \leq 1,\ \boldsymbol{w}_i^T \boldsymbol{w}_j = 0\ \forall\, j < i \right\}.$$

Sparse PCA for the Dialect data set

Applying our ADMM heuristic, we can approximately solve for the first few penalized principal component vectors.

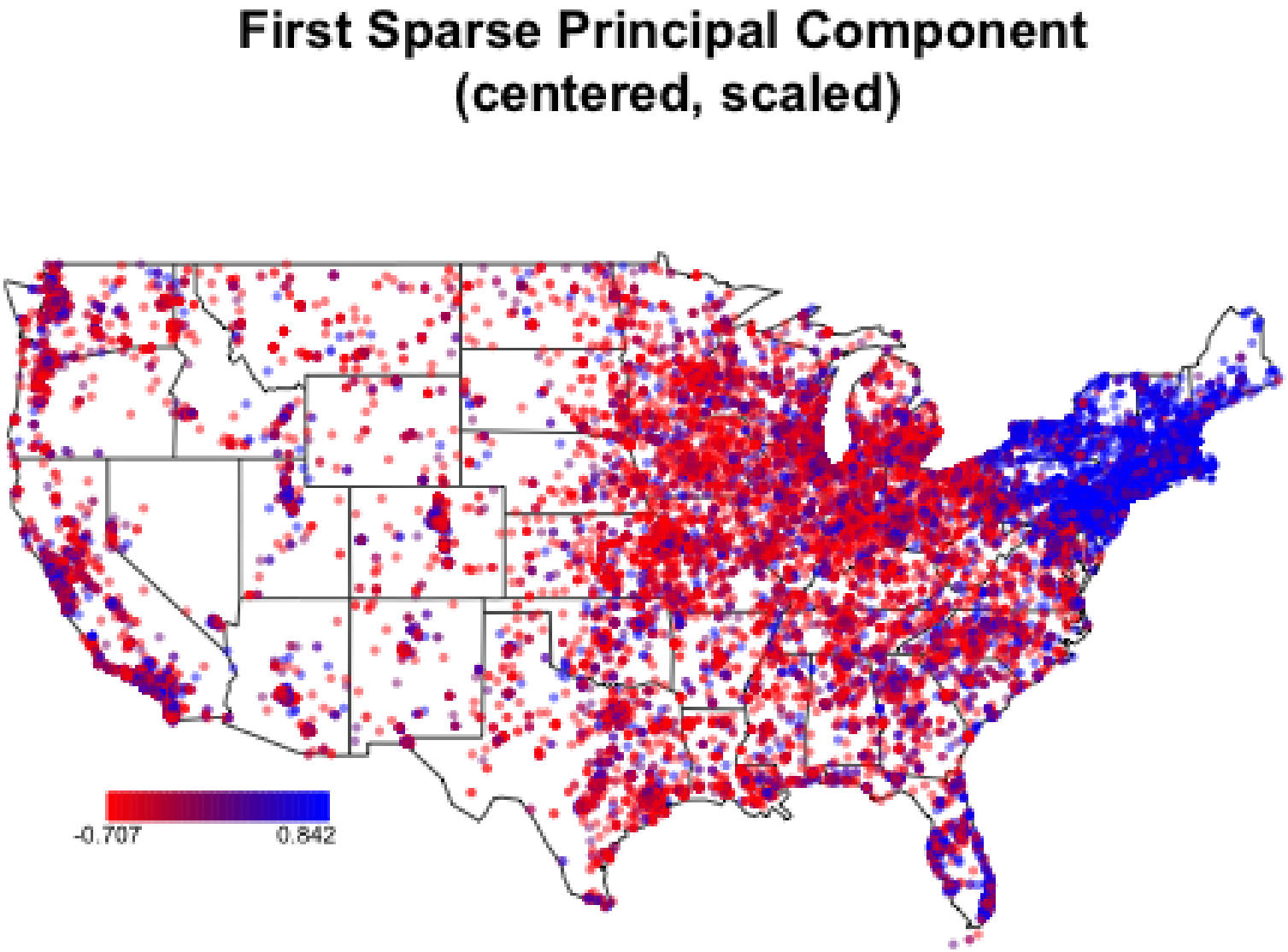
$\boldsymbol{w}_1$ has **3** nonzero entries out of **468**.

- ”Sneakers vs. Tennis shoes” (Q73(a), Q73(f)),
- ”Soda?” (Q105(a)).

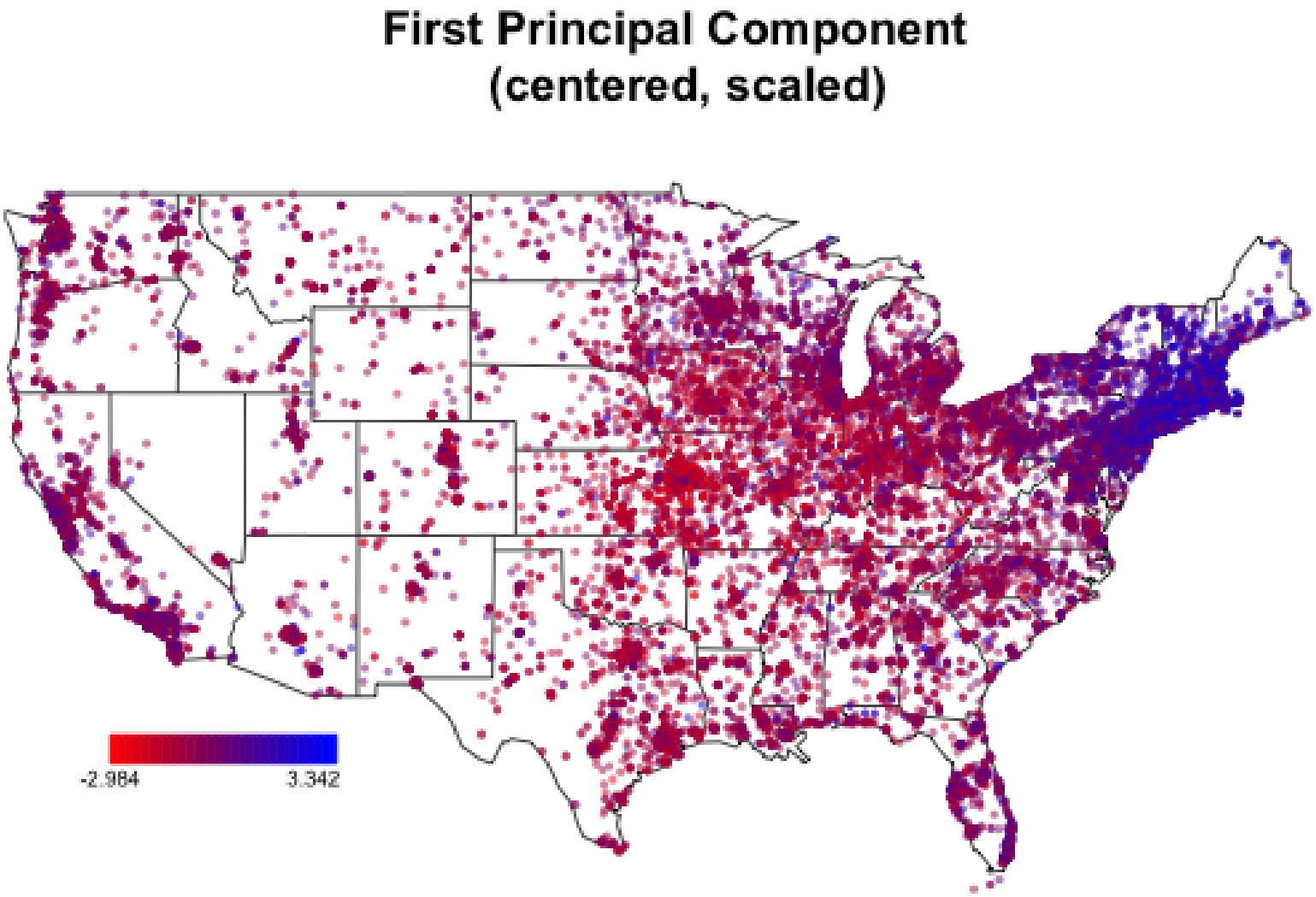
$\boldsymbol{w}_2$ has **8** nonzero entries:

- All variants of do you call your grandparents “Grandpa/Grandma” or “Something else”?

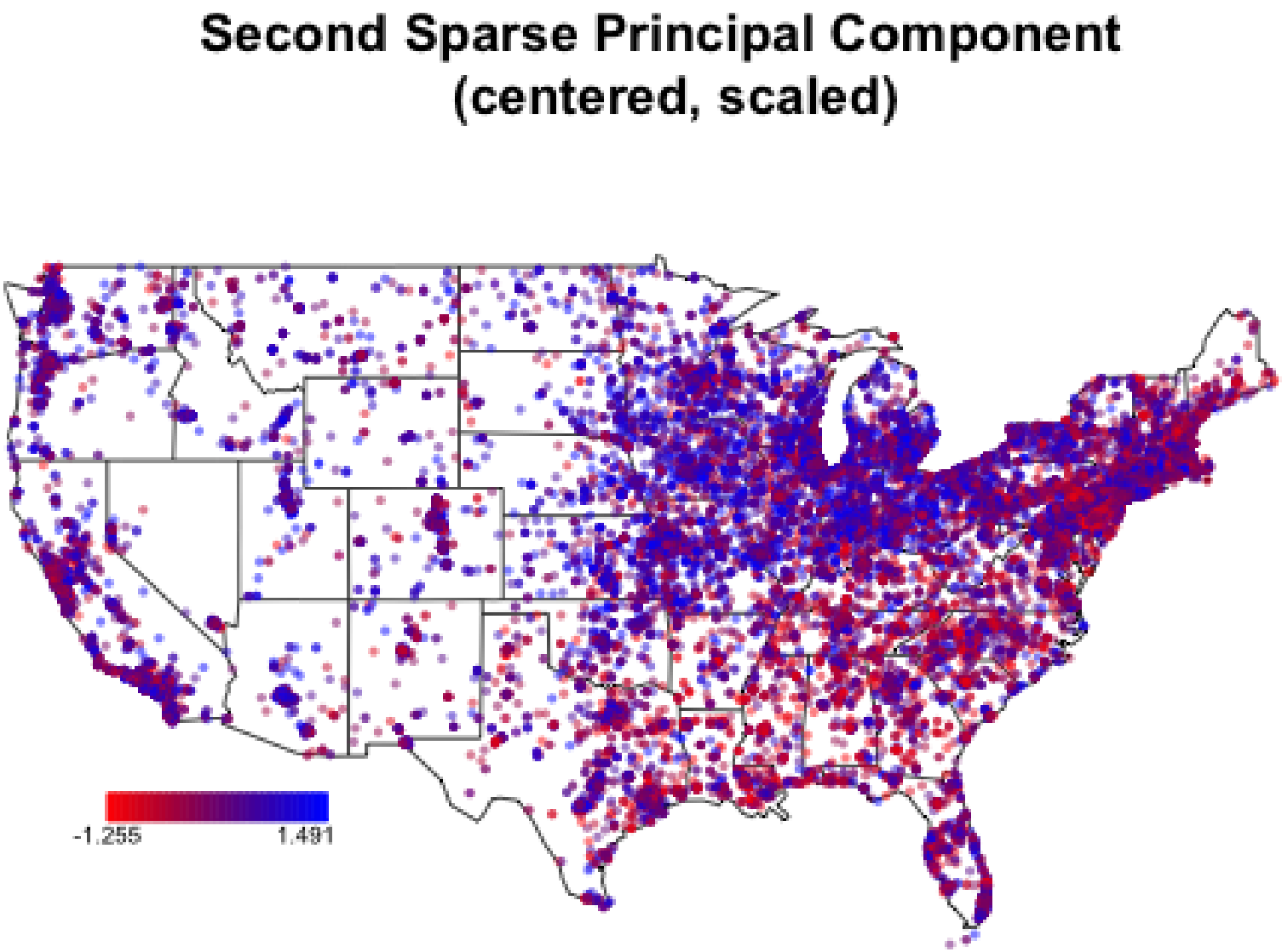
Projection onto the 1st Sparse PC



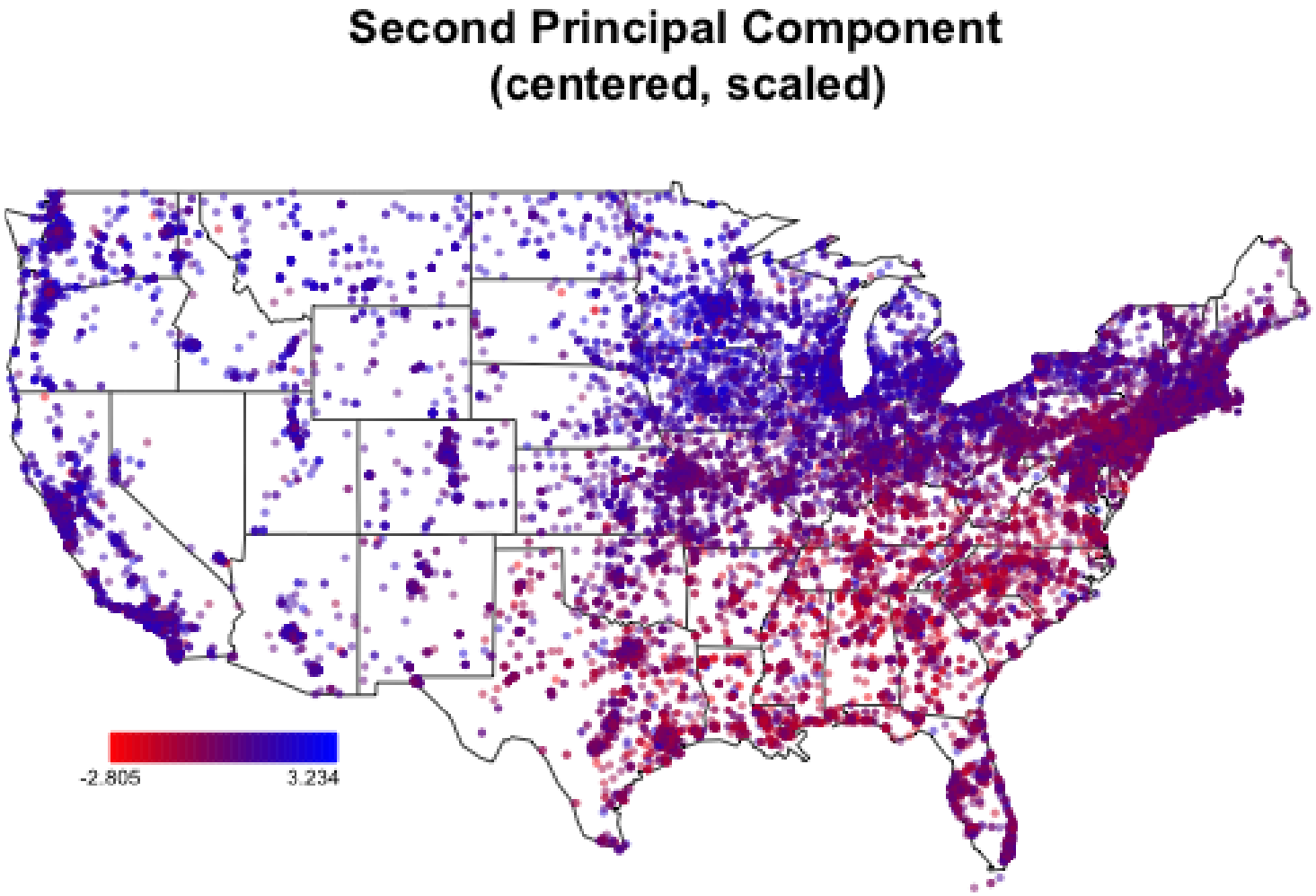
Projection onto the 1st Sparse PC



Projection onto the 2nd Sparse PC



Projection onto the 2nd Sparse PC



Open problems

Recoverability:

- Conditions under which “optimal” solutions obtained from relaxation?
- Can we derive bounds on classification error?

Numerical Linear Algebra issues: solution of large scale eigenvalue decomposition, linear system solving are main bottlenecks.

Application to similar problems? (optimal scoring formulations of LDA?)

Thank you!

Preprint available on arxiv:
arxiv.org/abs/1401.5492

Matlab and R implementations:
bpames.people.ua.edu/software.html

Research supported in part by the Institute for Mathematics and its Applications with funds from the NSF, and by Joel Tropp under ONR award N00014-11-1002.