Presentation to the Functional Data in New York (FDNY) working group

Linear Conditioning for Clustering Functional Data

Thaddeus Tarpey

Wright State University

September 18, 2013

Acknowledgements

This presentation is a "work-in-progress" and has benefited from discussions with Eva Petkova (NYU), and Todd Ogden (Columbia), Phil Reiss, Huaihou Chen, and Zhe Su at NYU and from Felix Essuman-Nelson at Wright State University.

This work is supported by grants R01 MH095836 and R01 MH099003 from the National Institute of Mental Health (NIMH)

Underlying Motivation for this Work

- Discover biosignatures for placebo (non-specific) response.
- Challenge: Drug treated patients can respond due to placebo effects.
- Several approaches will be used – today's presentation will focus on clustering functional data.

Illustration: 6-week Trial Comparing Fluoxetine to Placebo

Fit linear mixed effects model to longitudinal data using quadratic trajectories.
Outcome = Hamilton Rating Scale for Depression (HRSD)

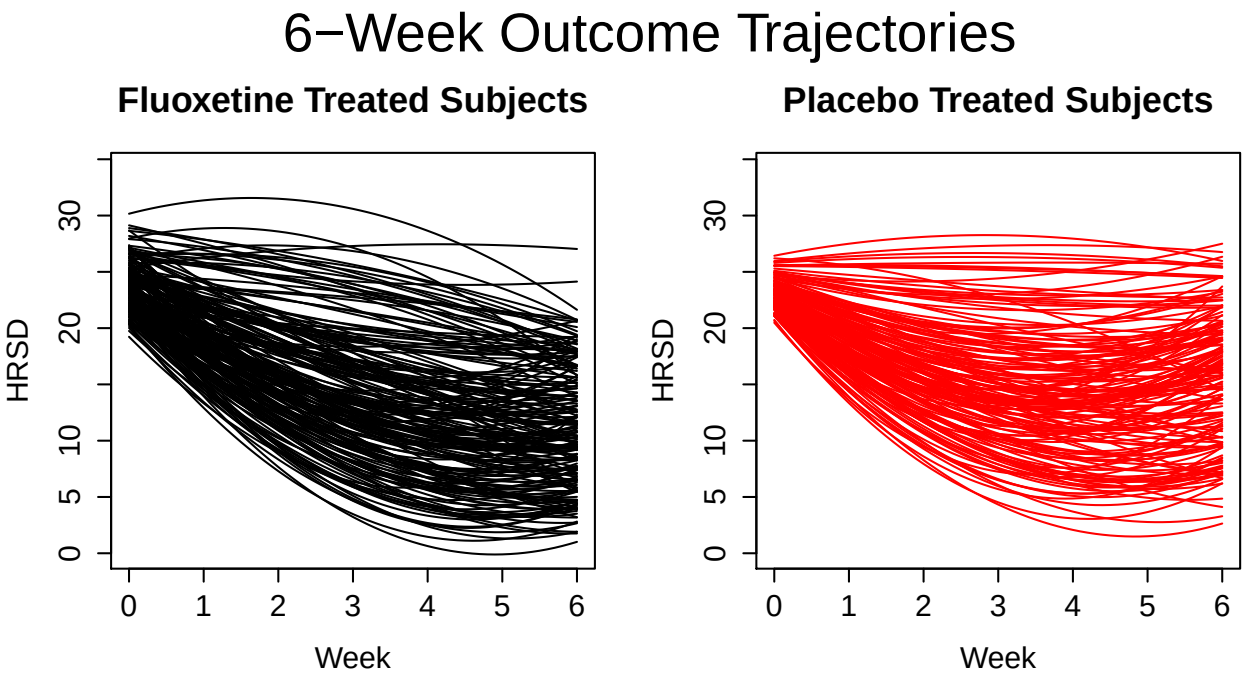
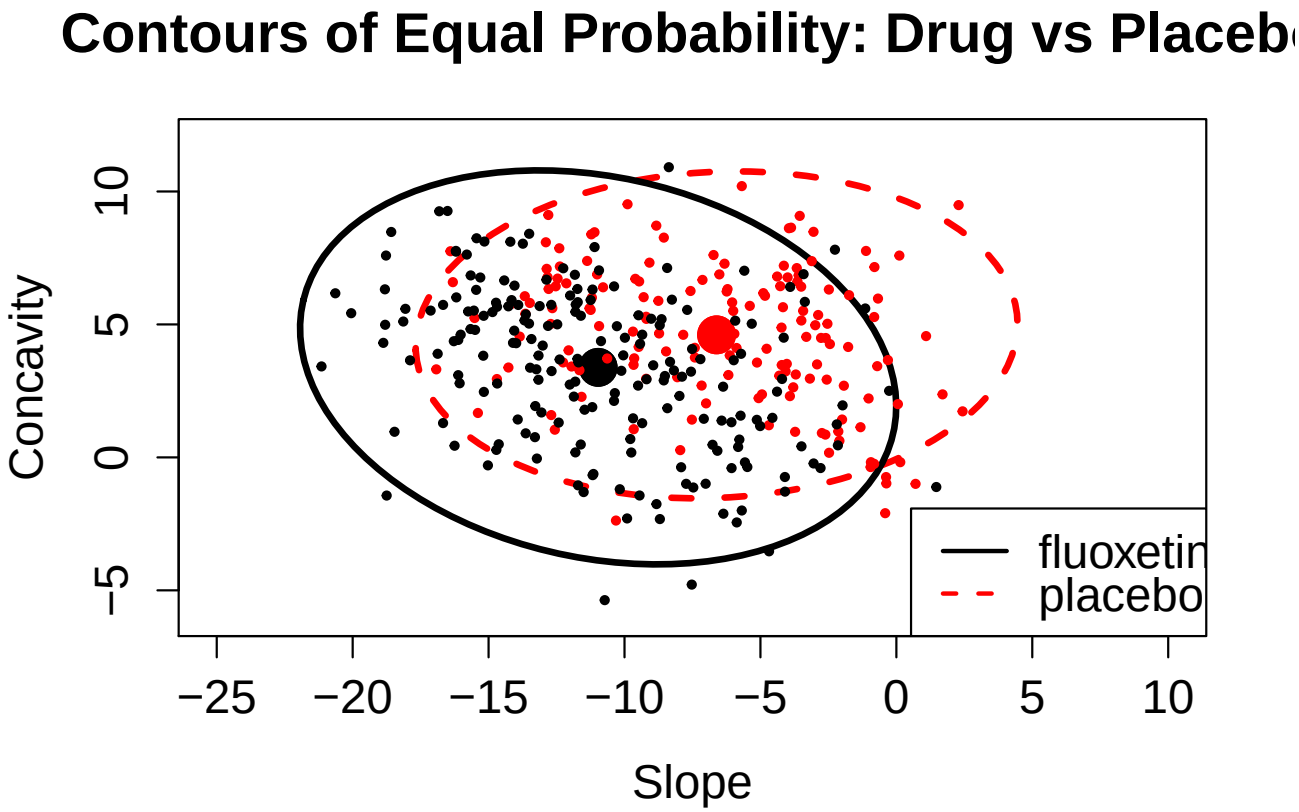


Illustration: 6-week Trial Comparing Fluoxetine to Placebo

Contours of equal density probability for fluoxetine and placebo treated subjects in coefficient space of trajectories.



Clustering Longitudinal Trajectories to Access Impact of Baseline Moderators

Setup: We have longitudinal trajectories for two or more treatment conditions.

Associate clusters with different types of outcomes: specific drug responder, placebo responder, non-responder.

Moderator Importance Plot: Plot the probability of cluster membership as a function of a baseline covariate x.

Linear Conditioning ... Background

"Preconditioning" is used in numerical linear algebra improve (e.g. speed up) algorithms to solve systems of equations $\boldsymbol{Ax} = \boldsymbol{b}$ when the columns of $\boldsymbol{A}$ are highly correlated (poorly conditioned).
Left multiply the equation by a matrix $\boldsymbol{C}$ to "pre-condition" the system and improve numerical algorithms: $\boldsymbol{CAx} = \boldsymbol{Cb}$.

The same idea can be applied to the traditional linear model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}.$$

Paul et al (2008) consider pre-conditioning only the outcome $\boldsymbol{y}$, by projecting it onto the top singular values of $\boldsymbol{X}$ before running the lasso.

Jia and Rohe (2013 preprint) propose "preconditioning" both sides of the equation via a linear transformation.

Clustering Functional Data

Functional Data: $\boldsymbol{y}_i, i = 1, \ldots, n$

Model: $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{X}_i \sim n_i \times p$ consisting of basis functions and ϵ_i a vector of random errors.

Goal: Cluster the functional trajectories

- Factor out the noise ($\boldsymbol{\epsilon}$) and cluster the coefficients ($\boldsymbol{\beta}_i$) instead of the raw data ($\boldsymbol{y}_i$).
- Data reduction: Work with a small number of regression coefficients instead of a potentially large vector of outcome values in $\boldsymbol{y}_i$.

Linear Conditioning for Clustering Functional Data

If $\boldsymbol{A}$ is a non-singular matrix, then
the model

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i,$$

is identical to

$$\boldsymbol{y}_i = [\boldsymbol{X}_i\boldsymbol{A}^{-1}][\boldsymbol{A}\boldsymbol{\beta}_i] + \boldsymbol{\epsilon}_i$$

- The linearly transformed design matrix $\boldsymbol{X}_i\boldsymbol{A}^{-1}$ can be regarded as a change in the basis representation of the functional data.

Cluster results based on the original $\boldsymbol{\beta}_i$ can differ dramatically from cluster results using $\boldsymbol{A}\boldsymbol{\beta}_i$.

Goal: Determine a linear transformation $\boldsymbol{A}$ to optimize the clustering.

Types of Linear "Conditioning" for Clustering

- Feature Selection: Multiply by a diagonal matrix with 0 or 1's on the main diagonal.
- Clustering derivatives: Derivatives often correspond to linear transformation of the coefficient vector. Often a major source of variability in functional data is attributable to variation in intercepts which may not be of interest - differentiating the functions gets rid of this variation.
- Principal Component Analysis – cluster first few PC's.
- Independent Component Analysis (ICA) transformations - steer the clustering algoirthm in the direction, not of primary variance (PCA), but in "non-normal" directions.
- Weighting, e.g. standardizing the variables to unit variance.

Linear Conditioning Example: A Weighting Function

Chen et al. (2013) propose a weight function $w^2(t)$ to improve functional data analysis methods, such as clustering.

Let $y_i(t)$ and $y_{i'}(t)$ denote to functional data points. Functional clustering methods are often based on a distance metric between functions, such as L^2 distance:

$$\|y_i(t) - y_{i'}(t)\|^2 = \int (y_i(t) - y_{i'}(t))^2 dt.$$

Proposal: use a weighted L^2 distance:

$$\|y_i(t) - y_{i'}(t)\|_w^2 = \int w^2(t)(y_i(t) - y_{i'}(t))^2 dt$$

Linear Conditioning Example: A Weighting Function continued ...

Using a basis representation of the functional data $y_i(t) = \sum_{j=1}^{p} z_{ij}\alpha_j(t)$, it follows that

$$\|y_i(t) - y_{i'}(t)\|_w^2 = (\boldsymbol{z}_i - \boldsymbol{z}_{i'})'\boldsymbol{A}(\boldsymbol{z}_i - \boldsymbol{z}_{i'}),$$

where $\boldsymbol{A} = [a_{jj'}]$ with

$$a_{jj'} = \int \alpha_j(t)\alpha_{j'}(t)w^2(t)dt.$$

Thus, weighted L^2 clustering corresponds to clustering the linearly transformed coefficients

$$A^{1/2}\boldsymbol{z}_i,$$

where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})'$ is the coefficient vector.

Canonical Linear Transformation for Clustering

Idea: Stretch the data (via a linear transformation) that separates the groups as much as possible.

From **Canonical Discriminant Analysis**: Transform in order to maximize the between cluster variability relative to the within cluster variability.
$\boldsymbol{W}$ = within cluster covariance matrix
$\boldsymbol{B}$ = between cluster covariance matrix

$$\text{Total Variance} = \boldsymbol{\Psi} = \boldsymbol{W} + \boldsymbol{B}$$

Simultaneously diagonalize $\boldsymbol{W}$ and $\boldsymbol{B}$:

$$\boldsymbol{W}^{-1/2}\boldsymbol{B}\boldsymbol{W}^{-1/2} = \boldsymbol{H}\boldsymbol{D}\boldsymbol{H}' \text{ (=Spectral Decomposition)},$$

where $\boldsymbol{H}$ is orthogonal and $\boldsymbol{D}$ is diagonal.
Let

$$\boldsymbol{\Gamma} = \boldsymbol{W}^{-1/2}\boldsymbol{H}$$

Columns of $\boldsymbol{\Gamma}$ correspond to the canonical transformation for clustering.

Canonical Linear Transformation for Clustering continued ...

Clustering coefficient vectors $\boldsymbol{\beta}_i$ for functional data: The covariance matrix of $\boldsymbol{\Gamma}'\boldsymbol{\beta}_i$ is

$$\boldsymbol{\Gamma}'\boldsymbol{\Psi}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}'(\boldsymbol{W} + \boldsymbol{B})\boldsymbol{\Gamma} = \boldsymbol{I} + \boldsymbol{D}.$$

Inflate the between-cluster variability relative to the within-cluster, one can further transform using a *canonical* transformation for clustering

$$\boldsymbol{C}\boldsymbol{\Gamma}'\boldsymbol{\beta}_i$$

where

$$\boldsymbol{C} = \text{Stretching Matrix} = \text{diag}(c_1, c_2, \ldots, c_p)$$

- $c_j > 1$ stretches the distribution in the jth direction.
- $0 \leq c_j < 1$ constricts the distribution.

Illustration: $K = 3$ Finite Normal Mixture Simulation

- Primary variability direction does not coincide with the "between" group variability (denoted by the mixture component means)
- Marginal distributions show no indication of a mixture distribution.

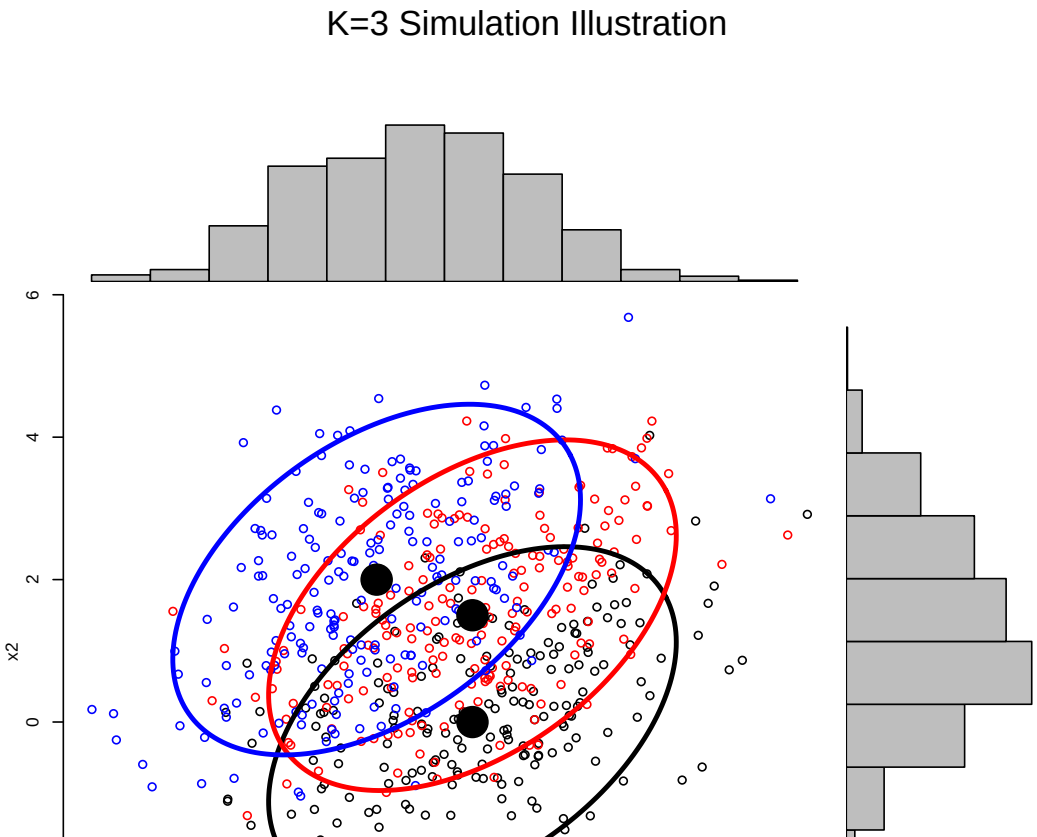


Illustration: $K = 3$ Finite Normal Mixture Simulation

Run the k-means algorithm: cluster means completely miss the true means.

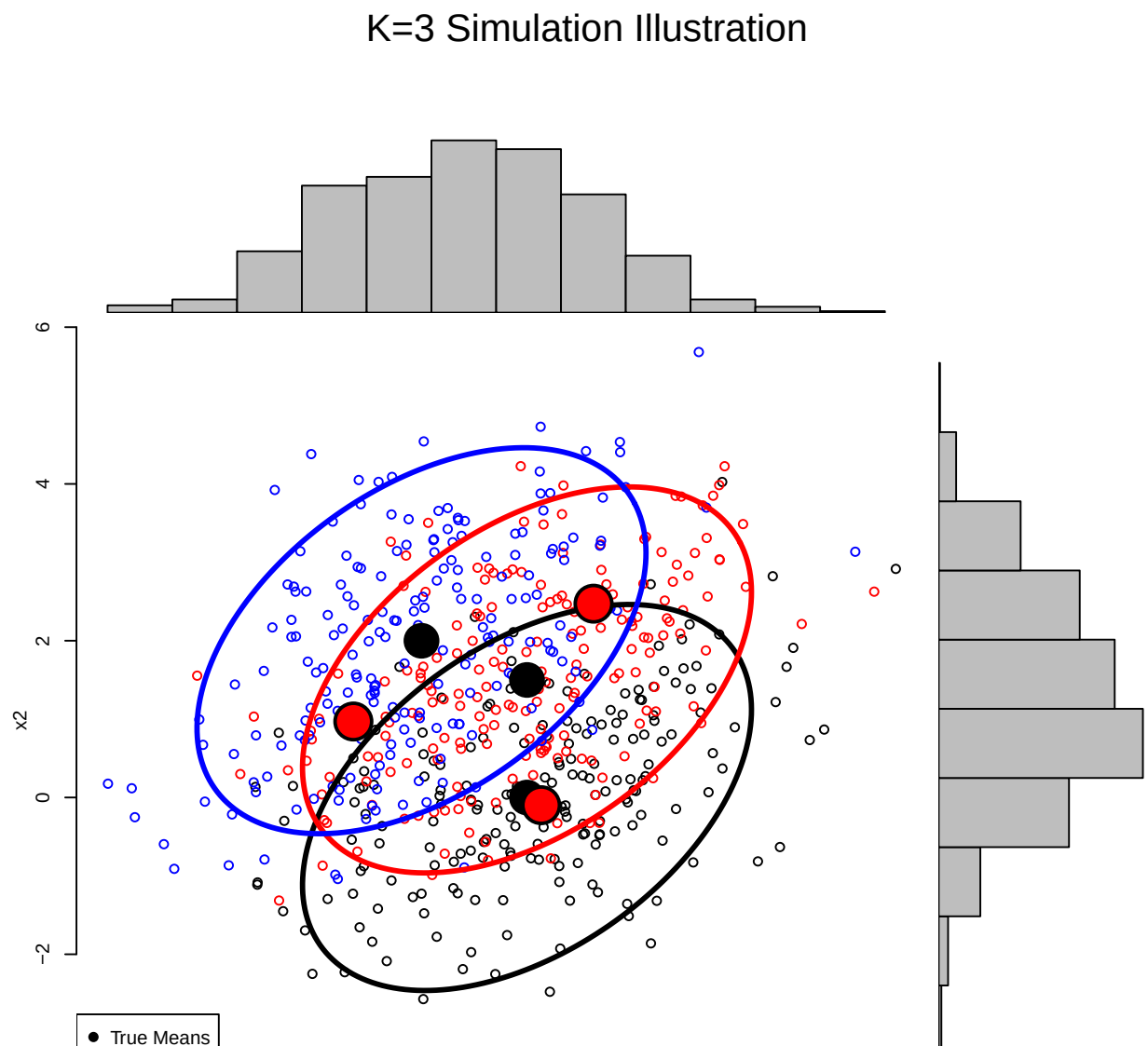
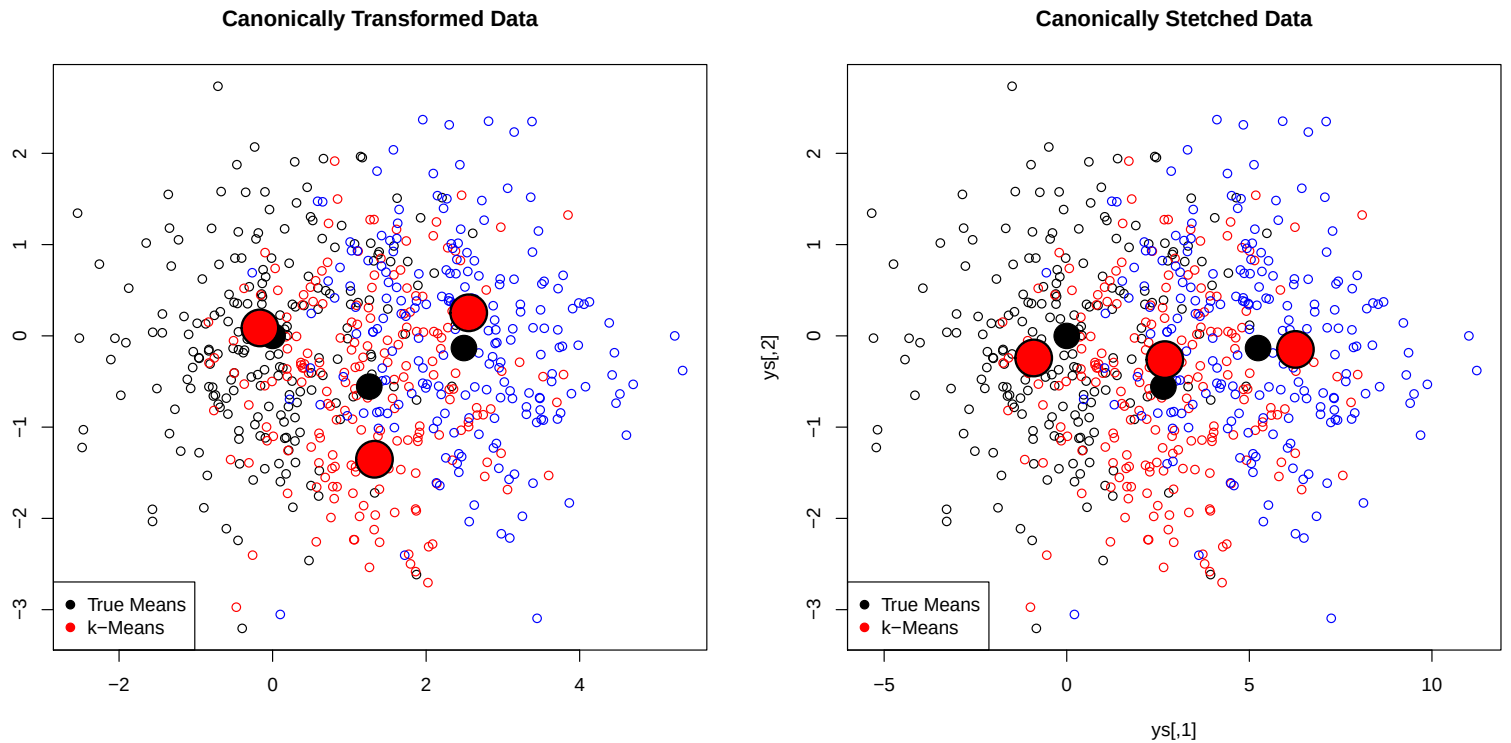


Illustration: $K = 3$ Finite Normal Mixture Simulation continued ...

Canonically transformed data (left panel) and stretched (right panel)



Clustering Quality Indices: R^2

Measure the proportion of total variability explained by the clustering.

Similar to the coefficient of determination R^2 in regression.

$$R^2 = 1 - \frac{\text{within sum-of-squares}}{\text{total sum-of-squares}}.$$

R^2 scale invariant – use it to compare clusterings resulting from different linear transformations of the data.

Clustering Quality Indices: Variation of Information (VI)

A metric comparing how well two-different clusterings of a data set "match-up" is the *variation of information* (Meilă 2007).

Given two clusterings of the same data, $\mathcal{C}_1$ and $\mathcal{C}_2$, let

$$P(j, j') = \frac{|C_j \cap C_{j'}|}{n},$$

for cluster C_j in $\mathcal{C}_1$ and cluster $C_{j'}$ in $\mathcal{C}_2$.

$$\text{Mutual Information} = I(\mathcal{C}_1, \mathcal{C}_2) = \sum_{j=1}^{k} \sum_{j'=1}^{k} P(j, j') \log\left(\frac{P(j, j')}{P_1(j) P_2(j')}\right).$$

$$\text{Entropy for Clustering } \mathcal{C} = H(\mathcal{C}) = -\sum_{j=1}^{k} P(j) \log(P(j))$$

Clustering Quality Indices: Variation of Information (VI) continued ...

Variation of Information (VI)

$$VI(\mathcal{C}_1, \mathcal{C}_2) = H(\mathcal{C}_1) + H(\mathcal{C}_2) - 2I(\mathcal{C}_1, \mathcal{C}_2).$$

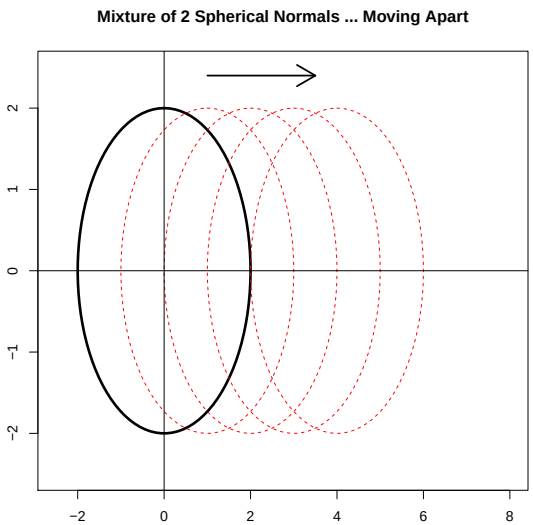
$VI = 0$ if the two clusterings produce identical clusters (up to a re-labeling); otherwise $VI > 0$.

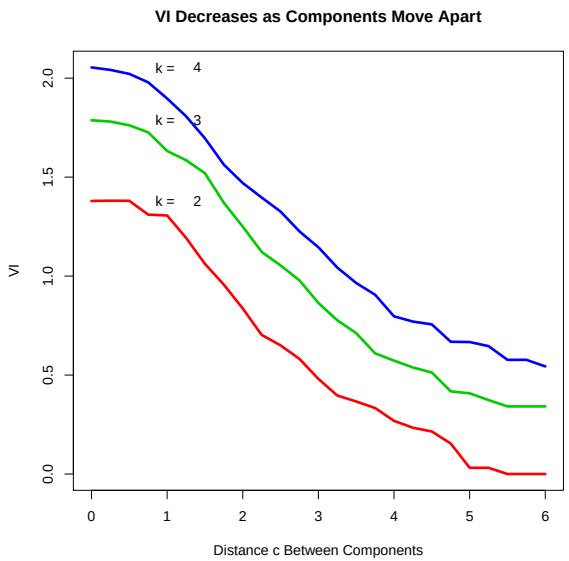
Variation of Information (VI) Illustration

Cluster data from 2 Spherical Bivariate Normal Distributions

$$N(\mathbf{0}, \boldsymbol{I}) \text{ and } N((c, 0)', \boldsymbol{I})$$

As the mixture components move apart ... VI decreases





Fluoxetine vs Placebo Trial Revisited

Determine a canonical transformation using the two treatments:

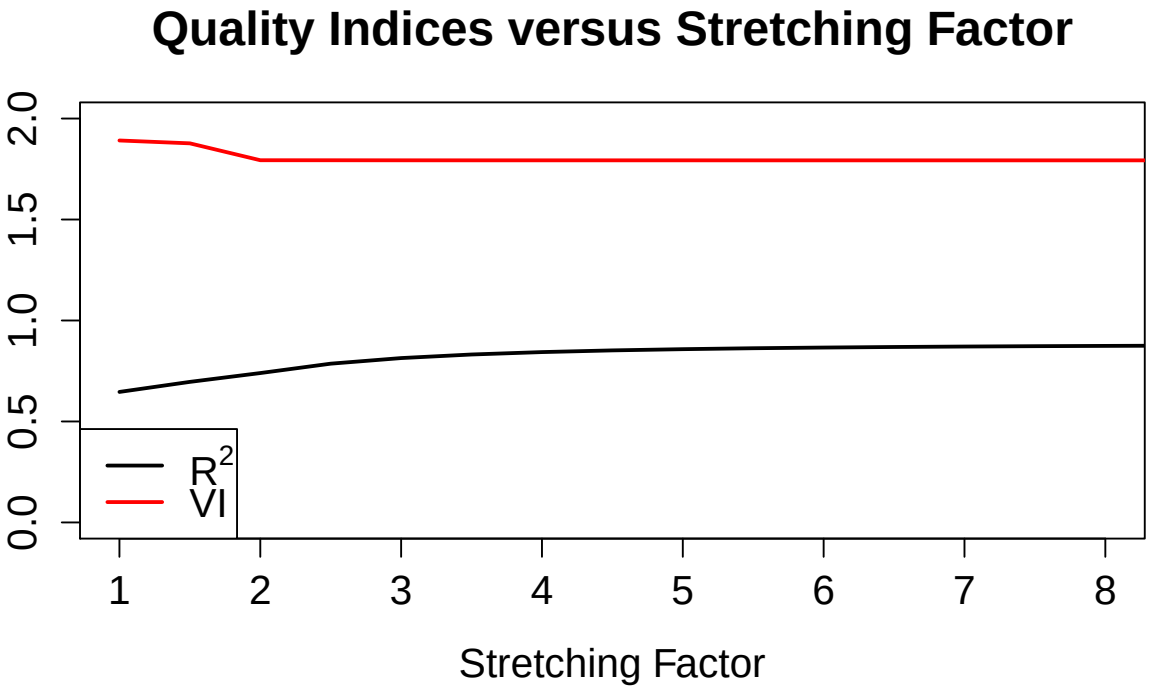
$\boldsymbol{B}$ = Between Group (Drug & Placebo) covariance matrix with two treatment groups

$\boldsymbol{W}$ = Within Group (Drug & Placebo) covariance matrix.

Use these two matrices to determine the canonical transformation for clustering.

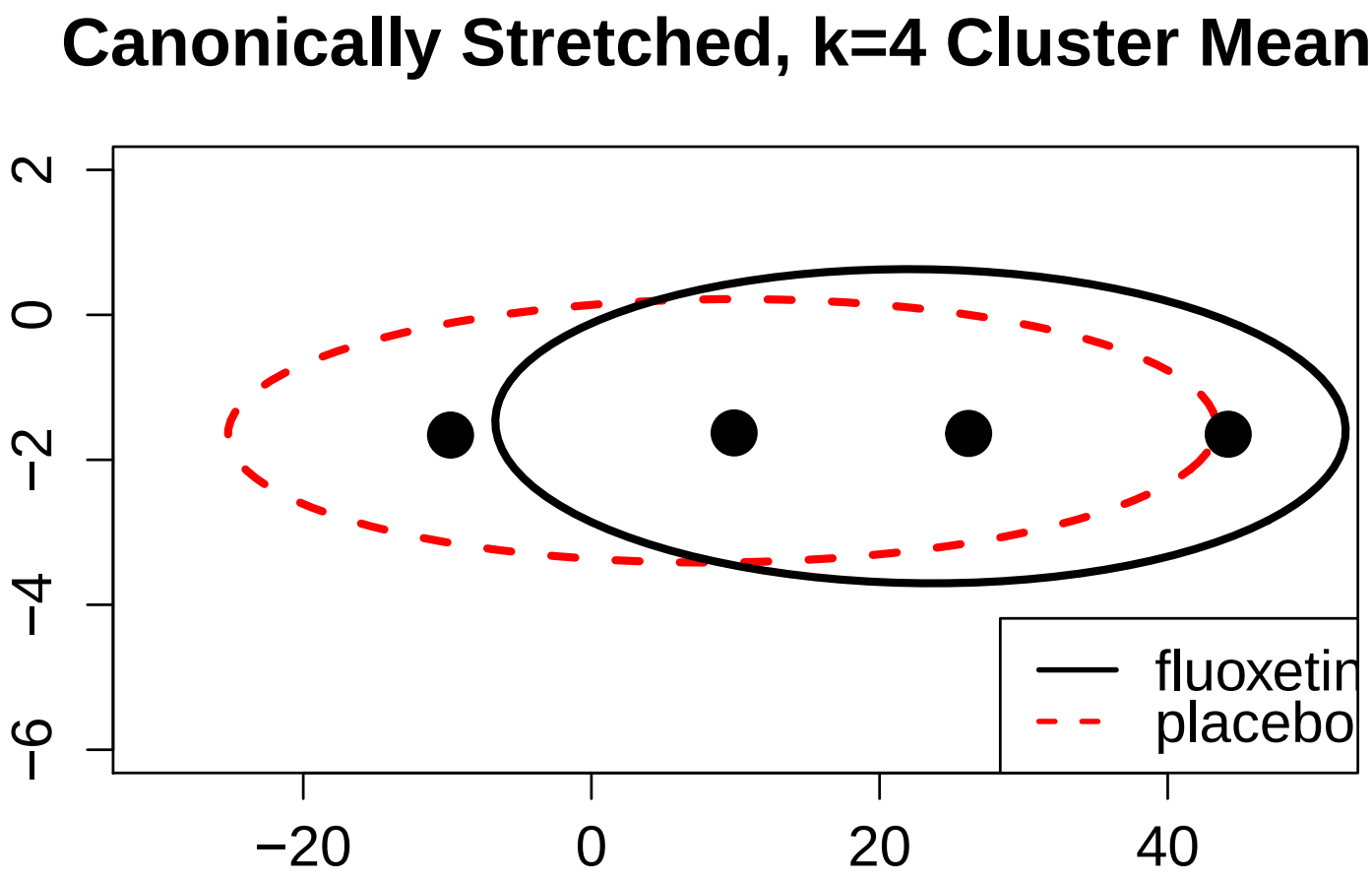
Fluoxetine vs Placebo Trial: Canonical Transformation of Coefficient Distribution

Jointly cluster the longitudinal trajectories into $k = 4$ clusters to determine a regions primarily drug-treated or placebo-treated and regions of overlap.

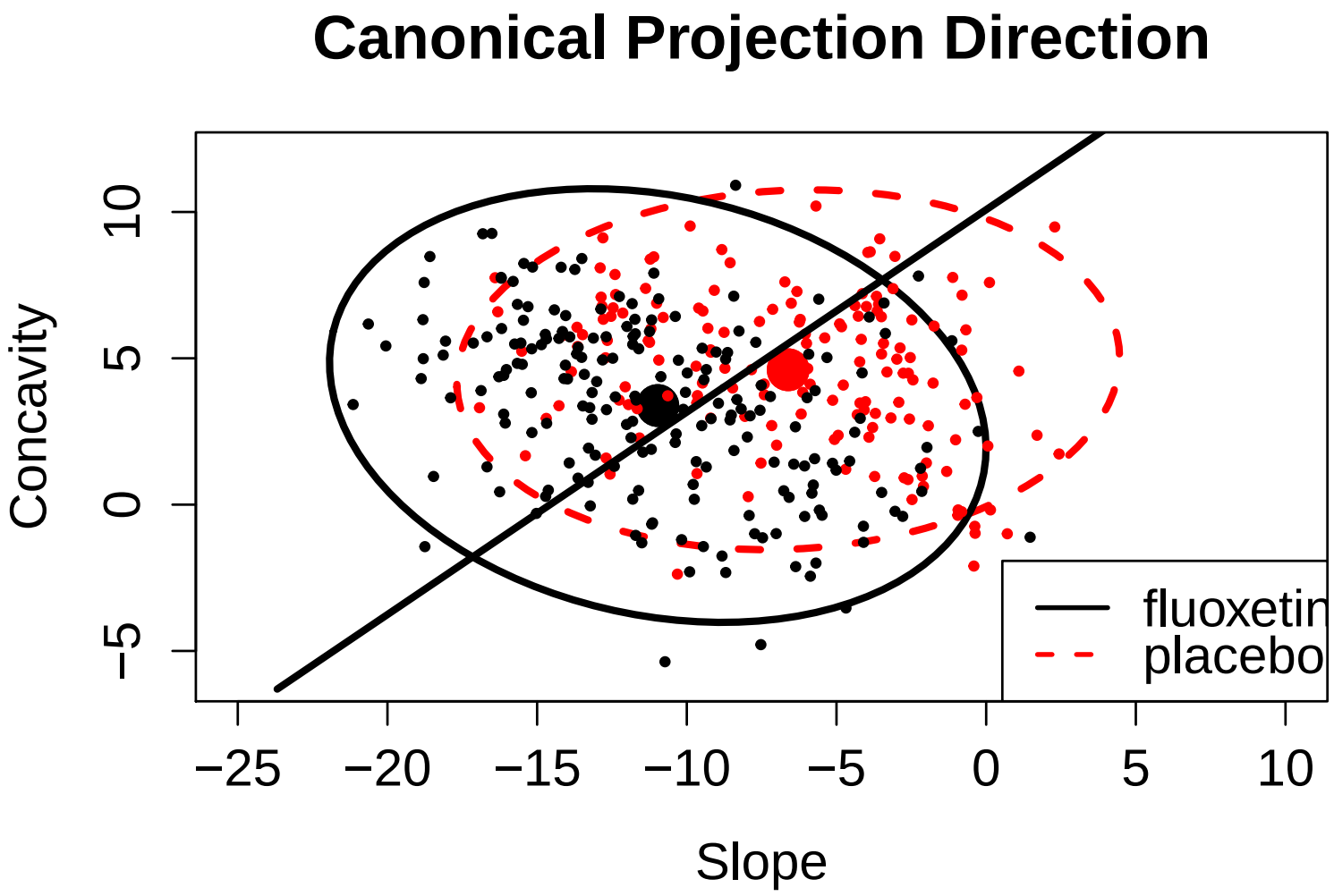


- Optimal Transformation: Appears to be a projection

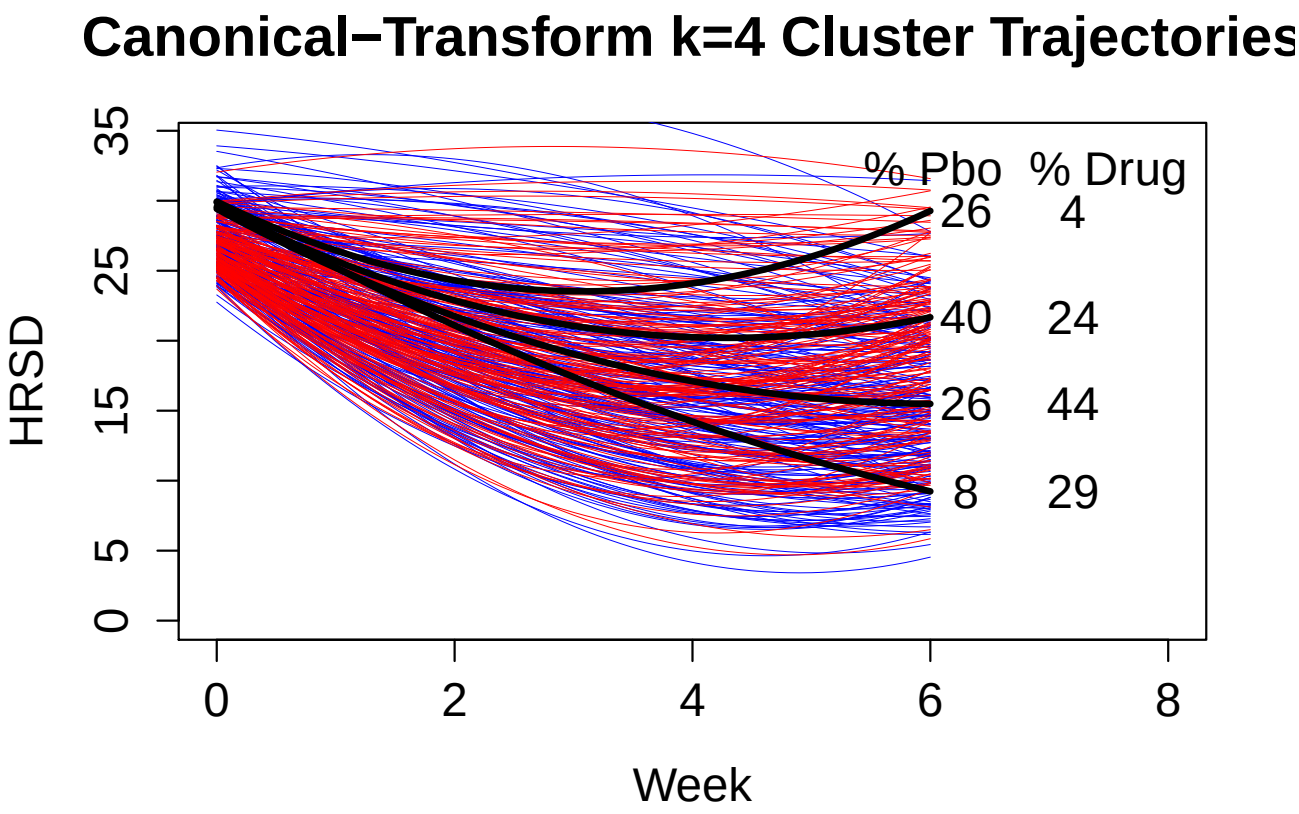
Clustering Results from Canonical Transformation



Canonical Projection Direction



Canonical Cluster Mean Curves: $k = 4$



Moderator Importance Plots

- Jointly cluster functional outcomes from two treatment groups.
- Use a canonical transformation for clustering to produce clusters that are homogeneous as possible with respect to one or the other treatment groups.
- Some clusters will necessarily be heterogeneous due to substantial overlap of outcomes from the different treatments.
- Clustering the functional (or longitudinal) outcomes will allow us to tease apart prototypical trajectories that are primarily associated with one or the other treatment groups.

Moderator Importance Plots continued ...

x = candidate moderator

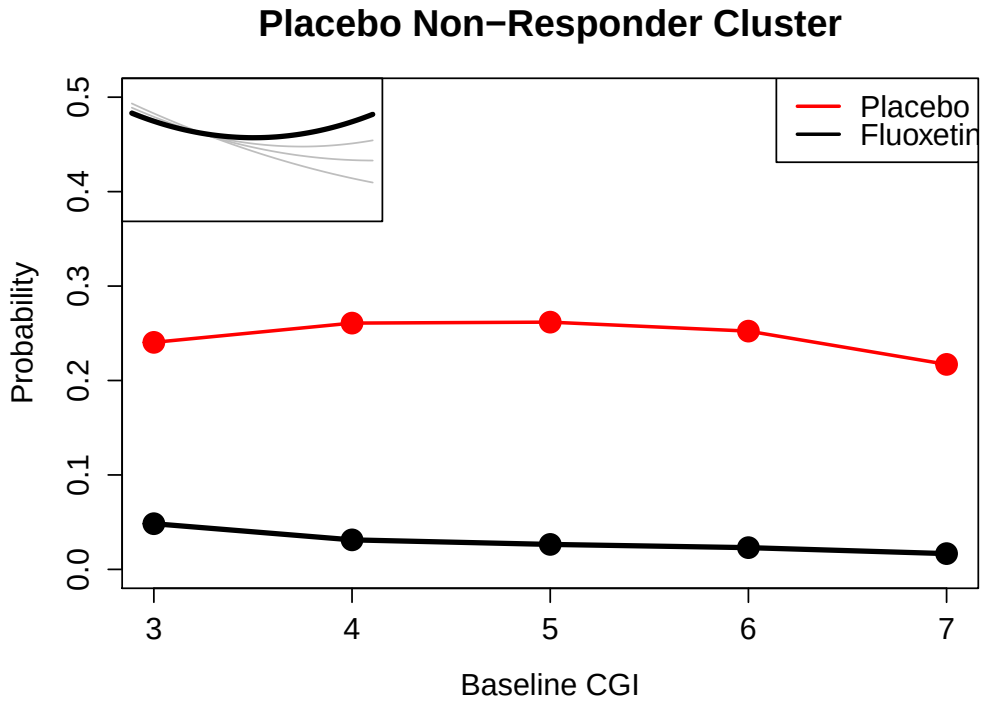
Idea: Estimate $P(\text{Belonging to Cluster } C_j|x)$

Monte Carlo simulation used to estimate this conditional probability using estimated parameters from the linear mixed effects model.

Plot $\hat{P}(\text{Belonging to Cluster } C_j|x)$ versus x.

Moderator Importance Plots: Is Baseline CGI a Moderator?

Clinical Global Impression (CGI): with values from $1 - 7$, higher scores corresponding to higher depression severity.

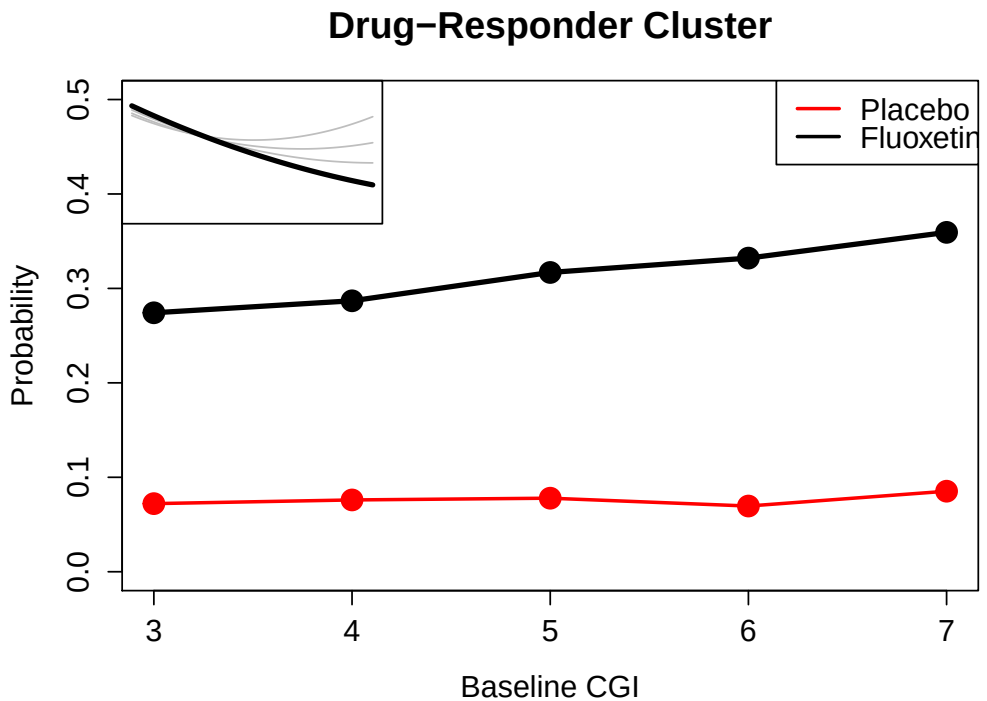


Does CGI predict being a placebo-treated non-responder?

Weak effect seen here.

Moderator Importance Plots: Is Baseline CGI a Moderator?

Clinical Global Impression (CGI): with values from $1 - 7$, higher scores corresponding to higher depression severity.



Does CGI predict being a specific drug responder?

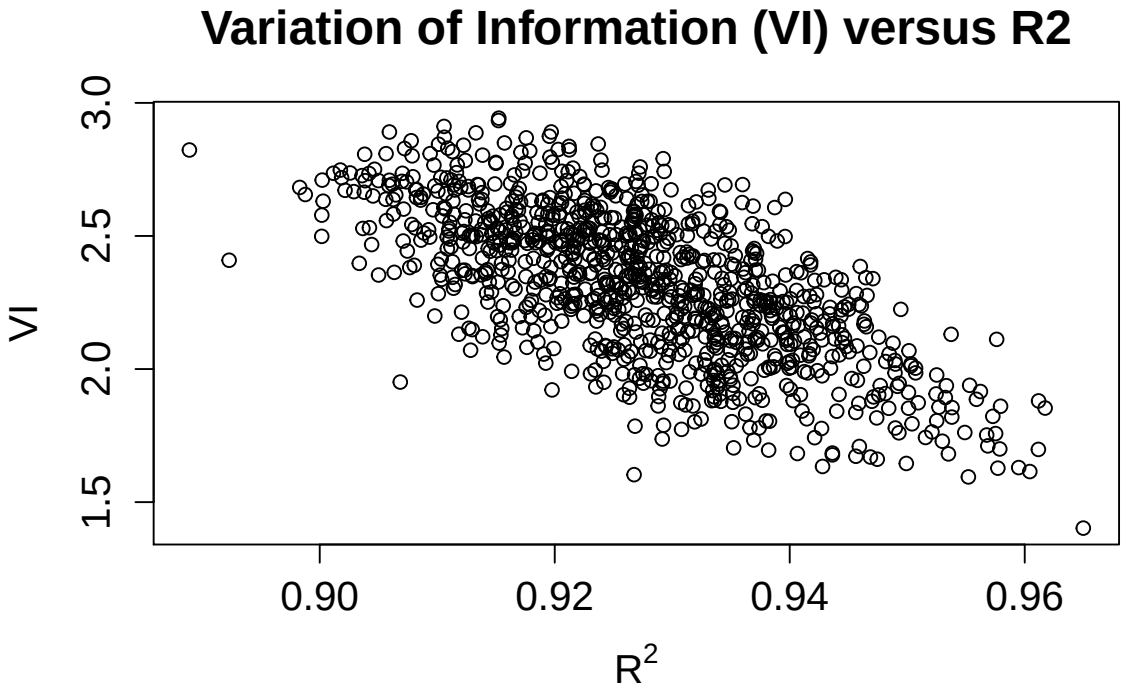
Likelihood of specific drug responder increases with higher baseline severity.

Look at Projections of the Coefficient Data for Clustering

- In the canonical transformation depression example, the optimal transformation in terms of the clustering R^2 corresponded basically to a 1-dimensional projection.
- Explore this further via a simulation: A normal mixture with $K = 5$ components and dimension $p = 5$ was simulated using randomly generated means and covariance matrices.
- $n = 100$ observations were simulated from each mixture component.
- 1000 random projections (onto a line) were also generated and the k-means algorithm was run on each projection specifying $k = 5$ cluster means.

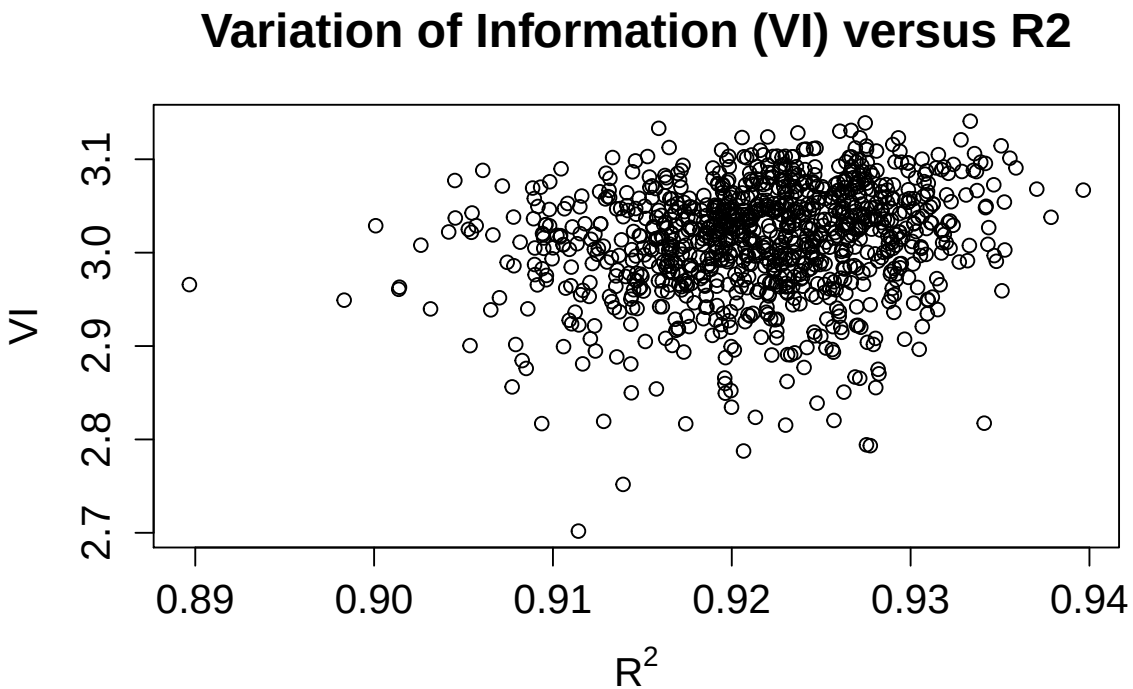
Simulation Illustration: Clustering 1-d Projections

A plot of the variation of information versus the clustering R^2 for 1000 random projections of the data: **Clustering R^2 for projected data tends to increase as VI decreases.**



Clustering Projections Simulation continued ...

Repeat the simulation experiment except increase the dimension from $p = 5$ to $p = 50$: Now there is no clear relation between the clustering R^2 and VI.

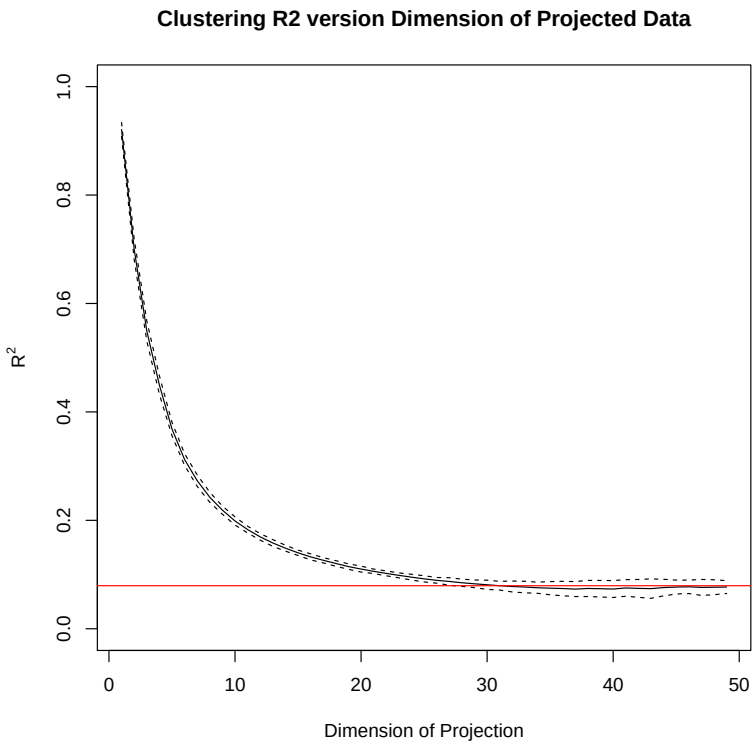


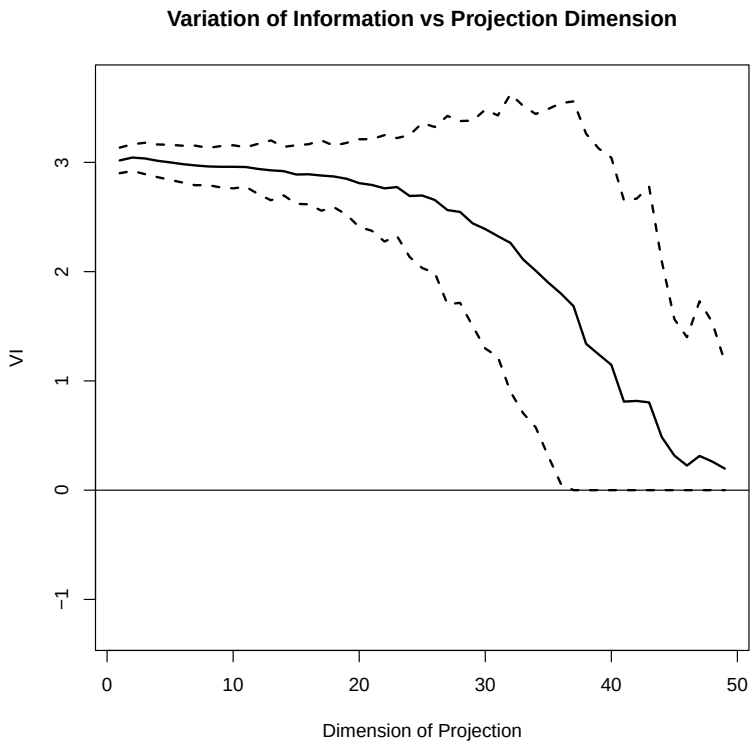
Optimal Dimension for Projections?

- Previous illustration looked at clustering coef. data projected onto a 1-dimensional line.
- Look at clustering results when projecting onto lower-dimensional planes $q < p$ for this 50 dimensional simulated data.

Optimal Dimension for Projections continued ...

Take 100 random projections of dimension $q = 1, 2, \ldots, 49$.

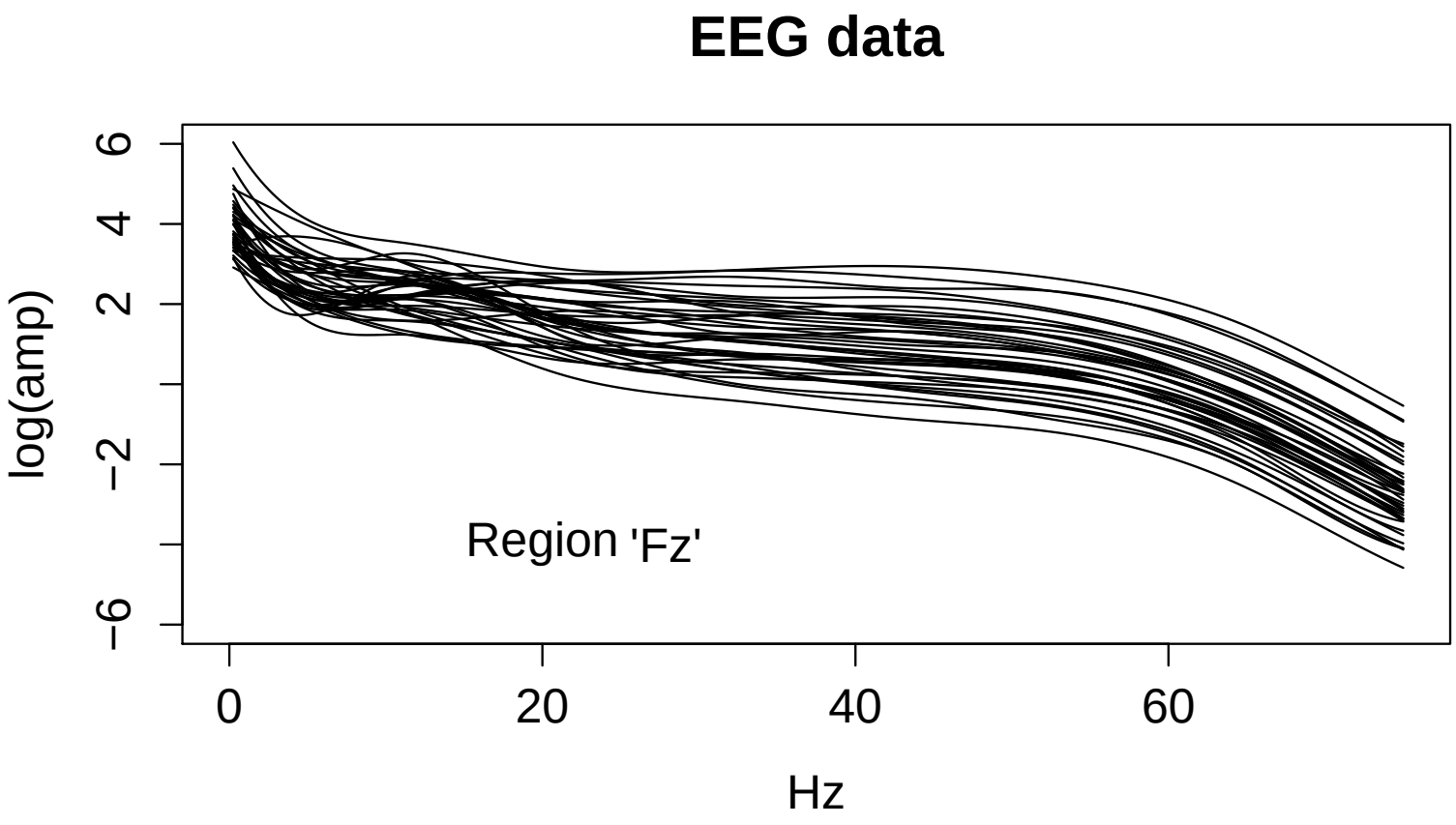




Data from a Depression EEG Study (Bruder et al. 2009)

- $n = 56$ depressed subjects treated with Buproprion alone or SSRI alone or a combination of these two treatments.
- Each subject had EEG recordings from 67 regions on the scalp.
- Preprocessed data consists of $y =$ amplitude measures versus frequence (Hz)
- Penalized cubic B-splines where fit to each subject's data for each region using a 10-dimensional basis.
- Use the "vows" package in R to quickly fit the penalized splines to all subjects and regions ("Massively Parallel Nonparametric Regression, with an Application to Developmental Brain Mapping," by Reiss et al, 2013 *Journal of Computational and Graphical Statistics*).

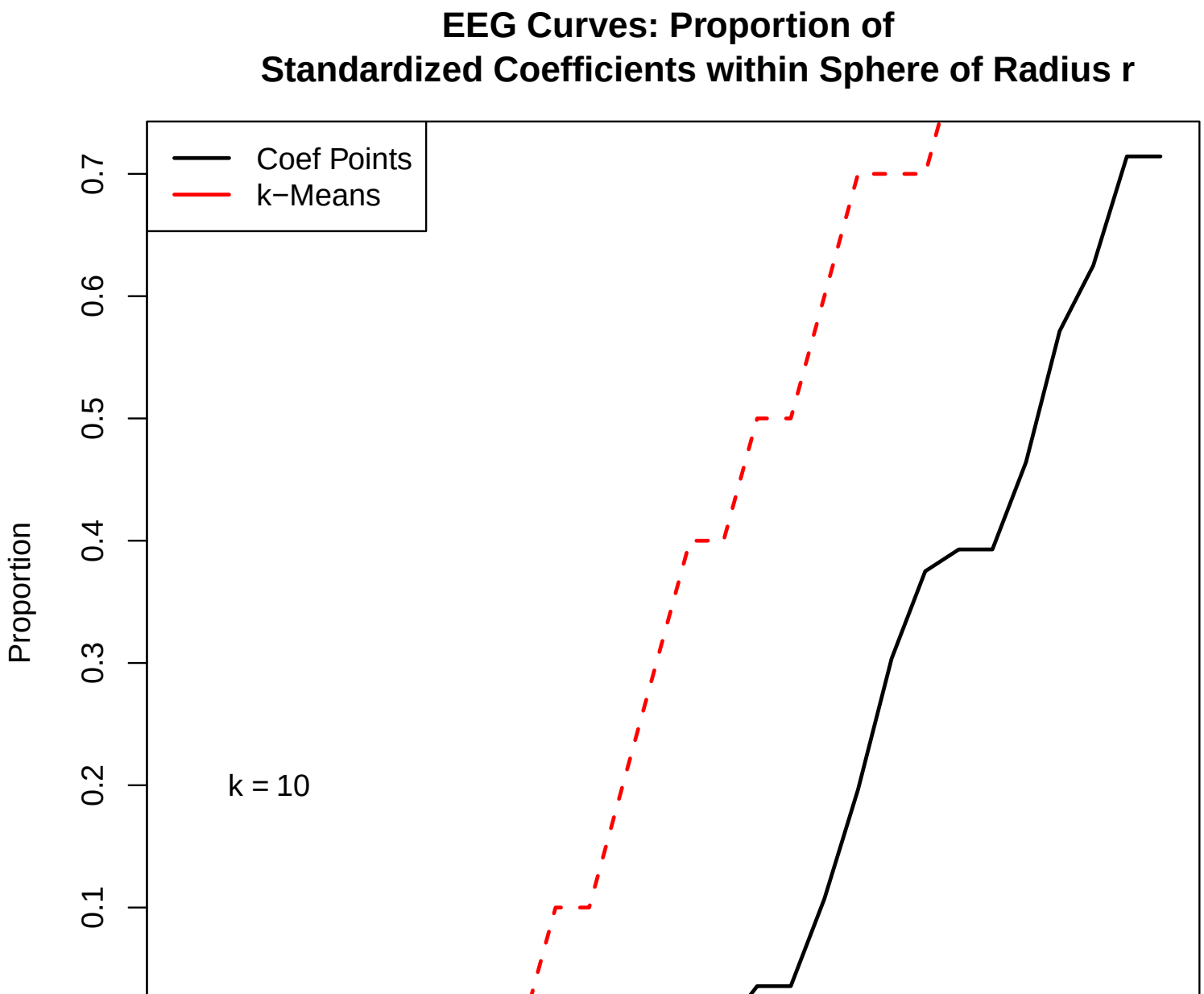
Illustration One EEG Region



Cluster High-Dimensional EEG Curve Data

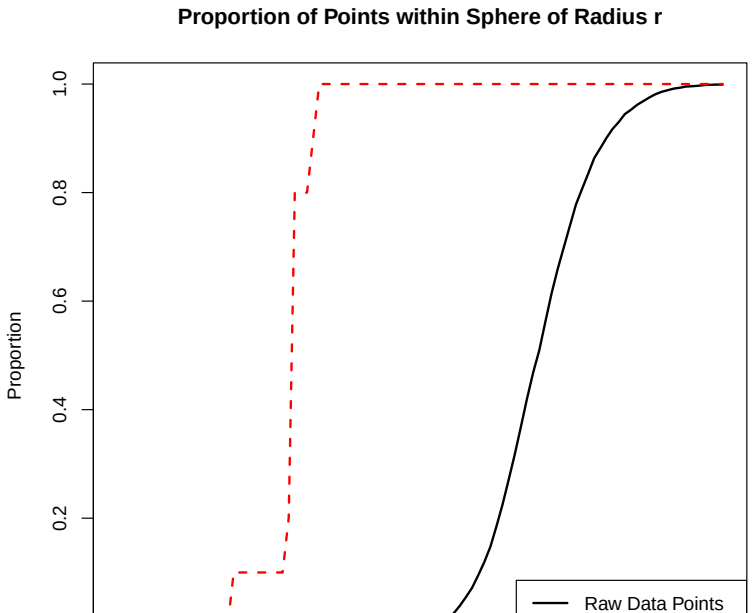
- Each subject has an 11-dimensional B-spline coefficient vector at each of 67 regions. For the sake of illustration, we shall concatenate the data across regions yielding a 737-dimensional coefficient data matrix.
- Center and standardize this matrix
- Cluster the curves and see what happens.

Illustration: Clustering Across all EEG Regions



High-d Cluster Simulation Illustration

- Use a simulation to illustrate the phenomenon seen in clustering high-dimensional EEG data.
- Simulate data from a $K = 10$ normal mixture in $p = 50$ dimensional space (randomly selected means and covariance matrices) wiht $n = 500$ data points simulated from each mixture component.
- The k-means algorithm for $k = 10$



ICA Transformation Illustration

12-week open-label acute phase treatment of depression ($n = 429$) with fluoxetine.

Longitudinal outcome: Hamilton Rating Scale for Depression (HRSD) at 11 time points including baseline.

Use B-splines to fit curves - 5 dimensional coefficient distribution.

A crude check of normality of the coefficient distribution using the Shapiro-Wilks test.

- None of the coefficient distributions deviated from normality ($p > 0.05$).

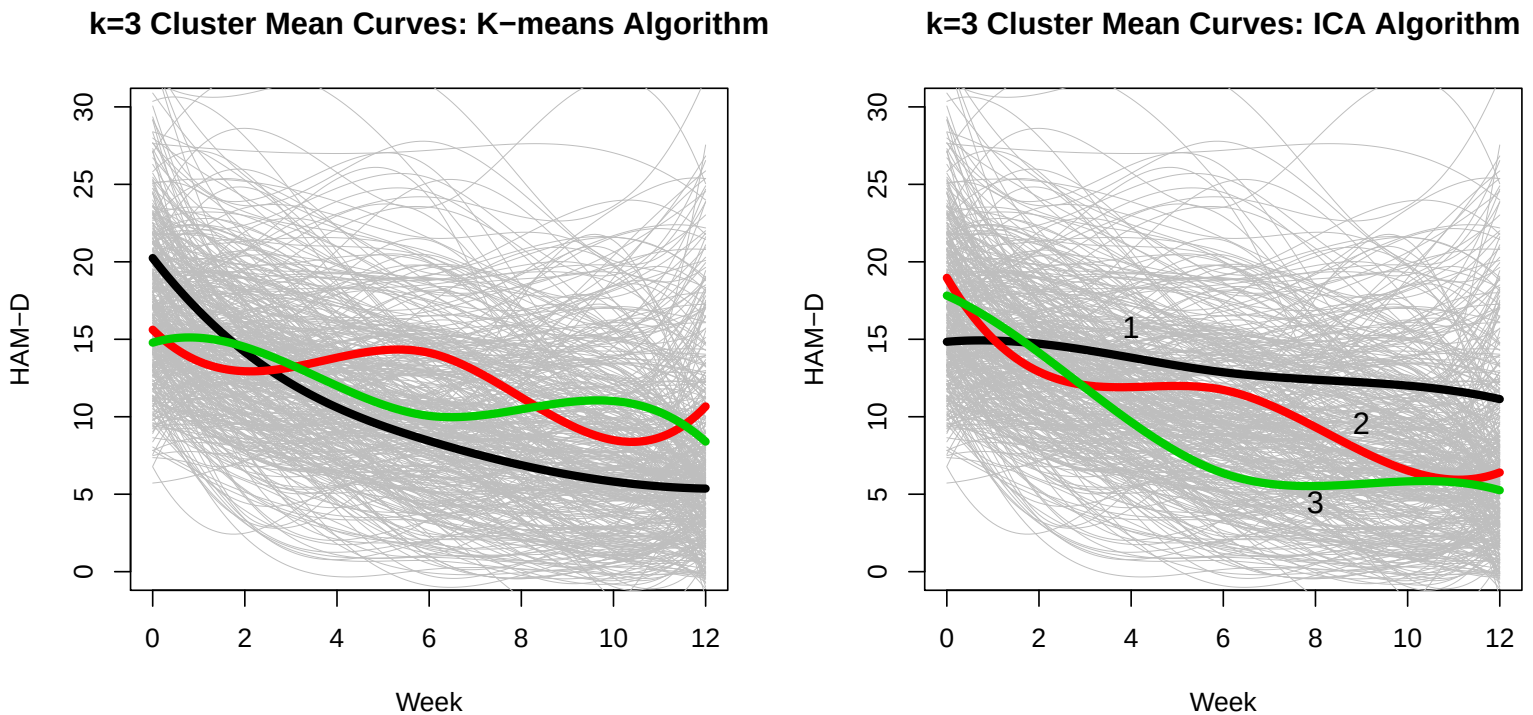
ICA Transformation Illustration continued ...

The FastICA algorithm (Hyvarinen and Oja 2000) was run on the coefficient distribution in R

- Three of the five independent components yielded extremely small Shapiro-Wilks p-values ($p < 0.00001$).
- There exist strongly non-normal directions in the coefficient distribution.
- Linearly transform the coefficient distribution data by artificially inflating the variability in the 3-non-normal directions (by a factor of 1000 say).

ICA Transformation Illustration continued ...

$k = 3$ cluster mean curves fit to the B-spline curves using the usual k-means algorithm (left panel) and using the ICA clustering algorithm in the right panel. Individual-level curves are plotted in grey.



References

Bruder, J., Jürgen, K., and Tenke, C., (2012), "Event-Related Brain Potentials in Depression: Clinical, Cognitive and Neurophysiologic Implications," in n S. J. Luck & E. S. Kappenman (Eds.), The Oxford Handbook of Event-Related Potential Components (pp. 563-592) New York: Oxford University Press.

Chen, H., Reiss, P. and Tarpey, T. (2013), "Weighted L^2 Distance for Functional Data," in revision for *Biometrics*.

Jia, J. and Rohe, K. (2013), "Preconditioning to Comply with the Irrepresentable Condition," Preprint.

Meilă, M. (2007), "Comparing clusteringsan information based distance," *Journal of Multivariate Analysis*, **98**, 873-895.

D. Paul, E. Bair, T. Hastie, and R. Tibshirani (2008), "Preconditioning for feature selection and regression in high-dimensional problems," *The Annals of Statistics*, **36**, 15951618.

Reiss, P., Huang, L. Chen Y-H, Huo, L., Tarpey, T. and Mennes, M. (2013), "Massively Parallel Nonparametric Regression, with an Application to Developmental Brain Mapping," *Journal of Computational and Graphical Statistics*, in press.

Tarpey, T. (2007), "Linear Transformations and the k-Means Clustering Algorithm: Applications to Clustering Curves," *The American Statistician*, **61**, 34–40.