Model Selection and Dimension Reduction: Dealing with High-Dimensional Analysis

Gregory Vaughan

Department of Statistics, University of Connecticut

Statistical Consulting Services

May 10, 2017

Outline

High Dimensional Data

Benefits

- As technology advances, it is becoming easier to collect lots of data
- More options than ever before for investigating interesting questions
- Modern research often calls for analysis of large data sets; e.g. genetics, medical imaging, health sciences, etc (Fan et al. 2014)

Drawbacks

- Very quickly there can be too many characteristics/variables to study
- Interpretation becomes very difficult
- Computation becomes intractable

Challenges

Supervised vs. Unsupervised Learning

- Two main types of problem for high dimensional analysis
- Imagine a child

Challenges

Supervised vs. Unsupervised Learning

Supervised Learning

Unsupervised Learning

Challenges

Supervised vs. Unsupervised Learning

- Supervised Learning
  - There is a clear target to aim for, often called a response, or dependant variable
  - This target helps to guide the analysis
- Unsupervised Learning
  - There is no target
  - Instead, the analysis focuses on finding an overarching structure

Goals

Dimension Reduction (Unsupervised and Supervised)

- Reduce the number of variables being considered in a way that maximizes the amount of information being included in the analysis
- Ease computational difficulties by reducing the number of variables being analyzed

Model Selection (Supervised)

- Identify a subset of the variables that optimally describes the relationship being studied
- Find a parsimonious model that is easy to understand

Goals

Ockham chooses a razor

Everything should be made as simple as possible, but not simpler.
- Albert Einstein

Workshop Objectives

- Cover the key ideas behind popular high-dimensional analysis techniques
  - Factor Analysis
  - LASSO
- Present walkthroughs in `R` of these techniques to give hands on experience with the software

Table of Contents

Unsupervised Learning

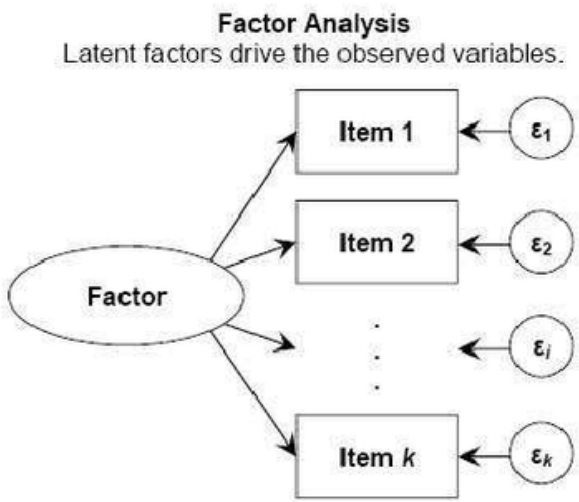
$$\overbrace{\begin{matrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{matrix}}^{p \text{ Variables}} \left\{ \vphantom{\begin{matrix} a \\ a \\ a \\ a \end{matrix}} \right. n \text{ Observations}$$

- Want to understand relationship between variables
- Identify overarching structure to the data

Dimension Reduction

Factor Analysis

- A popular unsupervised learning approach for dimension reduction
- General Idea
  - Assume there are underlying latent quantities called factors that are unobserved, but determine the observed variables to a large degree
  - We express the individual variables as linear combinations of these underlying factors plus an error term



Factor Analysis

- Suppose we assume there are m different factors, $F_1, \ldots F_m$
- Then formally we assume the i^{th} variable X_i can be expressed as

$$X_i = \underbrace{\alpha_{i1}F_1 + \ldots + \alpha_{im}F_m}_{\text{communality}} + \underbrace{\epsilon}_{\text{uniqueness}}$$

- Communality: the representation of the amount of variability explained by the latent factors
- Uniqueness: the variability unique to the particular variable
- ϵ is assumed to be normally distributed with a mean of 0.

Factor Analysis

$$X_i = \quad \alpha_{i1}F_1 + \ldots + \alpha_{im}F_m \quad + \quad \epsilon$$

- The α values are called the factor loadings
  - α_{ij} is the loading between the i^{th} variable and the j^{th} factor
- The loadings are used to measure the amount of variability explained by the factors
  - $\sum_{i=1}^{p} \alpha_{ij}^2 \implies$ impact of factor j on data
  - $\left(\sum_{i=1}^{p} \alpha_{ij}^2\right)/p \implies$ proportion of variance in all variables attributed to factor j
  - $1 - \sum_{j=1}^{m} \alpha_{ij}^2 \implies$ amount of variability in variable i not explained by underlying factors

Exploratory Factor Analysis

- In practice, we often do not know the appropriate number of factors
- There is a Hypothesis test that can indicate if additional factors are needed
- Try multiple numbers of factors
- For each number of factors that are tried, we aim to label the factor based on the factor loadings
  - Examine which variables load heavily from the factor
  - Identify some commonality to these variables

Example: Personality Study

- Participant self-rating on 32 personality traits
  - i.e. distant, talkative, lazy, cooperative, etc...
- How we represent the "self"

```
persDat <- read.table("personality.csv", sep =",")

head(persDat)
```

```
##   distant talkatv carelss hardwrk anxious agreebl
## 1       2       7       1       4       7       8
## 2       3       8       2       7       5       8
## 3       6       6       2       5       1       8
```

Example: Personality Study

```
FAfit1 <- factanal(persDat, factors = 5, rotation = "none")

FAfit1
```

```
## Loadings:
##          Factor1 Factor2 Factor3 Factor4 Factor5
## distant   0.618           -0.103            0.229
## talkatv  -0.581  -0.140    0.516
## carelss   0.353  -0.448    0.348   0.130
## hardwrk  -0.349   0.573                     0.150
## anxious   0.401   0.365    0.374   0.418
## agreebl  -0.265           -0.212   0.527    0.195
## tense     0.450   0.464    0.450   0.292
## kind     -0.418   0.256   -0.229   0.449    0.188
```

Example: Personality Study

```
##                    Factor1 Factor2 Factor3 Factor4 Factor5
## SS loadings          6.668   4.057   2.727   1.951   1.370
## Proportion Var       0.208   0.127   0.085   0.061   0.043
## Cumulative Var       0.208   0.335   0.420   0.481   0.524
```

$$X_i = \quad \alpha_{i1} F_1 + \ldots + \alpha_{im} F_m \quad + \quad \epsilon$$

- SS Loadings $\Longrightarrow \sum_{i=1}^{p} \alpha_{ij}^2$
- Proportion Var $\Longrightarrow \left(\sum_{i=1}^{p} \alpha_{ij}^2 \right) / p$

Example: Personality Study

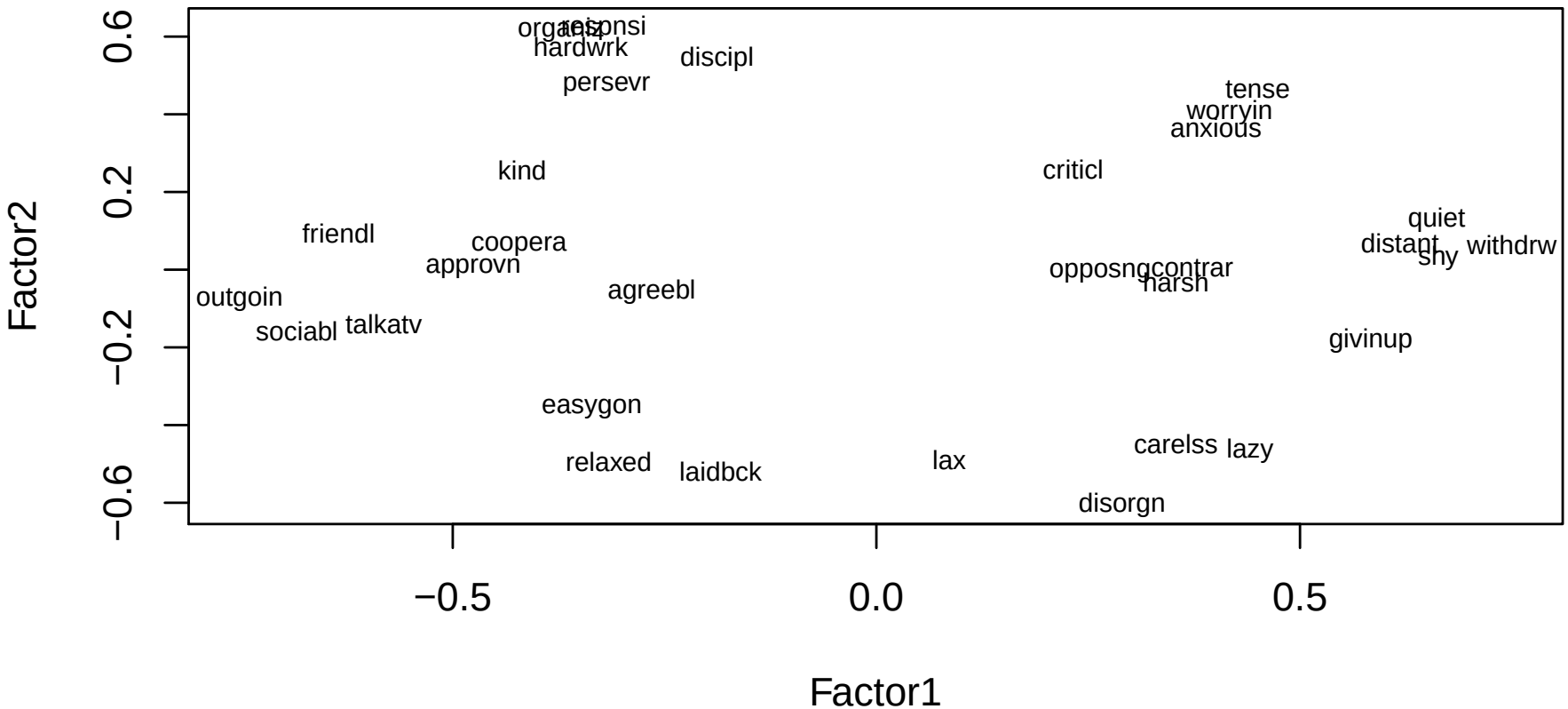
```
## Uniquenesses:
## distant talkatv carelss hardwrk anxious agreebl   tense
##   0.545   0.371   0.530   0.518   0.391   0.566   0.293
##    kind opposng relaxed disorgn outgoin approvn     shy
##   0.470   0.560   0.388   0.423   0.259   0.635   0.391
## discipl   harsh persevr friendl worryin respnsi contrar
##   0.603   0.509   0.641   0.395   0.395   0.416   0.428
## sociabl    lazy coopera   quiet organiz criticl     lax
##   0.401   0.504   0.561   0.294   0.386   0.603   0.672
## laidbck withdrw givinup easygon
##   0.527   0.363   0.594   0.596
```

$$X_i = \quad \alpha_{i1}F_1 + \ldots + \alpha_{im}F_m \quad + \quad \epsilon$$

- Uniqueness $\implies 1 - \sum_{j=1}^{m} \alpha_{ij}^2$

Example: Personality Study

```
load = FAfit1$loadings[,1:2]
plot(load, type="n") # set up plot
text(load,labels=names(persDat),cex=.7) # add variable names
```

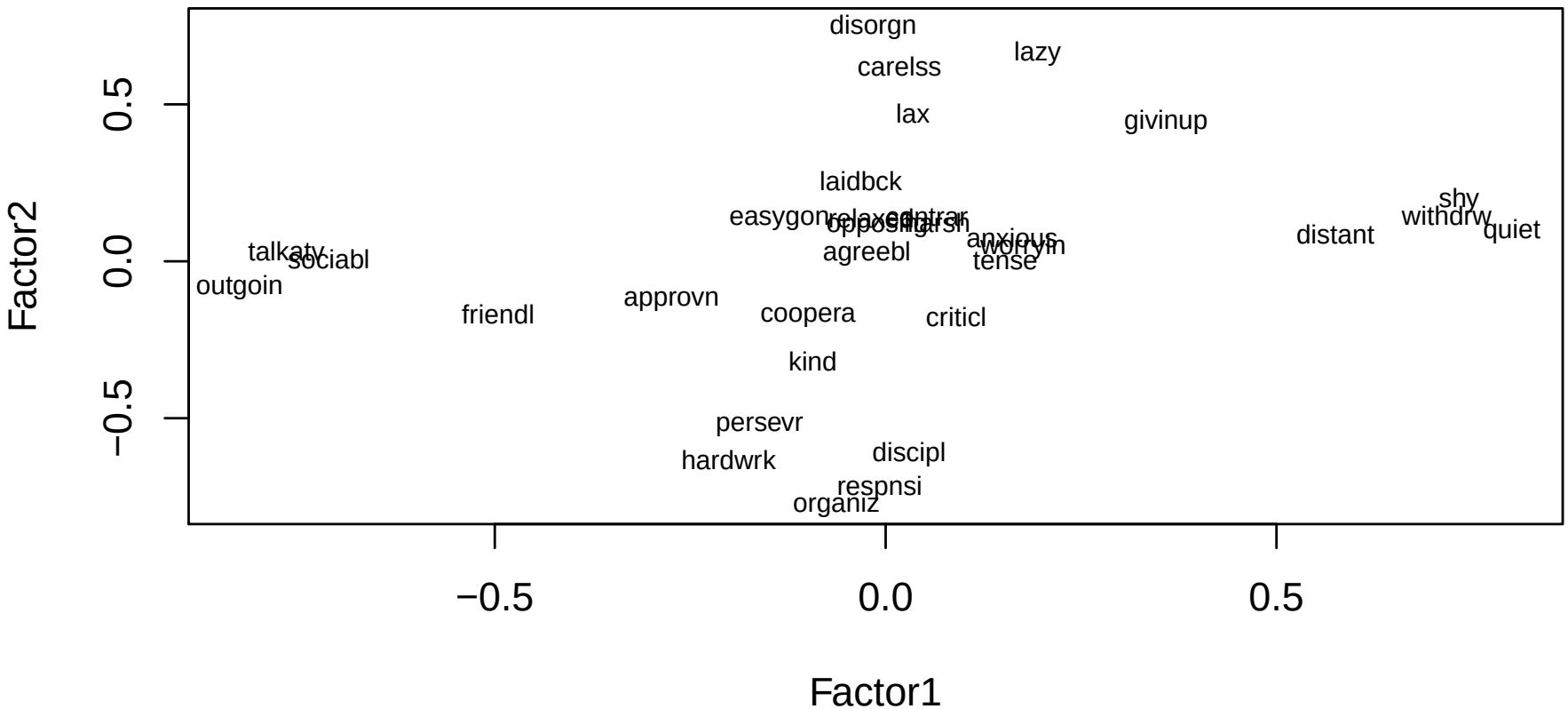


Rotation of Factor Loadings

- Rotation $\implies$ creating linear combinations of factors, resulting in new factors
- Rotations are beneficial!
- `varimax` rotation forces each variable to load more heavily on a single factor, less on others.
  - Eases interpretation
- No drawback! Any rotation of a best fitting answer is also best fitting!

Example: Personality Study

```
FAfit2 <- factanal(persDat, factors = 5,
                                rotation = "varimax")
```



Example: Personality Study

Test of Sufficient Factors

```
FAfit2
```

```
Test of the hypothesis that 5 factors are sufficient.
The chi square statistic is 707.22 on 346 degrees of freedom.
The p-value is 5.42e-27
```

Confirmatory Factor Analysis

- When we have a previous factor structure, we can test if the model is a good fit
- Pre-specify our factors, and identify some of the loadings as 0
- Perform a fitting to estimate the remaining loadings

Example: Holzinger and Swineford

- Famous study by Holzinger and Swineford (1939)
- Mental ability test scores of 7^{th} and 8^{th} graders
- Two different schools
- 9 different tests

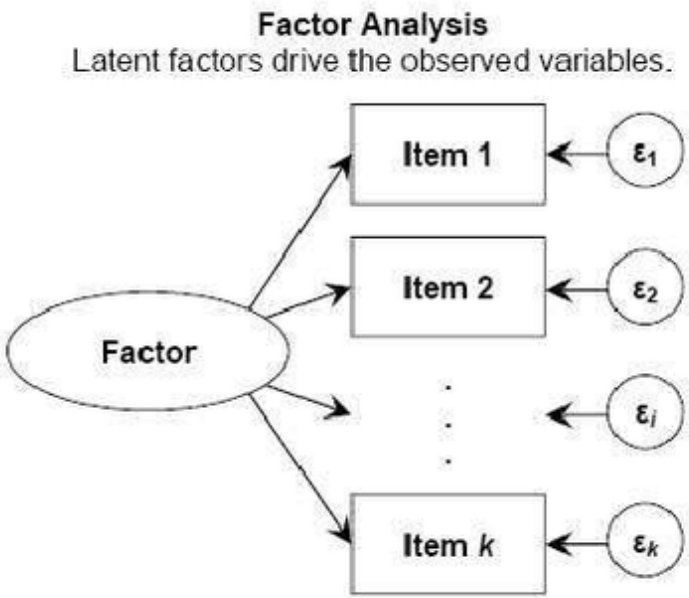
Example: Holzinger and Swineford

```
require(lavaan)

HS.model <- ' visual  =~ x1 + x2 + x3
              textual =~ x4 + x5 + x6
              speed   =~ x7 + x8 + x9 '

CFAfit <- cfa(HS.model, data = HolzingerSwineford1939)
```

```
##   Estimator                                         ML
##   Minimum Function Test Statistic                85.306
##   Degrees of freedom                             24
##   P-value (Chi-square)                           0.000
```

EFA vs. PCA

- Another popular technique $\Longrightarrow$ Principal Component Analysis (PCA)
- In PCA, construct components w/ linear combinations of the variables
- Many refer to EFA and PCA interchangeably $\Longrightarrow$ NOT THE SAME!



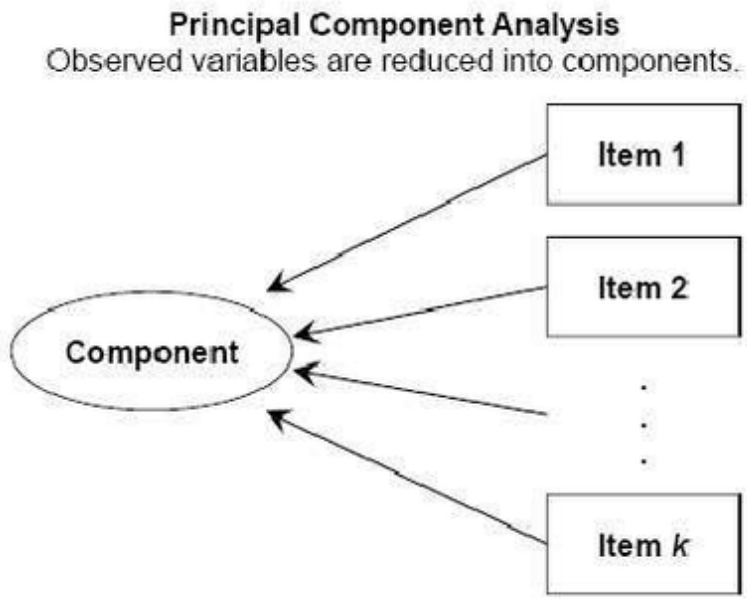


Table of Contents

Supervised Learning

Regression

$$
\underbrace{\begin{Bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{Bmatrix}}_{n \text{ Observations}}
\qquad
\overbrace{\begin{matrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{matrix}}^{p \text{ Covariates}}
$$

- Aiming for (predicting) future response values
- Many covariates $\implies$ hard to interpret
- Want best possible model $\implies$ some covariates may hurt the model
- $p > n \implies$ traditional estimation won't work

Model Selection

- Suppose we have come criterion C that measures model's relative effectiveness
- Can compare models based on C
- For example, C = Mean Squared Error (MSE), i.e.

$$C(\text{model}) = \text{MSE}(\text{model}) = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 / n,$$

$Y \to$ response $\qquad \widehat{Y} \to$ corresponding predictions

- To compare

$$\text{MSE}(\text{Model 1}) = 10 \quad \text{Model 1}$$
$$\text{MSE}(\text{Model 2}) = 20 \quad \text{is better!}$$

Model Selection

Model Selection

- Having such a criterion, we could compare all possible models (combinatorial approach)
- p covariates $\implies$ each one is in or out $\implies$ 2^p models!
  Many models to compare $\implies$ lots of computation
- Possibly infeasible amount of computation time

LASSO

Least absolute shrinkage and selection operator (LASSO) reduces the number of candidate models to make problem feasible

- Traditional Regression $\Longrightarrow \min_{\beta} f(\beta)$

  $f \rightarrow$ loss function $\qquad \beta \rightarrow p \times 1$ vector

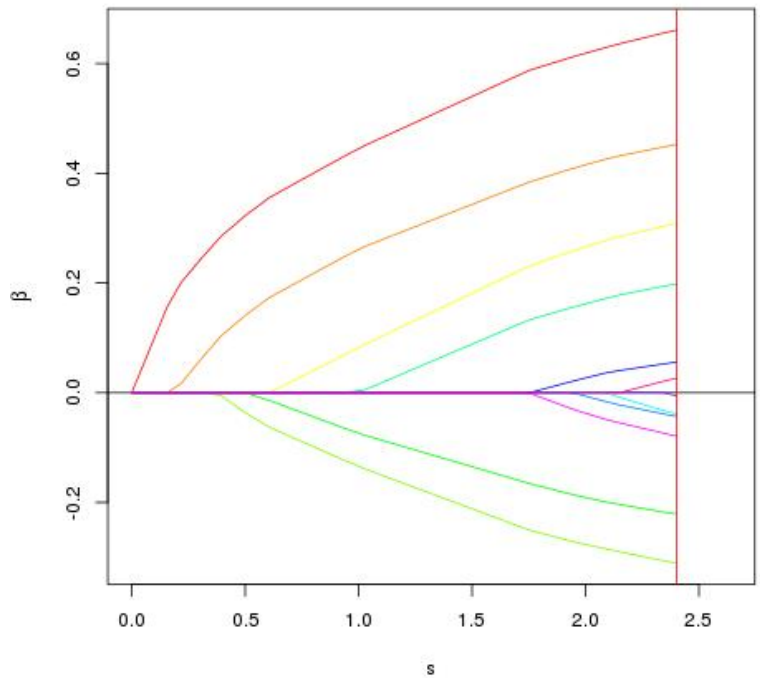
  i.e. Multiple Linear Regression
  - Assume $Y_i = \beta_0 + X_{i1}\beta_1 + \ldots + X_{ip}\beta_p + \epsilon$
  - ϵ is normally distributed with mean of 0
  - $f(\beta) = \sum_{i=1}^{n} (Y_i - (\beta_0 + X_{i1}\beta_1 + \ldots + X_{ip}\beta_p))^2$
- LASSO

$$\min_{\beta} f(\beta) \text{ subject to } \sum_{i=1}^{p} |\beta_i| \leq s,$$

  - $s \rightarrow$ tuning parameter

LASSO

- $s \rightarrow$ controls complexity
  - $s = 0 \implies$ empty model
  - Large $s \implies$ full model
- Path dictates subset of models to choose from using criterion C



LASSO Advantages

- Honey, I shrunk the estimates!

- All estimates are shrunken towards zero
- Makes our estimates more precise in exchange for a slight bias
- Estimates below a certain threshold are forced to exactly 0
  - This effectively removes the less important variables from the model

LASSO in Practice

- The value of s needs to be appropriately selected
- Best value will depend on Criterion C and data

Naive Approach

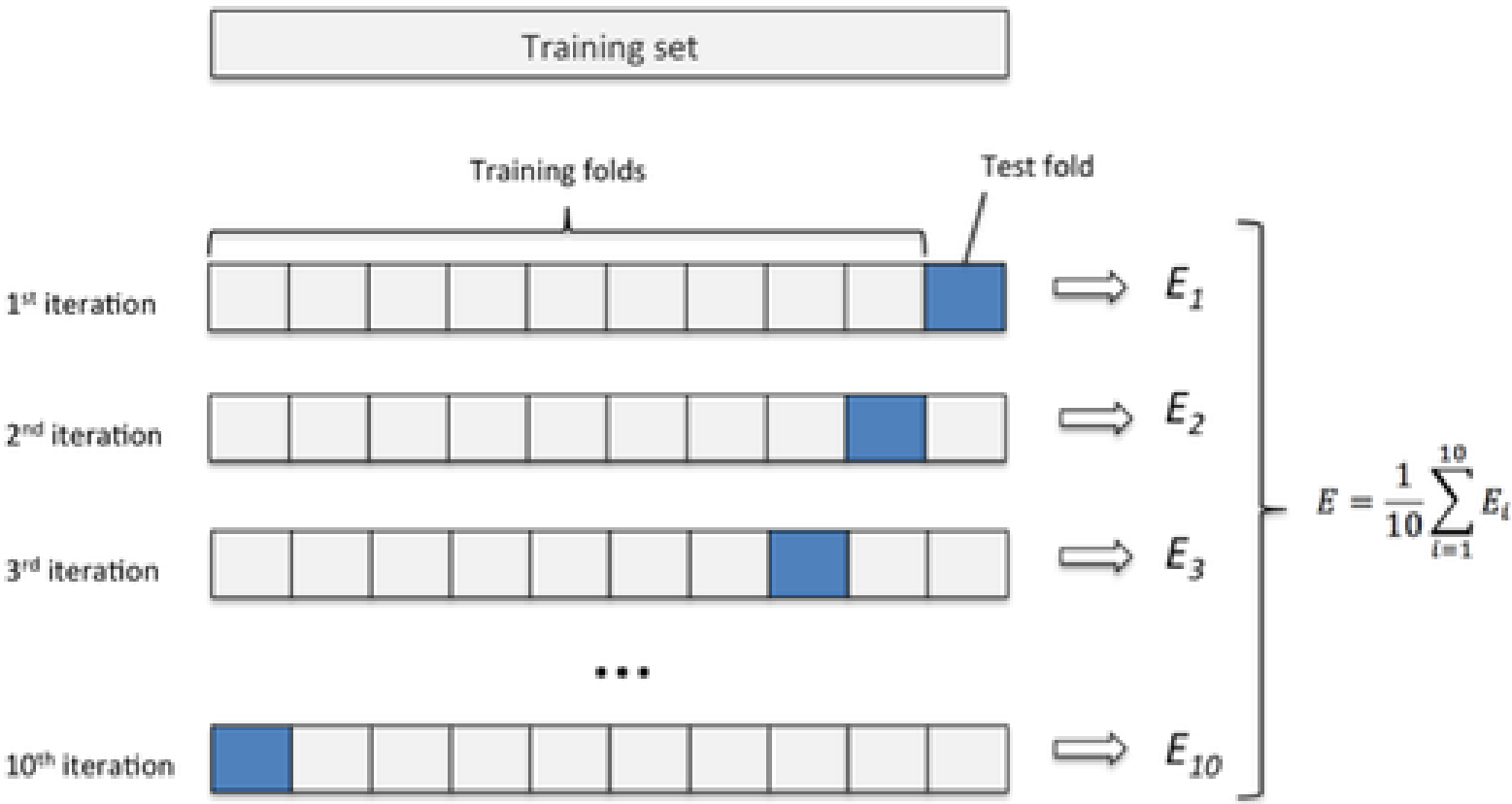
- Divide Data into a training data and testing data
- Several values of s are considered, generate estimates for each using training data
- Test performance using C with the testing data for each value of s
- Value of s with the best performance is selected

Cross-Validation

There is a better way!

1. Divide Data into multiple (k) groups of folds
2. For $i = 1, 2, \ldots, k$
   - 2.1 The i^{th} fold is the current testing fold
   - 2.2 Build model for each candidate value of s w/ all other folds
   - 2.3 Calculate criterion C (usually predictive Mean Squared Error) using testing fold for each s
3. Calculate average criterion value across all values of i for each value of s
4. Value of s with best average performance is chosen
5. Model is re-fit using entire dataset for optimal value of s

Cross-Validation



Example: Brain Injury

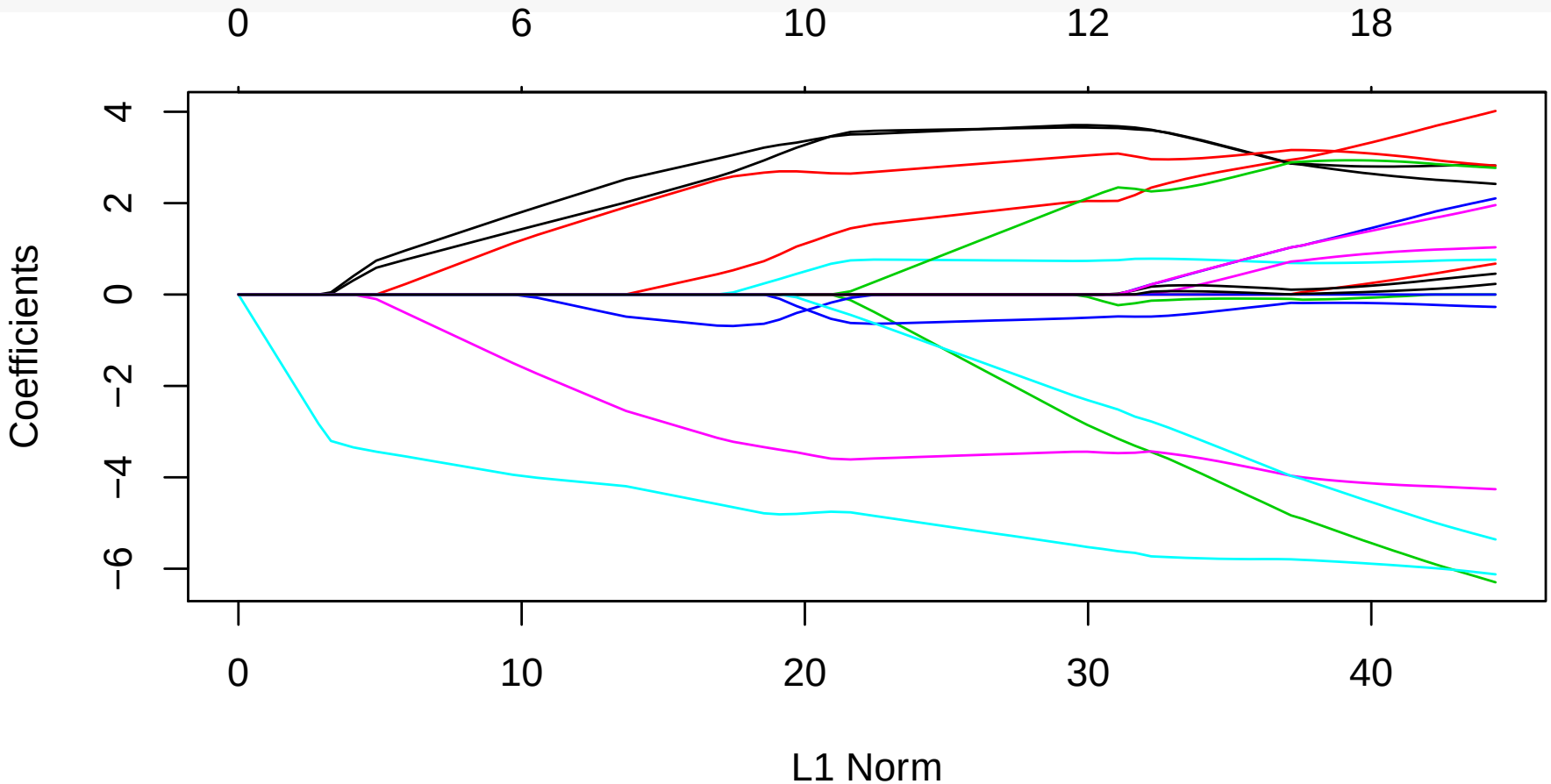
- Subject had suffered a traumatic brain injury
- Studying word recall performance compared to population without brain damage
- Many covariates/tests evaluating subjects
- Not many subjects

```
require(glmnet)
memDat <- read.table("memory.csv", sep =",")
head(memDat)
```

```
##    Y NART_FS NART_V NART_P WAB_AQ WAB_Rdg WMS_VRI_Rec
## 1 60       3      2      4      2       4           5
## 2 82       4      3      1      3       6           3
## 3  4       3      5      2      2       7           7
```

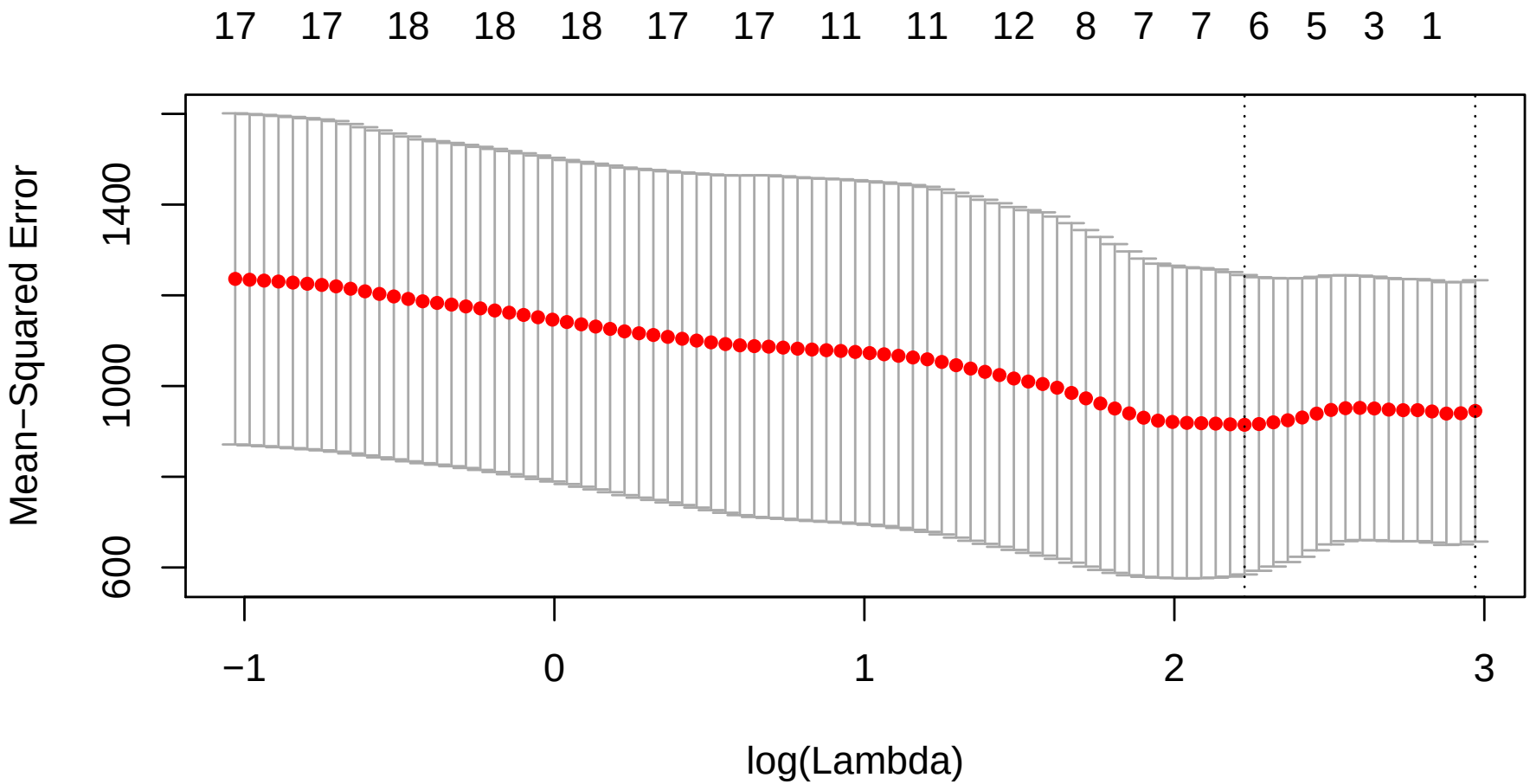
Example: Brain Injury

```
y <- memDat$Y
x <- as.matrix(memDat[,-1])
LASSOpath <- glmnet(y = y, x = x)
plot(LASSOpath)
```



Example: Brain Injury

```
LASSOfit <- cv.glmnet(y = y, x = x, nfolds = 5)
plot(LASSOfit)
```



LASSO vs. Stepwise Regression

- Another popular and well know model selection technique is Stepwise Regression
  1. Start with an empty model, constants only
  2. For a fixed number of iterations, do the following
     - 2.1 For each variable not in the most recent model, perform a fitting where the variable is included
     - 2.2 Calculate criterion of fit for each model generated in this way
  3. Variable with optimal criterion is selected and added to the model
  4. If criterion is not improved in a given iteration process terminates

LASSO vs. Stepwise Regression

- While stepwise regression is popular, it has many issues that lasso can successfully navigate
  - The resulting model's R squared values may be biased to be high
  - Suffers severely from correlation between the covariates
  - Does not benefit much from increased sample sizes
  - Does not benefit from estimate shrinkage

Summary

- Supervised vs. Unsupervised Learning
- Dimension Reduction & Model Selection
- Factor Analysis (Unsupervised, Dimension Reduction)
  - Exploratory Factor Analysis: `R` function `factanal`
  - Confirmatory Factor Analysis: `R` package `lavaan`
  - EFA $\neq$ PCA
- LASSO (Supervised, Dimension Reduction & Model Selection)
  - `R` package `glmnet`
  - LASSO $>$ stepwise regression

Thank You For Listening!

Fan, J., Han, F., and Liu, H. (2014). Challenges of big data analysis. *National Science Review* **1,** 293–314.

Holzinger, K. J. and Swineford, F. (1939). A study in factor analysis: the stability of a bi-factor solution. *Supplementary Educational Monographs* .