Measuring Ideology in Text and Networks

John Myles White

September 24, 2012

Ideal Points: Measure of Political Ideology

- Ideal points are a continuous scale for measuring ideology
- How can we relate them to text and networks?

Harvest Congressional Text and Match with Ideal Points

- Manually gather text from each senator
  - Floor speeches
  - Press releases
  - Op-eds
- Assign senator's ideal point to text written by them
- Predict ideal points using text

Importing Text into R

- Use the ‘tm’ package
- Store all documents in one large CSV file
- Use ‘read.csv’ to turn CSV file into a data.frame
- Create a ‘tm’ corpus from the data.frame
- Build a document-term matrix

Example Document A from the Corpus

i want to talk about jobs lately it seems that everyone says they want to talk about jobs and that we'll get around to tackling jobs next week or the week after

Example Document B from the Corpus

there was a major legislative accomplishment in washington last week and it's getting less attention than it deserves because it isn't national health care reform

Example Document-Term Matrix

Document	I	Want	Talk	Jobs	Week
A	1	2	2	3	2
B	0	0	0	0	1

Creating a Corpus with 'tm'

```
corpus <- Corpus(DataframeSource(documents))

corpus <- tm_map(corpus,
                 tolower)

corpus <- tm_map(corpus,
                 removeWords,
                 stopwords('english'))
```

Creating a Document Term Matrix with 'tm'

```
document.term.matrix <- DocumentTermMatrix(corpus)

x <- as.matrix(document.term.matrix)

y <- authors$IdealPoint
```

Text Regression

- Turn text analysis into a standard regression problem
- Outcomes: senators' ideal points
- Predictors: document words counts

Regularized Regression in R

- Use the 'glmnet' package
- Fits any form of Elastic Net
- 'alpha' shifts between LASSO and ridge
- 'alpha' $= 1$ by default, which gives LASSO
- 'lambda' controls the prediction error / regularization tradeoff
- 'glmnet' fits many values of 'lambda' automatically

Simple glmnet Example

```
x1 <- c(1, 2, 3)
x2 <- c(1, 1, 3)
x3 <- c(1, 4, 3)
x <- cbind(x1, x2, x3)

a <- 1
b <- 2
c <- 3

y <- a * x1 + b * x2 + c * x3 + rnorm(3, 0, 1)
```

Simple glmnet Example

```
library('glmnet')

fit <- glmnet(x, y)

fit
```

Simple glmnet Example

```
Call:  glmnet(x = x, y = y)

      Df   %Dev Lambda
 [1,]  0 0.0000 4.3300
 [2,]  1 0.1550 3.9460
...
[38,]  2 0.9989 0.1385
[39,]  2 0.9991 0.1262
```

Fit Model to Text Data

```
fit <- glmnet(training.x, training.y)
```

Find Most Biased Terms

```
term.weights <- coef(fit, s = optimal.lambda)

sorted.terms <- sort(term.weights[,1])

n <- length(sorted.terms)

most.democratic.terms <- sorted.terms[1:10]
most.republican.terms <- sorted.terms[(n - 9):n]
```

Top 10 Most Republican Terms

Term	Value
okla	1.23
bailey	0.647
johnny	0.588
administering	0.561
neb	0.556
sam	0.542
986	0.532
texans	0.493
patriotism	0.466
demint	0.417

Top 10 Most Democratic Terms

Term	Value
sherrod	-0.367
sheldon	-0.249
dec	-0.196
possess	-0.168
salaries	-0.158
tom	-0.152
debbie	-0.151
dark	-0.148
lautenberg	-0.133
fought	-0.106

Debugging Our Results

- Too many names of senators in our list
- Strip out all the names from corpus
- Run analysis from scratch on clean corpus

Top 10 Most Republican Terms excluding Names

Term	Value
okla	1.13
neb	0.726
bailey	0.674
2415	0.638
986	0.578
kansans	0.543
administering	0.516
texans	0.467
profoundly	0.459
patriotism	0.430

Top 10 Most Democratic Terms excluding Names

Term	Value
cedar	-0.224
chaired	-0.197
dec	-0.158
dark	-0.146
blocked	-0.138
reverses	-0.134
1960s	-0.125
insurers	-0.0958
fought	-0.0926
possess	-0.0923

Assessing Our Predictive Power

```
predicted.y <- predict(fit,
                       newx = test.x,
                       s = optimal.lambda)

predicted.y <- as.numeric(predicted.y[,1])

predictions <- data.frame(Predicted = predicted.y,
                          Empirical = test.y,
                          Residual = predicted.y - test.y)
```

Assessing Our Predictive Power

```
RMSE <- with(predictions, sqrt(mean(Residual ^ 2)))
```

Final Model Comparison Results

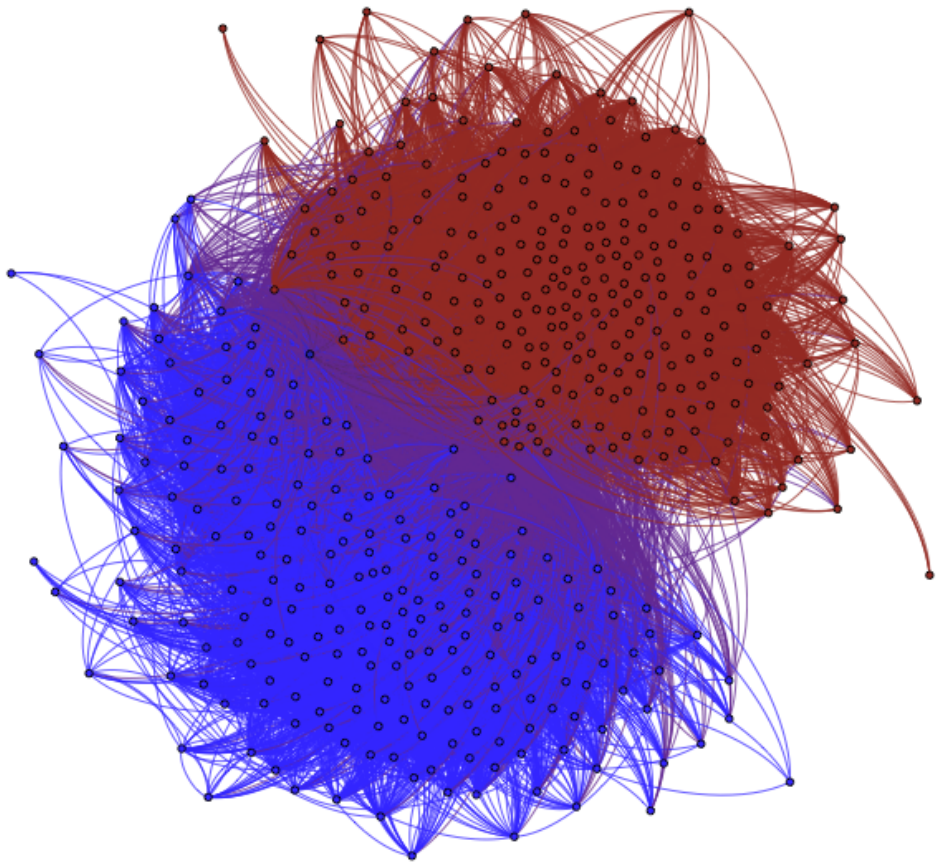
Model	RMSE
Pure Intercept Regression	1.02
Lasso Text Regression excluding Senators' Names	0.879
Lasso Text Regression including Senators' Names	0.805

Network Analysis in R

Many tools:

- igraph
- statnet
- latentnet
- eigenmodel

Political Networks are Polarized

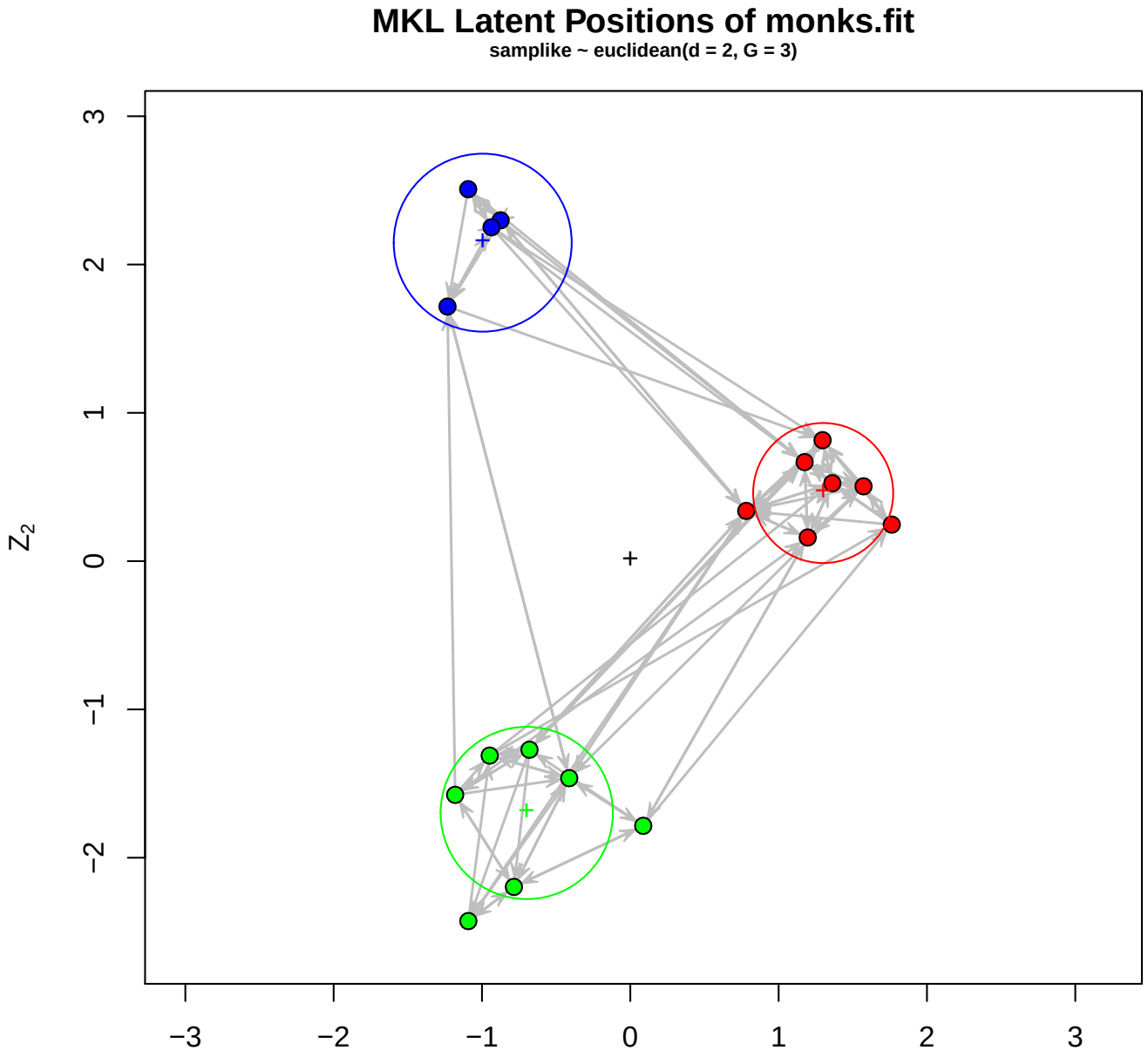


Latent Space Models for Networks

```
library("latentnet")

data(sampson)
monks.fit <- ergmm(samplike ~ euclidean(d = 2, G = 3))
plot(monks.fit)
```

Sample Latent Space Model Fit



CRAN Packages Used

- plyr
- ggplot2
- glmnet
- latentnet
- ProjectTemplate