Homework 1

Jordan Giebas

Due Wednesday, November 1 at 5:30 PM

Part 1:

Reconsider the data set that was presented in the final exam for Mini 1. Create a new data frame that gives the day-to-day changes in all of the rates. Run a principal components analysis on these vectors. Is PCA effective in reducing the dimensionality of the rate change vectors? Try it with and without scaling the variables first and describe how the results change.

```
library(quantmod)
library(ggplot2)
library(plyr)

## Get the data from the final exam
######################################################################
fileheader = read.table("commercial_paper_rates.csv",sep=",",nrows=6,stringsAsFactors=FA
cprfullnames = as.character(c("Time.Period",fileheader[1,-1]))
cprdata = read.table("commercial_paper_rates.csv",sep=",", header=T,
                     skip=5,stringsAsFactors = FALSE,
                     na.strings=c("ND","NA"))

cprdata$Time.Period = as.Date(cprdata$Time.Period, format="%Y-%m-%d")
SP500 = getSymbols("SP500",src="FRED", auto.assign=FALSE)
```

```
## 'getSymbols' currently uses auto.assign=TRUE by default, but will
## use auto.assign=FALSE in 0.5-0. You will still be able to use
## 'loadSymbols' to automatically load data. getOption("getSymbols.env")
## and getOption("getSymbols.auto.assign") will still be checked for
## alternate defaults.
##
## This message is shown once per session and may be disabled by setting
## options("getSymbols.warning4.0"=FALSE). See ?getSymbols for details.
```

```
keep = match(time(SP500), cprdata$Time.Period)
fulldata = cbind(cprdata[keep,],data.frame(SP500))
fulldata = fulldata[complete.cases(fulldata),]
######################################################################

dates = fulldata$Time.Period
dates = dates[2:length(dates)]

data = fulldata[,c(2:ncol(fulldata))]
```

```r
data = -1*colwise(diff)(data)

df = cbind(dates, data)
df_dates = df[,1]
df_data = df[,c(2:ncol(df))]

df_pca_noscale = prcomp(df_data)
summary(df_pca_noscale)
```

```
## Importance of components%s:
##                            PC1     PC2     PC3     PC4     PC5     PC6
## Standard deviation     31.8318 0.50191 0.20963 0.18422 0.14854 0.13777
## Proportion of Variance  0.9995 0.00025 0.00004 0.00003 0.00002 0.00002
## Cumulative Proportion   0.9995 0.99977 0.99981 0.99985 0.99987 0.99989
##                            PC7     PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.13573 0.12042 0.11811 0.10987 0.09836 0.08618
## Proportion of Variance 0.00002 0.00001 0.00001 0.00001 0.00001 0.00001
## Cumulative Proportion  0.99991 0.99992 0.99993 0.99995 0.99996 0.99996
##                           PC13    PC14    PC15    PC16    PC17    PC18
## Standard deviation     0.07699 0.07573 0.07142 0.06231 0.05713 0.05471
## Proportion of Variance 0.00001 0.00001 0.00001 0.00000 0.00000 0.00000
## Cumulative Proportion  0.99997 0.99997 0.99998 0.99998 0.99999 0.99999
##                           PC19    PC20   PC21    PC22    PC23    PC24
## Standard deviation     0.05057 0.04637 0.0434 0.03668 0.03283 0.02953
## Proportion of Variance 0.00000 0.00000 0.0000 0.00000 0.00000 0.00000
## Cumulative Proportion  0.99999 0.99999 1.0000 1.00000 1.00000 1.00000
##                           PC25
## Standard deviation     0.02719
## Proportion of Variance 0.00000
## Cumulative Proportion  1.00000
```

```r
df_pca_scale = prcomp(df_data, scale=TRUE)
summary(df_pca_scale)
```

```
## Importance of components%s:
##                          PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     3.286 1.7211 1.25874 1.08177 1.02858 0.94911
## Proportion of Variance 0.432 0.1185 0.06338 0.04681 0.04232 0.03603
## Cumulative Proportion  0.432 0.5505 0.61386 0.66067 0.70299 0.73902
##                            PC7     PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.89489 0.86337 0.77500 0.75889 0.73070 0.68482
## Proportion of Variance 0.03203 0.02982 0.02402 0.02304 0.02136 0.01876
## Cumulative Proportion  0.77105 0.80087 0.82489 0.84793 0.86929 0.88805
##                           PC13    PC14    PC15    PC16    PC17    PC18
## Standard deviation     0.62838 0.60601 0.55324 0.52020 0.51393 0.48022
## Proportion of Variance 0.01579 0.01469 0.01224 0.01082 0.01057 0.00922
## Cumulative Proportion  0.90384 0.91853 0.93077 0.94160 0.95216 0.96139
##                           PC19    PC20    PC21   PC22    PC23    PC24
## Standard deviation     0.45402 0.43800 0.41772 0.3640 0.34223 0.32064
## Proportion of Variance 0.00825 0.00767 0.00698 0.0053 0.00468 0.00411
## Cumulative Proportion  0.96963 0.97731 0.98429 0.9896 0.99427 0.99838
##                           PC25
## Standard deviation     0.20107
## Proportion of Variance 0.00162
## Cumulative Proportion  1.00000
```

From the PCA summary, we can see that about 99.95% of the variability in the data is captured from the first principal component and 99.97% from the first two. Moroever, instead of working with each a predictor consisting of 25 features, we are able to just look at the first one or two and we can rest assured that we have captured the variability within the data.

When the principal componenets were scaled, the amount of variability captured in each of the components was much lower. Whereas the first pc in the unscaled case resulted in 99.95% of the variability, we don't see that level of precision until the 24th (out of 25) pc.

Part 2:

As described in lecture, one approach to the time series dimension reduction situation we were facing would be to first smooth the time series, and then use these smoothed time series as the input to a dimension reduction algorithm. We will try this here.

In particular, smooth each time series using `loess()`. I will leave the choice of the smoothing parameter up to you. After smoothing, you should use the `predict()` function to evaluate the fitted model on a regular grid of x values. These smoothed time series are what should be utilized in the dimension reduction. Use Isomap. Explore the first two-dimensions to see if there is meaningful low-dimensional structure in the plot.

Comment: If you watched lecture, this will make a lot more sense. Be sure to watch the lecture prior to attempting this.