

41201-01: Data Mining

Professor Taddy

**Problem Set 7**

Joanne Chen

James Hardiman

Evan Johnson

Honor code: We pledge our honor that we have not violated the Honor Code during the completion of this assignment.

**[1] Discuss correlation amongst dimensions of fx. How does this relate to the applicability of factor modelling?**

Because there are a number of positive correlations amongst all dimensions of fx that vary between 0 and 1 (see graph and code used to generate below) we suspect that there may be some underlying factors of which each dimension in fx is composed. This implies this data set maybe a good candidate for PCA. If were no correlations (i.e., dimensions of fx are already orthogonal) then there would be no need for PCA.

*corfx <- cor(fx)*

*hist(corfx)*



**[2] Fit, plot, and interpret principal components.**

Once we conduct PCA we can plot the variance of each component to obtain the following plot:



The first component accounts for most of the variance (~44% if we look at the summary output in R). The principal component seems to be geography with China, Hong Kong, and Japan having lower amounts of PC1 than the other countries.

**[3] Regress SP500 returns onto currency movement factors, using both ‘glm on ﬁrst K’ and lasso techniques. Use the results to add to your factor interpretation.**

Running glm on first K and lasso techniques we also get sparse solutions. Running the code below for a glm on first K we obtain a AICc and BIC solutions of K = 3. Using a lasso technique suggests ~4 (see plot).

*kfx <- lapply(1:20, function(K) glm(return ~., data=zfxdf[,1:K,drop=FALSE]))  
aiccfx <- sapply(kfx, AICc)  
which.min(aiccfx) #returns K = 3*

*kbic <- sapply(kfx, BIC)  
which.min(kbic)  
#min.bic also returns K = 3*

*lassofx <- cv.gamlr(x=zfx, y=return)  
plot(lassofx)*



**[4] Fit lasso to the original covariates and describe how it differs from PCR here.**

Running a simple lasso in R we obtain coefficients for each of the covariates. This differs from PCR in that in PCR we are trying to determine an orthogonal set of underlying components of which all the covariates are made (or at least a meaningful proportion of them). A simple lasso does not assume an orthogonal set of underlying components and chooses coefficients for the covariates given that minimizes the OOS deviance of the prediction.

Output of simple lasso

seg23

intercept 0.0009587991

exalus -0.0932269562

exbzus -0.0477638186

excaus .

exchus .

exdnus .

exhkus .

exinus .

exjpus .

exkous .

exmaus .

exmxus -0.4812048989

exnzus .

exnous .

exsius .

exsfus .

exslus .

exsdus -0.0795373350

exszus .

extaus .

exthus .

exukus .

exvzus .

exeuus .