Big Data: Homework 7

Will Clark & Matthew DeLio
41201-01

May 28, 2015

1 Foreign Exchange Factor Modeling

2 Principal Components Analysis

At first glance, our data set of monthly foreign exchange rates seems to have a lot of noise and very little signal. The goal of principal components analysis is to see if there are any patterns connecting variables in our data set, and if there are, use these relationships to reduce the dimensionality of our data. To this end, we use `prcomp` to find the principal components of foreign exchange movements. For each observation i, which is a month of foreign exchange rates for 23 currencies, the method estimates:

$$E[x_i] = \varphi_i v_{i,1} + \varphi_2 v_{i,2} + ... + \varphi_k v_{i,k} \tag{1}$$

We can now represent the data along the new set of dimensions $v_{i,j}$, which should reveal any latent patterns that were not observable when we were looking at the set of original dimensions $x_{i,j}$.

We can start by looking at the scree plot of our PCA, shown in Figure 1. This shows us the sorted eigenvalues of the covariance matrix of the scaled data; the highest eigenvalue is the principal component that explains most of the variation in the data. The steep drop off after the first bar tells us that the first principal component explains a large degree of the variability in our data.

We can look at the rotations on the first principal component and see if there is any obvious interpretation. Figure 2 shows the rotations of PC1 on each country; countries in red have floating exchange rates and countries in blue have fixed exchange rates (Venezuela, China, Hong Kong, and Sri Lanka). Given that all the pegged exchange rates are on one side and all the floating rates are on the other, we can tentatively conclude that the first principal component is really telling us about the fixed/floating divide. It makes sense that most of the variation in exchange rates would occur between those that are allowed to move freely and those that are not, and since PCA is supposed to find the latent sources of variation, it follows that this is the first dimension on which it chooses to sort the data.

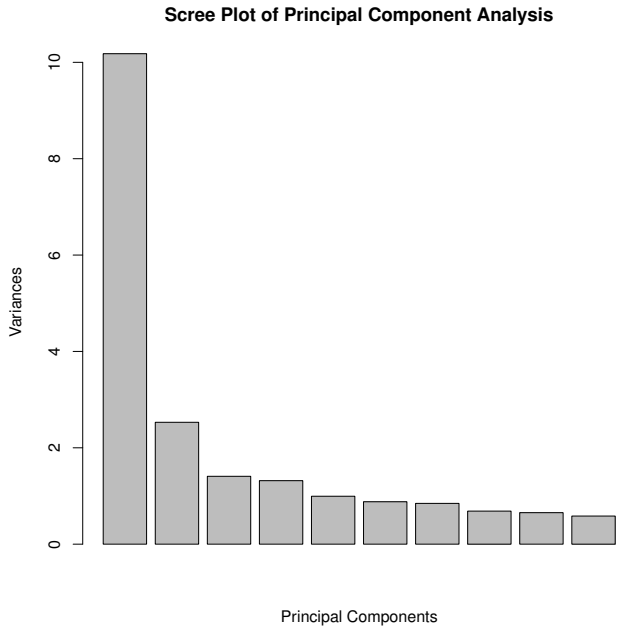

Figure 1: Foreign Exchange PCA Scree Plot

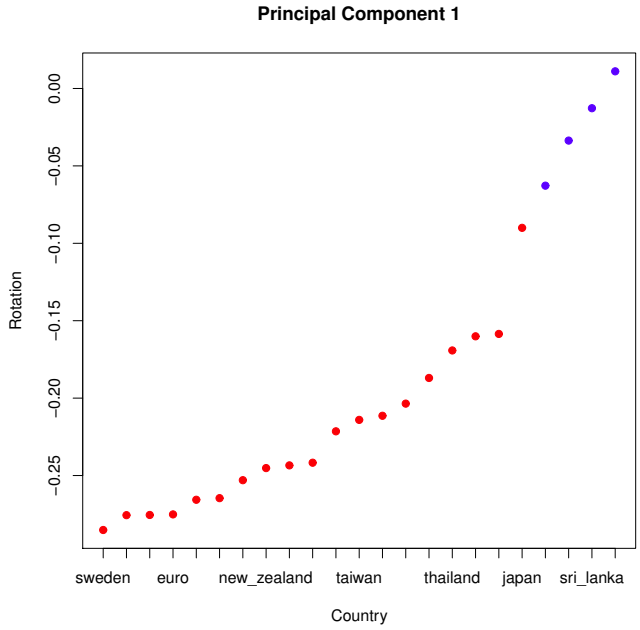

Figure 2: Distribution of First Principal Component

3 S&P 500 Returns on Currency Factors

We can use our principal components from Section 2 to estimate monthly returns of the S&P 500. First, we use only the first principal component in a standard regression model:

$$R_{SP,i} = \alpha + \beta_1 z_{1,i} + \varepsilon_i \tag{2}$$

where $z_{1,i}$ is each observation projected in $v_{i,j}$ space. The in-sample R^2 of this regression is around 16 percent. Instead of running a standard linear regression on only one component, we can throw all the principal components into a lasso and see which coefficients are chosen. The model selected by the various decision criteria are shown in Table 1. The model chosen by CV.min includes 16 of the possible 23 covariates, which is a fairly complex model given that PCA is intended to be a dimension reduction exercise.

	$\log(\lambda)$	R^2	Covariates Selected
AICc	-5.83	0.49	13
AIC	-6.20	0.52	16
BIC	-5.41	0.43	10
CV.Min	-6.20	0.37	16
CV.1se	-5.08	0.26	9

Table 1: ICs for S&P500 Returns Regressed on FX Principal Components

The average out-of-sample R^2 of the cross-validated lasso regressions are 31 percent (for the CV.min model) and 21 percent (for the CV.1se model), so adding the extra set of principal components does increase the predictive accuracy of the model relative to the simple linear regression on only one principal component. Interestingly, the principal component with the highest coefficient in the CV.min lasso regression is PC23, the one that explains the least variation in the data set. If we look at the rotations on PC23, we see that all 21 are basically equal to zero, and one is very high and one is very low. The two outliers are the Euro and the Danish krone, which is pegged to the Euro. This means that the strongest predictor of S&P returns in the currency data is just the US/EUR exchange rate, even though this is the dimension that explains the least amount of variation in the data to begin with.

4 Regression on All Covariates

In this section, we estimate a lasso regression of S&P 500 returns on all the foreign exchange covariates. The models selected by each decision criterion are listed in Table 2. We can see that there is a lot of variation in the models selected; the CV.min model includes 18 covariates while the CV.1se model only includes 3. Note that the average out-of-sample R^2 in this case 28 percent for the CV.min and 17 percent for the CV.1se model; this means we have lost predictive accuracy relative to the lasso regressions run on the principal components in Section 3.

The difference between this set of regressions and the set in the previous section is that here we are regressing S&P returns on the variables in $x_{i,j}$ space (the original set of dimensions). In the prior case, we were regressing market returns on the variables projected into $v_{i,j}$ space, which we estimated with our `prcomp` method. The former case should allow us to include only covariates that are independent of each other, because the principal components aggregate all sources of common variation in the data. In the latter set of regressions, we are including all the raw covariates, some of which we know from our heat map in Section 1 move quite closely with each other. Our expectation is that the PCA regression would give us better results, and this is validated by the better out-of-sample prediction generated by these regressions.

Another possibility is to run a lasso on all raw covariates and all principal components and see

	$\log(\lambda)$	R^2	Covariates Selected
AICc	-7.20	0.50	18
AIC	-7.20	0.50	18
BIC	-5.01	0.33	5
CV.Min	-7.01	0.28	18
CV.1se	-4.13	0.17	3

Table 2: ICs for S&P500 Returns Regressed on FX Movements

which coefficients are selected. The results of this lasso are presented in Table 3. The model selected by CV.min has almost the same out-of-sample R^2 as the CV.min model of the regression on the raw covariates; we have not gained anything by including the raw covariates along with our principal components even though only 13 of the 20 covariates it selects are principal components. This suggests that if we were to use forex rates as a predictor of market returns, we would be better off using only a model that included principal components.

	$\log(\lambda)$	R^2	Covariates Selected
AICc	-5.39	0.44	11
AIC	-6.69	0.54	20
BIC	-4.32	0.23	3
CV.Min	-6.18	0.27	20
CV.1se	-4.18	0.15	3

Table 3: ICs for S&P500 Returns Regressed on FX Movements and Principal Components