QBUS6830 Financial Time Series and Forecasting
S1, 2019

Solutions to Lab Sheet 5

Q1 (PCA and Factor modelling)
We use the data from the text by Tsay in Chapter 9, being monthly returns on IBM, HPQ, Intel, JP Morgan and Bank of America, from January, 1990 to December, 2008.

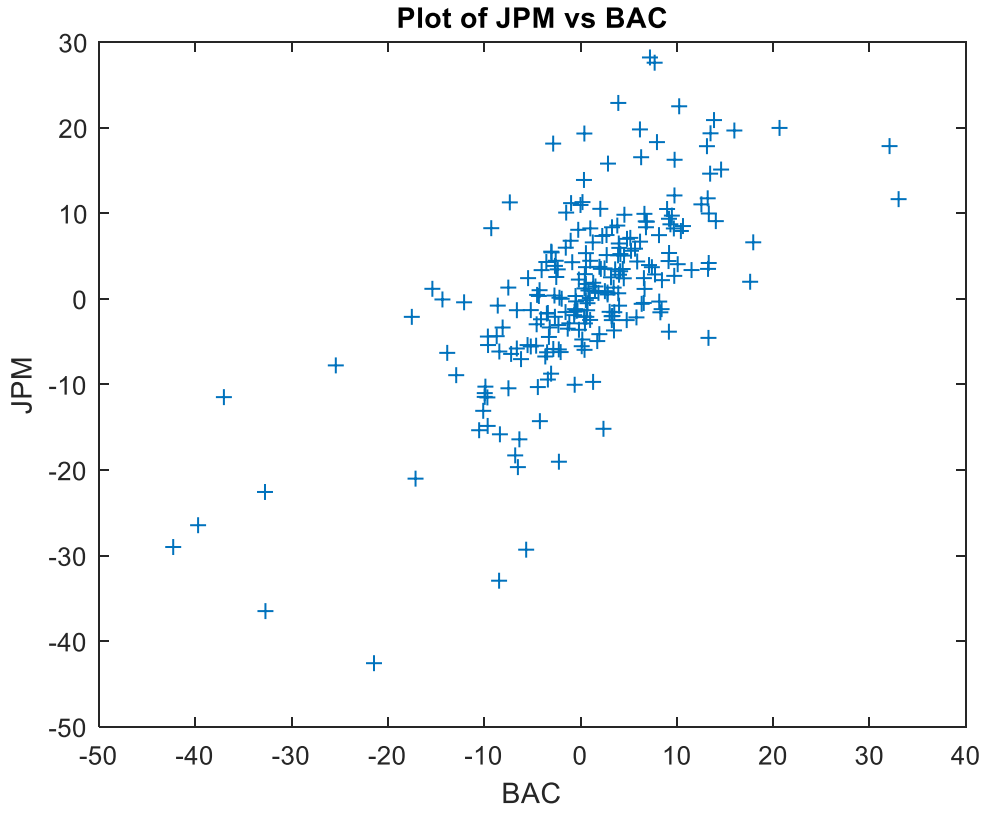
(a) Find the sample correlation matrix and assess whether the data is suitable for a PCA

The sample correlation matrix is:

	BAC	HPQ	IBM	INTC	JPM
BAC	1.0000	0.2591	0.2545	0.2521	0.6836
HPQ	0.2591	1.0000	0.4620	0.5495	0.3889
IBM	0.2545	0.4620	1.0000	0.4593	0.3384
INTC	0.2521	0.5495	0.4593	1.0000	0.3578
JPM	0.6836	0.3889	0.3384	0.3578	1.0000

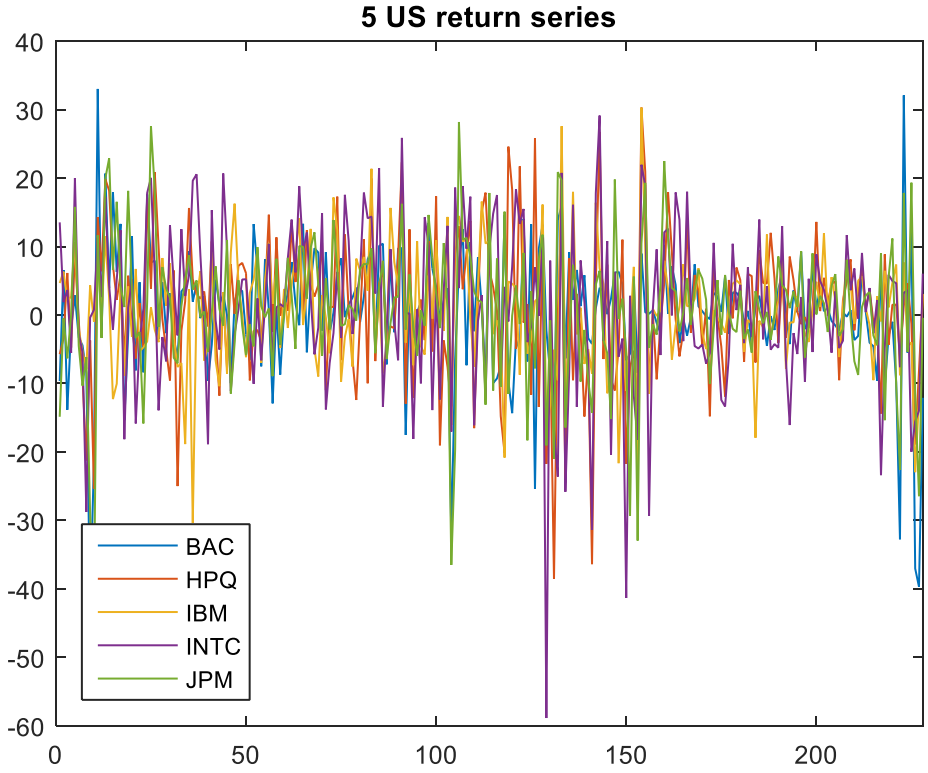
Clearly none of these are close to 0 or to 1: pairwise correlations range from 0.25 up to 0.68. Thus the asset return series here are all positively correlated and none are collinear (correlation = 1). All correlations are significantly different to 0. A hypothesis test for H_0: $\rho = 0$ H_0: $\rho \neq 0$ for the smallest correlation of 0.2521 (between BAC and INTC) was found to have a p-value of 0.00011885 and the null is clearly rejected. P-values for the test on the other larger correlations would be even smaller. So these variables are not independent and seem to show some relationships.

A scatterplot of the two highest correlated variables indicates that they appear to show a roughly linear relationship. This seems like an ideal candidate dataset for PCA.



(b) Perform a PCA on this data and report the results. Form the 2D and 3D biplots of the first 2 and 3 components. Discuss.

The 5 times series are plotted here:



The low volatility period leading up to 2008 is readily apparent, as is the increase in volatility marking the GFC period.

The table below shows the weights applied to each individual asset series to form each PC (i.e. the eigenvectors). The PCs are listed in order of importance regarding how much overall variability they capture in the data as a whole.

Principal components

	1	2	3	4	5
BAC	0.3474	0.6097	-0.1188	-0.0093	0.7024
HPQ	0.4826	-0.2786	0.7009	-0.4298	0.1159
IBM	0.3298	-0.1393	0.2643	0.8954	0.0144
INTC	0.5808	-0.4781	-0.6516	-0.0962	0.0163
JPM	0.4476	0.5502	-0.0128	-0.0642	-0.7019

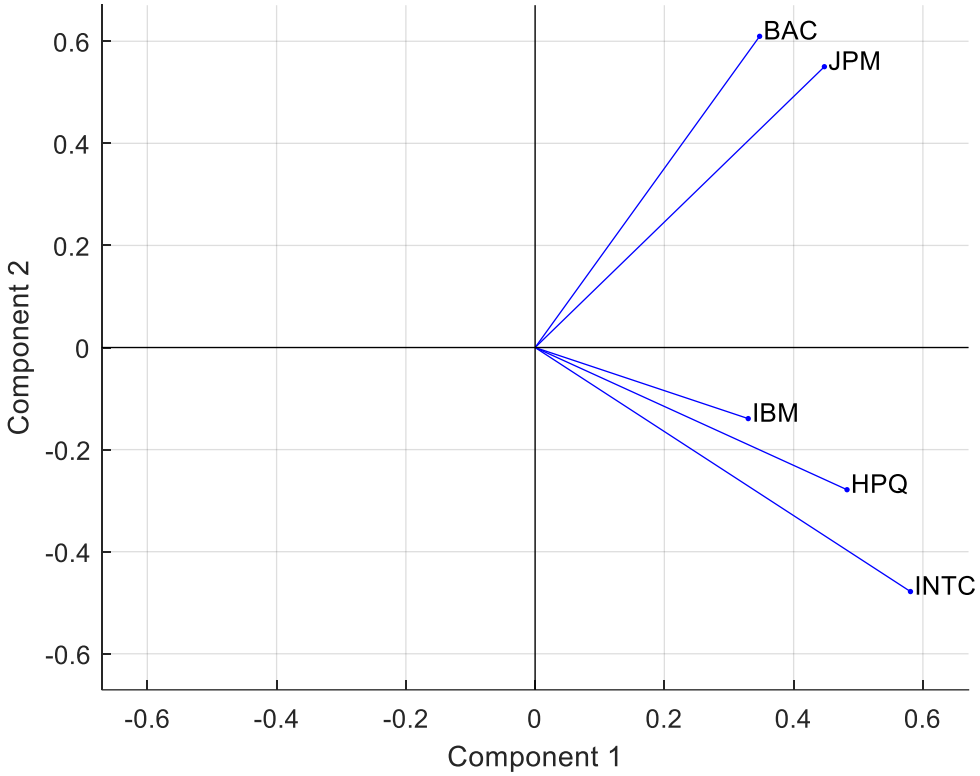
Lambdas

	1	2	3	4	5
Value	0.5349	0.2126	0.1081	0.0881	0.0562
Cumulative	0.5349	0.7475	0.8557	0.9438	1.0000

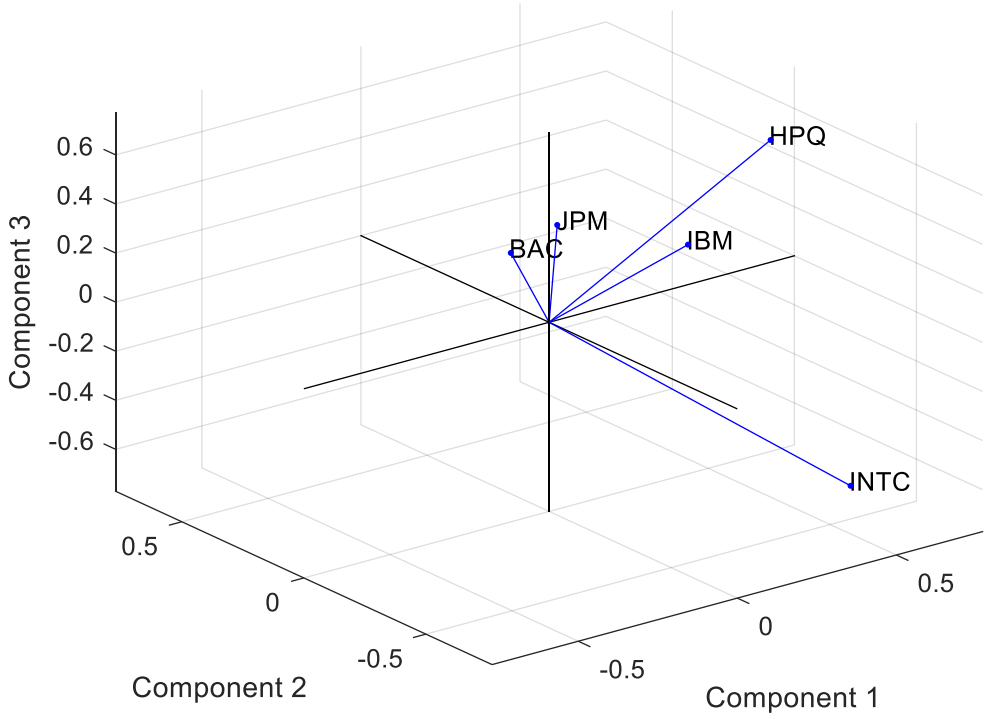
The 1st component captures 53.5% of the total variance. The 2nd captures a further 21.3% of total variance, the 3rd captures 10.8% etc. The first three PCs combined capture 86% of the total variance in the data, the 1st 4 components capture 94% of that variance.

3 or 4 components seem needed, and most analysts would likely choose one of these two options.

A biplot of the first 2 components is:



The 2D biplot shows that all stock series load positively on factor 1 and that the second factor splits the Banking stocks (BAC and JPM) from the (tech stocks (IBM, HPQ and INTC)
A biplot of the first 3 components is:



The 3D biplot is harder to examine in 2D! In the lab if you should pass the mouse over the circular arrow on the figure window. It will say "Rotate 3D". If you click it you can then rotate the figure as you like to get a better perspective of each component represented.

(c) How many principal components do you think are adequate to explain these variables? Describe the (relevant) PCs found, do they make sense or have a relevant or useful interpretation?

3 or 4 components seem needed to adequately represent the variance in the data, and most analysts would likely choose one of these two options. Usually analysts like to capture at least 80% of the variance in the data.

The 1st component weights close to equally on all assets: it thus seems like a market factor. (Note that the sign of the weights is not really relevant, so sometimes all will be negative on the "market" factor, but the interpretation is the same. Since the sum of squares of weights add to 1, signs may not be unique in components). This component weights lowest on IBM, highest on INTC. This may indicate that IBM is least affected by the market OR it may be because IBM has the lowest variance (which it does here, being 74.6; remember this component wants to maximise the variance, so down-weighting low varying return series is natural and intuitive) OR a bit of both. The highest weighted component is INTC, which also has the highest sample variance (of 146.5) among these assets (again this component maximises the variance, so highest weight on highest varying return series makes sense).

Variance-Covariance matrix estimate

	BAC	HPQ	IBM	INTC	JPM
BAC	91.8261	26.2997	21.0695	29.2428	67.4529
HPQ	26.2997	112.2189	42.2812	70.4519	42.4237
IBM	21.0695	42.2812	74.6379	48.0269	30.1030
INTC	29.2428	70.4519	48.0269	146.4969	44.5942
JPM	67.4529	42.4237	30.1030	44.5942	106.0369

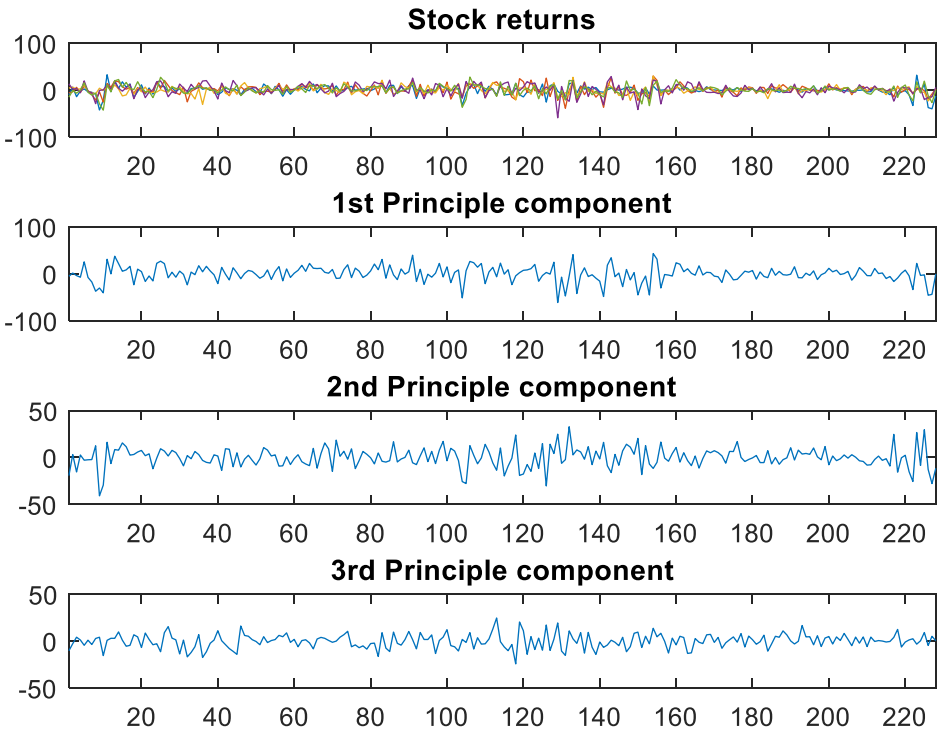
The 2nd component seems to contrast IBM, HPQ and INTC against JPM and BAC: i.e. a tech-stock vs financial stock contrasting component. (Again signs are only relevant in groups, so multiplying all weights in this component by -1 leads to exactly the same interpretation).

The 3rd component seems to mainly contrast HPQ (and IBM) with INTC, a within tech-stock contrasting component.

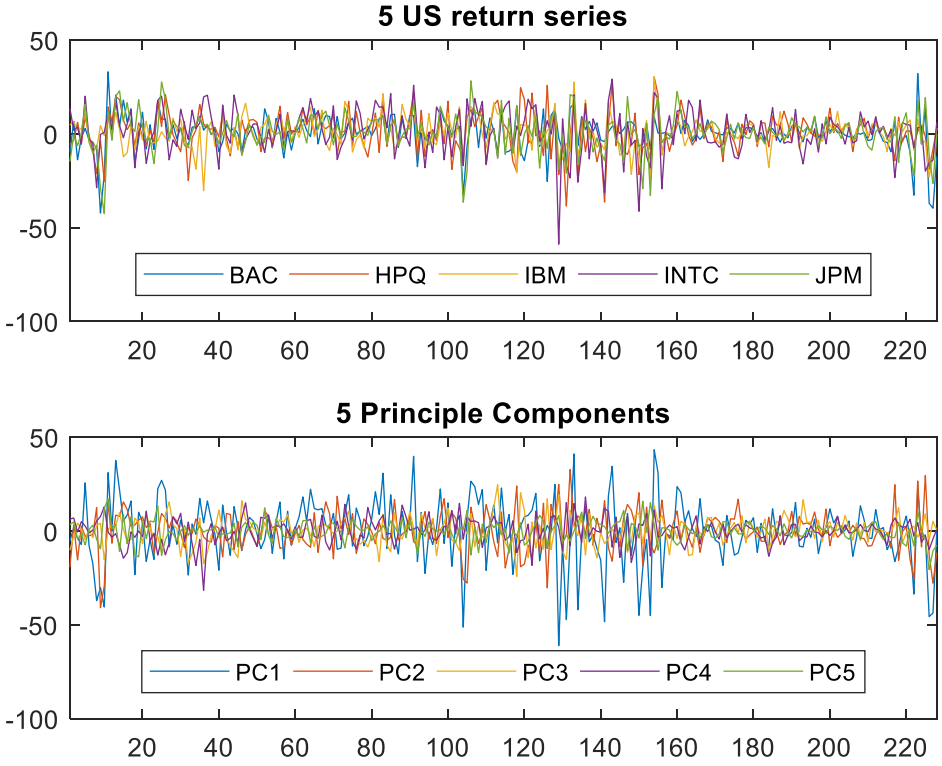
The 4th component seems to contrast IBM and HPQ, another within tech-stock contrasting component.

Finally, the 5th component seems to contrast JPM and BAC, a within financial-stock contrasting component.

The plot below shows the 5 asset return series on the same plot. Below that are plot the 1st three PCs over time. Remember that these PCs are uncorrelated with each other.

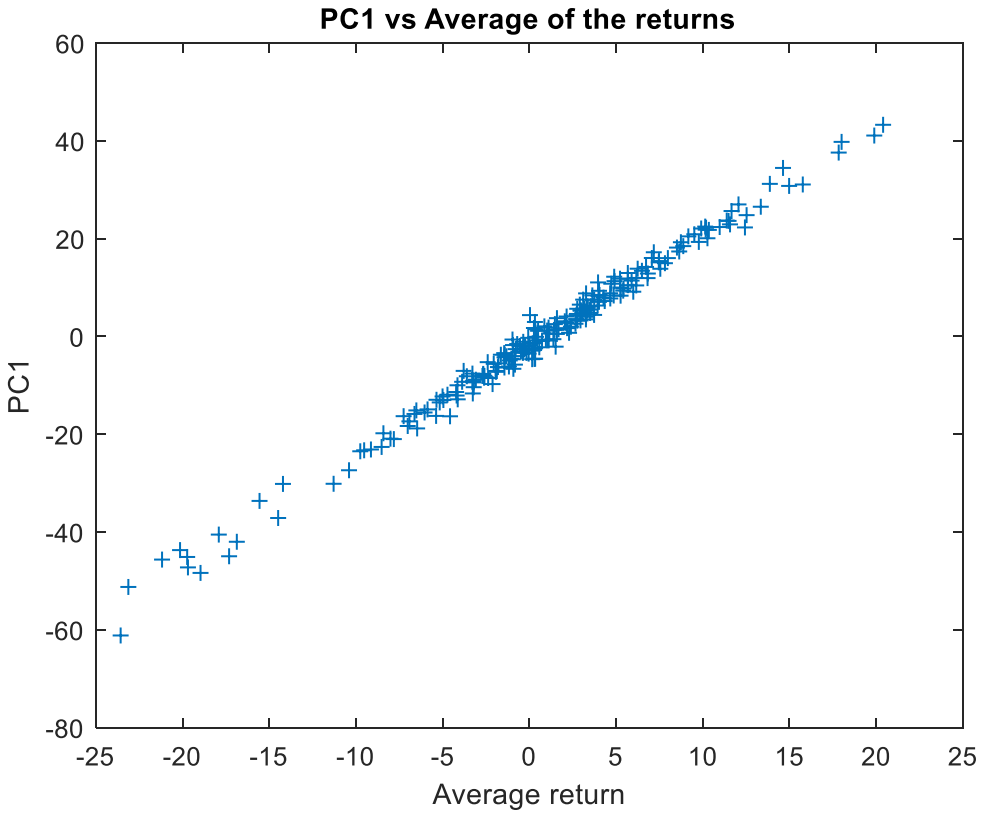


The first component seems to capture the average overall return movements across these 5 assets. It is clear why it captures the most variation in the 5 series. It seems hard to see exactly what the 2nd and 3rd PCs are capturing, though the 2nd is more active than PC1 during 2008 and the GFC.



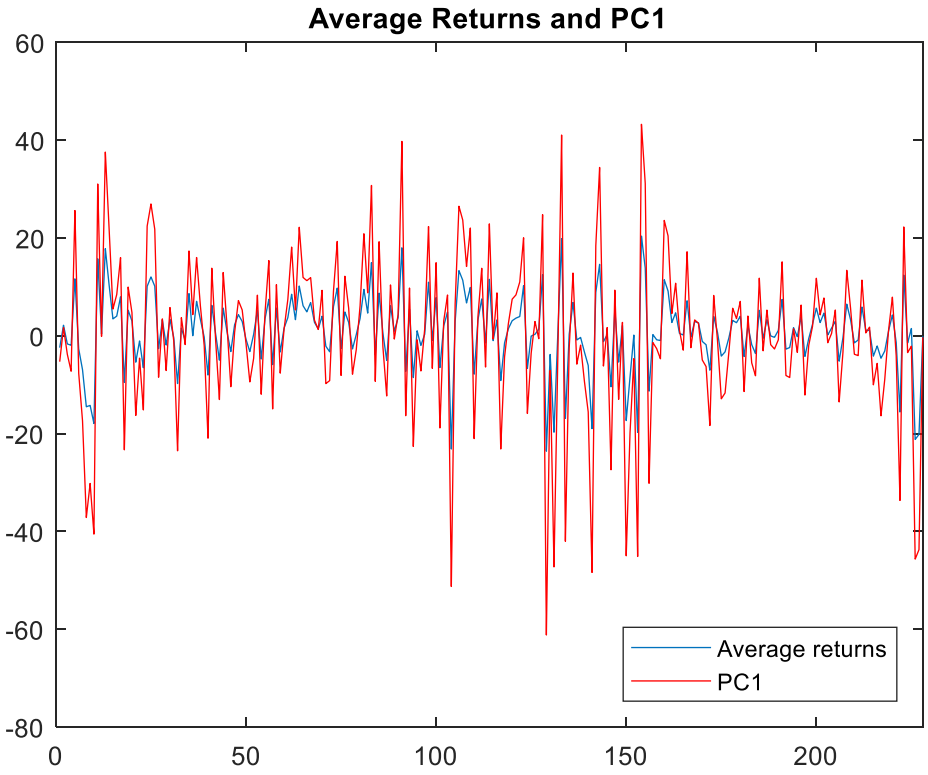
The plot above again shows the five asset returns series on top and then below shows all 5 PCs overlaid and over time. The 1st PC is in blue, and clearly captures most of the up and down movements across the 5 asset series, with emphasis on the big down movements in 2008. The 2nd PC is in green and it is prevalent during the up-swings in 2008, clearly affected by the positive weights on BAC and JPM which rose sharply in price at those times. The 3rd PC is in red, and again it is difficult to see what that one is capturing.

The plot below shows the average over the 5 assets for each month versus the 1st PC. The correlation here is 0.995. The final plot shows the monthly mean of asset returns and the 1st PC (in red).



z

Clearly PC1 captures the average or mean monthly return across the assets as well as most of the variance of these assets.



(d) Perform a Factor Analysis with m=1 factor on these return series.

Matlab first standardises the data by subtracting off the vector of means and then dividing each series by its standard deviation. The initial factor output is thus for the standardised series and must be transformed. For m=1, the estimated standardised data factor loadings are:

BAC	HPQ	IBM	INTC	JPM
0.7261	0.5196	0.4708	0.4969	0.8426

After transforming, the actual estimated factor loadings are:

BAC	HPQ	IBM	INTC	JPM
6.958	5.505	4.067	6.014	8.677

Remember that these are regression coefficients that multiply the estimated factor to estimate the conditional average of each series of returns. The factor has an assumed mean of 0 and a variance of 1, though the estimated factor has a sample mean of 4.6×10^{-17}, and a sample variance of 1.22. Thus, these factor loadings are large since the asset returns themselves have variances close to 100 (i.e. standard deviations close to 10%) and these size loadings are required to multiply a mean 0, variance 1 factor to estimate the magnitudes of returns in the five series.

The estimated error variances, and sample variances, for each asset are

	BAC	HPQ	IBM	INTC	JPM
SER^2	43.42	81.92	58.10	110.33	30.75
s^2	91.83	112.22	74.64	146.50	106.04

From these, we can work out the SER and R^2 for this factor model for each asset.

	BAC	HPQ	IBM	INTC	JPM
SER	6.5893	9.0510	7.6221	10.5038	5.5454
R^2	0.5272	0.2700	0.2216	0.2469	0.7100
s	9.5826	10.5933	8.6393	12.1036	10.2974

The last row shows the sample standard deviation in each series. Note, the R^2s are all between 22 and 71% and are quite variable across the series! The SERs in each case are between 5.5% and 10.5%, showing reasonable but not ideal accuracy and medium level error for monthly returns, compared to their original standard deviations, that go from 8.6% to 12.1%!!

Overall the model R^2, across all five assets combined, is 38.9%. Note that this is comparable to the 1st PC above which captured 53% of the variance in these returns. Surprisingly, the linear 1st PC has captured more variance than the nonlinear factor here.

(e) Describe the factor loadings and factor found: do they make sense or have a relevant or useful interpretation?

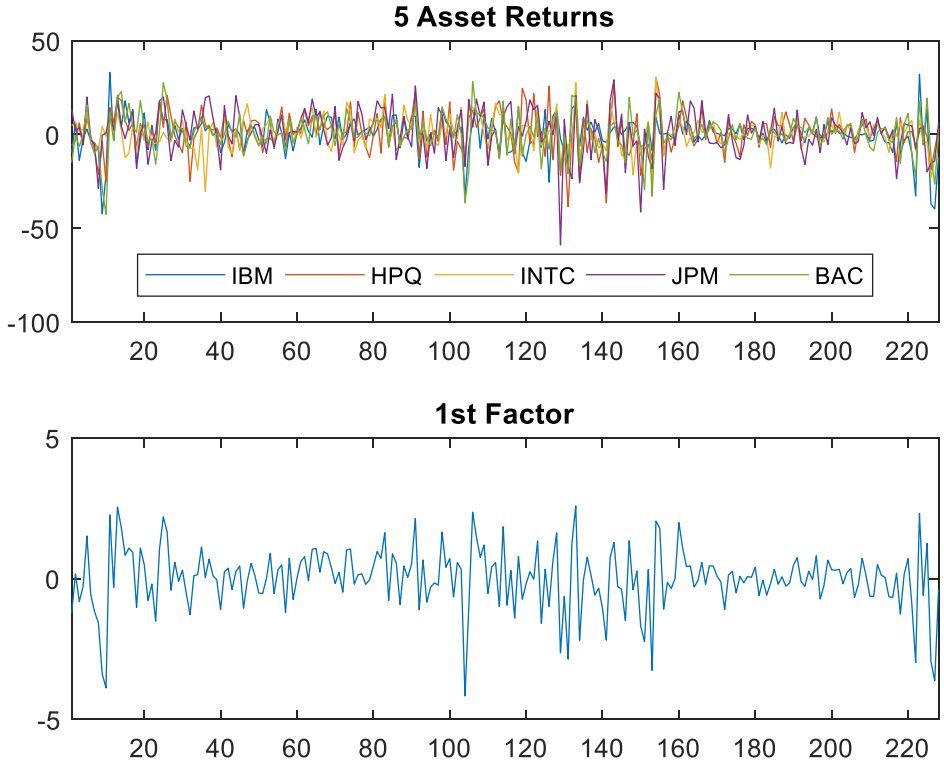
For m = 1, the estimated factor loadings are:

BAC	HPQ	IBM	INTC	JPM
6.958	5.505	4.067	6.014	8.677

This could be described as a "market" factor, since all loadings are positive and are not too spread out from each other, i.e. the assets seem similarly positively affected by the unknown factor. However, the loadings on the financial stocks, JPM and BAC, do seem somewhat higher than those on the tech-related stocks. So, the assets could be described as loading roughly equally on this 1st factor, similar to a market

factor, but with slightly higher loadings for financials and slightly lower loadings on tech stocks; i.e. this factor affects all assets positively, but more strongly influences financial assets.

The plot below shows the five asset return series and their estimated underlying factor below.



This factor captures 38.9% of the variation in the five assets combined, and the conformance here is obvious from the plot.

(f) Assess whether this 1 factor model is appropriate for this data.

This single factor here is positively correlated with each asset, the sample correlations are below:

	BAC	HPQ	IBM	INTC	JPM
Correlation	0.8020	0.5739	0.5200	0.5488	0.9307
Loading	6.9575	5.5045	4.0671	6.0139	8.6767
R^2	0.5272	0.2700	0.2216	0.2469	0.7100

Naturally these correlations are proportional to the factor loadings and the R^2s, by design. Clearly, the financial stocks are best predicted by this factor, which is not a good predictor of the tech-related stock returns. Overall the model R^2, across all five assets combined, is only 38.9%, so a two factor model seems warranted. Note that this is comparable to the 1st PC above which captured 53% of the variance in these returns. Surprisingly, the linear 1st PC has captured more variance than the nonlinear factor here. Note the correlation between the 1st PC above and this single factor is 0.904.

The chi-squared test of model fit to the data, which has a null hypothesis of: The single factor model (m=1) is adequate for this data, has a p-value of $2.5 \times 10^{-18} \sim= 0$. The chi-squared statistic value is 92.02, and the p-value is the chance of getting a chi-squared (with 5 df) that large. We can thus strongly reject the single factor model. Again, a 2 factor model seems called for.

(g) Perform a Factor Analysis with m=2 factors on these return series. Describe the factor loadings and factors found: do they make sense or have a relevant or useful interpretation? Assess whether this 2 factor model is appropriate for this data. How many factors should we choose to use?

For *m*=2 factors, the estimated standardised factor loadings are:

	BAC	HPQ	IBM	INTC	JPM
Factor 1	0.1273	0.7334	0.5937	0.7093	0.3605
Factor 2	0.9577	0.1739	0.1868	0.1680	0.6659

We can see factor 1 loads highly on the tech stocks and much lower on the financials. Factor 2 loads highly on the financials and low on the techs.

The actual factor loadings on the untransformed data are:

	BAC	HPQ	IBM	INTC	JPM
Factor 1	1.2196	7.7696	5.1296	8.5856	3.7117
Factor 2	9.1773	1.8423	1.6141	2.0338	6.8567

The estimated error variances for each asset are

	BAC	HPQ	IBM	INTC	JPM
m=1	43.419	81.920	58.097	110.330	30.751
m=2	6.116	48.458	45.720	68.648	45.246

The 2 factor model seems to have added some more explanatory power, except for the JPM series.

From these, we can work out the SER and R^2 for this factor model for each asset.

	BAC	HPQ	IBM	INTC	JPM
s	9.5826	10.5933	8.6393	12.1036	10.2974
SER (*m*=1)	6.5893	9.0510	7.6221	10.5038	5.5454
SER (*m*=2)	2.4730	6.9612	6.7616	8.2854	6.7265
R^2 (*m*=1)	0.5272	0.2700	0.2216	0.2469	0.7100
R^2 (*m*=2)	0.9334	0.5682	0.3874	0.5314	0.5733

The first row shows the sample standard deviation in each series. Note that the individual R^2s are all increased from *m*=1 to 2, except for JPM and are now all above 38%, going up to 93% for BAC. Note that these are ADJUSTED R^2 measures, i.e. they could reduce as *m* increases, as has happened for JPM.

The SER in each case for *m* = 2 is now between 2.5% and 8.3%, again showing more reasonable level of accuracy and lowish standard error for monthly returns, compared to their original standard deviations.

Overall the model adjusted R^2, across all five assets combined, is 59.7%. Once again this is surprisingly only just better than a single linear PC, and worse than the 2 component PCA, which captured about 75% of the total variations in these series.

The chi-squared test examining the adequacy of the fit for the 2 factors has a p-value of 0.599 and a chi-squared value of only 0.276; i.e. the chance of getting a chi-squared (1) as high as 0.276 is 0.599. Thus, we cannot reject the two factor model and this suggests that the two factor could be preferred to the single factor model in this case.

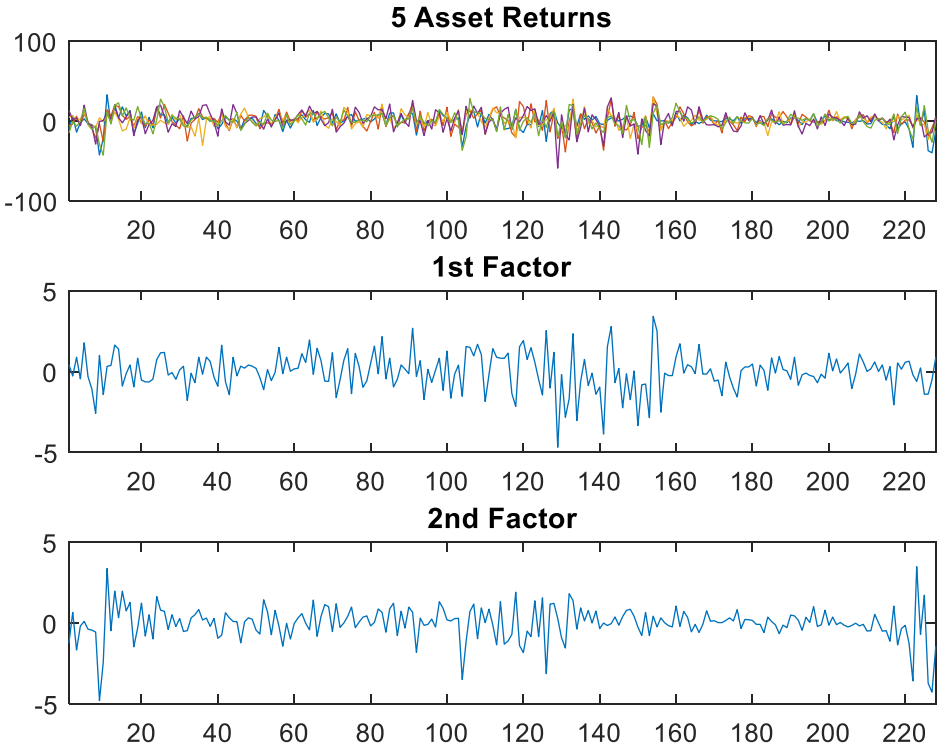
For this *m*=2 factor model, the estimated factor loadings are:

	BAC	HPQ	IBM	INTC	JPM
Factor 1	1.2196	7.7696	5.1296	8.5856	3.7117
Factor 2	9.1773	1.8423	1.6141	2.0338	6.8567

The first factor again loads positively on each of the five series, but now the three tech assets are much more highly affected or loaded than JPM and BAC (BAC has a very small comparative loading). This is more like a factor that drives returns on tech stocks, with some smaller loading on JPM. Not really a market-like factor, more a tech-dominated plus JPM factor.

The second factor puts smallish loadings on the tech stocks and much larger loadings on JPM and largest load on BAC. This factor is driving the financial asset returns.

The plot below shows the five asset return series and their two estimated underlying factors below.



These factors capture 59.7% of the variation in the five assets combined, and the conformance here is again obvious from the plot. In particular, we see that in the middle part of the sample, around the time of the tech-bubble bursting, the first factor (which loads highly on tech assets) is capturing the movements in the returns, but not much is happening for this factor at the time of the crisis. However, the 2nd factor (which loads highly on financial assets) dominates and captures the variance better at the start of the sample, 1990s and the end of the sample, 2008 and GFC period, when financial assets did especially poorly.

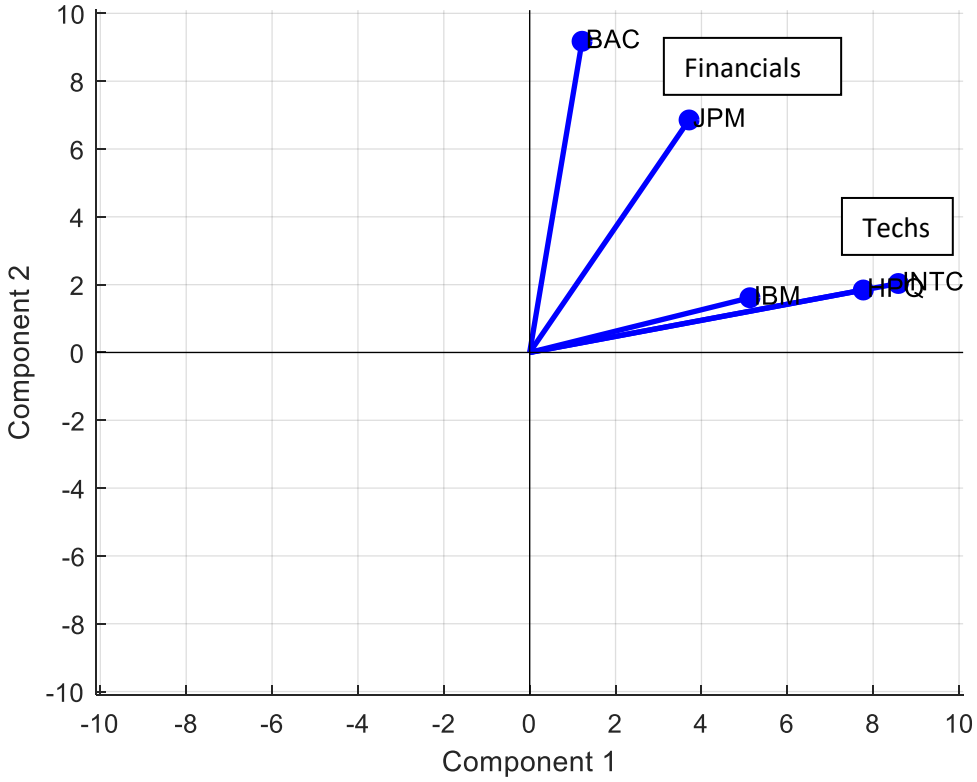
Both factors are positively correlated with each asset, the sample correlations are below:

	BAC	HPQ	IBM	INTC	JPM	Other factor
1st factor	0.0870	0.8436	0.6800	0.8159	0.3775	-0.0615
2nd factor	0.9872	0.1283	0.1518	0.1239	0.6669	-0.0615

Naturally these are proportional to the factor loadings, by design.

The two factors are meant to be uncorrelated, but the sample correlation between them is slightly negative and not 0 as intended (it is meant to be a restriction).

For the two-factor model, a "bi-plot" shows the factors and loadings graphically:



It shows the separation of three series (the tech assets) and the two financials in each factor (component)

(h) How about the un-rotated 2 factors? Do these have a better interpretation?

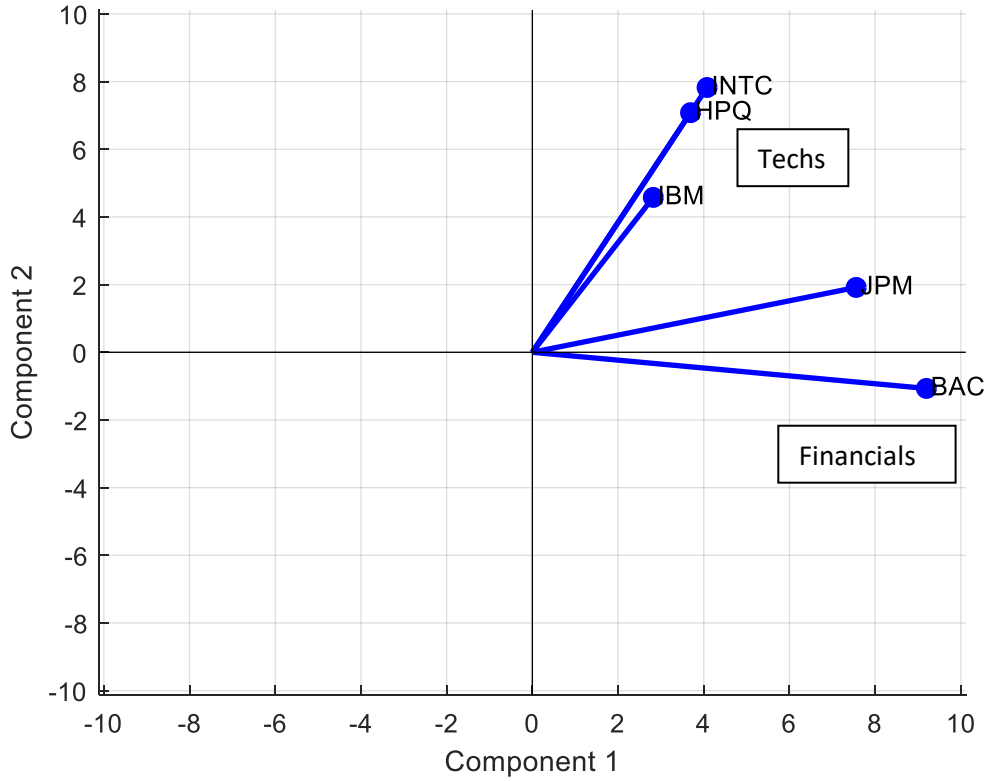
A single factor cannot be rotated.

For *m*=2, the 'unrotated' factors are

	BAC	HPQ	IBM	INTC	JPM
Factor 1	9.1961	3.6917	2.8230	4.0776	7.5576
Factor 2	-1.0686	7.0804	4.5770	7.8245	1.9165

These unrotated factors are roughly simply the original two factors swapped. The first factor is now loading highly on the financial assets, while the 2nd is loading highly on the tech stocks.

The unrotated factors have the following bi-plot, essentially showing the same thing as above.



Other rotations are possible.

The *m*=3 factor model could not be estimated by Matlab, which informed me that 3 was too many factors for it to estimate.

(i) Compare the PCA, 1 and 2 factor models in terms of model fit, adequacy and usefulness.

It seems the PCA has done a better job at explaining these asset returns than the Factor model has. In particular, the 1 and 2 component PCA model captured more of the total variation than the *m*=1 and 2 factor models, respectively. Further, the 3 and 4 component PCA models seemed to do best overall, but a 3 factor model could not be estimated.

Since PCA components can be used as factors, it may be most useful to employ the 3 component PCA as a factor model. The PCA components also seem more useful here because they are easy to calculate and replicate into the future (they are simply weighted sums of these asset returns series) and they have clear interpretations that intuitively make some sense.