Module 1, section 3 : Factor and Multivariate Modelling

Chapter 9 in Tsay
Chapter 3 in McNeil, Frey, & Embrechts
Chapter 3 in Brooks (Appendix 3.2)
Chapter 5 in Campbell, Lo and MacKinlay

1 Unknown factor regression models

- Consider the multiple regression model:

$$y_t = \alpha + \beta_1 f_{1,t} + \beta_2 f_{2,t} + \epsilon_t,$$

which is a 2 factor model.

- OR in general

$$y_t = \alpha + \beta_1 f_{1,t} + \beta_2 f_{2,t} + \ldots + \beta_m f_{m,t} + \epsilon_t,$$

which is an m-factor model.

- When factors are observed, this is just multiple regression.

- However, financial economists don't believe they know all the relevant factors that might explain stock returns.

- And/or they don't think these factors can all be observed, e.g. investor sentiment, mood of the market, etc.

- Can we estimate BOTH the unobserved factors and their coefficient βs?

- Not from a single series of data! *Why not?*

- Turns out that we need to consider a multivariate setting to do so.

- i.e. we need to consider more than 1 asset return series as joint dependent variables to estimate unknown factor models.

- We need to learn a bit about multivariate settings first.

Multivariate Random Variables

- Consider that the vector $(y_{1,t}, y_{2,t}, \ldots, y_{n,t})'$ represents the realisation of a vector random variable $\mathbf{Y} = (Y_1, Y_2, .., Y_n)'$ at time t.

- The $n \times 1$ vector $E(\mathbf{y}) = \mu$ contains each $E(y_i) = \mu_i$, so that $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^{'}$.

- Also the $n \times n$ variance-covariance matrix $\text{Var}(\mathbf{y}) = \Sigma$, means that $\text{Var}(y_i) = \sigma_i^2$ and $\text{Cov}(y_i, y_j) = \sigma_{ij}$, where

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \cdot & \cdot & \ldots & \cdot \\ \sigma_{1n} & \sigma_{2n} & \ldots & \sigma_n^2 \end{bmatrix}$$

- A correlation matrix can always be formed via a matrix decomposition

$$\Sigma = DRD$$

where $D = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ so that $R_{ij} = \sigma_{i,j}/(\sigma_i \sigma_j)$.

- R is thus the matrix of pairwise correlations between the variables.

- The matrix Σ is usually assumed to be a positive definite matrix, i.e. $\mathbf{a}'\Sigma\mathbf{a} > 0$, for all vectors $\mathbf{a} \neq \mathbf{0}$.

- This implies that Σ is invertible, so that Σ^{-1} exists and is unique, which is required in many statistical and econometric estimation techniques.

- This basically implies that the rvs in $\mathbf{Y} = (Y_1, Y_2, .., Y_n)'$ are not perfectly correlated with each, are not perfect linear combinations of each other, and none of them have 0 variance (are constant).

- These are pretty much the same conditions that the regression matrix $\mathbf{X}$ had to satisfy in multiple regression.

3 Principal Component Analysis (PCA)

- The mathematics in this section will NOT be assessed. I am presenting it FYI only and to assist with your understanding and comprehension of what factor modeling is actually about.

- PCA is a simpler, special case of factor modeling. In fact, principal components are one example of a set of unknown factors.

- PCA is also a general method used to reduce the dimensionality of a set of multivariate data.

- Consider again that the vector $\mathbf{y}_t = (y_{1,t}, y_{2,t}, \ldots, y_{n,t})$, represents the time t realisation of a vector random variable $\mathbf{Y} = (Y_1, Y_2, .., Y_n)'$.

- We observe this vector of n variables for times $t = 1, \ldots, T$.

- This is a panel data setting.

- The basic idea of PCA is to take a set of m linear combinations of the original variables $\mathbf{Y}$, in a way that ensures most (or all) of the variation in $\mathbf{Y}$ is retained in these m factors.

- Dimensionality is then reduced, if $m < n$.

- PCA is appropriate only when $n > 1$.

- It tries to understand the structure of a set of **related** variables $\mathbf{Y}$, while reducing their dimension.

- If the variables in $\mathbf{Y}$ are not correlated with each other, then there is NO point doing PCA. *why?*

- We search for m "optimal" linear combinations of $\mathbf{Y}$, where $m < n$.

- Say we have T observation vectors $\mathbf{y}_{\cdot t} = (y_{1,t}, y_{2,t}, \ldots, y_{n,t})^{'}$ which represent realizations of a vector random variable $\mathbf{Y} = (Y_1, Y_2, .., Y_n)'$.

- In PCA we find a linear combination of the n observed vectors $\mathbf{y}_{i\cdot} = (y_{i,1}, y_{i,2}, \ldots, y_{i,T})^{'}$, $i = 1, \ldots, n$, i.e.

$$\begin{aligned} \mathbf{w}_1 &= a_{11}\mathbf{y}_{1\cdot} + a_{12}\mathbf{y}_{2\cdot} + \ldots + a_{1n}\mathbf{y}_{n\cdot} \\ &= \mathbf{a}_1^{'}\mathbf{y} \end{aligned}$$

  so that Var($\mathbf{w}_1$) is a **maximum**.

- Note that Var($\mathbf{w}_1$) $= \mathbf{a}_1^{'}$Var($\mathbf{y}$)$\mathbf{a}_1$. If we set Σ =Var($\mathbf{y}$), then Var($\mathbf{w}_1$) $= \mathbf{a}_1^{'}\Sigma\mathbf{a}_1$.

- This is a non-linear maximisation problem, but it is an illogical one at present.

- Why?

- The variance $\mathbf{a}_1^{'}\Sigma\mathbf{a}_1$ would be maximised by setting any of the $a_{1j} = \infty$.

- The scale of $\mathbf{a}_1$ is thus arbitrary here. So, we fix the scale s.t. $\mathbf{a}_1'\mathbf{a}_1 = 1$. This makes sense, right? Why?

- We now have a nonlinear maximisation problem, subject to a constraint. One way to solve this is called the Lagrange multiplier method.

- Instead of maximising $\text{Var}(\mathbf{w}_1)$ directly, we maximise $L(\mathbf{a}_1) = \mathbf{a}_1^{'}\Sigma\mathbf{a}_1 - \lambda(\mathbf{a}_1'\mathbf{a}_1 - 1)$.

- Mathematically this makes sense, since we want $\mathbf{a}_1'\mathbf{a}_1 = 1$ for our estimated vector $\mathbf{a}_1$

- Maximising can be done by setting the 1st partial derivative to 0, i.e.

$$\frac{\delta L}{\delta \mathbf{a}_1} = 2\Sigma\mathbf{a}_1 - 2\lambda\mathbf{a}_1 = \mathbf{0}$$

which implies that

$$(\Sigma - \lambda I)\mathbf{a}_1 = \mathbf{0}$$

where I is the identity matrix, i.e. for any square matrix A, $AI = IA = A$.

- One solution is to set $\mathbf{a}_1 = \mathbf{0}$. Why is this invalid?
- We must find λ and $\mathbf{a}_1$ so that $\Sigma\mathbf{a}_1 = \lambda\mathbf{a}_1$ when $\mathbf{a}_1 \neq \mathbf{0}$
- In general, the solution to $B\mathbf{a} = \mathbf{b}$ would be $\mathbf{a} = B^{-1}\mathbf{b}$, which depends upon B^{-1} existing.
- But, if $(\Sigma - \lambda I)^{-1}$ does exist, the only answer possible would be $\mathbf{a}_1 = (\Sigma - \lambda I)^{-1}\mathbf{0} = \mathbf{0}$.
- So, to satisfy BOTH $(\Sigma - \lambda I)\mathbf{a}_1 = \mathbf{0}$ and $\mathbf{a}_1'\mathbf{a}_1 = 1$, the only possibility is that $(\Sigma - \lambda I)^{-1}$ **does not exist**.

- In the rules for matrix algebra, there is a special function called a determinant, labelled $|A|$.

- The matrix A is singular, so that A^{-1} does not exist, if, and only if, $|A| = 0$.

- Thus, if $(\Sigma - \lambda I)^{-1}$ does not exist, then

$$|(\Sigma - \lambda I)| = 0$$

- The solutions (values of λ) to $|(\Sigma - \lambda I)| = 0$ are called the **eigenvalues** of the matrix Σ.

- Eigenvalues are very, very useful in many areas of mathematics and statistics in general. Here, they are pivotal.

- To solve for the linear combination $\mathbf{a}_1$ that maximises $\text{Var}(\mathbf{w}_1)$ s.t. $\mathbf{a}_1'\mathbf{a}_1 = 1$, we need to find the eigenvalues of Σ, which are the values of λ for which $|\Sigma - \lambda I| = \mathbf{0}$.

- The formula $|\Sigma - \lambda I| = \mathbf{0}$ is actually an order-n polynomial. If solvable, it can have up to n solutions.

- If Σ is a positive definite matrix, there will be n eigenvalues, λ_i, that solve $|\Sigma - \lambda I| = \mathbf{0}$.

- Note that:

$$\begin{aligned} \mathrm{Var}(\mathbf{w}_1) &= \mathbf{a}_1^{'}\Sigma\mathbf{a}_1 \\ &= \mathbf{a}_1^{'}\lambda I\mathbf{a}_1 \\ &= \lambda \end{aligned}$$

- Thus, if a solution for $\mathbf{a}_1$ exists, and it maximises the variance of $\mathbf{w}_1$, this variance must equal the *largest eigenvalue* of Σ, $\lambda_i = \max(\lambda_1, \ldots, \lambda_n)$.

- For $n > 3$, these eigenvalues λ_i must be found by a numerical search procedure, built into most software packages, including Matlab.

- Once these are found, they are placed in order and labelled so that $\lambda_1 > \lambda_2 > \ldots > \lambda_n$.

- Then, we choose $\mathbf{a}_1$ so that $\mathrm{Var}(\mathbf{w}_1) = \lambda_1$.

- The solution for $\mathbf{a}_1$, which again must be found by solving a set of n equations(!!), is then called an **eigenvector** of the matrix Σ.

- Thus, to find $\mathbf{a}_1$, we need to find the largest eigenvalue of Σ and then $\mathbf{a}_1$ is the corresponding eigenvector of Σ.

- Most software packages can find eigenvalues and eigenvectors (if they exist) quite quickly and accurately.

- $\mathbf{w}_1 = \mathbf{a}_1'\mathbf{y}$ is the called the 1st principal component of the variables in $\mathbf{y}$.

- How can we determine how much of the variance in $\mathbf{y}$ has been captured in $\mathbf{w}_1$?

- The common choice is to take:

$$\frac{\text{Var}(\mathbf{w}_1)}{\sum_{i=1}^{n} \text{Var}(\mathbf{y}_i)} = \frac{\lambda_1}{\sum_{i=1}^{n} \sigma_i^2} = \frac{\lambda_1}{\sum_{i=1}^{n} \lambda_i}$$

- Usually, the 1st principal component will not capture a high enough proportion of the variance in $\mathbf{y}$.

- Then, a 2nd principal component will be found.

- So as to be different to the 1st PC, we find the linear combination of $\mathbf{y}$ that has the 2nd highest possible variance.

- PCA makes the 2nd PC uncorrelated with the 1st PC. i.e. choose $\mathbf{a}_2$ s.t.

$$\begin{aligned} \mathbf{w}_2 &= a_{21}\mathbf{y}_{1\cdot} + a_{22}\mathbf{y}_{2\cdot} + \ldots + a_{2n}\mathbf{y}_{n\cdot} \\ &= \mathbf{a}_2^{'}\mathbf{y} \end{aligned}$$

  has $\text{Var}(\mathbf{w}_2)$ as a maximum, $\mathbf{a}_2'\mathbf{a}_2 = 1$ AND $\text{Cov}(\mathbf{w}_1, \mathbf{w}_2) = 0$.

- This is equivalent to $\mathbf{a}_1'\mathbf{a}_2 = 0$, which means that $\mathbf{a}_1$ and $\mathbf{a}_2$ are orthogonal.

- It turns out that maximising $\text{Var}(\mathbf{w}_2)$ subject to $\mathbf{a}_2'\mathbf{a}_2 = 1$ and $\mathbf{a}_1'\mathbf{a}_2 = 0$, has $\text{Var}(\mathbf{w}_2)$ as the 2nd largest eigenvalue λ_2

- and that the linear combination weights $\mathbf{a}_2$ are given by the corresponding eigenvector of Σ.

- How can we determine how much of the variance in $\mathbf{y}$ has been captured in $\mathbf{w}_1$ and $\mathbf{w}_2$?

- The common choice is to take:

$$\frac{\text{Var}(\mathbf{w}_1) + \text{Var}(\mathbf{w}_2)}{\sum_{i=1}^{n} \text{Var}(\mathbf{y}_i)} = \frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{n} \sigma_i^2} = \frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{n} \lambda_i}$$

- When can we stop? What proportion of the variance is high enough? Good question!

- It turns out that the n solutions to $|\Sigma - \lambda I| = \mathbf{0}$ provide the n highest variance, orthogonal linear combinations of the variables $\mathbf{y}$. These are the n eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_n$.

- The actual linear combinations $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ are the corresponding n eigenvectors of Σ.

- In practice, analysts will take as many of these (m) as they feel necessary, balancing between parsimony (small m) and capturing as much of the variance as possible.

- The actual choices made (for m) are often closer to art than science.

Example

- Again we employ Kenneth French's data library and consider the 5 industry sector asset portfolios in the US: Consumer, Manufacturing, HiTech, Health and Other.

- We consider these together in the PCA framework.

- We first use the sample covariance matrix to estimate Σ.

- The 1st principal component is estimated as:

$$\begin{aligned} w_{1,t} &= 0.42 \times \text{Cnsmr}_t + 0.47 \times \text{Manuf}_t + 0.52 \times \text{Hitech}_t \\ &+ 0.42 \times \text{Health}_t + 0.38 \times \text{Other}_t \end{aligned}$$

- What might this component represent?

- What is a possible or logical interpretation here? Does it relate to the CAPM or factor models? Does this make sense so far?

- This component captures 89% of the variation in the 5 portfolio returns by itself. Could we have guessed this figure already??

- The correlation coefficient between this 1st PC and the market excess returns is 0.92 (!)

- Figures 1 and 2 show a scatterplot and time plots of the 1st PC and the market excess returns.

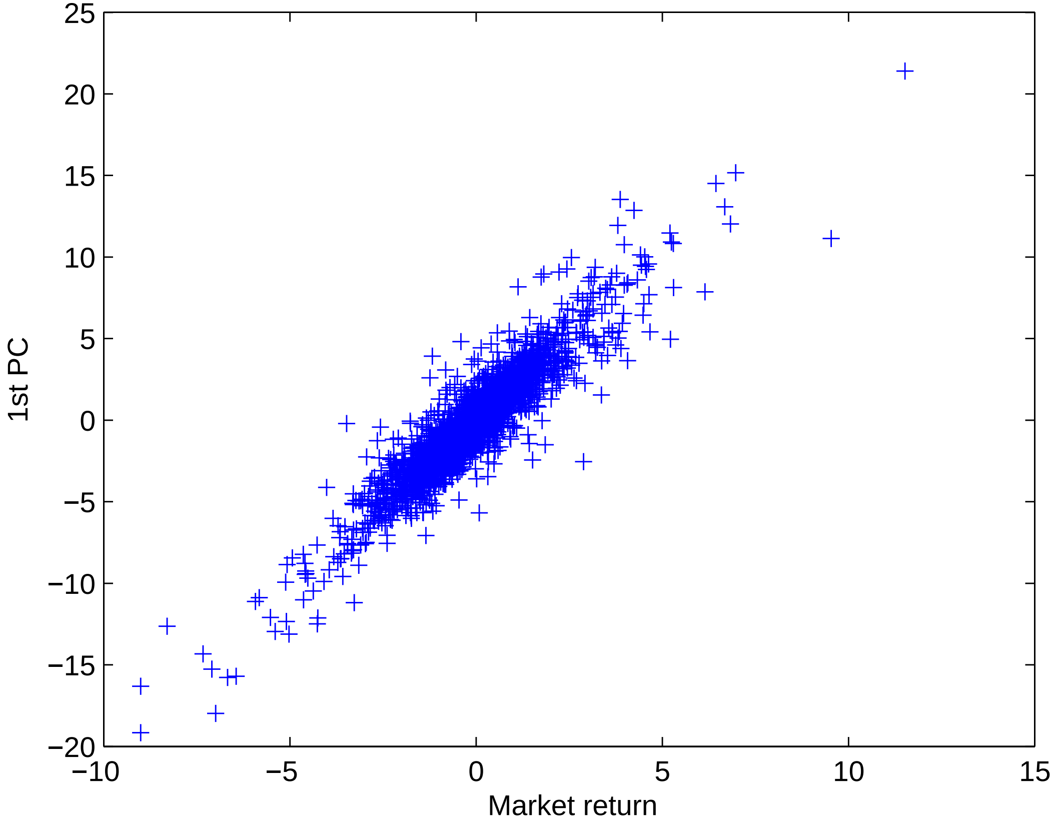

Figure 1: Scatterplot of 1st PC vs Market return

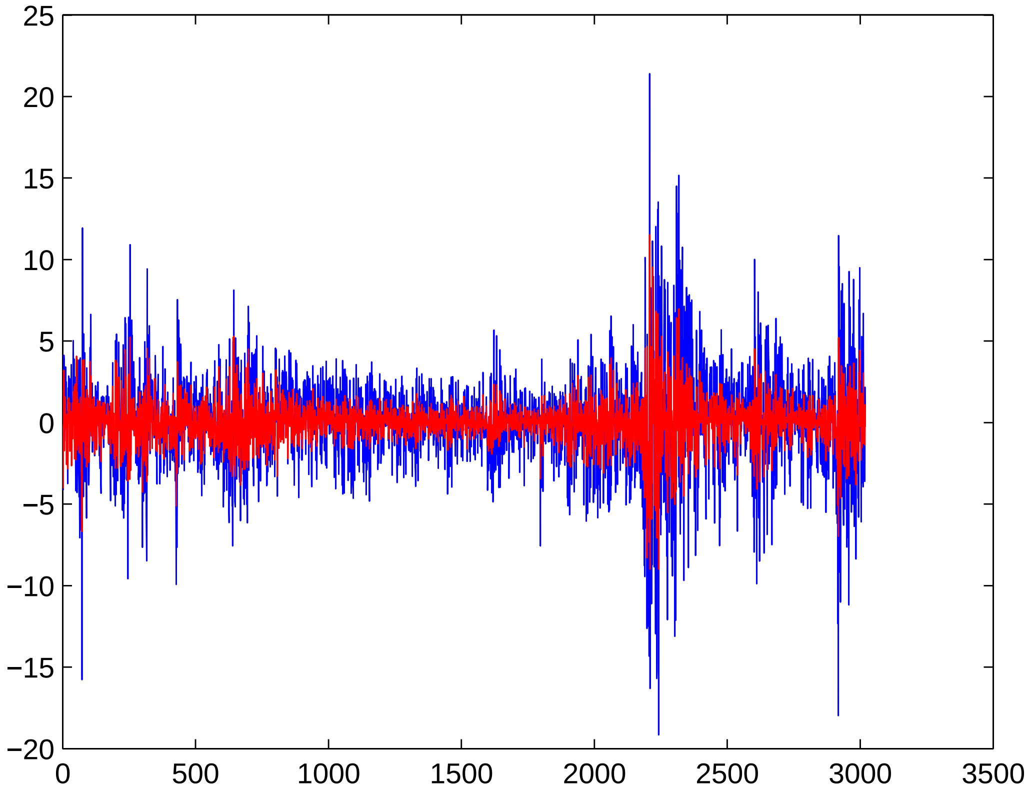

Figure 2: Time plot of 1st PC (blue) and Market return (red)

- The 2nd PC captures an extra 6% of the variability in the 5 portfolios combined, i.e. the 1st two PCs capture 95% of the total variation in these return portfolios.

- It is:

$$\begin{aligned} w_{2,t} &= -0.28 \times \text{Cnsmr}_t - 0.53 \times \text{Manuf}_t + 0.60 \times \text{Hitech}_t \\ &+ 0.42 \times \text{Health}_t - 0.31 \times \text{Other}_t \end{aligned}$$

- Interpretation?

- Table 1 gives the 5 PCs and their respective capturing of the variance, individually and in total.

Table 1: Estimates of the principal components of the 5 daily industry portfolios

	1st PC	2nd PC	3rd PC	4th PC	5th PC
Industry	Weights	Weights	Weights	Weights	Weights
Consumer	0.422	-0.285	0.082	0.509	-0.690
Manufacturing	0.473	-0.533	-0.036	-0.700	-0.012
Hi-Tech	0.525	0.600	0.588	-0.133	0.045
Health	0.423	0.420	-0.803	0.008	-0.005
Other	0.380	-0.315	0.036	0.482	0.723
Percentage (%) variance	89%	6%	2.5%	1.7%	0.7%
Cumulative % variance	89%	95%	97.6%	99.3%	100%

- Most analysts would stop at $m = 1$ or $m = 2$ (more likely) here. *Comments?*

- A "market" factor and an "industry group distinguishing" (Hi-tech/Health vs Cnsmr/Manuf/Other) factor account for 95% of the variation across the five portfolios.

- We could now focus analysis on these two components, instead of the original five.

- Or we could use this understanding of how the portfolios move or vary to build fundamental observed factors for a market model

- And/or to build an investment or risk management strategy.

- Biplots are graphical representations to visualize two or three components in a PCA.

- Figure 3 gives two components

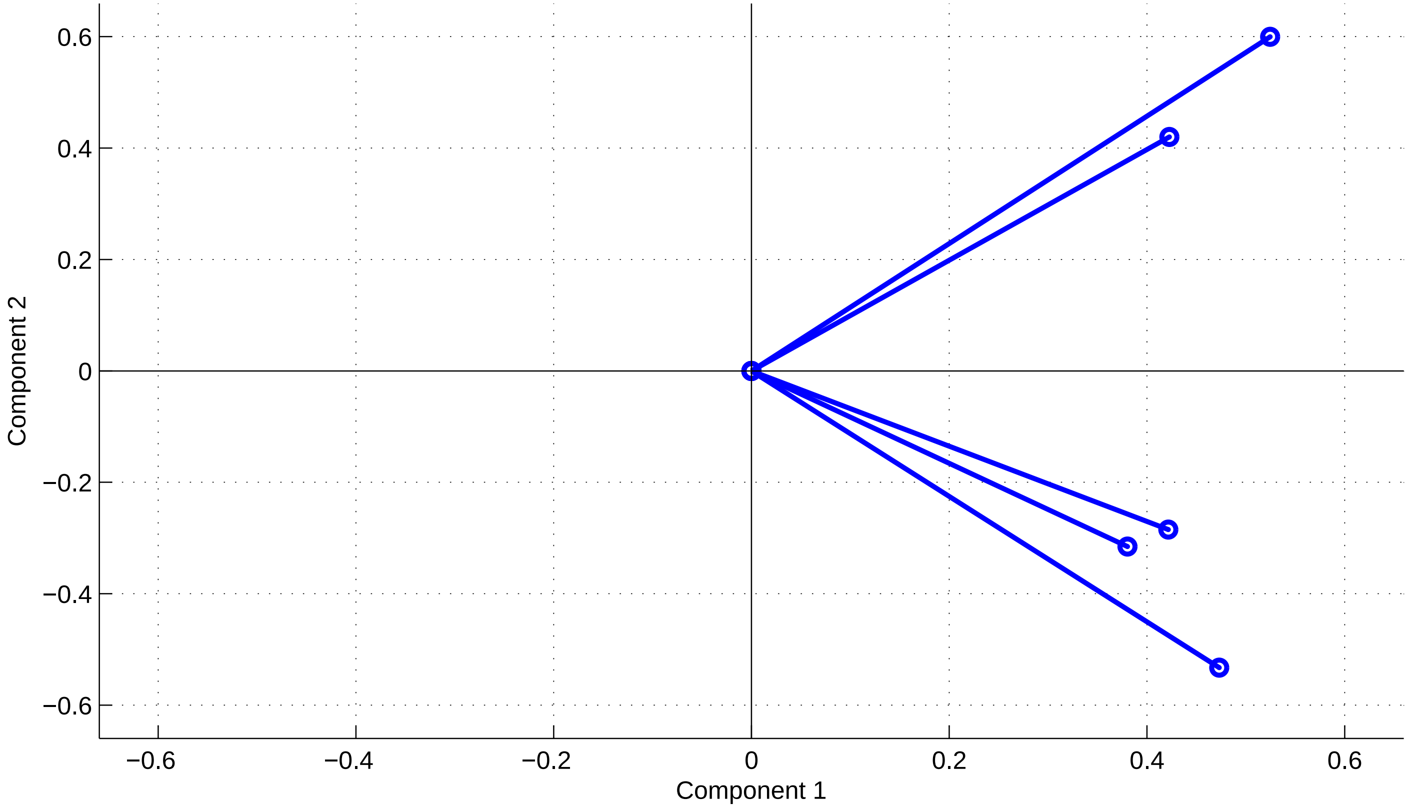

Figure 3: Biplot of 1st and 2nd sets of PC weights

- Figure 4 gives three components

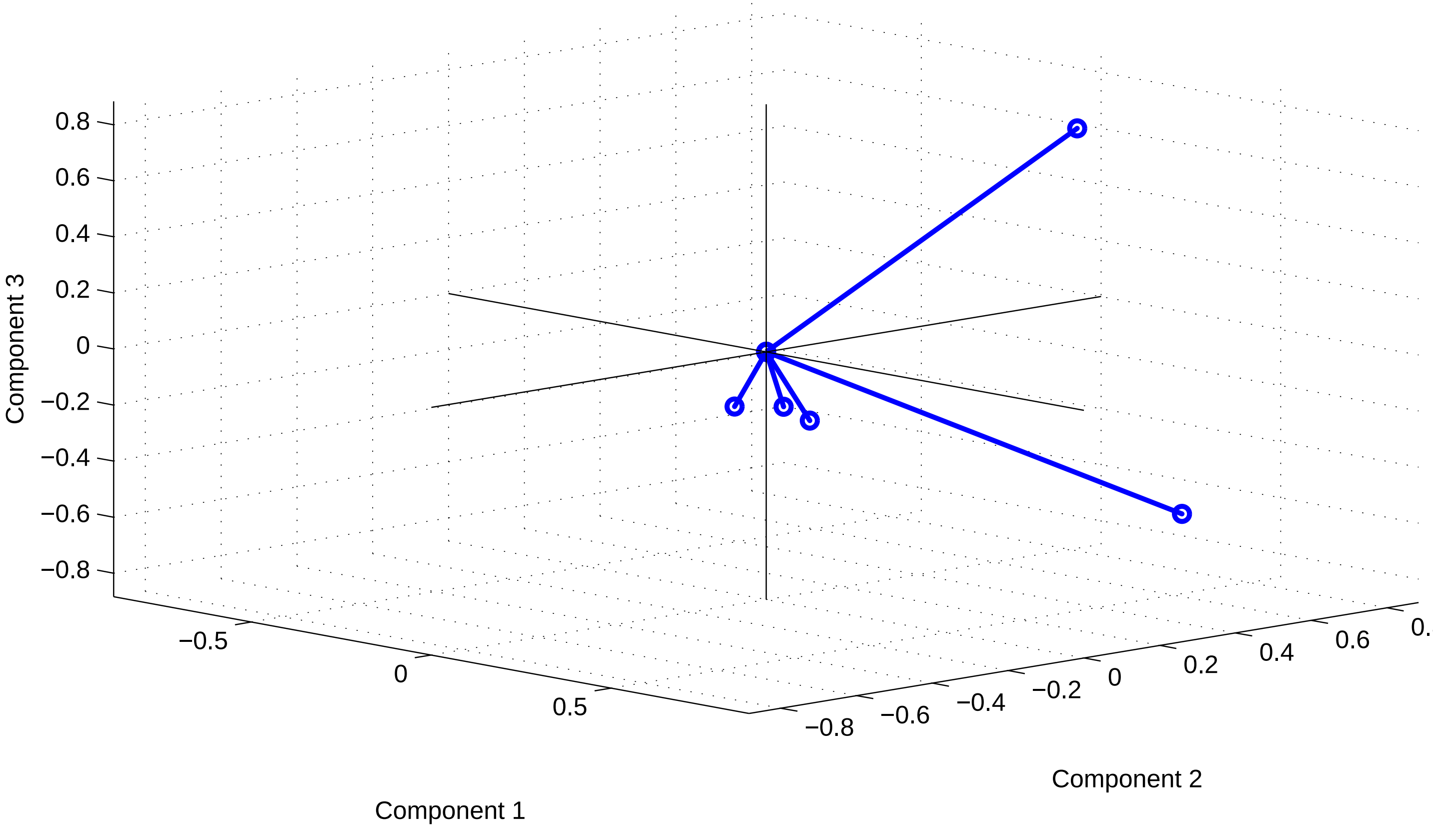

Figure 4: 3D biplot of 1st, 2nd and 3rd sets of PC weights

3.1 Principal Component Analysis (PCA, ctd)

- In fact, in general any positive definite matrix can be written as the product decomposition of its eigenvectors and eigenvalues, as follows:

  $$\Sigma = A^{'}\Lambda A$$

  where Λ is a diagonal matrix with ith diagonal element the ith eigenvalue, λ_i.

- A is the matrix where the ith column contains the ith eigenvector $\mathbf{a}_i$.

- $\Sigma = A^{'}\Lambda A$ is called the eigenvalue decomposition, or a *canonical* decomposition or a *spectral* decomposition of the matrix Σ.

- In practice PCA can either be done on the sample var-cov matrix Σ OR on the sample correlation matrix Ω.

- Unfortunately, different (though related and usually highly similar) principal components result depending on which matrix, Σ or Ω, is chosen.

- In other words, PCA is dependent on the scale of the variables in $\mathbf{y}$. If the scale changes, the principal components change. This is because the eigenvalues are also scale dependent, they depend directly on Σ.

- For that reason, many analysts prefer to use PCA only on the correlation matrix, Ω, so that each of the variables has the same standardised scale, of 1.

- Repeating the PCA but using the correlation matrix instead yields the results in Table 2:

Table 2: Principal components of the 5 daily industry portfolios, using the correlation matrix

	1st PC	2nd PC	3rd PC	4th PC	5th PC
Industry	Weights	Weights	Weights	Weights	Weights
Consumer	0.459	-0.282	0.111	-0.379	0.745
Manufacturing	0.447	-0.426	-0.179	0.766	-0.021
Hi-Tech	0.439	0.520	0.702	0.194	-0.079
Health	0.436	0.584	-0.680	-0.089	0.009
Other	0.455	-0.358	0.037	-0.474	-0.663
Percentage (%) variance	89%	6%	2.3%	1.8%	0.8%
Cumulative % variance	89%	95%	97.4%	99.2%	100%

- The interpretations and analysis and conclusions are much the same as before.

- Despite many statisticians preferring this latter method, using the correlation matrix, financial analysts usually prefer using the variance-covariance matrix for financial returns in PCA.

- This is because an average of standardised returns does not really have a market factor interpretation, since the market return does not employ mean-variance standardised returns.

- **Assumptions** of PCA:

  1. The variance of any linear combination of $\mathbf{y}$ is positive, i.e. Σ is positive-definite so that $\mathbf{a}'\Sigma\mathbf{a} > 0$ always, for any vector of weights $\mathbf{a}(\neq \mathbf{0})$.

- This assumption simply implies that the variables in $\mathbf{y}$ are not linear combinations, or replications, of each other. It also implies that the variables in $\mathbf{y}$ are not perfectly correlated with each other, so that each variable represents different information to the others, both individually and as a group.

- Assumption 1 further implies that the matrix Σ =Var($\mathbf{y}$) is non-singular, i.e. Σ^{-1} exists.

- In fact, it suggests that:

$$\Sigma = A^{'}\Lambda A \ ; \ \Sigma^{-1} = A^{'}\Lambda^{-1}A$$

- Though this last result is slightly beyond this unit, it is important in multivariate quantitative analysis.

Factor Analysis and Factor Modelling

- Factor analysis (FA) or Factor modelling (FM) has many of the same goals as PCA

- The main goal is to find a small number of underlying factors that explain most of the variation in the original variables $\mathbf{y}$.

- i.e. What underlying factors drive or explain the data $\mathbf{y}$?

- There are two main distinctions between FM and PCA:

    1. FM specifies and utilises a full statistical model.
    2. The factors identified by FA are **not** (i.e. never!) unique. PCA's factors are unique (by design).

- Point 1 is an advantage, and point 2 is a disadvantage, for FM compared to PCA.

- Because PCA has no statistical model, PCA factors are not subject to uncertainty calculations (i.e. standard errors are not usually available, prediction is not obvious).

- FM, however, can be subsequently tested for statistical significance (i.e. standard errors can be estimated) and used for generating predictions and forecasts, etc. This is because FM is based on a statistical model.

- The typical m-factor FM can be written as:

$$y_{it} = \alpha_i + \beta_{i1} f_{1t} + \beta_{i2} f_{2t} + \ldots + \beta_{im} f_{mt} + \epsilon_{it}$$

  where $\mathbf{y}_i$ is the ith original variable,for $i = 1, \ldots, n$, each observed from $t = 1, \ldots, T$.

- The model specifies that there are m underlying factors, $\mathbf{f}_1, \ldots, \mathbf{f}_m$ that explain the original variables $\mathbf{y}$, in a regression sense.

- There are three common situations with this model:

  1. The m factors are known exogenous variables, as in Fama and French (1982)'s extended CAPM.
  2. The m factors are unknown variables, that must be estimated from the observed data.

  The 2nd case can be further classified into two categories:

2(a) The parameters α, β are unknown and must also be estimated from the data; or

2(b) The parameters α, β are known or set to fixed values. e.g. the "Barra" factor model.

- We focus on case 2(a) here.

- In practice m is unknown and must either be set or also estimated or inferred from the data, as in PCA.

- The FM above can be written in multiple linear vector-matrix form:

$$\mathbf{y}_t = \alpha + \beta \mathbf{f}_t + \epsilon_t$$

where $\mathbf{y}_t$ is a vector containing all the n variables observed at time t.

- It is common to make the following error assumptions:

$$\begin{aligned} E(\epsilon_t) &= \mathbf{0} \; ; \; \text{Cov}(\epsilon_{it}, f_{jt}) = 0 \\ \text{Var}(\epsilon_t) &= \Psi \; = \text{diag}(\psi_1^2, \psi_2^2, \ldots, \psi_n^2) \end{aligned}$$

In other words, each error vector ϵ_t has 0 mean, is uncorrelated with each factor f_{jt}, including when $i = j$ and also uncorrelated within itself, $\text{Cov}(\epsilon_{it}, \epsilon_{jt}) = 0$, $(j \neq i)$ and over time.

- This means cross-sectional correlation is assumed to be 0, as is time correlation, for the errors.

- Further, each error variable has its own variance ψ_i^2, independent of time.

- These are simply the standard assumptions made in multivariate linear regression as estimated by statistical software.

- The ith variable error series ϵ_i is often called the ith **specific** factor, since it $\ldots$?.

- The underlying factors $\mathbf{f}_j$ are often called the **common** factors, since each affects every variable $\mathbf{y}_i$.

- If we stack all the time points together, we can write the model as:
$$\mathbf{y} = \theta \mathbf{X} + \epsilon$$
where $\mathbf{X}$ contains a $T \times 1$ vector of $\mathbf{1}$ (for each intercept term) for each of the n variables, plus each of the m factors $\mathbf{f}_j$ observed over $t = 1, \ldots, T$, in each subsequent column.

- $\mathbf{y}$ has dimension $Tn \times 1$, $\mathbf{X}$ has dimension $Tn \times (m+1)$, θ has dimension $(m+1) \times 1$.

- This is simply a multivariate linear regression model.

- If the m factors are known (i.e. case 1. above), then the OLS estimate of the unknown parameters is:

$$\hat{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- In this case, the model can be estimated and analysed as in standard OLS regression.

- If the parameters β are known, then a LS estimate of the unknown factors is:

$$\hat{\mathbf{f}} = (\beta'\Psi^{-1}\beta)^{-1}\beta'\Psi^{-1}\mathbf{y}$$

- See the Barra factor model.

- The statistically more interesting case is where BOTH the factors and the regression parameters are unknown.

- This is what quant analysts usually mean when they say "factor analysis" or "statistical factor analysis".

- Here, the full matrix model above implies that

$$\begin{aligned}\text{Var}(\mathbf{y}) &= \Sigma \\ &= \text{Var}(\theta\mathbf{X}) + \text{Var}(\epsilon) \\ &= \theta\text{Var}(\mathbf{X})\theta^{'} + \Psi\end{aligned}$$

  since the errors and factors are all uncorrelated.

- If we now assume that the factors are mean-variance stationary and that:

$$\begin{aligned}E(\mathbf{f}_t) &= \mu_f = \mathbf{0} \\ \text{Var}(\mathbf{f}_t) &= \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)\end{aligned}$$

  which enforces the factors to have mean 0, variances λ_j and to be uncorrelated, then this implies that:

$$\begin{aligned}\Sigma &= \beta \text{Var}(\mathbf{f})\beta^{'} + \Psi \\ &= \beta\Lambda\beta^{'} + \Psi\end{aligned}$$

- Since Ψ is a diagonal matrix, the covariances among the variables $\mathbf{y}$ are **completely** determined by the parameters β and factor variances λ_j, $j = 1, \ldots, m$.

- The variances of the variables $\mathbf{y}$ are determined by the specific variances $\Psi_{ii} = \psi_i^2$, plus the parameters and factor variances.

- This is another type of decomposition of a matrix, which again happens to be a variance-covariance matrix.

- As with the PCA problem, there is an arbitrary scale issue in the variances of the unknown factors. As such, it is common to assume that $\lambda_1 = \lambda_2 = \ldots = \lambda_m = 1$, so that $\Lambda = I_m$.

- Thus, it is common to set $\Lambda = I_m$, so that $\mathrm{Var}(f_{it}) = 1$, and $\mathrm{Cov}(f_{it}, f_{jt}) = 0$ and thus:
$$\Sigma = \beta\beta^{'} + \Psi$$

- The parameters β are called the **factor loadings**.

- These factor loadings thus completely determine the covariances for $\mathbf{y}$.

- Statistical estimation of this factor model has the goal to estimate $\beta, \mathbf{f}$ and also Ψ

- Is this possible? Are the unknown parameters uniquely identified and estimable from data?

- Note that:

$$\begin{aligned}\mathrm{Var}(y_{it}) &= \psi_i^2 + \sum_{j=1}^{m} \beta_{ij}^2 \\ \mathrm{Cov}(y_{it}, y_{kt}) &= \sum_{j=1}^{m} \beta_{ij}\beta_{kj} \\ \mathrm{Cov}(y_{it}, f_{kt}) &= \beta_{ik}\end{aligned}$$

- The parameters are identified and directly estimable if, and only if, this set of equations has a solution.

- These equations are non-linear in the parameters. As such it is **very** difficult to know whether a solutions exists.

- For most data sets a solution will exist for some values of $m < n$ but not for others!

- For some data sets no solution will exist for any m!

- The portion of the variance of $\mathbf{y}_i$ due to the common factors is $\sum_{j=1}^{m} \beta_{ij}^2$ and is called the **communality**.

- It is like a regression adjusted R^2 for each variable $\mathbf{y}_i$.

- The error variance part is called the **uniqueness** or **specificity** or **specific variance**, i.e. ψ_i^2.

- It is like SER^2 from regression, for $\mathbf{y}_i$.

- From above, there are n variance and $n(n-1)/2$ covariance equations above, making up Σ.

- There are also m factor loadings, each with n elements, and n uniqueness variances to estimate.

- A total of $n + n(n-1)/2$ equations to solve with $mn + n$ parameters.

- Further, there are m factors with T unknowns (i.e. mT) to estimate

- There are Tn observations in total. Clearly m must be somewhat less than n for any chance at identifiable estimation.

- However, the nonlinearity of the equations to be solved mean that the usual rules (e.g. $m < n$) for identification do not apply.

Example

- Again we employ Kenneth French's data library and consider the 5 daily industry sector asset portfolios in the US: Consumer, Manufacturing, HiTech, Health and Other, from January 1, 2000 to December 30, 2011.

- We consider these together, now in a factor modeling framework.

- First, consider a 1 factor model:

  $$y_{it} = \alpha_i + \beta_i f_t + \epsilon_{it}$$

  where $\mathbf{y}_i$ represent each industry portfolio,for $i = 1, \ldots, 5$, each observed from $t = 1, \ldots, 3019$ and $\text{Var}(\epsilon_{it}) = \sigma_i^2$.

- Matlab first estimates the parameter α_i by the sample mean of each return series. *why??*.

- Matlab then returns the estimated model:

$$\begin{aligned}
\widehat{\text{Cnsmr}}_t &= \bar{y}_1 + 1.28 \times f_t \\
\widehat{\text{Manuf}}_t &= \bar{y}_2 + 1.41 \times f_t \\
\widehat{\text{Hitech}}_t &= \bar{y}_3 + 1.45 \times f_t \\
\widehat{\text{Health}}_t &= \bar{y}_4 + 1.17 \times f_t \\
\widehat{\text{Other}}_t &= \bar{y}_5 + 1.16 \times f_t
\end{aligned}$$

  where the unknown factor f is assumed to have mean 0 and variance 1.

- Each portfolio seems to "load" roughly equally on the single unknown factor. What does that imply the unknown factor could be?

- Figure 5 shows a scatterplot of the single common factor and the market excess returns. The correlation between the estimated factor and market returns is (again) 0.92!

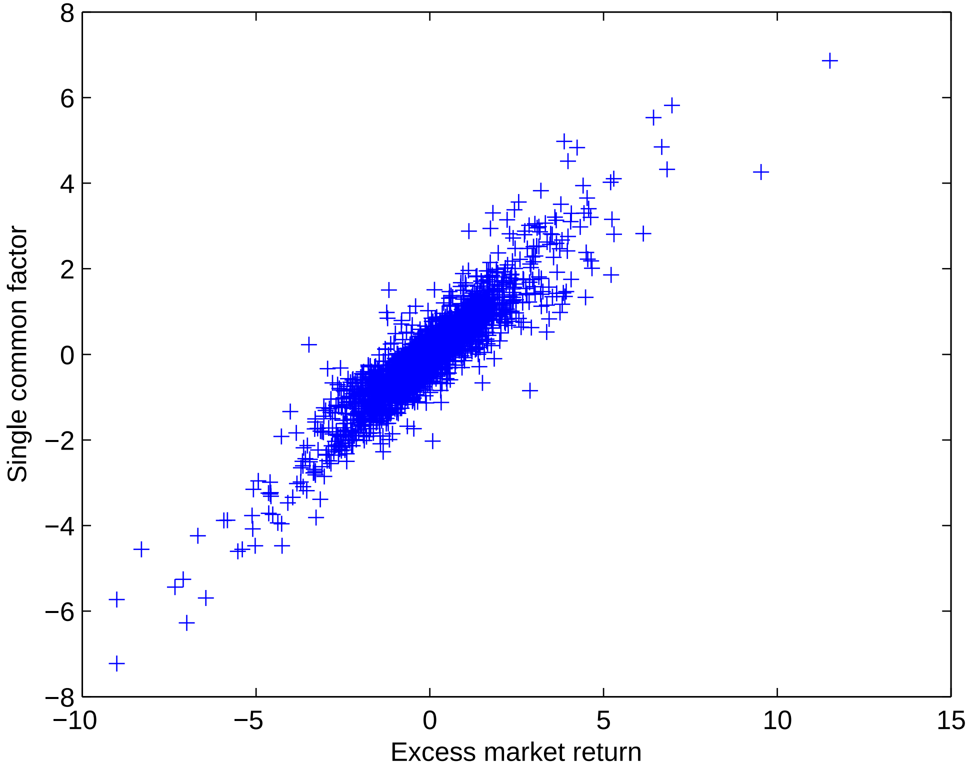

Figure 5: Scatterplot of single common factor vs Market return

- Figures 6 shows a time plot of the single common factor and the market excess returns.

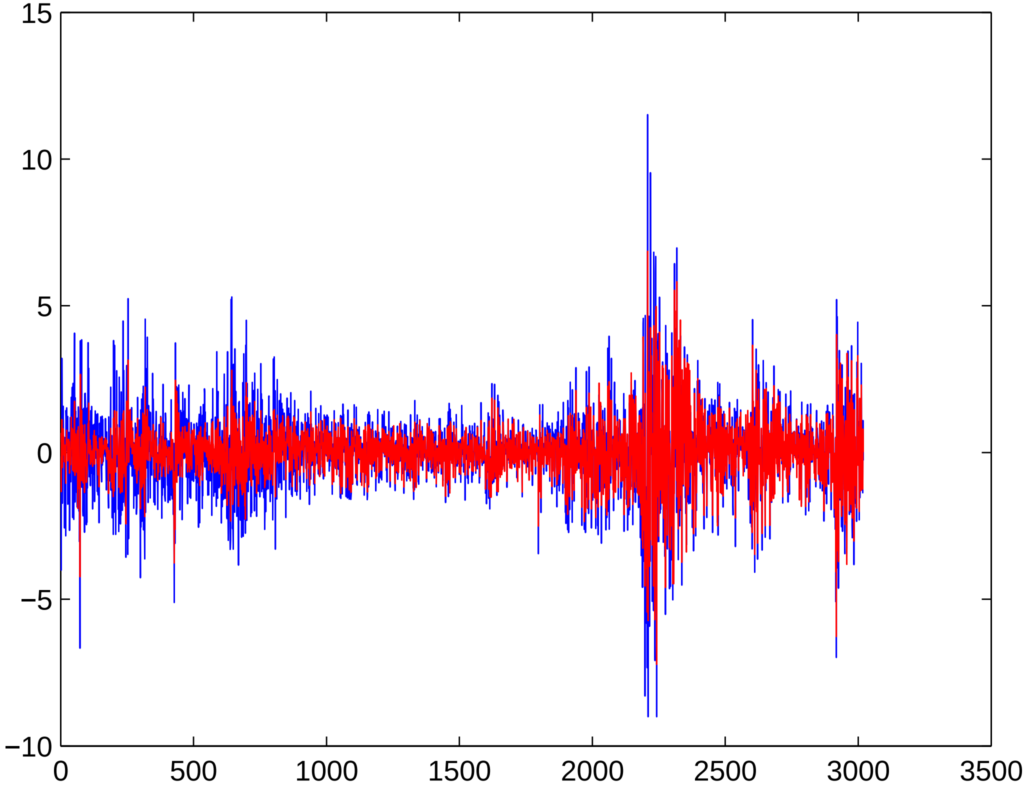

Figure 6: Time plot of common factor (blue) and Market return (red)

- A summary of the 1 factor model output is below. Table 3 gives the single factor loadings and their respective capturing of the variance, individually.

Table 3: Estimates of the factor loadings and specific variances of the 5 daily industry portfolios

	Factor	Specific	SER	R^2
Industry	Loadings	Variance		adjusted
Consumer	1.279	0.063	0.25	0.963
Manufacturing	1.407	0.260	0.51	0.884
Hi-Tech	1.449	0.647	0.80	0.764
Health	1.172	0.478	0.69	0.742
Other	1.157	0.077	0.28	0.946
Percentage (%) variance	85%			

- Since the 1st common factor is highly correlated to the market return series, have we improved over the original CAPM here?

- Compare these results to the original CAPM results:

Table 4: CAPM estimates for 5 daily industry portfolios

Industry	β	CI for β	SER	R^2
Consumer	0.832	(0.817,0.847)	0.59	0.80
Manufacturing	0.959	(0.942,0.976)	0.66	0.80
Hi-Tech	1.022	(1.0001,1.043)	0.84	0.74
Health	0.788	(0.768,0.809)	0.80	0.66
Other	0.761	(0.748,0.775)	0.53	0.80

- The improvements seem large in each case, higher R^2 and lower SER for the factor model over the CAPM.

- How many factors should we employ? Here we have $m = 1$.

- Next we try an $m = 2$ factor model. Results are in Table 5.

Table 5: Estimates of the factor loadings and specific variances of the 5 daily industry portfolios

	1st Factor	2nd Factor	Specific	SER	R^2
Industry	Loadings	Loadings	Variance		adjusted
Consumer	1.076	0.688	0.069	0.262	0.959
Manufacturing	1.207	0.725	0.255	0.505	0.886
Hi-Tech	0.791	1.452	0.014	0.117	0.995
Health	0.772	0.955	0.342	0.585	0.815
Other	1.004	0.585	0.064	0.252	0.955
Percentage (%) variance	92.5%				

- Is there a different interpretation on the 1st factor now?

- The correlation between the 1st estimated factor and the market returns is now 0.69.

- Interpretation on 2nd factor?

- Can the factors be interpreted together in a meaningful way?

- Does this model improve on the single factor model, for this data?

- Matlab informs me that it cannot estimate a 3 factor model for this data set: there are too many unknowns to estimate.

- Choice between 1 and 2 factor model?

- There is a statistical test comparing the sample covariance matrix to the var-cov matrix estimated using the loadings and specific variances, under the ML estimation method (described below).

- The test statistic is:

$$LR(m) = -[T - 1 - \frac{2n+5}{6} - \frac{2m}{3}] \left(\ln \left| \hat{\Sigma} \right| - \ln \left| \hat{\beta}\hat{\beta}' + \hat{\Psi} \right| \right)$$

  which, under a null hypothesis that the number of factors equals m, follows a chi-squared distribution with $0.5[(n-m)^2 - n - m]$ degrees of freedom.

- For the 1 factor model, the p-value on this test is 0. The conclusion of the test is that $m = 1$ factor can be strongly rejected.

- For the 2 factor model, the p-value on the hypothesis that $m = 2$ cannot be calculated, since one of the specific variances is too close to 0.

- Thus, $m = 2$ may be preferable, since we can almost predict Hi-Tech perfectly (with ≈ 0 specific variance) and since we reject $m = 1$.

- The 2 factor model factors and loadings are not unique. They can be 'rotated' without changing the estimates of the industry portfolios or their variance-covariance matrix (or the specific variances).

- Another possible answer then, is this:

Table 6: Estimates of the factor loadings and specific variances of the 5 daily industry portfolios

	1st Factor	2nd Factor	Specific	SER	R^2
Industry	Loadings	Loadings	Variance		adjusted
Consumer	1.175	0.499	0.069	0.262	0.959
Manufacturing	1.280	0.586	0.255	0.505	0.886
Hi-Tech	1.644	-0.167	0.014	0.117	0.995
Health	1.224	0.098	0.342	0.585	0.815
Other	1.050	0.498	0.064	0.252	0.955
Percentage (%) variance	92.5%				

- Are these factor (loading)s more interpretable?

- Why don't the values for SER, R^2 or overall variance captured change here, after the rotation?

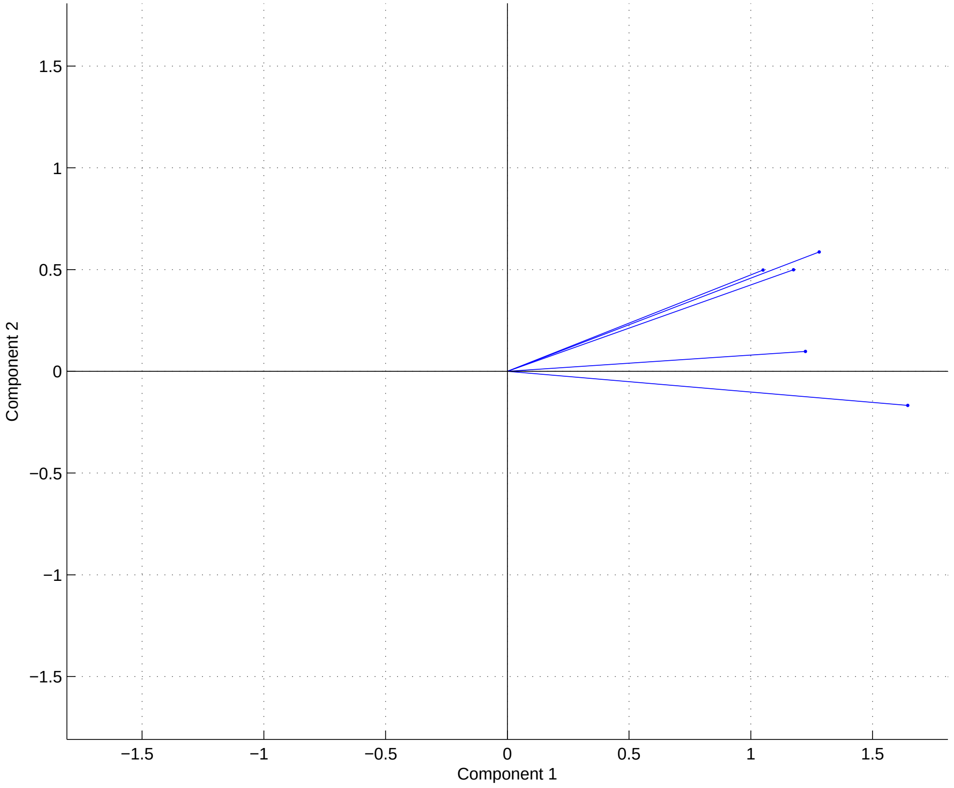

Figure 7: Biplot of 1st and 2nd sets of PC weights

- Three methods are mainly used for estimation of the parameters β, Ψ and then $\mathbf{f}$ (if needed):

  1. Maximum likelihood, assuming the factors are Gaussian AND the observations are jointly Gaussian.
  2. Exploiting a PCA analysis decomposition.
  3. Using nonlinear search methods to solve the nonlinear equations above, perhaps by minimising the sums of squares of the differences between sample variances and covariances and their factor model equation.

- If we assume

$$\begin{aligned} \mathbf{f} &\sim N(\mathbf{0}, I) \text{ and} \\ \epsilon_i|\Psi &\sim N(\mathbf{0}, \Psi) \text{ so that} \\ \mathbf{y}_i|\Psi, \beta &\sim N(\alpha, \Sigma = \beta\beta' + \Psi) \end{aligned}$$

  then it may be possible to estimate β, Ψ by maximising the Gaussian likelihood for $p(\mathbf{y}|\Sigma, \beta)$.

- A solution still has to exist to find the mle! Often the search algorithm will have numerical problems finding ml estimates.

- Matlab employs this option 1.

- However, Matlab first standardises the data by subtracting the sample mean from each series $\mathbf{y}_i$ and then dividing this by the standard deviation of each series $\mathbf{y}_i$, creating observations with mean 0 and variance 1, so that:

$$\hat{\beta}^s \hat{\beta}^{s'} + \hat{\Psi}^s \approx R$$

  where R is the sample correlation matrix. We need to re-scale to get the true communalities and specific variances, via:

$$\begin{aligned} \hat{\beta}_i &= s_i \hat{\beta}_i^s \\ \hat{\Psi}_i &= s_i^2 \hat{\Psi}_i^s \end{aligned}$$

  where s_i^2 is the sample variance of $\mathbf{y}_i$.

- Unlike PCA, ML estimates of factor loadings and specific variances are equivalent, via the transformation above, regardless of whether the data (i.e. sample covariance matrix) or standardised data (i.e. sample correlation matrix) are used in estimation.

- Alternatively, under method 2. (PCA), note that if $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ are the eigenvalues of Σ and A is the matrix whose columns are the eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$, then:
$$\Sigma = A^{'}\Lambda A$$

- However, if we take only the m highest eigenvalues, $\Lambda_m = \text{diag}(\lambda_1, \ldots, \lambda_m)$ and their corresponding eigenvectors, $A_m = (\mathbf{a}_1 \ \ldots \ \mathbf{a}_m)$ then
$$\Sigma = A_m^{'}\Lambda_m A_m + D \ = \beta\beta^{'} + D$$
where $D = \Sigma - A_m^{'}\Lambda_m A_m = A^{'}\Lambda A - A_m^{'}\Lambda_m A_m$

- Here the estimates are:

$$\begin{aligned}\hat{\beta} &= A_m^{'}\sqrt{\Lambda_m}\\ \hat{\Psi} &= \text{diag}(A^{'}\Lambda A - A_m^{'}\Lambda_m A_m)\end{aligned}$$

- i.e. the specific variances are estimated by $\hat{\psi}_i^2 = \hat{\Sigma}_{ii} - \sum_{j=1}^{m} \hat{\beta}_{ij}^2$

- One advantage of the PCA method is that changing the value of m does not change any of the factors. i.e. if $m = 2$ and $m = 3$ are considered separately, in BOTH cases the first two factors will be exactly the same.

- This is NOT the case for the ML factor estimation method 1, or method 3, in general.

- Under method 3. it is usual to minimise:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(\Sigma_{ij} - S_{ij})^2$$

where $\Sigma_{ij} = (\beta\beta^{'} + \Psi)_{ij}$ and S is the sample covariance matrix.

- Usually method 1. (ML) is more numerically optimal, i.e. able to find solutions for a particular choice of m, than method 3.

Factor rotation and non-uniqueness

- Even if a solution to the factor estimation problem is possible, uniqueness in estimation, for $\beta, \mathbf{f}$ is **not possible** for such a factor model.

- Let the matrix P satisfy $P^{'}P = PP^{'} = I$.

- Then, if $\beta^{*}, \mathbf{f}^{*}, \Psi^{*}$ is a solution to $\Sigma = \beta\beta^{'} + \Psi$

- then so is $\beta^* P, P^{'}\mathbf{f}^*, \Psi^*$, since:

$$E(\mathbf{f}^* P(\mathbf{f}^* P)^{'}) = E(\mathbf{f}^* P P^{'} \mathbf{f}^{*'}) = E(\mathbf{f}^* \mathbf{f}^{*'}) = I_m$$

  and

$$(\beta^* P)(\beta^* P)^{'} = \beta^* P P^{'} \beta^{*'} = \beta^* \beta^{*'}$$

- Thus

$$\Sigma = \beta^* \beta^{*'} + \Psi^* = \beta^* P(\beta^* P)^{'} + \Psi^*$$

- Thus, any and EVERY orthogonal rotation of the factors, described by $P\mathbf{f}$, where $P^{'}P = PP^{'} = I$, is ALSO a solution leading to **exactly the same** variance-covariance matrix Σ.

  Often, rotations are considered until a suitable interpretation of the factor loadings is obtained.

- Even though factors might be estimated by PCA, they can always be rotated as above!

- Once β and Ψ are estimated, the factors $\mathbf{f}$ can be estimated, conditional upon $\hat{\beta}$ and $\hat{\Psi}$, via the GLS method, as:

$$\hat{\mathbf{f}}_t = \left(\hat{\beta}'\hat{\Psi}^{-1}\hat{\beta}\right)^{-1}\hat{\beta}'\hat{\Psi}^{-1}(\mathbf{y} - \hat{\mu})$$

  where $\hat{\mu}$ is usually estimated as the sample mean of the observations.