Chapter 3.

Principal Components and Statistical Factor Models

This chapter of introduces the principal component analysis (PCA), briefly reviews statistical factor models. PCA is among the most popular statistical tools applied in finance and many other disciplines. Principal components were initially invented by Pearson around 1900. We may simply understand the method as reducing the number of existing variables to fewer variables that maintain as much information in the data as possible. PCA is typical of the of the low rank and independent representations of data and has more profound implications.

Since the commonly used financial or time series factor models, such as CAPM and Fama-French three factor models are actually regression models with *observed* covariates, which are very different from PCA or statistical factor models, we leave the discussion of them into the future chapters.

3.1. The probabilistic view of the principal component analysis.

Let

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$$

be the random vector of p dimension that we are concerned with. For example, X may represent the returns of p stocks. As before, we use

$$\Sigma = \text{var}(X) = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \vdots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{pmatrix}, \qquad \text{where } \sigma_{kl} = \text{cov}(X_k, X_l).$$

to denote the variance matrix X. The mean of X plays no role in PCs, and we assume here $E(X) = 0$ for convenience. By matrix singular value decomposition, we know

$$\Sigma = \mathbf{e}\Lambda\mathbf{e}'$$

where

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} \quad \text{with } \lambda_1 \geq \cdots \geq \lambda_p > 0$$

and

$$\mathbf{e} = (\mathbf{e}_1 \vdots \cdots \vdots \mathbf{e}_p) = \begin{pmatrix} e_{11} & \cdots & e_{1p} \\ \vdots & \vdots & \vdots \\ e_{p1} & \cdots & e_{pp} \end{pmatrix} \quad \text{is an orthonormal matrix,}$$

i.e., $\mathbf{e}\mathbf{e}' = I_p$. Note that $\mathbf{e}$ is a $p \times p$ matrix and $\mathbf{e}_k$ is its k-th column and therefore is a p-vector. And $(\lambda_k, \mathbf{e}_k)$, $k = 1, ..., p$, are the eigenvalue-eigenvector pairs of the matrix Σ.

The variation of a one dimensional random variable can be quantified by its variance. For a random variable X of p-dimension, its variation, fully described by its variance matrix Σ. One commonly used quantification of the total "amount" of variation of X is the trace of Σ, $trace(\Sigma)$.

Suppose we wish to use one single variable (1-dim) to maximumly quantify the variation of all p variables, say through linear combination of the components of X. We may try to construct it so that its variance is the largest. Let $Y_1 = \mathbf{a}^T X$, where $\mathbf{a}$ is a p-dimensional constant vector. Then

$$\text{var}(Y_1) = \mathbf{a}^T \text{var}(X)\mathbf{a} = \mathbf{a}^T \Sigma \mathbf{a}.$$

We wish to identify a Y_1, so that its variance is the largest. This variance depends on the scale of $\mathbf{a}$, which can be measured by its Euclidean norm. A fair comparison should require $\mathbf{a}$ to be of a fixed norm, say norm 1. The problem becomes searching for $\mathbf{a}$ of unit norm such that $\text{var}(Y_1)$ is the largest, i.e.,

$$\mathbf{a} = \text{argmax}\{\mathbf{b}^T\Sigma\mathbf{b} : \|\mathbf{b}\| = 1\}.$$

Recall the singular value decomposition, $\Sigma = \mathbf{e}\mathbf{\Lambda}\mathbf{e}^T$ and that $(\lambda_i, \mathbf{e}_i)$ are eigenvalue-eigenvalue pairs, such that $\lambda_1 \geq ... \geq \lambda_p$. Notice that $\mathbf{e}_i$ are orthogonal to each other with unit norms.

It follows that the solution is $\mathbf{e}_1$. That is, $Y_1 = \mathbf{e}_1^T X$ achieves the maximum variance which is λ_1. And we call this Y_1 the first principal component.

After finding the first principal component that is the "most important", one can mimic the procedure to find the "second most important" variable: $Y_2 = \mathbf{a}^T X$, such that

$$\mathbf{a} = \text{argmax}\{\mathbf{b}^T\Sigma\mathbf{b} : \|\mathbf{b}\| = 1, \mathbf{b}\perp\mathbf{e}_1\}$$

Note that $\mathbf{b}\perp\mathbf{e}_1$ is equivalent to the zero correlation between Y_1 and the search space of of Y_2. This implies we are looking the "second most important" in an attempt to explain most of the variation in X not explained by Y_1.

Along the line, one can formulate the second, third ... p-th principal components. Mathematically, the solutions are:

$$Y_k = \mathbf{e}_k^T X, \qquad k = 1, ..., p$$

where Y_k is the k-th principle component with variance λ_k. Note again that

$$\text{var}(\mathbf{y}) = \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix} \tag{1}$$

implies the principle components are orthorganal to each other, and the first being most important, second being the second most important, ..., with the importance measured by their variances.

Remark. (This remark is more advanced than required) There are two views of principal components that rest on the same mathematical fact. The first is the analysis view, which, along the line of the above argument, is to maximize $trace(UX)$ subject to U being a $r \times p$ orthonormal matrix, i.e., $UU^T = I_r$. Then, it can be proved that

$$U = \begin{pmatrix} \mathbf{e}_1^T \\ \vdots \\ \mathbf{e}_r^T \end{pmatrix}$$

and UX is just the $(Y_1, ...Y_r)^T$. This is maximizing the information of variation of the existing variables X using lower dimensional variables.

The other view is called synthesis view: One tries to use a $p \times p$ matrix A of rank r such that $trace(X - AX)$ is the smallest. Then it can be shown that

$$A = (\mathbf{e}_1 \vdots \cdots \vdots \mathbf{e}_r) \begin{pmatrix} \mathbf{e}_1^T \\ \vdots \\ \mathbf{e}_r^T \end{pmatrix}$$

and so that

$$AX = (\mathbf{e}_1 \vdots \cdots \vdots \mathbf{e}_r) \begin{pmatrix} \mathbf{e}_1^T \\ \vdots \\ \mathbf{e}_r^T \end{pmatrix} X = (\mathbf{e}_1 \vdots \cdots \vdots \mathbf{e}_r) \begin{pmatrix} Y_1 \\ \vdots \\ Y_r \end{pmatrix} = \sum_{j=1}^{r} Y_j \mathbf{e}_j$$

. The underlying interpretation is: if one tries to reconstruct X with a low dimensional variable by linearly combining X, and the low dimensionality is retricted to be smaller than r; then such a variable must be a linear combination of the first r principal components.

A brief proof (Not required). *The Analysis View*: Recall the decomposition of Σ. Let A be $r \times p$ orthonormal matrix with rank $r \leq p$, which maximizes

$$\begin{aligned} trace\{\mathrm{var}(AX)\} &= trace\{A\Sigma A^T\} = trace\{AU\Lambda U^T A^T\} = trace\{\Lambda U^T A^T AU\} \\ &= +trace(\Lambda D^T D) \qquad D = AU = (d_{ij}) \text{ an orthonormal } r \times p \text{ matrix} \end{aligned}$$

$D^T D$ are one of the $p \times p$ matrix with diagonal elements at most 1 and trace being r. The maximization of $trace(\Lambda D^T D)$ then must be $\sum_{j=1}^r \lambda_i$; and $D^T D$ must be I_r, and thus choosing $A = (\mathbf{e}_1, ..., \mathbf{e}_r))^T$ reaches the maximum, and $AX = Y$.

The Synthesis view: Let A be $p \times p$ matrix with rank $r \leq p$. Recall the decomposition of $\sum_{j=1}^r d_{ji}^2 \leq 1$ Suppose A matrix minimizes the following trace. Write

$$\begin{aligned} trace\{\mathrm{var}(X - AX)\} &= trace\{\Sigma - 2A\Sigma + A\Sigma A^T\} \propto trace\{-2AU\Lambda U^T + AU\Lambda U^T A^T\} \\ &= +trace(-2\Lambda D + \Lambda D^T D) \qquad D = U^T AU = (d_{ij}) \\ &\propto \sum_{i=1}^p \lambda_i (d_{ii} - 1)^2 + \sum_{i\not j} \lambda_i d_{ij}^2 \end{aligned}$$

Then, $d_{ij} = 0$ for all $i\ /j$, implying D must be diagonal. Since D must be of rank r, r of the diagnoal elements are positive. Then, it follows that $d_{ii} = 1$ for $i \leq r$, and $d_{ii} = 0$ for $i > r$. Then $A = UDU^T = \sum_{i=1}^r \mathbf{e}_i \mathbf{e}_i^T$, and $AX = \sum_{i=1}^r Y_i$.

Now we can summarize as follows.

Set

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix} = \mathbf{e}'(X - \mu) = \begin{pmatrix} \mathbf{e}_1' \\ \vdots \\ \mathbf{e}_p' \end{pmatrix} (X - \mu).$$

Clearly, $Y_j = \mathbf{e}_j'(X - \mu)$. By a simply calculation,

$$\mathrm{var}(Y) = \mathbf{e}'\mathrm{var}(X)\mathbf{e} = \mathbf{e}'\Sigma\mathbf{e} = \mathbf{e}'\mathbf{e}\Lambda\mathbf{e}'\mathbf{e} = \Lambda.$$

In particular, $\mathrm{var}(Y_j) = \lambda_j$, $j = 1, ..., p$, and $\mathrm{cov}(Y_k, Y_l) = 0$, for $1 \leq k \neq l \leq p$. Then, Y_j is called the j-th population P.C. The interpretation of the P.C.s is presented in the following. To make it clearer, we call a linear combination of X, $b'X$ with $\|b\| = 1$ a unitary linear combination.

(1). The first P.C. Y_1 explains the most variation among all unitary linear combinations of X. Namely,

$$\mathrm{var}(Y_1) = \lambda_1 = \max\{\mathrm{var}(b'X) : \|b\| = 1,\ b \in R^p\}.$$

The fraction of total variation of X explained by Y_1 is

$$\frac{\mathrm{var}(Y_1)}{\mathrm{var}(Y_1) + \cdots + \mathrm{var}(Y_p)} = \frac{\lambda_1}{\lambda_1 + \cdots + \lambda_p}.$$

Note that $\lambda_1 + \cdots + \lambda_p = trace(\Sigma)$ is used to measure total variation of X.

(2). The k-th P.C. Y_k explains the most variation not explained by the previous $k - 1$ P.C.s $Y_1, ..., Y_{k-1}$ among all unitary linear combination. Specifically,

$$\mathrm{var}(Y_k) = \lambda_k = \max\{\mathrm{var}(b'X) : \|b\| = 1,\ b'X \perp Y_1,\ ...,\ b'X \perp Y_{k-1},\ b \in R^p\}$$

Here and throughout, $\perp$ means 0 correlation. The fraction of total variation of X explained by Y_k is

$$\frac{\mathrm{var}(Y_k)}{\mathrm{var}(Y_1) + \cdots \mathrm{var}(Y_p)} = \frac{\lambda_k}{\lambda_1 + \cdots \lambda_p}.$$

We may summarize the P.C.s in the following table.

		eigenvalue (variance)	eigenvector (combination coefficient)	percent of variation explained	P.C.s as linear combination of $X-\mu$
1st P.C.	Y_1	λ_1	$\mathbf{e}_1$	$\lambda_1/(\lambda_1+\cdots+\lambda_p)$	$Y_1=\mathbf{e}_1'(X-\mu)$
2nd P.C.	Y_2	λ_2	$\mathbf{e}_2$	$\lambda_2/(\lambda_1+\cdots+\lambda_p)$	$Y_2=\mathbf{e}_2'(X-\mu)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p-th P.C.	Y_p	λ_p	$\mathbf{e}_p$	$\lambda_p/(\lambda_1+\cdots+\lambda_p)$	$Y_1=\mathbf{e}_p'(X-\mu)$

Note that $\mathbf{e}_j=(e_{1j},...,e_{pj})'$ is the j-th column of $\mathbf{e}$. As the P.C.s are orthogonal to each other (0 correlated), the part of variation explained by each P.C.s are distinct or non-overlapping with each other.

The relative size of the variance of a P.C. or the percentage of total variation explained measures the importance of the P.C.. Thus the 1st P.C. is the most important, the 2nd P.C. the 2nd important, and so on.

It is often desired to reduce the number of variables, especially when the number of variables in concern are too many. But the reduction must be done without much loss of information. P.C.s provide an ideal way of such reduction. One may retain the first k P.C.s, which altogether explains

$$\frac{\lambda_1+\cdots+\lambda_k}{\lambda_1+\cdots+\lambda_p}$$

of the total variation.

An understanding from the point of view of autoencoder. The autoencoding is an important part of the deep learning technology. It involves representing the variables in two-steps: encoding and decoding; and the PCA serves as a basic example. For X to Y (the principal components) is the encoding step and from Y back to X is the decoding step.

$$X \overset{\text{encoding}}{\Longrightarrow} Y=\mathbf{e}^T X \overset{\text{decoding}}{\Longrightarrow} X=\mathbf{e}$$

Or, specifically,

$$X \overset{\text{encoding}}{\Longrightarrow} Y_k=\mathbf{e}_k^T X,\ k=1,...,p. \overset{\text{decoding}}{\Longrightarrow} X=\sum_{k=1}^{p} Y_k\mathbf{e}_k$$

Such a representation is only mathematically useful. In actual applications, only the first few, say r, important principle compoents are retained. And the process becomes

$$X \overset{\text{encoding}}{\Longrightarrow} Y_k=\mathbf{e}_k^T X, k=1,...,r. \overset{\text{decoding}}{\Longrightarrow} X^*=\sum_{j=1}^{r} Y_j\mathbf{e}_j.$$

A simpler view is that encoding is the zipping of original variales or data, and decoding is the unzipping of the encoded variables or data.

3.2. Statistical view of principal components

The population P.C.s are only theoretical, in data analysis we need to work with their sample analogues: the sample P.C.s. Suppose there are n observations of p variables presented as

$$\mathbf{X}=\left(X_{(1)}\vdots X_{(2)}\vdots\cdots\vdots X_{(p)}\right)=\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}_{n\times p}.$$

Then $X_{(k)}$, an n-vector, contains all n observations of the k-th variable. Let $\mathbf{S}$ be the sample variance matrix. By decomposition,

$$\mathbf{S}=\hat{\mathbf{e}}\hat{\Lambda}\hat{\mathbf{e}}'$$

Let

$$\begin{aligned}\mathbf{Y}_{n\times p} &= \left(Y_{(1)} \vdots Y_{(2)} \vdots \cdots \vdots Y_{(p)}\right)\\ &= \left(X_{(1)} - \bar{X}_1 \vdots X_{(2)} - \bar{X}_2 \vdots \cdots \vdots X_{(p)} - \bar{X}_p\right)\hat{\mathbf{e}}\end{aligned}$$

where $\bar{X}_k = (1/n)\sum_{i=1}^n x_{ik}$ is the sample average of the n observations of the k-th variable. We summarize the sample P.C.s as follows.

		eigenvalue (variance)	eigenvector (combination coefficient)	percent of variation explained	P.C.s as linear combination of $X - \mu$
1st P.C.	$Y_{(1)}$	$\hat{\lambda}_1$	$\hat{\mathbf{e}}_1$	$\hat{\lambda}_1/(\hat{\lambda}_1 + \cdots + \hat{\lambda}_p)$	$Y_{(1)} = \sum_{j=1}^p \hat{e}_{j1}(X_{(j)} - \bar{X}_1)$
2nd P.C.	$Y_{(2)}$	$\hat{\lambda}_2$	$\hat{\mathbf{e}}_2$	$\hat{\lambda}_1/(\hat{\lambda}_1 + \cdots + \hat{\lambda}_p)$	$Y_{(2)} = \sum_{j=1}^p \hat{e}_{j2}(X_{(j)} - \bar{X}_1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p-th P.C.	$Y_{(p)}$	$\hat{\lambda}_p$	$\hat{\mathbf{e}}_1$	$\hat{\lambda}_p/(\hat{\lambda}_1 + \cdots + \hat{\lambda}_p)$	$Y_{(p)} = \sum_{j=1}^p \hat{e}_{jp}(X_{(j)} - \bar{X}_1)$

Interpretations analogous to the population P.C.s applies to the sample P.C.s. We omit the details.

3.3 Statistical factor models

Factor analysis may be viewed as a refinement of the principal component analysis. The objective is, like the P.C. analysis, to describe the relevant variables in study in terms of a few underlying variables, called factors.

Let $X = (X_1, \cdots, X_p)'$ be the random variables in study with

$$\text{mean } E(X) = \mu = \begin{pmatrix}\mu_1\\ \vdots\\ \mu_p\end{pmatrix} \quad \text{and variance} \quad \text{var}(X) = \Sigma = \begin{pmatrix}\sigma_{11} & \cdots & \sigma_{1p}\\ \vdots & \vdots & \vdots\\ \sigma_{p1} & \cdots & \sigma_{pp}\end{pmatrix}.$$

The orthogonal factor model is

$$X_{p\times 1} - \mu_{p\times 1} = L_{p\times m}F_{m\times 1} + \epsilon_{p\times 1}, \tag{$*$}$$

where $m \le p$, $\mu = E(X)$,

$$L = \begin{pmatrix}l_{11} & \cdots & l_{1m}\\ \vdots & \vdots & \vdots\\ l_{p1} & \cdots & l_{pm}\end{pmatrix}$$ is called the *factor loading matrix* (which is non-random),

$$F = \begin{pmatrix}F_1\\ \vdots\\ F_m\end{pmatrix}$$ are called the factors or common factors,

$$\epsilon = \begin{pmatrix}\epsilon_1\\ \vdots\\ \epsilon_p\end{pmatrix}$$ are called errors or specific errors.

The model can be re-expressed as

$$X_i - \mu_i = \sum_{j=1}^m l_{ij}F_j + \epsilon_i, \qquad i = 1, ..., p.$$

And l_{ij} is called the loading of X_i on the factor F_j.

The assumptions of the orthogonal model are:

(1). $E(F) = 0_{m\times 1}$ and $\text{var}(F) = I_m$.

(2). $E(\epsilon) = 0_{p\times 1}$ and $\text{var}(\epsilon) = \Psi$, a diagonal matrix with diagonal elements: $\psi_1, ..., \psi_p$.

(3). $\text{cov}(F, \epsilon) = 0_{m\times p}$.

Remark The above model assumption implies that

$$\text{cov}(F_i, F_j) = \begin{cases} 1 & if\ i = j \\ 0 & if\ i \neq j \end{cases} \qquad \text{cov}(F_i, \epsilon_j) = 0 \text{ and } \text{cov}(\epsilon_i, \epsilon_j) = \begin{cases} \psi_i & if\ i = j \\ 0 & if\ i \neq j. \end{cases}$$

Moreover,

$$\text{cov}(X, F) = \text{cov}(LF, F) = L \qquad \text{cov}(X_i, F_j) = l_{ij}, \quad i = 1, ..., p;\ j = 1, ..., m.$$

Under the orthogonal factor model, the variance matrix of X, Σ, can be written as

$$\begin{aligned} \Sigma &= \text{var}(X) = \text{var}(LF + \epsilon) = \text{var}(LF) + \text{var}(\epsilon) \\ &= L\text{var}(F)L' + \Psi = LL' + \Psi. \end{aligned}$$

In particular,

$$\begin{aligned} \sigma_{ii} &= \text{var}(X_i) = \sum_{j=1}^{m} l_{ij}^2 + \psi_i \equiv h_i^2 + \psi_i \\ \sigma_{ij} &= \text{cov}(X_i, X_j) = \sum_{k=1}^{m} l_{ik} l_{jk}, \qquad i \neq j. \end{aligned}$$

Here $h_i^2 \equiv l_{i1}^2 + \cdots + l_{im}^2$ is called *communality*, which is the portion of the variance of X_i explained by common factors. ψ_i, called *specific variance* or *uniqueness*, is the portion of the variance of X_i explained by specific factor, the error pertained to the i-th variable X_i only.

In orthogonal factor model, the factors or common factors are supposed to be important underlying factors that significantly affect all variables. Besides these factors, the remaining ones are those only pertained to the relevant variables. Specifically, ϵ_i, the error pertained to the i-th variable X_i, explains the part of the variation of X_i that cannot be explained by common factors or by other errors.

Remark The orthogonal factor model is essentially different from linear regression model, although there is certain formality resemblance. The key difference is the common factor F, which seemingly plays the role of covariates in linear regression model, is not observable.

Remark The orthogonal factor model has unidentifiable L and F, up to a rotation, in the sense that

$$X - \mu = LF + \epsilon = L^* F^* + \epsilon,$$

where $L^* = LT$ and $F = T'F$ with T being any orthonormal $m \times m$ matrix. Then, (F^*, ϵ) still satisfies the assumptions (1)-(3).

The estimation of the statistical factor models is mainly about estimating the loading matrix. Here the maximimum likelihood method or principal component approach can be applied. The PC approach simple takes as the factors the first r eigenvectors corresponding to the largest r eigenvalues of the sample variance matrix, namely the renormalized prinpical componets. The maximum likelihood method assume $X_1, ..., X_n$ are iid $\sim MN(\mu, \Sigma)$, and then maximizes the the likelihood

$$lik(\mu, \Sigma) \equiv (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\{-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)'\Sigma^{-1}(X_i - \mu)\}.$$

Under the orthogonal factor model, $\Sigma = LL' + \Psi$. Then, the likelihood becomes

$$lik(\mu, L, \Psi) \equiv (2\pi)^{-np/2} |LL' + \Psi|^{-n/2} \exp\{-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)'(LL' + \Psi)^{-1}(X_i - \mu)\}.$$

With certain restriction, the MLE of L and Ψ can be computed. The actual computation can be quite complex.

There are various other issues including:

1. The choice of the number of factors in the model. Some statistical generalized likelihood ratio tests can be applied.

2. Factor rotation. The orthogonal factor model is not identifiable up to a rotation of the common factors or factor loading matrix. In other words,

$$X - \mu = LF + \epsilon = L^* F^* + \epsilon$$

for $L^* = LT$ and $F^* = T'F$ where T is any $m \times m$ orthonormal matrix. Therefore, up to a rotation, it is legitimate and often desirable to choose a pair (L, F) so that it may achieve better interpretability. A criterion called *varimax criterion* can be applied to find an optimal rotation. Let $\hat{L}^*$ be the $p \times m$ *rotated* factor loading matrix with elements $\hat{l}^*_{ij}$. Define $l^*_{ij} = \hat{l}^*_{ij}/\hat{h}_i$, and

$$V = \sum_{j=1}^{m} \Big[\frac{1}{p} \sum_{i=1}^{p} (l^*_{ij})^4 - \{ \frac{1}{p} \sum_{i=1}^{p} (l^*_{ij})^2 \}^2 \Big],$$

which is the sum of the column-wise variance of the squares of scaled factor loadings. Find the optimal l^*_{ij} such that V achieves maximum. Then, the optimal rotated factor loading matrix is $\hat{l}^*_{ij} = \hat{h}_i \times$ (the optimal l^*_{ij}).

3. Factor scores. Let $x_1, ..., x_n$ be a sample of n observations of X that follows the orthogonal factor model. Write

$$x_i - \mu = Lf_i + e_i$$

Then, f_i and e_i may be regarded as the realized but unobserved values of the common factors and the errors that produced the i-th observation x_i. Note that x_i is $p \times 1$, f_i is $m \times 1$ and e_i is $p \times 1$. Factor scores refer to estimator of f_j, denoted by $\hat{f}_j$ or $\tilde{f}_j$. There are two commonly used methods of estimation. Factor scores can be estimated using weighted/unweighted least squares method or using Regression method.

3.4. Examples

Example 3.1. (Sample principle components from standardized data in Example 1.2.) We consider **X** is the daily returns of six stocks (601398, 601939, 601288, 600028, 601857, 601088) in the period July 16, 2010 through July 21, 2016. The 6 observations in 1462 successive days seems to be independent, but the returns of pairs of stocks are correlated since the stocks tend to move together when some conditions vary. (see pair plot figures in Example 1.2).

Suppose the covariate matrix **X** have been standardized by column. Then the sample mean is

$$\bar{\mathbf{X}} = [3.5690, 4.2365, 4.1656, 1.6030, .30679, 1.3279] \times 10^{-4}.$$

The first 6 PCs are shown in Table 1. The rotation of the fist PC is counted as the weight of the six stocks. We could see that the first PC is highly correlated with the market average. Thus the first PC is also called the market component with the proportion of variance 73.2%. Likewise the second PC represents a contrast between the banking stocks and oil stocks, which could be called as the industry component. It can be seen from Table 2 that 85.26 percentage of the variance is explained by the first 2 PCs.

The remaining PCs are not that easy to be interpreted, which might be caused by certain stocks. Anyway, they only explain small percentage of the variance.

As a result, most of the variance of the six stocks return is due to the market activity and independent the industry contrast.

	PC1	PC2	PC3	PC4	PC5	PC6
601398	0.4111359	0.4516954	-0.113085840	0.6849229	-0.23359365	0.3007737286
601939	0.4281655	0.3361847	0.008669585	-0.6417251	0.17558664	0.5108200793
601288	0.4280175	0.3830204	-0.007170309	-0.1435821	0.04009217	-0.8048737374
600028	0.4017174	-0.4882749	-0.274536684	-0.1952141	-0.69747350	-0.0162045745
601857	0.4010982	-0.4463199	-0.417857256	0.2062420	0.65021406	0.0002229649
601088	0.3770755	-0.3140584	0.858551070	0.1334519	0.06122370	0.0226636198

Table 1: The first 6 PCs of six stocks.

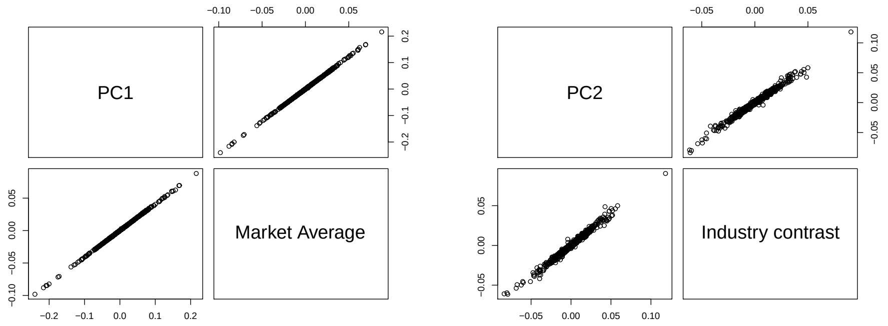

Figure 1: Pair plots of the PCs and market average, industry contrast.

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.096	0.8506	0.63822	0.42249	0.40847	0.36272
Proportion of Variance	0.732	0.1206	0.06789	0.02975	0.02781	0.02193
Cumulative Proportion	0.732	0.8526	0.92052	0.95026	0.97807	1.00000

Table 2: Summary of PCA results.

Example 3.2.
Part 1

In this example, let us consider more stocks. The source data is about the daily returns of the constituent stocks of CSI 300 index, from Jan 1st, 2001 to July 21st, 2016. Note that this dataset contains lots of missing values (NAs) because of issues like trading halt. In this case, common used PCA functions in R such as `prcomp` may raise error. To handle this issue, for each stock we replace the missing values with the mean value so they don't contain any information about the final PCA results since afterwards all data will be centralized. Part of the loading matrix is shown below:

	PC1	PC2	PC3	PC4	PC5	PC6	···
000001	-0.062	-0.005	0.006	-0.084	0.036	-0.065	···
000002	-0.060	-0.007	0.001	-0.104	0.030	-0.038	···
000009	-0.077	-0.004	-0.013	0.006	-0.0001	0.053	···
000027	-0.074	-0.005	-0.014	-0.020	0.008	0.006	···
000039	-0.077	-0.007	-0.005	-0.027	0.0003	-0.013	···
000046	-0.079	-0.006	-0.003	-0.074	0.027	-0.034	···
···	···	···	···	···	···	···	···

Table 3: Part of loading matrix of Example 3.2

Next table is about the importance of the first 7 P.C. Note that though P.C.1 and P.C.2 explained 26.7% and 11.5% of the total variation respectively, other P.C.s contribute little.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	0.253	0.166	0.100	0.086	0.081	0.076	0.065
Proportion of Variance	0.267	0.115	0.042	0.031	0.028	0.024	0.018
Cumulative Proportion	0.267	0.383	0.425	0.456	0.483	0.507	0.525

Table 4: The importance matrix for the first 7 P.C.s

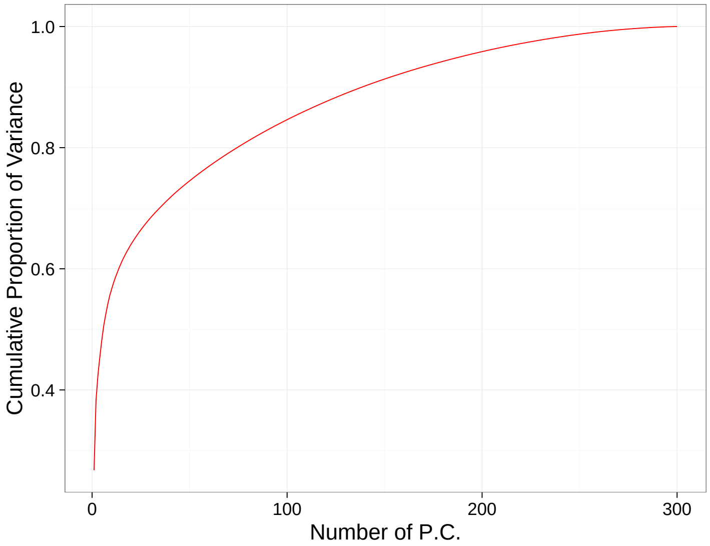

Figure 2: Cumulative proportion of variance explained by PCs

From Figure 2, we found that to explain 70% of the total variation, one need to consider about the first 50 P.C.s. This implies for large stock pool, the price variation may not be explained well by only a few factors. To dig out more information, here we draw a pairwise plots of the first 5 P.C.s, which is shown in Figure 3.

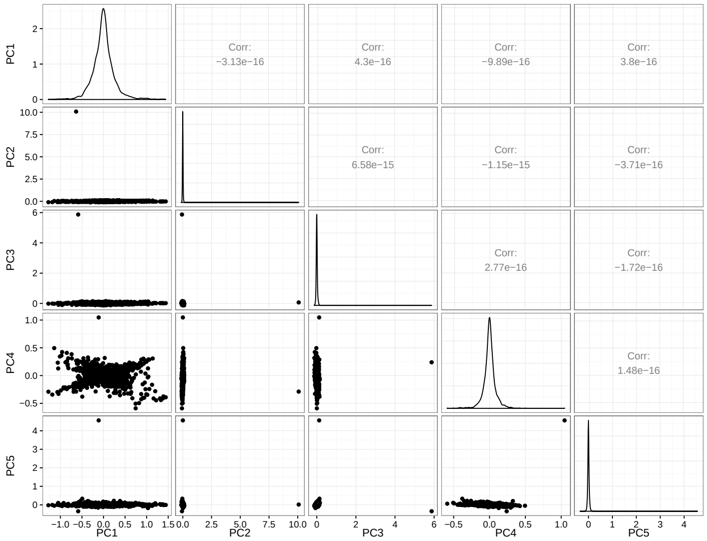

Figure 3: Pairwise plots of the first 5 P.C.s

From Figure 3, we found some strange things. For P.C.2, P.C.3 and P.C.5, there are some outliers which is extremely far away from the majority. After checking the coefficients of these P.C.s, we figured out that these outliers are due to some special events of some special stocks. For example, in P.C.2, the weight of stock 000156 is unusually high compared with other stocks. Actually, 000156 were suspended from 2006 to 2012. On the first day of its resumption (Oct 19th, 2012), its price rocketed from 1.25 to 14.42. This event affected the whole PCA result.

To handle this issue, we replace those extreme values (daily return larger than 0.1 or less than −0.1) with NAs and do the same analysis again. The loading matrix and importance matrix are shown in the following two tables.

	PC1	PC2	PC3	PC4	PC5	PC6	···
000001	-0.063	-0.090	0.074	-0.067	0.087	-0.054	···
000002	-0.060	-0.110	0.048	-0.038	0.085	-0.142	···
000009	-0.079	0.006	-0.057	-0.091	-0.005	0.020	···
000027	-0.073	-0.024	-0.006	0.012	-0.082	-0.036	···
000039	-0.078	-0.022	0.026	0.045	-0.037	-0.017	···
000046	-0.078	-0.074	0.034	-0.098	0.028	-0.118	···
···	···	···	···	···	···	···	···

Table 5: Part of loading matrix of Example 3.2 (after removing extreme data points)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	0.242	0.081	0.071	0.051	0.047	0.044	0.042
Proportion of Variance	0.356	0.039	0.031	0.016	0.013	0.011	0.011
Cumulative Proportion	0.356	0.395	0.426	0.442	0.455	0.466	0.477

Table 6: The importance matrix for the first 7 P.C.s (after removing extreme data points)

Note that this time only P.C.1 can explain more than 5% of the total variation. As shown in Figure 4, if we want to explain 80% of the total variation, we need consider the first 100 P.C.s which is kind of useless.

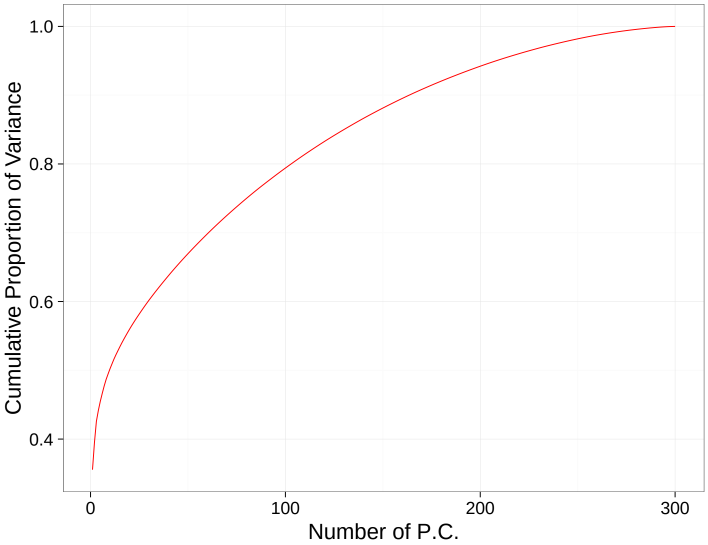

Figure 4: Cumulative proportion of variance explained by PCs (after removing extreme data points)

From Figure 5, we found that now the value distributions of P.C.2, P.C.3 and P.C.5 are much normal, as extreme daily returns have been removed.

Next plot shows the relation between P.C.1 and CSI 300 index daily return. We found that they are nearly perfect negative correlated.

Part 2

In this part, we consider *weighted* daily return of the component stocks of CSI 300 index. Here weight is judged by stock's capitalization. Then we do the same analysis as Part 1.

The importance matrix for the first 7 P.C.s is shown in Table 8. We observe that now the first 5 P.C.s can explain 84.1% of the total variance. Now we only use these 5 P.C.s and the loading matrix to reconstruct

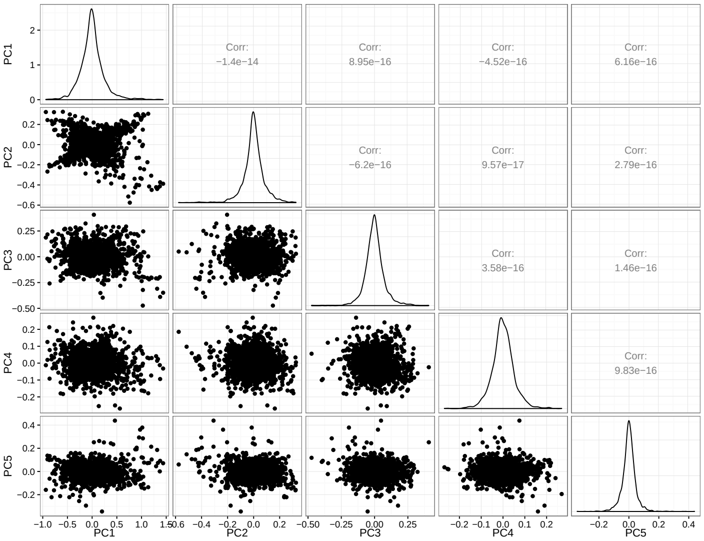

Figure 5: Pairwise plots of the first 5 P.C.s (after removing extreme data points)

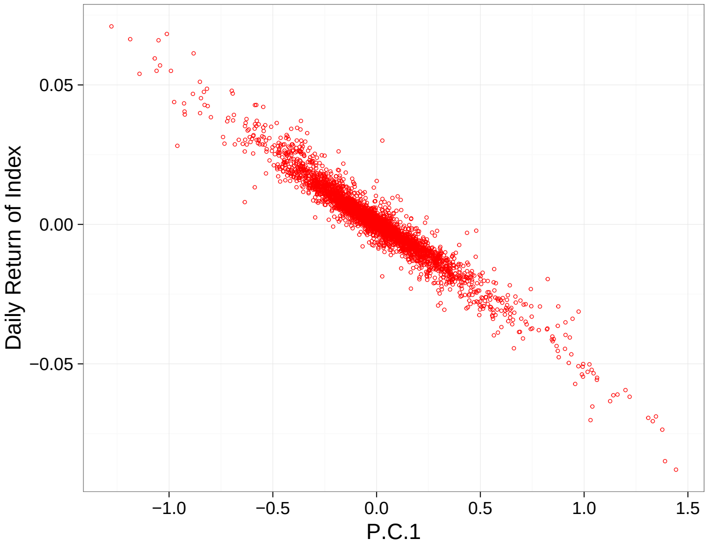

Figure 6: The relation between P.C.1 and CSI 300 index daily return

	PC1	PC2	PC3	PC4	PC5	PC6	· · ·
000001	-0.029	0.004	-0.004	-0.055	0.099	-0.053	· · ·
000002	-0.031	-0.009	-0.007	-0.047	0.031	-0.032	· · ·
000009	-0.005	0.003	0.0003	-0.010	0.018	-0.018	· · ·
000027	-0.011	0.007	0.001	-0.018	0.033	-0.035	· · ·
000039	-0.017	0.005	-0.005	-0.025	0.027	-0.045	· · ·
000046	-0.007	-0.004	0.001	-0.014	0.009	-0.009	· · ·
· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·

Table 7: Part of loading matrix of Example 3.2 (weighted daily return)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	0.003	0.002	0.001	0.001	0.001	0.001	0.001
Proportion of Variance	0.438	0.234	0.109	0.036	0.024	0.022	0.019
Cumulative Proportion	0.438	0.672	0.781	0.817	0.841	0.863	0.881

Table 8: The importance matrix for the first 7 P.C.s (weighted daily return)

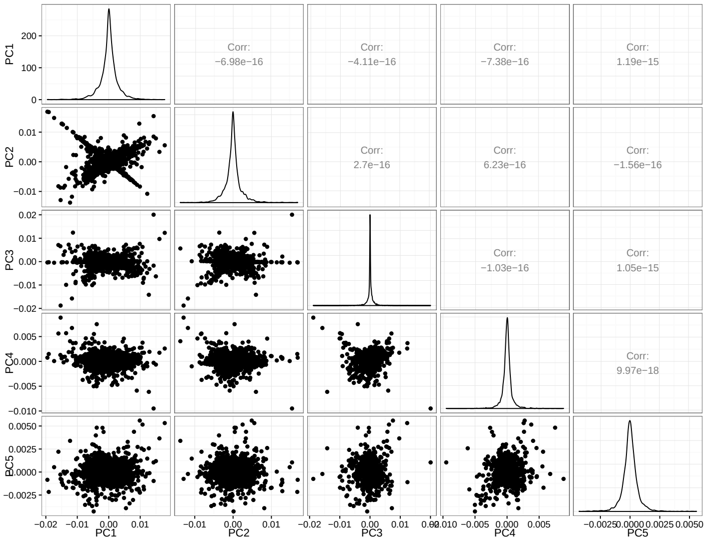

Figure 7: Pairwise plots of the first 5 P.C.s (weighted daily return)

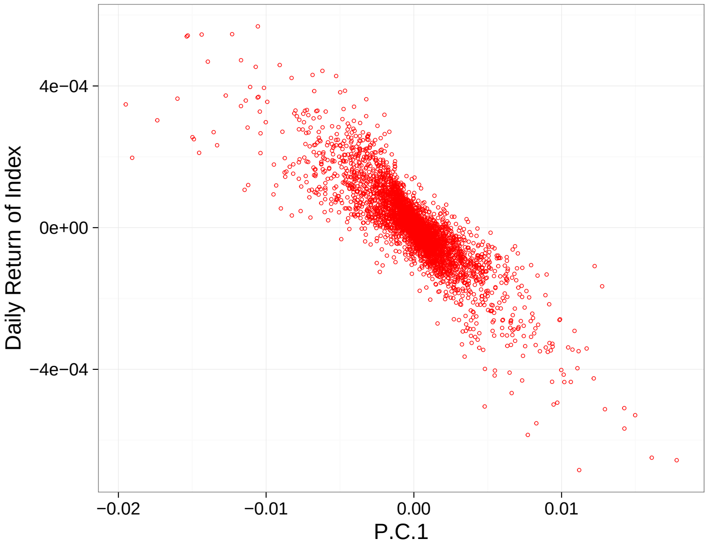

Figure 8: The relation between P.C.1 and CSI 300 index daily return (weighted daily return)

the stocks' daily returns. In the next plot, we choose 4 stocks (601318, 002202, 000977 and 601016, the ranks of their capitalization in the total 300 stocks are 1, 100, 200 and 300 respectively) and compare their reconstructed daily return with the original return.

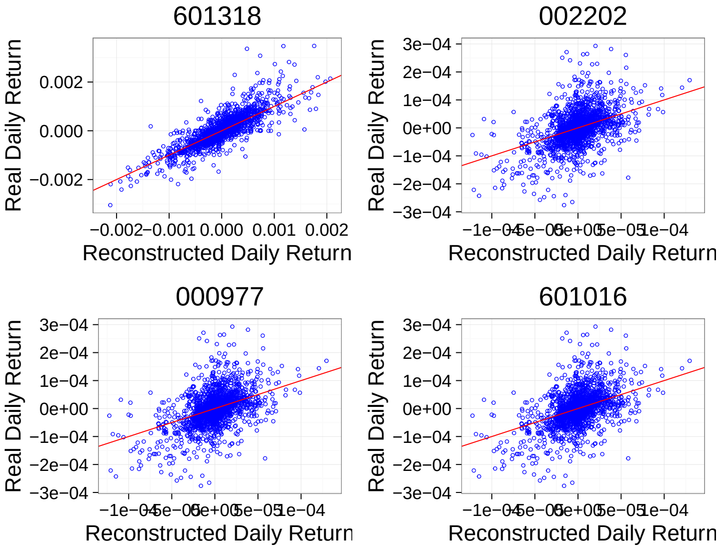

Figure 9: Comparison between reconstructed return and real return

From the above figure, we observe that the reconstruction works better for stocks with larger capitalization. This is understandable since in Part 2 we focus on weighted daily return.

Part 3

Following Part 2, now we separate the whole dataset (weighted daily return) into two parts: the training set is from 2001 to 2015; the testing set is all data points from 2016.

The importance matrix of the first 5 P.C.s is shown in the following table

	PC1	PC2	PC3	PC4	PC5
Standard deviation	0.003	0.002	0.001	0.001	0.001
Proportion of Variance	0.438	0.234	0.109	0.036	0.024
Cumulative Proportion	0.438	0.672	0.781	0.817	0.841

Table 9: The importance matrix for the first 5 P.C.s (training set)

Using the loading matrix, we reconstruct daily return in testing set. Similar to Figure 9, we pick up four stocks and compare their reconstructed daily return with the original daily return, as shown in Figure 10.

Again, the reconstruction is more successful for large stock such as 601318.

Example 3.3. (Factor analysis of daily returns.) Let us recall the return analysis in example 3.1. The data with 1462 daily returns has been introduced. We now use principle component analysis and maximum likelihood method to do factor analysis of the daily return. Taking $m = 2$, it is straightforward to obtain the solutions to the orthogonal factor analysis. To make a comparison, we also provide the ML solution for the factor loadings. The factor loadings, specific variances, and proportion of total (standardized) sample variance explained by each factor are provided in Table 10& 11.

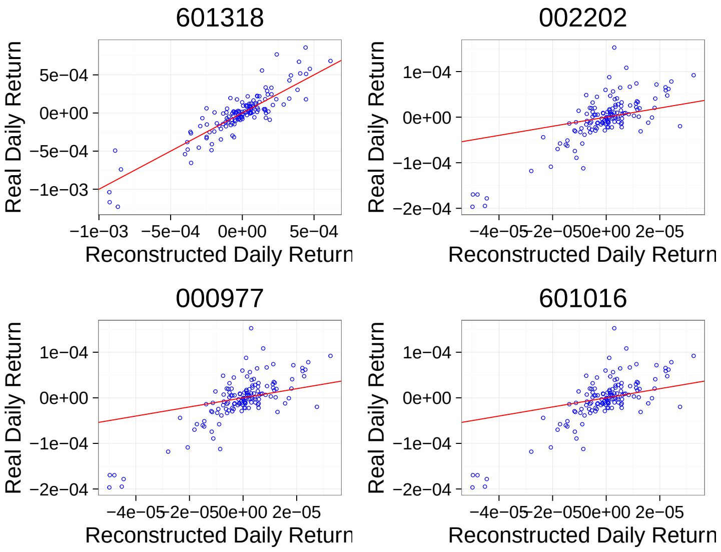

Figure 10: Comparison between reconstructed return and real return (testing set)

Table 10: Factor analysis results by PCA

Variable	Estimated factor loadings			Estimated rotated factor loadings		
	F_1	F_2	$\hat{\psi}_i = 1 - \hat{h}_i^2$	$\tilde{F}_1$	$\tilde{F}_2$	$\tilde{\psi}_i = 1 - \tilde{h}_i^2$
601398	0.84	-0.30	0.20	0.82	0.35	0.20
601939	0.89	-0.24	0.16	0.81	0.43	0.16
601288	0.90	-0.30	0.11	0.86	0.39	0.11
600028	0.84	0.43	0.11	0.32	0.89	0.11
601857	0.82	0.33	0.22	0.37	0.80	0.22
601088	0.72	0.13	0.46	0.44	0.59	0.46
Cumulative proportion of total variance explained	0.88	0.12		0.53	0.47	

The residual matrix of PCA factor analysis is

$$\hat{\mathbf{R}} = \begin{pmatrix} 0.203 & 0.001 & 0.003 & 0.002 & 0.008 & -0.017 \\ & 0.158 & 0.005 & -0.003 & -0.004 & 0.0138 \\ & & 0.105 & 0 & -0.002 & 0.004 \\ & & & 0.111 & 0.002 & 0.003 \\ & & & & 0.223 & -0.005 \\ & & & & & 0.456 \end{pmatrix}.$$

Actually, the results of PCA and maximum likelihood is rather similar. It seems fairly clear that the first factor, F_1 represents general economic conditions and might be called a market factor. All of the stocks load highly on this factor, and the loadings are about equal. The second factor contrast the. banking stocks with the oil stocks. Thus, F2 seems to differentiate stocks in different industries and might be called an industry factor. To sum, rates of return appear to be determined by general market conditions and activities that are unique to the different industries, as well as a residual or firm specific factor. This conclusion is coincident with Example 3.1.

Exercises.

3.1. Exercise 17.10 problem 1 of RU.

3.2. Exercise 17.10 problem 2 of RU.

3.3. Exercise 17.10 problem 3 of RU.

Table 11: Factor analysis results by maximum likelihood

Variable	Estimated factor loadings			Estimated rotated factor loadings		
	F_1	F_2	$\hat{\psi}_i = 1 - \hat{h}_i^2$	$\tilde{F}_1$	$\tilde{F}_2$	$\tilde{\psi}_i = 1 - \tilde{h}_i^2$
601398	0.84	-0.31	0.20	0.83	0.34	0.20
601939	0.88	-0.25	0.16	0.82	0.41	0.16
601288	0.89	-0.31	0.11	0.87	0.38	0.11
600028	0.85	0.43	0.09	0.33	0.89	0.09
601857	0.82	0.31	0.23	0.39	0.78	0.23
601088	0.73	0.12	0.46	0.45	0.58	0.46
Cumulative proportion of total variance explained	0.88	0.12		0.54	0.46	

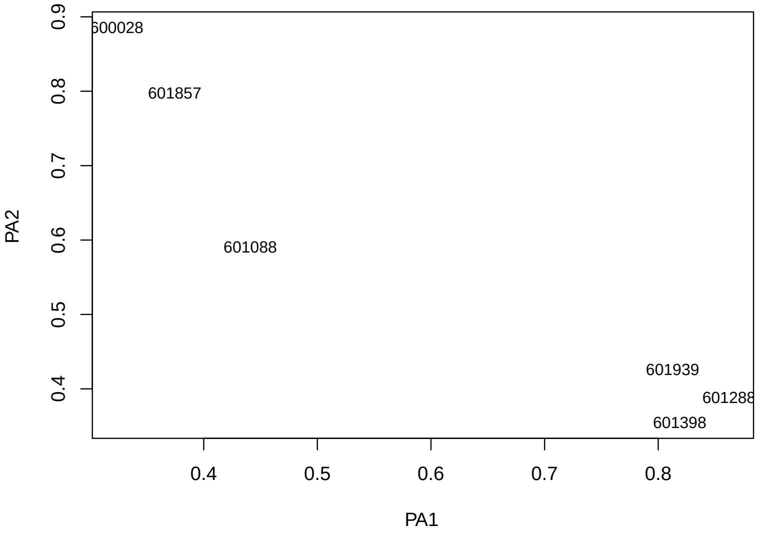

Figure 11: Plot of F_1 by F_2.