FE670 Algorithmic Trading Strategies

Factor Models and Estimation

Steve Yang

Stevens Institute of Technology

steve.yang@stevens.edu

09/16/2021

Overview

- Despite their apparent simplicity and widespread use, factor models entail conceptual subtleties that are not immediate to grasp.

  Static factor models represent a large number N of *random variables* in terms of a small number K of different *random variables* called factors.

  Dynamic factor models represent a large number N of *time series* in terms of a small number K of different *time series* called dynamic factors.

- In everyday language, we use the term factor to indicate something that has a causal link with an event. We apply the term factor both to identifiable exogenous events and to characteristics of different events.

- There are two important aspects of the formal factors:

  First, in science we call factors those variables that provide a common explanation of many other variables.

  Second, factors as observable variables might be used to predict additional observations but often hidden non-observable factors are the really important variables, and observations are used only to estimate them.

* For example, factor models were first introduced in psychometrics to find a common cause for many different psychometric observations coming from tests. Causes such as intelligence or personality traits are the important variables one wants to ascertain, while observations such as the results of psychometric tests are only used to determine hidden personality factors.

Static Factor Models

- Static factor models are factor models where factors do not have any dynamics. We will consider only *linear factor models* as they represent the vast majority of models used in finance.
- **Linear Factor Models**

  A linear factor model has the form

$$x_{it} = \alpha_i + \sum_{j=1}^{K} \beta_{ij} f_{jt} + \epsilon_{it}, i = 1, 2, ..., N, j = 1, 2, ..., K, t = 1, 2, ..., T$$

  where

  x_i = the i-th variable to explain

  α_i = the average of the i-th variable

  β_{ij} = the proportionality constant of the i-th variable to the j-th factor

  f_j = the j-th factor

  ϵ_i = the i-th residual term

Static Factor Models

- We can write the previous linear factor model in matrix form

$$\boldsymbol{\alpha} + \boldsymbol{\beta} f + \boldsymbol{\epsilon}$$

or explicitly

$$\begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} + \begin{bmatrix} \beta_{11} & \cdots & \beta_{1K} \\ \vdots & \ddots & \vdots \\ \beta_{N1} & \cdots & \beta_{NK} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

where

$\mathbf{x}$ = the N-vector of variables

$\boldsymbol{\alpha}$ = the N-vector of means of $\mathbf{x}$

$\boldsymbol{\beta}$ = the $N \times K$ constant matrix of factor loadings

$\mathbf{f}$ = the K-vector of factors

$\boldsymbol{\epsilon}$ = the N-vector of residuals

Static Factor Models

- Given M samples, we can rewrite this model in an explicit regression form.

$$\mathbf{X} = \mathbf{F}\mathbf{B}' + \mathbf{E}$$

where

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,N} \\ \vdots & \ddots & \vdots \\ x_{M,1} & \cdots & x_{M,N} \end{bmatrix}, \mathbf{F} = \begin{bmatrix} 1 & f_{1,1} & \cdots & f_{1,K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & f_{M,1} & \cdots & f_{M,K} \end{bmatrix}$$

$$\mathbf{E} = \begin{bmatrix} \epsilon_{1,1} & \cdots & \epsilon_{1,N} \\ \vdots & \ddots & \vdots \\ \epsilon_{M,1} & \cdots & \epsilon_{M,N} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \alpha_1 & \beta_{1,1} & \cdots & \beta_{1,K} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_N & \beta_{N,1} & \vdots & \beta_{N,K} \end{bmatrix}$$

* For the inference of a factor model of returns, we have only one realization of the process to rely on -namely, one observation from each point in time from past history.

Empirical Indeterminacy of the Model and Factor Rotation

We can always form linear combinations of variables: without restrictions the linear factor model states that residuals are a linear combination of observations and factors.

- A powerful restriction consists in requiring that residuals are zero-mean variables mutually uncorrelated and uncorrelated with factors.
- Such a model is called a strict factor model. We can write the model as follows:

$$
\begin{aligned}
\mathbf{x} &= \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{f} + \boldsymbol{\epsilon} \\
E(\boldsymbol{\epsilon}) &= 0 \\
E(\mathbf{f}) &= 0 \\
cov(\mathbf{f}, \boldsymbol{\epsilon}) &= 0 \\
cov(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) &= \mathbf{D}, \mathbf{D} = (\sigma_1^2, ...\sigma_N^2)\mathbf{I}_N \\
\mathbf{D} &= diag(\sigma_1^2, ...\sigma_N^2)
\end{aligned}
$$

Empirical Indeterminacy of the Model and Factor Rotation

If the model is formulated with explicit time dependence, then we assume that the same strict factor model holds at each time

$$\begin{aligned} \mathbf{x}_t &= \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{f}_t + \boldsymbol{\epsilon} \\ E(\boldsymbol{\epsilon}_t) &= 0 \\ E(\mathbf{f_t}) &= 0 \\ cov(\mathbf{f}_t, \boldsymbol{\epsilon}_t) &= 0 \\ cov(\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_t) &= \mathbf{D}, \mathbf{D} = (\sigma_1^2, ...\sigma_N^2)\mathbf{I} \end{aligned}$$

and that observations and residuals in different times are independent and identically distributed (i.i.d) variables.

* Note that requiring that the residuals are mutually uncorrelated and uncorrelated with factors is different from requiring that the residuals are i.i.d. variables.

** The former is an assumption on the model, the latter is an assumption on how different samples are distributed.

The Covariance Matrix of Observations

- The strict factor model conditions do not uniquely identify the factors. Given any nonsingular matrix A, we have

$$\mathbf{x} = \boldsymbol{\alpha} + \boldsymbol{\beta}^{*}\mathbf{f}^{*} + \boldsymbol{\epsilon} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{A}^{-1}\mathbf{A}\mathbf{f} + \boldsymbol{\epsilon} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{f} + \boldsymbol{\epsilon}$$

- We can use this fact to transform factors in a set of orthonormal variables (i.e. variables whose covariance matrix is the identify matrix). If we denote $\mathbf{Q}$ as the covariance matrix of factors $\mathbf{f}$, the covariance matrix $\mathbf{Q}^{*}$ of the transformed factors $\mathbf{f}^{*}$ can be written as:

$$\mathbf{Q}^{*} = E[\mathbf{f} * \mathbf{f}^{*'}] = E[(\mathbf{A}\mathbf{f})(\mathbf{A}\mathbf{f})'] = E[\mathbf{A}\mathbf{f}\mathbf{f}'\mathbf{A}] = \mathbf{A}\mathbf{Q}\mathbf{A}$$

- As $\mathbf{Q}$ is symmetric and positive semidefinite because it is a covariance matrix, its eigenvalues will be real non-negative numbers and the relationship $\mathbf{Q} = \mathbf{Q}^{-1}$ holds.

- Factor models are one of the tools that can be used to reduce the dimensionality of a covariance matrix of observations.

- In financial applications, the covariance matrix of observations is too large to be correctly estimated. Consider Russell 1000, the covariance matrix of $1,000$ time series of returns has $1,000 \times 999/2 = 499,500$ different entries.
- If we estimate covariances using four years of daily return data (approximately 1,000 days), we have a total of 1 million data points to estimate for about half a million entries, that, two data points per estimate.
- Using a strict factor model, we only need to determine the covariance matrix of factors plus the matrix of betas and the variances of the residuals.
- Assume a 10 factor model, we only need to estimate $15,950$ numbers, and we have an average of 62 data points per estimate (30 times improvement).

- Consider a factor model: $\mathbf{x} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{f} + \boldsymbol{\epsilon}$. we can write the covariance matrix of returns as

$$\boldsymbol{\Sigma} = E[(\mathbf{x}-\boldsymbol{\alpha})(\mathbf{x}-\boldsymbol{\alpha})'] = E[(\boldsymbol{\beta}\mathbf{f}+\boldsymbol{\epsilon})(\boldsymbol{\beta}\mathbf{f}+\boldsymbol{\epsilon})'] = \boldsymbol{\beta}\mathbf{Q}\boldsymbol{\beta}' + \mathbf{V} + 2\boldsymbol{\beta}E(\mathbf{f}\boldsymbol{\epsilon}')$$

where $\mathbf{V} = E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}']$. The last term is zero as factors and residuals are assumed to be independent. Therefore, we have

$$\boldsymbol{\Sigma} = \boldsymbol{\beta}\mathbf{Q}\boldsymbol{\beta}' + \mathbf{V}$$

- This formula simplifies further if we apply a factor rotation that transforms factors into standardized orthonormal factors whose covariance matrix is the identity matrix

$$\boldsymbol{\Sigma} = \boldsymbol{\beta}\boldsymbol{\beta}' + \mathbf{V}$$

- If our factor model is a strict factor model then the matrix $\mathbf{V}$ becomes a diagonal matrix $\mathbf{D}$ and the covariance matrix becomes

$$\boldsymbol{\Sigma} = \boldsymbol{\beta}\boldsymbol{\beta}' + \mathbf{D}, \text{ or } \boldsymbol{\Sigma} = \boldsymbol{\beta}\boldsymbol{\beta}' + \sigma^2\mathbf{I}$$

Using Factor Models

- In financial applications, portfolios are constructed performing a risk-return optimization that requires computing a covariance matrix.
- Computing an unrestricted covariance matrix is not feasible for large universe. Factor models reduce the calculations to computing the exposures to each factor plus the small factor covariance matrix and idiosyncratic error variances.
- Factor models thus used are risk models. They explain returns at time t as linear combinations of factors given at time t.
- Risk measured by variance of returns, due to exposures to common factors is the residual undiversifiable risk. This risk cannot be diversified away regardless of how large a portfolio we choose.
- If factors can be forecast-ed, or if factors that explain returns at time t are known at time $t-1$, then factor models can be used for forecasting returns.

Factor Analysis and Principal Components Analysis

What is factor analysis? A process to determine statistical factors. Factors are not observed but must be determined together with the model. The result of the factor analysis will be:

1. A (multivariate) time series of factors.
2. A (multivariate) time series of residuals.
3. The covariance matrix of factors.
4. The factor loadings for each return process.
5. The variances of each residual term.

- The factor loadings represent the exposures of returns to each factor. We might use these numbers in sample (backward looking), for example to evaluate the risk associated with a fund manager's performance.
- If we need to optimize a portfolio, we have to use our estimates out of sample (forward looking).

Factor Analysis and Principal Components Analysis

There are two basic techniques for estimating factor models: factor analysis and principal component analysis.

- **Factor Analysis:** Let's consider a strict factor model of returns with standard (orthonormal) factors. Without loss of generality we make the additional assumption $\alpha = 0$. if we call the return $\mathbf{r}_t$, the model can be written as

$$\begin{aligned}
\mathbf{r}_t &= \beta \mathbf{f}_t + \epsilon_t \\
E(\epsilon_t) &= 0 \\
E(\mathbf{f_t}) &= 0 \\
cov(\mathbf{f}_t, \epsilon_t) &= 0 \\
cov(\epsilon_t, \epsilon_t) &= \mathbf{D}, \mathbf{D} = (\sigma_1^2, ...\sigma_N^2)\mathbf{I} \\
cov(\mathbf{f}_t, \mathbf{f}_t) &= \mathbf{I}
\end{aligned}$$

- Under the above assumptions, we have formulation:

$$\mathbf{R} = \mathbf{F}\beta^{'} + \mathbf{E}$$

Factor Analysis via Maximum Likelihood

- There is a complete and rigorous procedure for linear strict factor models under the assumptions that returns, factors, and residuals are multivariate normal variables. Therefore, we can use the maximum likelihood estimation (MLE) principle to estimate the model parameters.

$$\begin{aligned} \mathbf{r}_t &\sim N(0, \mathbf{\Sigma}) \\ \mathbf{f}_t &\sim N(0, \mathbf{I}_K) \\ \epsilon_t &\sim N(0, \mathbf{D}) \end{aligned}$$

Given the assumption that our model is a strict factor model with orthonormal factors, the covariance matrix of return can be represented as:

$$\mathbf{\Sigma} = \beta\beta' + \mathbf{D}, \mathbf{D} = (\sigma_1^2, ...\sigma_N^2)\mathbf{I}$$

Factor Analysis via Maximum Likelihood

- The MLE principle estimates the parameters of a model by maximizing the *likelihood* of the model. The likelihood of a model is the product of the densities of the model's variables estimated on all samples. Under the assumption of joint normality and zero means, we can explicitly write the joint distribution of returns:

$$\mathbf{r}_t \sim N(0, \mathbf{\Sigma}) = \left[(2\pi)^N |\mathbf{\Sigma}|\right]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{r}_t' \mathbf{\Sigma}^{-1} \mathbf{r}_t\right\}$$

and the likelihood is given by

$$L(\mathbf{\Sigma}) = \left[(2\pi)^N |\mathbf{\Sigma}|\right]^{-\frac{1}{2}} \prod_{t=1}^{T} \exp\left\{-\frac{1}{2}\mathbf{r}_t' \mathbf{\Sigma}^{-1} \mathbf{r}_t\right\}$$

$$\ell(\mathbf{\Sigma}) = -\frac{NK}{2} log(2\pi) - \frac{K}{2} log(|\mathbf{\Sigma}|) - \frac{1}{2} \sum_{t-1}^{T} \left\{\mathbf{r}_t' \mathbf{\Sigma}^{-1} \mathbf{r}_t\right\}$$

Factor Analysis via Maximum Likelihood

- The maximum likelihood expectation (MLE) method maximizes the log-likelihood $\ell(\mathbf{\Sigma})$ as a function of the covariance matrix:

$$\mathbf{\Sigma} = \arg\max_{\Sigma} \ell(\mathbf{\Sigma})$$

  using the restriction

$$\mathbf{\Sigma} = \beta\beta' + \mathbf{D}, \mathbf{D} = (\sigma_1^2, ...\sigma_N^2)\mathbf{I}$$

- A numerical method for MLE is based on the *expectation maximization* EM algorithm. The EM algorithm is an iterative procedure for determining the log-likelihood when some variables are missing or when there are hidden variables.

The Expectation Maximization Algorithm

- The EM algorithm assumes that besides observed data there may be missing or so-called hidden data. The observed data are the returns $\mathbf{r}_t$ while the missing data are hidden, non-observed factors $\mathbf{f}_t$.
- The observed returns $\mathbf{r}_t$ and the unobserved factors $\mathbf{f}_t$ are called complete data. Call $\mathbf{z}_t$ the vector of complete data at time t

$$\mathbf{z}_t = \begin{bmatrix} \mathbf{r}_t \\ \mathbf{f}_t \end{bmatrix}$$

- If factors were observed, we could apply the MLE principle in a straightforward way by maximizing the likelihood given all observed data. However, when factors are not observed, we need iterative methods to compute the model parameters.

The Expectation Maximization Algorithm

- Intuitively speaking, the EM method is an iterative Bayesian method that alternates between two steps, the E step and the M step.
- The E step assumes that the model parameters are known and makes the best estimate of the hidden variables given the observed data the the model parameters estimated at the previous step.
- The M step computes new model parameters via ML estimates using the hidden data estimated in the previous step. The new parameters are then used to form new estimates of the hidden data and a new cycle is performed.
* This simplified description highlights the Bayesian nature of the EM method in the E step where hidden factors are estimated given actual observed data.

Dempster, Laird and Rubin (DLR) described that, in the case of factor analysis with normal distributions, the EM algorithm simplifies as:

1. The E step computes the expectation of the sufficient statistics of the log-likelihood given the observed data and the matrices β_p and $\mathbf{D}_p$ computed in the previous M step.
2. The M step computes new matrices β_{p+1} and $\mathbf{D}_{p+1}$ using the sufficient statistics computed in the previous E step.

We have to maximize the quantity:

$$\ell_C = -\frac{T}{2} log(|\mathbf{D}|) - \frac{1}{2}\sum_{t-1}^{T}\left\{\mathbf{r}_t^{'}\mathbf{D}^{-1}\mathbf{r}_t - 2\mathbf{r}_t^{'}\mathbf{D}^{-1}\beta\mathbf{f}_t + Tr\left[\beta^{'}\beta^{'}\mathbf{D}^{-1}\mathbf{f}_t^{'}\mathbf{f}_t\right]\right\}$$

as the other terms of the complete likelihood do not depend on β and $\mathbf{D}$.

The Expectation Maximization Algorithm

- **The E-Step**

  The E-step computes the expectation of the log-likelihood of the complete data given the observed data. The joint distribution $p(\mathbf{r}_t, \mathbf{f}_t|\beta, \mathbf{D})$ of the complete variables is normal given the linearity of the model and yields

$$E(\mathbf{z}_t|\beta, \mathbf{D}) = E\left(\begin{bmatrix} \mathbf{r}_t \\ \mathbf{f}_t \end{bmatrix}\right) = 0$$

$$cov(\mathbf{z}_t\mathbf{z}_t|\beta, \mathbf{D}) = \Lambda$$

$$= E\begin{bmatrix} cov(\mathbf{r}_t\mathbf{r}_t|\beta, \mathbf{D}) & cov(\mathbf{r}_t\mathbf{f}_t|\beta, \mathbf{D}) \\ cov(\mathbf{r}_t\mathbf{f}_t|\beta, \mathbf{D}) & cov(\mathbf{f}_t\mathbf{f}_t|\beta, \mathbf{D}) \end{bmatrix}$$

$$= E\begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$$

The Expectation Maximization Algorithm

- As $E(\mathbf{r}_t, |\beta, \mathbf{D}) = 0$, the complete data sufficient statistics are $\Lambda_{11}, \Lambda_{12}, \Lambda_{22}$. The E-step at step p replaces the complete data sufficient statistics with their expectations given the data. Following Rubin and Thayer, the sufficient statistics are

$$
\begin{aligned}
E(\mathbf{\Lambda}_{11}|\mathbf{r}_t, \beta_p, \mathbf{D}_p) &= E\left[cov(\mathbf{r}_t\mathbf{r}_t|\beta_p, \mathbf{D})_p\right] \\
&= cov(\mathbf{r}_t\mathbf{r}_t) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_t'\mathbf{r}_t \\
E(\mathbf{\Lambda}_{12}|\beta_p, \mathbf{D}_p) &= E\left[cov(\mathbf{r}_t\mathbf{f}_t|\mathbf{r}_t)\right] = \beta_p\gamma_{p+1} \\
E(\mathbf{\Lambda}_{22}|\beta_p, \mathbf{D}_p) &= E\left[cov(\mathbf{f}_t\mathbf{f}_t|\mathbf{r}_t)\right] = \gamma_p'\mathbf{\Lambda}_{22}\gamma_{p+1} + \mathbf{\Delta}_{p+1}
\end{aligned}
$$

where γ_{p+1} and $\mathbf{\Delta}_{p+1}$ are the regression coefficient matrix and the residual covariance matrix of the regression of the factors $\mathbf{f}_t$ on the returns $\mathbf{r}_t$, respectively.

The Expectation Maximization Algorithm

- Compute the expectation of ℓ_C given the data:

$$E[\ell_C] = -\frac{T}{2} log(|\mathbf{D}|)$$

$$-\frac{1}{2}\sum_{t-1}^{T}\left\{\mathbf{r}_t^{'}\mathbf{D}^{-1}\mathbf{r}_t - 2\mathbf{r}_t^{'}\mathbf{D}^{-1}\beta E[\mathbf{f}_t|\mathbf{r}_t,\beta_p,\mathbf{D}_p]\right\}$$

$$-\frac{1}{2}\sum_{t-1}^{T}\left\{Tr\left[\beta^{'}\beta^{'}\mathbf{D}^{-1}E[\mathbf{f}_t^{'}\mathbf{f}_t|\mathbf{r}_t,\beta_p,\mathbf{D}_p]\right]\right\}$$

* Given the joint normality of the model, we need to compute only the means $E(\mathbf{f}_t|\mathbf{r}_t,\beta_p,\mathbf{D}_p])$ and the covariances $E[\mathbf{f}_t^{'}\mathbf{f}_t|\mathbf{r}_t,\beta_p,\mathbf{D}_p]$. In order to compute these (sufficient) statistics, we need the distribution of factors given the data, but the model prescribes the distribution of data given the factors. Here we need to employ Bayes' theorem $p(\mathbf{f}_t|\mathbf{r}_t,\beta,\mathbf{D}) \propto p(\mathbf{r}_t|\mathbf{f}_t,\beta,\mathbf{D})p(\mathbf{f}_t)$

The Expectation Maximization Algorithm

- **The M-Step**

  First we replace the sufficient statistics of the log-likelihood with their expectations computed in the E-step. Then we maximize the log-likelihood equating to zero its partial derivatives with respect to β and $\mathbf{D}$:

$$\frac{\partial}{\partial \beta} E[log(L)] = 0, \frac{\partial}{\partial \mathbf{D}} E[log(L)] = 0$$

* Given that the complete data distribution is normal, we can use a simplified process based on the regressions outlined in DLR and described in detail in Rubin and Thayer.

How to Determine the Number of Factors

- Many different theoretical and heuristic solutions have been proposed.
  1. Heuristic solutions are based on estimating the incremental gain in model quality in going from p to $p+1$ factors. Perhaps the best known of these heuristics is the *scree test* proposed by Cattell. The scree test is based on plotting the eigenvalues of the covariance matrix in order of size. Cattell observed that this plot decreases sharply up to a certain point and then slows down. The point where it starts to slow down is an approximate estimate of the number of factors.
  2. Theoretical solutions are often based on Information Theory criteria. Information-based criteria introduce a trade-off between the size of residuals and model complexity.

* Note that if returns could be represented by a correctly specified strict factor model, the number of factors would be determined and factors would be empirically determined up to a linear transformation.

Factor Analysis Example Python

- In Python, this package is made available with
  $>$ *library*(*sklearn.decomposition.FactorAnalysis*).

- The observations are assumed to be caused by a linear transformation of lower dimensional latent factors and added Gaussian noise. Without loss of generality the factors are distributed according to a Gaussian with zero mean and unit covariance. The noise is also zero mean and has an arbitrary diagonal covariance matrix.

- FactorAnalysis performs a maximum likelihood estimate of the so-called loading matrix, the transformation of the latent variables to the observed ones, using SVD based approach.