A Guide to Factor Analysis

Pranav Nanga

1 Overview

Consider a universe of p stocks, and suppose we have a price return series of n ticks for each. We can mathematically view this as n multivariate p-dimensional random variables $X_1, X_2, \ldots, X_n$, each drawn i.i.d from a distribution with mean 0 and some covariance matrix Σ, where the sample at time t is given by

$$X_{*t} = \begin{bmatrix} X_{1t} \\ X_{2t} \\ \vdots \\ X_{pt} \end{bmatrix}, \qquad t = 1, 2, \ldots, n$$

Our goal, as described in the case packet, is to construct an estimate of the covariance matrix $\widehat{\Sigma}$. The naive way of doing this is by calculating the covariance matrix of our random sample, via the formula:

$$\widehat{\Sigma}_{sam} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{*i} - \overline{X})(X_{*i} - \overline{X})^T$$

Note that we are trying to estimate $\frac{1}{2}(p^2 - p) + p$ parameters (the diagonal and one triangle of the covariance matrix) with t samples. When t is small compared to p, this estimate breaks down. The purpose of factor modeling is to decompose the return series into a smaller set of factors.

2 General framework of factor models

Now suppose that there are m common features called *factors* that explain the returns of all the stocks in the market. Let $X_{1*}, \ldots, X_{p*}$ be time series of each stocks' returns and $F_{1*}, \ldots, F_{m*}$ be time series of each individual factor. More specifically, let

$$X_{i*} = \begin{bmatrix} X_{i1} & X_{i2} & \cdots & X_{in} \end{bmatrix}, \qquad F_{j*} = \begin{bmatrix} F_{j1} & F_{j2} & \cdots & F_{jn} \end{bmatrix}$$

Then, we assume that the return series of a given stock are

$$\begin{aligned} X_{1*} &= \mu_1 + \beta_{11} F_{1*} + \beta_{12} F_{2*} + \cdots + \beta_{1m} F_{m*} + \epsilon_1 \\ X_{2*} &= \mu_2 + \beta_{21} F_{1*} + \beta_{22} F_{2*} + \cdots + \beta_{2m} F_{m*} + \epsilon_2 \\ &\vdots \\ X_{p*} &= \mu_p + \beta_{p1} F_{1*} + \beta_{p2} F_{2*} + \cdots + \beta_{pm} F_{m*} + \epsilon_p \end{aligned}$$

where μ_i, ϵ_i are constants representing the mean and noise per stock series, respectively, and β_{ij} are coefficients (called factor *loadings*) representing how stock i is affected by factor j. We can convert this set of equations into matrix form, via

$$X = \begin{bmatrix} - & X_{1*} & - \\ - & X_{2*} & - \\ & \vdots & \\ - & \vdots & - \\ - & X_{p*} & - \end{bmatrix}, \quad F = \begin{bmatrix} - & F_{1*} & - \\ - & F_{2*} & - \\ & \vdots & \\ - & \vdots & - \\ - & F_{m*} & - \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pm} \end{bmatrix}$$

to get the expression

$$X = \mu + \beta F + \epsilon$$

Note that this is a general framework of using a smaller set of factors to explain the returns of a larger set of stocks. The exact factors can either be statistically or fundamentally determined depending on the thesis of the modeler. The loadings can be time-invariant and estimated from the data, or even varying over time! We will present one such interpretation of the general factor model that may help you approach the case.

3 The BARRA approach

The BARRA model is an example of a fundamental factor model – that is, the factors affecting stock returns are observable and fundamentally determined features of the stocks. For example, the market capitalization of a stock or the GDP could be factors affecting returns. But since factors like market capitalization are not common to all stocks, we slightly modify our interpretation of the previous framework. Rather than letting the matrix F contain a time series of each individual factor, we store these features in β and determine F via multiple linear regression. Here, $[\beta]_{ij}$ is measurement of factor j on stock i, and entries of F represent the weighting of features affecting a given stock.
Spefically, suppose that R_t represents the excess returns of the stocks and F_t the weightings at time t. Then the factor model becomes

$$R_t = \beta F_t + \epsilon, \qquad R_t = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_p \end{bmatrix}$$

Once we have estimated F_t, we can estimate the covariance matrix of excess returns.

4 Application to the case

Note that the factors comprising the matrix β will not simply be the features we provide at each timestamp. Your job is to use the training data we provide to derive, from the provided features, a set of factors that explain the training data. To simplify things, you can assume that while the data inside β will change, the factors that comprise the matrix will be consistent across the training data and the competition. β can easily be expanded to include more data than time t, but the tradeoff is that you will have to estimate more parameters and the factor weightings in F may be less accurate. You can assume that stock returns will only rely on data from the current timestep. You can also assume that the loadings will change every 21 timesteps. There are still several questions you must answer, such as:

- How do you estimate $\widehat{\Sigma}_R$?
- How will you know if a set of factors (β) properly models the training data provided?
- How will your algorithm determine covariance matrices at test time?
- How will your algorithm determine the proper weight allocation of the portfolio?