An Incomplete Guide to Factor Models

Haochen Wang

1 Factor Models

- Let us first discuss population factor models. It is important to distinguish between population and sample analysis. The former focuses on random variables (or vectors) while the latter attends to actual observations. The difference is more than notational.
- A factor model attempts to explain market returns using the following framework

$$\begin{aligned} R_1 &= \mu_1 + l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \epsilon_1 \\ R_2 &= \mu_2 + l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \epsilon_2 \\ &\vdots \\ R_p &= \mu_p + l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \epsilon_p \end{aligned}$$

  where $R_1 \ldots R_p$, $F_1 \ldots F_m$, $\epsilon_1 \ldots \epsilon_p$ are scalar random variables, while $\mu_1 \ldots \mu_p$ and l_{ij} are constants. Note that there is no notion of time series here at all, since we are concerned with random variables. There is no data/observations so far.
- Putting this into vector-matrix notation, a factor model may be written as

$$\mathbf{R} = \mu + L\mathbf{F} + \epsilon$$

  where $\mu \in R^p$ and $L \in R^{p \times m}$.
- Fama-French variants, MSCI Barra, and Statistical Factor Models all fit into this framework. Usually m is much smaller than p, and this enables us to analyze a large set of random variables R_i through a small set of random variables F_i. The philosophies of the three models differ on what is observed in the market.
- The constants μ are colloquially referred to as alphas, and the constants L_{ij} the betas. The log return of an asset is controlled by log factor returns through alpha and beta. Note that if our factor model is sufficiently powerful as to be able to explain the market in its entirety, then alpha disappears. That what we meant by alpha exhaustion in the case packet.
- The idiosyncratic noise random vector ϵ are assumed to have expectation zero and are uncorrelated across assets i.e. a diagonal covariance matrix. The log asset return random vector $\mathbf{R}$ usually have non-zero expectation and non-diagonal covariance matrix. The same is true of log factor returns random vector $\mathbf{F}$.
- The random vectors/variables are observed in samples through time series, and so we introduce an additional index t.

$$\begin{aligned} R_{1t} &= \mu_1 + l_{11}F_{1t} + l_{12}F_{2t} + \ldots + l_{1m}F_{mt} + \epsilon_{1t} \\ R_{2t} &= \mu_2 + l_{21}F_{1t} + l_{22}F_{2t} + \ldots + l_{2m}F_{mt} + \epsilon_{2t} \\ &\vdots \\ R_{pt} &= \mu_p + l_{p1}F_{1t} + l_{p2}F_{2t} + \ldots + l_{pm}F_{mt} + \epsilon_{pt} \end{aligned}$$

  or

$$\mathbf{R_t} = \mu + L\mathbf{F_t} + \epsilon_\mathbf{t}$$

- Our aim in this case is to understand the mean and covariance structure of $\mathbf{R}$, and it is up to you to find out the story the data tells.

2 Fama-French vs Barra vs Statistical Factor Models

We see that the motivation of a factor model is primarily one of dimensionality reduction. It is up to the practitioners to decide the exact implementations. The main tension arises as to what in this factor model is actually observable.

2.1 Fama-French

Fama and French assume that the factor returns $\mathbf{F_t}$ are observed. They constructed intricate long-short portfolios SMB, HML etc to approximate the returns of abstract factors such as Size and Value. Given these factor return series, they estimate L_{ij} and ϵ, and thus attempt to explain asset returns.

The downside (maybe?) of this approach is that there is no reason why the return a factor such as Value must be captured by the particular formulation of Kenneth French's SMB portfolio. I can do mine, you yours. Different ways of proxy constructions could make the results drastically different. In fact, Kenneth French is diligently publishing the returns of his carefully constructed portfolios on his Dartmouth homepage. You may want to take a look.

For a trading competition like ours, if we were to generate data using Fama-French, and therefore assume a few ground truth portfolios that are factor return proxies, then we must demand all participants to conform to the same. However, there is no reason why my SMB long-shorting top and bottom quintiles must be superior to yours long-shorting top and bottom thirds. Not very fair and hard to test.

2.2 MSCI Barra

MSCI Barra factor models are not created by academics, and are a little more usable (a bigger maybe?). Here it is the loadings matrix L_{ij} that is observed. As unintuitive as it may sound, if you think of the loadings as the exposure to a particular factor, then we can just treat asset properties such as Price-to-Earnings ratio, etc as loadings. In this case, even though L are constants, they may evolve over time. That does not make L a random variable(matrix). To actually use these properties, you should do some simple transformations and mean-std normalizations. Given L_{ij}, we estimate the distributions of F and ϵ, and go from there.

Asset properties can be directly observable, e.g. those we see on Morningstar, Google Finance, but they can be more derived and involved. In fact there are people selling their composite asset loadings which are supposed to be highly explanatory. Guess who? Of course, MSCI. I am surprised to learn that it is a thriving business.

2.3 Statistical Factor Models

Throw away the pretense of finance! It is but dimensionality reduction. Nothing is observed, for nothing is needed. We just treat both L and F as unobserved and do some estimation so long as there is some internal consistency. You may try this on our data. Note that in this case a curious property of rotational invariance occurs to the learnt loadings.

3 Conclusion

Factor Modeling, if viewed through the lens of machine learning, is fundamentally a kind of regularization. The fact that the condition number of a matrix can be linked to overfitting is a rather intriguing insight.

Don't be excited when you see that your vanilla sample covariance estimation does really well on the training data. Exercising restraints prepares you to confront the unknown.