IEOR 4739
Factor models

Suppose we have a collection of n assets. A *factor model* for the asset returns is a statistical model of the form:

$$r \;=\; \mu + \epsilon + V^T f \tag{1}$$

where

- μ is the n-vector of expected returns (computed from historical data),
- r is the n-vector of returns (a random variable),
- ϵ is an n-vector of *idiosyncratic errors* (a random variable),
- f is an p-vector of *factors* (a random variable) and V is an $p \times n$-matrix of "factor exposures".

The factors are, usually, major economic factors (inflation, etc.) that can be readily measured and for which historical data is available. These factors are all measured relative to their means (see below). Typically, p is much smaller than n. The matrix V is obtained through a multivariate regression, to correlate assets to factors.
It is assumed that:

1. $E(f) = 0$ (we simply measure deviation from the mean of each factor),
2. $E(\epsilon) = 0$. For an asset j, ϵ_j measures the "variation" of r_j that is "intrinsic" to asset j. Zero expectation means that this variation adds noise to the return which tends to correct over time.
3. For any two *different* assets j and k, $E(\epsilon_j \epsilon_k) = 0$. Same as in the previous item.
4. For an asset j and factor i, $E(\epsilon_j f_i) = 0$.

Denote by F the factor covariance matrix, i.e. $F_{ik} = E(f_i f_k)$. This matrix can again be estimated from historical data. Further, for any asset j let σ_j^2 be the variance of ϵ_j, i.e. $\sigma_j^2 = E(\epsilon_j^2)$. This is the part of the variance for the return of asset j that is entirely due to asset j itself. Below we will discuss how this is computed.

From equation (1) and the assumptions, we have that $E(r) = \mu$, as desired. Furthermore, for any assets j and k, the covariance between r_j and r_k equals:

$$
\begin{aligned}
E((r_j - \mu_j)(r_k - \mu_k)) &= E[(\epsilon_j + \sum_i V_{ij} f_i)(\epsilon_k + \sum_h V_{hk} f_h)] = && (2) \\
&= E(\epsilon_j \epsilon_k) + \sum_i \left(V_{ij} \sum_h V_{hk} E(f_i f_h) \right) = && (3) \\
&= E(\epsilon_j \epsilon_k) + \sum_i \left(V_{ij} \sum_h V_{hk} F_{ih} \right). && (4)
\end{aligned}
$$

It is seen that the second term in the last equation equals

the j, k-term of $V^T F V$,

whereas the first term equals 0 if $j \neq k$, and it equals σ_j^2 if $j = k$. In summary, denoting by Q the return covariance matrix, we have that:

$$
Q = diag(\sigma^2) + V^T F V \qquad (5)
$$

where $diag(\sigma^2)$ is the $n \times n$ diagonal matrix whose j, j entry equals σ_j^2.

Now suppose we have a typical mean-variance optimization problem, e.g. a problem of the form

$$
\begin{aligned}
& \text{minimize } \lambda x^T Q x - \mu^T x \\
\text{s.t.} & \\
& \sum_j x_j = 1, \\
& Ax \geq b. \\
& 0 \leq x.
\end{aligned}
$$

using (5) the problem becomes:

$$
\begin{aligned}
& \text{minimize } \lambda \sum_j \sigma_j^2 x_j^2 + \lambda y^T F y - \mu^T x \\
\text{s.t.} & \\
& y - Vx = 0 \\
& \sum_j x_j = 1, \\
& Ax \geq b. \\
& 0 \leq x.
\end{aligned}
$$

Here, the y are additional variables (p of them).

A remaining issue is how to set the σ_j^2. One approach for doing this is to simply construct the statistical model

$$r - \mu - V^T f \quad = \quad \epsilon. \tag{6}$$

The quantity in the left-hand side has mean zero, and its covariance matrix is $diag(\sigma^2)$. In other words, we compute the variance of the part of return *not* explained by μ and $V^T f$. From a practical perspective, some of the computed σ_j^2 may be too small – they may amount to "noise". Practitioners would simply use zero instead, with the result that we are now just approximating (though closely) the matrix in the left-hand side of (5).

A heuristic approach.

As an alternative to using economic factors, the technique of "principal components analysis" constructs a model of the form (1) where the factors are "synthetic" and are derived from the data itself. We will return to this later in the course, but instead let us first consider a heuristic sometimes used by practitioners.

Suppose we want to build a model of the form (5), using synthetic factors, and further, suppose we have a "good" estimate of the matrix $diag(\sigma^2)$. We have Q (computed from historical data) and let

$$\tilde{Q} = Q - diag(\sigma^2).$$

Since $\tilde{Q}$ is symmetric, positive-semidefinite, it can be written as

$$\tilde{Q} = LDL^T,$$

where D is a diagonal matrix with nonnegative entries, and L is a lower triangular matrix (this is a fact from Linear Algebra). Suppose the assets have been numbered so that the entries of D satisfy: $d_1 \geq d_2 \geq \ldots \geq d_n$. Typically, many of the smaller d_i will be close to zero.

Suppose we **approximate** D with the diagonal matrix $\bar{D}$ that only keeps the "largest" entries of D, that is to say $\bar{d}_j = d_j$ for $j = 1, 2, \ldots p$ (where p is small compared to n); and $\bar{d}_j = 0$ for $j > p$. Then

$$\bar{Q} = L\bar{D}L^T,$$

is an approximation to Q. If the values $d_{r+1}, d_{r+2}, \ldots, d_n$ are indeed very small, then $\bar{Q} \approx Q$. Furthermore, if we define

$$\bar{L} = \text{submatrix of } L \text{ consisting of the first } p \text{ columns}$$

then it is not hard to see that

$$\bar{Q} \quad = \quad \bar{L}\bar{F}\bar{L}^T, \tag{7}$$

where $\bar{F} = diag(d_1, d_2, \ldots, d_r)$. This is our approximate "factor model", whose explanatory power needs to be tested – the testing may result in our having to update the quantities σ_j^2.

Another heuristic approach.

Suppose that we believe that a certain (not too large) subset of K assets indirectly control (or approximately control) the behavior of all other assets. We can then use these assets to generate a factor model with K factors. Some questions that arise then are:

(i) How do we choose the K assets? Assuming (for now) that we have selected a value for K then one could choose those K assets for which the sum of absolute values of correlations with the other assets is largest. Or, one might have an idiosyncratic choice in mind (i.e., we strongly feel that Coca-Cola should be in the list).

(ii) Having chosen the assets, how do we find the matrix V? This amounts to setting up a linear regression model; how would you do it?

(iii) And, finally, how do we pick K?

A common approach.

This involves using the so-called 'principal components model' – we saw a bit of this today, more next lecture.