5. Multi-Factor Models and Ross' Arbitrage Pricing Theory

5.1. **Multi-Factor Models.** The Capital Asset Pricing Model, as discussed in the last lecture, has been generalized by many authors. Perhaps the most important class of generalizations are the so-called *multi-factor models* which really originated in the landmark paper of Ross (1976), which is among the most highly cited papers in finance. See also Roll and Ross (1980) for related empirical work. There are also some excellent textbooks and reference sources geared to practitioners; see Fabozzi, Focardi, and Kolm (2010), Grinold and Kahn (2000), and Connor, Goldberg, and Korajczyk (2010). For the mathematics of linear regression, read Chapter 3 of Friedman, Hastie, and Tibshirani (2001).

Multi-factor models assume a linear functional form

$$R_{t+1} = X_t f_{t+1} + \epsilon_{t+1}, \quad \mathbb{E}[\epsilon] = 0, \quad \mathbb{V}[\epsilon] = D \tag{5.1}$$

where R_{t+1} is an n-dimensional random vector containing the cross-section of returns in excess of the risk-free rate over some time interval $[t, t+1]$, and X_t is a (non-random) $n \times p$ matrix that can be calculated entirely from data known before time t. The variable f in (5.1) denotes a p-dimensional random vector process which cannot be observed directly.

Definition 5.1. A *strict factor model* is one in which the variance-covariance matrix of the residuals is diagonal:

$$D := \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2) \text{ with all } \sigma_i^2 > 0. \tag{5.2}$$

The factor model is said to be *noiseless* if $D = 0$.
We will henceforth restrict attention to strict factor models. Eq. (5.2) entails that all significant sources of correlation are already captured by factors, represented as columns of X_t. We henceforth suppress the implicit time index whenever our discussions concern a single time interval. Note, for later use, that D^{-1} exists and can be computed in $O(n)$ time because it is diagonal.

Definition 5.2. For a portfolio with holdings vector $h \in \mathbb{R}^n$, the vector $h'X \in \mathbb{R}^p$ is called the *exposure vector* of the portfolio. The j-th element of $h'X$ is called the *exposure* of h to the j-th factor.

In our discussion of the CAPM, we proved that the beta of a portfolio $h \in \mathbb{R}^n$ is given by $\beta_h = \sum_i h_i \beta_i$. Definition 5.2 is a direct generalization of this. In fact, it's a very reasonable thing to choose one of the columns of X to be a cross-section of estimated betas, in which case the exposure of portfolio h to the beta factor will exactly equal β_h from the CAPM.

5.2. **Statistically-estimated factor returns.** Since the variable f in (5.1) denotes a p-dimensional random vector process which cannot be observed directly, information about the f-process must be obtained via statistical inference. We assume that the f-process has finite first and second moments given by

$$\mathbb{E}[f] = \mu_f, \quad \text{and} \quad \mathbb{V}[f] = F. \tag{5.3}$$

The primary outputs of a statistical inference process are the parameters μ_f and F mentioned in (5.3), and other outputs one might be interested in include estimates of the daily realizations $\hat{f}_{t+1}$.

The simplest way of estimating historical daily realizations of $\hat{f}_{t+1}$ is by least-squares (ordinary or weighted, as appropriate), viewing the defining model equation (5.1) as a regression problem. We have not yet specified the distributional family that the ϵ_{t+1} come from, just the first two moments. Let's for now suppose it's normal; then (5.1) states:

$$R_{t+1} \,|\, f_{t+1} \sim N(X_t f_{t+1}, D)$$

We now drop the time subscript, and let y denote one particular realization of R_{t+1}. The likelihood function is

$$\frac{1}{\sqrt{(2\pi)^k |D|}} \exp\left(-\frac{1}{2} (y - Xf)^{\mathrm{T}} D^{-1} (y - Xf) \right),$$

The maximum-likelihood estimator of f_{t+1} comes from maximizing the likelihood function. One can equivalently maximize the log-likelihood since the log function is monotone increasing. The log-likelihood is usually easier to work with since many likelihood functions involve products and exponentials. In this case, we solve the following problem:

$$\max_f \big[-(y - Xf)^{\mathrm{T}} D^{-1} (y - Xf) \big] \tag{5.4}$$

Assuming that $X'D^{-1}X$ is stably invertible, the solution to (5.4) is

$$\hat{f} = (X'D^{-1}X)^{-1} X'D^{-1}y \tag{5.5}$$

Eq. (5.5) is one you'll want to memorize. Note that an overall scalar in D would cancel, so for example if $D = \sigma^2 I$ for some $\sigma > 0$ then $\hat{f}$ doesn't depend on σ.

Statistically estimated realizations $\hat{f}$ are called *factor returns* by practitioners (Menchero, Morozov, and Shepard, 2008). This nomenclature arises because they can be viewed as returns on certain portfolios, called *factor portfolios*. If we take D proportional to the identity matrix for simplicity, and assume that $X'X$ is invertible, then the factor returns are

$$\hat{f} = X^+ y, \qquad \text{where} \quad X^+ = (X'X)^{-1} X'. \tag{5.6}$$

Note that X^+ has dimensions $p \times n$, like the transpose. Since $y \in \mathbb{R}^n$ is a cross-sectional vector of returns, the j-th factor return $\hat{f}_j$ is the return on a long-short portfolio whose holdings are given by the j-th row of X^+.

The j-th row of X^+, when viewed as a portfolio, has rather special properties. For example, it has unit exposure to the j-th factor, and 0 exposure to all other factors as is proven by the following identity:

$$X^+X = (X'X)^{-1}X'X = I.$$

This is actually quite important for building intuition about the factor returns and factor portfolios. For example, if one of the columns of X is CAPM beta, then the other factor portfolios are β-neutral. If X contains an industry classification, then the other factors are industry-neutral, etc.

5.3. **Basic portfolio risk calculations.** The model (5.1), (5.2) and (5.3) entails associated reductions of the first and second moments of the asset returns:

$$\mathbb{E}[R] = X\mu_f, \quad \text{and} \quad \Sigma := \mathbb{V}[R] = D + XFX' \tag{5.7}$$

where X' denotes the transpose. Eq. (5.7) is quite useful for portfolio construction and for analyzing existing portfolios. For example, it says that

$$h'\Sigma h = h'Dh + h'XFX'h$$

which expresses the portfolio's variance in terms of the *idiosyncratic variance* $h'Dh$ and a second term computable from only the exposure vector.

Almost anything can now be reduced to a simpler form using the factor structure. For example, we know from an earlier lecture that the CAPM beta of an asset is

$$\beta_i = \frac{\text{cov}(r_i, r_M)}{\text{var}(r_M)}$$

The model (5.1), (5.2) and (5.3) entails ways of computing both the numerator and denominator in this expression:

$$\begin{aligned} \text{cov}(r_i, r_M) &= \text{cov}(x_i \cdot f, x_M \cdot f) = x_i' F x_M \\ \text{var}(r_M) &\approx x_M' F x_M \end{aligned}$$

where x_i is the i-th row of X, which is to say the exposure vector of the i-th security, and where x_M is the exposure vector of the market.

What do we mean by "the exposure vector of the market"? In practice, one could estimate x_M by choosing a basket, such as the S&P 500, as a proxy for the market portfolio, and taking $x_M = h_M'X$ where h_M is the holdings vector of the basket.

5.4. **An Empirical Example.** In this section, we present a detailed empirical example. We study the United States equity market over the period 1992-2015 through the lens of an APT model (5.1)–(5.3).

For each day t in our sample, we take $n = 2000$ and select the top n stocks in the US market, sorted by market capitalization. We chose this value because stocks falling below the top 2000 by market cap tend to be illiquid and have wide spreads, making them difficult to trade for institutional investors. We restrict the study to common stock, hence excluding closed-end funds, REITs, ETFs, unit trusts, depository receipts, warrants, etc. Our only data sources for this study were CRSP and IBES which we access via the Wharton Research Data Service.

We construct our model to contain five of the most commonly-studied and well-known sources of systematic risk: market beta, size, value, momentum, and volatility, as well as a classification of the stocks into industries. For further discussions of the five risk premia mentioned here, see (Fama and French, 1993; Connor, Goldberg, and Korajczyk, 2010; Menchero, Morozov, and Shepard, 2008; Asness, Moskowitz, and Pedersen, 2013).

There are likely other sources of systematic risk which could be considered in a more complete model. Our goal is only to illustrate the techniques developed earlier in the lecture, and this model will do nicely for that purpose. With the exception of the volatility premium, our model is intentionally very similar to that of Fama and French (1992), which is one of the most-cited papers in finance. The classic paper on the volatility premium (Ang et al., 2006) was not until 2006, but all other factors had been discussed in the academic literature prior to the beginning of our sample period.

In this example, the number of factors is about 75 due to the 5 risk premia and about 70 industries. Our industry classification is based on the "major group" of the Standard Industrial Classification (SIC) system. The model is

$$\boldsymbol{r}_{t+1} = \mathbf{X}_t \boldsymbol{f}_{t+1} + \epsilon_{t+1}, \quad \mathbb{E}[\epsilon] = 0, \quad \mathbb{V}[\epsilon] = D$$

where $\boldsymbol{r}_{t+1} \in \mathbb{R}^n$ denotes a cross-section of asset returns over the interval $[t, t+1]$, and $\mathbf{X}_t$ is calculated using data knowable before day t. Specifically, $\boldsymbol{r}_{t+1}$ are close-to-close total returns, and t will denote a day, or when more precision is required, the exact time of the US equity market close on day t.

The first 5 columns of $\mathbf{X}_t$ represent *exposures* to the five risk premia mentioned above. With the exception of market beta, which we do not transform in any way, the exposures are calculated as a "raw" value which is then "gaussianized" in the cross section. Gaussianization refers to a robustification procedure for extremely fat-tailed data by which the data is converted to ranks and passed through the

inverse CDF of a normal. The latter preserves the order, but reshapes the data to appear normally distributed.

- **Market beta**: each asset's daily excess return time series is winsorized and regressed against the S&P 500 excess return time series, over a trailing two year window with an intercept. The beta is the slope coefficient in this regression. The results are further improved using the Vasicek (1973) Bayesian adjustment in the cross section.
- **Size:** market capitalization.

- **Volatility:** as in the market beta calculation, each asset's daily excess return time series is winsorized and regressed against the S&P 500 excess return time series, over a trailing two year window. The mean-square error (MSE) from this regression is the raw exposure.
- **Momentum:** trailing compound returns over the last 12 months with the last 1 month excluded. For this and the next, see Asness, Moskowitz, and Pedersen (2013) and references therein.
- **Value:** the raw exposure is e_t/p_t where e_t represents the sum of the trailing 4 quarters' earnings-per-share (EPS), adjusted for any splits which occurred between the earnings announcement date and date t, and p_t is the close price on day t.

We define a stock's exposure to an industry as 1 if the stock is classified into that industry by SIC, and zero otherwise. Hence the remaining 70 columns of $\mathbf{X}_t$ are populated by 1s and 0s, with one column per industry.

Next we calculate the OLS factor returns

$$\hat{\boldsymbol{f}}_{t+1} = (\mathbf{X}_t'\mathbf{X}_t)^{-1}\mathbf{X}_t'\boldsymbol{r}_{t+1}. \tag{5.8}$$

Since our example concerns industry-neutral and market-neutral portfolios, we will not ascribe any persistent risk premia to the market-beta factor, nor to any of the industry factors. The cumulative factor returns of the other risk premia is shown in Fig. 5.1.

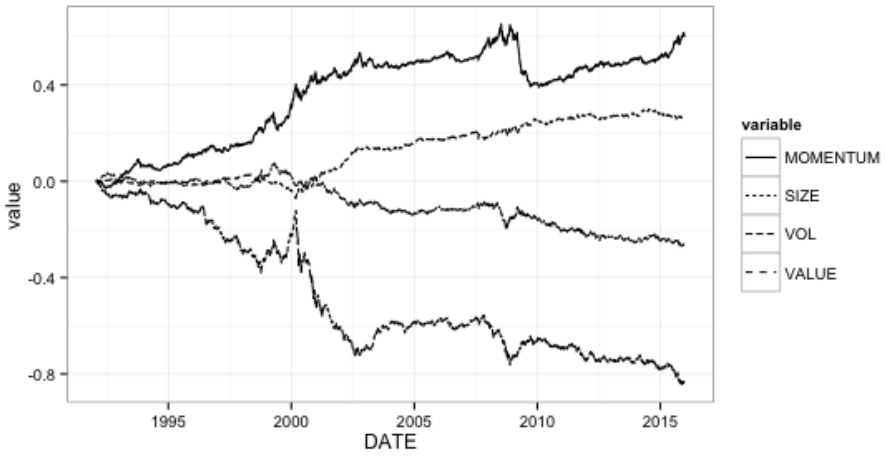

FIGURE 5.1. Cumulative factor returns to risk premia $\hat{\boldsymbol{f}}_{t+1}$ as computed by (5.8). The two positive-drift risk premia are momentum and value; the two negative-drift risk premia are size and volatility.

Fig. 5.1 corresponds well with our intuition and established consensus from the academic literature: the two positive risk premia are momentum and value; the two negative risk premia are size and volatility. For example, a negative premium to the size factor indicates that low market capitalization stocks tend to outperform high market capitalization stocks after controlling for market beta, industry, and other systematic sources of risk.

A similar study produced by Barra shows returns to a similar set of factors in a manner analogous to Fig. 5.1, over a much earlier time period which ends around the time my study begins. We show their results below as Fig. 5.2, Fig. 5.3, Fig. 5.4.

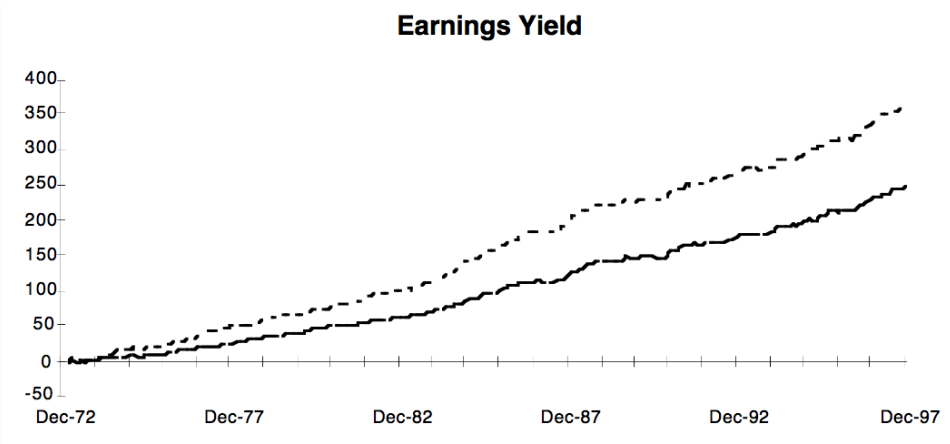

FIGURE 5.2. Cumulative factor returns to the Earnings Yield factor as computed by Barra.

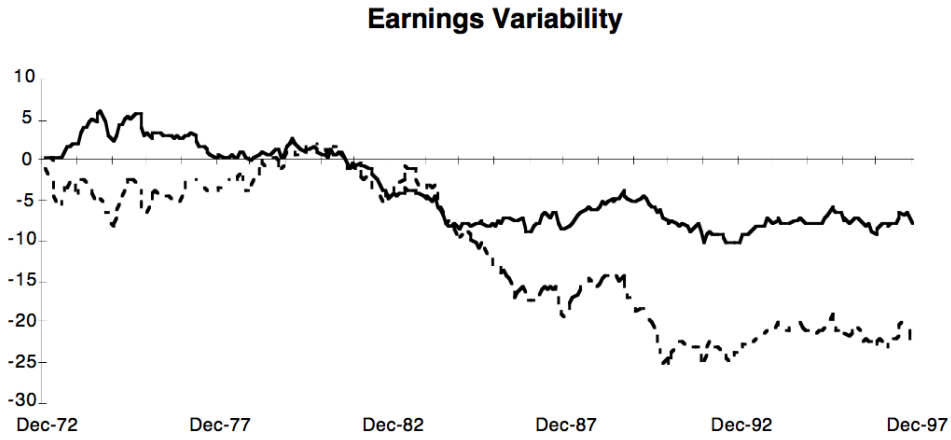

FIGURE 5.3. Cumulative factor returns to the Earnings Variability factor as computed by Barra.

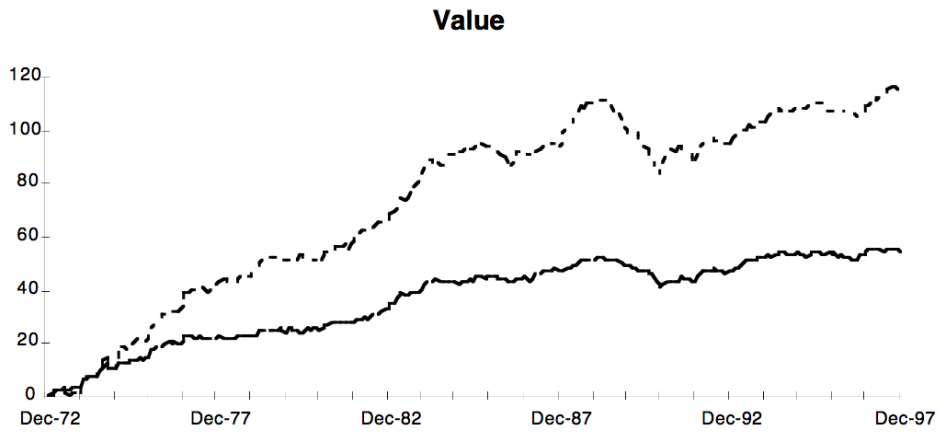

FIGURE 5.4. Cumulative factor returns to the Value factor as computed by Barra.

They have two different lines for each factor, corresponding to two different ways to run the regression. The difference is: when analyzing style factor i, do we include style factors $j \neq i$ in the model? (The industry factors are always included). The "fully multivariate" approach corresponds to the solid line. The fact that these are very different is food for thought. Does the very existence of a "value premium" depend on how it is measured?

5.5. **Identifiability.**

Definition 5.3. The model (5.1) is said to be *identifiable* if any of the following equivalent conditions hold:

(a) $\text{rank}(X) = p$, ie. X is full rank,
(b) $X'X$ is invertible, where X' denotes the transpose
(c) The function $\ell(f) = \|R - Xf\|^2$ has a unique minimizer.
(d) The ordinary least-squares (OLS) estimator of f exists.

We call a model *barely identifiable* (or approximately collinear) if conditions (b) or (c) are close to being violated; e.g. if $X'X$ has a very small eigenvalue, or equivalently, if $\ell(f)$ has a direction of near-zero curvature, or the Hessian of $\ell(f)$ is nearly degenerate.

Here is an example of what can happen in barely identifiable models, which my old boss used to call "regression nastiness."

Example 5.1. Consider estimating the coefficients from data which was actually generated from the model $y = 3+x_1+x_2$, and observe that the coefficient estimates for x_1 and x_2 are huge in magnitude relative to their true values, and of opposite sign to each other.

```
N <- 20;
x1 <- rnorm(N);
x2 <- rnorm(N, mean = x1, sd = .01);
y <- rnorm(N, mean = 3 + x1 + x2)
lm(y ~ x1 + x2)$coef
(Intercept)          x1          x2
   2.994764  -35.833870   37.642712
```

Identifiability might seem an important condition to require when building a factor model, but, in practice, unidentifiable and barely identifiable models arise naturally and often.

Example 5.2. Consider a unified model for the European equity market. Significant drivers of asset return covariance include market beta, industry membership, country membership, and others. Define

$$N_{i,j} = \begin{cases} 1, & \text{stock } i \text{ is a member of industry } j \\ 0 & \text{otherwise} \end{cases}$$

with C defined similarly for countries. Let B denote an $n \times 1$ column vector of market betas. The full design matrix is $X = \begin{bmatrix} B & C & N \end{bmatrix}$ and is of course not identifiable.

An exact colinearity, such as in Example 5.2, can be rectified by choosing a basis for the column space of X and replacing X with a (thinner) matrix having the elements of this basis as columns. This is less than ideal for two reasons: it loses the economic meanings of the factors, and doesn't easily generalize to the case of approximate colinearity, ie. barely identifiable models.

Example 5.3. Suppose we augment model (5.2) with two closely-related alpha forecasts, which we call α_1 and α_2 (such as earnings yield with two different types of earnings). Collect these into an $n \times 2$ matrix $A = \begin{bmatrix} \alpha_1 & \alpha_2 \end{bmatrix}$. The augmented design matrix is then

$$X = \begin{bmatrix} A & B & C & N \end{bmatrix}$$

We now have an approximate and an exact colinearity in the same model!

Identifiability is not necessary for the model (5.1) to be a correct description of the world. Referring to Example 5.2, there can, of course exist latent, unobservable stochastic processes f_j corresponding to industries and countries which drive returns according to eq. (5.1). However, lack of identifiability complicates estimation of $\hat{f}$ and by extension, μ_f and F.

If the model is identifiable, then reasonable estimates for the factor returns are

$$\hat{f}_{OLS} = (X'X)^{-1}X'R,$$

and μ_f can be estimated by the time-series mean of $\hat{f}_{OLS}$. In the unidentifiable case, $\hat{f}_{OLS}$ doesn't exist, and in the barely identifiable case, $\hat{f}_{OLS}$ misleadingly contains large opposing coefficients for strongly-correlated factor pairs.

Several well-known methods exist for obtaining estimates of $\hat{f}$ in the unidentifiable case. Intuitively, when there are multiple solutions to the first-order condition, some constraint or prior must be introduced to prefer one solution over another.

One may remove exact colinearities by imposing $p - \operatorname{rank}(X)$ linear constraints on the coefficients. This *restricted least squares* method (Greene and Seaks, 1991) is simple and classical, but ultimately incomplete. It would handle Example 5.2 but not Example 5.3, and the constraints must be decided arbitrarily.

A more general class of inference procedures are provided by Bayesian regression, including the popular ridge, lasso, and elastic-net estimators as special cases; see Zou and Hastie (2005). Such procedures have the added benefit of providing an explicit model complexity parameter to be used in cross-validation; see Friedman, Hastie, and Tibshirani (2001) for details.

The simplest of the Bayesian regression estimators is the *ridge* estimator, defined by

$$\hat{f}_\lambda = \arg\min_f \left[\|R - Xf\|^2 + \lambda\|f\|^2\right] = (X'X + \lambda I)^{-1}X'R, \quad \lambda > 0. \tag{5.9}$$

For any $\lambda > 0$ and any real matrix X, $\hat{f}_\lambda$ exists. Moreover, the limit

$$\lim_{\lambda \to 0+} \hat{f}_\lambda = X^+R \tag{5.10}$$

also exists, where X^+ is the Moore-Penrose pseudoinverse of X.

5.6. **The singular value decomposition and the pseudoinverse.** For $A \in \mathrm{M}(n, p; \mathbb{R})$ there exists a factorization of the form

$$A = USV'$$

where U is an $n \times n$ orthogonal matrix, S is an $n \times p$ diagonal matrix with non-negative real numbers on the diagonal, and V is $p \times p$ orthogonal. Such a factorization is called a *singular value decomposition* of A, and the diagonal entries of S are known as the *singular values* of A. The convention is to list the singular values in descending order, and with this convention, the diagonal matrix S is uniquely determined by A (though the matrices U and V are not).

The SVD arises from finding an orthogonal basis $\{\mathbf{v}_i\}$ for the row space that gets transformed into an orthogonal basis for the column space:

$$A\mathbf{v}_i = \sigma_i \mathbf{u}_i.$$

These bases become the columns of U and V. It follows that

$$A = \sum_i \sigma_i \mathbf{u}_i \otimes \mathbf{v}^*{}_i$$

where $\mathbf{u}_i$ and $\mathbf{v}^*{}_i$ are the i-th columns of the corresponding SVD matrices and σ_i are the ordered singular values, and $\otimes$ denotes the outer product.

In applications a full basis of the null-space of A is usually not required. Instead, it is often sufficient (as well as faster, and more economical for storage) to compute a reduced version of the SVD. Suppose $p \ll n$; then the *thin SVD* is the decomposition

$$A = U_p S_p V'$$

where U_p is defined as the first p columns of U, S_p is the $p \times p$ matrix with the singular values on the diagonal (ie. the upper $p \times p$ block of S), and V is as before.

Various truncated SVDs can be considered in which only the t column vectors of U and t row vectors of V' corresponding to the t largest singular values are calculated. The resulting decomposition is written

$$\tilde{A}_t = U_t S_t V_t',$$

and $\tilde{A}_t$ is the closest approximation to A that can be achieved by a matrix of rank t. In particular, for $t = r = \text{rank}(A)$, one has the compact (or "economical") SVD, which is still an exact representation of the matrix A.

Not every matrix has an inverse, but every matrix does have a *pseudoinverse*. The pseudoinverse was discovered independently by Moore (1920) and Penrose (1955), and there is a nice monograph by Albert (1972) for background. The Moore-Penrose pseudoinverse X^+ is defined so that $X^+ y$ is the minimum-norm vector among all minimizers of the least-squares function $\ell(f) = \|y - Xf\|$. If there is only one minimizer, ie. the identifiable case, then X^+ is given by $(X'X)^{-1}X'$, but X^+ is well-defined by the minimum norm criterion for any matrix X.

If $A = USV'$ is the SVD of A, the Moore-Penrose pseudoinverse is given by

$$A^+ = VS^+U' \tag{5.11}$$

where S^+ is formed by replacing every non-zero diagonal entry by its reciprocal and transposing the resulting matrix. The thin or compact SVD can be used in (5.11) for efficiency.

Using the SVD in this manner provides an automatic way to ensure numerical stability. Computing S^+ is nothing more than computing the multiplicative inverses

of a sequence of real numbers. One should treat those numbers as zero if they are within floating-point precision! More aggressive regularization can be obtained by further increasing this threshold away from the floating-point epsilon.

As mentioned above, in the special case when $X'X$ is invertible, the pseudoinverse is $X^+ = (X'X)^{-1}X'$. Furthermore, b^* minimizes $\|y - Xb\|^2$ if and only if

$$b^* = X^+ y + (I - X^+ X)v \quad \text{for some } v.$$

Thus we can characterize all least-squares minimizers; the minimum-norm element corresponds to $v = 0$. Note however that the minimum-norm solution will not, in general, correspond to stepwise elimination of columns which are linear combinations of the others.

Now consider a linear regression problem in standard form, $y = Xf + \epsilon$. We can write $X = USV'$ and then from (5.11), the minimum-norm least-squares solution is given by

$$\hat{f}_{\text{ls}} = VS^+U'y.$$

Similarly, the ridge solution with ridge parameter λ is given by

$$\hat{f}_{\text{ridge}} = V\frac{S}{S^2 + \lambda I}U'y.$$

When using the *thin* SVD, it is no longer the case that U_p is unitary/orthogonal. In the thin case, $U_p'U_p = I_p$ but $U_pU_p' \neq I_n$, and in fact the latter is related to the predicted values in the multiple regression. Specifically,

$$\hat{y} := X\hat{f} = U_pU_p'y.$$

The last few equations show that, once one has computed the singular value decomposition of X, a long list of useful calculations related to linear regression can be immediately calculated from the SVD, including the coefficients for OLS or ridge regression, the Moore-Penrose pseudoinverse and the predicted (or fitted) response $\hat{y}$. All ridge solutions for any ridge parameter λ can be computed from the same SVD.

5.7. **Arbitrage Pricing.** Ross (1976) had the insight that a strict factor model is a useful device for separating common risk from diversifiable risk in asset returns. In addition to imposing a strict factor model, Ross assumes that the number of assets n is large and imposes an upper bound on the individual elements of D. This makes economic sense; if a company's σ_i were arbitrarily large, it could quickly take over the economy or go out of business!

For a portfolio with dollar holdings $\boldsymbol{h} \in \mathbb{R}^n$, define the p-norm as

$$\|\boldsymbol{h}\|_p := \left(\sum_{i=1}^{n} |h_i|^p \right)^{1/p}.$$

The L^1 norm is also called *gross market value (GMV)*. Define the *weight vector* to be $\boldsymbol{w} := \boldsymbol{h}/\|\boldsymbol{h}\|_1$. There are various measures of diversification in the literature, but the simplest one is

$$\boldsymbol{w}'\boldsymbol{w} = \|\boldsymbol{w}\|_2^2. \tag{5.12}$$

Note that (5.12) is bounded between $1/n$ and 1.

Define a *well-diversified* portfolio as one in which $\boldsymbol{w}'\boldsymbol{w} \to 0$ as $n \to \infty$. Ross shows that in a strict factor model, well-diversified portfolios have approximately zero asset-specific risk:

$$\boldsymbol{w}' D \boldsymbol{w} \leq \boldsymbol{w}'\boldsymbol{w} \max\{\sigma_i^2 : i = 1 \ldots, n\}$$

Ross (1976) gives a restriction on expected returns analogous to that of the CAPM. In the CAPM, there was only one risky portfolio, and all expected returns were found to be linearly related to the expected return on that portfolio. Arbitrage Pricing Theory (APT) gives a similar relation, but with multiple distinct sources of expected return.

Suppose for simplicity that the actual market returns (excess above the risk-free rate) obey a noiseless factor model, so there is some matrix X such that $R = Xf$, as in (5.1) with $D = 0$. We assume without loss of generality that X contains a column of 1's, or a collection of columns that sums to the vector $\mathbf{1} = (1, 1, \ldots, 1)$. A portfolio with holdings vector $\boldsymbol{h}$ is said to be *self-financing* if $\boldsymbol{h}'\mathbf{1} = 0$, or in other words the total dollar amount of the long positions equals (with opposite sign) the dollar amount of the short positions. Self-financing means, equivalently, that one can create the portfolio with no net infusion of cash.

Now suppose all agents trading in this market have the same expected return vector, $\boldsymbol{\mu} := \mathbb{E}[R]$. Consider the projection of expected returns on the column span of matrix X:

$$\boldsymbol{\mu} := X\hat{\boldsymbol{\pi}} + \boldsymbol{\eta}$$

where $\hat{\boldsymbol{\pi}} = X^+\boldsymbol{\mu}$, and $\boldsymbol{\eta}$ is the residual of the projection operation, so $X'\boldsymbol{\eta} = 0$. Since $\mathbf{1} \in \mathbb{R}^n$ is an element of the column span of X, the condition $X'\boldsymbol{\eta} = 0$ implies that $\mathbf{1}'\boldsymbol{\eta} = 0$, ie. that $\boldsymbol{\eta}$ is self-financing. The portfolio with holdings $\boldsymbol{\eta}$ has no factor exposure, and no idiosyncratic risk either (since $D = 0$), hence it is riskless.

The expected return of portfolio $\boldsymbol{\eta}$ is

$$\boldsymbol{\mu}'\boldsymbol{\eta} = \|\boldsymbol{\eta}\|^2$$

So either $\boldsymbol{\eta} = 0$, or there is a portfolio with zero risk and non-zero expected excess return. This would represent an arbitrage opportunity, so in the absence of arbitrage, expected returns are a linear combination of the columns of X.

If we allow $D \neq 0$, hence remove the noiseless assumption, then a similar argument can be constructed to show that the absence of statistical arbitrage implies that all well-diversified portfolios have mean returns that are linear in the factor exposures.

Problem 5.1. Use the Sherman-Morrison-Woodbury matrix inversion lemma to derive a simple expression for the inverse of the covariance matrix in an APT model. In other words, derive an expression for Σ^{-1} where

$$\Sigma = \mathbb{V}[R] = X'FX + D$$

where D is diagonal and X is $n \times p$ and as usual we assume $p \ll n$. In your answer, any matrices being inverted should be either diagonal or $p \times p$.

Problem 5.2. Show that, for any $n \times p$ real matrix X (not necessarily full rank) and n-vector Y, the following are equal:

(1) $\lim_{\delta \to 0+} (X'X + \delta I)^{-1} X'Y$
(2) The smallest-norm element of $\operatorname{argmin}_b \|Y - Xb\|$.
(3) $VS^{+}U'Y$ where $X = USV'$ is the SVD of X.

References

Albert, Arthur (1972). *Regression and the Moore-Penrose pseudoinverse*. Academic Press.

Ang, Andrew et al. (2006). "The cross-section of volatility and expected returns". In: *The Journal of Finance* 61.1, pp. 259–299.

Asness, Clifford S, Tobias J Moskowitz, and Lasse Heje Pedersen (2013). "Value and momentum everywhere". In: *The Journal of Finance* 68.3, pp. 929–985.

Connor, Gregory, Lisa R Goldberg, and Robert A Korajczyk (2010). *Portfolio risk analysis*. Princeton University Press.

Fabozzi, Frank J, Sergio M Focardi, and Petter N Kolm (2010). *Quantitative equity investing: Techniques and Strategies*. John Wiley & Sons.

Fama, Eugene F and Kenneth R French (1992). "The cross-section of expected stock returns". In: *the Journal of Finance* 47.2, pp. 427–465.

– (1993). "Common risk factors in the returns on stocks and bonds". In: *Journal of financial economics* 33.1, pp. 3–56.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin.

Greene, William H and Terry G Seaks (1991). "The restricted least squares estimator: a pedagogical note". In: *The Review of Economics and Statistics*, pp. 563–567.

Grinold, Richard C and Ronald N Kahn (2000). *Active portfolio management*. McGraw Hill New York, NY.

Menchero, Jose, Andrei Morozov, and Peter Shepard (2008). "The Barra Global Equity Model (GEM2)". In: *MSCI Barra Research Notes*, p. 53.

Moore, E. H. (1920). "On the reciprocal of the general algebraic matrix". In: *Bulletin of the American Mathematical Society* 26, pp. 394–395.

Penrose, Roger (1955). "A generalized inverse for matrices". In: *Mathematical proceedings of the Cambridge philosophical society* 51.3, pp. 406–413.

Roll, Richard and Stephen A Ross (1980). "An empirical investigation of the arbitrage pricing theory". In: *The Journal of Finance* 35.5, pp. 1073–1103.

Ross, Stephen A (1976). "The arbitrage theory of capital asset pricing". In: *Journal of economic theory* 13.3, pp. 341–360.

Vasicek, Oldrich A (1973). "A Note on Using Cross-sectional Information in Bayesian Estimation of Security Betas". In: *The Journal of Finance* 28.5, pp. 1233–1239.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320.