Research Program in Finance
WORKING PAPER SERIES

WORKING PAPER NO. 173
ESTIMATING PERVERSIVE ECONOMIC FACTORS WITH MISSING OBSERVATIONS

BY
GREGORY CONNOR
ROBERT A. KORAJCZYK

Research Program in Finance Working Papers are preliminary in nature, their purpose is to stimulate discussion and comment. Therefore, they should not be cited or quoted in any publication without the permission of the author. Single copies of a paper may be requested from the Institute of Business and Economic Research.

Institute of Business and Economic Research
University of California Berkeley
RESEARCH PROGRAM IN FINANCE AT THE
WALTER A. HAAS SCHOOL OF BUSINESS,
UNIVERSITY OF CALIFORNIA, BERKELEY

The Research Program in Finance in the Walter A. Haas School of Business at the University of California has as its purpose the conduct and encouragement of research in finance, investments, banking, securities markets, and financial institutions. The present reprint and working paper series were established in 1971 in conjunction with a grant from the Dean Witter Foundation.

INSTITUTE OF BUSINESS AND ECONOMIC RESEARCH
Carl Shapiro, Director

The Institute of Business and Economic Research is an organized research unit at the University of California, Berkeley, whose mission is to promote research by faculty and graduate students in the fields of business and economics. The Institute carries out its mission by organizing programs and activities that enrich the research environment, administering extramural research awards, publishing working papers, and making direct grants for research.
ESTIMATING PERVERSIVE ECONOMIC FACTORS

WITH MISSING OBSERVATIONS

Gregory Connor
School of Business Administration
University of California, Berkeley

and

Robert A. Korajczyk
Kellogg Graduate School of Management
Northwestern University

Working Paper 173

November 1986
Revised: April 1987
ESTIMATING PERVERSIVE FACTORS
WITH MISSING OBSERVATIONS

ABSTRACT

We suggest a technique for estimating pervasive economic factors which allows the use of all available security return data. The resulting factor estimates can be used in applications and tests of the Arbitrage Pricing Theory (APT). An obvious advantage of the technique is that more precise estimates of the factors are obtained while avoiding potential survivorship biases in factor construction. Empirically, the factor estimates using the entire data set outperform (in terms of asset pricing) estimates using only continuously traded assets.
A growing number of empirical studies of asset pricing rely on the use of large numbers of assets (cross-sections) to estimate the relevant parameters of the model under study. Shanken [16] provides an analysis of the asymptotic properties (as the number of assets becomes large) of standard cross-sectional regression tests. Scott [15] uses large cross-sections to estimate the marginal rate of substitution in an intertemporal asset pricing model. Many tests of linear multi-factor asset pricing models involve, either explicitly or implicitly, construction of portfolios which are highly (and as the number of assets approaches infinity, perfectly) correlated with the pervasive factors in the economy.¹ For a variety of reasons, it is common practice for empiricists to utilize returns on securities that are continuously traded over the relevant sample period.

In this paper, we suggest a basis portfolio construction technique which allows the use of virtually all available asset returns. The technique is an extension of the principal components based technique in Connor and Korajczyk [4]. The main advantage of the technique is that the dimensionality of the problem does not increase as the number of assets, n, increases. This allows us, under certain stationarity conditions, to utilize all available data.

Excluding assets with missing observations reduces the sample size and may induce various forms of selection bias. An indication of the potential increase in data points by using return series with missing observations can be obtained by considering some recent empirical studies. Lehmann and Modest [11] use returns on NYSE and AMEX stocks which have no missing daily data over five year subperiods. The average number of firms meeting this
requirement over their four subperiods is 1247 while the average number of firms with returns observed on a typical day is 2404. Connor and Korajczyk [5] use returns on NYSE and AMEX stocks which have no missing monthly data over five year subperiods. The average number of firms meeting this requirement over their four subperiods is 1672 while the average number of firms with returns observed in a typical month is 2383. Also, given that firms with missing observations are more likely to be small firms, current techniques (which restrict their analyses to continuously traded securities) may not be able to pick up small firm related factors due to insufficient number of observations. This might be one reason why the APT has been unable to totally explain small firm related anomalies (see [11] or [5]).

Our estimation procedure is presented in Section I. An empirical investigation of the properties of basis portfolios formed from data without missing observation versus those formed using the entire data set is in Section II. Our conclusions and suggestions for future extensions are given in section III.

I. Estimating Pervasive Factors With Missing Observations

Assume an exact multi-factor pricing relation such as the one implied by the equilibrium version of the APT in Connor [3]. That is,

\[ \bar{r}_{it} - \gamma_{0t} = b_i (\bar{r}_t + \bar{F}_t) + \bar{\varepsilon}_{it} \quad i = 1, 2, \ldots \]

\[ t = 1, 2, \ldots, T \]

where:

\[ \bar{r}_{it} \] - the return on asset i;

\[ \gamma_{0t} \] - the return on the riskless asset;

\[ b_i \] - the 1 x k vector of asset i's sensitivities (loadings) on the k factors:
\( \gamma_t \) - the \( k \times 1 \) vector of risk premia associated with the \( k \) factors;

\( \bar{\varepsilon}_t \) - the unexpected components of the factors, a \( k \times 1 \) vector, and;

\( \bar{\varepsilon}_{1t} \) - the idiosyncratic component of the return on asset 1.

We assume that returns follow an approximate factor structure. That is, \( \mathbb{E}(\bar{\varepsilon}) = 0 \); \( \mathbb{E}(\bar{\varepsilon}_1) = 0 \), for all \( i \); and \( \Sigma^2 \) has bounded eigenvalues as \( n \to \infty \), where \( \Sigma^2 \) is the covariance matrix of the idiosyncratic returns of the first \( n \) assets. (A more detailed discussion of the assumptions is given in [4].)

Assume that we observe returns on \( n \) risky assets and the riskless asset over \( T \) time periods. Define \( R^2 \) as the \( n \times T \) matrix of observed returns on the \( n \) assets in excess of the riskless rate; \( B^2 \) as the \( n \times k \) matrix of factor loadings; \( F \) as the \( k \times T \) matrix of realizations of \( \gamma + f \); and \( \varepsilon^2 \) as the \( n \times T \) matrix of realized idiosyncratic returns. Thus,

\[
R^2 = B^2 F + \varepsilon^2
\]  

The task at hand is to estimate the unobservable values of \( F \) given observations of \( R^2 \) and an assumed value of \( k \). We show, in [4] that the first \( k \) eigenvectors of the \( T \times T \) matrix \( \Omega^2 = (1/n) R^2 R^2 \) approach a transformation of \( F \) as \( n \) grows large. That is, if \( G^2 \) is the \( k \times T \) matrix of the first \( k \) eigenvectors of \( \Omega^2 \), then \( G^2 = L^2 F + \phi^2 \) where \( L^2 \) is a non-singular \( k \times k \) matrix and \( \text{plim} \phi^2 = 0 \), the null matrix.

The result relies on the fact that \( \Omega^2 = A^2 + Y^2 + Z^2 \) where

\[
A^2 = (1/n) F' B^2 F,
\]

\[
Y^2 = (1/n) (F' B^2 \varepsilon^2 + \varepsilon^2 B^2 F),
\]
\[ Z^n = (1/n) \epsilon^n, \epsilon^n \]

and that \( Y^n \) and \( Z^n \) approach the null matrix and \( \sigma^2 I_T \), respectively, where \( I_T \) is the \( T \times T \) identity matrix. A simple example which we have found useful in providing some intuition for this result is presented in the appendix.

Now let \( m (m > n) \) denote the total number of assets in the sample, including those which have some missing observations; \( R^m \) be the \( m \times T \) matrix of returns in which missing values are replaced by zeros; and \( n_{tr} \) be the number of assets with returns in both periods \( t \) and \( r \), \( (n \leq n_{tr} \leq m) \). Now define \( \Omega^m_{tr} = (R^{m'} R^m)_{tr}/n_{tr} \).

Given the assumption that the non-missing returns of those firms with missing observations still follow the process (1), (i.e., no nontrading or spillover effects) and the cross-sectional distribution of factor loadings and idiosyncratic variances are the same, the elements of \( A^n, Y^n, \) and \( Z^n \), which make up \( \Omega^m \) approach the same limits \( \lim(n_{tr}) \rightarrow \infty \) as \( A^n, Y^n, \) and \( Z^n \) do \( \lim(n \rightarrow \infty) \) since they are merely averages over a greater number of observations than the corresponding elements of \( A^n, Y^n, \) and \( Z^n \). Thus, \( G^m \) (the eigenvectors of \( \Omega^m \)) = \( L^m F + \phi^m \) with \( \lim \phi^m = 0 \).

If we could observe \( A^n \) or \( A^n + \sigma^2 I \) (in (3)) directly then the eigenvectors would equal some transformation of the factors, say \( L^n F \).

Define \( C^n = (Y^n + Z^n - \sigma^2 I_T) \) and \( C^m \) similarly. As \( n \) and \( m \) increase, both \( C^n \) and \( C^m \) converge to the null matrix. We can express the size of the difference between the estimated and "true" basis portfolios as a function of the norm of the respective \( C \) matrices. Our estimate of the \( i^{th} \) factor is \( G^n_i \), the \( i^{th} \) eigenvalue \( \Omega_i \) (corresponding to the \( i^{th} \) largest eigenvalue \( \lambda_i \)).

Let \( G^n_i \) denote the \( i^{th} \) eigenvector one would obtain from \( A^n + \sigma^2 I \). From
Theorem 4.2 of Stewart [18, pp. 295-6] we can obtain a bound on the magnitude of the difference between $G^*_i$ and $G_i$ in terms of the norm of $C^n$.

$$
\|G^*_i - G_i\| \leq \|C^n\|/\delta_i + O(\|C^n\|^2)
$$

(4)

where $\delta_i = \min(|\lambda_i - \lambda_j|: j \neq i)$ and $O(\cdot)$ represents order of magnitude. The level of approximation error depends on sample size indirectly through $\|C^n\|$. The approximate factor structure assumed in [4] is sufficient to guarantee that

$$
\lim_{n \to \infty} \|C^n\| = 0.
$$

However, $C^n$ need not approach the null matrix monotonically as $n$ increases due to cross-sectional differences in idiosyncratic variances and cross-sectional residual correlation. Since the elements of $C^n$ and $C^m$ are simple averages, we cannot guarantee that $\|C^m\|$, which uses the entire data set is expected to be smaller than $\|C^n\|$ which uses continuously traded assets only. Optimal weighting would require knowledge of $\nu_n$ and $\nu_m$. However, if the assets with missing observations are not "too" different from continuously traded assets, then one would generally expect $\|C^m\|$ to be smaller than $\|C^n\|$.

From (4) we also see that our approximation error in estimating the $i^{th}$ factor is a function of the minimum difference between the $i^{th}$ eigenvalue and all other eigenvalues of $C^n + \sigma^2 I$. A given reduction in $\|C\|$ will have a greater influence on the bound in (4) when $\min(|\lambda_i - \lambda_j|)$ is small. Empirically, the distance between adjacent eigenvalues tends to decrease as the number of factors extracted increases. This implies that we would expect a greater improvement in the approximation error, for a given improvement in $\|C\|$,
the larger the value of $i$.

The important assumption above is that the non-missing returns of assets follow the process in (1). Some examples of where this may or may not be true are in order. A case in which we would not expect there to be much problem with the assumption is one where the missing observations are caused by listing or delisting of the firm (which has non-missing observations while listed) and the data are monthly or weekly. A case where the assumption may be violated is where the missing observations are caused by suspension of trading. The manner in which returns are calculated in this case (by our source of returns data, the Center for Research in Security Prices (CRSP) at the University of Chicago) is that the "returns" for the missing periods are included in the first non-missing period return. That is, if we observe $P_{it}$ and $P_{it+r}$ only then $R_{it+1}$, $R_{it+2}$, ..., $R_{it+r-1}$ will be missing and $R_{it+r} = \frac{(P_{it+r}/P_{it}) - 1}{\text{plus any dividends paid in the interim}}$. In this case the reported return in period $t + r$ includes returns over a number of periods. Hence, $A^m$, $Y^m$ and $Z^m$ will not converge to the appropriate matrixes. This would not necessarily be a problem, asymptotically, if the number of firms with such spillovers is of smaller order than $\min(n_{cr})$. This particular type of non-synchronous trading problem can be handled by eliminating the first non-missing observation (i.e. $R_{it+r}$ in the example above). In our empirical work below we use this elimination scheme. This results in a reduction in observations of two-tenths of one percent (a typical month has 2378 firms with observed returns after eliminating the first non-missing return compared to 2383 for the full sample).

An additional case in which the simple approach outlined above is not appropriate is when there is a significant amount of non-synchronous trading (e.g., using daily data). In this case one might think of the observed return
on day \( t \) as a combination of current and past "true" returns (see Cohen, et al. [2]). This is a violation of the process in (1) assumed for the observed returns. This problem with non-synchronous trading is not peculiar to our approach. Shanken [17] shows that standard factor analytic approaches are very sensitive to non-synchronicity for daily data. We assume that the monthly observation interval is such that non-synchronous trading can be ignored.

II. Empirical Comparison of Alternative Basis Portfolios

In this section we use monthly data on NYSE and AMEX firms to compare estimates of the pervasive factors obtained using all available data versus estimates obtained using only those firms which have no missing monthly observations over five-year subperiods. The subperiods analyzed are 1964-68, 1969-73, 1974-78, and 1979-83. The numbers of firms available under the non-missing data restriction and without any restrictions are presented in Table 1. The typical month has 2378 firms with observable returns (this number excludes any occurrence of a first non-missing return). The average number of firms with non-missing data over the subperiods is 1672. Thus, using the entire data set provides an average increase in assets of 42%.

We present two sets of diagnostics on the relative performance of the two sets of factor estimates, or basis portfolios. The first is an analysis of the canonical correlations between the factor estimates. The second set of diagnostics is a comparison of the mispricing of assets relative to these alternative basis portfolios.

A. Canonical Correlations

Let \( G^H_t \) be the \( k \times 1 \) vector of period \( t \) factors extracted from the continuously traded sample and let \( G^M_t \) be the \( k \times 1 \) vector of period \( t \) factors
extracted from the full sample including assets with missing data. We cannot simply calculate the correlation between the corresponding elements of $G^*_t$ and $G^m_t$ because of a rotational indeterminacy problem common to factor analytic techniques. That is, the transformations $L^R$ and $L^M$ need not be the same. For this reason we employ canonical correlation analysis which calculates the maximal correlation between linear combinations of the two sets of variates. The $i^{th}$ canonical correlation is the correlation between the linear combinations $c_i' G^*_t$ and $m_i' G^m_t$, where $c_i$ and $m_i$ are $k \times 1$ vectors that solve

$$\max \ \text{SCORR}(c_i' G^*_t, m_i' G^m_t)$$

Subject to: \text{SCORR}(c_i' G^*_t, c_j' G^*_t) = 0

\text{SCORR}(m_i' G^m_t, m_j' G^m_t) = 0

$$1 \leq j < i$$

and SCORR is the sample correlation. The results of the canonical correlation analysis are given in Table 2. We present the first ten canonical correlations between the first ten factors extracted from the continuously traded data set and the first ten factors extracted from the entire data set.

All of the canonical correlations are significantly different from zero, say at the 5% level, and all are in the range of 0.8 to 1.0 with the exception of the tenth canonical correlation for three of the four subperiods. Thus, the two factor estimation methods (with and without noncontinuous returns) give highly correlated estimates. This may be due to the large number of assets
included in the continuously traded sample.

B. Tests of Exact Multifactor Pricing

High correlations between the two sets of basis portfolios are not sufficient to guarantee that the pricing implications will be the same or similar across the basis portfolios. Differences in factor means will not influence correlations but will, in general, lead to different inferences regarding pricing restrictions. For a given set of assets we can estimate the level of mispricing as the intercept in a regression of their excess return vector, \( R \), on the vector of basis portfolio excess returns, \( F \).

\[
R_t = a + BF_t + \varepsilon_t
\]  

(6)

Exact multifactor pricing implies that \( a = 0 \) (see (2)). Equivalently \( a = 0 \) implies that some linear combination of the basis portfolios is on the minimum variance frontier formed by \( R \) and \( F \), (see Huberman and Kandel [7]).

Since our basis portfolios represent the true factors plus the noise term, \( \phi \) (or equivalently, combinations of our basis portfolios only intersect the minimum variance frontier in the limit as \( n \to \infty \)) ordinary least squares (OLS) estimates of \( a \) are biased away from zero when the true value of \( a \) is zero. This bias remains even if the number of time periods, \( T \), is allowed to grow to infinity. The size of the bias is directly related to the magnitude of our estimation error, \( \phi \). If we let \( Q_n^\phi \) represent the \( k \times k \) covariance matrix of basis portfolio estimation errors, \( \gamma \) represent the vector of expected returns on the true basis portfolios, and \( b_i \) represent the factor loadings for asset \( i \) (see (1)), then the asymptotic bias (now as \( T \to \infty \)) of the estimated mispricing parameter, \( a_i \) is
\[ \text{plim}(a_1 \hat{a}_1 - a_1) = -(1 - \gamma) L^{n'} L \gamma^{n'} L^n Q^n \phi b_i. \]
\[ T \to \infty \]

This bias goes to zero as \( Q^n \) approaches the null matrix. But \( Q^n \to 0 \) as the number of assets used to construct the basis portfolios increases. Thus, if the hypothesis of exact multifactor pricing is true and using all assets improves our estimates of the factors, we would expect our estimates of mispricing to be closer to zero when \( G^m \), rather than \( G^n \), is used in the regression (6). Thus we would reject the hypothesis that \( a = 0 \) less often when \( G^m \) is the regressor.

We estimate the mispricing of two sets of portfolios formed on the basis of instruments which have, in past empirical research, been associated with anomalous asset return behavior. These instruments are firm size (measured by market value of the firm's common stock) and dividend yield. (See Banz [1] and Keim [9, 10]). For each five year subperiod we form ten "size" portfolios and ten "dividend yield" portfolios. Our measure of size is the market value of common stock at the end of the month preceding the subperiod (e.g., for the 1964-68 subperiod size is calculated at the end of December 1963). Dividend yield is defined as the average monthly dividend yield over the year preceding the subperiod

\[ \bar{d}_i = \frac{\sum_{t=T-13}^{T-1} \frac{D_{it}}{P_{it-1}}}{13} \]

where,

\[ \bar{d}_i \quad \text{our measure of dividend yield for firm i;} \]
\[ T \quad \text{the first month of the subperiod;} \]
\[ D_{it} = \text{the dividend paid by firm } i \text{ in period } t; \]
\[ P_{it} = \text{the price of asset } i \text{ in period } t. \]

All NYSE and AMEX assets are ranked by size and dividend yield. We form ten size portfolios by allocating 10% of the assets to each portfolio. Each asset remains in this portfolio as long as there is return data for that asset. The portfolios are equally weighted although the number of assets may change through time as assets are delisted. This form of portfolio construction is not subject to the type of survivorship bias possibly induced by requiring continuous trading. The ten dividend yield portfolios are formed in an almost identical manner, with one slight exception. Assets with a dividend yield of zero (generally greater than 10% of the sample) are allocated to the first portfolio. The remaining assets are allocated to the remaining nine portfolios.

For each set of portfolios we initially test the joint null hypothesis that the mispricing is zero for all portfolios within a standard multivariate regression framework

\[
\begin{bmatrix}
R^s_1 \\
\vdots \\
R^s_{10}
\end{bmatrix} = \begin{bmatrix} I_{10} \otimes (\mathbf{1}^G) \end{bmatrix} + \begin{bmatrix} xz \\
\vdots \\
xz \\
\ldots \\
\ldots \\
xz \\
xz \\
\varepsilon_{10}
\end{bmatrix}
\]

(7)

where: \( x = (s \text{ for size portfolios, } d \text{ for yield portfolios}); \)
\[ z = \{ n \text{ for basis portfolios estimated with only continuously traded assets; } m \text{ for basis portfolios estimated from all assets}; \]
\[ a_{1}^{xz} = \text{mispriicing of portfolio } i, \text{ formed by ranking on instrument } x \]
\[ \text{relative to basis portfolios } z; \]
\[ \Theta = \text{the Kronecker product;} \]
\[ \lambda = \text{a vector of ones.} \]

We will write (7) more compactly as

\[ R^{x} = \left[ I \otimes (\lambda : G_{x}') \right] \beta^{xz} + \epsilon^{xz} \quad (8) \]

with the usual assumption that there is cross-sectional correlation between the \( \epsilon \)'s but not autocorrelation. That is, \( E(\epsilon^{xz} \epsilon^{xz'}) = (\Sigma^{xz} \otimes I_{10}) \). For the various possible combinations of portfolios and factor estimation techniques we test the hypothesis

\[ H_{0}: a_{1}^{xz} = \ldots = a_{10}^{xz} = 0 \quad (9) \]

versus

\[ H_{A}: a_{j}^{xz} \neq 0 \text{ for some } j = 1, 2, \ldots, 10 \]

In addition we estimate separate regressions which allow the level of mispricing to be different in January than in the rest of the year, (see Keim \([9, 10]\)). Under the null hypothesis, this January-specific mispricing, as well as the non-January-specific mispricing, should equal zero. The multivariate regression (8) is estimated (using monthly data) for both size and dividend yield portfolios and both types of basis portfolios over four separate five year subintervals; 1964-68, 1969-73, 1974-78, and 1979-83. To test the hypothesis (9) we compute the modified likelihood ratio (MLR)
statistic of Rao [14, p. 555]. This statistic has the advantage that its exact finite sample distribution is known when the regression errors are multivariate normally distributed. Rather than present the statistics for each subperiod we aggregate the tests across subperiods. These aggregate test statistics are reported in Table 3. The column denoted "a = 0" gives the aggregated test statistics for the hypothesis that the mispricing across portfolios is zero [eqn. (9)]. The column denoted "a_1 = 0" gives test statistics for the hypothesis that January specific abnormal returns are zero. Finally, the column "a_NJ" contains test statistics for the hypothesis that mispricing which is not specific to January is zero. Figures 1 through 6 graph the relation between size or yield and estimated mispricing.

From Table 3 there is not a dramatic advantage for either method of forming basis portfolios, although the set of basis portfolios formed from the entire data set does do better in terms of a lower number of rejections of the multifactor pricing model for the yield based portfolios. For comparison purposes we first consider typical tests of the "CAPM." The tests overwhelmingly reject mean/variance efficiency of standard market portfolio proxies (value-weighted and equal-weighted NYSE and AMEX stocks) for both January specific and non-January specific mispricing.

For either basis portfolio formation procedure there is sufficient seasonality in the basis portfolios so that we do not reject that January specific mispricing is zero, (for both size and yield portfolios). Thus the factor models are consistent with a scenario in which seasonal mispricing relative to the CAPM is caused by seasonal factor risk premia and constant betas.

A plot of the estimated January specific mispricing versus size and
yield are provided in Figures 2 and 5 respectively, [for the equal and value weighted market proxies, (CAPM-EW and CAPM-VW), the multifactor model with five factors estimated from the continuously traded sample (APT-5C), and from the entire sample (APT-5M)]. January specific mispricing relative to the versions of the CAPM exhibit the same type of behavior documented by others. There is a pronounced negative relation between firm size and January mispricing (as shown in Keim [9]) and a nonlinear (slightly u-shaped) relation between yield and January mispricing (as shown in Keim [10]). For the multifactor models there is no discernable size related January seasonal (except possibly for the first portfolio) while there is a slight (statistically insignificant) U-shaped yield related January seasonal.

For the five and ten-factor models we reject $a = 0$ and $a_{NJ} = 0$ for the size portfolios (see also Figures 1 and 3). When yield based portfolios are used as LHS variables we reject $a = 0$ and $a_{NJ} = 0$ when we use the continuously traded set of assets to form basis portfolios. However, when we use our proposed technique which allows use of the entire data set we do not reject the multifactor pricing model (with yield based portfolios).

Figures 4 and 6 are graphs of the relation between yield and mispricing ($a$ and $a_{NJ}$, respectively). There does seem to be a slight inverted U pattern in the non-January specific mispricing of yield based portfolios.

The results in Table 3 and Figures 1 through 6 indicate that there may be some improvement in pricing performance by using the basis portfolios constructed from the entire data set but that the difference in mispricing relative to $C^R$ or $C^M$ is not very large. This small difference between the two sets of basis portfolios may be due to a variety of reasons. The large
number of securities without missing observations, n (which averages 1672 over the four subperiods) may lead to values of $C^n$ close to the null matrix. Simulations in [5] indicate that the error in forming basis portfolios is likely to be quite small, hence the improvement expected from adding the assets with missing data may be small. However, in applications where the continuously traded sample size is smaller our technique will lead to greater proportionate improvement. Also, the idiosyncratic variability of the assets which do not trade continuously might be sufficiently larger than the continuously traded assets that the increase in precision due to increased sample size is partially offset by the higher variability (see footnote 4) or these assets might not satisfy the appropriate stationarity assumptions.

III. Conclusions

We have proposed a method of constructing factor mimicking portfolios that utilizes returns on all assets. In addition to providing a larger sample of assets, the technique is not subject to the survivorship bias inherent in methods that require each asset to have no missing data.

Even though continuously traded samples using 1500 to 1750 firms provide apparently good estimates of the factors, we find that the mimicking portfolios using the entire data set perform better, in terms of pricing assets. The improvement is most noticeable when we investigate assets sorted on the basis of dividend yield. The importance of techniques which can handle missing observations should grow with the increasing availability of security returns data for assets which are unlikely to have long continuous trading histories (e.g., over the counter stocks).
Appendix

In this appendix we give a simple example of the asymptotic principal components technique which provides some useful intuition. The simplest case with which we can deal is when there is one pervasive factor and two time periods, i.e.,

\[
\bar{R}_{it} = b_i(\gamma + \bar{F}_t) + \epsilon_{it} \quad i = 1, 2, 3, \ldots \\
\quad t = 1, 2.
\]

Note that the risk premia, \( \gamma_t \), can vary through time arbitrarily, but that the risk premia are not separately identifiable from the mean-zero factor, \( \bar{F}_t \). In this case \( \Omega^n \) is a 2 x 2 matrix whose \((t, r)\) element is equal to

\[
\sum_{i=1}^{n} R_{it} R_{ir} / n.
\]

The diagonal elements are given by

\[
\Omega_{rr}^n = (\gamma_r + \bar{F}_r)^2 \left( \sum_{i=1}^{n} \frac{b_i^2}{n} \right) + \left( \sum_{i=1}^{n} \frac{\epsilon_{ir}^2}{n} \right) + 2(\gamma_r + \bar{F}_r)(\sum_{i=1}^{n} b_i \epsilon_{ir} / n) \quad r = 1, 2
\]

(12)

The off diagonal terms in \( \Omega^n \) are given by
\[\Omega_{12} = \Omega_{21} = (\gamma_1 + \bar{z}_1)(\gamma_2 + \bar{z}_2)\left(\sum_{i=1}^{n} b_i^2/n\right)\]

\[+ \left(\sum_{i=1}^{n} \varepsilon_i \varepsilon_i/n\right) + (\gamma_1 + \bar{z}_1)\left(\sum_{i=1}^{n} b_i \varepsilon_i \varepsilon_i/n\right)\]

\[+ (\gamma_2 + \bar{z}_2)\left(\sum_{i=1}^{n} b_i \varepsilon_i \varepsilon_i/n\right)\]  

Under our assumptions, the (cross-sectional) average squared beta converges (as \(n \to \infty\)) to some value, say \(\bar{z}^2\) and the (cross-sectional) average \(\varepsilon_i^2\) converges to \(\sigma^2\). By the assumption of an approximate factor structure, and temporally independent \(\varepsilon\)'s, the last term in (12) and the last three terms in (13) converge (again as \(n \to \infty\)) to zero. Therefore, as \(n \to \infty\) \(\Omega^n\) converges to

\[\Omega = \bar{z}^2 \begin{bmatrix}
(\gamma_1 + \bar{z}_1)^2 & (\gamma_1 + \bar{z}_1)(\gamma_2 + \bar{z}_2) \\
(\gamma_1 + \bar{z}_1)(\gamma_2 + \bar{z}_2) & (\gamma_2 + \bar{z}_2)^2
\end{bmatrix} + \sigma^2 \mathbb{I}_2.\]  

The limit matrix, \(\Omega\) contains all of the information which we seek, (i.e., \((\gamma_1 + \bar{z}_1)\) and \((\gamma_2 + \bar{z}_2))\). We merely need a means of extracting this information. The reader can check that the first eigenvector of \(\Omega\) is proportional to the vector of realized factors plus their risk premia.


5. ______. "Risk and Return in an Equilibrium APT." Working paper #9, Department of Finance, Northwestern University, March 1987.


10. ______. "Dividend Yields and Stock Returns: Implications of Abnormal


1. See, for example, Grinblatt and Titman [6] and Huberman, Kandel, and Stambaugh [8].

2. $G^n_t$ represents the $t^{th}$ column of $G^n$ while $G^n_i$ represents the $i^{th}$ row of $G^n$. The former is the estimate of the $k$ factors at time $t$ while the latter is the estimated $T \times 1$ time series of factor $i$.

3. For vectors the norm $\| \cdot \|$ is the inner product. For symmetric positive semi-definite matrices the norm $\| \cdot \|$ is the largest eigenvalue of the matrix.

4. For example, assume we have a sample of $n$ observations on a random variable that is normal and independently distributed $NID(\mu, \sigma_1^2)$. Suppose we are given a sample of $n$ observations on a second random variable that is $NID(\mu, \sigma_2^2)$. The sample mean from the original sample has a variance of $\sigma_1^2/n$ while the sample mean from the combined sample has a variance of $(\sigma_1^2 + \sigma_2^2)/4n$. The latter is larger than the former if $\sigma_2^2 > 3\sigma_1^2$. If we knew the values of $\sigma_1$ and $\sigma_2$ a more efficient estimate of $\mu$ can be obtained by a weighted average where the weight for the first $n$ assets is $[n\sigma_1^2(1/\sigma_1^2 + 1/\sigma_2^2)]^{-1}$ and the weight for the second $n$ assets is $[n\sigma_2^2(1/\sigma_1^2 + 1/\sigma_2^2)]^{-1}$. The variance of this estimate of $\mu$ is not greater than than the variance of the original sample regardless of the relative sizes of the variances. Thus adding appropriately weighted observations will not decrease precision.

5. Let $\hat{\Sigma}_r$ and $\hat{\Sigma}_u$ be our estimate of the contemporaneous covariance matrix of $\varepsilon$ under the restricted and unrestricted models, respectively. The MLR statistic for our hypotheses is given by

$$[(\hat{\Sigma}_r/|\hat{\Sigma}_u|) - 1] \cdot (T - k - p)/p$$
where $|\cdot|$ denotes the determinant, $T$ is the number of time series observations, $k$ is the number of regressors, and $p$ is the number of LHS portfolios.

6. The MLR statistics have an F distribution. Several methods of aggregating F-tests have been suggested. We calculate the p-value (or right tail area) of our F-statistic and find the value of $\chi^2$ random variable which has the same p-value (the degrees of freedom of the $\chi^2$ variable are equal to the numerator degrees of freedom of the F variable). We then sum the $\chi^2$ statistics across the subperiods. The aggregate statistic is compared to a $\chi^2$ statistics with degrees of freedom equal to the sum of the subperiod degrees of freedom.

7. These tests amount to tests of mean-variance efficiency of the market proxy portfolio chosen. They tell us little about the validity of the CAPM (with the true market) unless we are willing to make some statements about the relation between the proxy and the true market.

<table>
<thead>
<tr>
<th>Period</th>
<th>Number of Firms w/o Missing Data</th>
<th>Average Number of Firms-Total Sample</th>
<th>Minimum Number of Firms-Total Sample</th>
<th>Maximum Number of Firms-Total Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1964-68</td>
<td>1487</td>
<td>2146</td>
<td>2050</td>
<td>2188</td>
</tr>
<tr>
<td>1969-73</td>
<td>1720</td>
<td>2460</td>
<td>2186</td>
<td>2705</td>
</tr>
<tr>
<td>1974-78</td>
<td>1734</td>
<td>2566</td>
<td>2283</td>
<td>2715</td>
</tr>
<tr>
<td>1979-83</td>
<td>1745</td>
<td>2340</td>
<td>2265</td>
<td>2460</td>
</tr>
<tr>
<td>Average</td>
<td>1672</td>
<td>2378</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2
Canonical Correlations for First Ten Factors - Continuously Traded Sample vs. Entire Sample

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>(0.0001)</td>
<td>(0.0001)</td>
<td>(0.0001)</td>
<td>(0.0001)</td>
</tr>
<tr>
<td>2</td>
<td>0.999</td>
<td>0.998</td>
<td>0.996</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>(0.0001)</td>
<td>(0.0005)</td>
<td>(0.0010)</td>
<td>(0.0005)</td>
</tr>
<tr>
<td>3</td>
<td>0.993</td>
<td>0.993</td>
<td>0.996</td>
<td>0.996</td>
</tr>
<tr>
<td></td>
<td>(0.0018)</td>
<td>(0.0017)</td>
<td>(0.0011)</td>
<td>(0.0010)</td>
</tr>
<tr>
<td>4</td>
<td>0.989</td>
<td>0.984</td>
<td>0.993</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>(0.0030)</td>
<td>(0.0042)</td>
<td>(0.0017)</td>
<td>(0.0018)</td>
</tr>
<tr>
<td>5</td>
<td>0.984</td>
<td>0.980</td>
<td>0.990</td>
<td>0.988</td>
</tr>
<tr>
<td></td>
<td>(0.0042)</td>
<td>(0.0051)</td>
<td>(0.0027)</td>
<td>(0.0031)</td>
</tr>
<tr>
<td>6</td>
<td>0.976</td>
<td>0.960</td>
<td>0.988</td>
<td>0.985</td>
</tr>
<tr>
<td></td>
<td>(0.0063)</td>
<td>(0.0103)</td>
<td>(0.0030)</td>
<td>(0.0038)</td>
</tr>
<tr>
<td>7</td>
<td>0.968</td>
<td>0.941</td>
<td>0.983</td>
<td>0.980</td>
</tr>
<tr>
<td></td>
<td>(0.0083)</td>
<td>(0.0149)</td>
<td>(0.0045)</td>
<td>(0.0051)</td>
</tr>
<tr>
<td>8</td>
<td>0.955</td>
<td>0.915</td>
<td>0.967</td>
<td>0.946</td>
</tr>
<tr>
<td></td>
<td>(0.0115)</td>
<td>(0.0212)</td>
<td>(0.0085)</td>
<td>(0.0137)</td>
</tr>
<tr>
<td>9</td>
<td>0.928</td>
<td>0.844</td>
<td>0.891</td>
<td>0.798</td>
</tr>
<tr>
<td></td>
<td>(0.0180)</td>
<td>(0.0375)</td>
<td>(0.0270)</td>
<td>(0.0476)</td>
</tr>
<tr>
<td>10</td>
<td>0.387</td>
<td>0.413</td>
<td>0.864</td>
<td>0.493</td>
</tr>
<tr>
<td></td>
<td>(0.1107)</td>
<td>(0.1079)</td>
<td>(0.0334)</td>
<td>(0.0994)</td>
</tr>
</tbody>
</table>

Note: Approximate standard errors in parentheses.
Table 3
Tests for the Absence of Mispicing
10 Portfolios Formed by Market Value or Dividend Yield.
Basis Portfolios Estimated from Continuously Traded Sample and Entire Sample.\textsuperscript{a}

\[ R^n = a^T e^n + b^n F + \epsilon^n \quad \text{and} \quad R^n = a_{NJ}^T e^n + a^D + b^n F + \xi^n \]

<table>
<thead>
<tr>
<th>ASSETS USED</th>
<th>CAPM - EW</th>
<th>CAPM - VW</th>
<th>APT - 5</th>
<th>APT - 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>LHS Basis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portfolio</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constr.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a = 0</td>
<td>69.42</td>
<td>105.81</td>
<td>72.00</td>
<td>66.88</td>
</tr>
<tr>
<td>a = 0</td>
<td>(0.003)</td>
<td>(&lt;.001)</td>
<td>(0.001)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>a = 0</td>
<td>76.61</td>
<td>40.66</td>
<td>73.68</td>
<td>78.54</td>
</tr>
<tr>
<td>a = 0</td>
<td>(&lt;.001)</td>
<td>(0.441)</td>
<td>(0.001)</td>
<td>(&lt;001)</td>
</tr>
<tr>
<td>a = 0</td>
<td>80.06</td>
<td>44.32</td>
<td>78.40</td>
<td>88.06</td>
</tr>
<tr>
<td>a = 0</td>
<td>(&lt;.001)</td>
<td>(0.294)</td>
<td>(&lt;.001)</td>
<td>(&lt;.001)</td>
</tr>
<tr>
<td>a = 0</td>
<td>66.78</td>
<td>113.11</td>
<td>72.74</td>
<td>82.53</td>
</tr>
<tr>
<td>a = 0</td>
<td>(0.005)</td>
<td>(&lt;.001)</td>
<td>(0.001)</td>
<td>(&lt;.001)</td>
</tr>
<tr>
<td>a = 0</td>
<td>62.67</td>
<td>47.98</td>
<td>59.82</td>
<td>65.20</td>
</tr>
<tr>
<td>a = 0</td>
<td>(0.012)</td>
<td>(0.181)</td>
<td>(0.023)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>a = 0</td>
<td>47.93</td>
<td>48.01</td>
<td>46.41</td>
<td>56.24</td>
</tr>
<tr>
<td>a = 0</td>
<td>(0.182)</td>
<td>(0.180)</td>
<td>(0.225)</td>
<td>(0.046)</td>
</tr>
</tbody>
</table>

Note: P-values in parentheses

\textsuperscript{a} Aggregation of subperiod results. Statistic has \( \chi^2 \) distribution with 40 degrees of freedom under the null hypothesis.
Figure 2

JANUARY MISPricing
SIZE BASED PORTFOLIOS

(% per annum)

Portfolio

CAPM-VW  APT-5M  APT-5C  CAPM-EW

s1  s2  s3  s4  s5  s6  s7  s8  s9  s10
Figure 3

NONJANUARY MISPRICING
SIZE BASED PORTFOLIOS

(\% per annum)

Portfolio

CAPM-VW  APT-5M  APT-5C  CAPM-EW
Figure 4

MISPRICING
YIELD BASED PORTFOLIOS

(% per annum)

Portfolio

CAPM-VW  +  APT-5M  ✦  APT-5C  △  CAPM-EW

y1  y2  y3  y4  y5  y6  y7  y8  y9  y10
Figure 5

JANUARY MISPRICING
YIELD BASED PORTFOLIOS

(% per annum)

Portfolio

CAPM-VW

APT-5M

APT-5C

CAPM-EW
Figure 6

NONJANUARY MISPRICING
YIELD BASED PORTFOLIOS

(% per annum)

Portfolio

CAPM-VW
APT-5M
APT-5C
CAPM-EW