A Performance Comparison of Large-n Factor Estimators

Zhuo Chen, Gregory Connor, and Robert A. Korajczyk*

November 6, 2014

Abstract

This paper uses simulations to evaluate the performance of various methods for estimating factor returns in an approximate factor model when the cross-sectional sample (n) is large relative to the time-series sample (T). We study the performance of the estimators under a variety of alternative specifications of the underlying factor structure. We find that 1) all of the estimators perform well, even when they do not accommodate the form of heteroskedasticity present in the data; 2) for the sample sizes considered here, accommodating heteroskedasticity does not deteriorate performance much when simple forms of heteroskedaticity are present; 3) estimators that handle missing data by substituting fitted returns from the factor model converge to the true factors more slowly than the other estimators.

*Chen: PBC School of Finance, Tsinghua University, 43 Chengfu Road, Haidian District, Beijing, 100083, P. R. China. Phone: +86-10-62781370; E-mail: chenzh@pbcsf.tsinghua.edu.cn. Connor: Department of Economics, Finance and Accounting, National University of Ireland, Maynooth County Kildare, Ireland. E-mail: gregory.connor@nuim.ie. Korajczyk: Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208-2001, USA. Phone: (847) 491-8336; E-mail: r-korajczyk@kellogg.northwestern.edu. We thank Patrick Gagliardini, Eric Ghysels, Alex Horenstein, Egon Kalotay, Maja Kos, Eric Renault, Stephen Satchell, Alvin Stroyny, and participants at the Society for Financial Econometrics conference on Large Scale Factor Models and the Multinational Finance Society conference for helpful comments.

In a linear factor model of returns, the return on each asset is the sum of a linear combination of a few systematic factors plus an idiosyncratic return. Ross (1976) shows that in an economy with many assets, a linear factor model provides a natural way to capture the diversifiable and nondiversifiable components of asset returns. In Ross's original specification, returns were assumed to follow a strict factor model, that is, one in which the idiosyncratic returns have zero covariance. Chamberlain and Rothschild (1983) generalize the model, allowing nonzero covariance but imposing the assumption that the eigenvalues of the idiosyncratic-return covariance matrix are bounded as the number of assets grows to infinity. This generalization is called an approximate factor model. The approximate factor model framework has been used in a wide range of applications. In addition to common stock return modelling (Ross's original motivation), the approximate factor model framework is now used in business-cycle forecasting, e.g., Stock and Watson (2002, 2006), large-scale macroeconomic time-series modeling, e.g., Forni, Hallin, Lippi and Reichlin (2005), and credit models, e.g., Gagliardini and Gourieroux (2009). Our focus here is on Ross's original application, to common stock returns, but our results have potential relevance to other approximate factor model applications as well.

There are several econometric methodologies for estimating approximate factor models when the cross-sectional sample size (n) is large relative to the time-series sample size (T); the chosen methodology often depends upon the application at hand. Connor and Korajczyk (1986) show that the factors in an approximate factor model with k factors can be consistently estimated by the first k eigenvectors of the cross-product matrix of excess returns. Many other papers similarly rely on eigenvector-based estimation of factor returns, including Connor and Korajczyk (1987, 1988), Stroyny (1992), Stock and Watson (1998, 2002), and Jones (2001). In this paper, we use simulation methods to compare the performance of various methods for estimating factor returns. We calibrate our simulation model based on the observed features of US common stock returns. We simulate panel data sets of returns under different assumptions on the factor model, including the degree of cross-sectional and time-series heteroskedasticity and the cross-sectional correlations of idiosyncratic returns.

We apply the estimators to both balanced and unbalanced panel data sets of simulated returns. We consider a variety of cross-sectional sample sizes, and this allows us to investigate the convergence properties of the estimators as the sample size grows, and their levels of precision for particular sample sizes. For each simulation sample, we compare the estimated factors to the true factors and evaluate the performance of the estimators by averaging across a large number of simulated samples.

For the sample sizes considered here, all of the estimators perform well, even when they do not accommodate the form of heteroskedasticity in the simulated data, although one estimator lags the others slightly in performance. Estimators that explicitly incorporate either cross-sectional or time-series heteroskedasticity outperform other estimators when those types of heteroskedasticity are present. For the sample sizes considered here, accommodating general types of heteroskedasticity does not deteriorate performance much when the data are generated with simpler forms of heteroskedasticity.

1 Large-n Estimators of Factor-Mimicking Portfolios

Assume that the data-generating process for returns on all securities is a k-factor asset pricing model. Let e^{n} be an n-vector of ones. Let B be an $n \times k$ matrix of factor loadings, or betas. Let $r_{f,\mathrm{t}}$ denote the zero-beta return for period t, f_{t} denote the k-vector of zero-mean factor shocks at period t, and μ_{t} denote the k-vector of factor risk premia at period t. Let ϵ_{t} be an n-vector of idiosyncratic returns and let r_{t} denote the n-vector of asset returns. The equilibrium asset pricing model implies

$$R_t = r_{\mathrm{t}} - e^{\mathrm{n}} r_{f,\mathrm{t}} = B(\mu_{\mathrm{t}} + f_{\mathrm{t}}) + \epsilon_{\mathrm{t}}, \tag{1}$$

where $E[f_{\mathrm{t}}] = 0$ and $E[\epsilon_{\mathrm{t}}] = 0$. We assume that zero expectation for residual returns holds conditional on f_{t}, $E[\epsilon_{\mathrm{t}} \mid f_{\mathrm{t}}] = 0$. A strict factor model is one in which the residual covariance matrix is diagonal, with bounded elements (i.e., $E[\epsilon_{\mathrm{t}}\epsilon_{\mathrm{t}}'] = D$, where $D_{i,i} < \infty$ and $D_{i,j} = 0$ for $i \neq j$). An approximate factor model allows for covariation in idiosyncratic returns across assets that is diversifiable in the limit as $n \to \infty$. This implies that the eigenvalues of $E[\epsilon_{\mathrm{t}}\epsilon_{\mathrm{t}}']$

$= \Sigma$ are bounded as $n \to \infty$. For a time-series sample over the periods $t = 1, 2, ... T$, define R to be the $n \times T$ matrix of realized excess returns on the n securities for T time periods: $R = [R_1 R_2 ... R_T]$. We write the data-generating process in matrix form as

$$R = BF + \epsilon, \tag{2}$$

where F is the $k \times T$ matrix of the realizations of the factors plus risk premia and ϵ is the $n \times T$ matrix of idiosyncratic returns. We wish to provide an estimate of the factor excess returns, F, in settings where n can be large relative to T. In the next four subsections we describe the estimation procedures that we evaluate and compare in our study.

1.1 Asymptotic principal components (APC)

For $n >> T$ (for example, 10,000 assets over a 240-month time period) the difficulty posed by standard factor analytic procedures is that for the estimation of the $k \times T$ matrix of factor realizations, F, one needs to estimate and invert a much larger $n \times n$ covariance matrix (in the example above, where $n = 10,000$ and $T = 240$, the $n \times n$ covariance matrix has over 50 million distinct entries, but we have only 1.2 million data points). Connor and Korajczyk (1986) suggest APC as an alternative method of estimating factor portfolio returns directly without needing to estimate and decompose the full covariance matrix. Let Ω denote the $T \times T$ cross-product matrix of excess returns:

$$\Omega = \frac{1}{n} R'R. \tag{3}$$

Let $\widehat{F}$ denote the $k \times T$ matrix of the k eigenvectors of Ω corresponding to the largest k eigenvalues of Ω. Connor and Korajczyk (1986) show that, for a k-factor approximate factor model, $\widehat{F}$ is an n-consistent estimate of F. They call this estimator the asymptotic principal components estimator. This estimator makes no assumptions about cross-sectional heteroskedasticity in the idiosyncratic returns (the diagonal elements of V) other than the boundedness discussed above. However, it does not attempt to utilize any such heteroskedasticity in estimation. The estimator makes fairly restrictive assumptions about any time-series

heteroskedasticity in idiosyncratic returns: any asset could have time variation in its idiosyncratic variance, but the average (across the n assets) idiosyncratic variance must be time invariant. The estimator also assumes that the econometrician has a balanced panel. That is, there are no missing data in the $n \times T$ matrix of returns, R. Restricting the sample to assets with a complete return history for T periods clearly induces survivorship bias into the factor estimates.

A number of subsequent studies have generalized the procedure to take into account cross-sectional and time-series heteroskedasticity as well as unbalanced panels.

1.2 Incorporating cross-sectional heteroskedasticity

Connor and Korajczyk (1988) propose estimating the diagonal idiosyncratic variance matrix by regressing asset returns on the initial APC factor estimates, $\widehat{F}$, and using the residuals to estimate the diagonal residual covariance matrix, D,

$$\widehat{\epsilon} = R - \widehat{B}\widehat{F} \tag{4}$$

$$\widehat{D} = Diag\left(\frac{\widehat{\epsilon}\widehat{\epsilon}'}{T}\right). \tag{5}$$

The return matrix is then rescaled by the estimated standard deviations of the idiosyncratic returns,

$$R^* = \widehat{D}^{-1/2}R, \tag{6}$$

and the factors are estimated by applying the APC procedure to R^*. We will refer to this estimator as APC-X to denote that it is a variant of the APC procedure designed to account for cross-sectional heteroskedasticity.

Stroyny (1992) proposes a large-n variant of maximum-likelihood factor analysis based on the EM algorithm (Dempster, Laird, and Rubin (1977) and Rubin and Thayer (1982)). A standard identification assumption in factor analysis is that the factors have a covariance

matrix equal to the identity matrix. Stroyny (1992) argues that applying this constraint significantly slows the convergence of the EM factor analysis procedure and advocates only applying the desired rotation of the factors after the procedure has converged. In simulations, Stroyny (1992, Table 1) finds that the modified procedure is significantly faster than the standard EM procedure. The number of iterations required actually decreases in n for the Stroyny procedure and is nonmonotonic for the standard EM procedure (for $n = 5{,}000$, EM requires 1,194 iterations and Stroyny's procedure requires 19, iterations while for $n = 10{,}000$, EM does not converge and Stroyny's procedure requires 18 iterations). The total CPU time is approximately linear in n for the Stroyny procedure. We refer to this procedure as MLFA-S to denote Maximum Likelihood Factor Analysis using the Stroyny (1992) procedure.

1.3 Incorporating time-series heteroskedasticity

Factor analysis generally assumes that each asset's idiosyncratic volatilities is constant through time, while the APC procedure assumes that the average idiosyncratic volatility across assets is constant through time. Given the evidence of time variation in volatility, in general (for example, Andersen, Bollerslev, and Diebold (2010)), and idiosyncratic volatility, in particular (for example, Campbell et al. (2001) and Connor, Korajczyk, and Linton (2006)), it seems that incorporating times-series heteroskedasticity into factor estimation is desirable. Jones (2001) proposes such an estimator, called heteroskedastic factor analysis (HFA). Jones (2001) assumes that the cross-sectional average idiosyncratic volatility is time dependent

$$\overline{\Sigma}_t =^{\text{plim}}_{n\to\infty} \frac{1}{n}\Sigma_{i,i,t},$$

where $\Sigma_{i,i,t}$ is the (i,i) element of the covariance matrix of idiosyncratic returns, $\Sigma_t = \mathrm{E}(\epsilon_t\epsilon_t')$. Define the $T \times T$ matrix, $\overline{\Sigma}$, to be the diagonal matrix with elements $(t,t) = \overline{\Sigma}_t$. Jones's procedure estimates factor returns by calculating eigenvectors of the scaled matrix,

$$\widehat{\overline{\Sigma}}^{-1/2}\Omega\widehat{\overline{\Sigma}}^{-1/2}. \tag{7}$$

These factor estimates are used to reestimate idiosyncratic returns and $\widehat{\widehat{\Sigma}}$, and the process is iterated until convergence.

1.4 Accommodating unbalanced panels

It is not unusual for empirical analyses of factor models to estimate factor-mimicking portfolios from balanced panels of data (e.g., Roll and Ross (1980), Connor and Korajczyk (1988), Lehmann and Modest (1988), and Jones (2001)). However, requiring a balanced panel induces survivorship bias into the sample. There are several alternative approaches to estimating factor-mimicking returns with missing data.

Connor and Korajczyk (1987) suggest a method of factor estimation with missing data. This procedure estimates the cross-product matrix Ω^u over the all observed data (the u superscript denotes an unbalanced panel). Define $\delta_{i,t} = 1$ if the $\{i,t\}$ element of R is observed and $\delta_{i,t} = 0$ otherwise, and define the $\{t,\tau\}$ element of Ω as

$$\Omega^u_{t,\tau} = \frac{\sum_{i=1}^n \delta_{i,t}\delta_{i,\tau}R_{i,t}R_{i,\tau}}{\sum_{i=1}^n \delta_{i,t}\delta_{i,\tau}}. \tag{8}$$

Factor mimicking portfolio returns are estimated from the eigenvectors of the redefined Ω^u. While Ω is guaranteed to be positive semidefinite for a balanced sample, Ω^u is not for an unbalanced sample. However, for the samples typically used in practice, we have never come across a case where Ω^u is not positive semidefinite. We will refer to this estimator as APC-M to denote that it is the APC estimator with missing data.

The APC-M estimator can be modified to accommodate cross-sectional heteroskedasticity by constructing Ω^u from the scaled observed returns defined in (6). That is, the factor estimates are the k eigenvectors of

$$\Omega^{u*}_{t,\tau} = \frac{\sum_{i=1}^n \delta_{i,t}\delta_{i,\tau}R^*_{i,t}R^*_{i,\tau}}{\sum_{i=1}^n \delta_{i,t}\delta_{i,\tau}} \tag{9}$$

associated with the k largest eigenvalues. We will refer to this estimator as APC-MX to denote that it is the APC estimator with missing data incorporating cross-sectional heteroskedasticity.

Stock and Watson (1998, 2002) extend the APC approach in a number of dimensions. We focus here on the extension to accommodate missing data. Under stronger assumptions than necessary for consistency of the APC estimator (i.e., $\epsilon_{i,t} \sim$ i.i.d. $N(0, \sigma^2)$) the MLE estimator of $\{B, F\}$ minimizes the nonlinear least squares objective function (see Stock and Watson (1998)):

$$\Lambda = (nT)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} \delta_{i,t} (R_{i,t} - B_{i,\bullet} F_{\bullet,t})^2 \tag{10}$$

where, for any matrix X, $X_{i,\bullet}$ denotes the i^{th} row of X and $X_{\bullet,t}$ denotes the t^{th} column of X. The first-order conditions are

$$F_{\bullet,t} = \left(\sum_{t=1}^{T} \delta_{i,t} B'_{i,\bullet} B_{i,\bullet} \right)^{-1} \left(\sum_{t=1}^{T} \delta_{i,t} B'_{i,\bullet} \; R_{i,t} \right) \tag{11}$$

and

$$B_{i,\bullet} = \left(\sum_{t=1}^{T} \delta_{i,t} R_{i,t} F'_{\bullet,t} \right) \left(\sum_{t=1}^{T} \delta_{i,t} F_{\bullet,t} \; F'_{\bullet,t} \right)^{-1}, \tag{12}$$

which correspond to the time-series and cross-sectional regressions (2) (which is a time-series regression when viewed as a regression of R on F and a cross-sectional regression when viewed as a regression of R on B) applied to the observed data in the unbalanced panel. They obtain the MLEs of F and B by iterating between the first-order conditions, Equations (11) and (12) (Stock and Watson, 1998). An alternative approach to obtaining the MLEs is to minimize Λ using the EM algorithm of Dempster, Laird, and Rubin (1977). Let Λ^* denote the negative complete-data log-likelihood function

$$\Lambda^*(B, F) = (nT)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} (R^*_{i,t} - B_{i,\bullet} F_{\bullet,t})^2, \tag{13}$$

where $R^*_{i,t}$ is the latent value of $R_{i,t}$. The EM algorithm iteratively maximizes the expected value of the complete-data likelihood (minimizes the expected value of $\Lambda^*(B, F)$), conditional on the estimates from the prior iteration. Let $B^j$and F^j denote the estimated factor loadings and factors after the j^{th} iteration of the algorithm. Under the assumed error structure, this amounts to minimizing, at iteration j,

$$(nT)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} (R^{*,j-1}_{i,t} - B^j_{i,\bullet} F^j_{\bullet,t}), \tag{14}$$

where $R_{i,t}^{*,j-1} = R_{i,t}$ if $\delta_{i,t} = 1$ and $R_{i,t}^{*,j-1} = B_{i,\bullet}^{j-1} F_{\bullet,t}^{j-1}$ if $\delta_{i,t} = 0$ (see Stock and Watson (1998, p. 11)). Thus, the missing data are filled in with the fitted values from the factor model obtained in the previous iteration. The factor portfolio returns obtained from minimizing (14) are equal to (up to a nonsingular rotation L) the APC estimate obtained from $R_{i,t}^{*,j-1}$. Applying the EM algorithm amounts to an iterative application of APC until convergence. In the case where there are no missing data, the Stock and Watson (1998, 2002) estimator is identical to the Connor and Korajczyk (1986) estimator.

Stroyny's MLFA-S approach can be extended similarly by replacing missing observations with the fitted returns from the previous iteration's estimates of F and B. We call this estimator MLFA-MS. Jones (2001) suggests extending the HFA procedure along the lines of Connor and Korajczyk (1987), in which Ω and $\overline{V}$ are estimated over the nonmissing sample. We call this estimator HFA-M to denote that it is the HFA estimator with missing data.

2 Empirical Analysis

We simulate asset returns using alternative specifications of factor models for increasing numbers of assets and study the convergence, as n increases, of the factor estimates to the true underlying simulated factor model. We consider four basic cases regarding the nature of the covariance matrix of idiosyncratic returns: 1) cross-sectional and time-series homoskedasticity: $E[\epsilon_t \epsilon_t'] = \sigma^2 I$, where I is an $n \times n$ identity matrix; 2) cross-sectional heteroskedasticity and time-series homoskedasticity: $E[\epsilon_t \epsilon_t'] = D$, where D is a $n \times n$ diagonal matrix that is time invariant; 3) cross-sectional homoskedasticity and time-series heteroskedasticity: $E[\epsilon_t \epsilon_t'] = \sigma_t^2 I$, where I is an $n \times n$ identity matrix; and 4) cross-sectional and time-series heteroskedasticity: $E[\epsilon_t \epsilon_t'] = D_t$, where D_t is a $n \times n$ diagonal matrix. For each of these cases, we consider both balanced and unbalanced panels to assess the effects of missing data. However, we maintain throughout the assumption that the data are missing at random (MAR). In addition to the strict factor models discussed above, we allow for diversifiable levels of correlation in idiosyncratic returns across assets (i.e., nondiagonal

idiosyncratic covariance matrices) by constructing idiosyncratic returns, $\epsilon_{i,t}$, as

$$\epsilon_{i,t} = \rho\epsilon_{i-1,t} + u_{i,t} \tag{15}$$

for $\rho \in \{0, 0.25, 0.50\}$. As long as $\rho < 1$, the idiosyncratic returns are diversifiable for large n. This gives us 24 different cases to simulate (4 cases regarding cross-sectional and time-series heteroskedasticity $\times$ 2 cases regarding missing data $\times$ 3 cases regarding cross-asset correlation in idiosyncratic returns).

2.1 Simulation design

Each simulation sets $T = 240$ to correspond to a twenty-year period of monthly data. The numbers of assets, n, used in the simulation are 250, 500, 750 and 1,000 to 10,000 in increments of 1,000. To give a sense for the cross-sectional sample sizes used here versus various equity markets, Table 1 lists the minimum, mean, and maximum number of companies, over the 1991 to 2010 period, included in the CRSP indices for the New York (NYSE), American (AMEX), and NASD Stock Exchanges. The Table also includes similar figures for various exchanges over the 1996 to 2010 period, which are obtained from the World Federation of Exchanges.[1] The combined NYSE, AMEX, and NASD markets have a minimum of 5663 and a maximum of 9047 firms. The equivalent figures are 4721 and 5798 for the Bombay Stock Exchange, 2362 and 2693 for the London Stock Exchange, 720 and 894 for Shanghai, 1266 and 3937 for Canada, 336 and 557 for Brazil, and 148 to 570 for Poland. Thus our range of 250 to 10,000 firm in the simulation covers the sizes of a large number of national exchanges. We simulate a three-factor model ($k = 3$). For each scenario, we run 5,000 iterations of the simulation. We apply each of the relevant estimators to obtain estimates of the k factors. We do not study the question of the appropriate tests for the (unknown) true number of factors (for example, Connor and Korajczyk (1993) and Bai and Ng (2002)). That is, we simulate a three-factor model and estimate three factors.

[1] Downloaded from http://www.world-exchanges.org/statistics/monthly-reports on 3 November, 2014.

For each simulation and each estimator, we regress the estimated factor mimicking portfolio returns on the true underlying factors and a constant. Because of the well-known rotational indeterminacy of factor estimates, we regress each estimated factor on all three true factors,

$$\widehat{F} = \alpha + bF + u. \tag{16}$$

For each iteration of the simulation, we tabulate the R^2 values and the values of the estimated intercepts (and associated t-statistics) of these k regressions. Perfect estimators would imply R^2 values equal to unity and intercepts equal to zero.

2.1.1 Simulating the Factors

We simulate the return matrix R (dimension $n \times T$) using a three-factor model. We construct the true factor excess return matrix, F, by drawing from a normal distribution, $N(\mu, \Sigma_F)$. The mean and standard deviation of the three factors correspond to those of the Fama–French (1993) three factors, R_m, HML, and SMB[2]. We use monthly observations from January 1991 to December 2010 to calculate the factors' means and standard deviations: $\mu \times 100 = [0.58; 0.31; 0.34]$, $\sigma_F \times 100 = [4.46; 3.51; 3.37]$. We ignore any correlation across the true Fama–French factors and simulate the factor model assuming zero correlation across the three factors.

2.1.2 Simulating the Factor Loadings

The simulated beta loading matrix B (dimension $n \times k$) is generated based on the empirical distribution of stocks' loadings on the Fama-French factors. For each common stock traded in NYSE/NASDAQ/AMEX with more than 36 months of observations (10,937 stocks in total) over the period from January 1991 to December 2010, we estimate a time-series regression of stock excess returns on the Fama–French factors and calculate the average factor loadings and standard deviation of factor loadings: $\overline{B}_{i,\cdot} = [1.001; 0.882; 0.210]$, $\sigma_B =$

[2] Available at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library/f-f_factors.html.

[0.824; 1.204; 1.315]. In the simulation we draw betas from a multivariate normal distribution with these parameters. As with the factor covariance matrix, we assume that the covariance across the factor betas is zero.

2.1.3 Simulating the Idiosyncratic Returns

Similarly, we rely on the estimated residuals from the Fama–French three-factor model applied to the CRSP sample stocks over the 1991 to 2010 period to define the properties of the simulated idiosyncractic returns. Let $\widehat{\sigma}_i$ be the estimated standard deviation of the idiosyncratic return of asset i. Also, let $\widehat{\sigma}_t$ be the cross-sectional average standard deviation of idiosyncratic returns in period t.

Heteroskedasticity Case 1: When we assume both cross-sectional and time-series homoskedasticity and no cross-correlation we construct idiosyncratic returns that are drawn from a normal distribution with a mean of zero and a standard deviation, $\overline{\sigma}$, equal to the average (across the 10,937 stocks) value of $\widehat{\sigma}_i$. When we have cross-correlation, the idiosyncratic return for asset 1 in period t, $\varepsilon_{1,t}$, is drawn from this distribution and the remaining idiosyncratic returns are constructed as $\varepsilon_{i,t} = \rho\varepsilon_{i-1,t} + u_{i,t}$, where $u_{i,t} \sim N(0, (1-\rho^2)\overline{\sigma}^2)$. This gives each asset an unconditional idiosyncratic standard deviation of $\overline{\sigma}$. This is done independently for each time period, t.

Heteroskedasticity Case 2: When we assume cross-sectional heteroskedasticity but time-series homoskedasticity, we randomly pick n empirical standard deviations calculated from the Fama-French three-factor regression residuals, $\widehat{\sigma}_i$, from the sample of 10,937 stocks. Then we generate the residual matrix from a normal distribution with mean 0 and standard deviation $\widehat{\sigma}_i$. If we have cross-sectional correlation, the idiosyncratic return for asset 1 in period t, $\varepsilon_{1,t}$, is drawn from a $N(0, \widehat{\sigma}_1^2)$ distribution, and the remaining idiosyncratic returns are constructed as $\varepsilon_{i,t} = \rho\varepsilon_{i-1,t} + u_{i,t}$, where $u_{i,t} \sim N(0, \widehat{\sigma}_i^2 - \rho^2\widehat{\sigma}_{i-1}^2)$.

Heteroskedasticity Case 3: When we assume time-series heteroskedasticity but cross-sectional homoskedasticity, every asset's idiosyncratic return in period t is drawn from a normal distribution with standard deviation equal to $\widehat{\sigma}_t$, the cross-sectional average standard

deviation of idiosyncratic returns in period t. Figure 1 plots the time series of $\widehat{\sigma}_t$ over our sample period. There is substantial variation in $\widehat{\sigma}_t$ through time. When we have cross-correlation, the idiosyncratic return for asset 1 in period t, $\varepsilon_{1,t}$, is drawn from a $N(0, \widehat{\sigma}_t^2)$ distribution, and the remaining idiosyncratic returns are constructed as $\varepsilon_{i,t} = \rho\varepsilon_{i-1,t} + u_{i,t}$, where $u_{i,t} \sim N(0, (1-\rho^2)\widehat{\sigma}_t^2)$.

Heteroskedasticity Case 4: when we assume both time-series and cross-sectional heteroskedasticity, we assume that the idiosyncratic variance of each of the 10,937 stocks is proportional to the cross-sectional average idiosyncratic variance in period t. For each stock, i, we estimate the constant of proportionality, θ_i,

$$\widehat{\theta}_i = \frac{1}{T_i} \sum_{t \in \Upsilon_i} \frac{\widehat{\varepsilon}_{i,t}^2}{\widehat{\sigma}_t^2}, \tag{17}$$

where Υ_i is the set of time periods for which asset i has observations, and T_i is number of elements in Υ_i. Thus, an asset's idiosyncratic risk has a common element (driven by $\widehat{\sigma}_t^2$) and an asset-specific element, consistent with the evidence in Connor, Korajczyk, and Linton (2006). Every asset's idiosyncratic return in period t is drawn from a $N(0, \widehat{\theta}_i\widehat{\sigma}_t^2)$ distribution. When we have cross-correlation, the idiosyncratic return for asset 1 in period t, $\varepsilon_{1,t}$, is drawn from a $N(0, \widehat{\theta}_i\widehat{\sigma}_t^2)$ distribution, and the remaining idiosyncratic returns are constructed as $\varepsilon_{i,t} = \rho\varepsilon_{i-1,t} + u_{i,t}$, where $u_{i,t} \sim N(0, \widehat{\theta}_i\widehat{\sigma}_t^2 - \rho^2\widehat{\theta}_{i-1}\widehat{\sigma}_t^2)$.[3]

For each case, we draw the true factor matrix F once. For each run of the simulation, we draw the $n \times k$ matrix of factor loadings and the $n \times T$ matrix of idiosyncratic returns from the distribution described above. Given these, we generate an $n \times T$ return matrix $R = FB + \varepsilon$ and apply the factor estimators to the returns for n = 250, 500, 750, 1,000, 2,000, 3,000, ..., and 10,000 for 5,000 simulations. For each of our 24 case combinations, and 5,000 simulations, we regress $\widehat{F}$ on the true F, and record the adjusted R^2 the estimated intercept, $\widehat{\alpha}$, and the associated t-statistic for the intercept.

For the estimators that require iteration to convergence, that is HFA, MLFA-S, HFA-M,

[3]For Heteroskedasticity Cases 2 and 4 with idiosyncratic cross-correlations, we impose the constraint that the standard deviation of the idiosyncratic return be at least 0.1%.

and APC-EM, we run the iterations until the minimum R^2 (across the $k = 3$ estimated factors) from the multivariate regression of the factors from iteration j on the factors from iteration $j-1$ is greater than or equal to 0.999.

To generate an unbalanced sample with missing observations, we randomly pick n stocks from the 10,937 CRSP sample stocks, with replacement, and use their pattern of missing observations during 1991 to 2010 to generate a simulated return series with missing values. When we pick the n stocks to generate the pattern of missing observations, we also use those stocks' idiosyncratic volatility, $\widehat{\sigma}_i$, in Cases 2 and 4 for which we have cross-sectional heteroskedasticity.

2.2 Balanced panel of asset returns

For the balanced panel case, we apply four estimators, APC, APC-X, HFA, and MLFA-S. Figure 2 shows the average minimum R^2 values for Case 1 for all three factors and all three cross-correlation structures. The factors change across columns, and the value of ρ changes across rows. Several points are clear from the figure. First, all four estimators perform comparably even though three of the estimators are estimating extra parameters. Second, accuracy falls as we estimate additional factors and as the idiosyncratic return correlation across assets increases. Third, all of the estimators are fairly accurate. The smallest mean R^2 values exceed 0.9, even for the estimates of the third factor, with $\rho = 0.5$, and with the smallest number of assets in the crosssection ($n = 250$). When we have 2,000 assets in the crosssection, almost all mean R^2 values equal 0.99 or higher.

Figure 3 shows the average R^2 values for Case 2 for all three factors and all three cross-correlation structures. In this scenario, idiosyncratic-return variance varies across assets but is constant through time. In this instance, one would expect that APC-X and MLFA-S would have superior performance since they explicitly take into account the differences in idiosyncratic risks across assets.

Again, there are several points that are clear from the figure. First, the procedure from Stroyny (1992), MLFA-S, dominates all of the other procedures, until we reach large values

of N (4,000 to 5,000). Second, APC-X dominates APC and HFA significantly for estimation of factors two and three, but actually underperforms them slightly for estimation of factor one. Third, with the exception of MLFA-S, cross-sectional heteroskedasticity significantly slows the convergence of the factor estimates to the true factors. While under Case 1 the R^2s are 0.9 and higher, under Case 2, the R^2 values are as low as 0.5 and need approximately 3,000 to 4,000 assets for the first factor; 4,000 to 5,000 assets for the second factor; and 5,000 (APC-X) and 9,000 to 10,000 (APC and HFA) assets for the third factor to attain R^2 values of 0.975. Fourth, APC and HFA are equivalent, which would be expected given that there is no time-series heteroskedasticity in the scenario. Fifth, aside from MLFA-S, the convergence of the R^2 values is no longer monotonic in n.

Figure 4, shows the average R^2 values for Case 3 for all three factors and all three cross-correlation structures. In this scenario, idiosyncratic-return variance varies across time but is identical across assets. First, as expected, HFA outperforms the other three estimators. The performance differential is very small for factor one but increases as we extract additional factors. Second, the performance of the other three estimators is indistinguishable. Third, the improvement due to taking into account time-series heteroskedasticity is much smaller than the improvement due to taking into account cross-sectional heteroskedasticity. While this is a function of our modeling choices for both cross-sectional and time-series heteroscedasticity, our choices are meant to replicate the sample characteristics.

Figure 5 shows the average R^2 values for Case 4, in which idiosyncratic-return variance varies across time and across assets. First, MLFA-S (for all three factors) and APC-X (for factors two and three) dominate APC and HFA, particularly for small samples. The superior performance of MLFA-S and APC-X may be a function of the dispersion of idiosyncratic return variance in the crosssection versus in the time series. That is, a sample with greater volatility of volatility in the time series might lead to relatively better performance for HFA. However, our sample period includes the "great moderation" and the recent financial crisis of 2008–2009 and should provide substantial variation in volatility. Second, HFA dominates APC by a small margin for the first factor and by a slightly larger margin for factors two

and three.

We also calculate the t-statistics for the significance of the difference in average R^2 for the 6 pairwise comparisons of the estimators for each of the 12 cases considered here and each of the 13 choices of n. With 5,000 simulations, apparently the power of the test is quite large. For each of the 936 comparisons (6 pairwise comparisons of the 4 estimators $\times$ 4 heteroskedasticity cases $\times$ 3 values of ρ $\times$ 13 choices of n), the difference in average R^2 is statistically significant.

2.3 Unbalanced panel returns

For the unbalanced panel case, we apply five estimators, APC-M, APC-MX, HFA-M, MLFA-MS, and APC-EM. Figure 6 shows the average R^2 values for Case 1 for all three factors and all three cross-correlation structures. As before, the factors change across columns and the value of ρ changes across rows. First, the APC-M, APC-MX, and HFA-M estimators perform well and comparably even though two of the estimators are estimating extra parameters. Second, the MLFA-MX performs as well as APC-M, APC-MX, and HFA-M for the second and third factors. For the first factor, MLFA-MX performs similarly to APC-M, APC-MX, and HFA-M estimators for small values of n, but then converges to lower R^2 values similar to APC-EM for large values of n. Third, the APC-EM estimator seems to converge more slowly than the others. With 10,000 assets, the R^2 values for the APC-EM estimator are around 0.95 for all three factors.

Figure 7 shows the average R^2 values for Case 2 for all three factors and all three cross-correlation structures. In this scenario, idiosyncratic-return variance varies across assets but is constant through time. In this instance, one would expect that APC-MX and MLFA-MS would have superior performance since they explicitly take into account the differences in idiosyncratic risks across assets.

First, as in Figure 6, the behavior of MLFA-MS differs between factor one, on one hand, and factors two and three on the other. Second, the APC-MX and MLFA-MS estimators dominate the other procedures for factors two and three. For factor one, MLFA-MS dom-

inates for small samples and APC-MX dominates for intermediate samples. Third APC-M and HFA-M yield similar values of R^2, with APC-M performing slightly better for factor one and HFA-M better for factors two and three. Fourth, APC-EM converges more slowly than the other estimators.

Figure 8 shows the average R^2 values for Case 3 for all three factors and all three cross-correlation structures. In this scenario, idiosyncratic-return variance varies across time but is identical across assets. First, as expected, HFA-M outperforms the other estimators, although the performance differential is often relatively small. The estimates provided by AMP-M and APC-MX provide fits to the true factors that are either identical to (for factor one) or only slightly lower than (for factors two and three) HFA-M. Second, APC-EM and MLFA-MS converge more slowly than the other three estimators. The fit provided by MLFA-MS slightly declines (rather than increases) with n over the range $1000 < n < 8000$ for factor one.

Figure 9, shows the average R^2 values for Case 4, in which idiosyncratic-return variance varies across time and across assets. First, APC-M, APC-MX, and HFA-M estimators perform well and comparably for factor one, and MLFA-MS does well for factor one for small values of n. Second, for factors two and three, the ordering of these four estimators (in terms of the highest average R^2 values), is MLFA-MS, APC-MX, HFA-M, and APC-M. Third, APC-EM underperforms the other four estimators.

We also calculate the t-statistics for the significance of the difference in average R^2 for the 6 pairwise comparisons of the estimators for each of the 12 cases considered here and each of the 10 choices of n. With 5,000 simulations, apparently the power of the test is quite large. For each of the 936 comparisons (6 pairwise comparisons $\times$ 4 heteroskedasticity cases $\times$ 3 values of ρ $\times$ 13 choices of n), the difference in average R^2 is statistically significant.

3 Conclusion

We study the performance of a number of estimators of factor-mimicking portfolios by simulating asset returns under a variety of assumptions about the nature of cross-sectional and time-series heteroskedasticity, cross-correlation of idiosyncratic returns, and whether the data are from a balanced or unbalanced panel. When the data are from a balanced panel, all of the estimators perform similarly when there is no heteroskedasticity (Case 1), and cross-sectional sample sizes as small as 250 assets provide very accurate factor estimates. Cross-sectional heteroskedasticity (Case 2) leads to superior performance of the MLFA-S and APC-X estimators. The former provides very accurate estimates for sample sizes as small as 250, while the latter requires sample sizes between 3,000 and 5,000 for the R^2 values, relative to the true factor, to be above 0.975. APC and HFA require much larger samples (7,000 to 10,000) for the second and third factors. Time-series heteroskedasticity of the magnitude observed in the monthly data (Case 3) leads to superior performance of the HFA estimator, although all four estimators are accurate with sample sizes as small as 250. When there is both cross-sectional and time-series heteroskedasticity (Case 4), APC-X and MLFA-S provide the most accurate factor estimates, followed by HFA, and then APC.

When the data are from an unbalanced panel, APC-M, APC-MX, and HFA-M perform similarly when there is no heteroskedasticity (Case 5), and cross-sectional sample sizes as small as 500 assets provide very accurate factor estimates. The APC-EM procedure slightly underperforms, relative to the other estimators. MLFA-MS performs well for the second and third factors but relatively poorly for the first. Cross-sectional heteroskedasticity (Case 6) leads to superior performance of the APC-MX estimator for smaller sample sizes and for the MLFA-MS estimator for the second and third factors. AMP-MX provides very accurate estimates for sample sizes of 500 to 1,000, while APC-M and HFA-M require sample sizes around 2,000 (for factor one) to 6,000 (for factor three) to match APC-MX. Time-series heteroskedasticity of the magnitude observed in the monthly data (Case 7) leads to slightly superior performance of the HFA estimator relative to APC-M and APC-MX for the second and third factors. When there is both cross-sectional and time-series heteroskedasticity (Case

8), AMPC-MX, APC-M, and HFA-M perform equally well for the first factor, while for the second and third factors the ranking of performance is MLFA-MS, APC-MX, and HFA-M, followed by APC-M and APC-EM.

All of the estimators perform relatively well. Those allowing for heteroskedasticity do well when that form of heteroskedasticity is present and do not underperform by much if that form of heteroskedasticity is not present. Accommodating the type of cross-sectional heteroskedasticity present in the data yields greater improvement in factor estimates than accommodating time-series heteroskedasticity. A fairly consistent result is that the estimators that substitute fitted values from the factor model for missing data, APC-EM and MLFA-MX, converge more slowly than the other estimators.

REFERENCES

Andersen, Torben G., Tim Bollerslev, and Francis X. Diebold, (2010) "Parametric and Nonparametric Volatility Measurement," Chapter 2 in *Handbook of Financial Econometrics, Volume 1*, edited by Yacine Aït-Sahalia and Lars Peter Hansen. Amsterdam: North-Holland.

Bai, Jushan and Serena Ng. (2002) "Determining the Number of Factors in Approximate Factor Models," *Econometrica* **70**, 191–221.

Bai, Jushan and Serena Ng. (2008) "Large Dimensional Factor Analysis," *Foundations and Trends in Econometrics* **3**, 89–163.

Brennan, Michael J., (1971) "Capital Asset Pricing and the Structure of Security Returns," Working paper, University of British Columbia.

Campbell, John Y., Martin Lattau, Burton G. Malkiel, and Yexiao Xu, (2001) "Have Individual Stocks Become More Volatile? An Empirical Exploration of Idiosyncratic Risk." *Journal of Finance* **56**, 1−43.

Chamberlain, Gary, and Michael Rothschild, (1983) "Arbitrage, Factor Structure and Mean-Variance Analysis in Large Asset Markets," *Econometrica* **51**, 1305–1324.

Connor, Gregory, and Robert A. Korajczyk, (1986) "Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis," *Journal of Financial Economics* **15**, 323−346.

Connor, Gregory, and Robert A. Korajczyk, (1987) "Estimating Pervasive Economic Factors with Missing Observations," Working paper, http://ssrn.com/abstract=1268954.

Connor, Gregory, and Robert A. Korajczyk, (1988) "Risk and Return in an Equilibrium APT: Application of a New Test Methodology," *Journal of Financial Economics* 21: 255−290.

Connor, Gregory, and Robert A. Korajczyk, (1993) "A Test for the Number of Factors in an Approximate Factor Model," *Journal of Finance* **48**, 1263−1291.

Connor, Gregory, Robert A. Korajczyk, and Oliver Linton, (2006) "The common and specific components of dynamic volatility," *Journal of Econometrics* **132**, 231−255.

Dempster, A. P., N. M. Laird, D. B. Rubin, (1977) "Maximum Likelihood from Incomplete Data via the *EM* Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1–38.

Fama, Eugene F., and Kenneth R. French, (1993) "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics* **33**, 5−56.

Forni, Mario, Marc Hallin, Marco Lippi, and Lucrezia Reichlin, (2005) "The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting," *Journal of the American Statistical Association* **100,** 830–840.

Gagliardini, Patrick, and Christian Gourieroux, (2009) "Efficiency in Large Dynamic Panel Models with Common Factor," Working paper 09-12, Swiss Finance Institute.

Jones, Christopher S., (2001) "Extracting Factors from Heteroskedastic Asset Returns," *Journal of Financial Economics* **62**, 293−325.

Lehmann, Bruce N., and David M. Modest, (1988) "The Empirical Foundations of the Arbitrage Pricing Theory," *Journal of Financial Economics* **21**, 213−254.

Roll, Richard, and Stephen A. Ross, (1980) "An Empirical Investigation of the Arbitrage Pricing Theory," *Journal of Finance* **35**, 1073−1103.

Ross, Stephen A., (1976) "The arbitrage theory of capital asset pricing," *Journal of Economic Theory* **13**, 341–360.

Rubin, Donald B., and Dorothy T. Thayer, (1982) "*EM* Algorithms for ML Factor Analysis," *Psychometrika* **47**, 69−76.

Stock, James H., and Mark W. Watson, (1998) "Diffusion Indices," Working paper:

http://ssrn.com/abstract=226366.

Stock, James H., and Mark W. Watson, (2002) "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association, Theory and Methods* **97**, 1−13.

Stock, James H., and Mark W. Watson, (2006) "Forecasting with Many Predictors," *Handbook of Economic Forecasting* Volume 1, edited by G. Elliott, C.W.J. Granger and A. Timmermann. Amsterdam: Elsevier, pp 515–554.

Stroyny, Alvin L., (1992) "Still More on *EM* Factor Analysis," Working Paper, University of Wisconsin, Milwaukee, WI.