A Performance Comparison of Large-*n* Factor Estimators

Zhuo Chen
PBC School of Finance, Tsinghua University

Gregory Connor
National University of Ireland, Maynooth

Robert A. Korajczyk
Kellogg School of Management, Northwestern University

We evaluate the performance of various methods for estimating factor returns in an approximate factor model. Differences across estimators are most pronounced when there is cross-sectional heteroscedasticity or when cross-sectional sample sizes, *n*, have fewer than 4,000 assets. Estimators incorporating either cross-sectional or time-series heteroscedasticity outperform the other estimators when those types of heteroscedasticity are present. The differences are most pronounced when the cross-sectional sample is small. (*JEL* G10, G12, C15, C23)

Received December 2, 2015; editorial decision May 16, 2017 by Editor Jeffrey Pontiff.

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 are 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

We thank Sara Filipe, Patrick Gagliardini, Eric Ghysels, Alex Horenstein, Egon Kalotay, Maja Kos, Jeffrey Pontiff, Eric Renault, Stephen Satchell, and Alvin Stroyny; an anonymous referee; and participants at the Society for Financial Econometrics conference on Large Scale Factor Models and the Multinational Finance Society conference for helpful comments. Send correspondence to Robert A. Korajczyk, Kellogg School of Management, Northwestern University, 2211 Campus Drive, Evanston, IL 60208; telephone: (847) 491-8336. E-mail: r-korajczyk@kellogg.northwestern.edu.

doi:10.1093/rapstu/rax017 Advance Access publication May 29, 2017

approximate factor model. The approximate factor model framework has been used in a wide range of applications. In addition to common stock return modeling (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 et al. 2005), and credit models (e.g., Gagliardini and Gourieroux 2014). 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 using large-n methods. We calibrate our simulation model based on the observed features of U.S. 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 heteroscedasticity 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.

Differences across estimators are most pronounced when there is cross-sectional heteroscedasticity and cross-sectional sample sizes, n, have fewer than 4,000 stocks. For very large n, the estimators generally perform similarly. Estimators that explicitly incorporate either cross-sectional or time-series heteroscedasticity outperform the other estimators when those types of heteroscedasticity are present for the balanced sample. With both cross-sectional and time-series heteroscedasticity, as well as an unbalanced panel, Connor and Korajczyk's (1988) and Jones's (2001) methods, which accommodate cross-sectional and time-series heteroscedasticity, respectively, provide the most accurate factor return estimates. Many empirical studies and simulations in the literature use cross-sectional sample sizes in the range for which estimators incorporating heteroscedasticity lead to improvements in the precision of the factor estimates.

1. Large-*n* Estimators of Factor-Mimicking Portfolios

We assume that the data-generating process for returns on all securities is an approximate k-factor model. We also assume that asset risk premiums are linear in factor betas. 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 premiums at period t. Let ϵ_{t} be an n-vector of idiosyncratic returns, and let r_{t} denote the n-vector of asset returns. We assume that a k-factor equilibrium asset pricing model holds so that

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

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}, $E[\epsilon_{\mathrm{t}}|f_{\mathrm{t}}] = 0$. 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 $t = 1, 2, \ldots T$, define R to be the $n \times T$ matrix of realized excess returns on the n securities for T time periods: $R = [R_1 R_2 \ldots R_T]$. 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 premiums, 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 implement and compare in our study.

1.1 Asymptotic principal components

For $n \gg T$ (e.g., 10,000 assets over a 60-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 $T = 60$, the $n \times n$ covariance matrix has over 50 million distinct entries, but we have only 600,000 data points). Connor and Korajczyk (1986) derive asymptotic principal components (APC) as a 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 $\hat{F}$ 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, $\hat{F}$ is an n-consistent estimate of F. They call this estimator the asymptotic principal components estimator. This estimator makes no assumptions about cross-sectional heteroscedasticity in the idiosyncratic returns (the diagonal elements of V) other than the boundedness discussed above. However, it does not attempt to utilize any such heteroscedasticity in estimation. The estimator makes fairly restrictive assumptions about any time-series heteroscedasticity 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 heteroscedasticity as well as unbalanced panels.

1.2 Incorporating cross-sectional heteroscedasticity

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

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

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

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

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

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 heteroscedasticity. The APC-X procedure is a variant of weighted principal components (Stock and Watson 2006, section 4.3). The APC-X procedure is also an example of feasible generalized principal components estimation (FGPCE) discussed by Choi (2012) (see example 1 on p. 286).

Stroyny (1992) proposes a large-n variant of maximum-likelihood factor analysis based on the *EM* algorithm (Dempster, Laird, and Rubin 1977; 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 at each iteration 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 Stroyny's procedure (for $n = 5{,}000$, the standard *EM* estimator 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, or maximum likelihood factor analysis, using Stroyny's (1992) procedure.

1.3 Incorporating time-series heteroscedasticity

Factor analysis generally assumes that each asset's idiosyncratic volatility 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 (e.g., Andersen, Bollerslev, and Diebold 2010), and idiosyncratic volatility, in particular (e.g, Campbell et al. 2001; Connor, Korajczyk, and Linton 2006), it seems that incorporating times-series heteroscedasticity into factor estimation is desirable. Jones (2001) proposes such an estimator, called heteroscedastic factor analysis (HFA). The HFA procedure is a variant of weighted principal components (Stock and Watson 2006, section 4.3; Boivin and Ng 2006, p. 186). Jones (2001) assumes that the cross-sectional average idiosyncratic volatility is time dependent:

$$\bar{\Sigma}_t = \underset{n\to\infty}{\text{plim}} \frac{1}{n}\Sigma_{i,i,t},$$

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, $\bar{\Sigma}$, to be the diagonal matrix with elements $(t, t) = \bar{\Sigma}_t$. Jones's procedure estimates factor returns by calculating eigenvectors of the scaled matrix,

$$\hat{\bar{\Sigma}}^{-1/2}\Omega\hat{\bar{\Sigma}}^{-1/2}. \tag{7}$$

These factor estimates are used to reestimate idiosyncratic returns and $\hat{\bar{\Sigma}}$, 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; Jones 2001). However, requiring a balanced panel may induce survivorship bias into the sample. Several alternative approaches are available for 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 all the 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 in which 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 heteroscedasticity 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 and which adjusts for cross-sectional heteroscedasticity.

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):

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

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

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

and

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

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

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

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,

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

where $R^{**,j-1}_{i,t} = R_{i,t}$ if $\delta_{i,t} = 1$ and $R^{**,j-1}_{i,t} = B^{j-1}_{i,\cdot} F^{j-1}_{\cdot,t}$ if $\delta_{i,t} = 0$ (see Stock and Watson 1998, page 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 non-singular rotation L) the APC estimate obtained from $R^{**,j-1}_{i,t}$. Applying the *EM* algorithm amounts to an iterative application of APC until convergence. In the case in which there are no missing data, Stock and Watson's (1998, 2002) estimator is identical to Connor and Korajczyk's (1986) estimator. We call Stock and Watson's (1998) estimator APC-EM to denote its use of the EM algorithm.

Jones (2001) suggests extending the HFA procedure along the lines of Connor and Korajczyk (1987), in which Ω and $\bar{V}$ 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 varying numbers of assets and study the behavior, as n increases, of the factor estimates relative 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 heteroscedasticity 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 heteroscedasticity: $E[\epsilon_t\epsilon_t'] = \sigma_t^2 I$, where I is an $n \times n$ identity matrix; and (4) cross-sectional and time-series heteroscedasticity: $E[\epsilon_t\epsilon_t'] = D_t$, where D_t is an $n \times n$ 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 analysis the assumption that the data are missing at random. 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. We simulate economies in which the idiosyncratic returns are normally and student-t distributed. This gives us 96 different cases to simulate (four cases regarding cross-sectional and time-series heteroscedasticity $\times$ two cases with complete or missing data $\times$ three cases regarding cross-asset correlation in idiosyncratic returns $\times$ two alternative lengths of time series, 60 and 120 months $\times$ two cases of normal and student-t idiosyncractic returns).

2.1 Simulation design

Our sample period is 1976 to 2015. Each simulation uses either $T = 60$ or 120 to correspond to five- and ten-year periods of monthly data. The parameters of the simulations are calibrated by choosing one of the five- or ten-year nonoverlapping time periods, resampling firms with available data on the Center for Research in Security Prices (CRSP) stock database from that time period, and computing simulation parameters based on the sampled stocks (the sampling is discussed in more detail below).

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 1976 to 2015 period, included in the CRSP indices for the New York (NYSE), American (AMEX), and NASD Stock Exchanges. The table also includes similar

Table 1
Numbers of listed equities by exchange

Exchange	Period	Minimum	Average	Maximum
NYSE	1976–2015	1,489	2,144	2,870
NYSE+AMEX	1976–2015	2,248	2,920	3,740
NYSE+AMEX+NASDAQ	1976–2015	4,675	6,538	9,047
Australia	1996–2014	1,147	1,624	1,988
Brazil	1996–2014	336	417	557
Canada (TMX Group)	1996–2014	1,266	3,031	3,937
China (Shanghai)	1996–2014	720	871	974
China (Shenzhen)	1996–2014	496	916	1,595
Deutsche Bőrse	1996–2014	610	701	1,043
India (Bombay)	1996–2014	4,721	5,060	5,798
India (NSE)	1996–2014	810	1,327	1,698
Hong Kong	1996–2014	649	1,130	1,631
Poland (Warsaw)	1996–2014	148	393	872
United Kingdom (London)	1996–2014	2,152	2,319	2,693

Observations are at monthly intervals chosen to correspond to the period to which parameters in the simulation are calibrated, subject to data availability. The table reports the minimum, average, and maximum number of firms listed. Data for the NYSE, AMEX, and NASD exchanges are from the Center for Research in Security Prices (CRSP). Data for the other exchanges are from the World Federation of Exchanges.

figures for various exchanges over the 1996 to 2014 period, which are obtained from the World Federation of Exchanges.[1] The combined NYSE, AMEX, and NASD markets have a minimum of 4,675 and a maximum of 9,047 firms. The equivalent figures are 4,721 and 5,798 for the Bombay Stock Exchange, 2,152 and 2,693 for the London Stock Exchange, 720 and 974 for Shanghai, 1,266 and 3,937 for Canada, 336 and 557 for Brazil, and 148 and 872 for Poland. Thus, our range of 250 to 10,000 firms 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 draws 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 (e.g., Connor and Korajczyk 1993; Bai and Ng 2002). That is, we simulate a three-factor model and estimate three factors.

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:

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

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

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

intercepts equal to zero. We simulate the return matrix R (dimension $n \times T$) using Fama and French's (1993) three-factor model.[2] The true factor matrix, F, consist of the three factors, R_m, HML, and SMB.

2.1.1 Factor loadings. The simulated beta loading matrix B (dimension $n \times k$) is generated based on the empirical distribution of stocks' loadings on Fama-French's factors. For each common stock traded in NYSE/NASDAQ/AMEX with more than 36 months of observations over the relevant five- or ten-year estimation period, we estimate a time-series regression of stock excess returns on Fama-French's factors and calculate the estimated factor loadings.

2.1.2 Idiosyncratic returns. Similarly, we rely on the estimated residuals from Fama-French's three-factor model applied to the CRSP sample stocks over the sample period to define the properties of the simulated idiosyncractic returns. Let $\hat{\sigma}_i$ be the estimated standard deviation of the idiosyncratic return of asset i over the relevant estimation period. Some of these estimates are implausibly large, especially for stocks with a short time series of observations (e.g., $\hat{\sigma}_i$ on the order of 300%). We winsorize the sample of estimated idiosyncratic risk at the 99% level. That is, for any stock, i, with $\hat{\sigma}_i > \hat{\sigma}_{99\%}$ (where $\hat{\sigma}_{99\%}$ is the 99th percentile of the cross-sectional distribution of idiosyncratic risk), we set $\hat{\sigma}_i$ equal to $\hat{\sigma}_{99\%}$. Also, we estimate the average idiosyncratic volatility in period t by

$$\hat{\sigma}_t^2 = \sum_{\substack{i=1,n_t \\ i \in W}} \frac{\hat{\epsilon}_{i,t}^2}{n_t},$$

where $\hat{\epsilon}_{i,t}$ is the estimated idiosyncratic return on asset i in period t, and $i \in W$ denotes that the squared idiosyncratic returns have been winsorized at the 99% quantile.

Heteroscedasticity Case 1: When we assume both cross-sectional and time-series homoscedasticity and no cross-correlation, we construct idiosyncratic returns that are drawn from a normal distribution (Case 1a) or a t distribution with degrees of freedom, ν, equal to five, the average across sample stocks (Case 1b). Idiosyncratic returns have a mean of zero and a standard deviation, $\bar{\sigma}$, equal to the average value (across sample stocks) of $\hat{\sigma}_i$. When we have cross-correlation, the idiosyncratic return for asset 1 in period t, $\epsilon_{1,t}$, is drawn from these distributions and the remaining idiosyncratic returns are constructed as $\epsilon_{i,t} = \rho\epsilon_{i-1,t} + u_{i,t}$, where $u_{i,t} \sim N(0, (1-\rho^2)\bar{\sigma}^2)$ or $u_{i,t} \sim t(0, (1-\rho^2)\bar{\sigma}^2, \nu = 5)$. This gives each asset an unconditional idiosyncratic standard deviation of $\bar{\sigma}$. This is done independently for each time period, t.

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

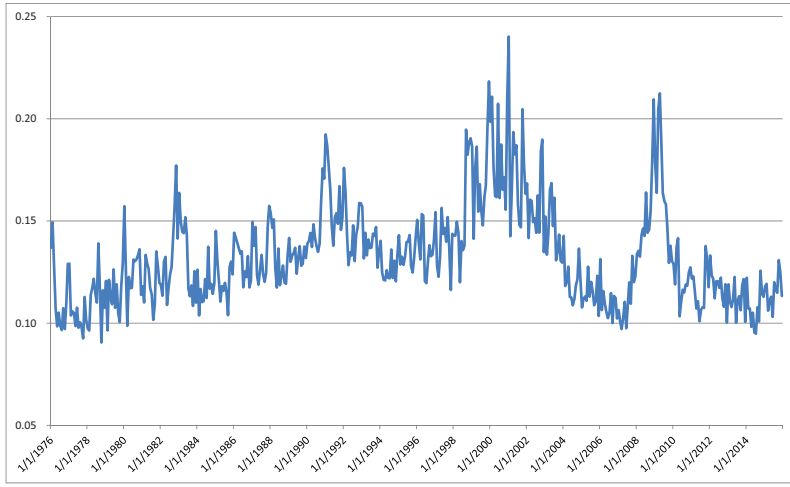

Figure 1
Time series of the cross-sectional average squared idiosyncratic return from Fama-French's model CRSP firms over the period from January 1976 to December 2015.

Heteroscedasticity Case 2: When we assume cross-sectional heteroscedasticity but time-series homoscedasticity, we randomly pick n empirical standard deviations, with replacement, calculated from Fama-French's three-factor regression residuals, $\hat{\sigma}_i$, from the sample stocks. We generate the residual matrix from a normal (Case 2a) and student-t distribution (Case 2b). The idiosyncratic returns have a mean 0 and standard deviation $\hat{\sigma}_i$. If we have cross-sectional correlation, the idiosyncratic return for asset 1 in period t, $\epsilon_{1,t}$, is drawn from a $N(0, \hat{\sigma}_1^2)$ or $t(0, \hat{\sigma}_1^2, \nu = 5)$ distribution, and the remaining idiosyncratic returns are constructed as $\epsilon_{i,t} = \rho\epsilon_{i-1,t} + u_{i,t}$, where $u_{i,t} \sim N(0, \hat{\sigma}_i^2 - \rho^2\hat{\sigma}_{i-1}^2)$ or $t(0, \hat{\sigma}_i^2 - \rho^2\hat{\sigma}_{i-1}^2, \nu = 5)$.

Heteroscedasticity Case 3: When we assume time-series heteroscedasticity but cross-sectional homoscedasticity, every asset's idiosyncratic return in period t is drawn from either a normal (Case 3a) or student-t (Case 3b) distribution with a standard deviation equal to $\hat{\sigma}_t$ as calculated above. Figure 1 plots the time series of $\hat{\sigma}_t$ over our sample period. There is substantial variation in $\hat{\sigma}_t$ through time. When we have cross-correlation, the idiosyncratic return for asset 1 in period t, $\epsilon_{1,t}$, is drawn from a $N(0, \hat{\sigma}_t^2)$ or $t(0, \hat{\sigma}_t^2, \nu = 5)$ distribution, and the remaining idiosyncratic returns are constructed as $\epsilon_{i,t} = \rho\epsilon_{i-1,t} + u_{i,t}$, where $u_{i,t} \sim N(0, (1-\rho^2)\hat{\sigma}_t^2)$ or $t(0, (1-\rho^2)\hat{\sigma}_t^2, \nu = 5)$.

Heteroscedasticity Case 4: When we assume both time-series and cross-sectional heteroscedasticity, we assume that the idiosyncratic variance of each of the sample 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,

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

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 $\hat{\sigma}_t^2$), consistent with the evidence of Connor, Korajczyk, and Linton (2006). Every asset's idiosyncratic return in period t is drawn from either a $N(0, \hat{\theta}_i \hat{\sigma}_t^2)$ (Case 4a) or $t(0, \hat{\theta}_i \hat{\sigma}_t^2, \nu = 5)$ (Case 4b) distribution. When we have cross-correlation, the idiosyncratic return for asset 1 in period t, $\epsilon_{1,t}$, is drawn from a $N(0, \hat{\theta}_1 \hat{\sigma}_t^2)$ or $t(0, \hat{\theta}_1 \hat{\sigma}_t^2, \nu = 5)$ distribution, and the remaining idiosyncratic returns are constructed as $\epsilon_{i,t} = \rho \epsilon_{i-1,t} + u_{i,t}$, where $u_{i,t} \sim N(0, \hat{\theta}_i \hat{\sigma}_t^2 - \rho^2 \hat{\theta}_{i-1} \hat{\sigma}_t^2)$ or $N(0, \hat{\theta}_i \hat{\sigma}_t^2 - \rho^2 \hat{\theta}_{i-1} \hat{\sigma}_t^2, \nu = 5)$.[3]

Over our sample period there are eight nonoverlapping 60-month periods, (1976–1980, 1981–1985, . . ., 2011–2015) and four nonoverlapping 120-month periods (1976–1985,1986–1995, 1996–2005, and 2006–2015). We discuss the results for the 60-month periods here and report the results for 120-month periods in the Internet Appendix. We require a stock to have 36 months of data within a subperiod to be included in the sample. Over the eight 60-month periods, there are 4,278, 4,606, 5,462, 5,433, 6,237, 4,700, 4,123, and 3,416 CRSP firms that meet the 36-month data requirement. Over the four 120-month periods, there are 5,849, 7,345, 7,679, and 4,735 CRSP firms that meet the 36-month data requirement. We simulate each hypothetical economy 5,000 times. For simulations of 60-month sample periods, we use Fama-French's factors from each of the eight 60-month periods for 625 (5,000/8) simulations, for a total of 5,000 simulations. For simulations of 120-month sample periods, we use Fama-French's factors from each of the 120-month periods for 1,250 (5,000/4) simulations. For each run of the simulation, we draw the $n \times k$ matrix of factor loadings and the corresponding idiosyncratic volatility from the estimated values, with replacement. Given the idiosyncratic volatility estimate, we generate the idiosyncratic returns by using either the normal or t distributions described above. Given these, we generate an $n \times T$ return matrix $R = FB + \epsilon$ and apply the factor estimators to the returns for cross-sectional samples of n = 250, 500, 750, 1,000, 2,000, 3,000, …, and 10,000. For each of our 96 case combinations, and 5,000 simulations, we regress $\hat{F}$ on the true F, and record the adjusted R^2, the estimated intercept, $\hat{\alpha}$, and the associated t-statistic for the intercept.

Cases 3 and 4, which allow for time-series heteroscedasticity, also preserve any conditional heteroscedasticty of idiosyncratic volatility, Table 2 shows the results of a regression of $\hat{\sigma}_t^2$ on the cross-products of Fama-French's factor realizations for each 60-month period as well as for the full 480-month

[3] For Heteroscedasticity 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%.

Table 2
Conditional heteroscedasticity, 60-month subsample regressions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Intercept	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	(23.44)	(25.07)	(27.32)	(34.76)	(18.39)	(14.36)	(9.13)	(31.85)	(42.23)
$MKTRF^2$	−0.10	0.31	0.04	0.20	0.03	0.23	0.27	0.21	0.21
	(−0.71)	(3.23)	(0.43)	(1.16)	(0.14)	(1.30)	(1.68)	(3.14)	(2.30)
SMB^2	0.24	1.37	−0.04	0.15	−0.30	1.21	1.87	−0.28	−0.28
	(0.70)	(4.73)	(−0.12)	(0.81)	(−1.00)	(3.43)	(2.14)	(−1.26)	(2.05)
HML^2	−0.98	−0.90	−9.65	2.61	−4.19	−3.84	1.94	−15.14	−15.14
	(−0.26)	(−0.34)	(−1.44)	(0.75)	(−2.47)	(−2.18)	(0.48)	(−2.54)	(5.56)
MKTRF*SMB	−0.01	1.37	−0.15	1.58	−0.83	−0.15	−0.62	−0.15	−0.15
	(−0.01)	(4.69)	(−0.79)	(5.66)	(−2.22)	(−0.5)	(−0.85)	(−0.80)	(−1.28)
MKTRF*HML	0.10	0.70	−0.05	0.01	−0.74	−0.55	0.89	0.08	0.08
	(0.31)	(2.72)	(−0.17)	(0.03)	(−2.05)	(−1.66)	(2.21)	(0.37)	(2.20)
SMB*HML	0.42	1.15	0.12	0.54	−0.63	−0.65	0.44	−0.86	−0.86
	(1.01)	(3.13)	(0.23)	(2.12)	(−1.23)	(−1.59)	(0.42)	(−32.69)	(0.00)
F-test	1.20	9.38	0.90	6.29	2.74	7.29	4.64	3.77	14.35
Prob(F>Ftest|H_0)	0.68	1.00	0.50	1.00	0.98	1.00	1.00	1.00	1.00
p-value	0.32	0.00	0.50	0.00	0.02	0.00	0.00	0.00	0.00
R^2	0.12	0.51	0.09	0.42	0.24	0.45	0.34	0.30	0.15
Adj R^2	0.02	0.46	−0.01	0.35	0.15	0.39	0.27	0.22	0.14
Sample	1976–1980	1981–1985	1986– 1990	1991–1995	1996–2000	2001–2005	2006–2010	2011–2015	1976–2015

We regress $\widehat{\sigma}_t^2$ on the squared values of the factors and their crossproducts ($F_{.,t}F'_{.,t}$) plus an intercept. *t*-statistics are reported in parentheses. The F-test (and associated *p*-value) is for the hypothesis that all slope coefficients are zero, that is, no conditional heteroscedasticity.

period (results for the 120-month periods are reported in the Internet Appendix). In six of the eight 60-month periods there is statistically significant conditional heteroscedasticity (at the 5% level), as can be seen from the F-test for joint significance of the factor cross-products. Thus, the simulation design preserves the conditional heteroscedasticity that exists in the data.

For the estimators that require iteration to convergence, that is HFA, MLFA-S, HFA-M, 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 use the same pattern of missing observations as observed in the data to generate a simulated return series with missing values. That is, we sample jointly from $B_{i,\cdot}$, $\hat{\sigma}_i$ and the pattern of missing observations.

2.2 Balanced panel of asset returns: Normally distributed idiosyncratic returns

For the balanced-panel case, we apply four estimators, APC, APC-X, HFA, and MLFA-S. We discuss the 60-month samples here and relegate the 120-month samples to the Internet Appendix, since the results are quite similar. We first discuss the results with normally distributed idiosyncratic returns and then turn to the results with t-distributed idiosyncratic returns. Figure 2 shows the average (across the 5,000 simulations) R^2 values for Case 1a for all three factors and two cross-correlation structures ($\rho = 0.0,0.5$). Our figures report the results for $n = 250$, 500, 1,000, 2,000, 5,000, and 10,000 (the full results appear in the Internet Appendix). The factors change across columns, and the value of ρ changes across rows. For ease of comparison, since scales can vary across graphs, the dashed horizontal line in each graph is at an R^2 value of 0.95. Several points are clear from the figure. First, all four estimators perform comparably even though three of the estimators are estimating extra parameters. In fact, it is often difficult to make out any difference between the estimators for samples of 500 or more stocks (see the Internet Appendix for exact numerical values for all charts). Second, accuracy falls for higher-order 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 0.8, even for the estimates of the third factor, with $\rho = 0.5$, and with the smallest number of assets in the cross-section ($n = 250$). When we have 2,000 assets in the cross-section, almost all mean R^2 values equal 0.98 or higher.

Figure 3 shows the average R^2 values for Case 2a for all three factors and two cross-correlation structures ($\rho = 0.0,0.5$). 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.

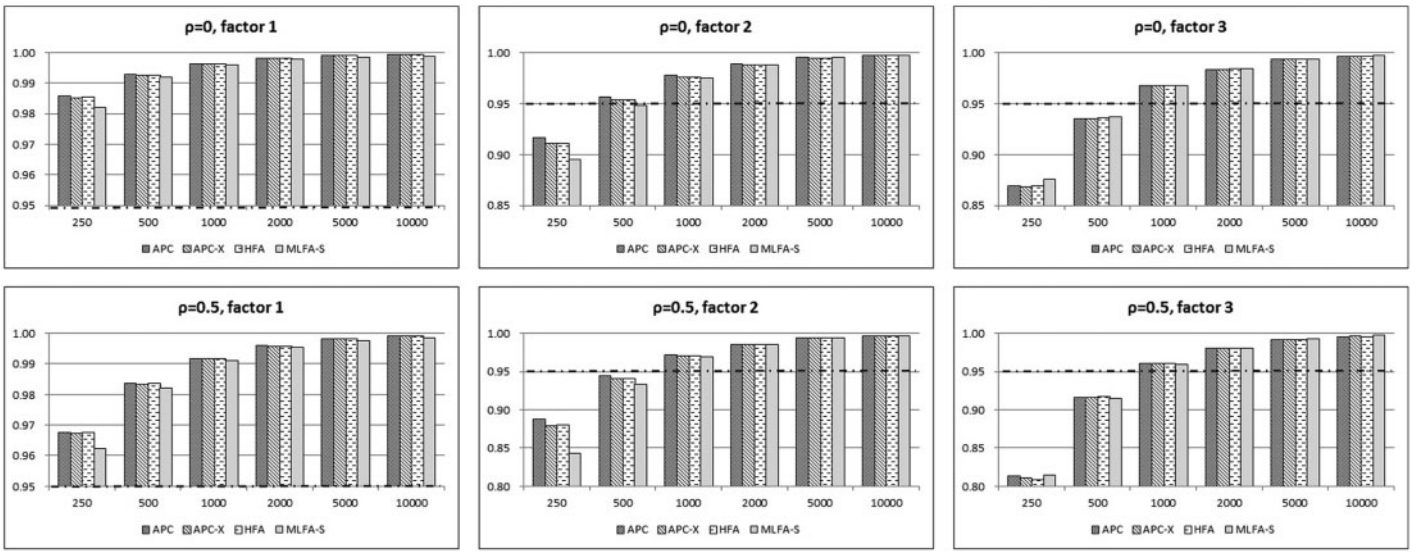

Figure 2
R^2 values from a regression of estimated factors on true factors: Case 1a, cross-sectional and time-series homoscedasticity
Balanced sample, normally distributed idiosyncratic errors, 60-month estimation period. The estimators are Connor and Korajczyk's (1986) asymptotic principal components (APC); Connor and Korajczyk's (1988) version of weighted principal components (APC-X) that accommodates cross-sectional heteroscedasticity; Jones's (2001) heteroskedastic factor analysis (HFA) that accommodates time series idiosyncratic heteroscedasticity; and Stroyny's (1992) maximum likelihood factor analysis (MLFA-S).

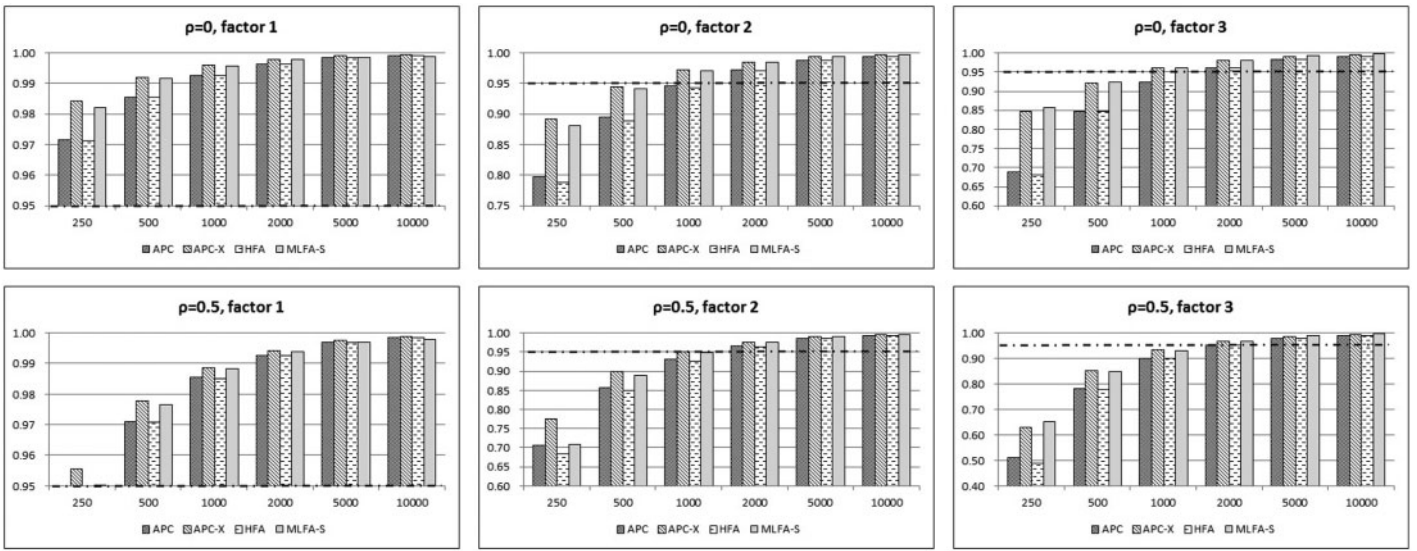

Figure 3
R^2 values from a regression of estimated factors on true factors: Case 2a, cross-sectional heteroscedasticity and time-series homoscedasticity
Balanced sample, normally distributed idiosyncratic errors, 60-month estimation period. The estimators are Connor and Korajczyk's (1986) asymptotic principal components (APC); Connor and Korajczyk's (1988) version of weighted principal components (APC-X) that accommodates cross-sectional heteroscedasticity; Jones's (2001) heteroskedastic factor analysis (HFA) that accommodates time series idiosyncratic heteroscedasticity; and Stroyny's (1992) maximum likelihood factor analysis (MLFA-S).

Again, there are several points that are clear from the figure. First, APC-X and the procedure from Stroyny (1992), MLFA-S, dominate the other procedures, until we reach values of n ranging from 2,000 to 5,000. Second, with the exception of APC-X and MLFA-S, cross-sectional heteroscedasticity significantly slows the convergence (in n) of the factor estimates to the true factors. While under Case 1a the R^2s are 0.8 and higher, under Case 2a, the R^2 values are as low as 0.5 and need approximately 2,000 to 3,000 assets for the second and third factors to attain minimum R^2 values above 0.975. Third, APC and HFA are essentially equivalent, which would be expected given that there is no time-series heteroscedasticity in the scenario.

Figure 4, shows the average R^2 values for Case 3a for all three factors and two cross-correlation structures ($\rho = 0.0,0.5$). In this scenario, idiosyncratic-return variance varies across time but is identical across assets. First, as expected, HFA outperforms the other three estimators for factors two and three. The performance differential is very small for factor one but increases slightly as we extract additional factors. Second, the performance of the other three estimators is indistinguishable.

Figure 5 shows the average R^2 values for Case 4a, in which idiosyncratic-return variance varies across time and across assets. MLFA-S (for all three factors) and APC-X (for factors two and three) dominate APC and HFA for small cross-sectional samples (n). The superior performance of MLFA-S and APC-X may be a function of the dispersion of idiosyncratic-return variance in the cross-section 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 five NBER dated recessions, including the recent financial crisis of 2008–2009, and should provide substantial variation in volatility.

2.3 Balanced panel of asset returns: *t*-distributed idiosyncratic returns

Figure 6 shows the average (across the 5,000 simulations) R^2 values for Case 1b for all three factors and two cross-correlation structures ($\rho = 0.0,0.5$). Several points are clear from the figure. As in the case with normally distributed idiosyncratic returns, all four estimators perform comparably even though three of the estimators are estimating extra parameters. In fact, it is often difficult to make out any difference between the estimators for samples of 500 or more stocks. Second, accuracy falls for higher-order factors and as the idiosyncratic-return correlation across assets increases. Third, having leptokurtic idiosyncratic returns reduces the accuracy of the estimators, particularly for smaller values of n. The smallest mean R^2 values exceed 0.65, even for the estimates of the third factor, with $\rho = 0.5$, and with the smallest number of assets in the cross-section ($n = 250$). When we have 2,000 assets in the cross-section, almost all mean R^2 values equal 0.95 or higher.

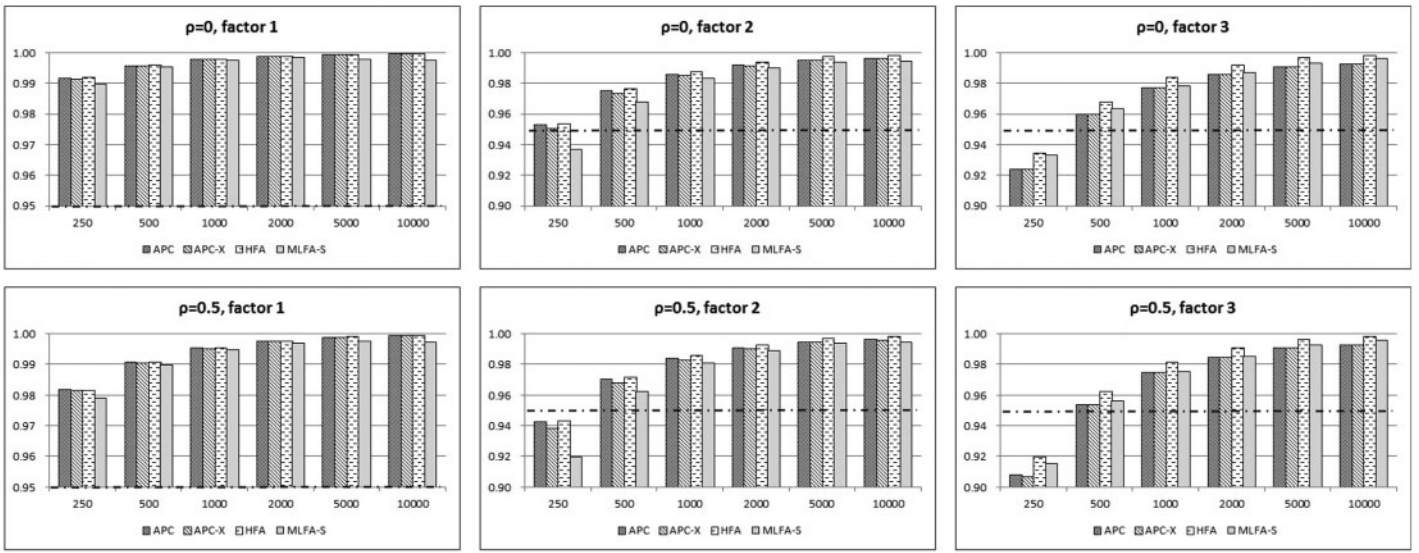

Figure 4
R^2 values from a regression of estimated factors on true factors: Case 3a, cross-sectional homoscedasticity and time-series heteroscedasticity
Balanced sample, normally distributed idiosyncratic errors, 60-month estimation period. The estimators are Connor and Korajczyk's (1986) asymptotic principal components (APC); Connor and Korajczyk's (1988) version of weighted principal components (APC-X) that accommodates cross-sectional heteroscedasticity; Jones's (2001) heteroskedastic factor analysis (HFA) that accommodates time series idiosyncratic heteroscedasticity; and Stroyny's (1992) maximum likelihood factor analysis (MLFA-S).

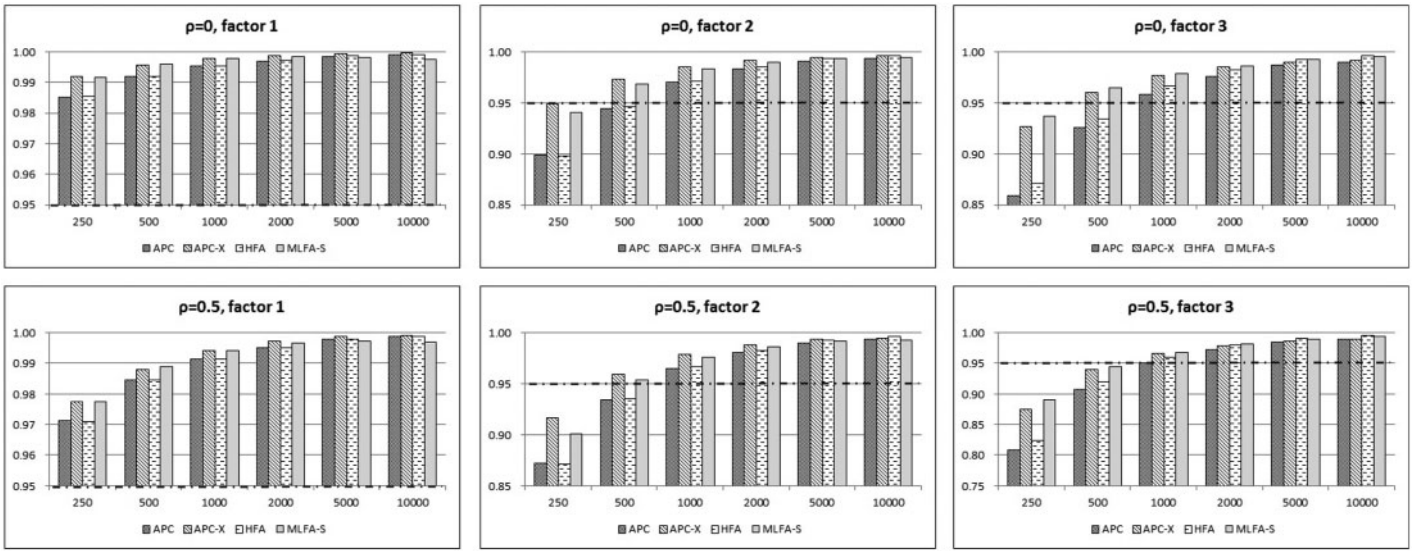

Figure 5
R^2 values from a regression of estimated factors on true factors: Case 4a, cross-sectional and time-series heteroscedasticity
Balanced sample, normally distributed idiosyncratic errors, 60-month estimation period. The estimators are Connor and Korajczyk's (1986) asymptotic principal components (APC); Connor and Korajczyk's (1988) version of weighted principal components (APC-X) that accommodates cross-sectional heteroscedasticity; Jones's (2001) heteroskedastic factor analysis (HFA) that accommodates time series idiosyncratic heteroscedasticity; and Stroyny's (1992) maximum likelihood factor analysis (MLFA-S).

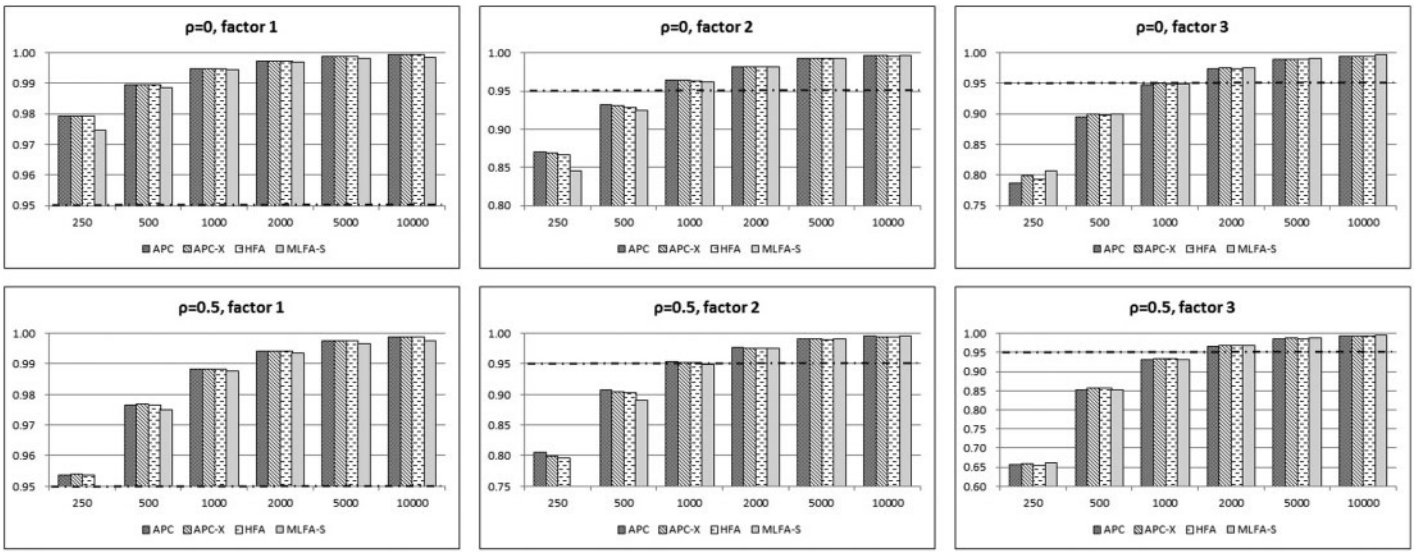

Figure 6
R^2 values from a regression of estimated factors on true factors: Case 1b, cross-sectional and time-series homoscedasticity
Balanced sample, *t*-distributed idiosyncratic errors, 60-month estimation period. The estimators are Connor and Korajczyk's (1986) asymptotic principal components (APC); Connor and Korajczyk's (1988) version of weighted principal components (APC-X) that accommodates cross-sectional heteroscedasticity; Jones's (2001) heteroskedastic factor analysis (HFA) that accommodates time series idiosyncratic heteroscedasticity; and Stroyny's (1992) maximum likelihood factor analysis (MLFA-S).

Figure 7 shows the average R^2 values for Case 2b for all three factors and two cross-correlation structures ($\rho = 0.0, 0.5$). 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.

As in Figure 3, APC-X and the procedure from Stroyny (1992), MLFA-S, dominate the other procedures, until we reach values of n around 5,000. Second, cross-sectional heteroscedasticity significantly slows the convergence of the factor estimates to the true factors. While under Case 1b, the R^2s are 0.65 and higher, under Case 2b, the R^2 values are as low as 0.3 and need approximately 1,000 to 5,000 assets for the second and third factors to attain minimum R^2 values above 0.975. Fourth, APC and HFA are essentially equivalent, which would be expected given that there is no time-series heteroscedasticity in the scenario.

Figure 8, shows the average R^2 values for Case 3b for all three factors and two cross-correlation structures ($\rho = 0.0, 0.5$). In this scenario, idiosyncratic-return variance varies across time but is identical across assets. First, as expected, HFA outperforms the other three estimators for factors two and three and outperforms MLFA-S for factor 1. Second, the performance of MLFA-S declines in n for the first factor.

Figure 9 shows the average R^2 values for Case 4b, in which idiosyncratic-return variance varies across time and across assets. MLFA-S (for all three factors) and APC-X (for factors two and three) dominate APC and HFA for small cross-sectional samples (n). For large values of n, HFA performs slightly better for factors two and three.

2.4 Unbalanced panel returns: Normally distributed idiosyncratic returns

The relative comparisons across the alternative cases of heteroscedasticity for balanced panels, discussed above, gives a good sense for the effect of changing assumptions about the form of heteroscedasticity on the performance of alternative estimators. To conserve space, we only discuss the most realistic case, in which there is both cross-sectional and time-series heteroscedasticity (Case 4a). The full results are available in the Internet Appendix. For the unbalanced-panel case, we apply four estimators, APC-M, APC-MX, HFA-M, and APC-EM. Figure 10 shows the average R^2 values for Case 4a for all three factors and two cross-correlation structures ($\rho = 0.0,0.5$). First, APC-MX outperforms the other three estimators for low values of n. For factors one and two, APX-MX outperforms for all values of n and both $\rho = 0.0$ and $\rho = 0.5$. For factor three, APC-MX outperforms for values of n less than or equal to 3, 000 ($\rho = 0.0$) or values of n less than or equal to 2, 000 ($\rho = 0.5$). Second, for factor three, HFA is the best estimator in those

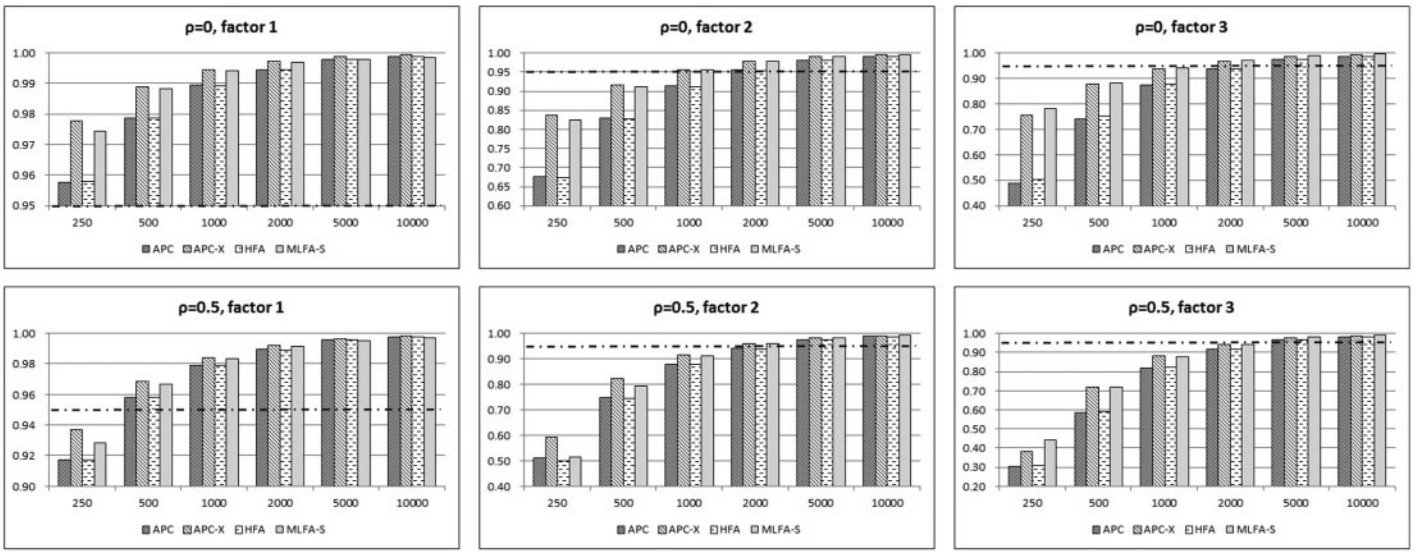

Figure 7
R^2 values from a regression of estimated factors on true factors: Case 2b, cross-sectional heteroscedasticity and time-series homoscedasticity
Balanced sample, *t*-distributed idiosyncratic errors, 60-month estimation period. The estimators are Connor and Korajczyk's (1986) asymptotic principal components (APC); Connor and Korajczyk's (1988) version of weighted principal components (APC-X) that accommodates cross-sectional heteroscedasticity; Jones's (2001) heteroskedastic factor analysis (HFA) that accommodates time series idiosyncratic heteroscedasticity; and Stroyny's (1992) maximum likelihood factor analysis (MLFA-S).

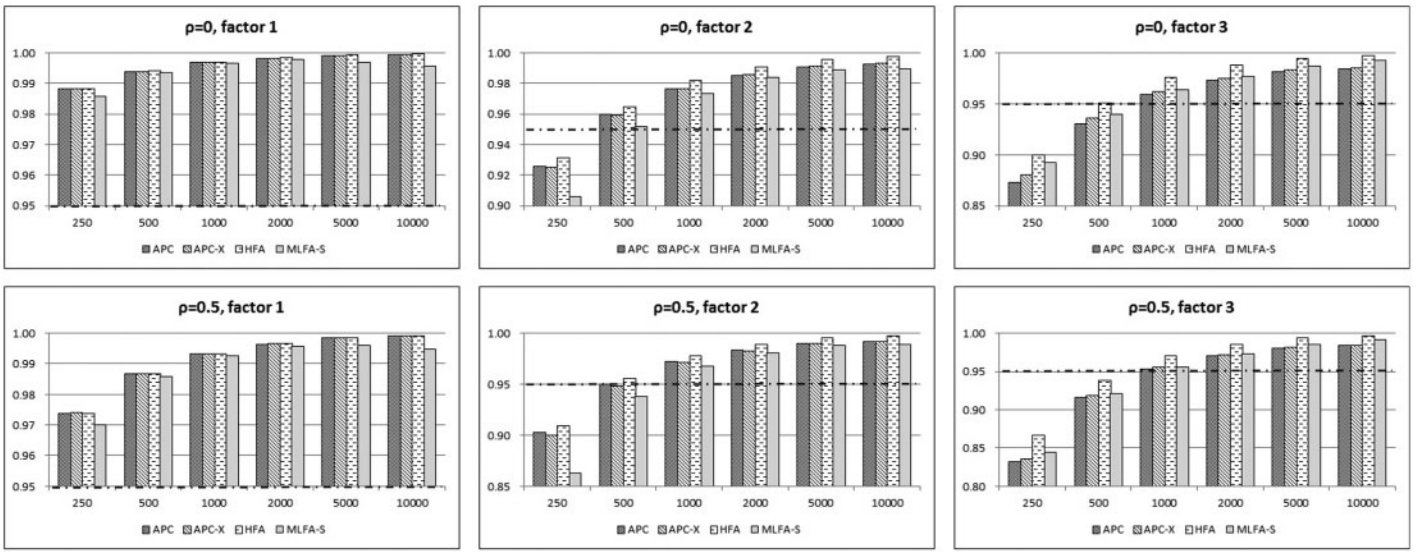

Figure 8
R^2 values from a regression of estimated factors on true factors: Case 3b, cross-sectional homoscedasticity and time-series heteroscedasticity
Balanced sample, *t*-distributed idiosyncratic errors, 60-month estimation period. The estimators are Connor and Korajczyk's (1986) asymptotic principal components (APC); Connor and Korajczyk's (1988) version of weighted principal components (APC-X) that accommodates cross-sectional heteroscedasticity; Jones's (2001) heteroskedastic factor analysis (HFA) that accommodates time series idiosyncratic heteroscedasticity; and Stroyny's (1992) maximum likelihood factor analysis (MLFA-S).

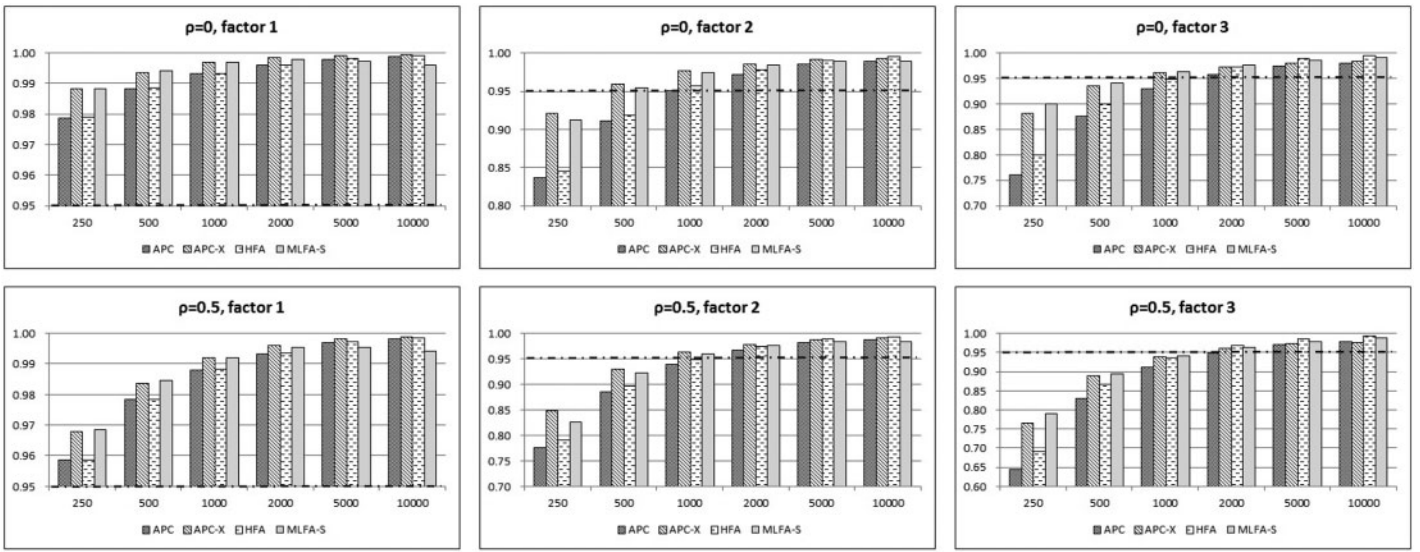

Figure 9
R^2 values from a regression of estimated factors on true factors: Case 4b, cross-sectional and time-series heteroscedasticity
Balanced sample, *t*-distributed idiosyncratic errors, 60-month estimation period. The estimators are Connor and Korajczyk's (1986) asymptotic principal components (APC); Connor and Korajczyk's (1988) version of weighted principal components (APC-X) that accommodates cross-sectional heteroscedasticity; Jones's (2001) heteroskedastic factor analysis (HFA) that accommodates time series idiosyncratic heteroscedasticity; and Stroyny's (1992) maximum likelihood factor analysis (MLFA-S).

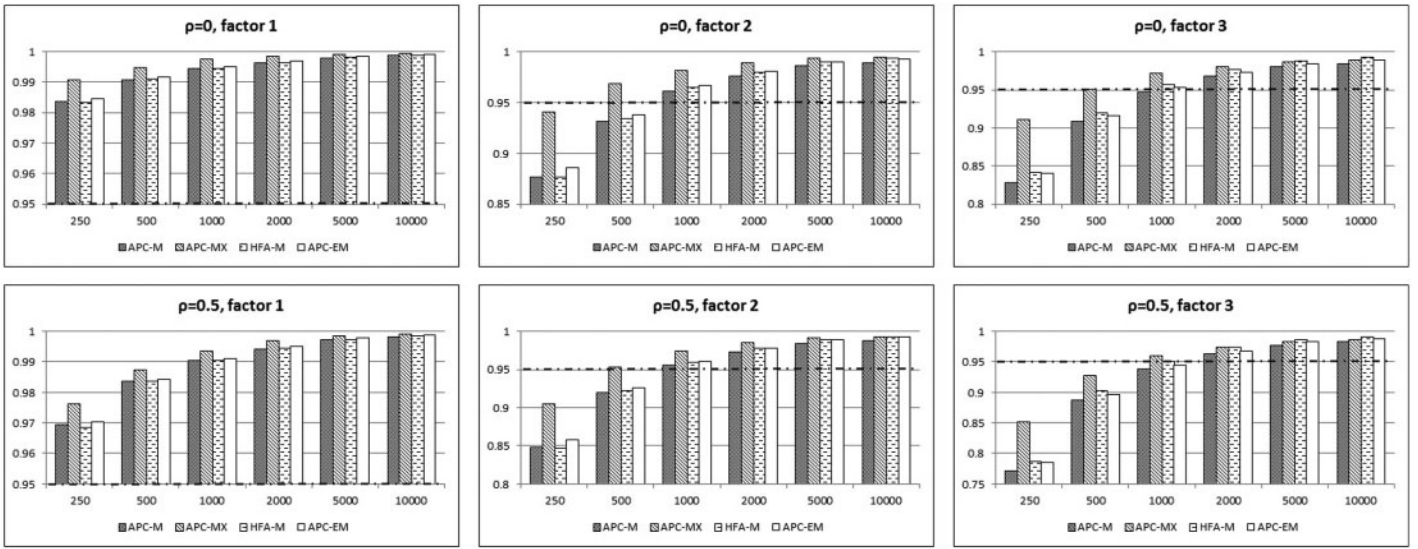

Figure 10
R^2 values from a regression of estimated factors on true factors: Case 4a, cross-sectional and time-series heteroscedasticity
Unbalanced sample, normally-distributed idiosyncratic errors, 60-month estimation period. The estimators Connor and Korajczyk's (1987) incomplete data asymptotic principal components (APC-M); Connor and Korajczyk's (1987, 1988) version of weighted principal components (APC-MX) that accommodates cross-sectional heteroscedasticity; Jones's (2001) heteroskedastic factor analysis (HFA-M) that accommodates time-series idiosyncratic heteroscedasticity and missing data; and Stock and Watson's (1998, 2002) estimator.

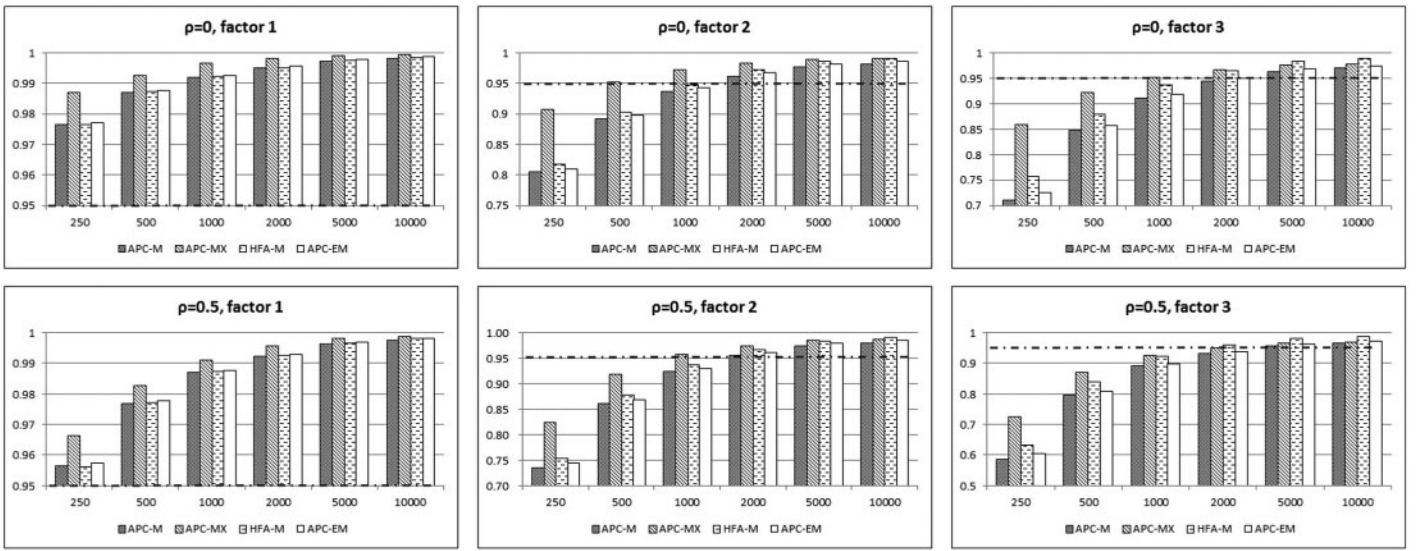

Figure 11
R^2 values from a regression of estimated factors on true factors. Case 4b, cross-sectional and time-series heteroscedasticity
Unbalanced sample, *t*-distributed idiosyncratic errors, 60-month estimation period. The estimators Connor and Korajczyk's (1987) incomplete data asymptotic principal components (APC-M); Connor and Korajczyk's (1987, 1988) version of weighted principal components (APC-MX) that accommodates cross-sectional heteroscedasticity; Jones's (2001) heteroskedastic factor analysis (HFA-M) that accommodates time-series idiosyncratic heteroscedasticity and missing data; and Stock and Watson's (1998, 2002) estimator.

instances in which APC-MX is not. Third, APC-EM and APC-M perform similarly.

2.5 Unbalanced panel returns: *t*-distributed idiosyncratic returns

Figure 11 shows the average R^2 values for Case 4b for all three factors and two cross-correlation structures ($\rho = 0.0,0.5$). First, APC-MX outperforms the other three estimators for low values of n. For factor one, APC-MX outperforms for all values of n and both $\rho = 0.0$ and $\rho = 0.5$. For factor two, APC-MX outperforms for values of n less than or equal to 10, 000 ($\rho = 0.0$) or values of n less than or equal to 5, 000 ($\rho = 0.5$). For factor three, APC-MX outperforms for values of n less than or equal to 2, 000 ($\rho = 0.0$) or values of n less than or equal to 1, 000 ($\rho = 0.5$). Second, for factors two and three, HFA is the best estimator in those instances in which APC-MX is not. Third, APC-EM and APC-M perform similarly.

3. Conclusion

In this paper, we document the performance of a number of estimators of factor returns using large-n methodologies. We simulate asset returns obeying an approximate factor model with a variety of assumptions about the nature of cross-sectional and time-series heteroscedasticity, the cross-correlation of idiosyncratic returns, the distribution of idiosyncratic returns, and with the data drawn from both balanced and unbalanced panels. The methods used for balanced panels include (1) APC, the asymptotic principal components estimator of Connor and Korajczyk (1986); (2) APC-X, the procedure of Connor and Korajczyk (1988) designed to accommodate cross-sectional heteroscedasticity in idiosyncratic returns and also a variant of weighted principal components (Stock and Watson 2006, section 4.3) and the feasible generalized principal components estimation (FGPCE) (Choi 2012); (3) Stroyny's (1992) maximum likelihood factor analysis, which also accommodates cross-sectional heteroscedasticity; and (4) Jones's (2001) heteroscedastic factor analysis (HFA), which incorporates time-series heteroscedasticity in idiosyncratic returns. The methods used for unbalanced panels include (1) APC-M, the missing data version of APC from Connor and Korajczyk (1987); (2) APC-MX, the missing data version of APC-X; (3) APC-EM, the *EM* algorithm-based estimator of Stock and Watson (1998); and (4) HFA-M, the missing data version of HFA from Jones (2001).

When the data are from a balanced panel and there is no heteroscedasticity, all the estimators perform similarly. In this case, cross-sectional sample sizes as small as 250 assets provide very accurate factor estimates. Idiosyncratic returns with fat tails require larger cross-sectional samples to achieve a given level of fit for the estimators. Cross-sectional heteroscedasticity leads to superior performance of the MLFA-S and APC-X estimators.

APC and HFA require much larger samples (3,000 to 6,000) to perform similarly to the estimators designed to accommodate cross-sectional heteroscedasticity. Time-series heteroscedasticity of the magnitude observed in monthly data leads to superior performance of the HFA estimator, particularly for factors two and three. When both cross-sectional and time-series heteroscedasticity are present, APC-X and MLFA-S provide the most accurate factor estimates for lower values of n, while HFA provides the most accurate factor estimates for higher values of n.

When the data are from an unbalanced panel and there is no heteroscedasticity, all estimators perform similarly . In this case, cross-sectional sample sizes as small as 500–1,000 assets provide very accurate factor estimates. t-distributed idiosyncratic returns lead to slightly less accurate estimators. With both cross-sectional and time-series heteroscedasticity (Cases 4a and 4b), the APC-MX estimator is most accurate for either all values of n (factors one and two with normal returns and factor one for t -distributed returns) or smaller values of n (all other cases). When APC-MX is not the best estimator, HFA-M is the best estimator. The results indicate that estimators that account for heteroscedasticity are preferred, particularly when the cross-sectional sample is small. The full U.S. market of traded equities over 60-month periods (and requiring at least 36 months of observations) provides cross-sectional sample sizes between 4,123 and 6,237 firms, so the differences across estimators are relevant for studies in most markets.

Our read on the recent literature is that most papers do not accommodate cross-sectional heteroscedasticity and almost none, except for Jones (2001), accommodate time-series heteroscedasticity. Goyal, Pérignon, and Villa (2008) study group-specific and cross-group factors. Their empirical work has cross-sectional samples varying from 2,942 to 4,023, split between groups of stocks that are traded on the NYSE (samples between 1,500 and 1,763) and NASDAQ (samples between 1,252 and 2,263). There is no adjustment for heteroscedasticity. Ando and Bai (2015) also study group-specific factor structures but accommodate the existence of observable and unobservable factors. Their sample has 1,039 stocks in group A and 102 in group B. The statistical factors are estimated without taking into account heteroscedasticity. Greenaway-McGrevy, Han, and Sul (2012) also study estimators that include observables and latent factors. Their procedure does not accommodate heteroscedasticity, although a variant allows for serial correlation in idiosyncratic returns. Their simulation analysis varies with the cross-sectional samples ranging from 25 to 4,000.

Westerlund and Urbain (2015) compare the performance of the APC estimator to simple cross-sectional averaging. Their estimators do not accommodate heteroscedasticity, and they simulate factor structures without heteroscedasticity with cross-sectional samples up to 1,200. Two papers by Ludvigson and Ng (2007, 2009) use scaled variables in the factor estimation, which implicitly corrects for cross-sectional heteroscedasticity, but not for

time-series heteroscedasticity. Su and Wang (2017) propose a factor estimator robust to time-varying factor loadings and simulate its performance in factor economies with either homoscedasticity or cross-sectional heteroscedasticity for cross-sectional samples of 100 or 200.

The cross-sectional sample sizes are typically in the range such that estimators taking heteroscedasticity into account would improve precision. We have not replicated those studies to see if their inferences would be overturned or strengthened, but our results suggest that those looking for more precise factor estimates should consider estimators that account for heteroscedasticity.

References

Andersen, T. G., T. Bollerslev, and F. X. Diebold. 2010. Parametric and nonparametric volatility measurement. In *Handbook of financial econometrics*, eds. Y. Aï t-Sahalia and L. P. Hansen, volume 1. Amsterdam: North-Holland.

Ando, T., and J. Bai. 2015. Asset pricing with a general multifactor structure. *Journal of Financial Econometrics* 13:556–604.

Bai, J., and S. Ng. 2002. Determining the number of factors in approximate factor models. *Econometrica* 70:191–221.

Boivin, J., and S. Ng. 2006. Are more data always better for factor analysis? *Journal of Econometrics* 132:169–94.

Campbell, J. Y., M. Lettau, B. G. Malkiel, and Y. Xu. 2001. Have individual stocks become more volatile? An empirical exploration of idiosyncratic risk. *Journal of Finance* 56:1–43.

Chamberlain, G., and M. Rothschild. 1983. Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica* 51:1305–24.

Choi, I. 2012. Efficient estimation of factor models. *Econometric Theory* 28:274–308.

Connor, G., and R. A. Korajczyk. 1986. Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics* 15:323–46.

———. 1987. Estimating pervasive economic factors with missing observations. Working Paper, http://ssrn.com/abstract=1268954.

———. 1988. Risk and return in an equilibrium APT: Application of a new test methodology. *Journal of Financial Economics* 21:255–90.

———. 1993. A test for the number of factors in an approximate factor model. *Journal of Finance* 48:1263–91.

Connor, G., R. A. Korajczyk, and O. Linton. 2006. The common and specific components of dynamic volatility. *Journal of Econometrics* 132:231–55.

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, E. F., and K. R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33:5–56.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin. 2005. The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association* 100:830–40.

Gagliardini, P., and C. Gourieroux. 2014. Efficiency in large dynamic panel models with common factors. *Econometric Theory* 30:961–1020.

Goyal, A., C. Pérignon, and C. Villa. 2008. How common are common return factors across the NYSE and Nasdaq? *Journal of Financial Economics* 90:252–71.

Greenaway-McGrevy, R., C. Han, and D. Sul. 2012. Asymptotic distribution of factor augmented estimators for panel regression. *Journal of Econometrics* 169:48–53.

Jones, C. S. 2001. Extracting factors from heteroskedastic asset returns. *Journal of Financial Economics* 62:293–325.

Ludvigson, S. C., and S. Ng. 2007. The empirical risk–return relation: A factor analysis approach. *Journal of Financial Economics* 83:171–222.

———. Macro factors in bond risk premia. *Review of Financial Studies* 22:5028–67.

Roll, R., and S. A. Ross. 1980. An empirical investigation of the arbitrage pricing theory. *Journal of Finance* 35:1073–103.

Ross, S. A. 1976. The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13:341–60.

Rubin, D. B., and D. T. Thayer. 1982. *EM* algorithms for ML factor analysis. *Psychometrika* 47:69–76.

Stock, J. H., and M. W. Watson. 1998. Diffusion indices. Working Paper, http://ssrn.com/abstract=226366.

———. 2002. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association, Theory and Methods* 97:1–13.

———. 2006. Forecasting with many predictors. In *Handbook of economic forecasting*, eds. G. Elliott, C. W. J. Granger, and A. Timmermann, pp. 515–54. Amsterdam: Elsevier.

Stroyny, A. L. 1992. Still more on EM factor analysis. Working Paper, University of Wisconsin.

Su, L., and X. Wang. 2017. On time-varying factor models: Estimation and testing. *Journal of Econometrics* 198:84–101.

Westerlund, J., and J.-P. Urbain. 2015. Cross-sectional averages versus principal components. *Journal of Econometrics* 185:372–77.