Which factor model? A systematic return covariation perspective

Shamim Ahmed* Ziwen Bu† Lazaros Symeonidis‡ Daniel Tsvetanov§¶

April 14, 2023

Abstract

We examine which factor model best captures systematic return covariation. Focusing on the economic implications for portfolio risk control, the pairwise variance equality test and the model confidence set procedure suggest that the Fama and French (2015) five-factor model, the Barillas and Shanken (2018) six-factor model, and the Fama and French (2018) six-factor model are the top performers for the factor model-implied minimum risk portfolios in the out-of-sample. When it comes to the minimum tracking error portfolios, the Barillas and Shanken (2018) six-factor model and the Fama and French (2018) six-factor model are the overall winners in the horse race.

JEL Classification: C51; C52; G11; G12

Keywords: Asset pricing model; Factor model; Model evaluation; Portfolio selection; Out-of-sample

*University of Liverpool Management School, University of Liverpool, Liverpool L69 7ZH, United Kingdom; Shamim.Ahmed@liverpool.ac.uk.

†Birmingham Business School, University of Birmingham, Birmingham B15 2TT, United Kingdom; Z.Bu.1@bham.ac.uk.

‡Essex Business School, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom; l.symeonidis@essex.ac.uk.

§Corresponding author. Norwich Business School, University of East Anglia, Norwich NR4 7TJ, United Kingdom; D.Tsvetanov@uea.ac.uk.

¶We thank Kees G. Koedijk (Editor), an anonymous referee, Raman Uppal, Seok Young Hong, and participants at the 2021 IAAE Conference for useful comments and suggestions.

1. Introduction

While much of asset pricing literature involves searching for factors or factor models that explain differences in expected returns, less focuses on analyzing their ability to capture systematic return covariation. For example, in a recent study, Harvey et al. (2016) catalog 313 papers that propose 316 different factors to explain the cross-section of expected returns. However, they do not attempt to relate the candidate factors or, equivalently, the factor models to the covariance matrix of returns. This paper seeks to contribute to this relatively underresearched area by formally examining which factor model best captures systematic return covariation. Knowing this has both academic and practical importance. Theory suggests that a necessary condition for any factor candidate is that it must be related to the covariance matrix of returns (Pukthuanthong et al., 2019). The condition can be generalized to a given set of factors that constitute an asset pricing model. If the common covariation in returns can be traced to a small set of underlying factors, then these factors serve as candidates for the sources of systematic risk and expected returns provide compensation for bearing such risk (Chan et al., 1998). In this respect, our study helps to validate different models as candidates for the sources of systematic risk. In fact, Chan et al. (1999, p. 968) reiterate, "[F]actor models of security returns were originally proposed as parsimonious ways to predict return covariances and simplify portfolio optimization." Thus, comparing the pricing ability of factor models in the cross-section of expected returns, which numerous studies have done in the existing literature (see, for example, Ahmed et al., 2019; Fama and French, 2016), is beyond the scope of this paper.

From the viewpoint of the investment practitioner, it is also immensely important to formally identify which factor model best captures the systematic component of return covariation, since this component is the main source of portfolio risk. In this context, Chan et al. (1998, p. 159) put it: "[T]he ability to identify which factors best capture systematic return covariation is central to applications of multifactor pricing models." Chan et al. (1998, 1999) further emphasize that some factors may not be priced, but can account for the common shared variation in asset returns, which makes them important for investors who wish to manage portfolio risk. More practically, factor models help handle high-dimensional sets of assets in the investment universe by significantly reducing the number of parameters in the estimation of the covariance matrix. Thus, analyzing the dynamics of the covariance matrix of asset returns captured by the factor models may aid developing better-performing investment strategies (Moskowitz, 2003). Indeed, there is ample evidence that factor pricing models are routinely employed for portfolio risk control, in which model-implied forecasts of the future covariance matrix of returns are the key input (see, for example, Ang, 2014; Brandt, 2010; Chincarini and Kim, 2006; Meucci, 2005).

To uncover which factor model best captures systematic return covariation, we focus on the economic implications for portfolio risk control. Specifically, we conduct examination through the lens of an investor who utilizes forecasts of the factor model-implied covariance matrix of excess returns to minimize the out-of-sample variance of her portfolio. The key point here is that if a factor model does a good job in capturing the systematic portion of the covariance matrix of asset returns, then it would lead to a low variance for the minimum variance portfolio implied by the model (see Section 2.3 for a formal justification to this assertion). To make our empirical analyses as realistic as possible, we consider the minimum variance portfolio (i.e., without any restriction on weights for individual assets), the restricted minimum variance portfolio (i.e., limiting long and short positions), and the long-only minimum variance portfolio (i.e., without short selling). Our out-of-sample model evaluation approach has a much broader appeal and offers two notable advantages. First, there is extensive evidence that the in-sample performance of a factor model tends to correlate poorly with its ability to generate satisfactory out-of-sample forecasts (see Chan, Karceski, and Lakonishok, 1999; Simin, 2008). Thus, an out-of-sample analysis is needed to uncover the true performance of a model. Second, it allows some of the issues well-documented in the literature to be circumvented, such as the useless factor bias (Kan and Zhang, 1999) and the data snooping biases (Foster et al., 1997; Linnainmaa and Roberts, 2018; Lo and MacKinlay, 1990; MacKinlay, 1995).

We consider minimum risk portfolios implied by factor models instead of mean-variance portfolios for several reasons. First, our objective is to formally assess the ability of factor models to capture the systematic return covariation. For this purpose, a minimum variance portfolio is better suited as it offers a "clean" way of evaluating the quality of a factor model-implied covariance matrix estimator by sidestepping the thorny issue of predicting expected returns (Ledoit and Wolf, 2017). It has long been recognized that expected returns are notoriously difficult to predict. As a result, the performance of the traditional mean-variance efficient portfolio is highly sensitive to the estimation error in the arithmetic mean of the returns (Chan et al., 1999; DeMiguel et al., 2009; Jagannathan and Ma, 2003). Second, it is well-documented that estimated minimum variance portfolios have desirable out-of-sample properties not only in terms of risk but also in terms of Sharpe ratio (see, among others, Jagannathan and Ma, 2003; Nielsen and Aylursubramanian, 2008; Olivares-Nadal and DeMiguel, 2018). Indeed, such portfolios are often included in the range of financial products sold by the mutual fund industry (Ledoit and Wolf, 2017). Third, there is mounting evidence that superior returns to investment performance are elusive and therefore emphasis on risk control is growing in the asset management industry (Ang, 2014; Chan et al., 1999). Our empirical framework is also aligned with this trend as it compares factor models based on their ability to capture the systematic component of return comovement, rather than their ability to explain expected returns.

Some studies have shown that the dominant influence of the market factor on return variation can cause the performance of minimum risk portfolios implied by different factor models to be very similar, especially when additional constraints on the weights are imposed (see Chan et al., 1999, p. 955). Consequently, differentiating between models can be very difficult. A way to address this issue is to remove the impact of the dominant market factor. As it turns out, this problem is a specific case of tracking a benchmark portfolio (Chan et al., 1999). This motivates us to extend our analyses to the minimum tracking error portfolio, which involves minimizing the portfolio's tracking error variance. To be consistent with the practice in the investment industry, we also consider the restricted minimum tracking error portfolio and the long-only minimum tracking error portfolio. Arguably, these portfolios are the most interesting to portfolio managers who are evaluated relative to some benchmark (Chan et al., 1999; Cremers and Petajisto, 2009). This is because institutional investors commonly instruct their managers to construct minimum tracking error portfolios, using a subset of stocks that have low transaction costs and high liquidity, to track certain benchmark indices that contain assets that are not actively traded (Jagannathan and Ma, 2003; Jorion, 2003). It is worth noting that to construct these portfolios, forecasts of the covariance matrix of returns in excess of the benchmark's return are required where factor models are frequently used to obtain these forecasts (Chan et al., 1999).

We examine the out-of-sample performance of a large array of both classic and new-generation models. Our array comprises: the capital asset pricing model of Sharpe (1964) and Lintner (1965), the Fama and French (1993) three-factor model, the Fama and French (1993) and Carhart (1997) four-factor model, the Asness and Frazzini (2013) three-factor model, the Hou et al. (2015) q-factor model, the Fama and French (2015) five-factor model, the four-factor model of Fama and French (2015) that excludes the "high minus low" value factor, the Stambaugh and Yuan (2017) four-factor model, the Barillas and Shanken (2018) six-factor model, the Fama and French (2018) six-factor model, and the Daniel et al. (2020) three-factor model. We confine our examination to these models for two reasons. First, they have survived as the workhorses if not the best models over the years. In fact, almost all of the models under investigation have become academic standards and are frequently used for risk-adjustment purposes in the asset pricing literature (see, for example, Ahmed, Bu, and Ye, 2023a,b; Bali, Engle, and Murray, 2016; Hirshleifer, Hsu, and Li, 2018). Second, our methodology (described in the next section) requires estimation of the covariance matrices at daily frequency. Hence, we consider models for which daily data are readily available in the public domain. Moreover, given the abundance of models in the literature, examining these 11 major factor models keeps the out-of-sample model comparison exercise reliably manageable.

One may think that models that are successful in explaining the cross-section of expected returns should also be good in capturing the systematic return covariation. This is not always true because there are factors that are reliably associated with expected returns but not risks (see Moskowitz, 2003; Pukthuanthong et al., 2019). As a result, whether in isolation or in combination with other factors forming a particular model, they do not describe return covariation. Also, in the presence of constraints on borrowing and short selling, a set of benchmarks that is correct for pricing is not necessarily correct for investing (Pástor and Stambaugh, 2000). We show, as other authors have (see Moskowitz, 2003, Table 2), that some of our models well-known for explaining return anomalies do not turn up as the best model in the horse race of capturing systematic return covariation. The Hou et al. (2015) q-factor model is an example despite the evidence of its superior pricing ability in the cross-section of expected returns (see Ahmed et al., 2019).

To statistically assess our competing factor models in a realistic setting, choosing test assets in which to invest poses a challenge. In this regard, we take guidance from prior studies (see, among others, Chan et al., 1999; Giglio and Xiu, 2021; Lewellen et al., 2010; Moskowitz, 2003), while keeping our analyses parsimonious. The test assets include: samples of 50 small stocks (randomly selected) and 100, 250 and 500 largest stocks (drawn with a defined rule) from NYSE-, AMEX-, and NASDAQ-listed nonfinancial and nonregulated ordinary common stocks; a set of value-weighted 48 industry portfolios; and a set of value-weighted 340 portfolios used in Giglio and Xiu (2021). All of them are easily accessible to researchers, as described in Section 3.2, and therefore serve as a reasonable starting point to conduct our empirical analyses.

We begin with computing the out-of-sample forecasts of the covariance matrix implied by each competing model, where the model-implied covariance matrix is filtered to focus solely on the systematic portion of return covariation (see Section 2 for details). Engle and Colacito (2006) show that when the problem at hand is to assess the accuracy of different covariance forecasts, ranking the alternative models based on the volatility of optimized portfolios offers a minimum loss of efficiency, even in the context of a mean-variance optimal portfolio. Therefore, the variance of the portfolio is a measure that should be able to pick the correct covariance estimator independently of the model of expected returns. Their analysis further reveals that using a Sharpe ratio criterion for ranking covariance matrix estimators may be misleading, since it can lead to the selection of the wrong model (Engle and Colacito, 2006, Corollary 2 of Proposition 1). Given these findings and our objective of comparing models in terms of their ability to capture systematic return covariation, we conduct pairwise tests of equality of the out-of-sample variances for the returns realized on the minimum variance portfolios of test assets implied by the competing models. In the case of a tracking error minimizing portfolio, the object of interest is the out-of-sample variance of the difference between the portfolio's return and the return on a given benchmark. As in Jagannathan and Ma (2003), our benchmark is the Standard & Poor's (S&P) 500 index.

While the pairwise test takes us well beyond the common practice of identifying the best model simply by comparing point estimates of various performance metrics, the underlying testing procedure may not always determine unambiguously the best-performing factor model when multiple models are involved in the horse race (Ahmed et al., 2019; Barillas et al., 2020; Gospodinov et al., 2013; Kan et al., 2013). To address this concern, we further conduct simultaneous comparison of factor pricing models by utilizing the model confidence set (MCS) procedure of Hansen et al. (2011). The advantage of the MCS procedure is that it enables us to determine the best-performing asset pricing model(s) from a collection of competing models, with a given level of confidence. In our case, "best-performing" is defined in terms of a lower out-of-sample variance (or tracking error variance) of the optimized portfolio implied by a given factor model.

Our results show that the Fama and French (2015) five-factor model, the Barillas and Shanken (2018) six-factor model, and the Fama and French (2018) six-factor model are the best for capturing the systematic return covariation, as reflected in a significantly lower out-of-sample variance of the corresponding minimum risk portfolios. When it comes to the different versions of the minimum tracking error portfolios, the Barillas and Shanken (2018) six-factor model and the Fama and French (2018) six-factor model once again turn up as the best models in the horse race. All these findings remain robust to: (1) accommodating the impact of transaction costs on model performance in the out-of-sample; (2) a recursive exercise in which factor models are compared as they are proposed against previously introduced models; (3) an evaluation of economic gains from using factor models; and (4) subperiod analyses and different market states.

In a recent work, Ahmed et al. (2019) show that the Hou et al. (2015) q-factor model, the Fama and French (2015) five-factor and four-factor models, and the Barillas and Shanken (2018) six-factor model are the top performers in explaining several well-known return anomalies. Our results, from comprehensive analyses, suggest that models that are successful in this context are not necessarily successful for capturing systematic return covariation of assets, which is aligned with Pástor and Stambaugh (2000) and Pukthuanthong et al. (2019), and justifies the goal of the paper.

Taken together, the robust empirical results in this paper contribute to the growing literature on evaluating asset pricing models in out-of-sample, especially focusing on their relation to covariance risk. They also have valuable implications for practical applications including portfolio risk optimization in the asset management industry. Furthermore, the optimization techniques that we demonstrate using factor pricing models should help professional money managers deal with the curse of dimensionality of assets in the ever-expanding investment universe.

The studies that are closely related to ours are Chan et al. (1998, 1999) and Moskowitz (2003), which explore the performance of factor models (or factors in isolation) in capturing the covariance structure of returns. But our paper differs from these studies on several grounds. First, and most importantly, we compare models using formal statistical procedures involving the pairwise variance equality test and the computation of the MCS. Second, we employ a wide variety of test assets spanning individual stocks and equity portfolios. Third, our comparison of factor models utilizes a broad set of alternative versions of minimum variance and minimum tracking error portfolios commonly encountered in practice. Fourth, we conduct additional analyses to account for the impact of transaction costs on model performance. Fifth, we perform a recursive exercise in which models are formally compared as they are proposed against previously introduced models in the literature. Sixth, we compute economic gains from using factor models. Lastly, we conduct sensitivity and subperiod analyses. All these features of the empirical design are beyond those adopted in the aforementioned studies and make our findings more reliable and robust, let alone uncovering the best models from a much larger array of both classic and new-generation models.

2. Framework

Suppose we have N risky test assets and r_t denotes the $N \times 1$ vector of returns on those assets during period t $(= 1, \ldots, T)$. Then a linear factor pricing model for r_t with K risk factors can be specified as

$$r_t - r_{RF,t}\mathbb{1}_N = \mu_r + \beta f_t + \epsilon_t, \tag{1}$$

where $r_{RF,t}$ is the risk-free rate at time t, $\mathbb{1}_N$ is an $N \times 1$ vector of ones, μ_r is an $N \times 1$ vector of asset risk premia, β is an $N \times K$ matrix of factor loadings (i.e., risk exposure) of the test assets, f_t is a $K \times 1$ vector of demeaned factor returns (i.e., factor return innovations or factor shocks), ϵ_t is an $N \times 1$ vector of residuals, and $E(\epsilon_t|f_t) = 0$. In line with most empirical studies comparing asset pricing models (see, for example, Barillas and Shanken, 2018; Fama and French, 2018; Hou, Mo, Xue, and Zhang, 2019), we assume that the factors are replicated by the returns on zero-investment portfolios. Moreover, we assume that at any given point in time t

$$f_t|\mathcal{F}_{t-1} \sim \mathcal{N}(0_K, \Omega_t) \quad \text{and} \quad \epsilon_t|\mathcal{F}_{t-1} \sim \mathcal{N}(0_N, V_t), \tag{2}$$

where $\mathcal{F}_{t-1}$ is the information set generated by the past values of f_t and ϵ_t; 0_K and 0_N are, respectively, the $K \times 1$ and $N \times 1$ vectors of zeros; and Ω_t and V_t are the positive definite matrices of dimensions $K \times K$ and $N \times N$, respectively. The above assumption is consistent with the ample evidence that volatility is predictable (see, among others, Bollerslev et al., 1992, 1994; Moskowitz, 2003). It is worth mentioning that the assumption of normality in equation (2) is not pivotal because even if it does not hold, the composite-likelihood estimation becomes the quasi composite-likelihood, which still has desirable properties (see Engle et al., 2019).

In light of equation (2), the total (conditional) covariance matrix of excess returns, denoted Σ_t, implied by equation (1), is decomposed into a systematic portion and a residual portion as

$$\Sigma_t = \beta \Omega_t \beta' + V_t. \tag{3}$$

Note that equation (3) is a generalization of the well-known total covariance matrix implied by an unconditional static factor model that assumes $\Sigma_t = \Sigma$.[1] In this regard, our framework is consistent with the unconditional dynamic factor models, which have been previously adopted in the literature under the assumption that the factors are latent (see, among others, Aguilar and West, 2000; Engle, Ng, and Rothschild, 1990; Nardari and Scruggs, 2007).

Charoenrook and Conrad (2008) and Pukthuanthong et al. (2019) argue that equation (3) is a necessary condition for any factor candidate. That is, the factor must be related to the covariance matrix of returns. The condition can be generalized to a particular model specification in the sense that the factors that constitute the model should explain the systematic return comovement.[2] This generalization is the motivation for our study to identify the model that best accounts for the common shared variation in returns, which is paramount to our understanding of the underlying forces that move asset prices. It is also of immense importance to practitioners who use factor models for portfolio risk optimization. Indeed, the only relevant risk component of the total covariance matrix decomposition given by equation (3), for which an

[1] By unconditional, we mean that the parameters in the multivariate regression given by equation (1) are assumed to be constant over time, which is in line with the model specification adopted by most prior studies that compare asset pricing models (see, among others, Barillas et al., 2020; Hou et al., 2019; Kan et al., 2013).

[2] The necessary condition implied by equation (3) does not distinguish between pervasive priced factors and unpriced factors (Pukthuanthong et al., 2019). But our interest lies in evaluating the ability of models to capture the systematic return covariation regardless of whether their constituent factors are priced or not.

investor demands a premium is $\beta\Omega_t\beta'$, since the idiosyncratic risk component can be diversified away.[3]

To quantify the ability of a factor model to capture the systematic component of asset return variation, we analyze the out-of-sample performance of optimized portfolios derived from the estimated covariance matrix of excess returns implied by the model. In doing so, we follow Moskowitz (2003, pp. 438−439) and concentrate on the systematic portion of return covariation (i.e., $\beta\Omega_t\beta'$) captured by the factors that constitute the model. If a factor model i captures the systematic component of asset return variation better than a factor model j, the covariance matrix implied by model i will be closer to the "true" covariance matrix. Therefore, by using the result of Theorem 1 in Engle and Colacito (2006, p. 239), we can conclude that model i will allow investors to achieve lower volatility, higher return, or both. We elaborate on this point in Section 2.3. Since $\beta\Omega_t\beta'$ is symmetric positive semidefinite, we add a constant (across models) diagonal matrix, denoted V_t^{full}, to each competing factor model-implied $\beta\Omega_t\beta'$ to ensure nonsingularity (see Moskowitz, 2003, p. 439). The matrix, V_t^{full}, contains only the conditional volatility estimates of the residuals from a "full" model (i.e., a model that employs all the distinct factors of the competing models). As a result, the optimal weights or, equivalently, the performance of the portfolios will vary across models only to the extent that the systematic component of return covariation captured by the factors that constitute a particular model varies (see equations (4) through (7) for an elaboration on this point). Stated alternatively, this will allow us to compare the models solely by their ability to capture the systematic component of the total covariance matrix given that the idiosyncratic component is the same across models.[4,5]

2.1 Portfolio optimization: Minimum variance portfolio

Consider a set, $\mathcal{M}^0$, that contains a finite number of competing factor models, where each model is indexed by $i = 1, \ldots, m_0$. To facilitate the comparison across the different factor models and uncover the "true" ability of each model to predict the future covariance matrices, we begin with a rolling out-of-sample forecasting approach similar to that in Chan et al. (1999), DeMiguel et al. (2009), and Ledoit et al. (2019). The covariance matrix forecasts are generated from a

[3] For an elaboration of this point, consider a vector of allocations, $w_t^i = (w_{1,t}^i, \ldots, w_{N,t}^i)'$, of the risky assets implied by the factor model i at time t and assume that $\iota_N' w_t^i = 1$. If the portfolio is well-diversified, that is, $\sum_{n=1}^{N}(w_{n,t}^i)^2 \to 0$ as $N \to \infty$ (see, for example, Chamberlain, 1983), the idiosyncratic component of its variance tends to zero as $N \to \infty$, provided that the eigenvalues of V_t are uniformly bounded away from zero.

[4] If we utilize V_t^i instead of V_t^{full}, then the performance of the portfolios will vary across models not only to the extent that the systematic component of return covariation captured by a particular model varies but also to the extent that the idiosyncratic component captured by the same particular model varies.

[5] Our overall findings remain qualitatively the same when we use V_t^i (i.e., residual covariance matrix from factor model i) instead of V_t^{full}. These results are available in Tables IA15 and IA16 of the Internet Appendix.

multivariate generalized autoregressive conditional heteroscedasticity (GARCH) model that we use to characterize the time-varying dynamics of the covariance matrix of test asset returns implied by each factor model. In particular, we employ the dynamic conditional correlation (DCC) model of Engle (2002) to estimate the conditional correlation matrix of the factors that belong to model i. We choose the DCC model for the following reasons. First, it is a rather popular and widely used multivariate volatility model, which has been shown to possess good out-of-sample forecasting properties (see Laurent et al., 2012). Second, it offers a great deal of computational advantages over other multivariate volatility models and avoids over-parameterization (see Engle, 2002). Following Engle et al. (2019), we use a GARCH(1,1) process to estimate the elements of Ψ_t, that is, the vector of conditional variances of the factor return innovations that is used to convert the conditional correlation matrix to a conditional covariance matrix (see Appendix A for more details). Similarly, we employ a GARCH(1,1) process to estimate the (n, n)th entry of V_t^{full}, that is, the diagonal matrix of residual volatility estimates from the "full" factor model.[6]

To improve the precision of the parameter estimates of the multivariate GARCH models and thus mitigate the concern that estimation error might add noise to our empirical analysis, we use daily data for both the test assets and the factors. Also, in line with the recent literature (see, among others, Bollerslev et al., 2018; Ledoit et al., 2019), we update all portfolios at the end of every 21-day period. The general set-up for the estimation of the minimum variance portfolio is as follows. Given a full sample of T observations, we choose an in-sample estimation window of length L. At each portfolio formation date, d, for model $i \in \mathcal{M}^0$ and conditional on the availability of data from day $d - L + 1$ to d, the vector of allocations of the risky assets, w_d^i, is determined by the solution to the following quadratic programming problem:

$$\min_{w_d^i} \frac{1}{2} \left(w_d^i\right)' \hat{\Sigma}_d^i \left(w_d^i\right) \qquad \text{s.t.} \quad \mathbb{1}_N' w_d^i = 1, \tag{4}$$

[6] We acknowledge that there are alternative univariate GARCH model specifications incorporating additional features such as asymmetries. But we choose the simplest and the most parsimonious alternative to model the time-varying variance dynamics. In their comprehensive study of 330 GARCH-type models, Hansen and Lunde (2005) show that a GARCH(1,1) process is sufficient to successfully forecast the volatility of different asset classes.

where

$$\hat{\Sigma}_d^i = \left(\hat{\beta}_i \bar{\Omega}_d^i \hat{\beta}_i' + \bar{V}_d^{full}\right), \tag{5}$$

$$\bar{\Omega}_d^i = \frac{1}{21} \sum_{z=d+1}^{d+21} \hat{\Omega}_z^i, \tag{6}$$

$$\bar{V}_d^{full} = \frac{1}{21} \sum_{z=d+1}^{d+21} \hat{V}_z^{full}, \tag{7}$$

$\hat{\beta}_i$ is the ordinary least squares estimate of β_i, the factor loadings of asset pricing model i, and $\bar{\Omega}_d^i$ and $\bar{V}_d^{full}$ are, respectively, the averaged (over z-step ahead) forecasts of Ω^i and V^{full}.[7] To conserve space, details of the estimation of $\hat{\Omega}_z^i$ and $\hat{V}_z^{full}$ are given in Appendix A. The forecasts in equations (6) and (7) are averaged to address the mismatch between the portfolio updating frequency and the covariance matrix forecasting frequency.

The solution to the optimization in equation (4) is given by

$$\hat{w}_d^i = \frac{\left(\hat{\Sigma}_d^i\right)^{-1} \mathbb{1}_N}{\mathbb{1}_N' \left(\hat{\Sigma}_d^i\right)^{-1} \mathbb{1}_N}. \tag{8}$$

Upon obtaining the optimized weights at day d, we then compute the returns on the minimum variance portfolio, denoted $r_{\mathrm{MVP},z}^i$, for the subsequent 21 days as

$$r_{\mathrm{MVP},z}^i = \left(\hat{w}_d^i\right)' r_z \quad \text{for} \quad z = d+1, \ldots, d+21. \tag{9}$$

At the end of day $d + 21$, the forecasting and optimization procedures are repeated.

The above procedure starts at day $d = L$ and is conducted until we reach the end of the full sample. We therefore end up with a time-series of $T - L$ daily out-of-sample returns realized on each portfolio constructed using asset pricing model i. This enables us to evaluate the performance of competing models in terms of their implications for portfolio choice, noting that investment differences across models will be driven only by the systematic component, $\hat{\beta}_i \bar{\Omega}_d^i \hat{\beta}_i'$.

To incorporate settings that correspond to actual practice for most investment managers, we consider two other scenarios. In the first scenario, as in Moskowitz (2003), we add the constraint (in equation (4)) that the weights on any asset are between −0.50 and 1.50. This averts generating extreme portfolio positions. In the second scenario, as in DeMiguel et al. (2014) and Jagannathan and Ma (2003), short selling and borrowing at the risk-free are not

[7] $\hat{\beta}_i$ is estimated using data from day $d - L + 1$ to d. In this sense, although we assumed β_i to be constant over time, the estimates that we obtain at each portfolio formation date, d, may vary due to sampling error.

allowed, so that the weight on any single asset lies between 0 and 1.[8] We refer to the portfolios incorporating these additional scenarios as the restricted minimum variance portfolio and the long-only minimum variance portfolio, respectively. Note that real-life portfolio managers also face transaction costs in addition to varying degrees of constraints on leverage and turnover. We accommodate possible transaction costs in our analyses in Section 5.1.

2.2 Portfolio optimization: Minimum tracking error portfolio

A strand of literature has shown through factor analysis of equity returns and analysis of eigenvalues that a major factor, namely, the market factor, is the dominant source of return variation (see Chan et al., 1999; Connor and Korajczyk, 1993). This dominant influence of the market factor, which tends to overpower other factors, can cause the performance of minimum variance portfolios implied by different models to be quite similar. Hence, differentiating between the models can be challenging, especially when additional constraints on the weights are imposed (Chan et al., 1999). To address the possible impact of a dominant factor and to design our experiment such that the differences across the factor model-implied portfolios (or, equivalently, the factor models) are more stark, we also consider the minimum tracking error portfolio. The objective in this case is to minimize the variance of the difference between the portfolio's return and the return on a benchmark. Given that the market factor is the dominant factor, if the benchmark's market exposure is not too unrepresentative of those of the underlying set of assets, then the difference between the market betas of the portfolio and the benchmark can be set close to zero (Chan et al., 1999). Thus, the incremental informativeness of any remaining factors becomes easier to detect, which in turn, leaves more room to differentiate between the models.[9]

To implement the tracking error minimization, we follow Chan et al. (1999) and Jagannathan and Ma (2003), and obtain the forecasts of model-implied covariance matrix of returns in excess of the benchmark's return. The forecasting scheme for covariance matrices and the procedure for portfolio updating are the same as those in Section 2.1. In addition, we consider two alternative versions of the minimum tracking error portfolio. These are the restricted minimum tracking error portfolio (i.e., limiting long and short positions) and the long-only minimum tracking error portfolio (i.e., no short selling). It is worth highlighting that these versions of the minimum tracking error portfolios are perhaps the most interesting to portfolio managers who are paid to outperform a benchmark. Accordingly, they are concerned with how their portfolios deviate

[8] With the additional constraints, the resulting optimized portfolios are not explicitly given and therefore the solutions need to be found numerically.

[9] A more detailed explanation on why adjusting by a benchmark matters, for differentiating between factor model-implied portfolios, can be found in Section 5.1 of Chan et al. (1999).

from the benchmark rather than the absolute variances of their portfolios (Chan et al., 1999).

2.3 Model comparison metric

Let Σ_τ be the true conditional covariance matrix of asset returns at time τ $(= L+1, \ldots, T)$, Σ_d be the average of Σ_τ over the period $d+1, \ldots, d+21$, where $d = \{21s - 20L\}$ and $s = L, \ldots, (T+20L)/21 - 1$, and w_d be the portfolio weights computed according to equation (8), but using Σ_d rather than $\hat{\Sigma}_d^i$. Using the law of iterated expectations, we can show that

$$\text{var}\left[w_d' \left(r_\tau - \mu_r\right)\right] := \sigma^2 = E\left[\sigma_\tau^2\right], \tag{10}$$

where var$[\cdot]$ is the unconditional variance operator, $E[\cdot]$ is the unconditional expectation operator, and

$$\sigma_\tau^2 := w_d' \Sigma_\tau w_d. \tag{11}$$

Since the objective in equation (4) is to find the combinations of assets that have a minimum variance (out-of-sample) under the constraint $\mathbb{1}_N' w_d$, we define the following function to quantify the quality of the covariance matrix estimator:

$$\mathcal{L}_i = \frac{1}{T-L} \sum_{\tau=L+1}^{T} \left(\sigma_{\tau,i}^2 - \sigma_\tau^2\right), \tag{12}$$

where $\sigma_{\tau,i}^2$ is defined analogously to equation (11), but using w_d^i (i.e., portfolio weights implied by model i) instead of w_d. A similar loss function is defined by Engle et al. (2019), who provide a detailed justification for its scientific usefulness. The following proposition shows that equation (12) possesses the desired property of a loss function, that is, the loss is strictly positive only if the covariance estimator has error in it and is zero otherwise.

Proposition 1. *$\mathcal{L}_i \geqslant 0$; if for each τ there exists an invertible matrix, Π, such that $\Pi' \Sigma_\tau \Pi = \Sigma_\tau^i$, where $\Sigma_\tau^i = \beta_i \Omega_\tau^i \beta_i' + V_\tau^{full}$, then $\mathcal{L}_i = 0$ if and only if $\Pi = I_N$ $\forall \tau$, where I_N is the $N \times N$ identity matrix.*

Proof. See Appendix B. ■

An immediate implication of the above result is that

$$E\left[\frac{1}{T-L} \sum_{\tau=L+1}^{T} \sigma_{\tau,i}^2\right] \geqslant E\left[\frac{1}{T-L} \sum_{\tau=L+1}^{T} \sigma_\tau^2\right].$$

Using the linearity of expectation and the relation specified in equation (10), which also holds for $\sigma_{\tau,i}^2$, we can show that the above inequality implies that $\sigma_i^2 \geqslant \sigma^2$. Thus, σ^2 is a lower bound of σ_i^2, the (unconditional) out-of-sample variance of a portfolio constructed using the estimator of the covariance matrix of asset returns implied by model i. So, if $\Sigma_\tau^i = \Sigma_\tau$, then $w_d^i = w_d$ and $\sigma_i^2 = \sigma^2$, whereas if $\Sigma_\tau^i \neq \Sigma_\tau$, then $\sigma_i^2 > \sigma^2$. In other words, the smaller the error in the estimator, the more precise the forecast of the covariance matrix, which in turn translates into a smaller out-of-sample variance of the constructed portfolio. So, if a model i does a better job in capturing the systematic portion of the covariance matrix of returns than model j, then Σ_τ^i will be a better estimator of Σ_τ than Σ_τ^j, which will eventually result in σ_i^2 being closer to σ^2 than σ_j^2. Therefore, σ_i^2 will be lower than σ_j^2. Note that the matrix Π in Proposition 1 is guaranteed to exist and is indeed unique positive definite, since both Σ_τ^i and Σ_τ are positive definite.[10]

In light of the above discussion, an appropriate statistic for comparing the competing models' ability to capture the systematic return covariation in an out-of-sample setting is

$$\triangle_{ij} := \sigma_i^2 - \sigma_j^2 \quad \text{for all } i, j \in \mathcal{M}^0, \tag{13}$$

where σ_i^2 and σ_j^2 are the (unconditional) variances of the out-of-sample returns on the minimum variance portfolios implied by factor models i and j, respectively. In the cases of the tracking error minimizing portfolios, σ_i^2 and σ_j^2 are, respectively, the (unconditional) variances of the out-of-sample excess returns (over the benchmark's return) realized on the portfolios implied by factor models i and j. Alternatively, one can consider using the logarithmic version of equation (13),

$$\mathit{\Delta}_{ij} = \ln(\sigma_i^2) - \ln(\sigma_j^2) \quad \text{for all } i, j \in \mathcal{M}^0. \tag{14}$$

It is worth highlighting that $\mathit{\Delta}_{ij}$ is the logarithmic transformation of the F-test for the equality of variances. As Efron (1982) stresses, the transformation is both normalizing and variance stabilizing, both of which are conducive to better finite-sample properties of our inference methods. This is the quantity that we use in our empirical application.

The choice of the above metric can further be justified by the fact that we consider the minimum variance portfolio, which is designed to minimize the variance rather than to maximize the expected return or the Sharpe ratio. Also, a large literature (see Engle, Ledoit, and Wolf, 2019; Ledoit and Wolf, 2017) emphasizes that the most important performance measure in the context of the minimum variance portfolio is the out-of-sample variance. Although a higher out-of-sample Sharpe ratio is desirable, it should be of secondary importance for evaluating the quality of the covariance matrix estimator. Furthermore, Engle and Colacito (2006, Corollary

[10] It can be shown that $\Pi = \Sigma_\tau^{-1/2}(\Sigma_\tau^{1/2}\Sigma_\tau^i\Sigma_\tau^{1/2})^{1/2}\Sigma_\tau^{-1/2}$ (see Bueno et al., 2007, and the references therein).

2 of Proposition 1) argue that even in the context of a Markowitz portfolio, the out-of-sample standard deviation (and thus the variance) is still a preferred performance measure, since the Sharpe ratio can potentially lead to the selection of the wrong estimator.

2.3.1 Pairwise test

We first compare models on a pairwise basis. Note that the statistic, Δ_{ij}, will be strictly negative (positive) only if a factor model i does a better (worse) job in capturing the covariance structure of test asset returns out-of-sample than a competing model, j. This entails testing $H_{0,ij} : \Delta_{ij} = 0$ against $H_{A,ij} : \Delta_{ij} \neq 0$. The associated test statistic is given by

$$\eta_{ij} = \frac{\hat{\Delta}_{ij}}{\hat{\sigma}_{\hat{\Delta}}},$$

where $\hat{\Delta}_{ij}$ is an estimate of Δ_{ij} and $\hat{\sigma}^2_{\hat{\Delta}}$ is a consistent estimate of the variance of $(T-L)^{-1/2}\hat{\Delta}_{ij}$. Under some suitable conditions, Ledoit and Wolf (2011) point out that $\hat{\Delta}_{ij}$ is approximately normally distributed and thus the limiting distribution of η_{ij} is a standard normal. To improve inference accuracy, we use the studentized bootstrap method of Ledoit and Wolf (2011) to generate B bootstrap samples of portfolio returns with length $\bar{T} = T - L$ (i.e., the out-of-sample period). But in the cases of the minimum tracking error portfolios implied by the factor models, we create B bootstrap samples of portfolio returns in excess of the benchmark's return.

The actual implementation of the bootstrap is identical to the studentized bootstrap procedure outlined in Section 3.2.2 of Ledoit and Wolf (2008). We employ $B = 9,999$ resamples and report the bootstrap p-values computed in accordance to Remark 3.2 of Ledoit and Wolf (2008). Also, as in Algorithm 3.1 of Ledoit and Wolf (2008), we pick a data-dependent block size for the circular block bootstrap, by fixing the selection of block sizes to $\{1, 2, 4, 6, 8, 10\}$ and generating 5,000 pseudo sequences from a bivariate GARCH model. Our empirical results remain robust if we go beyond the input block size of 10 (for example, 12).

2.3.2 Simultaneous test

The pairwise model comparison may not always determine unambiguously the best performing model when multiple models are involved in the horse race (Barillas et al., 2020; Gospodinov et al., 2013; Kan et al., 2013). This is due to the over-rejections of the null of equal model performance arising from the process of sequentially searching for the best model across several alternatives. To alleviate this concern, we utilize the MCS procedure of Hansen et al. (2011) to perform simultaneous comparison of factor models. The MCS procedure determines the best factor model(s) from a collection of models, $\mathcal{M}^0$, with a given level of confidence. Similar to Definition 1 in Hansen et al. (2011, p. 458), we define the set of superior models, $\mathcal{M}^*$, as

$$\mathcal{M}^* \equiv \{i \in \mathcal{M}^0 : \mathit{\Delta}_{ij} \leqslant 0 \text{ for all } j \in \mathcal{M}^0\}.$$

The MCS procedure determines $\mathcal{M}^*$ through a sequence of significance tests, where models that are found to be significantly inferior to other competing models in $\mathcal{M}^0$ are eliminated. The null hypotheses that are being tested take the following form

$$H_{0,\mathcal{M}} : \mathit{\Delta}_{ij} = 0 \text{ for all } i, j \in \mathcal{M},$$

where $\mathcal{M} \subset \mathcal{M}^0$. Thus, starting with $\mathcal{M} = \mathcal{M}^0$, $H_{0,\mathcal{M}}$ is tested using an equivalence test, $\xi_{\mathcal{M}}$, at significance level ϱ. If the null hypothesis is not rejected, then $\mathcal{M}^*$ is defined as $\widehat{\mathcal{M}}^*_{1-\varrho} = \mathcal{M}$. Otherwise, a model from $\mathcal{M}$ is eliminated using some elimination rule, $e_{\mathcal{M}}$, and the testing procedure is repeated. A factor model that survives all the significance tests is contained in $\widehat{\mathcal{M}}^*_{1-\varrho}$ and is said to be superior to the eliminated models at significance level ϱ.

Hansen et al. (2011) emphasize that when the MCS procedure is implemented in finite samples, it is desirable to have a certain coherency between the hypothesis test, $\xi_{\mathcal{M}}$, and the elimination rule, $e_{\mathcal{M}}$. However, hypothesis testing that relies on asymptotic results cannot guarantee such coherency, unless the test and the elimination rule are chosen so that $P(\xi_{\mathcal{M}} = 1, e_{\mathcal{M}} \in \mathcal{M}^*) \leqslant P_0(\xi_{\mathcal{M}} = 1)$, where P is the true probability measure and P_0 is a sample transformation of P that satisfies the null hypothesis (see Hansen et al., 2011, Definition 3, p. 461). Our implementation of the MCS procedure is based on a multiple t-statistic approach that simplifies the choice of $e_{\mathcal{M}}$ in order to satisfy the notion of coherency.

Recall from the preceding section that the t-statistic associated with the null hypothesis $H_{0,ij} : \mathit{\Delta}_{ij} = 0$ is given by η_{ij} (see Section 2.3.1). Since $H_{0,\mathcal{M}} \Leftrightarrow \{H_{0,ij}$ for all $i, j \in \mathcal{M}\}$, a natural statistic for testing the hypothesis $H_{0,\mathcal{M}}$ is

$$\Gamma_{\mathcal{M}} = \max_{i,j \in \mathcal{M}} |\eta_{ij}|. \tag{15}$$

An elimination rule for this test statistic that will satisfy the coherency principle is

$$e_{\mathcal{M}} \equiv \arg\max_{i \in \mathcal{M}} \sup_{j \in \mathcal{M}} \eta_{ij}. \tag{16}$$

Thus, to summarize, the algorithm for constructing the MCS is as follows:

Step 1: Set $\mathcal{M} = \mathcal{M}^0$.

Step 2: Test $H_{0,\mathcal{M}}$ using $\Gamma_{\mathcal{M}}$ at level ϱ. If $H_{0,\mathcal{M}}$ is "accepted", define $\widehat{\mathcal{M}}^*_{1-\varrho} = \mathcal{M}$. Otherwise, use $e_{\mathcal{M}}$ to eliminate an object from $\mathcal{M}$ and repeat Step 2.

The asymptotic distribution of $\Gamma_{\mathcal{M}}$ under both the null and the alternative is nonstandard. Therefore, to be able to implement the MCS procedure for simultaneous comparison of models, we apply the block bootstrap method of Hansen et al. (2003, 2011) to the portfolio returns (or to the portfolio returns in excess of the benchmark's return in the cases of the minimum tracking error portfolios) in order to estimate the distribution of $\Gamma_{\mathcal{M}}$. We construct 9,999 bootstrap samples and set the block length equal to 10, which is a standard choice in the literature (see, for example, Liu et al., 2015). Also, our estimation reveals that this is by far the most frequent value for the optimized block size in pairwise comparisons, based on the algorithm of Ledoit and Wolf (2011).[11] The bootstrap implementation is also convenient for calculating the MCS p-values for all models under consideration. The interpretation of the MCS p-value is analogous to that of a classical p-value. In particular, let $\hat{p}_i$ be the MCS p-value for model i, such that i is included in $\widehat{\mathcal{M}}^*_{1-\varrho}$ if and only if $\varrho \leqslant \hat{p}_i$ (see Hansen et al., 2011, Theorem 3, p. 462).

3. Factor models and test assets

3.1 Competing models

We investigate the ability of 11 different models to capture the systematic portion of return variation. These are: (1) the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965), which includes a market (excess return) factor (MKT); (2) the Fama and French (1993) three-factor (FF3) model, which augments the CAPM by including empirically motivated size (SMB*, small minus big) and value (HML, high minus low book-to-market equity) factors; (3) the Fama and French (1993) and Carhart (1997) four-factor (FFC) model, which adds a momentum (UMD, up minus down) factor to the FF3 model; (4) the Asness and Frazzini (2013) three-factor (FFAF) model, which adds a more "timely" version of the value factor (HML^{m}) to the market and size factors of the FF3 model; (5) the Hou et al. (2015) four-factor q (HXZ) model, comprising market, size (r_{ME}), investment ($r_{\text{I/A}}$), and profitability (r_{ROE}) factors; (6) the Fama and French (2015) five-factor (FF5) model, consisting of market, size (SMB), value, profitability (RMW, robust minus weak), and investment (CMA, conservative minus aggressive) factors;[12] (7) the four-factor (FF4) model, which drops the value factor from the FF5 model;

[11] In their Appendix, Hansen et al. (2011) suggest using different choices for the MCS block length and verify that the result is not sensitive to the choice. We therefore experiment with alternative values for the block size, such as 5, 12, and 15. Our results are very similar in all cases and are available upon request.

[12] Adopting a modification to the original size factor, SMB*, Fama and French (2015) create an alternative version of the size factor, denoted SMB, which is the difference between the average of the value-weighted returns

(8) the Stambaugh and Yuan (2017) four-factor (SY4) model, which comprises market, size (SMB_M), and two mispricing factors, management (MGMT) and performance (PERF); (9) the Barillas and Shanken (2018) six-factor (BS6) model, which embeds the market and size factors of the FF5 model, the profitability and investment factors of the HXZ model, the momentum factor of the FFC model, and the value factor of the FFAF model; (10) the Fama and French (2018) six-factor (FF6) model, which augments the FF5 model by incorporating the momentum factor of the FFC model; and (11) the Daniel et al. (2020) three-factor (DHS) model, comprising market, financing (FIN), and post-earnings announcement drift (PEAD) factors.

We obtain daily data on the risk-free rate, and the MKT, SMB*, SMB, HML, UMD, RMW, and CMA factors, from the Internet Data Library provided by Kenneth French.[13] The daily data on the r_{ME}, $r_{I/A}$, and r_{ROE} factors are sourced from Lu Zhang's website,[14] while the data on the HML^m factor are from the AQR Data Library.[15] We collect data on the size factor, SMB_M, and the mispricing factors, MGMT and PERF, from Robert Stambaugh's website.[16] Since daily data on the behavioral factors, FIN and PEAD, are not available in the public domain, we follow the sample criterion and factor construction procedure in Daniel et al. (2020) to reproduce FIN and PEAD. Our sample period for model comparison spans January 3, 1972, to December 31, 2018.

3.2 Test assets and general portfolio-construction rules

To assess the out-of-sample performance of our factor pricing models in a realistic setting, choosing test assets is a challenging task. On the one hand, the use of well-diversified portfolios as test assets reduces estimation error in the covariance matrix of returns but fails to fully characterize ex ante expected return dispersion in the economy (Moskowitz, 2003). On the other hand, individual stocks are prone to generating large estimation error in the return covariance matrix but better characterize the true cross-section of expected returns (Ang et al., 2020). Given this trade-off, we utilize both individual stocks and well-diversified portfolios, separately, as test assets. For the sets of individual stocks, we begin with a preliminary sample, including all NYSE-, AMEX-, and NASDAQ-listed ordinary common stocks (CRSP share codes 10 and 11). Following Jagannathan and Ma (2003), stocks with prices below \$5 and stocks with a market value of equity below the 20th percentile of the NYSE market capitalization are excluded from

on the nine small stock portfolios of the three independent 2×3 sorts (on size and book-to-market equity, size and operating profitability, and size and investment) and the average of the value-weighted returns on the nine big stock portfolios of the three independent 2×3 sorts. For details, see Fama and French (2015).

[13] See `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

[14] See `https://sites.google.com/site/theqfactormodel/?pli=1`.

[15] See `https://www.aqr.com/library/data-sets/the-devil-in-hmls-details-factors-monthly`.

[16] See `https://finance.wharton.upenn.edu/~stambaug/`. We generate daily data beyond December 30, 2016, by using the sample criterion and factor construction procedure described in Stambaugh and Yuan (2017).

this preliminary sample. Also, stocks of financial and heavily regulated firms (with primary standard industrial classification codes between 6000 and 6999 and between 4900 and 4999) are excluded. For the remaining stocks, we collect their daily return data from CRSP.[17]

At any portfolio construction date, we compute the out-of-sample (averaged) forecast of the covariance matrix implied by each competing factor model using a rolling window of most recent 1,260 daily excess returns. This roughly corresponds to using five years of past data.[18] The forecast is the input to a quadratic programming routine for portfolio optimization. As outlined in Section 2.1, we update all portfolios at the end of every 21-day period. The out-of-sample model comparison period starts on December 30, 1976, and ends on December 31, 2018. For a given combination of factor model and set of test assets, this results in a total of 504 forecasts of future covariance matrices of excess returns and 10,584 daily out-of-sample portfolio returns.

To further select the sets of stocks from the filtered preliminary sample, we follow Ledoit et al. (2019). At each portfolio construction date, we identify all stocks with a complete return history for the upcoming 21 trading days and at most 2.5% missing returns over the most recent 1,260 days.[19] We then look for pairs of stocks for which sample return correlation exceeds 0.95 over the past 1,260 days. When such a pair is detected, the stock with the lower market capitalization of the two is dropped from the sample considered for the specific updating date. Of the remaining stocks, we then: (1) construct a set of 50 small stocks that are randomly selected;[20] and (2) pick the largest N stocks based on their market capitalization on the portfolio construction date. In the case of the largest N stocks, three sets of individual stocks are selected for which N=100, 250, and 500. This rule of selecting stocks has several advantages. First, it does not involve selecting individual stocks at random and therefore is expected to give more stable results. Second, allowing for increasing values of N by a well-defined rule would uncover the ability of the competing models to capture the systematic component of return covariation when N varies. Third, stocks with a large market capitalization tend to have a low bid-ask spread and a high depth in the order book, which allows (large/institutional) investors to heavily invest in them without breaching standard safety guidelines of the Exchange regulator.

We also make use of daily returns on two sets of value-weighted portfolios: (1) 48 industry portfolios; and (2) 340 portfolios from those used in Giglio and Xiu (2021).[21] The inclusion of

[17] The use of daily data to characterize the covariance structure of returns offers a reasonable balance when it comes to circumventing the issues associated with microstructure effects and structural breaks.

[18] We also consider a 1000-day rolling estimation window and find that the best performing models are the same as those reported in this paper. These additional results are available from the authors upon request.

[19] As in Ledoit et al. (2019), missing values of returns are replaced by zero.

[20] Small stocks are stocks with a market value of equity between the 20th and 50th percentiles of the NYSE market capitalization.

[21] We pick 340 portfolios from those used in Giglio and Xiu (2021) based on the criterion that they have a return history from January 3, 1972, to December 31, 2018.

these sets of portfolios is also in line with the advice of Lewellen et al. (2010) to avoid test assets with a strong factor structure. We source daily data on the 48 industry portfolios from Kenneth French's website, while the data on the 340 Giglio-Xiu portfolios from Lu Zhang's Global-q Data Library.[22] In the cases of the minimum tracking error portfolios, following Jagannathan and Ma (2003), we assume that the investor tracks the return of the S&P 500 index.

4. Empirical results[23]

4.1 Minimum variance portfolios[24]

We summarize the pairwise variance equality test results for all the asset pricing models under consideration in Table 1.[25] Our criterion for overall performance evaluation is the number of times a competing factor model generates a significantly lower out-of-sample return variance, of the corresponding optimized portfolio, than does any other model. For the minimum variance portfolios in Panel A, it can be seen that the FF5, BS6, and FF6 models offer the best overall performance, in that they statistically outperform competing asset pricing models the most number of times (i.e., 46 in total for each of the three models) across the six different sets of test assets for investment. A similar finding emerges for the restricted minimum variance portfolios in Panel B.[26] Analyzing the results for the long-only minimum variance portfolios in Panel C, we once again find that the FF5 and BS6 models secure the top position. But this time the HXZ and FF6 models turn out jointly to be the next best models. Aside from these models, the FF4 model shows a somewhat commendable performance. It is worth highlighting that the BS6 model is never statistically outperformed in any of our pairwise variance equality tests. Moreover, the CAPM exhibits the worst out-of-sample performance, followed by the DHS model. Although these findings are quite informative about the relative performance of the competing factor models to capture the systematic return covariation, they are subject to the criticism that the pairwise model comparisons do not take into account the process of

[22] See http://global-q.org/testingportfolios.html.

[23] Throughout this paper, we reject $H_{0,ij} : \Delta_{ij} = 0$ if the corresponding bootstrap p-value is at most 0.05.

[24] In the context of the minimum variance portfolios, the most important performance metric is the out-of-sample variance of returns, which we utilize to run a horse race of the 11 competing factor models. Of course, the other measures, such as average return, Sharpe ratio, leverage, and concentration, provide valuable information. But they are of secondary importance when judging the ability of a factor model to capture the systematic component of future covariance matrix that aids in creating more efficient portfolios ex post. Thus, in the interest of streamlining the empirical analyses, we limit the discussion to the variance equality test results and delegate the estimates of the other metrics for the optimized portfolios to Tables IA7–IA9 of the Internet Appendix.

[25] For brevity, we report detailed results of the pairwise tests in Tables IA1–IA3 of the Internet Appendix.

[26] This is not surprising because for a given combination of test assets and factor model, the restricted minimum variance portfolio weights turn out to be identical to those for the minimum variance portfolio counterpart (see Tables IA7 and IA8 of the Internet Appendix).

searching across alternative factor models. As a result, the underlying testing procedure may not always determine unambiguously the best-performing model when multiple models are involved in the competition. To account for this possibility and to provide a more robust analysis of the best-performing asset pricing model(s) out-of-sample, we rely on the MCS results.

Table 2 reports the MCS p-values for each of the competing factor models using a given set of test assets. To be consistent with our preceding analyses, we consider $\varrho = 5\%$ to make inferences on the models that end up in the MCS with a 95% level of confidence, $\widehat{\mathcal{M}}^*_{95\%}$. The results in Panel A for the minimum variance portfolios show that only the FF5, BS6, and FF6 models are contained in $\widehat{\mathcal{M}}^*_{95\%}$ for every set of test assets and therefore can be regarded as the best-performing models. The FF5, BS6, and FF6 models are also included in $\widehat{\mathcal{M}}^*_{95\%}$ for every set of test assets when we consider the restricted minimum variance portfolios in Panel B. The results for the long-only minimum variance portfolios in Panel C show that the HXZ, FF5, FF4, BS6, and FF6 models reside in $\widehat{\mathcal{M}}^*_{95\%}$ for every set of test assets. Although five out of the 11 competing models end up in $\widehat{\mathcal{M}}^*_{95\%}$ regardless of the set of test assets, the FF5 model appears to have the largest MCS p-value for three out of the six sets. Informally, the FF5 model can be viewed as the best among these best-performing models. The findings in Panel C are not unusual given the fact that nonnegativity constraints on portfolio weights are in place, which potentially strengthens the dominance of the market factor. As a result, the competing models do not differ much from each other in terms of minimizing portfolio variance and the MCS contains several of them. This is consistent with Hansen et al. (2011) who point out that the MCS may ends up having several (or possibly all) models when the data are less informative to distinguish between models. The results in Panels A, B, and C also show that a majority of the competing factor models end up in $\widehat{\mathcal{M}}^*_{95\%}$ when the investment universe comprises small stocks, regardless of which variant of portfolio is considered. Yet the FF5, BS6, and FF6 models have much higher probabilities than the rest of the factor models. Although the findings in Table 2 are slightly different from those summarized in Table 1, any differences can be rationalized by the fact that the pairwise testing procedure can lead to an overstatement of statistical significance. We acknowledge that although model comparison can be sensitive to the test assets, our findings are robust across the diverse sets of test assets.

4.2 Minimum tracking error portfolios[27]

Table 3 summarizes the pairwise variance equality test results. Similar to the criterion set out in Section 4.1, we determine the overall performance of a factor model by the number of

[27] We report detailed results of the pairwise tests for the minimum tracking error portfolios in Tables IA4–IA6 of the Internet Appendix. Also, estimates of other metrics can be found in Tables IA10–IA12.

times it generates a significantly lower out-of-sample variance for the portfolio excess returns (over the benchmark's return) than does any other model. We find that the FF6 model offers the best overall performance regardless of the set of test assets and the version of the minimum tracking error portfolios. The only exception to this is the sample of 340 Giglio-Xiu portfolios as test assets. The BS6 model, whose relative performance falls short marginally in pairwise tests compared to the FF6 model, turns out to be the next best factor model in the horse race. As noted earlier that the pairwise testing can lead to an overstatement of statistical significance, we further conduct empirical analyses based on the MCS procedure of Hansen et al. (2011).

Table 4 presents the results of applying the MCS criterion to determine the best model. Panels A, B, and C consider the minimum tracking error, the restricted minimum tracking error, and the long-only minimum tracking error portfolios implied by the factor models, respectively. In every panel, we see that only the BS6 and FF6 models end up in the MCS with a 95% level of confidence, $\widehat{\mathcal{M}}^*_{95\%}$, irrespective of the set of test assets for investment. Thus, the BS6 and FF6 models can be regarded as the best-performing models. Recall that when forming the long-only minimum variance portfolios in Panel C of Table 2, five out of the 11 competing models end up in $\widehat{\mathcal{M}}^*_{95\%}$ for every set of test assets. Now, only the BS6 and FF6 models are contained in $\widehat{\mathcal{M}}^*_{95\%}$ for every set of test assets. This empirical finding clearly justifies our motivation to formally conduct model comparison using the tracking error minimizing portfolios. By design, these portfolios are expected to mitigate the possible dominant influence of the market factor and thus to generate starker differences across the models in terms of the tracking error variance. In Table 4, the MCS results also show that most of the factor pricing models perform equally well when the investment universe is based on the sample of small stocks.

4.3 Summary of model performance

In the preceding sections, we have carried out extensive statistical analyses on the ability of 11 factor models to capture the systematic return covariation, by focusing on the economic implications for optimized portfolios. The results based of the MCS procedure in Table 2 show that the FF5, BS6, and FF6 models are contained in $\widehat{\mathcal{M}}^*_{95\%}$ regardless of the set of test assets to form the minimum variance, the restricted minimum variance, and the long-only minimum variance portfolios. The well-known HXZ and FF4 models end up in $\widehat{\mathcal{M}}^*_{95\%}$ irrespective of test assets but only for the long-only minimum variance portfolios. Given these findings, the FF5, BS6, and FF6 models can be regarded as the best models, in that they reside in $\widehat{\mathcal{M}}^*_{95\%}$ for all cases. That is, the three factor models above demonstrate the best overall performance among all competing models in capturing the systematic return covariation, as reflected by a significantly lower out-of-sample variance of the corresponding portfolios. We caution that our results do not imply that the remaining eight models are not capable of capturing the systematic return covariation. Instead, they simply suggest that these models are not as successful as the FF5, BS6, and FF6 models. When we consider the results for the minimum tracking error, the restricted minimum tracking error, and the long-only minimum tracking error portfolios implied by the factor models in Table 4, we find that the BS6 and FF6 models always end up in $\widehat{\mathcal{M}}^*_{95\%}$. But the FF5 model fails to appear as one of the top performing models. Recently, Ahmed et al. (2019) compare major factor models including the CAPM, FF3, FFC, FFPS (which combines a traded liquidity factor of Pástor and Stambaugh (2003) with those of the FF3 model), FFAF, HXZ, FF5, FF4, SY4, and BS6 models. The authors find that the HXZ, FF5, FF4, and BS6 models are the top performers in explaining several well-known return anomalies. Our results confirm that models that are successful in this context are not necessarily successful for capturing systematic return covariation, which is central to applications of asset pricing models.

5. Further analysis

5.1 Transaction costs

In this section, we conduct additional analyses using the out-of-sample returns net of transaction costs for the portfolios implied by the competing models. To do so, we begin with the approach of Ledoit et al. (2019). Specifically, during a month, from one day to the next, we hold the number of shares fixed rather than the portfolio weights. In this manner, there are no transactions at all during a month.[28] Transaction costs arise only from updating the portfolio at the beginning of a new month. Given the above setup, let $r^{SF}_{p,d+1}$ be the return on the minimum variance portfolio at day $d+1$. We now follow Brandt et al. (2009) to adjust transaction costs. At day $d+1$, the return to the minimum variance portfolio net of trading costs is computed as

$$r_{p,d+1} = r^{SF}_{p,d+1} - c\|\hat{w}_d - \hat{w}^{hold}_{d-1}\|_1,$$

where c denotes the constant proportional transaction costs, $\|\cdot\|_1$ denotes the ℓ_1-norm, and $\hat{w}^{hold}_d$ denotes the vector of the "hold" portfolio weights at the end of period d, which is determined by the initial vector of portfolio weights, $\hat{w}_d$, together with the evolution of the prices of the N assets in the portfolio during period d. The excess return (over the benchmark) to the minimum tracking error portfolio net of transaction costs, at day $d+1$, is estimated in a similar fashion. As in Brandt et al. (2009), we assume transaction costs to be at 0.5%.

[28] We are grateful to Michael Wolf for sharing the MATLAB code on return estimation.

Table 5 reports the MCS p-values for the minimum variance (Panel A), the restricted minimum variance (Panel B), and the long-only minimum variance (Panel C) portfolios net of trading costs. Consistent with our preceding finding in Table 2, we see that only the FF5, BS6, and FF6 models are contained in $\widehat{\mathcal{M}}^*_{95\%}$ regardless of the set of test assets to form the factor model-implied portfolios. This suggests that the FF5, BS6, and FF6 models continue to be the top performers when we account for the impact of transaction costs on portfolio returns. That is, the FF5, BS6, and FF6 models perform best in capturing the systematic return covariation. We report the MCS p-values for the minimum tracking error (Panel A), the restricted minimum tracking error (Panel B), and the long-only minimum tracking error (Panel C) portfolios net of transaction costs in Table 6. As in Table 4, it is observable that the BS6 and FF6 models are the best models, in that they reside in $\widehat{\mathcal{M}}^*_{95\%}$ for all cases. Reassuringly, our findings in Tables 5 and 6 remain robust to proportional transaction costs of 1% (not tabulated).

5.2 Comparing models recursively

So far, we have compared models, on a pairwise basis as well as simultaneously, using data ending on December 31, 2018. Given that most of our models have been introduced only in the recent years, researchers and practitioners back in the 1980s and 1990s would not consider investment and profitability factors as part of the true model, let alone mispricing factors of Stambaugh and Yuan (2017) or behavioral factors of Daniel et al. (2020). To uncover the best model(s) over time, we follow Feng et al. (2020) and conduct a recursive exercise in which models are compared as they are introduced against previously proposed models. Specifically, we compare the out-of-sample variances of the portfolios implied by each new model with those implied by the model(s) published in same or previous years, using only the data up to the publication year. For example, the FF3 model was published in 1993. So, we begin with comparing the CAPM and FF3 models by using data from January 3, 1972 to December 31, 1993. In this case, the out-of-sample period is from December 30, 1976, to December 31, 1993. We then compare the CAPM, FF3, and FFC models by using data from January 3, 1972, to December 31, 1997, since the FFC model was published in 1997. In this case, the out-of-sample period spans December 30, 1976, to December 31, 1997. We continue model comparison recursively until we reach the end of the full sample period. In the end, we have all but the DHS model in the horse race, since it was published in 2020.

Table IA13 in the Internet Appendix reports the MCS p-values for the sequential equality tests of the out-of-sample variances of the optimized portfolios implied by the factor models. The results suggest that the FF5, BS6, and FF6 models remain as the best-performing models since their introduction in the literature, which is consistent with our earlier finding in Section 4.1. We present the MCS p-values for the tracking error minimizing portfolios, implied by the competing factor pricing models, in Table IA14 of the Internet Appendix. In all cases, the results show that only the BS6 and FF6 models survive as the best models since their publication, which is clearly in line with our preceding empirical finding in Section 4.2.

5.3 Economic gains

We next evaluate the economic gains from using the factor models under consideration in an out-of-sample Value-at-Risk (VaR) context. The VaR is widely used in practice to measure the maximum expected loss of an investment (expressed in monetary value or in percent), over a certain period of time, and with a given level of confidence. For example, a 1-day 95% VaR equal to −2%, means that the investor is 95% confident that the loss of her portfolio will not exceed −2% over the next day.

The VaR is calculated as the value of the α% level of the return distribution (where α is the chosen level of significance). Our VaR implementation follows closely Engle (2002). Specifically, the $(100 - \alpha)$ percentile VaR for model i on day d is computed as follows:

$$VaR_d^i = F_\alpha^{-1}\sqrt{\hat{w}_d^{i\prime}\hat{\Sigma}_d^i\hat{w}_d^i}, \tag{17}$$

where F_α^{-1} is the critical value from an inverse normal distribution corresponding to the α% level, while $\hat{w}_d^{i\prime}$ are the optimized portfolio weights from factor model i on day d (see equation (4)), and $\hat{\Sigma}_d^i$ is the matrix of day-ahead out-of-sample covariance forecasts from factor model i. We focus on the 1-day VaR as this is a very common choice by academics and practitioners (see, for example, Engle, 2002; Ferreira and Lopez, 2005).[29] We then calculate for each model i, the number of VaR exceedances, often called *Hit*, as follows

$$Hit_d^i = \mathbb{I}_{\{r_d^i < VaR_d^i\}}, \tag{18}$$

where r_d^i is the return of the portfolio based on factor model i on day d, computed as $r_d^i = (\hat{w}_d^i)'r_d$ (where r_d are the asset returns on day d). The above indicator variable takes the value of 1, if the return of portfolio constructed based on factor model i is lower than the specified VaR figure and is zero otherwise. Essentially, it measures the number of $(100 - \alpha)$ percentile VaR exceedances.

Intuitively, a model is better in economic terms than a competing model, if its average

[29] As an additional robustness check, we consider the 10- and 21-day VaR estimates. The results are qualitatively similar to those of the 1-day VaR estimates and are available from the authors upon request.

number of exceedances, that is, $\sum_{d=1}^{D} Hit_d/D$ (where D is the number of out-of-sample VaR estimates), is lower than that of the competing model and Hit_d is as close to α as possible (ideally lower). Tables IA17 and IA18 of the Internet Appendix report the frequency of exceedances (Hit) for the 95% VaR under the minimum variance portfolios and the tracking error minimizing portfolios, respectively.[30] Overall, the results suggest that for the minimum variance portfolios, the FF5, BS6, and FF6 models are the best, whereas the BS6 and FF6 models are the top performers for the tracking error portfolios. These findings are consistent with those reported in Section 4.

5.4 Sensitivity and subperiod analyses

To verify that our results are not driven by specific episodes, we conduct empirical analyses in which data for the global financial crisis period (i.e., from August 1, 2008, to December 31, 2009) are excluded from the estimation of each model-implied covariance matrix forecasts. The global financial crisis is by far the most extreme event during our sample period. Table IA19 in the Internet Appendix reports the MCS p-values for the sequential equality tests of the out-of-sample variances of the portfolios implied by the factor models. Consistent with the results in Table 2, we see that only the FF5, BS6, and FF6 models are contained in $\widehat{\mathcal{M}}^*_{95\%}$ regardless of the set of test assets used to form the model-implied portfolios. When we consider the results for the tracking error minimizing portfolios implied by the competing factor models (in Table IA20 of the Internet Appendix), we find that the BS6 and FF6 models always end up in $\widehat{\mathcal{M}}^*_{95\%}$.

Finally, we conduct robustness checks of the results in Section 4 to specific subperiods. Specifically, we examine the ability of factor models to capture the systematic return covariation in the bull and bear markets. In the spirit of Daniel and Moskowitz (2016), we define all days of a given month in the out-of-sample period as the bear (bull) market observations if at the end of the previous month the cumulative CRSP value-weighted index (daily) return in the past 2 years is negative (positive or zero). The results in Tables IA21 and IA22 of the Internet Appendix, respectively, for the bull and bear markets show that the FF5, BS6, and FF6 models are always contained in $\widehat{\mathcal{M}}^*_{95\%}$ regardless of the set of test assets to form the model-implied variance minimizing portfolios. We report the MCS p-values of the tracking error minimizing portfolios for the bull and bear markets in Tables IA23 and IA24 of the Internet Appendix, respectively. Consistent with our preceding finding in Table 4, we see that the BS6 and FF6 models are the best models, in that they are included in $\widehat{\mathcal{M}}^*_{95\%}$ for all cases.[31]

[30] We also repeat the exercise for the 90% and 99% VaR and obtain qualitatively similar findings.

[31] Our empirical findings in Tables IA21 through IA24 also remain robust to recessionary and expansionary periods identified by the National Bureau of Economic Research. For brevity, these additional results are omitted

6. Conclusion

This paper empirically investigates which factor model best captures the common shared variation in asset returns. To do so, we focus on the economic implications for portfolio risk minimization. Logically, the higher the ability of a factor pricing model to capture the systematic portion of the covariance matrix of asset returns, the lower would be the variance of the minimum variance portfolio implied by the model. Our list of competing models comprises the capital asset pricing model of Sharpe (1964) and Lintner (1965), the Fama and French (1993) three-factor model, the Fama and French (1993) and Carhart (1997) four-factor model, the Asness and Frazzini (2013) three-factor model, the Hou et al. (2015) q-factor model, the Fama and French (2015) five-factor model, the four-factor model of Fama and French (2015) that excludes the value factor, the Stambaugh and Yuan (2017) four-factor model, the Barillas and Shanken (2018) six-factor model, the Fama and French (2018) six-factor model, and the Daniel et al. (2020) three-factor model. We assess these models' relative performance in generating a significantly low out-of-sample variance of the model-implied portfolios.

Using diverse sets of test assets for investment, alternative versions of the minimum variance portfolios, and formal statistical procedures, including the pairwise variance equality test and the simultaneous comparison of models, we find that the Fama and French (2015) five-factor model, the Barillas and Shanken (2018) six-factor model, and the Fama and French (2018) six-factor model are the best for capturing the systematic return covariation, as reflected in the magnitudes of the out-of-sample portfolio variances. By extending our empirical analyses to the different versions of the minimum tracking error portfolios, we also observe that the Barillas and Shanken (2018) six-factor model and the Fama and French (2018) six-factor model maintain their performance as the best models. All these findings remain qualitatively the same when we address the effect of transaction costs on model performance, conduct a recursive comparison of models, evaluate economic gains from using factor models, and conduct sensitivity and subperiod analyses. Taken together, our empirical results have important implications not only for academic researchers but also for investors and professional fund managers who may consider using asset pricing models as a parsimonious way to evaluate real-time uncertainty when making investment management decisions and/or modeling future risk exposure.

in this paper but are available from the authors upon request.

Appendix A. Forecasting covariance matrices

In what follows, we use $Diag\{\cdot\}$ to represent the function that transforms a vector into a diagonal matrix. We also omit the factor model subscript i from our notation for convenience.

Let $\Psi_t = (\psi^2_{1,t}, \ldots, \psi^2_{K,t})' \in \mathbb{R}^K$ be a vector of conditional variances of the factor return innovations, f_t, and $Q_t \in \mathbb{R}^{K\times K}$ be the conditional covariance matrix of the devolatized factor return innovations. To characterize the dynamics of the factor covariance matrix, Ω_t, we employ the DCC(1,1) model of Engle (2002). In particular, for each $\psi^2_{k,t}$ ($k = 1, \ldots, K$) we use the GARCH(1,1) specification

$$\psi^2_{k,t} = \lambda_{k,0} + \lambda_{k,1} f^2_{k,t-1} + \lambda_{k,2}\psi^2_{k,t-1}, \tag{A1}$$

where $\lambda_{k,0}$, $\lambda_{k,1}$, and $\lambda_{k,2}$ are the (factor-specific) model parameters. The evolution of Q_t over time is modeled as

$$Q_t = (1 - \gamma_1 - \gamma_2)\,\bar{Q} + \gamma_1 \tilde{f}_{t-1}\tilde{f}'_{t-1} + \gamma_2 Q_{t-1}. \tag{A2}$$

In equation (A2), γ_1 and γ_2 are scalars, and $\bar{Q}$ is the unconditional covariance matrix of $\tilde{f}_t$, where $\tilde{f}_t = (f_{1,t}/\psi_{1,t}, \ldots, f_{K,t}/\psi_{K,t})' \in \mathbb{R}^K$ is the vector of devolatized factor return innovations. Generating z-step ahead forecast of Ω_t from this representation involves the following steps:

Step 1: Using data from day $t - L + 1$ to t, estimate the DCC model parameters in a standard fashion, through a two-step maximum likelihood (ML) procedure.

Step 2: Obtain one-step ahead forecast for each factor k as

$$\hat{\psi}^2_{k,t+1} = \hat{\lambda}_{k,0} + \hat{\lambda}_{k,1} f^2_{k,t} + \hat{\lambda}_{k,2}\psi^2_{k,t}, \tag{A3}$$

where $\hat{\lambda}_{k,0}$, $\hat{\lambda}_{k,1}$, and $\hat{\lambda}_{k,2}$ are, respectively, the ML estimates of $\lambda_{k,0}$, $\lambda_{k,1}$, and $\lambda_{k,2}$. Then, generate z-step ahead factor variance forecast ($z \geqslant 2$) as follows:

$$\hat{\psi}^2_{k,t+z} = \sum_{\zeta=0}^{z-2} \hat{\lambda}_{k,0}(\hat{\lambda}_{k,1} + \hat{\lambda}_{k,2})^{\zeta} + (\hat{\lambda}_{k,1} + \hat{\lambda}_{k,2})^{z-1}\hat{\psi}^2_{k,t+1}. \tag{A4}$$

Step 3: Compute one-step ahead forecast for the conditional covariance matrix of the devolatized factor return innovations as

$$\hat{Q}_{t+1} = (1 - \hat{\gamma}_1 - \hat{\gamma}_2)\bar{Q} + \hat{\gamma}_1 \tilde{f}_t\tilde{f}'_t + \hat{\gamma}_2 Q_t, \tag{A5}$$

where $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are the ML estimators of γ_1 and γ_2, respectively. Next, using the approximation proposed by Engle and Sheppard (2001), that is, $\hat{Q}_{t+1} \approx \hat{R}_{t+1}$ and $\bar{Q} \approx \bar{R}$, we forecast the z-day ahead ($z \geqslant 2$) matrix of conditional quasicorrelations as

$$\hat{R}_{t+z} = \sum_{\zeta=0}^{z-2}(1-\hat{\gamma}_1-\hat{\gamma}_2)\bar{R}(\hat{\gamma}_1+\hat{\gamma}_2)^{\zeta} + (\hat{\gamma}_1+\hat{\gamma}_2)^{z-1}\hat{R}_{t+1}. \tag{A6}$$

Note that in practice, the diagonal elements of $\hat{Q}_{t+1}$ can deviate slightly from one, so before it is used for the estimation of $\hat{R}_{t+z}$ every row and every column has to be divided by the square root of the corresponding diagonal entry. The same adjustment has to be applied to $\bar{Q}$ as well.

Step 4: Equipped with the variance and correlation forecasts from Steps 2 and 3, obtain z-step ahead ($z \geqslant 1$) factor covariance matrix forecast as

$$\hat{\Omega}_{t+z} = Diag\{\hat{\Psi}_{t+z}\}^{1/2}\hat{R}_{t+z}Diag\{\hat{\Psi}_{t+z}\}^{1/2}. \tag{A7}$$

For the constant (across factor models) diagonal matrix V_t^{full}, we obtain z-step ahead forecast for each (n,n)th entry by following Steps 1 and 2 above.

Appendix B. Proof of Proposition 1

The proof is the same as that given for Theorem 1 in Engle and Colacito (2006), with some minor modifications. In particular, we can rewrite equation (12) as

$$\begin{aligned}\mathcal{L}_i &= \frac{1}{T-L}\sum_d\sum_z\left(\frac{\mathbb{1}_N'\left(\Sigma_d^i\right)^{-1}\Sigma_z\left(\Sigma_d^i\right)^{-1}\mathbb{1}_N}{\left[\mathbb{1}_N'\left(\Sigma_d^i\right)^{-1}\mathbb{1}_N\right]^2} - \frac{\mathbb{1}_N'\left(\Sigma_d\right)^{-1}\Sigma_z\left(\Sigma_d\right)^{-1}\mathbb{1}_N}{\left[\mathbb{1}_N'\left(\Sigma_d\right)^{-1}\mathbb{1}_N\right]^2}\right)\\ &= \frac{21}{T-L}\sum_d\left(\frac{\mathbb{1}_N'\left(\Sigma_d^i\right)^{-1}\Sigma_d\left(\Sigma_d^i\right)^{-1}\mathbb{1}_N}{\left[\mathbb{1}_N'\left(\Sigma_d^i\right)^{-1}\mathbb{1}_N\right]^2} - \frac{1}{\mathbb{1}_N'\left(\Sigma_d\right)^{-1}\mathbb{1}_N}\right),\end{aligned} \tag{B1}$$

where Σ_d^i is the covariance matrix estimator implied by model i. Let ϑ_d be a vector of random variables such that $E[\vartheta_d\vartheta_d'] = \Sigma_d$ and define

$$u_d = \mathbb{1}_N'\left(\Sigma_d^i\right)^{-1}\vartheta_d - \mathbb{1}_N'\left(\Sigma_d^i\right)^{-1}\mathbb{1}_N\left[\mathbb{1}_N'\left(\Sigma_d\right)^{-1}\mathbb{1}_N\right]^{-1}\mathbb{1}_N'\left(\Sigma_d\right)^{-1}\vartheta_d. \tag{B2}$$

Because

$$E\left[u_d^2\right] = \mathbb{1}_N'\left(\Sigma_d^i\right)^{-1}\Sigma_d\left(\Sigma_d^i\right)^{-1}\mathbb{1}_N - \left[\mathbb{1}_N'\left(\Sigma_d^i\right)^{-1}\mathbb{1}_N\right]^2\left[\mathbb{1}_N'\left(\Sigma_d\right)^{-1}\mathbb{1}_N\right]^{-1} \geqslant 0, \tag{B3}$$

it follows that $\mathcal{L}_i \geqslant 0$. When $\Pi = I_N$, we have

$$\begin{aligned} \mathbb{1}_N' \left(\Sigma_d^i\right)^{-1} \Sigma_d \left(\Sigma_d^i\right)^{-1} \mathbb{1}_N &= \left[\mathbb{1}_N' \left(\Sigma_d^i\right)^{-1} \mathbb{1}_N\right]^2 \left[\mathbb{1}_N' \left(\Sigma_d\right)^{-1} \mathbb{1}_N\right]^{-1} \\ &= \mathbb{1}_N' \left(\Sigma_d^i\right)^{-1} \mathbb{1}_N \\ &= \mathbb{1}_N' \left(\Sigma_d\right)^{-1} \mathbb{1}_N, \end{aligned} \tag{B4}$$

so that $\mathcal{L}_i = 0$. This completes the proof of Proposition 1. ■

References

Aguilar, O., West, M., 2000. Bayesian dynamic factor models and portfolio allocation. Journal of Business & Economic Statistics 18, 338–357.

Ahmed, S., Bu, Z., Tsvetanov, D., 2019. Best of the best: A comparison of factor models. Journal of Financial and Quantitative Analysis 54, 1713–1758.

Ahmed, S., Bu, Z., Ye, X., 2023a. Illiquidity, r&d investment, and stock returns. Journal of Money, Credit and Banking (forthcoming).

Ahmed, S., Bu, Z., Ye, X., 2023b. Product market competition, labor mobility, and the cross-section of stock returns. Review of Asset Pricing Studies (forthcoming).

Ang, A., 2014. Asset Management: A Systematic Approach to Factor Investing. Oxford University Press, New York.

Ang, A., Liu, J., Schwarz, K., 2020. Using stocks or portfolios in tests of factor models. Journal of Financial and Quantitative Analysis 55, 709–750.

Asness, C., Frazzini, A., 2013. The devil in HML's details. Journal of Portfolio Management 39, 49–68.

Bali, T. G., Engle, R. F., Murray, S., 2016. Empirical Asset Pricing: The Cross Section of Stock Returns. John Wiley & Sons, Inc., New Jersey.

Barillas, F., Kan, R., Robotti, C., Shanken, J., 2020. Model comparison with Sharpe ratios. Journal of Financial and Quantitative Analysis 55, 1840–1874.

Barillas, F., Shanken, J., 2018. Comparing asset pricing models. Journal of Finance 73, 715–754.

Bollerslev, T., Chou, R. Y., Kroner, K. F., 1992. ARCH modeling in finance: A review of the theory and empirical evidence. Journal of Econometrics 52, 5–59.

Bollerslev, T., Engle, R. F., Nelson, D. B., 1994. ARCH models. In: Engle, R. F., McFadden, D. L. (Eds.), Handbook of Econometrics. North-Holland, Amsterdam, pp. 2959–3038.

Bollerslev, T., Hood, B., Huss, J., Pedersen, L. H., 2018. Risk everywhere: Modeling and managing volatility. Review of Financial Studies 31, 71–91.

Brandt, M. W., 2010. Portfolio choice problems. In: Aït-Sahalia, Y., Hansen, L. P. (Eds.), Handbook of Financial Econometrics: Tools and Techniques. North-Holland, San Diego, pp. 269–336.

Brandt, M. W., Santa-Clara, P., Valkanov, R., 2009. Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. Review of Financial Studies 22, 3411–3447.

Bueno, M. I., Furtado, S., Johnson, C. R., 2007. Congruence of Hermitian matrices by Hermitian matrices. Linear Algebra and its Applications 425, 63–76.

Carhart, M. M., 1997. On persistence in mutual fund performance. Journal of Finance 52, 57–82.

Chamberlain, G., 1983. Funds, factors, and diversification in arbitrage pricing models. Econometrica 51, 1305–1323.

Chan, L. K. C., Karceski, J., Lakonishok, J., 1998. The risk and return from factors. Journal of Financial and Quantitative Analysis 33, 159–188.

Chan, L. K. C., Karceski, J., Lakonishok, J., 1999. On portfolio optimization: Forecasting covariances and choosing the risk model. Review of Financial Studies 12, 937–974.

Charoenrook, A., Conrad, J., 2008. Identifying risk-based factors, Working Paper, Vanderbilt University, Tennessee.

Chincarini, L. B., Kim, D., 2006. Quantitative Equity Portfolio Management: An Active Approach to Portfolio Construction and Management. McGraw-Hill Education, New York.

Connor, G., Korajczyk, R. A., 1993. A test for the number of factors in an approximate factor model. Journal of Finance 48, 1263–1291.

Cremers, K. J. M., Petajisto, A., 2009. How active is your fund manager? A new measure that predicts performance. Review of Financial Studies 22, 3329–3365.

Daniel, K., Hirshleifer, D., Sun, L., 2020. Short and long horizon behavioral factors. Review of Financial Studies 33, 1673–1736.

Daniel, K., Moskowitz, T. J., 2016. Momentum crashes. Journal of Financial Economics 122, 221–247.

DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? Review of Financial Studies 22, 1915–1953.

DeMiguel, V., Nogales, F. J., Uppal, R., 2014. Stock return serial dependence and out-of-sample portfolio performance. Review of Financial Studies 27, 1031–1073.

Efron, B., 1982. Transformation theory: How normal is a family of distributions? Annals of Statistics 10, 323–339.

Engle, R. F., 2002. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. Journal of Business & Economic Statistics 20, 339–350.

Engle, R. F., Colacito, R., 2006. Testing and valuing dynamic correlations for asset allocation. Journal of Business & Economic Statistics 24, 238–253.

Engle, R. F., Ledoit, O., Wolf, M., 2019. Large dynamic covariance matrices. Journal of Business & Economic Statistics 37, 363–375.

Engle, R. F., Ng, V. K., Rothschild, M., 1990. Asset pricing with a factor-ARCH covariance structure: Empirical estimates for treasury bills. Journal of Econometrics 45, 213–237.

Engle, R. F., Sheppard, K., 2001. Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH, NBER Working Paper 8554.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–56.

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Fama, E. F., French, K. R., 2016. Dissecting anomalies with a five-factor model. Review of Financial Studies 29, 69–103.

Fama, E. F., French, K. R., 2018. Choosing factors. Journal of Financial Economics 128, 234–252.

Feng, G., Giglio, S., Xiu, D., 2020. Taming the factor zoo: A test of new factors. Journal of Finance 75, 1327–1370.

Ferreira, M. A., Lopez, J. A., 2005. Evaluating interest rate covariance models within a Value-at-Risk framework. Journal of Financial Econometrics 3, 126–168.

Foster, F. D., Smith, T., Whaley, R. E., 1997. Assessing goodness-of-fit of asset pricing models: The distribution of the maximal R^2. Journal of Finance 52, 591–607.

Giglio, S., Xiu, D., 2021. Asset pricing with omitted factors. Journal of Political Economy 129, 1947–1990.

Gospodinov, N., Kan, R., Robotti, C., 2013. Chi-squared tests for evaluation and comparison of asset pricing models. Journal of Econometrics 173, 108–125.

Hansen, P. R., Lunde, A., 2005. A forecast comparison of volatility models: Does anything beat a GARCH (1,1)? Journal of Applied Econometrics 20 (7), 873–889.

Hansen, P. R., Lunde, A., Nason, J. M., 2003. Choosing the best volatility models: The model confidence set approach. Oxford Bulletin of Economics and Statistics 65, 839–861.

Hansen, P. R., Lunde, A., Nason, J. M., 2011. The model confidence set. Econometrica 79, 453–497.

Harvey, C. R., Liu, Y., Zhu, H., 2016. . . . and the cross-section of expected returns. Review of Financial Studies 29, 5–68.

Hirshleifer, D., Hsu, P.-H., Li, D., 2018. Innovative originality, profitability, and stock returns. Review of Financial Studies 31, 2553–2605.

Hou, K., Mo, H., Xue, C., Zhang, L., 2019. Which factors? Review of Finance 23, 1–35.

Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: An investment approach. Review of Financial Studies 28, 650–705.

Jagannathan, R., Ma, T., 2003. Risk reduction in large portfolios: Why imposing the wrong constraints helps. Journal of Finance 58, 1651–1683.

Jorion, P., 2003. Portfolio optimization with tracking-error constraints. Financial Analysts Journal 59, 70–82.

Kan, R., Robotti, C., Shanken, J., 2013. Pricing model performance and the two-pass cross-sectional regression methodology. Journal of Finance 68, 2617–2649.

Kan, R., Zhang, C., 1999. Two-pass tests of asset pricing models with useless factors. Journal of Finance 54, 203–235.

Laurent, S., Rombouts, J. V. K., Violante, F., 2012. On the forecasting accuracy of multivariate GARCH models. Journal of Applied Econometrics 27, 934–955.

Ledoit, O., Wolf, M., 2008. Robust performance hypothesis testing with the Sharpe ratio. Journal of Empirical Finance 15, 850–859.

Ledoit, O., Wolf, M., 2011. Robust performance hypothesis testing with the variance. Wilmott 2011, 86–89.

Ledoit, O., Wolf, M., 2017. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. Review of Financial Studies 30, 4349–4388.

Ledoit, O., Wolf, M., Zhao, Z., 2019. Efficient sorting: A more powerful test for cross-sectional anomalies. Journal of Financial Econometrics 17, 645–686.

Lewellen, J., Nagel, S., Shanken, J., 2010. A skeptical appraisal of asset pricing tests. Journal of Financial Economics 96, 175–194.

Linnainmaa, J. T., Roberts, M. R., 2018. The history of the cross-section of stock returns. Review of Financial Studies 31, 2606–2649.

Lintner, J., 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. Review of Economics and Statistics 47, 13–37.

Liu, L. Y., Patton, A. J., Sheppard, K., 2015. Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. Journal of Econometrics 187, 293–311.

Lo, A. W., MacKinlay, A. C., 1990. Data-snooping biases in tests of financial asset pricing models. Review of Financial Studies 3, 431–467.

MacKinlay, A. C., 1995. Multifactor models do not explain deviations from the CAPM. Journal of Financial Economics 38, 3–28.

Meucci, A., 2005. Risk and Asset Allocation. Springer Finance, Berlin.

Moskowitz, T. J., 2003. An analysis of covariance risk and pricing anomalies. Review of Financial Studies 16, 417–457.

Nardari, F., Scruggs, J. T., 2007. Bayesian analysis of linear factor models with latent factors, multivariate stochastic volatility, and APT pricing restrictions. Journal of Financial and Quantitative Analysis 42, 857–891.

Nielsen, F., Aylursubramanian, R., 2008. Far from the madding crowd – volatility efficient indices, Research Insights, MSCI Barra.

Olivares-Nadal, A. V., DeMiguel, V., 2018. Technical note – a robust perspective on transaction costs in portfolio optimization. Operations Research 66, 733–739.

Pástor, Ľ., Stambaugh, R. F., 2000. Comparing asset pricing models: An investment perspective. Journal of Financial Economics 56, 335–381.

Pástor, v., Stambaugh, R. F., 2003. Liquidity risk and expected stock returns. Journal of Political Economy 111, 642–685.

Pukthuanthong, K., Roll, R., Subrahmanyam, A., 2019. A protocol for factor identification. Review of Financial Studies 32, 1573–1607.

Sharpe, W. F., 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. Journal of Finance 19, 425–442.

Simin, T., 2008. The poor predictive performance of asset pricing models. Journal of Financial and Quantitative Analysis 43, 355–380.

Stambaugh, R. F., Yuan, Y., 2017. Mispricing factors. Review of Financial Studies 30, 1270–1315.

Table 1
Summary of tests for equality of variances

The table reports the summary of the pairwise tests for equality of the out-of-sample variances of the daily returns realized on the optimized portfolios implied by 11 different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) three-factor (FF3) model; the Fama and French (1993) and Carhart (1997) four-factor (FFC) model; the Asness and Frazzini (2013) three-factor (FFAF) model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) five-factor (FF5) model; the four-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2017) four-factor (SY4) model; the Barillas and Shanken (2018) six-factor (BS6) model; the Fama and French (2018) six-factor (FF6) model; and the Daniel et al. (2020) three-factor (DHS) model. The test assets for investment include: samples of 50 small, and 100, 250 and 500 largest stocks drawn from eligible NYSE-, AMEX-, and NASDAQ-listed nonfinancial and nonregulated ordinary common stocks; a set of value-weighted 48 IND (industry) portfolios; and a set of value-weighted 340 portfolios used in Giglio and Xiu (2021). Panels A, B, and C summarize the results for the minimum variance, the restricted minimum variance, and the long-only minimum variance portfolios, respectively. For a given set of test assets, each cell contains the number of times a given factor model produces a significantly (at the 5% level) lower out-of-sample variance, for the daily returns realized on the portfolio, relative to a competing model. The out-of-sample evaluation period starts on December 30, 1976, and ends on December 31, 2018.

Model	Small stocks	100 stocks	250 stocks	500 stocks	48 IND portfolios	Giglio-Xiu portfolios
Panel A: Minimum variance portfolios						
CAPM	0	0	0	0	0	0
FF3	2	1	3	3	3	3
FFC	3	4	4	4	3	4
FFAF	2	1	1	1	1	1
HXZ	2	6	6	6	3	3
FF5	7	7	8	8	8	8
FF4	2	4	4	3	4	3
SY4	2	3	6	6	6	3
BS6	6	7	8	9	6	10
FF6	7	7	8	8	8	8
DHS	0	1	1	1	1	1
Panel B: Restricted minimum variance portfolios						
CAPM	0	0	0	0	0	0
FF3	2	1	3	3	3	3
FFC	3	4	4	4	3	4
FFAF	2	1	1	1	1	1
HXZ	2	6	6	6	3	3
FF5	7	7	8	8	8	8
FF4	2	4	4	3	4	3
SY4	2	3	6	6	6	3
BS6	6	7	8	9	6	10
FF6	7	7	8	8	8	8
DHS	0	1	1	1	1	1
Panel C: Long-only minimum variance portfolios						
CAPM	0	0	0	0	0	0
FF3	2	2	0	2	0	1
FFC	2	0	0	2	0	0
FFAF	2	2	2	2	0	1
HXZ	1	3	5	6	1	1
FF5	5	4	4	6	0	2
FF4	2	2	5	4	0	0
SY4	0	0	2	2	0	0
BS6	3	1	5	6	6	0
FF6	3	1	6	5	2	0
DHS	0	0	0	0	0	0

Table 2
Model confidence set results

The table reports the model confidence set (MCS) p-values for the sequential equality tests of the out-of-sample variances of the daily returns realized on the optimized portfolios implied by 11 different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) three-factor (FF3) model; the Fama and French (1993) and Carhart (1997) four-factor (FFC) model; the Asness and Frazzini (2013) three-factor (FFAF) model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) five-factor (FF5) model; the four-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2017) four-factor (SY4) model; the Barillas and Shanken (2018) six-factor (BS6) model; the Fama and French (2018) six-factor (FF6) model; and the Daniel et al. (2020) three-factor (DHS) model. The test assets for investment include: samples of 50 small, and 100, 250 and 500 largest stocks drawn from eligible NYSE-, AMEX-, and NASDAQ-listed nonfinancial and nonregulated ordinary common stocks; a set of value-weighted 48 IND (industry) portfolios; and a set of value-weighted 340 portfolios used in Giglio and Xiu (2021). Panels A, B, and C report the MCS p-values for the minimum variance, the restricted minimum variance, and the long-only minimum variance portfolios, respectively. For a given set of test assets, * denotes the factor model that ends up in the MCS containing the best model(s) with a 95% level of confidence (i.e., $\widehat{\mathcal{M}}^*_{95\%}$). The MCS p-values are computed using the bootstrap procedure of Hansen et al. (2011), using 9,999 replications. The out-of-sample period starts on December 30, 1976, and ends on December 31, 2018.

Model	Small stocks	100 stocks	250 stocks	500 stocks	48 IND portfolios	Giglio-Xiu portfolios
Panel A: Minimum variance portfolios						
CAPM	0.000	0.000	0.000	0.000	0.000	0.000
FF3	0.082*	0.000	0.000	0.000	0.000	0.000
FFC	0.082*	0.000	0.000	0.000	0.000	0.020
FFAF	0.082*	0.000	0.000	0.000	0.000	0.000
HXZ	0.004	0.053*	0.001	0.003	0.000	0.020
FF5	0.765*	0.288*	0.861*	0.108*	0.152*	0.120*
FF4	0.056*	0.000	0.000	0.000	0.013	0.001
SY4	0.082*	0.002	0.001	0.003	0.013	0.004
BS6	0.765*	1.000*	0.861*	1.000*	0.134*	1.000*
FF6	1.000*	0.288*	1.000*	0.122*	1.000*	0.120*
DHS	0.001	0.000	0.000	0.000	0.000	0.000
Panel B: Restricted minimum variance portfolios						
CAPM	0.000	0.000	0.000	0.000	0.000	0.000
FF3	0.082*	0.000	0.000	0.000	0.000	0.000
FFC	0.082*	0.000	0.000	0.000	0.000	0.020
FFAF	0.082*	0.000	0.000	0.000	0.000	0.000
HXZ	0.004	0.053*	0.001	0.003	0.000	0.020
FF5	0.765*	0.288*	0.861*	0.108*	0.157*	0.120*
FF4	0.056*	0.000	0.000	0.000	0.013	0.001
SY4	0.082*	0.002	0.001	0.003	0.013	0.004
BS6	0.765*	1.000*	0.861*	1.000*	0.132*	1.000*
FF6	1.000*	0.288*	1.000*	0.122*	1.000*	0.120*
DHS	0.001	0.000	0.000	0.000	0.000	0.000
Panel C: Long-only minimum variance portfolios						
CAPM	0.000	0.000	0.001	0.010	0.473*	0.720*
FF3	0.907*	0.636*	0.004	0.007	0.536*	0.965*
FFC	0.900*	0.038	0.006	0.010	0.536*	0.029
FFAF	0.372*	0.636*	0.010	0.073*	0.536*	0.965*
HXZ	0.372*	0.698*	0.368*	0.764*	0.572*	0.965*
FF5	1.000*	1.000*	0.368*	0.103*	0.536*	1.000*
FF4	0.851*	0.698*	0.457*	0.764*	0.536*	0.965*
SY4	0.014	0.182*	0.210*	0.031	0.536*	0.768*
BS6	0.907*	0.698*	0.534*	1.000*	1.000*	0.720*
FF6	0.907*	0.636*	1.000*	0.764*	0.536*	0.720*
DHS	0.001	0.038	0.004	0.009	0.488*	0.720*

Table 3
Summary of tests for equality of variances: Tracking error portfolios

The table reports the summary of pairwise tests for equality of the out-of-sample variances of the daily excess returns (over the benchmark's return) realized on the tracking error portfolios implied by 11 different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) three-factor (FF3) model; the Fama and French (1993) and Carhart (1997) four-factor (FFC) model; the Asness and Frazzini (2013) three-factor (FFAF) model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) five-factor (FF5) model; the four-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2017) four-factor (SY4) model; the Barillas and Shanken (2018) six-factor (BS6) model; the Fama and French (2018) six-factor (FF6) model; and the Daniel et al. (2020) three-factor (DHS) model. The test assets for investment include: samples of 50 small, and 100, 250 and 500 largest stocks drawn from eligible NYSE-, AMEX-, and NASDAQ-listed nonfinancial and nonregulated ordinary common stocks; a set of value-weighted 48 IND (industry) portfolios; and a set of value-weighted 340 portfolios used in Giglio and Xiu (2021). The benchmark portfolio for the tracking error is the Standard & Poor's 500 index. Panels A, B, and C summarize the results for the minimum tracking error, the restricted minimum tracking error, and the long-only minimum tracking error portfolios, respectively. For a given set of test assets, each cell contains the number of times a given factor model produces a significantly (at the 5% level) lower out-of-sample variance, for the daily excess returns (over the benchmark's return) realized on the tracking error portfolio, relative to a competing model. The out-of-sample period starts on December 30, 1976, and ends on December 31, 2018.

Model	Small stocks	100 stocks	250 stocks	500 stocks	48 IND portfolios	Giglio-Xiu portfolios
Panel A: Minimum tracking error portfolios						
CAPM	0	6	0	0	0	0
FF3	2	0	1	2	3	1
FFC	3	2	3	5	7	2
FFAF	2	2	3	4	3	4
HXZ	2	4	4	4	2	8
FF5	3	2	4	4	7	2
FF4	2	5	6	4	2	4
SY4	2	0	0	2	3	1
BS6	2	4	8	9	9	8
FF6	4	6	8	9	9	5
DHS	0	4	0	1	1	1
Panel B: Restricted minimum tracking error portfolios						
CAPM	0	6	0	0	0	0
FF3	2	0	1	2	3	1
FFC	3	2	3	5	7	2
FFAF	2	2	3	4	3	4
HXZ	2	4	4	4	2	8
FF5	3	2	4	4	7	2
FF4	2	5	6	4	2	4
SY4	2	0	0	2	3	1
BS6	2	4	8	9	9	8
FF6	4	6	8	9	9	5
DHS	0	4	0	1	1	1
Panel C: Long-only minimum tracking error portfolios						
CAPM	0	6	0	0	0	0
FF3	2	0	1	3	3	1
FFC	3	2	4	4	7	2
FFAF	2	2	3	4	2	2
HXZ	2	2	4	4	2	4
FF5	3	2	4	4	7	2
FF4	2	5	7	5	2	4
SY4	2	0	0	2	3	1
BS6	3	4	8	9	9	7
FF6	7	6	8	9	9	5
DHS	0	4	0	1	1	1

Table 4
Model confidence set results: Tracking error portfolios

The table reports the model confidence set (MCS) p-values for the sequential equality tests of the out-of-sample variances of the daily excess returns (over the benchmark's return) realized on the tracking error minimizing portfolios implied by 11 different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) three-factor (FF3) model; the Fama and French (1993) and Carhart (1997) four-factor (FFC) model; the Asness and Frazzini (2013) three-factor (FFAF) model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) five-factor (FF5) model; the four-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2017) four-factor (SY4) model; the Barillas and Shanken (2018) six-factor (BS6) model; the Fama and French (2018) six-factor (FF6) model; and the Daniel et al. (2020) three-factor (DHS) model. The test assets for investment include: samples of 50 small, and 100, 250 and 500 largest stocks drawn from eligible NYSE-, AMEX-, and NASDAQ-listed nonfinancial and nonregulated ordinary common stocks; a set of value-weighted 48 IND (industry) portfolios; and a set of value-weighted 340 portfolios used in Giglio and Xiu (2021). The benchmark portfolio for the tracking error is the Standard & Poor's 500 index. Panels A, B, and C report the MCS p-values for the minimum tracking error, the restricted minimum tracking error, and the long-only minimum tracking error portfolios, respectively. For a given set of test assets, * denotes the factor model that ends up in the MCS containing the best model(s) with a 95% level of confidence (i.e., $\widehat{\mathcal{M}}^*_{95\%}$). The MCS p-values are computed using the bootstrap procedure of Hansen et al. (2011), using 9,999 replications. The out-of-sample evaluation period starts on December 30, 1976, and ends on December 31, 2018.

Model	Small stocks	100 stocks	250 stocks	500 stocks	48 IND portfolios	Giglio-Xiu portfolios
Panel A: Minimum tracking error portfolios						
CAPM	0.010	1.000*	0.002	0.000	0.000	0.000
FF3	0.695*	0.000	0.000	0.000	0.000	0.000
FFC	1.000*	0.001	0.010	0.002	0.001	0.000
FFAF	0.737*	0.000	0.002	0.001	0.000	0.220*
HXZ	0.032	0.217*	0.010	0.002	0.000	1.000*
FF5	0.737*	0.008	0.005	0.000	0.001	0.005
FF4	0.385*	0.599*	0.214*	0.002	0.000	0.220*
SY4	0.904*	0.000	0.000	0.000	0.000	0.000
BS6	0.695*	0.217*	0.214*	1.000*	0.702*	0.933*
FF6	0.999*	0.599*	1.000*	0.270*	1.000*	0.206*
DHS	0.012	0.415*	0.000	0.000	0.000	0.220*
Panel B: Restricted minimum tracking error portfolios						
CAPM	0.010	1.000*	0.002	0.000	0.000	0.000
FF3	0.695*	0.000	0.000	0.000	0.000	0.000
FFC	1.000*	0.001	0.010	0.002	0.001	0.000
FFAF	0.737*	0.000	0.002	0.001	0.000	0.220*
HXZ	0.032	0.217*	0.010	0.002	0.000	1.000*
FF5	0.737*	0.008	0.005	0.000	0.001	0.005
FF4	0.385*	0.599*	0.214*	0.002	0.000	0.220*
SY4	0.904*	0.000	0.000	0.000	0.000	0.000
BS6	0.695*	0.217*	0.214*	1.000*	0.702*	0.933*
FF6	0.999*	0.599*	1.000*	0.270*	1.000*	0.206*
DHS	0.012	0.415*	0.000	0.000	0.000	0.220*
Panel C: Long-only minimum tracking error portfolios						
CAPM	0.000	1.000*	0.000	0.000	0.000	0.000
FF3	0.255*	0.000	0.000	0.000	0.000	0.000
FFC	0.719*	0.003	0.023	0.004	0.001	0.039
FFAF	0.458*	0.000	0.001	0.000	0.000	0.146*
HXZ	0.255*	0.113*	0.023	0.000	0.000	0.746*
FF5	0.719*	0.049	0.006	0.000	0.001	0.146*
FF4	0.255*	0.739*	0.162*	0.004	0.000	0.292*
SY4	0.719*	0.000	0.000	0.000	0.000	0.024
BS6	0.719*	0.152*	0.162*	1.000*	0.192*	1.000*
FF6	1.000*	0.768*	1.000*	0.390*	1.000*	0.292*
DHS	0.000	0.170*	0.000	0.000	0.000	0.146*

Table 5
Model confidence set results (adjusted for transaction costs)

The table reports the model confidence set (MCS) p-values for the sequential equality tests of the out-of-sample variances of the daily returns realized on the optimized portfolios (net of trading costs) implied by 11 different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) three-factor (FF3) model; the Fama and French (1993) and Carhart (1997) four-factor (FFC) model; the Asness and Frazzini (2013) three-factor (FFAF) model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) five-factor (FF5) model; the four-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2017) four-factor (SY4) model; the Barillas and Shanken (2018) six-factor (BS6) model; the Fama and French (2018) six-factor (FF6) model; and the Daniel et al. (2020) three-factor (DHS) model. The test assets for investment include: samples of 50 small, and 100, 250 and 500 largest stocks drawn from eligible NYSE-, AMEX-, and NASDAQ-listed nonfinancial and nonregulated ordinary common stocks; a set of value-weighted 48 IND (industry) portfolios; and a set of value-weighted 340 portfolios used in Giglio and Xiu (2021). As in Brandt et al. (2009), the proportional transaction costs are 0.5%, and are constant across assets and over time.The details of adjusting trading costs are available in Section 5.1. Panels A, B, and C report the MCS p-values for the minimum variance, the restricted minimum variance, and the long-only minimum variance portfolios, respectively. For a given set of test assets, * denotes the factor model that ends up in the MCS containing the best model(s) with a 95% level of confidence (i.e., $\widehat{\mathcal{M}}^*_{95\%}$). The MCS p-values are computed using the bootstrap procedure of Hansen et al. (2011), using 9,999 replications. The out-of-sample period starts on December 30, 1976, and ends on December 31, 2018.

Model	Small stocks	100 stocks	250 stocks	500 stocks	48 IND portfolios	Giglio-Xiu portfolios
Panel A: Minimum variance portfolios						
CAPM	0.004	0.000	0.000	0.000	0.000	0.000
FF3	0.149*	0.000	0.000	0.000	0.000	0.015
FFC	0.149*	0.000	0.000	0.000	0.000	0.246*
FFAF	0.100*	0.000	0.000	0.000	0.000	0.000
HXZ	0.019	0.177*	0.002	0.013	0.000	0.243*
FF5	0.556*	0.411*	0.926*	0.113*	0.253*	0.246*
FF4	0.100*	0.000	0.000	0.000	0.030	0.015
SY4	0.149*	0.001	0.002	0.013	0.030	0.015
BS6	0.556*	1.000*	0.926*	1.000*	0.200*	1.000*
FF6	1.000*	0.411*	1.000*	0.115*	1.000*	0.246*
DHS	0.100*	0.000	0.000	0.000	0.000	0.000
Panel B: Restricted minimum variance portfolios						
CAPM	0.004	0.000	0.000	0.000	0.000	0.000
FF3	0.149*	0.000	0.000	0.000	0.001	0.015
FFC	0.149*	0.000	0.000	0.000	0.001	0.246*
FFAF	0.100*	0.000	0.000	0.000	0.000	0.000
HXZ	0.019	0.177*	0.002	0.013	0.001	0.243*
FF5	0.556*	0.411*	0.926*	0.113*	0.256*	0.246*
FF4	0.100*	0.000	0.000	0.000	0.029	0.015
SY4	0.149*	0.001	0.002	0.013	0.029	0.015
BS6	0.556*	1.000*	0.926*	1.000*	0.201*	1.000*
FF6	1.000*	0.411*	1.000*	0.115*	1.000*	0.246*
DHS	0.100*	0.000	0.000	0.000	0.000	0.000
Panel C: Long-only minimum variance portfolios						
CAPM	0.000	0.000	0.003	0.013	0.493*	0.717*
FF3	0.877*	0.645*	0.010	0.009	0.561*	0.957*
FFC	0.856*	0.029	0.010	0.013	0.561*	0.017
FFAF	0.382*	0.645*	0.021	0.112*	0.561*	0.957*
HXZ	0.382*	0.712*	0.472*	0.898*	0.561*	0.957*
FF5	1.000*	1.000*	0.472*	0.175*	0.561*	1.000*
FF4	0.811*	0.712*	0.616*	0.898*	0.561*	0.957*
SY4	0.027	0.185*	0.246*	0.032	0.561*	0.717*
BS6	0.877*	0.712*	0.616*	1.000*	1.000*	0.705*
FF6	0.877*	0.645*	1.000*	0.898*	0.561*	0.717*
DHS	0.003	0.029	0.010	0.013	0.561*	0.705*

Table 6
Model confidence set results:
Tracking error portfolios (adjusted for transaction costs)

The table reports the model confidence set (MCS) p-values for the sequential equality tests of the out-of-sample variances of the daily excess returns (over the benchmark's return) realized on the tracking error minimizing portfolios (net of trading costs) implied by 11 different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) three-factor (FF3) model; the Fama and French (1993) and Carhart (1997) four-factor (FFC) model; the Asness and Frazzini (2013) three-factor (FFAF) model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) five-factor (FF5) model; the four-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2017) four-factor (SY4) model; the Barillas and Shanken (2018) six-factor (BS6) model; the Fama and French (2018) six-factor (FF6) model; and the Daniel et al. (2020) three-factor (DHS) model. The test assets for investment include: samples of 50 small, and 100, 250 and 500 largest stocks drawn from eligible NYSE-, AMEX-, and NASDAQ-listed nonfinancial and nonregulated ordinary common stocks; a set of value-weighted 48 IND (industry) portfolios; and a set of value-weighted 340 portfolios used in Giglio and Xiu (2021). The benchmark portfolio for the tracking error is the Standard & Poor's 500 index. As in Brandt et al. (2009), the proportional transaction costs are 0.5%, and are constant across assets and over time.The details of adjusting trading costs are available in Section 5.1. Panels A, B, and C report the MCS p-values for the minimum tracking error, the restricted minimum tracking error, and the long-only minimum tracking error portfolios, respectively. For a given set of test assets, * denotes the factor model that ends up in the MCS containing the best model(s) with a 95% level of confidence (i.e., $\widehat{\mathcal{M}}^*_{95\%}$). The MCS p-values are computed using the bootstrap procedure of Hansen et al. (2011), using 9,999 replications. The out-of-sample period starts on December 30, 1976, and ends on December 31, 2018.

Model	Small stocks	100 stocks	250 stocks	500 stocks	48 IND portfolios	Giglio-Xiu portfolios
Panel A: Minimum tracking error portfolios						
CAPM	0.024	1.000*	0.009	0.000	0.000	0.000
FF3	0.619*	0.000	0.000	0.000	0.000	0.000
FFC	0.972*	0.000	0.011	0.009	0.007	0.000
FFAF	0.619*	0.000	0.009	0.002	0.000	0.172*
HXZ	0.124*	0.210*	0.011	0.005	0.000	1.000*
FF5	0.619*	0.021	0.011	0.000	0.007	0.001
FF4	0.325*	0.598*	0.363*	0.009	0.000	0.260*
SY4	0.972*	0.000	0.000	0.000	0.000	0.000
BS6	0.619*	0.210*	0.363*	1.000*	0.977*	0.311*
FF6	1.000*	0.392*	1.000*	0.278*	1.000*	0.123*
DHS	0.013	0.248*	0.000	0.000	0.000	0.260*
Panel B: Restricted minimum tracking error portfolios						
CAPM	0.024	1.000*	0.009	0.000	0.000	0.000
FF3	0.619*	0.000	0.000	0.000	0.000	0.000
FFC	0.972*	0.000	0.011	0.009	0.007	0.000
FFAF	0.619*	0.000	0.009	0.002	0.000	0.172*
HXZ	0.124*	0.210*	0.011	0.005	0.000	1.000*
FF5	0.619*	0.021	0.011	0.000	0.007	0.001
FF4	0.325*	0.598*	0.363*	0.009	0.000	0.260*
SY4	0.972*	0.000	0.000	0.000	0.000	0.000
BS6	0.619*	0.210*	0.363*	1.000*	0.977*	0.311*
FF6	1.000*	0.392*	1.000*	0.278*	1.000*	0.123*
DHS	0.013	0.248*	0.000	0.000	0.000	0.260*
Panel C: Long-only minimum tracking error portfolios						
CAPM	0.001	1.000*	0.002	0.000	0.000	0.000
FF3	0.123*	0.000	0.000	0.000	0.000	0.000
FFC	0.682*	0.001	0.041	0.004	0.010	0.027
FFAF	0.468*	0.000	0.002	0.000	0.000	0.208*
HXZ	0.224*	0.129*	0.041	0.000	0.000	1.000*
FF5	0.682*	0.129*	0.041	0.000	0.010	0.208*
FF4	0.232*	0.905*	0.271*	0.004	0.000	0.296*
SY4	0.682*	0.000	0.000	0.000	0.000	0.013
BS6	0.682*	0.155*	0.271*	1.000*	0.382*	0.786*
FF6	1.000*	0.905*	1.000*	0.371*	1.000*	0.208*
DHS	0.002	0.155*	0.000	0.000	0.000	0.208*