THE JOURNAL OF FINANCE • VOL. LXXIII, NO. 2 • APRIL 2018

Comparing Asset Pricing Models

FRANCISCO BARILLAS and JAY SHANKEN*

ABSTRACT

A Bayesian asset pricing test is derived that is easily computed in closed form from the standard F-statistic. Given a set of candidate traded factors, we develop a related test procedure that permits the computation of model probabilities for the collection of all possible pricing models that are based on subsets of the given factors. We find that the recent models of Hou, Xue, and Zhang (2015a, 2015b) and Fama and French (2015, 2016) are dominated by a variety of models that include a momentum factor, along with value and profitability factors that are updated monthly.

GIVEN THE VARIETY OF PORTFOLIO-BASED factors that have been examined by researchers, it is important to understand how best to combine them in a parsimonious asset pricing model for expected returns, one that excludes redundant factors. Although there are standard econometric techniques for evaluating the adequacy of a single model, a satisfactory statistical methodology for identifying the best factor pricing model(s) is conspicuously lacking in investment research applications. In this paper, we develop an easy-to-implement Bayesian procedure that allows us to compute model probabilities for the collection of all possible pricing models that can be formed from a given set of factors.

Beginning with the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965), the asset pricing literature has attempted to understand the determination of risk premia on financial securities. The central theme of this literature is that a security's risk premium should depend on the security's market beta or other measure(s) of systematic risk. In a classic test of the CAPM that builds on Jensen (1968), Black, Jensen, and Scholes (1972) examine the intercepts in time-series regressions of excess test portfolio returns on market excess returns. Given the CAPM implication that the market portfolio is efficient, these intercepts or "alphas" should be zero. A joint F-test of this hypothesis is later developed by Gibbons, Ross, and Shanken (1989), henceforth

*Barillas and Shanken are from the Goizueta Business School at Emory University. Shanken is also from the National Bureau of Economic Research. Thanks to two anonymous referees, Rohit Allena, Doron Avramov, Mark Fisher, Amit Goyal, Lubos Pastor, Seth Pruitt, Tim Simin, Ken Singleton (the Editor), Rex Thompson, Pietro Veronesi, and seminar participants at the Northern Finance Association meeting, Southern Methodist University, the Universities of Arizona, Geneva, Hong Kong, Lausanne, Luxembourg, Michigan, Pennsylvania, and Singapore, Hong Kong University of Science and Technology, Imperial College, National University of Singapore, Ohio State University, and Florida International University. We have read the *Journal of Finance*'s disclosure policy and have no conflicts of interest to disclose.

DOI: 10.1111/jofi.12607

GRS, who also explore the relation between the test statistic and standard portfolio geometry.[1]

In recent years, a variety of multifactor asset pricing models have been explored. Although tests of the individual models are routinely reported, these tests often suggest "rejection" of the implied restrictions, especially when the data sets are large (e.g., Fama and French (2016)). However, a relatively large p-value may say more about imprecision in estimating a particular model's alphas than the adequacy of that model.[2] Simple statistical tools with which to analyze the various models *jointly* in a model-comparison framework are thus sorely needed. The information that our methodology provides about relative model likelihoods complements that obtained from classical asset pricing tests and is more in the spirit of the adage "it takes a model to beat a model." [3]

Like other asset pricing analyses based on alphas, we require that the benchmark factors are traded portfolio excess returns or return spreads. For example, in addition to the market excess return, *Mkt*, the influential three-factor model of Fama and French (1993), hereafter FF3, includes a book-to-market or "value" factor, *HML* (high-low), and a size factor, *SMB* (small-big), based on stock market capitalization. Although consumption growth and intertemporal hedge factors are not traded, one can always substitute (maximally correlated) mimicking portfolios for the nontraded factors.[4] Although this introduces additional estimation issues, simple spread-portfolio factors are often viewed as proxies for the relevant mimicking portfolios (e.g., Fama and French (1996)).

We begin by analyzing the joint alpha restriction for a set of test assets in a Bayesian setting.[5] Prior beliefs about the extent of model mispricing are economically motivated and accommodate traditional risk-based views as well as more behavioral perspectives. The posterior probability that the zero-alpha restriction holds is then shown to be an easy-to-calculate function of the GRS F-statistic. Our related model-comparison methodology is likewise computationally straightforward. This procedure builds on results in Barillas and Shanken (2017), who highlight the fact that, for several widely accepted criteria, model comparison with traded factors only requires examination of each model's ability to price the factors in the other models.

[1] See related work by Treynor and Black (1973) and Jobson and Korkie (1982).

[2] De Moor, Dhaene, and Sercu (2015) suggest a calculation that highlights the extent to which differences in p-values may be influenced by differences in estimation precision across models, but they do not provide a formal hypothesis test.

[3] Avramov and Chao (2006) also explore Bayesian model comparison for asset pricing models. As we explain in the next two sections, their methodology is quite different from that developed here. A recent paper by Kan, Robotti, and Shanken (2013) provides asymptotic results for comparing model R^2s in a cross-sectional regression framework. Chen, Roll, and Ross (1986) nest the CAPM in a multifactor model with betas on macro-related factors included in cross-sectional regressions. In other Bayesian applications, Malatesta and Thompson (1993) apply methods for comparing multiple hypotheses in a corporate finance event-study context.

[4] See Merton (1973) and Breeden (1979), especially footnote 8.

[5] See earlier work by Shanken (1987b), Harvey and Zhou (1990), and McCulloch and Rossi (1991).

The observation that all models are necessarily simplifications of reality and hence must be false in a literal sense motivates an evaluation of whether a model holds approximately, rather than as a sharp null hypothesis. Additional motivation comes from recognizing that the factors used in asset pricing tests are generally *proxies* for the relevant theoretical factors.[6] With these considerations in mind, we extend our results to obtain simple formulas for testing the more plausible approximate models. As a warm up, we consider all models that can be obtained using subsets of the FF3 factors *Mkt, HML,* and *SMB*. A nice aspect of the Bayesian approach is that it permits comparison of nested models like CAPM and FF3, as well as the nonnested models {*Mkt HML*} and {*Mkt SMB*}. Moreover, we are able to *simultaneously* compare all of the models, as opposed to standard classical approaches that involve pairwise model comparison (e.g., Vuong (1989)). Over the period 1972 to 2015, the alphas for *HML* when regressed on *Mkt* or *Mkt* and *SMB* are highly "significant," whereas the alphas for *SMB* when regressed on *Mkt* or *Mkt* and *HML* are modest. Our procedure aggregates all of this evidence, arriving at posterior probabilities with our benchmark prior of 60% for the two-factor model {*Mkt HML*} and 39% for FF3, with the remaining 1% split between CAPM and {*Mkt SMB*}.

In our main empirical application, we compare models that combine many prominent factors from the literature. In addition to the FF3 factors, we consider the momentum factor, *UMD* (up minus down), introduced by Carhart (1997) and motivated by Jegadeesh and Titman (1993). We also include factors from the recently proposed five-factor model of Fama and French (2015), hereafter FF5.[7] These are *RMW* (robust minus weak), which is based on the profitability of firms, and *CMA* (conservative minus aggressive), which is related to firms' net investments. Hou, Xue, and Zhang (2015a, 2015b), henceforth HXZ, propose their own versions of the size (*ME*), investment (*IA*), and profitability (*ROE*) factors, which we also examine. In particular, *ROE* incorporates the most recent earnings information from quarterly data. Finally, we consider the value factor HML^m from Asness and Frazzini (2013), which is based on book-to-market rankings that use the most recent monthly stock price in the denominator. In total, we have 10 factors in our analysis.

Rather than mechanically apply our methodology with all nine of the nonmarket factors treated symmetrically, we structure the prior so as to recognize that several of the factors are different versions of the same underlying construct. Therefore, to avoid overfitting, we only consider models that contain at most one of the factors in each of the following categories: size (*SMB* or *ME*), profitability (*RMW* or *ROE*), value (*HML* or HML^m), and investment (*CMA* or *IA*). The extension of our procedure to accommodate such "categorical

[6] Kandel and Stambaugh (1987) and Shanken (1987a) analyze pricing restrictions based on proxies for the market portfolio or other equilibrium benchmark.

[7] We use the *SMB* factor from the FF5 model in this paper. Our findings are not sensitive to whether we use the original *SMB* factor, as the correlation between the two size measures is over 0.99.

factors" amounts to averaging results over different versions of the factors, with weights that reflect the likelihood that each version contains the relevant factors.

Using data from 1972 to 2015, we find that the model with the highest posterior probability is the six-factor model {*Mkt IA ROE SMB* HML^m *UMD*}. Thus, in contrast to previous findings by HXZ and FF5, value is no longer a redundant factor when the more timely HML^m is considered; and whereas HXZ also find that momentum is redundant, this is no longer true with the inclusion of HML^m. The timeliness of the HXZ profitability factor turns out to be important as well. The other top models are closely related to our best model, replacing *SMB* with *ME*, replacing *IA* with *CMA*, or excluding the size factors entirely. There is also overwhelming support for the six-factor model (or the five-factor model that excludes *SMB*) in direct tests of the model against the HXZ and FF5 models. These model-comparison results are qualitatively similar for priors motivated by a market efficiency perspective and others that allow for large departures from efficiency. The models identified as best in the Bayesian analysis perform well out of sample, but the HXZ and FF5 models do as well.

Model comparison assesses the *relative* performance of competing models. We also examine *absolute* performance for the top-ranked model. This test considers the extent to which the model does a good job pricing a set of test assets. Although we examine various test assets, we present results for 25 portfolios based on independent rankings by either size and momentum or book-to-market and investment. The evidence casts strong doubt on the validity of our six-factor model. The "rejection" is less overwhelming, however, when we consider an approximate version that allows for relatively small departures (average absolute value 0.8% per annum) from exact pricing.

The rest of the paper is organized as follows. Section I considers the classic case of testing a pricing model against a general alternative. Section II then compares nested pricing models and the relation between "relative" and "absolute" tests. We conduct Bayesian model comparison in Section III. Section IV extends the model-comparison framework to accommodate analysis with multiple versions of some factors. Section V provides results for various pricing models based on 10 prominent factors, while Section VI explores the out-of-sample performance of our high-probability models, along with several other models from the literature. Section VII examines the absolute performance of the model with the highest probability. Section VIII concludes. Several proofs of key results are provided in the Appendix, along with an extension that treats the market symmetrically with the other factors. An Internet Appendix with additional technical details is also available.[8]

[8] The Internet Appendix is available from the online version of the article on the *Journal of Finance* website.

I. Testing a Pricing Model Against a General Alternative

Traditional tests of factor pricing models compare a single restricted asset pricing model to an unrestricted alternative return-generating process that nests the null model. We explore the Bayesian counterpart of such a test in this section.

A. Statistical Assumptions and Portfolio Algebra

First, we lay out the factor model notation and assumptions. The factor model is a multivariate linear regression with N test-asset excess returns, r_t, and K factors for each of T months:

$$r_t = \alpha + \beta f_t + \varepsilon_t \; \varepsilon_t \sim N(0, \Sigma), \tag{1}$$

where r_t, ε_t, and α are $N \times 1$, β is $N \times K$, and f_t is $K \times 1$. The normal distribution of the ε_t is assumed to hold conditional on the factors and the ε_t are independent over time.

We assume that the factors are zero-investment returns such as the excess return on the market or the spread between two portfolios, like the Fama-French (1993) value-growth factor. Under the null hypothesis, $H_0 : \alpha = 0$, we have the usual simple linear relation between expected returns and betas:

$$E(r_t) = \beta E(f_t), \tag{2}$$

where $E(f_t)$ is a $K \times 1$ vector of factor premia.

The GRS test of this null hypothesis is based on the F-statistic with degrees of freedom N and $T{-}N{-}K$, which equals $(T{-}N{-}K)/(NT)$ times the Wald statistic:

$$W = T \frac{\hat{\alpha}' \hat{\Sigma}^{-1} \hat{\alpha}}{1 + Sh(F)^2}, \tag{3}$$

where

$$\hat{\alpha}' \hat{\Sigma}^{-1} \hat{\alpha} = (Sh(F, R)^2 - Sh(F)^2). \tag{4}$$

Here, $Sh(F)^2$ is the maximum squared sample Sharpe ratio over portfolios of the factors, where $\hat{\Sigma}$ is the maximum likelihood estimate (MLE) for the covariance matrix Σ. The term $Sh(F, R)^2$ is the squared Sharpe measure based on both factor and asset returns. The population counterpart of the identity in (4) implies that the tangency portfolio corresponding to the factor and asset returns equals that based on the factors alone under the null $\alpha = 0$. Thus, the expected return relation in (2) is equivalent to this equality of tangency portfolios and their associated squared Sharpe ratios.

B. A Bayesian Test of Efficiency

Bayesian tests of the zero-alpha restriction have been developed by Shanken (1987b), Harvey and Zhou (1990), and McCulloch and Rossi (1991). The test that we develop here takes as a starting point a prior specification considered in the Harvey and Zhou (1990) paper. Although they comment on the computational challenges of implementing this approach, we are able to derive a simple closed-form formula for the required Bayesian probabilities. The specification is appealing in that standard "diffuse" priors are used for the betas and residual covariance parameters. Thus, the data dominate beliefs about these parameters, freeing the researcher to focus on informative priors for the alphas, the parameters that are restricted by the models.[9] The details follow.

B.1. Prior Specification

The diffuse prior for β and Σ is

$$P(\beta, \Sigma) \propto |\Sigma|^{-(N+1)/2} \tag{5}$$

as in Jeffreys (1961). The prior for α is concentrated at zero under the null hypothesis. Under the alternative, we assume a multivariate normal informative prior for α conditional on β and Σ:

$$P(\alpha|\beta, \Sigma) = MVN(0, k\Sigma), \tag{6}$$

where the parameter $k > 0$ reflects our belief about the potential magnitude of deviations from the expected return relation.

Asset pricing theory provides some motivation for linking beliefs about the magnitude of alpha to residual variance. For example, Dybvig (1983) and Grinblatt and Titman (1983) derive bounds on an individual asset's deviation from a multifactor pricing model that are proportional to the asset's residual variance. From a behavioral perspective, Shleifer and Vishny (1997) argue that high idiosyncratic risk can be an impediment to arbitraging away expected return effects due to mispricing. Pastor and Stambaugh (2000) also adopt a prior for α with covariance matrix proportional to the residual covariance matrix. Building on ideas in MacKinlay (1995), they stress the desirability of a positive association between α and Σ in the prior, which makes extremely large Sharpe ratios less likely, as implied by (4).[10]

A very high Sharpe ratio can be viewed as an approximate arbitrage opportunity (see Shanken (1992)) in that a substantial return is expected with relatively little risk. It can also be interpreted as an opportunity to significantly "beat the market" insofar as expected return can be enhanced while retaining the total risk level of investing in the market. That researchers have prior beliefs about these issues is evident in the following quote from Cochrane and Saa-Requejo (2000, p. 82–83):

[9] Using improper (diffuse) priors for "nuisance parameters" like betas and residual covariances that appear in both the null and alternative models, but proper (informative) priors under the alternative for parameters like alpha, is in keeping with Jeffreys (1961) and others.

[10] Also see related work by Pastor and Stambaugh (1999) and Pastor (2000).

> There is a long tradition in finance that regards high Sharpe ratios as "good deals" that are unlikely to survive, since investors would quickly grab them up ... we assume that the investor would take any opportunity with a Sharpe ratio twice that of the S&P 500 ... Since most fund managers seem desperate for average returns a few percent above the S&P 500 index, this value seems conservative.

In this spirit, we think about plausible values for the maximum Sharpe ratio and use this to specify the value of k in the prior.

For a single asset, equation (6) implies that k is the prior expectation of the squared alpha divided by residual variance, or the square of the asset's *information ratio*. By equation (4), this is the expected increment to the maximum squared Sharpe ratio from adding the asset to the given factors. In general, with a vector of N returns, the quadratic form $\alpha'(k\Sigma)^{-1}\alpha$ is distributed as chi-square with N degrees of freedom, so the prior expected value of the increase in the squared Sharpe ratio, $\alpha'\Sigma^{-1}\alpha$, is k times N. Therefore, given a target value Sh_{max} for the square root of the maximum expected squared Sharpe ratio under the alternative, the required k is

$$k = \left(Sh_{max}^2 - Sh(f)^2\right)/N\,. \tag{7}$$

For example, in a test of the CAPM, Sh_{max}= 2.0 times the market ratio of 0.115 implies that k is $(2.0^2-1)(0.115^2)/10 = 0.004$ when there are 10 test assets.[11] As an alternative to this investment perspective on the prior, we can focus directly on the expected return relation and our assessment of plausible deviations from that relation. This is similar to the approach in Pastor (2000).

One may wonder why we do not consider a diffuse prior for alpha, to avoid having to make an assumption about the prior parameter k. In some contexts, a diffuse prior can be obtained in the limit by letting the prior variance approach infinity. However, allowing $\sigma_\alpha \to \infty$ here would amount to letting $k \to \infty$, which is not economically sensible since, by equation (7), this would imply that the maximum squared Sharpe ratio expected under the alternative is itself infinite. That some form of *informative* prior is required for alpha follows more generally from observations in a widely cited article on Bayes factors by Kass and Raftery (1995, p. 782). They discuss the relevance of "Bartlett's (1957) paradox," a situation whereby an estimate may be far from its null value, but even more unlikely under the alternative, thus yielding a Bayes factor that favors the null H_0. A consequence, they note, is that a prior under the alternative with a large variance will "force the Bayes factor to favor H_0." They attribute this point to Jeffreys (1961), noting that, "to avoid this difficulty, priors on parameters being tested must be proper and not have too big a spread." What this means in our application is that, in evaluating an asset pricing model, a researcher needs to think about how large the deviations from the model might plausibly be if the model is in fact false.

[11] Note that using the factor data to inform the prior is appropriate in this context since the entire statistical analysis is conditioned on f. We obtain similar results using a bias-adjusted estimate of the squared Sharpe ratio.

One general method that can be used to obtain a proper prior is to update a diffuse prior with a "minimal training sample," that is, a subset of the data that is just large enough to identify all the model parameters. The resulting posterior distribution can then play the role of the prior in analyzing the remaining data.[12] Avramov and Chao (2006) adopt a variant of this approach in comparing asset pricing models. Although computationally convenient, a potential concern is that such a prior for alpha could have a very large standard deviation, one that would be judged *economically* implausible, and might conflict with Jeffreys's recommendation that the prior spread not be "too big."

B.2. Bayes Factor and Posterior Model Probability

The Bayes factor BF measures the relative support for the null hypothesis $H_0 : \alpha = 0$ versus the alternative $H_1 : \alpha \neq 0$ in the data. Formally, BF is the ratio of the marginal likelihoods, $ML(H_0)/ML(H_1)$, where each marginal likelihood is a weighted average of the likelihoods over various parameter values. The weighting is by the prior densities associated with the different hypotheses. Since the parameters are integrated out, the marginal likelihood can be viewed as a function of the data (factor and test-asset returns):

$$ML = P(R|F) = \int\int P(R|F, \alpha, \beta, \Sigma)\, P(\alpha|\beta, \Sigma)\, P(\beta, \Sigma)\; d\alpha\, d\beta\, d\Sigma. \tag{8}$$

Here, the likelihood function is the joint conditional density $P(R|F, \alpha, \beta, \Sigma)$ viewed as a function of the parameters, $ML(H_1)$ is computed using the priors given in (5) and (6), and $ML(H_0)$ also uses (5) but substitutes the zero vector for α.

We can also view the test of $H_0 : \alpha = 0$ versus $H_1 : \alpha \neq 0$ in terms of the proportionality constant in the prior covariance matrix for α. Thus, we have a test of the value zero versus the value k. More generally, the null hypothesis can be modified to accommodate an approximate null that allows for small deviations from the exact model, as captured by the prior parameter $k_0 < k$. The usual exact null is obtained with $k_0 = 0$. We can now state a key result.

PROPOSITION 1: *Given the factor model in* (1) *and the prior in* (5) *and* (6), *the unrestricted marginal likelihood* $ML(H_1)$ *is proportional to*[13]

$$|F'F|^{-N/2} |S|^{-\frac{(T-K)}{2}} Q, \tag{9}$$

[12] Berger and Pericchi (1996) suggest averaging such results across all minimal training samples to increase stability of the procedure.

[13] By "unrestricted ML" we mean that the corresponding regression density does not restrict alpha to be zero. Of course, the informative prior under the alternative restricts our view about different values of alpha.

and the restricted marginal likelihood $ML(H_0)$ is proportional to

$$|F'F|^{-N/2}|S_R|^{-\frac{(T-K)}{2}}, \tag{10}$$

where S and S_R are the $N \times N$ cross-product matrices of the OLS residuals with α unconstrained or constrained to zero, respectively. The scalar Q is given by

$$Q = \left(1 + \frac{a}{(a+k)}(W/T)\right)^{-(T-K)/2}\left(1 + \frac{k}{a}\right)^{-N/2}, \tag{11}$$

where $a = (1 + Sh(F)^2)/T$ and W, given in (3), equals the GRS F-statistic times $NT/(T-N-K)$. Therefore, the Bayes factor for $H_0 : \alpha = 0$ versus $H_1 : \alpha \neq 0$ equals

$$BF = \frac{ML(H_0)}{ML(H_1)} = \frac{1}{Q}\left(\frac{|S|}{|S_R|}\right)^{(T-K)/2}. \tag{12}$$

Letting Q_{k_0} be the value of Q obtained with prior value k_0, the Bayes factor for k_0 versus k is

$$BF_{k_0,k} = Q_{k_0}/Q. \tag{13}$$

PROOF: See Appendix B.[14,15]

Consistent with intuition, the marginal likelihoods increase with the fit of each model, which is inversely related to the OLS residual sum of squares (unrestricted or restricted). However, it is well known from classical statistics that the fit implied by the OLS estimates overstates the true fit.[16] In Proposition 1, the presence of the $F'F$ term in the marginal likelihoods can be traced back algebraically to this fact. It amounts to a prior-weighted adjustment of fit to reflect measurement error in the OLS estimates and beliefs about the true regression parameters. Although $F'F$ drops out when comparing nested models, these terms play a role in our general model comparison below in Proposition 3. As to the remaining term, a large Q indicates a relatively small "distance" between the alpha estimates and the values of alpha anticipated under the prior for the unrestricted model (see Appendix A). It can be shown that Q is always between zero and one, thereby reducing $ML(H_1)$ and adjusting for the overstatement of fit obtained with the OLS alpha estimates.

[14] Harvey and Zhou (1990) derive (12) in the univariate case, with a complicated integral expression for Q. The function of W in (11) is our simplification, while (13) is both a further simplification and a generalization of (12).

[15] The formula is identical, apart from minor differences in notation, to the Bayes factor that Shanken (1987b) derives by conditioning directly on the F-statistic (rather than on all the data) for simplicity. Thus, surprisingly, it turns out that this simplification entails no loss of information under the diffuse prior assumptions made here.

[16] In a frequentist context, this gives rise to the familiar OLS degrees-of-freedom adjustment.

It can be verified that BF is a decreasing function of W; moreover, the larger the test statistic, the stronger is the evidence against the null that α is zero. When $N = 1$, W equals $T/(T-K-1)$ times the squared t-statistic for the intercept in the factor model. Other things equal, the greater the magnitude and precision of the intercept estimate, the bigger is that statistic, the lower is BF, and the weaker is the support for the null. For $N > 1$, the same conclusion applies to the maximum squared t-statistic over all portfolios of the test assets. In terms of the representation in (12), BF naturally decreases as the determinant of the matrix of restricted OLS sums of squared residuals increases relative to that for unrestricted OLS, suggesting that the zero-alpha restriction does not fit the data. As the ratio of determinants is less than one, a BF favoring the null ($BF > 1$) occurs when Q is sufficiently low, that is, the prior for alpha under the alternative is "inconsistent" with the estimate. The simple-to-compute formula for Q will also facilitate the model comparison calculations in Section III.

II. Relative versus Absolute Model Tests

In the previous section we analyze a test of a factor pricing model against a more general alternative. We refer to this as an absolute test of the "fit" of a model, that is, the extent to which the model's zero-alpha restriction matches up with the empirical estimates.[17] In this section we address the testing of one factor pricing model against another such model, what we call a relative test. Assume, as in the previous section, that there are K factors and N test assets of interest. In general, we consider models corresponding to all subsets of the factors, with the stipulation that the market factor, *Mkt*, is always an included factor (we relax this requirement later). This requirement is motivated by the fact that the market portfolio represents the aggregate supply of securities and therefore holds a unique place in portfolio analysis and the equilibrium pricing of assets, for example, the Sharpe-Lintner CAPM and the Merton (1973) intertemporal CAPM.[18]

In the present setting, the vector f corresponds to an $L - 1$ subset of the $K - 1$ nonmarket factors. The model associated with the L factors, (Mkt, f), is denoted by M, and the $K - L$ factors excluded from M are denoted by f^*. A valid model M will price the factor returns f^* as well as the test-asset returns r. Thus, the alphas of f^* and r regressed on (Mkt, f) equal zero under the model. However, the statistical analysis is greatly facilitated by using an equivalent representation of M. Let

$$f^* = \alpha^* + \beta^* \left[Mkt,\, f\right] + \varepsilon^* \tag{14}$$

and

$$r = \alpha_r + \beta_r \left[Mkt,\, f,\, f^*\right] + \varepsilon_r \tag{15}$$

[17] This should not be confused with fit in the sense of the factor model time-series R^2 for test assets. A high R^2 indicates that the factors explain much of the variation in realized returns, but does not rule out large alphas, which measure the model's *expected* return errors.

[18] See Fama (1996) for an analysis of the role of the market portfolio in the ICAPM.

be multivariate regressions for f^* and r. Note that the model in (15), which we call M_a, includes all K factors. Our analysis builds on the following pricing result, which is Proposition 1 in Barillas and Shanken (2017):

RESULT: The model M, which is nested in M_a, holds if and only if $\alpha^* = 0$ and $\alpha_r = 0$, that is, if and only if M "prices" the excluded factors and M_a "prices" the test assets.

In other words, given that the excluded-factor alphas on M are zero, the additional requirement that the test-asset alphas on M be zero is equivalent to the test-asset alphas being zero in the regressions on all the factors. Thus, M can be characterized as a constrained (the excluded-factor constraint) version of the all-factors model, with the latter model's restriction on test assets common to both models. As discussed in Barillas and Shanken (2017), this is the key to "test-asset irrelevance" based on the likelihood criterion. It implies that the impact of test assets on a model's likelihood is the same for each model and thus cancels out in model comparison.

More formally, the result above implies that the restricted joint density of returns under M is a product of three terms—an unrestricted term corresponding to the model's factors, a term corresponding to the factors that are excluded from the model, and a term corresponding to the test-asset returns. The second and third terms both impose the model's zero-alpha restrictions. However, the third term is the same for all models because the corresponding test-asset restriction does not depend on which factors are included in the model. The approach below extends this logic to the *marginal* likelihoods obtained by averaging likelihoods in accordance with priors over the parameters in each model. We focus on nested models in this section; in the following section we use the fact that the test-asset irrelevance conclusion holds for all factor-based models, nested or nonnested.[19]

For example, consider CAPM as nested in the Fama-French (1993) three-factor model (FF3). In this case, the usual alpha restrictions of the single-factor CAPM are equivalent to the one-factor CAPM intercept restriction for the excluded-factor returns, *HML* and *SMB*, and the FF3 intercept restriction for the test-asset returns. As the test-asset restrictions are common to both models, the models differ only with respect to the excluded-factor restrictions. If those restrictions hold, CAPM is favored over FF3 in the sense that the same pricing is achieved with fewer factors—a more parsimonious model.[20] Thus, the factors *SMB* and *HML* are *redundant* in this case. Otherwise, FF3 is preferred since it does not impose the additional restrictions that fail to hold. Similarly, if $\alpha^* = 0$, the tangency portfolio (and associated Sharpe ratio) based on all the factors, that is, spanned by *Mkt*, *HML*, and *SMB*, can be achieved

[19] Barillas and Shanken (2017) note that irrelevance does not require normality and holds under elliptical distributions that allow for heteroskedasticity of the residuals conditional on the factors and certain forms of serial correlation.

[20] See Corollary 1 in Barillas and Shanken (2017).

through investment in Mkt alone. If $\alpha^* \neq 0$, a higher (squared) Sharpe ratio can be obtained by exploiting all of the factor investment opportunities.

Three asset pricing tests naturally present themselves in connection with the nested-model representation of M. We can conduct a test of the all-inclusive model M_a with factors (Mkt, f, f^*) and left-hand-side returns r. Also, we can test M with factors f and left-hand-side returns consisting of test assets r plus the excluded factors f^*. These *absolute* tests pit the models (M_a or M) against more general alternatives for the distribution of the left-hand-side returns, with nonzero alphas. Finally, we can perform a *relative* test of M versus M_a with factors f and left-hand-side returns f^*. There is a simple relation between the Bayesian versions of these tests. We denote the Bayes factor for M_a in the first test as $BF^{abs}_{M_a}$, for M in the second test as BF^{abs}_{M}, and for M (versus M_a) in the third test as BF^{rel}.

PROPOSITION 2: *Assume that the multivariate regressions of f^* on (Mkt, f) in* (14) *and r on (Mkt, f, f^*) in* (15) *satisfy the condition that the residuals are independently distributed over time as multivariate normal with mean zero and constant residual covariance matrix. The prior for the regression parameters is of the form in* (5) *and* (6)*, with the priors for* (14) *and* (15) *independent. Then the BFs are related as follows:*

$$BF^{abs}_{M} = BF^{rel} \times BF^{abs}_{M_a}. \tag{16}$$

PROOF: The marginal likelihood is the expectation under the prior of the likelihood function. Write the joint density (likelihood function) of excluded-factor and test-asset returns, conditional on (Mkt, f), as the density for f^* given (Mkt, f), times the conditional density for r given (Mkt, f, f^*). Using the prior independence assumptions, the prior expectation of the product is the product of the expectations. By the earlier discussion, both densities are restricted (zero intercepts) under M, whereas only the density for r is restricted under M_a. These restrictions affect the marginal likelihoods in the numerators of the absolute-test BFs for these models. Therefore, letting the subscripts R and U stand for restricted and unrestricted densities,

$$BF^{abs}_{M} = \left\{ML_R\left(f^*\right) \times ML_R\ (r)\right\} / \left\{ML_U\left(f^*\right) \times ML_U\,(r)\right\}$$

and

$$BF^{abs}_{M_a} = \left\{ML_U\left(f^*\right) \times ML_R\ (r)\right\} / \left\{ML_U\left(f^*\right) \times ML_U\,(r)\right\},$$

where the conditioning variables have been suppressed to simplify the notation. Given that

$$BF^{rel} = \left\{ML_R\left(f^*\right) \times ML_R(r)\right\} / \left\{ML_U\left(f^*\right) \times ML_R(r)\right\},$$

the equality in (16) is easily verified.

Proposition 2 tells us that the absolute support for the nested model M equals the relative support for M compared to the larger (less-restrictive) model M_a

times the absolute support for M_a. Equivalently, the relative support for M versus M_a can be backed out from the absolute BFs as $BF_M^{abs}/BF_{M_a}^{abs}$. Thus, whether we compare the models directly or whether we relate the absolute tests for each model, the result is the same. This reflects the fact that the impact of the test-asset returns r on the absolute tests, $ML_R(r)$, is the same for each model and so cancels out in the model comparison—a Bayesian extension of the argument in Barillas-Shanken (2017).

III. Simultaneous Comparison of All Models Based on a Set of Factors

In the previous section, we saw how to compare two nested factor pricing models. Now suppose we wish to simultaneously compare a collection of asset pricing models, both nested and nonnested.[21] One question that arises in model comparison is how to accommodate the possibility that none of the models under consideration is exactly true. In our context, while the characterization of what it means for a model to hold requires that the test-asset and excluded-factor alphas be equal to zero, we do not presume that any model under consideration exactly satisfies this requirement as an empirical matter. It is therefore possible (and likely) that some relevant factors have not been identified. Nonetheless, we wish to compare the given models. One approach in this case would be to explore approximate versions of the models, as discussed earlier. Another is to interpret the marginal likelihoods and implied posterior probabilities as measures of the *relative success* of the exact models at predicting the data or the *comparative support* the data provide for the models.[22]

Since test-asset returns drop out of the model comparison, however, our Bayesian analysis can also be interpreted as providing the probability that a given model is best in terms of pricing the factor returns. Fama (1998) considers a related hypothesis in identifying the number of priced state variables in an intertemporal CAPM setting. Of course, a model that includes all of the factors will always perfectly price the factor returns and will yield the highest Sharpe ratio. Therefore, consistent with a desire for parsimony that is evident in the literature, our notion of best also requires that the model not include redundant factors. Thus, a model that does contain such factors need not be assigned much probability, even though it may generate a high Sharpe ratio.

A. The Model Comparison Methodology

As discussed in Section II, our methodology exploits the fact that each model can be viewed as a restricted version of the model that includes all of the factors under consideration. This leads to a convenient decomposition of the marginal likelihood for each model and the observation that the test-asset returns drop out in comparing models. Thus, inference about model comparison ends up

[21] Simultaneous inference about model comparison in a classical framework could potentially be based on the approach of Hansen, Lunde, and Nason (2011).

[22] See Kass and Raftery (1995) and Berger and Pericchi (1996) for discussion of these issues.

being based on an aggregation of the evidence from all possible multivariate regressions of excluded factors on factor subsets.

We use braces to denote models, which correspond to subsets of the given factors. For example, starting with the FF3 factors, there are four models that include *Mkt*: CAPM, FF3, and the nonnested two-factor models {*Mkt HML*} and {*Mkt SMB*}. Given the ML_j for each model M_j with prior probability $P(M_j)$, the posterior probabilities conditional on the data D are given by Bayes's rule as

$$P(M_j|D) = \{ML_j \times P(M_j)\} \times \left\{\sum_i ML_i \times P(M_i)\right\}, \tag{17}$$

where D refers to the sample of *all* factor and test-asset returns, F and R.

One distinctive feature of our approach, as compared to Avramov and Chao (2006), is that the factors that are not included as right-hand-side *explanatory* variables for a given model play the role of left-hand-side *dependent* returns whose pricing must be explained by the model's factors.[23] This is important from a statistical standpoint, as well as an asset pricing perspective, since (17) requires that the posterior probabilities for all models be conditioned on the same data. Thus, each model's restrictions are imposed on the excluded factors f^* as well as the test assets r in calculating the marginal likelihood, whereas the marginal likelihood for the included factors f is based on their *unrestricted* joint density.[24] Therefore, we also need to consider the multivariate regression

$$f = \alpha + \beta Mkt + \varepsilon, \tag{18}$$

where the residuals are again independently distributed over time as multivariate normal with mean zero and constant residual covariance matrix. Thus, by an argument similar to that used in deriving Proposition 2, we obtain the following result.

PROPOSITION 3: *Assume that the multivariate regressions of f on Mkt in* (18), *f^* on (Mkt, f) in* (14), *and r on (Mkt, f, f^*) in* (15) *satisfy the distributional conditions discussed previously. The prior for the parameters in each regression is of the form in* (5) *and* (6), *with independence between the priors for* (18), (14),

[23] The phrase "you're either part of the problem or part of the solution" comes to mind.

[24] That all marginal likelihoods must, in principle, be conditioned on the same data is a direct consequence of Bayes's theorem (e.g., Kass and Raftery (1995), equation 1). In traditional model-comparison applications such as Avramov (2002), who examines subsets of predictors for returns in a linear regression framework, conditioning a model's likelihood on all the data reduces to conditioning on the predictors that are included in the model. In that setting, the excluded predictors drop out of the likelihood function and thus can be ignored in evaluating the given model. This occurs since imposing the model restrictions amounts to placing slope coefficients of zero on those predictors. The same is not true for the excluded factors in our application, as their pricing by the included factors *does* affect the model likelihood.

and (15) *conditional on the sample of Mkt returns. Then the marginal likelihood for a model M with nonmarket factors f is of the form*

$$ML = ML_U\,(f|Mkt) \times ML_R\left(f^*|Mkt,\, f\right) \times ML_R\left(r|Mkt,\, f,\, f^*\right), \tag{19}$$

where the unrestricted and restricted (alpha = 0) regression marginal likelihoods are obtained using (9) *and* (10)*, respectively.*[25]

Here, $ML_U(f\ |Mkt)$ is calculated by letting f play the role of r (the left-hand-side returns) and *Mkt* the role of f (the right-hand-side returns) in (9). Similarly, $ML_R(f^*\ |Mkt, f)$ is computed by letting f^* play the role of r and (Mkt, f) the role of f in (10), etc.[26] The value of k in the prior for the intercepts in the unrestricted regressions is determined as in (7), but using the number of non-*Mkt* factors K–1 in the denominator, with Sh_{max} corresponding to all K factors and $Sh(Mkt)$ substituted for $Sh(f)$. It follows from the discussion there that k is the expected (under the alternative prior) increment to the squared Sharpe ratio at each step from the addition of one more factor. By concavity, therefore, the increase in the corresponding Sh_{max} declines as more factors are included in the model. The main conclusions are not sensitive to alternative methods that we tried for distributing the total increase in the squared Sharpe ratio.

Given (19), the posterior model probabilities in (17) can now be calculated by substituting the corresponding marginal likelihood for each model. We use uniform prior model probabilities to avoid favoring one model or another, which seems desirable in this research setting. Thus, we highlight the impact of the *data* on beliefs about the models. Other prior assumptions could easily be explored, however. For example, one may want to give greater weight to models that are judged to have stronger theoretical foundations. Note that, since the last term in (19) conditions on *all* of the factors, it is the same for all models and so cancels out in the numerator and denominator of (17). Thus, test assets are irrelevant for this Bayesian model comparison, as in the nested case of Proposition 2.

Before we move on to the empirical analysis, it is worth highlighting an important difference between our approach to model comparison and more conventional asset pricing tests. The latter, whether it be a variation on the classical GRS test or a Bayesian test like that in Proposition 1, is concerned solely with the restriction that alpha is zero when excess returns are regressed on the model factors. In our approach, however, there is an additional requirement that all of the model's factors actually be needed for pricing. This gives rise to the unrestricted component of a model's marginal likelihood, based on a prior for alpha that assumes each included factor increases the attainable Sharpe ratio. The overall joint measure of model likelihood is then the product of the restricted and unrestricted components. Therefore, it is not simply a

[25] Note that the entire statistical analysis of model comparison is conditioned on *Mkt*.

[26] It can be shown that, in this context, the total impact of the proportionality constants given in (A1) and (A2) of Appendix A is the same for each model and thus may be ignored in model comparison.

matter of identifying which set of factors produces the highest Sharpe ratio, but also whether a model does so in a parsimonious manner, given our prior beliefs about alphas.

In the specification test of Proposition 1, we imposed the same improper prior on the "nuisance parameters" β and Σ under the null and the alternative hypotheses (see footnote 9). But in our model comparison, the parameterizations differ across models. Berger and Pericchi (2001) mention that, in such cases, choosing the same multiplicative constant in the improper priors for models whose parameter spaces have different dimensions can be a bad idea. However, the number of nuisance parameters is the same for all models in our application.[27] Berger and Pericchi (2001) further note that it is reasonable to choose the same multiplicative constant in the improper priors when the parameters in the models are "essentially similar," and it is common practice to do so when they are the same. In the Internet Appendix, we show how to compute posterior model probabilities using a more complicated and computationally intensive variation on our basic methodology. Under this approach, the parameters for which improper priors are imposed are either identical across models or identical after a permutation of the nonmarket factors. Thus, it is natural to take all of the multiplicative constants as equal with this method. Empirically, the simpler approach presented in the text and the more complicated method yield nearly the same model probabilities. See the Internet Appendix for further details.

B. Extension When the Market Is Not Automatically Included in the Model

Given its unique role in asset pricing theory and investment analysis, the market factor is virtually always included in empirical analysis of models with traded factors. However, in some contexts, it may be desirable to explore models with mimicking portfolios for nontraded factors like the growth rates of aggregate consumption or industrial production. Presumably, the market would still be considered a potential factor in this context, but we may not want to insist a priori that it is included in the model. Therefore, in this section, we describe an extension of the methodology to accommodate this scenario.

The approach that we adopt randomizes the factor that plays the "anchor" role of *Mkt* in Proposition 3. We implement this by the formal device of placing a uniform prior over which factor is first in parameterizing the joint density of returns (different probabilities can also be incorporated). Conditional on *Mkt* being selected, the analysis proceeds exactly as before. But there is an equal probability that another factor, say *HML*, will play this role. Given that, the computations are the same except that *HML* and *Mkt* are exchanged everywhere. In this way, we can derive *conditional* posterior probabilities for each

[27] To see this, let K be the total number of factors and L be the number of factors in model M. There are L–1 betas for f in (19) and $(K-L)L$ for f^*. The number of residual variance/covariance parameters is $(L-1)L/2$ for f and $(K-L)(K-L+1)/2$ for f^*. One can verify that the sum of these terms is $K(K+1)/2 - 1$, independent of L.

anchor scenario. We then aggregate these probabilities to obtain the ones of interest—model probabilities that are "unconditional" in the sense of not being conditioned on which factor plays the anchor role. The details are provided at the end of Appendix C. Using this method, every factor has the same prior probability of being in the model, but no factor is guaranteed to be included.

IV. Comparing Models with Categorical Factors

In empirical work, several of the available factors often amount to different implementations of the same underlying concept (different ways of measuring the same factor), for example, size or value. In such cases, to avoid overfitting, it may be desirable to structure the prior so that it only assigns positive probability to models that contain at most one version of the factors in each category. In this section we extend our analysis to accommodate this perspective.

Our main empirical application presented in Section V below will include data for four factor categories. These data are available over the period 1972 to 2015. In the present illustration of the methodology, we examine a subset of those factors over the same period: two size factors, *SMB* from FF5 and *ME* from HXZ, along with *Mkt* and *HML*. The size factors differ in terms of the precise sorts used to construct the "small" and "big" sides of the return spreads. We refer to *Size* as a *categorical factor*, in this context, in contrast to the actual factors *SMB* and *ME*. Similarly, models in which some of the factors are categorical and the rest are standard factors are termed *categorical models*. We assign equal prior probabilities that *Size* and *HML* are in the model, while splitting the probability equally between the *SMB* and *ME* versions of the size factor.

To demonstrate the basic idea, consider categorical models based on the standard factors *Mkt* and *HML*, and the categorical factor *Size*. There are four categorical models—CAPM, {*Mkt HML*}, {*Mkt Size*}, and {*Mkt HML Size*}—each with prior probability 1/4.[28] We have two versions of the factors, $w_1 = (Mkt\ HML\ SMB)$ and $w_2 = (Mkt\ HML\ ME)$, and we can conduct separate model-comparison analyses with each over the 1972 to 2015 period. These separate analyses employ the methodology of Section III with all standard factors. The posterior model probabilities conditional on w_1 are {*Mkt HML*} 59.8%, {*Mkt HML SMB*} 39.2%, CAPM 0.6%, and {*Mkt SMB*} 0.4%. Conditional on w_2, we have {*Mkt HML*} 51.9%, {*Mkt HML ME*} 47.1%, {*Mkt ME*} 0.5%, and CAPM 0.5%. Note that the probabilities for models that include *SMB* are similar to those for models that include *ME*. This makes sense since the correlation between the two size factors is very high (0.98). The question then becomes, how should we aggregate these two sets of probabilities to obtain posterior probabilities for all six models?

[28] The value of k in the prior corresponds to a potential 50% increase in the Sharpe ratio relative to that of the market when the categorical model contains six factors, as discussed in the next section. For the three-factor model, the implied Sh_{max} is $1.27 \times Sh(Mkt)$, which is equal to the square root of $2k + Sh(Mkt)^2$ (see (7)).

First, suppose we assign a prior probability of 1/2 to each w, that is, to each version of the factors, and conditional prior probabilities of 1/4 for the four models in each w. The *unconditional* prior probabilities for CAPM and {*Mkt HML*}, the two models common to versions w_1 and w_2, are then (1/2)(1/4) + (1/2)(1/4) = 1/4 in each case. In contrast, the probability for {*Mkt SMB*}, which is associated with just one version of the factors, is (1/2)(1/4) + (1/2)(0) = 1/8, and likewise for {*Mkt ME*} and the three-factor models. Thus, this simple prior specification effectively splits the categorical model probabilities for {*Mkt Size*} and {*Mkt HML Size*} equally between the two versions of these categorical models, as desired.

Proposition C1 in Appendix C derives a formula for the posterior probability of each version of the factors and shows that the final model probabilities can be obtained by applying these weights to the conditional model probabilities above. The weights in this case are 47.4% for w_1 and 52.6% for w_2. The probability for {*Mkt HML*} is then (47.4%)(59.8%) + (52.6%)(51.9%) = 55.6%. For {*Mkt HML ME*}, which is associated only with the second version of the factors, it is (47.4%)(0) + (52.6%)(47.1%) = 24.8%. The probability is 18.6% for {*Mkt HML SMB*} and less than 1% for the remaining models. From the categorical model perspective, we have probability 55.6% for {*Mkt HML*}, 24.8% + 18.6% = 43.4% for {*Mkt HML Size*}, and less than 1% for the other models.

Thus, in this case the data, taken together with the prior, favor a parsimonious two-factor model over retaining the maximum number of factors. Intuitively, this reflects the basic regression evidence: the *HML* alpha on *Mkt* is highly "significant" (5.50% annualized with a t-statistic of 3.69), while the alphas for *SMB* and *ME* on {*Mkt HML*} are modest with t-statistics of around one.

V. Model Comparison Results with 10 Prominent Factors

A. The Factors

We now consider a total of 10 candidate factors. First, there are the traditional FF3 factors *Mkt*, *HML*, and *SMB* plus the momentum factor *UMD*. To these we add the investment factor *CMA* and the profitability factor *RMW* of Fama and French (2015). We also include the size (*ME*), investment (*IA*), and profitability (*ROE*) factors of Hou, Xue, and Zhang (2015a). Finally, we have the value factor HML^m from Asness and Frazzini (2013). The size, profitability, and investment factors differ based on the type of stock sorts used in their construction.

Fama and French (2015) create factors in three different ways. We use what they refer to as their "benchmark" factors. Similar to the construction of *HML*, these are based on independent (2 × 3) sorts, interacting size with operating profitability for the construction of *RMW* and with investment to create *CMA*. The factor *RMW* is the average of the two high-profitability portfolio returns minus the average of the two low-profitability portfolio returns. Similarly, *CMA* is the average of the two low-investment portfolio returns minus the average of the two high-investment portfolio returns. Finally, *SMB* is the average of the returns on the nine small-stock portfolios from the three separate 2×3 sorts minus the average of the returns on the nine big-stock portfolios.

Hou, Xue, and Zhang (2015a) construct their size, investment, and profitability factors from a triple ($2 \times 3 \times 3$) sort on size, investment-to-assets, and *ROE*. More importantly, the *HXZ* factors use different measures of investment and profitability. Fama and French (2015) measure operating profitability as NI_{t-1}/BE_{t-1}, where NI_{t-1} is earnings for the fiscal year ending in calendar year $t-1$, and BE_{t-1} is the corresponding book equity. HXZ use a more timely measure of profitability, ROE, which is income before extraordinary items taken from the most recent public quarterly earnings announcement divided by one-quarter-lagged book equity. The factor *IA* is the annual change in total assets divided by one-year-lagged total assets, whereas investment used by Fama and French (2015) is the same change in total assets from the fiscal year ending in year $t-2$ to the fiscal year ending in $t-1$, divided by total assets from the fiscal year ending in $t-1$, rather than $t-2$. As to value factors, HML^m is based on book-to-market rankings that use the most recent monthly stock price in the denominator. This is in contrast to Fama and French (1993), who use annually updated lagged prices in constructing *HML*. The sample period for our data is January 1972 to December 2015. Some factors are available at an earlier date, but the HXZ factors start in January of 1972 due to the limited coverage of earnings announcement dates and book equity in the Compustat quarterly files.

Rather than mechanically apply our methodology with all nine of the nonmarket factors treated symmetrically, we apply the framework of Section IV, which recognizes that several of the factors are just different ways of measuring the same underlying construct. Specifically, we only consider models that contain at most one version of the factors in each of the following categories: size (*SMB* or *ME*), profitability (*RMW* or *ROE*), value (*HML* or HML^m), and investment (*CMA* or *IA*). We refer to *Size, Profitability, Value, and Investment* as the *categorical factors*. The standard factors in this application are *Mkt* and *UMD*. Since each categorical model has up to six factors and *Mkt* is always included, there are 32 (2^5) possible categorical models. Given all the possible combinations of *UMD* and the different types of size, profitability, value, and investment factors, there are a total of 162 models under consideration.

Our benchmark scenario assumes that $Sh_{max} = 1.5 \times Sh(Mkt)$, that is, the square root of the prior expected squared Sharpe ratio for the tangency portfolio based on all six factors is 50% higher than the Sharpe ratio for the market only. Given the discussion in Section III.A, this is sufficient to determine the implied Sh_{max} values as we expand the set of included factors from one to six, leaving the intercepts unrestricted. We think of the 1.5 six-factor choice of multiple as a prior with a risk-based tilt, assigning relatively little probability to extremely large Sharpe ratios. Below, we examine the sensitivity of posterior beliefs to this assumption, as we also explore multiples corresponding to a more behavioral perspective (more mispricing) and one with a lower value.[29]

B. Empirical Results on Model Comparison

In this section we present model-comparison evidence for the sample period 1972 to 2015. Model probabilities are shown at each point in time to provide a historical perspective on how posterior beliefs would have evolved as the series of available returns has lengthened. Thus, we use all monthly data from January 1972 up to the given point in time in the (recursive) analysis and plot the corresponding model probabilities in Figure 1. Since we start with equal prior probabilities for each model, and given the substantial volatility of stock returns, it can take quite a while for a persistent spread in the posterior probabilities to emerge.

The top panel in Figure 1 shows posterior probabilities for the individual models. The models (and individual factors in the next panel) are ordered in the legend from the highest probability at the end of the sample to lowest. We find that quite a few of these models receive nontrivial probability, the best (highest-probability) model being the six-factor model {*Mkt SMB ROE IA* HML^m *UMD*}. This model has ranked first since 2000. The second-best model replaces *IA* with *CMA*, the third-best uses *ME* instead of *SMB,* and the sixth one uses *CMA* and *ME*, as opposed to *IA* and *SMB*. The fourth- and fifth-ranked are five-factor models that do not have a size factor and differ only in their investment factor choice. The top seven models all include *ROE*, HML^m, and *UMD*. All of these models fare better than FF5 and the four-factor model of HXZ, as do several other four-factor models, all of which contain *ROE*. The bottom panel of Figure 1 gives cumulative factor probabilities, that is, the sum of the posterior probabilities for models that include that factor. The probabilities are close to one for *ROE* and HML^m, with *UMD* around 97%. The probabilities drop substantially after that.

Figure 2 provides another perspective on the evidence, aggregating results over the different versions of each categorical model. Similar to the findings in the previous figure, the six-factor categorical model {*Mkt Value Size Profitability Investment UMD*} comes in first in the more recent years, with the posterior probability over 70% at the end of the sample. The five-factor model that excludes *Size* is next, with a probability of 20%. The third-best categorical model replaces the investment factor with size, while the fourth consists of the same five categories as in FF5. However, it is essential that the more timely versions of value and profitability are employed in these models. Specifically, the probability share for HML^m in the FF5 categorical model is 84.0%. This is the sum of the probabilities over versions of the categorical FF5 model that include HML^m divided by the total probability for that

[29] MacKinlay (1995) analyzes Sharpe ratios under risk-based and non-risk-based alternatives to the CAPM.

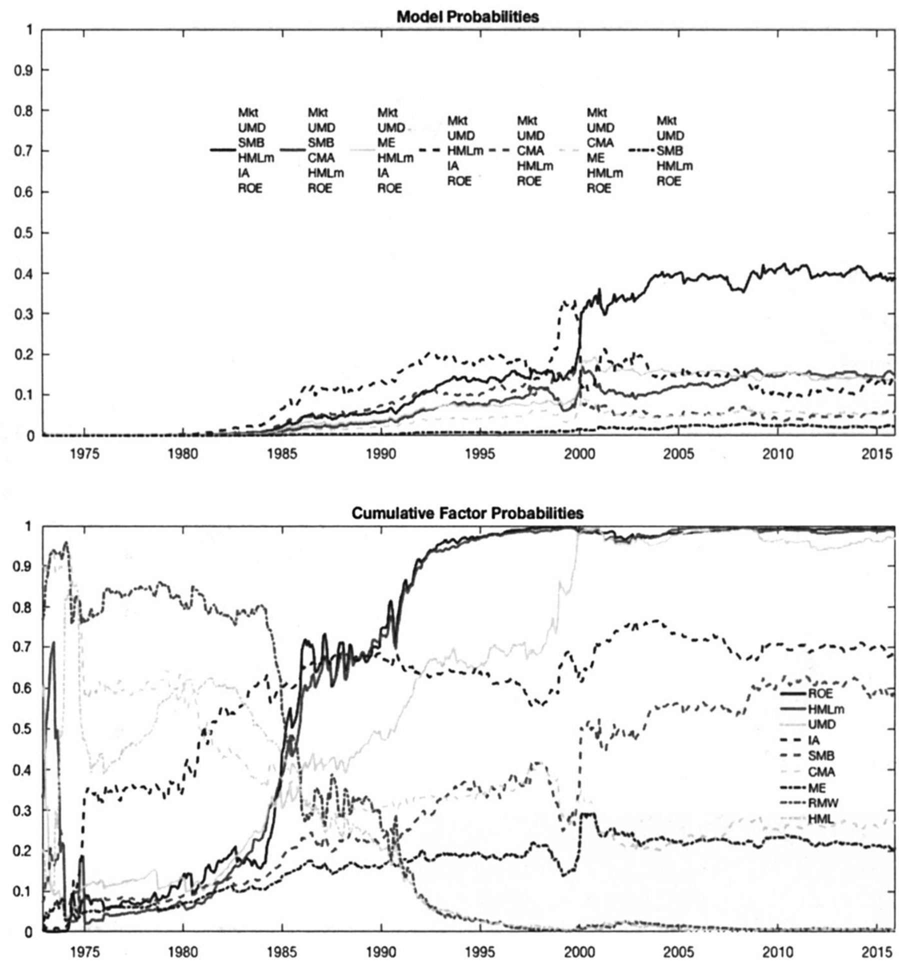

Figure 1. Model probabilities and cumulative factor probabilities. The top panel plots the time series of posterior model probabilities for the seven models with highest probability (ranked at the end of the sample). The sample periods are recursive, beginning in January 1972 and ending each month up to December 2015. Models are based on 10 prominent factors: the Fama and French (2015) factors (*Mkt, HML, SMB, CMA, RMW*), the HXZ (2015a) factors (*Mkt, ME, ROE, IA*), the Asness and Frazzini (2013) value factor HML^m, and the Carhart (1997) momentum factor *UMD*. Each model contains at most one factor from the following categories: size (*SMB* or *ME*), value (*HML* or HML^m), investment (*CMA* or *IA*), and profitability (*RMW* or *ROE*). The bottom panel plots the time series of cumulative posterior probabilities for each of the 10 factors. The prior is set so that $Sh_{max} = 1.5 \times Sh(Mkt)$, where $Sh(Mkt)$ is the sample *Mkt* Sharpe ratio and Sh_{max} is the square root of the expectation of the maximum squared Sharpe ratio with six factors included, taken with respect to the prior under the alternative that the nonmarket factor alphas are nonzero. (Color figure can be viewed at wileyonlinelibrary.com)

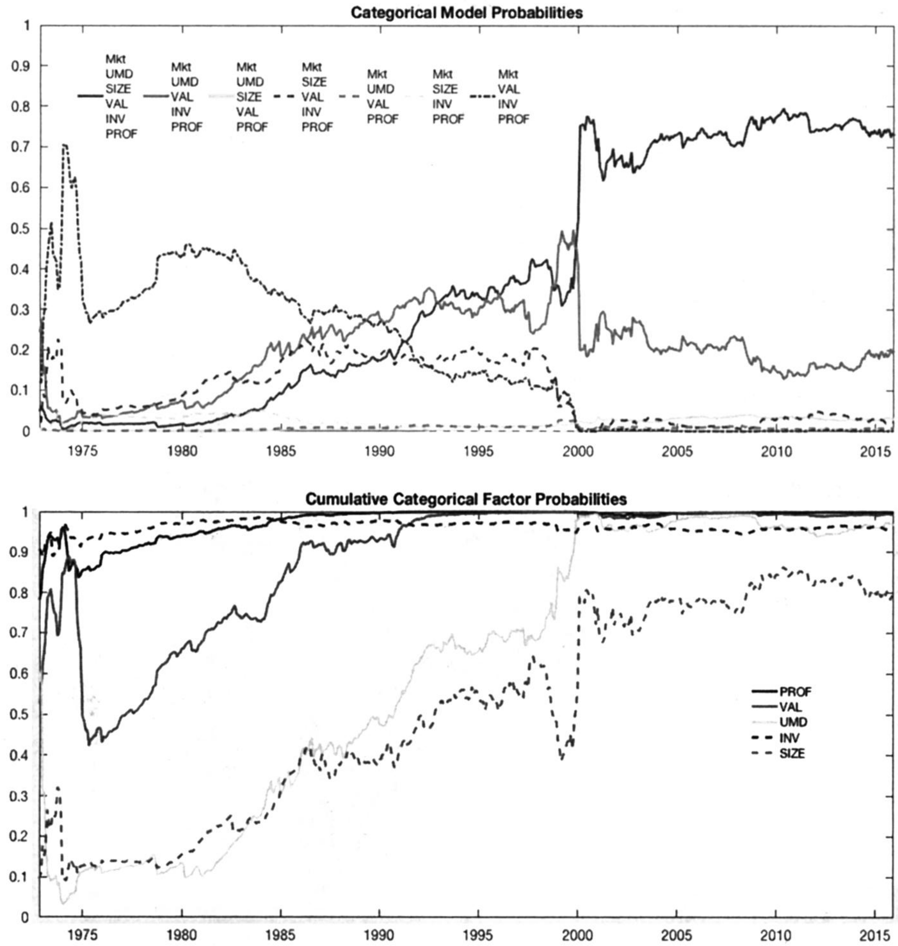

Figure 2. Categorical model probabilities and cumulative categorical factor probabilities. The top panel plots the times series of posterior model probabilities for the seven categorical models with highest probability (ranked at the end of the sample). The sample periods are recursive, beginning in January 1972 and ending each month up to December 2015. The factors consist of the market (*Mkt*) and momentum (*UMD*) factors, along with the categorical factors (two versions of each) size, value (*VAL*), investment (*INV*), and profitability (*PROF*). The bottom panel plots the time series of cumulative factor probabilities for each of the five categorical-model factors. The prior is set so that $Sh_{max} = 1.5 \times Sh(Mkt)$, where $Sh(Mkt)$ is the sample *Mkt* Sharpe ratio and Sh_{max} is the square root of the expectation of the maximum squared Sharpe ratio with six factors included, taken with respect to the prior under the alternative that the alphas are nonzero. (Color figure can be viewed at wileyonlinelibrary.com)

categorical model. Similarly, the share for *ROE* is 99.9% in the categorical FF5 model.

In terms of cumulative probabilities aggregated over *all* models, we see from the bottom panel of Figure 2 that the profitability category ranks highest since the mid-1980s. Interestingly, value is second with over 99% cumulative probability. Consistent with the findings in Figure 1, the *categorical share* for HML^m, that is, the proportion of the cumulative probability for value from models that include HML^m as opposed to *HML*, is 99.5%. Similarly, the categorical share of profitability is 99.4% for *ROE*. There is less dominance in the size and investment categories, with shares of 73.5% for *SMB* and 72.0% for *IA*.

B.1. Direct Model Comparison Results

While the analysis above simultaneously considers all 162 possible models, we also conduct direct tests that compare one model to another. In particular, we test our six-factor model against the recently proposed models of HXZ and Fama and French (2015). Such a test is easily obtained by working with the union of the factors in the two models and computing the marginal likelihood for each model as in (19). Assuming a prior probability of 0.5 for each model and zero probability for all other models, the posterior probability for model 1 in (17) is just $ML_1/(ML_1 + ML_2)$. Comparing the top individual model found above, $\{Mkt\ IA\ ROE\ SMB\ HML^m\ UMD\}$, to the four-factor model of HXZ, the direct test assigns 96.1% probability to the six-factor model. The probability is greater than 99% when compared to FF5, even if the size factor is deleted from the six-factor model.

B.2. Prior Sensitivity

The model comparison above is based on the prior assumption that $Sh_{\max} = 1.5 \times Sh(Mkt)$ when working with six factors. We next examine sensitivity to prior Sharpe multiples of 1.25, 1.5, 2, and 3. Tables I and II present the results for the individual and categorical models, respectively. Both tables show probabilities for the top seven models under the 1.5 multiple specification. The two best models, $\{Mkt\ SMB\ ROE\ IA\ HML^m\ UMD\}$ and $\{Mkt\ SMB\ ROE\ CMA\ HML^m\ UMD\}$, are also the two best under the more behavioral priors that allow for increases in the Sharpe ratio of two and three times the market ratio. Their probabilities rise from 24.8% to 50.0% and from 10.5% to 16.2%, respectively, as the multiple increases from 1.25. These two models are among the top four under the lower-multiple specification, although the probabilities of the different models are less spread out in this case.

The top model rankings for the categorical models in Table II are also fairly stable across the different priors. The six-factor categorical model {*Mkt SIZE PROF INV VAL MOM*} is always the best and the model that excludes *Size* comes in second, regardless of the prior. As the multiple rises, however, the posterior probability of the best model increases substantially, while the probability of the model that excludes *Size* declines. Below, we will see in Table IV that the *SMB* alpha on the other factors in the top-ranked six-factor model is 4.7% per annum, with a t-statistic of 3.0. Intuitively, as the prior multiple rises, the large *SMB* alpha becomes more consistent with the magnitude anticipated under the prior for the alternative, providing greater support for the conclusion

Table I

Prior Sensitivity for the Model Probabilities with 10 Factors

The table explores sensitivity to the prior of posterior model probabilities (in percent) for the seven models with highest probability (ranked at the end of the sample) when the prior multiple is set to 1.5 times the sample market Sharpe ratio, *Sh(Mkt)*. Four prior multiples are considered: 1.25, 1.5, 2.0, and 3.0. The sample period is January 1972 to December 2015. Models are based on 10 prominent factors: the Fama and French (2015) factors (*Mkt*, *HML*, *SMB*, *CMA*, *RMW*), the HXZ (2015a) factors (*Mkt*, *ME*, *ROE*, *IA*), the Asness and Frazzini (2013) value factor HML^m, and the Carhart (1997) momentum factor *UMD*. Each model contains at most one factor from the following categories: size (*Mkt* or *ME*), value (*HML* or HML^m), investment (*CMA* or *IA*), and profitability (*RMW* or *ROE*). The prior is set so that Sh_{max} = prior multiple × *Sh*(*Mkt*), where Sh_{max} is the square root of the expectation of the maximum squared Sharpe ratio with six factors included, taken with respect to the prior under the alternative that the nonmarket factor alphas are nonzero. *Sh*(*Mkt*) is 0.113 and the sample Sharpe ratio for the top-ranked model is 0.51 or 4.5 times *Sh*(*Mkt*).

Model/Prior Multiple	1.25	1.5	2	3
Mkt SMB ROE IA HML^m *UMD*	24.8	38.5	46.8	50.0
Mkt SMB ROE CMA HML^m *UMD*	10.5	14.8	16.3	16.2
Mkt ME ROE IA HML^m *UMD*	11.0	13.7	13.4	12.2
Mkt ROE IA HML^m *UMD*	17.5	13.7	9.3	8.0
Mkt ROE CMA HML^m *UMD*	7.9	6.0	4.0	3.3
Mkt ME ROE CMA HML^m *UMD*	4.8	5.4	4.8	4.1
Mkt SMB ROE HML^m *UMD*	1.5	2.3	3.0	4.1

Table II

Prior Sensitivity for the Categorical Model Probabilities with 10 Factors

The table explores sensitivity to the prior of posterior model probabilities (in percent) for the seven categorical models with highest probability (ranked at the end of the sample) when the prior multiple is set to 1.5 times the sample market Sharpe ratio, *Sh*(*Mkt*). Four prior multiples are considered: 1.25, 1.5, 2.0, and 3.0. The sample period is January 1972 to December 2015. The factors consist of the market (*Mkt*) and momentum (*UMD*) factors, along with the categorical factors (two versions of each) size, value (*VAL*), investment (*INV*), and profitability (*PROF*). The prior is set so that Sh_{max} = prior multiple × *Sh*(*Mkt*), where Sh_{max} is the square root of the expectation of the maximum squared Sharpe ratio with six factors included, taken with respect to the prior under the alternative that the nonmarket factor alphas are nonzero. *Sh*(*Mkt*) is 0.113 and the sample Sharpe ratio for the top-ranked model is 0.51 or 4.5 times *Sh*(*Mkt*).

Model/Prior Multiple	1.25	1.5	2	3
Mkt SIZE PROF INV VAL MOM	53.0	72.8	81.4	82.4
Mkt PROF INV VAL MOM	27.2	20.0	13.3	11.3
Mkt PROF SIZE VAL MOM	2.2	3.1	3.8	5.1
Mkt PROF SIZE VAL INV	8.7	2.4	0.7	0.3
Mkt PROF VAL MOM	1.2	0.9	0.7	0.8
Mkt SIZE PROF INV	3.1	0.5	0.1	0.0
Mkt PROF INV VAL	2.3	0.2	0.0	0.0

Table III

Categorical Factor Shares in the 10-Factor Analysis

The table explores sensitivity to the prior of the probability share (in percent) for each categorical factor in the 10-factor analysis. The share refers to the cumulative posterior probability for the given version of the categorical factor divided by the sum of cumulative probabilities for both versions of the factor. For each categorical factor, the version with the highest probability is shown. The remaining *PROF*, *INV*, *SIZE*, and *VAL* posterior probability goes to *RMW*, *CMA*, *ME*, and *HML*, respectively. The sample period is January 1972 to December 2015. The prior is set so that Sh_{max} = prior multiple $\times$ $Sh(Mkt)$, where Sh_{max} is the square root of the expectation of the maximum squared Sharpe ratio with six factors included, taken with respect to the prior under the alternative that the nonmarket factor alphas are nonzero.

Factor/Prior Multiple	1.25	1.5	2	3
ROE	96.1	99.4	99.9	100
IA	71.3	72.0	73.6	74.9
SMB	68.6	73.4	77.6	80.2
HML^m	94.9	99.5	99.9	100

that the alpha is not zero and that *SMB* is not redundant. Thus, the probability of the categorical model that includes (excludes) *Size* increases (decreases).

As noted above, the more timely *ROE* and HML^m factors account for most of the cumulative probability of the profitability and value categories. Table III shows that timely profitability and value remain responsible for the lion's share of the cumulative probability across the different priors, especially at higher Sharpe multiples. Results for *IA* and *SMB* are likewise not very sensitive to varying the prior.

C. Are Value and Momentum Redundant?

As demonstrated in Section III, all that matters when comparing two asset pricing models is the extent to which each model prices the factors in the other model. Hou, Xue, and Zhang (2015b) and Fama and French (2015) regress *HML* on models that exclude value and cannot reject the hypothesis that *HML*'s alpha is zero, thus concluding that HML is redundant. In addition, HXZ show that their model renders the momentum factor, *UMD*, redundant. On the other hand, our results above show that the model {*Mkt SMB ROE IA UMD* HML^m}, which receives highest posterior probability, contains both value and momentum factors.

To shed further light on this finding, Table IV reports the annualized intercept estimates for each factor in the top model when it is regressed on the other five factors. We observe that the intercepts for HML^m and *UMD* are large and statistically significant, rejecting the hypothesis of redundancy by conventional standards. HML^m has an alpha of 5.59% (t-statistic = 5.10) and *UMD* has an alpha of 6.44% (t-statistic = 4.05). When we regress the standard value factor, *HML*, on the non-value factors {*Mkt SMB ROE IA UMD*} in our top model, we find, as in the earlier studies, that it *is* redundant: the intercept is 0.95% with

Table IV

Intercepts for Each Factor in the Six-Factor Model on the Other Factors

The table presents annualized alphas from regressions of each factor on the other five factors in the model {*Mkt SMB ROE IA UMD* HML^m}. This model has the highest posterior probability in the 10-factor analysis. The sample period is from January 1972 to December 2013. *t*-statistics are reported in parentheses.

Factor	*SMB*	*ROE*	*IA*	*UMD*	HML^m
Alpha	4.70	6.51	1.17	6.44	5.59
(*t*-statistic)	(3.01)	(5.90)	(1.51)	(4.05)	(5.05)

a *t*-statistic of 0.81. The different results for the two value factors are driven largely by the fact that HML^m is strongly negatively correlated (–0.65) with *UMD*, whereas the correlation is only –0.17 for *HML*.[30] The negative loading for HML^m when *UMD* is included lowers the model expected return and raises the HML^m alpha, so that this timely value factor is not redundant.

We now evaluate the hypothesis that HML^m is redundant from a Bayesian perspective. Figure 3 shows results for the Bayesian intercept test on the other factors. As discussed earlier, the prior under the alternative follows a normal distribution with zero mean and standard deviation σ_α. The larger the value of σ_α, the higher the increase in the Sharpe ratio that one can expect to achieve by adding a position in HML^m to investment in the other factors. The horizontal axis in each panel of the figure gives the prior multiple. This is the Sh_{max} for the alternative, expressed as a multiple of the Sharpe ratio for the five factors in the null model that excludes HML^m.

The left panel of the figure gives the posterior probability for the null model. It quickly decreases to zero as the prior Sharpe multiple under the alternative increases, inconsistent with the view that HML^m is redundant. The right panel of Figure 3 provides information about the implied value of σ_α. This gives an idea of the likely magnitude of the α's envisioned under the alternative and should be helpful in identifying the range of prior multiples that one finds reasonable.[31] For example, to get a 25% increase in the Sharpe ratio from the already-high sample level of 0.42 for the null model, we would need a very large σ_α of about 7.5% per year.

The Bayesian analysis for *UMD* on the five other factors in the best model (not shown) looks much the same as Figure 3, with the probability for redundancy close to zero. To highlight the role of HML^m in this finding, we exclude that factor and examine the *UMD* alpha on the remaining factors, {*Mkt SMB IA ROE*}. In Figure 4, we see that the posterior probability for the null hypothesis of redundancy (*UMD* alpha is zero) is always above 50%, with values

[30] Asness and Frazzini (2013) argue that the use of less timely price information in HML "reduces the natural negative correlation of value and momentum."

[31] In general, the plot of σ_α is based on the average residual variance estimate for the left-hand-side assets.

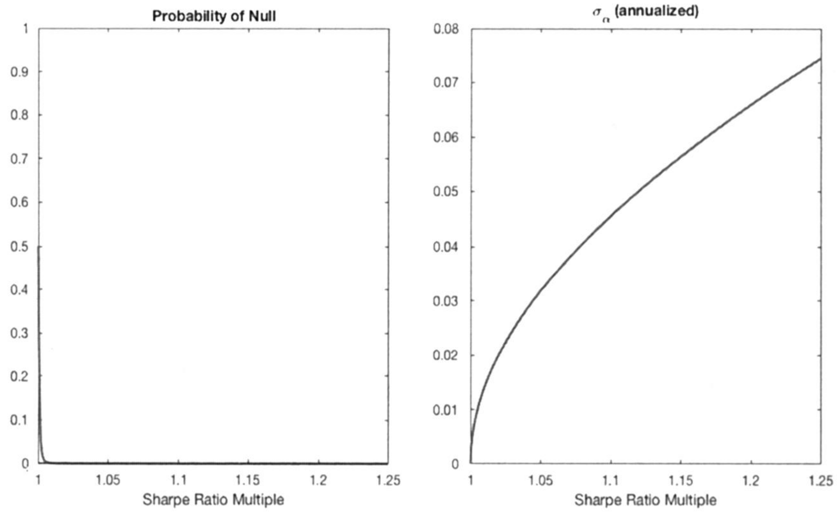

Figure 3. *HML^m* is not redundant in relation to the other factors {*Mkt SMB ROE IA UMD*} in the best model. A Bayesian intercept test is conducted for *HML^m* over the sample period January 1972 to December 2015. The left plot gives the posterior probability for the null that the *HML^m* alpha is zero as a function of the Sharpe ratio multiple. This is a multiple of the Sharpe ratio for the five factors (0.42 or 3.97 times the market Sharpe ratio) and gives the square root of the expectation of the maximum squared Sharpe ratio, taken with respect to the prior under the alternative that the *HML^m* alpha is nonzero. The plot on the right gives the implied standard deviation of the prior for alpha as a function of the Sharpe ratio multiple. (Color figure can be viewed at wileyonlinelibrary.com)

over 80% for Sharpe ratio multiples around 1.2. The conventional p-value exceeds 35% here, as indicated by the horizontal line in the figure. Thus, the evidence supports the view that *UMD is* redundant when *HML^m* is not in the model.

VI. Out-of-Sample Model Performance

Having completed our Bayesian comparison of models, we now examine descriptive evidence on the performance of models ranked highly by the Bayesian procedure. We present results for our top three models as well as the four prominent models considered earlier—CAPM, FF3, FF5, and HXZ. Keep in mind, however, that the model probabilities shown in Table I never exceed 50% for the priors considered and are generally much lower. The first column of Table V gives Sharpe ratios for each of these models estimated over the full sample, 1972 to 2015, denoted T. The top three models over this period all have ratios around 0.48, which is higher than those for the four benchmark models. The HXZ model comes closest, with a ratio of 0.38.

These results show that our procedure tends to "pick" models that have performed well on the data and thus have relatively high in-sample Sharpe ratios.

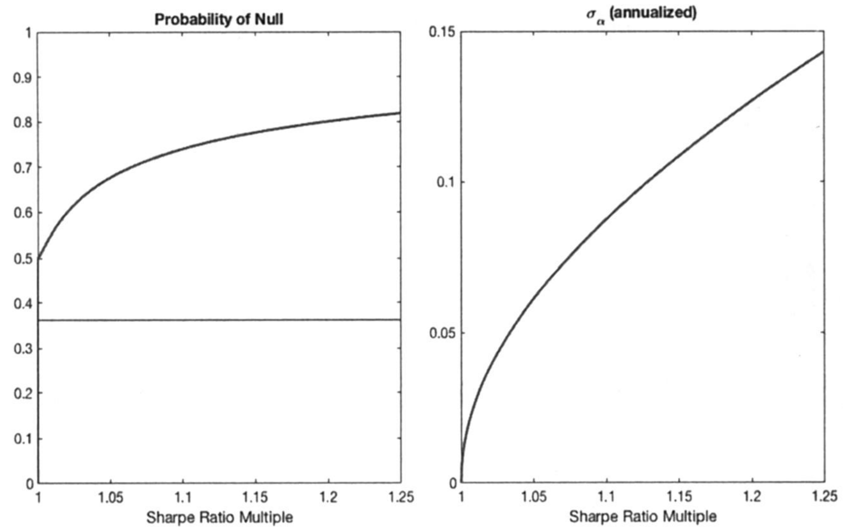

Figure 4. ***UMD* is redundant when *HML*^{m} is left out of the best model.** A Bayesian intercept test is conducted for *UMD* on the model {*Mkt SMB ROE IA*} over the sample period January 1972 to December 2015. The left plot gives the posterior probability for the null that the *UMD* alpha is zero as a function of the Sharpe ratio multiple. This is a multiple of the Sharpe ratio for the four factors (0.42 or 3.95 times the market Sharpe ratio) and gives the square root of the expectation of the maximum squared Sharpe ratio, taken with respect to the prior under the alternative that the *UMD* alpha is nonzero. The plot on the right gives the implied standard deviation of the prior for alpha as a function of the Sharpe ratio multiple. (Color figure can be viewed at wileyonlinelibrary.com)

Bayes's law is used to derive rational assessments of the model probabilities in light of the data and our priors. However, it is likely that there is an upward selection bias in the sample ratios of the models identified as best. It is therefore of interest to evaluate the out-of-sample (OOS) performance of the models. This requires that we specify an estimation period for the model comparison; the remaining data will then be used to evaluate the top models. If the estimation period is too short, the procedure will not have much power to identify good models. Similarly, if the performance period is too limited, the measures of performance may be noisy. With this trade-off in mind, we choose estimation periods corresponding to one-half (denoted $T/2$) and two-thirds (denoted $2T/3$) of the monthly data. Of course, our earlier Bayesian comparison conditions on *all* of the data, so analysis based on estimation over a subset may produce different model rankings. Nonetheless, we evaluate OOS performance for whichever models are ranked highly based on the earlier data.

We begin by applying our procedure over the estimation period. The top three models are identified and Sharpe ratios for each are then calculated over both the estimation and performance periods. The results are presented in the columns labeled *EST* and *PERF*, respectively, along with ratios for the benchmark models calculated over the same periods. Performance is calculated two ways. The first method estimates the tangency-portfolio weights and

Table V
Out-of-Sample Model Performance

The table presents in- and out-of-sample results on the performance of the three top-ranked models using our Bayesian procedure with prior Sharpe multiple 1.5, as well as four prominent models: CAPM, FF3, FF5, and HXZ. The first column gives the in-sample Sharpe ratio over the full sample, 1972 to 2015, denoted *T*. Columns labeled EST report the in-sample Sharpe ratio for the estimation periods that, correspond to half (*T*/2) and two-thirds (2*T*/3) of the sample. Columns labeled *PERF* and *PERFw* report the out-of-sample Sharpe ratios. PERF uses the out-of-sample period to compute the tangency-portfolio weights while PERFw uses the estimation-period returns to determine the tangency-portfolio weights and applies those weights to the out-of-sample returns. The three best models always include *Mkt*, *ROE*, HML^m, and *UMD*. Those based on the full sample also include, respectively (*SMB IA*), (*SMB CMA*), and (*ME IA*). Those based on half of the sample (*T*/2) include (*IA*), (*SMB IA*), and *CMA*, and the first two-thirds of the sample (2*T*/3) include (*SMB IA*), *IA*, and (*ME IA*).

	T	*T*/2			2*T*/3		
Model	Sample SR	*EST*	*PERF*	*PERFw*	*EST*	*PERF*	*PERFw*
1	0.488	0.778	0.345	0.235	0.596	0.312	0.276
2	0.486	0.815	0.370	0.248	0.567	0.284	0.256
3	0.483	0.779	0.350	0.237	0.592	0.298	0.269
CAPM	0.134	0.094	0.138	0.138	0.118	0.109	0.109
FF3	0.169	0.277	0.165	0.129	0.257	0.147	0.122
FF5	0.356	0.487	0.372	0.202	0.354	0.367	0.263
HXZ	0.373	0.613	0.359	0.250	0.489	0.288	0.271

the factor-return moments simultaneously in calculating sample Sharpe ratios for a model from the OOS data. The second method calculates the weights from the estimation-period returns and applies them to the returns in the performance period. Sharpe ratios are then computed from the returns on this implementable portfolio strategy. A "*w*" after *PERF* indicates that the weights are derived in this manner.

First, we summarize the *PERF* results with simultaneous estimation. Using one-half of the data (*T*/2), the top three models have estimation-period Sharpe ratios close to 0.8. The HXZ model follows at 0.61, then FF5 at 0.49, with FF3 and CAPM trailing at lower levels. The OOS performance-period ratio for the second-ranked model and the ratio for FF5 in that period are much lower, at about 0.37, with HXZ and the first/third-ranked models following close behind. All of these models easily beat CAPM and FF3. With the longer estimation period (2*T*/3), FF5 has the highest performance-period Sharpe ratio, 0.37, followed by our top model at 0.31. Taking estimation error into account, however, this difference is not statistically significant.[32] Now, the second/third models and HXZ follow the top-ranked model, again with substantially higher ratios than those for CAPM and FF3.

Next, we summarize the *PERFw* results with estimation-period weights. Naturally, with tangency weights estimated on separate data, the OOS ratios

[32] The asymptotic p-value based on Barillas et al. (2017) is 0.39.

are now lower, all below 0.30. In the $T/2$ scenario, our top three models and HXZ now perform similarly, followed by FF5. With two-thirds of the data used in the estimation period, our top model just barely edges out HXZ, FF5, and the other top models.

In Section V.B.2 we examine the prior sensitivity of our model comparisons to different assumptions about the multiple of $Sh(Mkt)$ that yields Sh_{max} when five factors are included along with *Mkt*. In particular, we consider priors with multiples of 1.25, 1.5 (our baseline prior), 2, and 3. A multiple of 1.25 corresponds to a relatively modest 25% increase obtained by exploiting fairly small alphas, whereas a multiple of 3 yields a more striking 200% increase in the Sharpe ratio and much larger alphas—a more behavioral prior. We find that there is considerable persistence in top-model rankings across priors, with occasional changes. The OOS model performance using each of these priors is similar to that in Table V.

With a Sharpe multiple of 1.25, the performance of the top three models is still very good, but slightly worse (by about 0.02) for the best model. This makes sense in that the 1.25 prior does not match up well with the substantial magnitude of the anomalies observed in the data. On the other hand, the larger prior multiples move in the direction of the ex post Sharpe multiple of 4.5 for the best model in Table II. Accordingly, we see slight improvements in the OOS performance of the top three models with these priors. For example, in the $T/2$ analysis with the multiple 3, the two highest-probability models are now virtually tied with FF5 in *PERF* and with HXZ in *PERFw* for the best OOS performance.

To summarize, the OOS results support the conclusion that models assigned relatively high probabilities by our Bayesian procedure tend to perform quite well out of sample. But HXZ and FF5 perform well too.

VII. Absolute Tests of the Best Model

In Section V we see that, over the sample period 1972 to 2015, the model with the highest posterior probability is the six-factor model {*Mkt IA ROE SMB* HML^m *UMD*}. Now we evaluate the overall performance of this model from the absolute perspective. Although a wide variety of test-asset portfolios have been examined, we present results for two representative sets that serve to illustrate some interesting findings. The first set of portfolios is based on independent stock sorts by size and momentum, whereas the second set is constructed by sorting stocks on book-to-market and investment.

The Bayesian test results for the six-factor model with the size/momentum portfolios are given in Figure 5. Similar to the redundancy tests, the horizontal axis in the figure shows the multiple of the Sharpe ratio for the factors in the null model, now the six-factor model. This is the multiple under the alternative that the left-hand-side assets are not priced by the model. The average absolute alpha is 1.65% per annum, while the GRS statistic is 3.32 with a p-value of nearly zero (1.9E−7), which strongly rejects the model in a classical sense. The Bayesian test also provides strong evidence against the null hypothesis. The

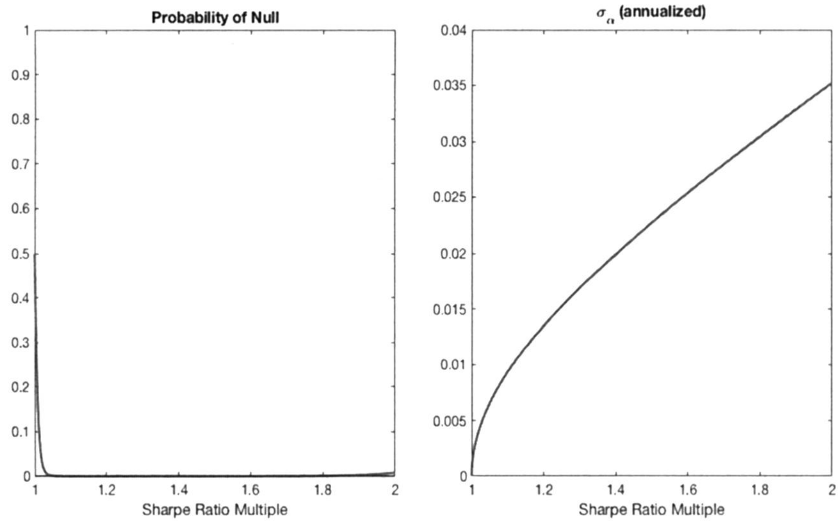

Figure 5. Absolute test of the model {*Mkt SMB IA ROE UMD HML^m*}. The test is conducted over the sample period January 1972 to December 2015. The test assets are 25 size/momentum portfolios. The left plot gives the posterior probability for the null that the test-asset alphas are zero as a function of the Sharpe ratio multiple. This is a multiple of the Sharpe ratio for the six factors (0.49 or 4.6 times the market Sharpe ratio) and gives the square root of the expectation of the maximum squared Sharpe ratio, taken with respect to the prior under the alternative that the alphas are nonzero. The plot on the right gives the implied standard deviation of the prior for alpha as a function of the Sharpe ratio multiple. (Color figure can be viewed at wileyonlinelibrary.com)

probability of the null quickly declines to an extended zero-probability range for the Sharpe multiples shown (up to two).

Now we turn to the results for the 25 portfolios formed on sorts by book-to-market and investment. The average absolute six-factor alpha for these test assets is 2.57%, which is much larger than the 1.65% for the size/momentum portfolios. Nonetheless, the GRS statistic is lower, 3.21, due to the larger residual variation in returns for the book-to-market/investment portfolios. The resulting p-value is still nearly zero (4.5E$-$7), however. Figure 6 plots the Bayesian test results. The message is similar to that in Figure 5, though the probability of the model (blue line) now turns up a bit more for Sharpe multiples less than 2.

An additional observation about the Bayesian analysis deserves emphasis. The probability for the model in Figure 6 not only rebounds from zero, but actually approaches one (not shown) as the Sharpe multiple and prior standard deviation for alpha increase to infinity.[33] This is an example of Bartlett's paradox, mentioned earlier. Roughly speaking, although the alpha estimates may deviate substantially from the null value of zero, they can be even further from the values of alpha envisioned under the alternative when the Sharpe multiple

[33] This is true in extensions of the probability plots in Figures 3 to 5 as well.

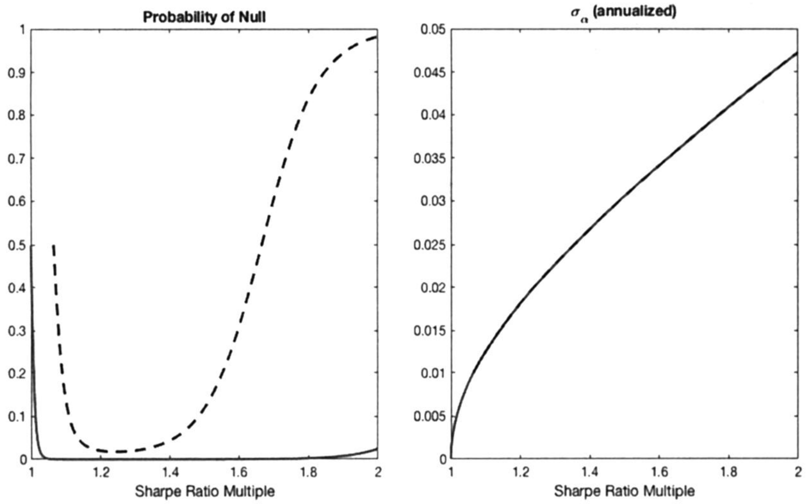

Figure 6. Absolute test of the model {*Mkt SMB IA ROE UMD HML^m*}. The test is conducted over the sample period January 1972 to December 2015. The test assets are 25 book-to-market/investment portfolios. The left plot gives the posterior probability for the null that the test-asset alphas are zero (blue curve) as a function of the Sharpe ratio multiple. This is a multiple of the Sharpe ratio for the six factors (0.49 or 4.2 times the market Sharpe ratio) and gives the square root of the expectation of the maximum squared Sharpe ratio, taken with respect to the prior under the alternative that the alphas are nonzero. The plot on the right gives the implied standard deviation of the prior for alpha as a function of the Sharpe ratio multiple. Results for the null that the alphas are approximately zero (annualized $\sigma_{\alpha 0} = 1\%$) are given by the black dashed curves. (Color figure can be viewed at wileyonlinelibrary.com)

is very large. As a result, the posterior probability favors the restricted model in such a case. Thus, in evaluating pricing hypotheses of this sort, it is essential to form a "reasonable" a priori judgment about the magnitude of plausible alphas (reflected in the choice of the parameter k).

A. Approximate-Model Results

We have seen that, for a range of prior Sharpe multiples that might be described as modest to quite large (given the already large Sharpe ratio of the six-factor model), the evidence in Figure 6 favors the statistical alternative with nonzero alphas over the sharp null hypothesis that test-asset alphas are zero for the model {*Mkt SMB IA ROE UMD HML^m*}. But this pricing model may nonetheless provide a good *approximation* to the data. After all, one might argue that models, by their nature, always leave out some features of reality and so cannot plausibly be expected to hold *exactly*. We address this issue in the Bayesian framework by modifying the prior under the null to accommodate relatively small deviations from the exact model specification. The modified *BF* formula was given earlier in Proposition 1.

The black dashed line in Figure 6 displays the results of an analysis in which the prior under the *null* now assumes that $\sigma_{\alpha 0} = 1\%$ (annualized). This allows for book-to-market/investment portfolio deviations from the exact version of the six-factor model, with an expected value of about 0.8% in magnitude. Such deviations imply a higher Sharpe ratio under the approximate null hypothesis, about 6% larger than that for the exact null. Thus, whereas the starting point for the null probability curve was previously at a Sharpe multiple of 1.0, the black dashed line now starts at a multiple of 1.06. As expected, the posterior probabilities in Figure 6 for the approximate model are higher than the probabilities obtained earlier for the exact model. We no longer have a "zero-probability range" for the approximate null and the even-odds breakpoint is now a multiple of about 1.6. Moreover, adapting the classical emphasis on confidence levels to the Bayesian setting, 95% confidence that the approximate model is false obtains only for prior multiples between (roughly) 1.15 and 1.45.

This extension to approximate models should enhance the appeal of the Bayesian approach for evaluating absolute model performance. It allows for more subtle and informative inferences in situations in which the sample size is large and models are routinely rejected at conventional levels by the GRS test, as in the application above or, for example, in Fama and French (2016).

VIII. Conclusion

In this paper, we derive a Bayesian asset pricing test that requires a prior judgment about the magnitude of plausible model deviations or "alphas" and is easily calculated from the GRS F-statistic. Given a set of candidate traded factors, we develop a related procedure that permits a Bayesian comparison of models, that is, the computation of model probabilities for the collection of all possible pricing models based on subsets of the given factors. The ability to simultaneously compare many models is an attractive aspect of the Bayesian approach.

Although our work is in the tradition of the literature on asset pricing tests, Bayesian analysis has also been used to address other kinds of questions in finance. For example, Pastor and Stambaugh (2000, p. 336) are interested in comparing models too, but from a different perspective. As they note, the objective of their study "is not to choose one pricing model over another." Rather, they examine the extent to which investors' prior beliefs about alternative pricing models (one based on stock characteristics and another on a stock's factor betas) impact the utility derived from the implied portfolio choices. Such utility-based metrics are undoubtedly important, but complementary to our focus on *inference* about models in this paper.

While we analyze the "classic" statistical specification with returns that are independent and identically normally distributed over time (conditional on the market), extensions to accommodate time-variation in parameters and conditional heteroskedasticity of returns would be desirable. We examine the factors in Asness and Frazzini (2013), Hou, Xue, and Zhang (2015a, 2015b), and Fama and French (2015, 2016) in our empirical analysis, but other factors related to short and long reversals, the levels of beta and idiosyncratic volatility, and various measures of liquidity could be considered in future work as well. Finally, the influence of data mining in the selection of candidate factors should be explored (see related work by Harvey, Liu, and Zhu (2016)).

Initial submission: November 23, 2015; Accepted: February 15, 2017
Editors: Bruno Biais, Michael R. Roberts, and Kenneth J. Singleton

Appendix A: Marginal Likelihoods and the Bayes Factor

In this section we provide the expression for the Bayes factor for $H_0 : \alpha = 0$ versus $H_1 : \alpha \neq 0$ in (12), which is equal to the ratio of marginal likelihoods of the restricted to the unrestricted models. The Internet Appendix derives the expressions for both marginal likelihoods.

ML for the restricted model, $\alpha = 0$:

$$P(R|F) = \left(\frac{1}{\pi}\right)^{\frac{N(T-K)}{2}} \Gamma_N\left(\frac{T-K}{2}\right) |F'F|^{-\frac{N}{2}} |S^R|^{-\frac{T-K}{2}}. \tag{A1}$$

ML for the unrestricted model, $\alpha \neq 0$:

$$P(R|F) = \left(\frac{1}{\pi}\right)^{\frac{N(T-K)}{2}} \Gamma_N\left(\frac{T-K}{2}\right) |F'F|^{-\frac{N}{2}} |S|^{-\frac{T-K}{2}} Q, \tag{A2}$$

where Q is defined as

$$Q = \iint \exp\left(-\frac{1}{2a}(\alpha - \hat{\alpha})'\Sigma^{-1}(\alpha - \hat{\alpha})\right) P(\alpha|\Sigma)P(\Sigma|F, R)d\alpha d\Sigma. \tag{A3}$$

Note that Q is always less than one since the exponent in (11) is uniformly negative. Also, a large Q indicates a relatively small "distance" between the alpha estimate and the values of alpha anticipated under the prior for the alternative model. With expressions for the marginal likelihoods of the restricted and unrestricted models at hand, we can now write the Bayes factor for $H_0 : \alpha = 0$ versus $H_1 : \alpha \neq 0$ as

$$BF = \frac{1}{Q}\left(\frac{|S|}{|S^R|}\right)^{(T-K)/2}. \tag{A4}$$

This is a multifactor version of the Harvey and Zhou (1990) Bayes factor. In Appendix B we show how to integrate α and Σ in Q analytically.

Appendix B: Derivation of the Formulas for Q and BF

If a random N-vector Y is distributed as multivariate normal $MVN(\mu, V)$ with V invertible, then the quadratic form $Y'V^{-1}Y$ has a noncentral chi-squared distribution with degrees of freedom N and noncentrality parameter $\mu' V^{-1}\mu$.[34] We apply this result to the shifted random vector α and its prior distribution. Let $x = (\alpha - \hat{\alpha})'(k\Sigma)^{-1}(\alpha - \hat{\alpha})$. We treat $\hat{\alpha}$ as a fixed vector in this analysis. The inner integral in the definition of Q in (A3) can be viewed as the expectation of $\exp(tx)$, a function of the random variable α with $t = -k/(2a)$. Given the prior $P(\alpha|\Sigma) = MVN(0, k\Sigma)$, $\alpha - \hat{\alpha}$ is distributed as $MVN(-\hat{\alpha}, k\Sigma)$ and it follows that the conditional distribution of x given Σ is noncentral chi-square with N degrees of freedom and noncentrality parameter $\lambda_\alpha = k^{-1}\hat{\alpha}'\Sigma^{-1}\hat{\alpha}$. Thus, the inner integral is equal to the moment-generating function of this noncentral chi-square evaluated at t:[35]

$$\frac{\exp\left(\frac{\lambda_\alpha t}{1-2t}\right)}{(1-2t)^{N/2}} = \frac{\exp\left(\frac{-\hat{\alpha}'\Sigma^{-1}\hat{\alpha}/(2a)}{1+k/a}\right)}{(1+k/a)^{N/2}} = \frac{\exp\left(\frac{-\hat{\alpha}'\Sigma^{-1}\hat{\alpha}}{2(a+k)}\right)}{(1+k/a)^{N/2}}. \tag{B1}$$

Next, we need to evaluate the integral of the product of (B1) and the posterior density for Σ. Given the distributional assumptions and the prior in Section I, the posterior distribution of Σ is inverted Wishart, $W^{-1}(S, T-K)$, and so the posterior for Σ^{-1} is $W(S^{-1}, T-K)$. Therefore, by the result on p. 535 of Rao (1973), $\hat{\alpha}'\Sigma^{-1}\hat{\alpha}$ is distributed as $\hat{\alpha}'S^{-1}\hat{\alpha}$ times a chi-square variable with $T-K$ degrees of freedom. Thus, the desired integral is $(1+k/a)^{-N/2}$ times the expectation of $\exp(tx)$, where x now refers to the chi-square variable and $t = -\hat{\alpha}'S^{-1}\hat{\alpha}/[2(a+k)]$. Hence, we need to evaluate the moment-generating function of the (central) chi-square:[36]

$$(1-2t)^{-(T-K)/2} = \left(1 + \hat{\alpha}'S^{-1}\hat{\alpha}/(a+k)\right)^{-(T-K)/2}.$$

Since $W = \hat{\alpha}'(S/T)^{-1}\hat{\alpha}/a$ by (3), if follows that

$$Q = \left(1 + \frac{a}{(a+k)}(W/T)\right)^{-(T-K)/2}\left(1+\frac{k}{a}\right)^{-N/2}. \tag{B2}$$

Proof of (13) when $k_0 = 0$: It remains to evaluate the ratio of determinants, $|S|/|S_R|$, in (A4). It is a standard result in the multivariate literature that the likelihood ratio test statistic is $LR = T\ln(|S_R|/|S|)$. Gibbons, Ross, and

[34] See https://en.wikipedia.org/wiki/Noncentral_chi-squared_distribution, for example, when the covariance matrix is the identity matrix. Given V, the logic of generalized least squares yields an invertible $N \times N$ matrix P such that $P'P = V^{-1}$ and thus $PVP' = I$. Therefore, the N-vector PY is distributed as $MVN(P\mu, I)$ and so $(PY)'(PY) = Y'P'PY = Y'V^{-1}Y$ has a noncentral chi-squared distribution with noncentrality parameter $(P\mu)'(P\mu) = \mu'V^{-1}\mu$.

[35] See http://en.wikipedia.org/wiki/Noncentral_chi-squared_distribution#Moment_generating_function.

[36] See http://en.wikipedia.org/wiki/Chi-squared_distribution.

Shanken (1989) further note that $LR = T\ln(1 + W/T)$, so $\exp(LR/T) = 1 + W/T = |S_R|/|S|$.[37] Therefore,

$$(|S|/|S_R|)^{(T-K)/2} = (1 + W/T)^{-(T-K)/2} = Q_0 \tag{B3}$$

and (13) follows from (A4), (B2), and (B3).

Allowing for $k_0 \neq 0$: Let BF and BF_{k_0} be the Bayes factors for comparing the prior value zero with the values k and k_0, respectively. The Bayes factor for the approximate null, $BF_{k_0,k}$, is the ratio of marginal likelihoods based on the prior values k_0 and k. This is $(1/BF_{k_0})/(1/BF) = BF/BF_{k_0}$, where the marginal likelihood for the exact null cancels out since it is in the numerators of both BF and BF_{k_0}. By (A4), this is Q_{k_0}/Q, as given in equation (13).

Appendix C: Aggregation over Different Versions of the Categorical Factors

Assume that the categorical models consist of up to K factors, K_C of which are categorical factors. In the example of Section IV, with the standard factors *Mkt* and *HML* and the categorical factor *Size*, $K = 3$ and $K_C = 1$. There are two versions of the factors, $w_1 =$ (*Mkt HML SMB*) and $w_2 =$ (*Mkt HML ME*). In general, there are 2^{K_c} versions of the K factors, each with prior probability $1/2^{K_c}$. The 2^{K-1} models associated with a given version are assigned uniform conditional prior probabilities of $1/2^{K-1}$. Let M be a version of a categorical model M_C with L_C categorical factors. The number of factor versions that include model M is $2^{K_c-L_c}$ (L_C slots are taken), so the fraction of versions that include M is $2^{K_c-L_c}/2^{K_c} = 1/2^{L_c}$. Thus, the total probability of $1/2^{K-1}$ for M_C is split evenly between the 2^{L_c} versions of that categorical model, of which M is one. In the example, $L_C = 0$ for {*Mkt HML*} and $L_C = 1$ for {*Mkt ME*}. With $K = 3$, the prior probability for the first model is $1/2^{K-1} = 1/2^2 = 1/4$, while the probability of the second is 1/2 ($1/2^{L_c}$) of that, or 1/8.

As mentioned earlier, it is essential that the marginal likelihoods for the various models be based on the same data. This requires a simple extension of (19). Take the model {*Mkt ME*}, for example. In the present context, the excluded factors consist of the noncategorical factor *HML* and the other version of the categorical *size* factor, *SMB*. We write the marginal likelihood as the unrestricted marginal likelihood for *ME* given *Mkt*, times the restricted marginal likelihood for *HML* given *Mkt* and *ME*, times the restricted marginal likelihood for *SMB* given all three factors in the version $w_2 =$ (*Mkt HML ME*). Thus, the extra marginal likelihood term at the end is for the second version of the size factor, which is treated like a test asset conditional on w_2. Similarly, for w_1, the extra marginal likelihood term at the end is for *ME* given (*Mkt HML SMB*).

In general, given a K-factor version w, let w^* denote the K_C alternate versions of the categorical factors ($w_2{}^*$ would consist of *SMB* above). Hence, there are

[37] This also follows from Stewart (1995) by formulating the zero-alpha restriction in terms of his equation (7) with $q = 1$. Let C be a $1 \times K$ vector with one in the (1,1) position and zeroes elsewhere, M an $N \times N$ identity matrix, and D a $1 \times N$ zero vector.

$K + K_C$ factors in all ($3 + 1 = 4$ in the example). The term F now refers to all of these factor data. For a model M with nonmarket factors f, all contained in w, let f^* now denote the factors in w that are excluded from M.[38] Let the prior for the parameters in the regression of w^* on (Mkt, f, f^*) take the usual form based on (5) and (6), again independent of the other components. The test assets r are now regressed on (Mkt, f, f^*, w^*), that is, *all* of the factors, and so the corresponding marginal likelihood will cancel out in all probability calculations, as earlier. This term will be ignored going forward. Therefore, the counterpart of (19) is now

$$ML(M|w) = ML_U\,(f|Mkt) \times ML_R\left(f^*|Mkt,\, f\right) \times ML_R\left(w^*|Mkt,\, f,\, f^*\right). \quad \text{(C1)}$$

For each w, the conditional posterior probabilities can then be obtained as in (17). Note that the marginal likelihood term corresponding to w^* will drop out of these computations that are conditional on w. The w^* term will differ across the various w's, however, and so will affect the posterior probabilities for the w's, as seen in Proposition C1.

PROPOSITION C1: *The unconditional (not conditional on w) posterior model probabilities are obtained as follows. First, calculate ML(M | w) in* (C1) *for each version w of the factors and each model in w. Then calculate the marginal likelihood for each as follows:*

$$ML(w) = E_{M|w}\{ML(M|w)\}, \quad \text{(C2)}$$

where $M\,|\,w$ refers to the uniform prior over the models in w. Next, calculate the unconditional probability of the data,

$$P(F) = E_w\left\{ML(w)\right\}, \quad \text{(C3)}$$

by averaging $ML(w)$ over the uniform prior $P(w)$. The posterior probability for each w is then

$$P(w|F) = ML(w)\,P(w)\,/P(F). \quad \text{(C4)}$$

Finally, the unconditional posterior probability for M is

$$P(M\,|F) = E_{w|F}\left\{P\left(M|w, F\right)\right\}, \quad \text{(C5)}$$

where $P(M\,|\,w, F)$ is the conditional posterior probability for M given factor version w and the expectation is taken with respect to the versions posterior $P(w\,|\,F)$.

PROOF: A general principle that we use repeatedly is $P(Y) = E_X\{P(Y\,|\,X)\}$. Also, by definition, $ML(M\,|\,w) = P(F\,|\,M, w)$ and $ML(w) = P(F\,|\,w)$. In (C2), M plays the role of X and F the role of Y, while we condition on w throughout in the "background." In (C3), w plays the role of X and F the role of Y. Equation (C4)

[38] The term f^* will differ across the different w's that include the factors f (a w subscript is implicit).

is just Bayes's theorem. Finally, w plays the role of X and M the role of Y in (C5), while we condition on F throughout in the background.

Allowing Each Factor to Be Included or Not in the Basic Methodology

This generalization of the basic methodology for a single version of the factors is based on Proposition C1, with a few modifications. Now, instead of the "hyperparameter" w denoting a version of the factors, it refers to the factor that plays the anchor role that *Mkt* plays in (18), (19), and (C1) above. Since the *entire* analysis is no longer conditioned on *Mkt* or any other factor, the density of the first factor must now be modeled. Thus, the counterpart of (C1) is

$$ML(M|w) = ML_U(w) \times ML_U(f|w) \times ML_R\left(f^*|w, f\right), \tag{C6}$$

where $ML_U(w)$ is the marginal likelihood for factor w assuming it is normally distributed independently over time, with unknown mean and variance, both of which have standard diffuse priors. Now f denotes the factors other than w that are in M, while f^* consists of the factors that are excluded from M. One approach is to let the prior parameter k be specified for each anchor in such a way that the corresponding Sh_{max} is the same for all anchors and equal to the multiple of the sample *Mkt* Sharpe ratio used in the main analysis. See the discussion below (19). This introduces an "empirical Bayesian" element that is common in other Bayesian asset pricing papers. The aggregation of conditional probabilities then proceeds exactly as in (C2) to (C5).

REFERENCES

Asness, Cliff, and Andrea Frazzini, 2013, The devil in HML's details, *Journal of Portfolio Management* 39, 49–68.

Avramov, Doron, 2002, Stock return predictability and model uncertainty, *Journal of Financial Economics* 64, 423–458.

Avramov, Doron, and John Chao, 2006, An exact Bayes test of asset pricing models with application to international markets, *Journal of Business* 79, 293–323.

Barillas, Francisco, Raymond Kan, Cesare Robotti, and Jay Shanken, 2017, Model comparison with Sharpe ratios, Working paper, Goizueta Business School, Emory University.

Barillas, Francisco, and Jay Shanken, 2017, Which alpha? *Review of Financial Studies* 30, 1316–1338.

Bartlett, Maurice, 1957, A comment on D. V. Lindley's statistical paradox, *Biometrika* 44, 533–534.

Berger James, and Luis Pericchi, 1996, The intrinsic Bayes factor for model selection and prediction, *Journal of the American Statistical Association* 91, 109–122.

Berger James, and Luis Pericchi, 2001, Objective Bayesian methods for model selection: Introduction and comparison, *IMS Lecture Notes: Monograph Series* 38, 137–193.

Black, Fisher, Michael C. Jensen, and Myron Scholes, 1972, The capital asset pricing model: Some empirical tests, in Michael C. Jensen, ed.: *Studies in the Theory of Capital Markets* (Praeger, New York).

Breeden, Douglas, 1979, An intertemporal asset pricing model with stochastic consumption and investment opportunities, *Journal of Financial Economics* 7, 265–296.

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

Chen, Nai-Fu, Richard Roll, and Stephen A. Ross, 1986, Economic forces and the stock market, *Journal of Business* 59, 383–403.

Cochrane, John H., and Jesus Saa-Requejo, 2000, Beyond arbitrage: Good-deal asset price bounds in incomplete markets, *Journal of Political Economy* 108, 79–119.

De Moor, Lieven, Geert Dhaene, and Piet Sercu, 2015, On comparing zero-alpha tests across multifactor asset pricing models, *Journal of Banking & Finance* 61, 235–240.

Dybvig, Phillip H., 1983, An explicit bound on individual assets' deviations form APT pricing in a finite economy, *Journal of Financial Economics* 12, 483–496.

Fama, Eugene F., 1996, Multifactor portfolio efficiency and multifactor asset pricing, *Journal of Financial and Quantitative Analysis* 31, 441–465.

Fama, Eugene F., 1998, Determining the number of priced state variables in the ICAPM, *Journal of Financial and Quantitative Analysis* 33, 217–231.

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

Fama, Eugene F., and Kenneth R. French, 1996, Multifactor explanations of asset pricing anomalies, *Journal of Finance* 51, 55–84.

Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

Fama, Eugene F., and Kenneth R. French, 2016, Dissecting anomalies with a five-factor model, *Review of Financial Studies* 29, 69–103.

Gibbons, Michael, Stephen Ross, and Jay Shanken, 1989, A test of the efficiency of a given portfolio, *Econometrica* 57, 1121–1152.

Grinblatt, Mark, and Sheridan Titman, 1983, Factor pricing in a finite economy, *Journal of Financial Economics* 12, 497–507.

Hansen, Peter, Asger Lunde, and James Nason, 2011, The model confidence set, *Econometrica* 79, 453–497.

Harvey, Campbell R., and Guofu Zhou, 1990, Bayesian inference in asset pricing tests, *Journal of Financial Economics* 26, 221–254.

Harvey, Campbell, Yan Liu, and Heqing Zhu, 2016, ... and the cross-section of expected returns, *Review of Financial Studies* 29, 5–68.

Hou, Kewei, Chen Xue, and Lu Zhang, 2015a, Digesting anomalies: An investment approach, *Review of Financial Studies* 28, 650–705.

Hou, Kewei, Chen Xue, and Lu Zhang, 2015b, A comparison of new factor models, Working paper, Ohio State University.

Jeffreys, Harold, 1961, *Theory of Probability* (Oxford University Press, New York).

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65–91.

Jensen, Michael C., 1968, The performance of mutual funds in the period 1945-1964, *Journal of Finance* 23, 389–416.

Jobson, Dave, and Robert Korkie, 1982, Potential performance and tests of portfolio efficiency, *Journal of Financial Economics* 10, 433–466.

Kan, Raymond, Cesare Robotti, and Jay Shanken, 2013, Pricing model performance and the two-pass cross-sectional regression methodology, *Journal of Finance* 68, 2617–2649.

Kandel, Shmuel, and Robert F. Stambaugh, 1987, On correlations and inferences about mean-variance efficiency, *Journal of Financial Economics* 18, 61–90.

Kass, Robert E., and Adrian E. Raftery, 1995, Bayes factors, *Journal of the American Statistical Association* 90, 773–795.

Lintner, John, 1965, The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, *Review of Economics and Statistics* 47, 13–27.

Malatesta, Paul H., and Rex Thompson, 1993, Government regulation and structural change in the corporate acquisitions market: The impact of the Williams Act, *Journal of Financial and Quantitative Analysis* 28, 363–379.

McCulloch, Robert, and Peter E. Rossi, 1991, A Bayesian approach to testing the arbitrage pricing theory, *Journal of Econometrics* 49, 141–168.

MacKinlay, A. Craig, 1995, Multifactor models do not explain deviations from the capital asset pricing model, *Journal of Financial Economics* 38, 3–28.

Merton, Robert, 1973, An intertemporal capital asset pricing model, *Econometrica* 41, 867–887.

Pastor, Lubos, 2000, Portfolio selection and asset pricing models, *Journal of Finance* 55, 179–223.
Pastor, Lubos, and Robert F. Stambaugh, 1999, Costs of equity capital and model mispricing, *Journal of Finance* 54, 67–121.
Pastor, Lubos, and Robert F. Stambaugh, 2000, Comparing asset pricing models: An investment perspective, *Journal of Financial Economics* 56, 335–381.
Rao, Radhakrishna, 1973, *Linear Statistical Inference and its Applications*, Wiley Series in Probability and Statistics (John Wiley & Sons, Hoboken, NJ).
Shanken, Jay, 1987a, Multivariate proxies and asset pricing relations: Living with the Roll critique, *Journal of Financial Economics* 18, 91–110.
Shanken, Jay, 1987b, A Bayesian approach to testing portfolio efficiency, *Journal of Financial Economics* 19, 195–216.
Shanken, Jay, 1992, The current state of the arbitrage pricing theory, *Journal of Finance* 67, 1569–1574.
Sharpe, F. William, 1964, Capital asset prices: A theory of market equilibrium under conditions of risk, *Journal of Finance* 19, 425–442.
Shleifer, Andrei, and Robert Vishny, 1997, The limits of arbitrage, *Journal of Finance* 52, 35–55.
Stewart, Kenneth, 1995, The functional equivalence of the W, LR, and LM statistics, *Economics Letters* 49, 109–112.
Treynor, Jack, and Fischer Black, 1973, How to use security analysis to improve portfolio selection, *Journal of Business* 46, 66–88.
Vuong, Quang, 1989, Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica* 57, 307–333.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: Internet Appendix.