NBER WORKING PAPER SERIES

COMPARING ASSET PRICING MODELS

Francisco Barillas
Jay Shanken

Working Paper 21771
http://www.nber.org/papers/w21771

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2015

Thanks to Doron Avramov, Mark Fisher, Amit Goyal, Lubos Pastor, Tim Simin, Rex Thompson, Pietro Veronesi and seminar participants at the Northern Finance Association meeting, Southern Methodist University, the Universities of Geneva, Luxembourg and Lausanne, Imperial College and Ohio State University. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Comparing Asset Pricing Models
Francisco Barillas and Jay Shanken
NBER Working Paper No. 21771
December 2015
JEL No. G11,G12

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 an analysis of model comparison, i.e., 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,b) and Fama and French (2015a,b) are both dominated by five and six-factor models that include a momentum factor, along with value and profitability factors that are updated monthly.

Francisco Barillas
Goizueta Business School
Emory University
1866 Brockton Glen NE
Atlanta, GA 30329
francisco_barillas@bus.emory.edu

Jay Shanken
Goizueta Business School
Emory University
1300 Clifton Road
Atlanta, GA 30322
and NBER
jay.shanken@emory.edu

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. There are standard econometric techniques for evaluating the adequacy of a single model, but a satisfactory statistical methodology for identifying the best factor-pricing model(s) is conspicuously lacking in investment research applications. We develop a Bayesian procedure that is easily implemented and 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 in finance has attempted to understand the determination of risk premia on financial securities. The central theme of this literature is that the risk premium should depend on a security's market beta or other measure(s) of systematic risk. In a classic test of the CAPM, Black, Jensen and Scholes (1972), building on the earlier insight of Jensen (1968), 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 GRS, who also explore the relation of the test statistic to standard portfolio geometry.[2]

In recent years, a variety of multifactor asset pricing models have been explored. While 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 (2015b). On the other hand, a relatively large p-value may say more about imprecision in estimating a particular model's alphas than the adequacy of that model.[3] Sorely needed are simple statistical

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

[3] 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.

tools with which to analyze the various models *jointly* in a model-comparison framework. This is an area of testing that has received relatively little attention, especially in the case of non-nested models. 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." [4]

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 non-traded factors.[5] While 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.[6] 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 (2015), who highlight the fact that for widely-accepted criteria, model comparison with traded

[4] Avramov (2006) also explores Bayesian model comparison for asset pricing models. As we explain in the next two sections, his 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 as well in cross-sectional regressions. In other Bayesian applications, Malatesta and Thompson (1993) apply methods in comparing multiple hypotheses in a corporate finance event study context.

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

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

factors only requires an examination of each model's ability to price the factors in the other models.

It is sometimes observed that all models are necessarily simplifications of reality and hence must be false in a literal sense. This 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.[7] With these considerations in mind, we extend our results to obtain simple formulas for testing approximate models. Implementation of this approach allows us to go beyond the usual test of an exact model and to obtain insight into a model's goodness of fit.

As an initial application of our framework, 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 the non-nested models {Mkt HML} and {Mkt SMB}. Over the period 1927-2013, alphas for HML when regressed on either 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 of 50% for the two-factor model {Mkt HML} and 40% for FF3.

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 the work of Jegadeesh and Titman (1993). We also include factors from the recently proposed five-factor model of Fama and French (2015a), hereafter FF5. These are RMW (robust minus weak), based on the profitability of firms, and CMA (conservative minus aggressive), related to firms' new net investments. Hou, Xue and Zhang (2015a, 2015b), henceforth HXZ, have proposed their own

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

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 ten factors in our analysis.

Rather than mechanically applying our methodology with all nine of the non-market factors treated symmetrically, we structure the prior so as to recognize that several of the factors are just different versions of the same underlying construct. Therefore, to avoid overfitting, we only consider models that contain at most one version of the factors in each category: 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 factors" amounts to averaging results over the different versions of the factors, with weights that reflect the likelihood that each version contains the relevant factors.

Using data from 1972 to 2013 we find that the individual model with 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 version HML^{m} is considered; and whereas HXZ also found momentum redundant, this is no longer true with 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, IA with CMA, or excluding 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.

Model comparison results assess the *relative* performance of competing models. We also examine *absolute* performance for the top-ranked model and for the HXZ model, which fares better than the FF5 model. These tests consider the extent to which the models do a good job of pricing a set of test assets and any excluded factors. Although various test assets were examined, results are presented for two sets: 25 portfolios based on independent rankings by either size and momentum or by book-to-market and investment. This evidence casts strong doubt on the validity of both models. The "rejection" of the six-factor model is less overwhelming, however, when an approximate version is considered that allows for relatively small departures (average absolute value 0.8% per annum) from exact pricing. With average deviations of 1.2%, the approximate model is actually favored for a range of reasonable priors.

The rest of the paper is organized as follows. Section 1 considers the classic case of testing a pricing model against a general alternative. Section 2 then considers the comparison of nested pricing models and the relation between "relative" and "absolute" tests. Bayesian model comparison is analyzed in Section 3 and Section 4 extends this framework to accommodate analysis with multiple versions of some factors. Section 5 provides empirical results for various pricing models and Section 6 concludes. Several proofs of key results are provided in an appendix.

1. 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.

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:

$$\mathrm{r_t} = \alpha + \beta \mathrm{f_t} + \varepsilon_\mathrm{t} \quad \varepsilon_\mathrm{t} \sim \mathrm{N}(0, \Sigma),$$

where r_t, ε_t and α are Nx1, β is NxK and f_t is Kx1. The normal distribution of the ε_t is assumed to hold conditional on the factors and the ε_t are independent over time. In matrix form,

$$\mathrm{R} = \mathrm{XB} + \mathrm{E}$$

where

$$\mathrm{R} = \begin{bmatrix} r_1' \\ r_2' \\ \vdots \\ r_T' \end{bmatrix}, \quad \mathrm{X} = \begin{bmatrix} 1\ f_1' \\ 1\ f_2' \\ \vdots \\ 1\ f_T' \end{bmatrix}, \quad \mathrm{B} = \begin{bmatrix} \alpha' \\ \beta' \end{bmatrix} \text{ and } E = \begin{bmatrix} \varepsilon_1' \\ \varepsilon_2' \\ \vdots \\ \varepsilon_T' \end{bmatrix}. \tag{1.1}$$

Here, R is TxN, X is Tx(K+1), B is (K+1)xN and E is TxN. The TxK matrix of factor data is denoted by F .

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 value-growth factor. Under the null hypothesis, $H_0 : \alpha = 0$, we have the usual simple linear relation between expected returns and betas:

$$E(r_t) = \beta E(f_t), \tag{1.2}$$

where $E(f_t)$ is the Kx1 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:

$$W = T\frac{\hat{\alpha}'\hat{\Sigma}^{-1}\hat{\alpha}}{1+\mathrm{Sh}(F)^2} = T\frac{\left(\mathrm{Sh}(F,R)^2 - \mathrm{Sh}(F)^2\right)}{1+\mathrm{Sh}(F)^2}. \tag{1.3}$$

Here, $\mathrm{Sh}(F)^2 = \overline{F}'\hat{\Sigma}_F^{-1}\overline{F}$ is the maximum squared sample Sharpe ratio over portfolios of the factors, where $\overline{F}$ is the vector of factor sample means and $\hat{\Sigma}$ and $\hat{\Sigma}_F$ are maximum likelihood estimates (MLE's) for the covariance matrices Σ and Σ_f. The term $\mathrm{Sh}(F,R)^2$ is the corresponding sample measure based on both factor and asset returns. One can also show that W is T/(T-K-1) times the maximum squared t-statistic for the regression intercept, taken over all possible portfolios of the test assets. The population Sharpe ratios, $\mathrm{sh}(\mathrm{f})^2$ and $\mathrm{sh}(\mathrm{f,r})^2$, are based on the true means and covariance matrices.

Under the alternative hypothesis, $\mathrm{H}_1 : \alpha \neq 0$, the F-statistic has a noncentral F distribution with noncentrality parameter λ such that

$$\lambda\left(1+\mathrm{Sh}(\mathrm{F})^2\right)/\,\mathrm{T} \;=\; \alpha\,'\Sigma^{-1}\alpha \;=\; \mathrm{sh}(\mathrm{f},\,\mathrm{r})^2-\mathrm{sh}(\mathrm{f})^2\,. \qquad (1.4)$$

See Gibbons, Ross and Shanken (1989). Under the null hypothesis $\lambda = 0$, the tangency portfolio corresponding to the factor and asset returns, $\tau(f,r)$, equals that based on the factors alone, $\tau(f)$. Thus, the expected return relation in (1.2) is equivalent to this equality of tangency portfolios and their associated squared Sharpe ratios.

A Bayesian F-test

Bayesian tests of the zero-alpha restriction have been developed by Shanken (1987), 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 paper. Although they comment on the computational challenges of implementing this approach, we are able to derive a simple 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.[8] The details are as follows.

The diffuse prior for β and Σ is

[8] 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.

$$\mathrm{P}(\beta,\, \Sigma) \propto |\Sigma|^{-(N+1)/2}, \tag{1.5}$$

as in Jeffreys (1961). The prior for α is concentrated at 0 under the null hypothesis. Under the alternative, we assume a multivariate normal informative prior for α conditional on β and Σ :

$$\mathrm{P}(\alpha \mid \beta, \Sigma) = \mathrm{MVN}(0,\, \mathrm{k}\Sigma), \tag{1.6}$$

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 McKinlay (1995), they stress the desirability of a positive association between α and Σ in the prior, which makes extremely large Sharpe ratio less likely, as implied by (1.4).[9]

For a single asset, (1.6) 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 (1.4), 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 $\alpha'(\mathrm{k}\Sigma)^{-1}\alpha$ is distributed as chi-square with N degrees of freedom, so the prior expected value of $\alpha'\Sigma^{-1}\alpha$ is k times N. Therefore, given a target value $\mathrm{Sh}_{\max}$ for the square root of the expected maximum, the required k is

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

$$\mathrm{k} = \left(\mathrm{Sh}_{\max}^2 - \mathrm{Sh}(\mathrm{f})^2\right) / \mathrm{N} \tag{1.7}$$

Note that using the factor data to inform the prior is appropriate in this context since the entire statistical analysis is conditioned on f. Alternatively, we can just think about the expected return relation and our assessment of plausible deviations from that relation. This is similar to the approach in Pastor (2000). If our subjective view is, say, that alphas should be less than 6% (annualized) with probability 95%, then we would want to choose k such that σ_α is about 3% (annualized). Given a residual standard deviation of 10% per annum, for example, the implied k would be $0.03^2/0.10^2 = 0.09$.

The Bayes factor BF measures the relative support for the null hypothesis in the data. Formally, BF is the ratio of the marginal likelihoods: $\mathrm{ML(H_0)/ML(H_1)}$, where each ML 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 ML can be viewed as a function of the data (factor and test-asset returns):

$$ML = P(R\,|\,F) = \int \int P(R\,|\,F,\,\alpha,\beta,\Sigma)\,P(\alpha\,|\,\beta,\Sigma)P(\beta,\,\Sigma)\,\mathrm{d}\,\alpha\,\mathrm{d}\,\beta\,\mathrm{d}\,\Sigma. \tag{1.8}$$

Here, the likelihood function is the joint conditional density $P(R\,|\,F,\,\alpha,\beta,\Sigma)$ viewed as a function of the parameters. $\mathrm{ML(H_1)}$ is computed using the priors given in (1.5) and (1.6); $\mathrm{ML(H_0)}$ also uses (1.5), but substitutes the zero vector for α.

We can also view the test of $\mathrm{H_0}: \alpha = 0$ vs $\mathrm{H_1}: \alpha \neq 0$ in terms of the proportionality constant in the prior covariance matrix for α. Thus, we have a test of the value 0 vs. the value k. More generally, the null hypothesis can be modified to accommodate an approximate null that allows for small average 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 our main result.

Proposition 1. Given the factor model in (1.1) and the prior in (1.5)-(1.6), the Bayes factor for $H_0 : \alpha = 0$ vs $H_1 : \alpha \neq 0$ equals

$$\mathrm{BF} = \frac{\mathrm{ML}(\mathrm{H}_0)}{\mathrm{ML}(\mathrm{H}_1)} = \frac{1}{\mathrm{Q}} \left(\frac{|\mathrm{S}|}{|\mathrm{S}_\mathrm{R}|} \right)^{(\mathrm{T}-\mathrm{K})/2} \tag{1.9}$$

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

$$\begin{aligned} \mathrm{Q} &= \int \int \exp\left(\frac{-1}{2\mathrm{a}} (\alpha - \hat{\alpha})' \Sigma^{-1} (\alpha - \hat{\alpha}) \right) \mathrm{P}(\alpha \,|\, \Sigma) \mathrm{P}(\Sigma \,|\, \mathrm{F}, \mathrm{R}) \, \mathrm{d}\alpha \, \mathrm{d}\Sigma \\ &= \left(1 + \frac{\mathrm{a}}{(\mathrm{a}+\mathrm{k})} (\mathrm{W}/\mathrm{T}) \right)^{-(\mathrm{T}-\mathrm{K})/2} \left(1 + \frac{\mathrm{k}}{\mathrm{a}} \right)^{-\mathrm{N}/2}, \end{aligned} \tag{1.10}$$

where $P(\Sigma \,|\, F, R)$ is the posterior density for Σ and $\mathrm{a} = \left(1 + \mathrm{Sh}(\mathrm{F})^2\right)/\mathrm{T}$. W is given in (1.3) and equals the GRS F-statistic times NT/(T-N-K). Letting Q_{k0} be the value of Q obtained with prior value k_0, the BF for k_0 vs k is

$$\mathrm{BF}_{\mathrm{k}_0,\mathrm{k}} = \mathrm{Q}_{\mathrm{k}_0} / \mathrm{Q} \tag{1.11}$$

Proof. See Appendix B.[10, 11]

It is easy to verify that BF is a decreasing function of W; the larger the test statistic, the stronger is the evidence against the null that α is (approximately) 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

[10] Harvey and Zhou derive (1.9) and the integral expression for Q. The function of W in (1.10) is our simplification, while (1.11) is both a simplification and generalization of (1.9).

[11] 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.

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 (1.9), the BF 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. BF also decreases as Q increases, where 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. Q is always less than one since the exponent in (1.10) is uniformly negative. As the ratio of determinants is likewise less than one, a BF favoring the null (BF > 1) occurs when Q is sufficiently low, i.e., the prior for alpha under the alternative is "inconsistent" with the estimate. The simple formula for Q will also facilitate the model comparison calculations in Section 3.

One may wonder why we don't consider a diffuse prior for alpha, so as to avoid having to make an assumption about the prior parameter k. For some specifications, a diffuse prior can be obtained in the limit by letting the prior variance approach infinity. Allowing $\sigma_\alpha \rightarrow \infty$ would amount to letting $k \rightarrow \infty$, which is not sensible economically since it implies, by (1.7), that the maximum 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). They discuss the relevance of "Bartlett's (1957) paradox, a situation in which 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), who recognizes 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.

One general method that can be used to obtain a proper prior is to update a diffuse prior with a "minimal training-sample," i.e., 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 (2006) adopts 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' recommendation that the prior spread not be "too big."

2. Relative versus Absolute Model Tests

In the previous section, we analyzed 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. 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 in all 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. This 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, e.g., the Sharpe-Lintner CAPM and the Merton (1973) intertemporal CAPM.[13]

In the present setting, the vector f corresponds to a *subset* of L-1 of the K-1 non-market 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

[12] Berger and Pericchi (1996) suggest averaging such results across all minimal training samples as a means of increasing stability of the procedure.

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

model. However, the statistical analysis is greatly facilitated by using an equivalent representation of M. Let

$$f^* = \alpha^* + \beta^*[Mkt, f] + \varepsilon^* \tag{2.1a}$$

and

$$r = \alpha_r + \beta_r[Mkt, f, f^*] + \varepsilon_r \tag{2.1b}$$

be multivariate regressions for f^* and r. Note that the model in (2.1b), which we will call M_a, includes all K factors. Barillas and Shanken (2015) show that the model M, which is nested in M_a, holds if and only if $\alpha^* = 0$ and $\alpha_r = 0$, i.e., if and only if $\alpha^* = 0$ and M_a holds for the test assets. We use this characterization of M below.

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 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. Otherwise, FF3 is preferred since it does not impose the additional restrictions that are violated. Alternatively, as Barillas and Shanken note, we can think about comparison in terms of a modified Hansen-Jagannathan (1997) distance, which ends up being equivalent to a factor-portfolio efficiency criterion.[14] If $\alpha^* = 0$, the tangency portfolio (and associated Sharpe ratio) based on all the factors, i.e., spanned by Mkt, HML and SMB, can be achieved 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 (classical or Bayesian) naturally present themselves in connection with the nested model representation of M. We can conduct a test of the all-

[14] The modification, due to Kan and Robotti (2008), is suitable in the excess-return setting with traded factors.

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. Finally, we can perform a *relative* test of M vs 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}. Given some additional assumptions similar to those made earlier, we then have

Proposition 2. Assume that the multivariate regression of f^* on (Mkt, f) in (2.1a) and r on (Mkt, f, f^*) in (2.1b) 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 (1.5) and (1.6), with the priors for (2.1a) and (2.1b) independent. Then the BFs are related as follows:

$$BF^{abs}_{M} = BF^{rel} \times BF^{abs}_{M_a} \tag{2.2}$$

Proof. The ML is the expectation under the prior of the likelihood function. Write the joint density (likelihood function) of factor and test-asset returns 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, under M, both densities are restricted (zero intercepts) in the numerator, whereas only the density for r is restricted under M_a. Therefore, letting the subscripts R and U stand for restricted and unrestricted densities,

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

and

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

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

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

the equality in (2.2) 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 vs. 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 relate the absolute tests for each model, the result is the same. This reflects the fact that, extending the argument in Barillas-Shanken (2015), 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. We refer to this as test-asset irrelevance.

Testing CAPM vs. FF3: An Illustration

To illustrate these ideas, suppose we want to test whether M_a, here the FF3 model with K = 3, is superior to M = CAPM with L = 1. In this case, f is the empty set (no non-market factors in CAPM) and there are 2 (K–L) excluded factors, f^* = (HML, SMB). Since the test assets are irrelevant, the pertinent restriction is that the CAPM alphas of SMB and HML are both zero. Thus, 1 plays the role of K and 2 the role of N in the required application of Proposition 1. We evaluate the CAPM restriction from both the classical and Bayesian perspectives, using monthly factor data for the period 1927 to 2013 obtained from Ken French's website (T = 1004). The GRS statistic is 4.56 with associated p-value 0.01, statistically significant in the conventional sense. The corresponding Wald statistic is then 2(1044)/(1044 - 2 - 1) times 4.56 or W = 9.14.

To implement the Bayesian approach, we need to specify the value of k in the prior. Since the full model M_a = FF3 includes 3 factors and the nested model M consists of the market factor only, adaptation of the earlier formula for k in (1.7) gives

$$k = \left(Sh_{max}^2 - Sh(Mkt)^2\right) / (3-1). \qquad (2.3)$$

The divisor is 2 here since the two excluded factors are added to the market factor and play the role of left-hand-side assets. As in Section 1, using the Mkt data to inform the prior is appropriate here since the entire statistical analysis is conditioned on the Mkt returns. Similar remarks will apply to the general model-comparison analysis of the next section.

In our example, the question is, how big do we think the Sharpe ratio increase might be as a result of adding HML and SMB to the market index? We allow for a 25% increase in this illustration, i.e., $\mathrm{Sh}_{\mathrm{max}} = 1.25 \times \mathrm{Sh(Mkt)}$. More precisely, the square root of the prior expected squared Sharpe ratio is 1.25 time the market's squared ratio. With a value of 0.115 for $\mathrm{Sh(Mkt)}$, a = (1 + 0.115^2)/1044 = 0.00097 and (2.3) gives k = 0.0037. Using (1.6), this value of k translates into (annualized) prior standard deviations of 2.51% and a 2.23% for the HML and SMB CAPM alphas, respectively. The latter is smaller since the SMB residual variance in the regression on Mkt is lower than that of HML over the full period.

Given this prior specification and letting $k_0 = 0$, the BF for the null CAPM vs. the alternative FF3 is Q_0/Q in (1.10)-(1.11), which equals

$$\left(\frac{1+\frac{a}{(a+k)}(W/T)}{1+(W/T)}\right)^{(T-1)/2}\left(1+\frac{k}{a}\right)^{2/2} =$$

$$\left(\frac{1+\frac{0.097}{(0.097+0.37)}(9.14/1044)}{1+(9.14/1044)}\right)^{(1004-1)/2}\left(1+\frac{0.37}{0.097}\right)^{2/2}$$

or 0.13. Thus the data (viewed through the lens of the prior), strongly favor the conclusion that the two alphas are not both zero, by odds of more than 7 to 1. Using the fact that the probability for the alternative is one minus the probability for the null, it follows that the posterior probability that the null is true is BF/(1 + BF) when the prior probabilities for both models are

0.5. This gives a posterior probability of 11.6% for CAPM with the BF of 0.13. As the p-value calculation does not even consider the alternative hypothesis, the 1% p-value cannot meaningfully be compared to this posterior probability.[15] Later, we provide an example in which the posterior probability favors the null, even though the p-value is low by conventional standards.

3. General Model Comparison

In the previous section, we saw how to use an asset-pricing test against a general alternative to compare two nested factor-pricing models. Now suppose we wish to simultaneously compare a collection of asset pricing models, both nested and non-nested. From a portfolio perspective, it is clear that the squared Sharpe ratio will always be maximized with all factors included, as in model M_a. The question that is being addressed when we consider the various models is whether that maximum can still be attained with a proper subset of the factors, one as small as possible. In other words, are some of the factors in M_a redundant? Fama (1998) considers a related hypothesis in identifying the number of priced state variables in an intertemporal CAPM setting.

Our methodology exploits the fact that the marginal likelihood (ML) for each model is the product of an unrestricted ML for the included factors and a restricted (alpha = 0) ML for the factors that are excluded and must be priced by the model.[16] Given our analysis of the Bayesian F-test, these likelihood measures are easily calculated, with excluded factors playing the role of left-hand-side returns. Thus, inference about model comparison ends up being based on an

[15] Shanken (1987b) discusses this issue in detail.

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

aggregation of the evidence from all possible multivariate regressions of excluded factors on factor subsets.[17]

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 non-nested two-factor models {Mkt HML} and {Mkt SMB}. Given the $\mathrm{ML_j}$ for each model $\mathrm{M_j}$ with prior probability $\mathrm{P(M_j)}$, the posterior probabilities conditional on the data D are given by Bayes' rule as

$$\mathrm{P(M_j \mid D)} = \left\{\mathrm{ML_j} \times \mathrm{P}\left(\mathrm{M_j}\right)\right\} / \left\{\sum\nolimits_{\mathrm{i}} \mathrm{ML_i} \times \mathrm{P}\left(\mathrm{M_i}\right)\right\}, \tag{3.1}$$

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 (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.[18] This is important from the statistical standpoint, as well as the asset pricing perspective, since (3.1) requires that the posterior probabilities for all models are conditioned on the same data. Thus, each model's restrictions are imposed on the excluded factors $\mathrm{f^*}$ as well as the test assets r in calculating the ML, whereas the ML for the included factors f is based on their *unrestricted* joint density.[19] Therefore, we also need to consider the multivariate regression

$$\mathrm{f} = \alpha + \beta \mathrm{Mkt} + \varepsilon, \tag{3.2}$$

[17] A frequentist approach to asset-pricing model comparison might be developed along the lines of Vuong (1989), but we leave that to future work.

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

[19] That all marginal likelihoods must, in principle, be conditioned on the same data is a direct consequence of Bayes' theorem, e.g., Kass and Raftery (1995) equation (1). In traditional model comparison applications such as Avramov (2002), which 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 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 in this case. The same is not true for the excluded factors in our application, as their pricing by the included factors *does* affect the model likelihood.

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

Proposition 3. Assume that the multivariate regressions of f on Mkt in (3.2), f^* on (Mkt, f) in (2.1a) and r on (Mkt, f, f^*) in (2.1b) satisfy the distributional conditions discussed previously. The prior for the parameters in each regression is of the form in (1.5) and (1.6), with independence between the priors for (3.2), (2.1a) and (2.1b) conditional on the sample of Mkt returns. Then the ML for a model M with non-market factors f is of the form

$$ML = ML_U(f \mid Mkt) \times ML_R(f^* \mid Mkt, f) \times ML_R(r \mid Mkt, f, f^*), \tag{3.3}$$

where the unrestricted and restricted (alpha = 0) regression MLs are obtained using (A.1) and (A.2) of Appendix A.

The value of k in the prior for the intercepts in the unrestricted regressions is determined as in (2.3), but using the total number of factors K in the denominator, with Sh_{max} corresponding to all K factors as well. It follows from the discussion after (1.6) that k is the expected (under the alternative) 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. Although we think this is a reasonable way to specify the prior, the results are not sensitive to alternative methods we have tried for distributing the total increase in the squared Sharpe ratio.

Given (3.3), the posterior model probabilities in (3.1) can now be calculated by substituting the corresponding ML for each model. We use uniform prior model probabilities to avoid favoring one model or another, which seems desirable in this sort of research setting. Thus, the impact of the data on beliefs about the models is highlighted. Other prior assumptions could easily be explored, however. Note that, since the last term in (3.3) is the same for all models, it cancels out in the numerator and denominator of (3.1). Thus, test assets are irrelevant for model comparison, as in the nested case of Proposition 2.

Model Probabilities with the Fama-French (1993) Factors

Over the period 1927 to 2013, we saw in the earlier illustration, that CAPM is rejected in favor of FF3 based on the conventional GRS test with p-value 0.01. The Bayes factor for CAPM vs. FF3 was 0.13. Now we compare these models simultaneously with the two-factor models {Mkt HML} and {Mkt SMB}. As earlier, the prior assumes that Sh_{max} = 1.25xSh(Mkt) for the three factors, with prior probability of 1/4 assigned to each of the four models.

Before conducting the formal Bayesian analysis, we examine some additional regression evidence - annualized alpha estimates with t-statistics in parentheses. Recall that the BF in favor of the zero-null hypothesis for a single dependent return is a decreasing function of this t-statistic. The alpha of HML on Mkt is large at 3.65% (2.85), while the alpha of SMB on Mkt is 1.34% (1.18). For SMB on Mkt and HML, the alpha is just 1.18% (1.04) and for HML on Mkt and SMB, it is 3.57% (2.79). The large HML alphas are evidence against CAPM and {Mkt SMB}, but consistent with both the two-factor model {Mkt HML} and FF3. The modest SMB alphas point to the two-factor model, however. But how strongly should we view this suggestion?

The Bayesian approach aggregates all of the evidence reflected in the marginal likelihoods and summarizes the results in terms of posterior model probabilities. For example, using (3.3), the ML for the model {Mkt HML} equals the unrestricted ML for HML conditional on the market (the HML alpha is unconstrained) times the restricted ML for SMB conditional on Mkt and HML (the SMB alpha is constrained to be zero) times the restricted ML for the test assets conditional on all three factors (test-asset alphas are zero). A similar calculation for each model results in model probabilities of 50.6% for {Mkt HML}, 39.6% for FF3, 5.2% for CAPM and 4.6% for {Mkt SMB}. In this application, the restricted model {Mkt HML} that excludes SMB comes out on top, but in other situations the model that includes all of the factors can dominate the competition. Also note that the ratio of probabilities for CAPM and FF3 is 0.13, equal to the BF obtained earlier from the Bayesian F-test.

4. Comparing Models with Categorical Factors

Often, in empirical work, several of the available factors amount to different implementations of the same underlying concept, 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.

A Categorical Example

Our main empirical application presented later in Section 5 will include data for four factor categories. This data is available over the period 1972-2013. 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 spread. 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*.

To demonstrate the basic idea, consider categorical models based on the standard factors Mkt and HML, and the categorical factor Size. As earlier, there are four categorical models, CAPM, {Mkt HML} {Mkt Size} and {Mkt HML Size), each with prior probability 1/4.[20] We have two versions of the factors, w_1 = (Mkt HML SMB) and w_2 = (Mkt HML ME), and can conduct separate model comparison analyses with each over the 1972-2013 period. These separate analyses employ the methodology of Section 3 with all standard factors. The posterior model probabilities conditional on w_1 are {Mkt HML} 55.9%, {Mkt HML SMB} 43.5%, CAPM 0.4% and {Mkt SMB} 0.2%. Conditional on w_2, we have {Mkt HML ME} 50.7%, {Mkt HML}

[20] The value of k in the prior now corresponds to a potential 50% increase in the Sharpe ratio relative to that of the market when the categorical model contains all three factors.

48.6%, {Mkt ME} 0.3% and CAPM 0.3%. Now the question is, how should we aggregate these two sets of probabilities to obtain posterior probabilities for all six models?

First, suppose we assign prior probability 1/2 to each w, i.e., 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 both 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 different versions of these categorical models, as desired.

The proposition below 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 47.8% for w_1 and 52.2% for w_2. The probability for {Mkt HML} is then (47.8%)(55.9%) + (52.2%)(48.6%) = 52.1%. For **{**Mkt HML ME}, which is only associated with the second version of the factors, it is (47.8%)(0) + (52.2%)(50.7%) = 26.5%. The probability is 20.8% for {MKT HML SMB} and less than 1% for the remaining models. Note that the probabilities for models that include SMB are fairly similar to those for models that include ME. This makes sense since the correlation between the two size factors is very high (0.98). From the categorical model perspective, we have probability 52.1% for {Mkt HML}, 26.5% + 20.8% = 47.3% for {Mkt HML Size} and less than 1% for the remaining models.[21]

Aggregation over Different Versions of the Factors

Assume the categorical models consist of up to K factors, K_C of which are categorical factors. In the example above, K = 3 and K_C = 1. In general, there are 2^{K_C} versions of the K

[21] Readers less interested in the methodological details can skip to the empirical Section 5 at this point.

factors (w_1 and w_2 above), 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 MLs for the various models be based on the same data. This requires a simple extension of (3.3). Take the model {Mkt ME), for example. In the present context, the excluded factors consist of the non-categorical factor HML and the other version of the categorical Size factor, SMB. We write the ML as the unrestricted ML for ME given Mkt, times the restricted ML for HML given Mkt and ME, times the restricted ML for SMB given all three factors in the version w_2 = (Mkt HML ME). Thus, the extra ML 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 ML 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 K + K_C factors in all (3 + 1 = 4 in the example). F refers to all of this factor data. For a model M with non-market factors f, all contained in w, let f^* now denote the factors in w that are excluded from M.[22] Let the prior for the parameters in the regression of w^* on (Mkt, f, f^*) take the usual form based on (1.5) and (1.6), again independent of the other components. The test-assets r are now regressed on (Mkt, f, f^*, w^*), i.e., *all* the factors, and so the corresponding ML will cancel out in all probability

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

calculations, as earlier. This term will be ignored going forward. Therefore, the counterpart of (3.3) is now

$$ML(M \mid w) = ML_U(f \mid Mkt) \times ML_R(f^* \mid Mkt, f) \times ML_R(w^* \mid Mkt, f, f^*). \qquad (4.1)$$

For each w, the conditional posterior probabilities can then be obtained as in (3.1). Note that the ML 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 we now see.

Proposition 4. The unconditional (not conditional on w) posterior model probabilities are obtained as follows. First, calculate ML(M | w) in (4.1) for each version w of the factors and each model in w. Then calculate the ML for each w as

$$ML(w) = E_{M|w}\{ML(M \mid w)\}. \qquad (4.2a)$$

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

$$P(F) = E_w\{ML(w)\}, \qquad (4.2b)$$

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

$$P(w \mid F) = ML(w)P(w)/P(F). \qquad (4.2c)$$

Finally, the unconditional probability for M is

$$P(M \mid F) = E_{w|F}\{P(M \mid w, F)\}, \qquad (4.2d)$$

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 (4.2a), M plays the role of X and F the role of Y, while we condition on w throughout in the "background." In (4.2b), w plays the role of X and F the role of Y. (4.2c) is just Bayes' theorem. Finally, w plays the role of X and M the role of Y in (4.2d), while we condition on F throughout in the background. □

5 Empirical Results

In this section, we first present model-comparison evidence and then Bayesian F-test results.

5.1 Empirical Results on Model Comparison

Model probabilities are shown at each point in time to provide an historical perspective on how posterior beliefs would have evolved as the series of available returns has lengthened. Examining different sets of factors provides additional perspective, as the collection of factors considered in the research community has expanded over time. Thus it is interesting to see how this affects posterior beliefs about the models. First, we simultaneously compare all the models that can be formed using the FF3 factors Mkt, SMB and HML. This small example extends the results shown in Section 3 and serves as a good illustration of our methodology. We then conduct our main empirical analysis, which compares models that can be formed from ten prominent factors in the literature. Since there are four categorical factors with two versions each, an admissible model will have at most six factors.

Our benchmark scenario for all model comparison exercises in this section assumes that Sh_{max} = 1.5 x Sh(Mkt), i.e., 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 3, this is sufficient to determine the implied Sh_{max} values as we expand the set of included factors from one to all six, leaving the intercepts unrestricted. In particular, for the three-factor model below, the corresponding multiple of Sh(Mkt) is 1.27, slightly higher than the 1.25 value used earlier. 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. Later, we examine the sensitivity of posterior beliefs to this assumption, as we also explore multiples corresponding to a more behavioral perspective and one with a lower value.[23]

Model Probabilities with the Three Fama-French (1993) Factors

In previous sections, we presented results using these three factors to illustrate our methodology. Recall that there are four models in all: CAPM, FF3 and the two-factor models {Mkt HML} and {Mkt SMB}. We now provide evidence on the formal model comparison among these four competing models over time. As earlier, we employ data from January 1927 to December 2013. Figure 1 presents the results of this exercise. The top panel shows the model probabilities while the bottom panel gives cumulative factor probabilities, i.e., the probability that each factor is included in the best model.

Since we start with equal prior probabilities for each model, it is not surprising that it takes a while to see a substantial spread in the posterior probabilities. The best-performing model since the mid-1980s has been {Mkt HML}, followed closely by FF3. The probabilities for these models are 51.3% and 39.1%, respectively, at the end of the sample. It is also of interest to note that the full model (FF3) need not have the highest probability. The CAPM and {Mkt SMB} probabilities generally decline after 1980 and are quite low at the end. However, CAPM would have been perceived as the best-performing model in the 1950s and 1960s, which interestingly was a time when the Fama-French model ranked last. In related evidence, Fama and French (2006) and Ang and Chen (2007) find that CAPM works well for B/M-sorted portfolios before 1963.

The cumulative factor probabilities are shown in the bottom panel. For each factor, this is the sum of the posterior probabilities for models that include that factor. The probabilities at the end of the sample are 90.3% for HML, reflecting its inclusion in the two top models, and 43.6% for SMB. Of course, the probability is one for Mkt by assumption.

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

[Figure 1]

The empirical analysis above was based on the prior assumption Sh_{max} = 1.5 x Sh(Mkt) for six factors, including the market. Now we explore the sensitivity of our full-sample results to different prior assumptions - multiples of 1.25, 1.5, 2, and 3. Table 1 presents the full-sample results for the three Fama-French factors. The sample Sharpe ratio for the market is 0.115 over the 1927-2013 period, while the FF3 Sharpe multiple is 1.23, close to the 1.27 implied by our baseline prior scenario. With the three Fama-French factors, the top model is always {Mkt HML}, the posterior probabilities rising from 44.9% to 65.3% as Sh_{max} increases. At the same time, the probabilities for FF3 decline from 42.2% to 23.1%. Overall, although we see some variation in the model probabilities for different priors, the rankings of the models are consistent.

[Table 1]

Model Probabilities with Ten Prominent Factors

We now consider a total of ten 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 (2015a). Finally, we also include the size ME, investment IA and profitability ROE factors in Hou, Xue and Zhang (2015a, 2015b), as well as 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 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 (2x3) sorts, interacting size with operating profitability for the construction of RMW, and separately with investments to create CMA. 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 2x3 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 x 3 x 3) sort on size, investment-to-assets, and ROE. More importantly, the HXZ factors use different measures of investment and profitability. Fama and French (2015a) 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. IA is the annual change in total assets divided by one-year-lagged total assets, whereas investment used by Fama and French is the same change in total assets from the fiscal year ending in year t-2 to the fiscal year in t-1, divided by total assets from the fiscal year ending in t-1, rather than t-2. In terms of 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 2013. 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 non-market factors treated symmetrically, we apply the framework of Section 4, which recognizes that several of the factors are just different versions of the same underlying concept. Therefore, we only consider models that contain at most) one version of the factors in each category: 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 *categorical factors*, in this context, in contrast to the actual factors employed in the various models. Similarly, models in which *some* of the factors are categorical and the rest are standard factors are termed *categorical models*. 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, we have a total of 162 models under consideration.

The top panel in Figure 2 shows posterior probabilities for the individual models, which were obtained under our baseline prior that allows for a multiple of 1.5 times the market Sharpe ratio. We find that quite a few of the individual models receive non-trivial probability, the best (highest probability) model being the six-factor model {Mkt SMB ROE IA HML^m UMD}. The second-best individual model replaces IA with CMA, the third-best uses ME instead of SMB and the sixth one uses both CMA and ME, as opposed to IA and SMB. Both the fourth and fifth best models 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 UMD. All of these models fare better than FF5 and the four-factor model of HXZ.

Figure 3 provides another perspective on the evidence, aggregating results over the different versions of each categorical model. Similar to the findings in the previous figure, by the end of the sample, the six-factor categorical model {Mkt Value Size Profitability Investment UMD} comes in first with posterior probability close to 75% and the five-factor model that excludes size is next, but with probability a little below 20%. The third best categorical model consists of the same five categories as in FF5, while the fourth best replaces the investment factor with momentum. However, it is essential that the more timely versions of value and profitability are employed in these models. Specifically, in untabulated calculations, the probability share for HML^m in the FF5 categorical model is 89.0%. 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 categorical model. Similarly, the shares for ROE are 99.9% in the categorical FF5 model and 99.0% in the categorical four-factor model.

In terms of cumulative probabilities aggregated over *all* models, we see from the bottom panel of Figure 3 that the recently proposed category, profitability, ranks highest. Interestingly, value is second with over 99% cumulative probability. Consistent with the findings in Figure 2, the *categorical share* for HML^m, i.e., 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 99.5% for ROE. There is less dominance in the size and investment categories, with shares of 75.4% for SMB and 63.1% for IA.

While the analysis above simultaneously considered all 162 possible models, we have also conducted 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. 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 (3.3). Assuming prior probability 0.5 for each model and zero probability for all other models, the posterior probability for model 1 in (3.1) 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 96.6% probability to the six-factor model. The six-factor model probability is greater than 99% when compared to FF5, even if the size factor is deleted from the model.

The model comparison above was based on a prior assumption that $Sh_{max} = 1.5*Sh(Mkt)$ in (2.5) when working with six factors. We next examine sensitivity to prior Sharpe multiples of 1.25, 1.5, 2, and 3. Tables 2 and 3 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 top 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 2 and 3 times the market ratio. Their probabilities rise from 21.9% to 45.3% and from 12.6% to 25.7% as the multiple increases from 1.25. These two models are among the top four under the lower-multiple specification, though the posterior probabilities are more diffuse in this case.

The top model rankings for the categorical models are also stable across the different priors. The six-factor categorical model {Mkt SIZE PROF INV VAL MOM} is always the best, regardless of the prior, and its posterior probability increases substantially as the multiple increases. The categorical model that excludes size comes in second for all multiples as well. As noted above, the more timely HML^{m} accounts for 99.5% of the cumulative probability for the value category. Table 4 shows that timely value remains responsible for the lion's share of the cumulative value probability across the different priors, especially at higher multiples. Varying the prior also yields fairly similar results for timely profitability (ROE), as well as IA and SMB.

Relative Tests: Are Value and Momentum Redundant?

Barillas and Shanken (2015) show that when comparing two asset-pricing models, all that matters is the extent to which each model prices the factors in the other model. Hou, Xue and Zhang (2015b) and Fama and French (2015a) 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 a value (HML^{m}) and a momentum factor (UMD).

To shed further light on this finding, Table 5 shows 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. HML^{m} has an alpha of 6.1% (t-stat 5.26) and UMD has an alpha of 6.6% (t-stat 3.96). 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 0.99% with a t-stat of 0.81. The different results for the two value factors are largely driven by the fact that HML^{m} is strongly negatively correlated (-0.65) with UMD, whereas the correlation is only -0.15 for HML[24]. The negative loading for $\mathrm{HML^m}$ when UMD is included lowers the model expected return and raises the $\mathrm{HML^m}$ alpha, so that this timely value factor is not redundant.

We now evaluate the hypothesis that $\mathrm{HML^m}$ is redundant from a Bayesian perspective. Figure 4 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 $\mathrm{HML^m}$ to investment in the other factors. The horizontal axis in each panel of the figure shows the prior multiple. This is the Sharpe ratio for the alternative, expressed as a multiple of the Sharpe ratio for the factors in the null model that excludes $\mathrm{HML^m}$.

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, strongly rejecting the conclusion that $\mathrm{HML^m}$ is redundant. Although the inference is not sensitive to the prior here, in other cases it may well be. The right panel of Figure 4 provides information about the implied value of σ_α. This gives an idea of the likely magnitude of α's envisioned under the alternative and should be helpful in identifying the range of prior multiples that one finds reasonable.[25] For example, to get an increase in the Sharpe ratio of 25% from the already-high level of 0.44 for the null model, we would need a very large σ_α of about 7.5% per year.

[Figure 4]

The Bayesian analysis for UMD (not shown) looks much the same as Figure 4, strongly rejecting redundancy. To highlight the role of $\mathrm{HML^m}$ in this finding, we exclude that factor and

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

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

show that the evidence then favors the conclusion that UMD *is* redundant with respect to the remaining factors {Mkt SMB IA ROE}. This essentially confirms the earlier finding of Hou, Xue and Zhang (2015b), but with SMB as the size factor, rather than ME. In Figure 5, we see that the posterior probability for the null hypothesis of redundancy (UMD alpha is zero) is always above 50%, with values over 80% for Sharpe ratio multiples around 1.15. The conventional p-value also exceeds 50% here, as indicated by the horizontal line in the figure.

[Figure 5]

5.2. Absolute Test Results

In this section we apply Proposition 1 with a set of test assets - what we call an absolute test. We will see that failing to account for the excluded factors when conducting the absolute test can lead to the conclusion that an inferior model performs better than one that is actually superior. However, once we incorporate the excluded factors, comparing the absolute results for the two models is in line with our earlier results on model comparison.

We saw above that over the sample period 1972-2013, the model with the highest posterior probability is the six-factor model {Mkt IA ROE SMB HML^m UMD}. Now we evaluate this model, as well as the four-factor model of HXZ, from the absolute perspective. Although a wide variety of test-asset portfolios has 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. Strictly speaking, the two models considered in this section are not nested because the HXZ model uses ME, whereas our top model uses SMB. However, the results are similar whether one uses ME or SMB.

To test the HXZ model, we initially follow common practice and only employ the test-asset portfolios. Then we add in the excluded factors UMD and HML^m as left-hand-side assets.

Using the 25-size/momentum portfolios from January 1972 to December 2013, the GRS statistic for the HXZ model is 2.72 with p-value approximately zero, rejecting the model. A descriptive statistic that has also been used to judge model performance is the average of the absolute values of the test-asset alphas, e.g., Fama and French (2015a). The HXZ model produces an average absolute alpha of 1.42% per annum. When we add the excluded factors UMD and HML^m as left-hand-side assets, the GRS statistic is 10.5 with p-value virtually zero, but the average absolute alpha increases only slightly to 1.45%.

The Bayesian F-test results for the HXZ model with the size/momentum portfolios are given in Figure 6. Similar to the redundancy tests, the horizontal axis in the figure shows the multiple of the Sharpe ratio for the factors in the given null model. This is the multiple under the alternative that the left-hand-side assets are not priced by the model. The blue line in each panel shows the results without the excluded factors UMD and HML^m, whereas the red dashed line adds those factors as left-hand-side returns. We see in the left panel that the probability for the HXZ model is close to zero for Sharpe multiples in the range of 1.1 to 1.6 (blue), but when UMD and HML^m are added, there is even stronger evidence against the model, with the probability close to zero for a much wider range of priors (red).

[Figure 6]

Next, we examine the absolute performance of the six-factor model with the same 25 size/momentum portfolios. The GRS statistic is 3.5 and the corresponding p-value is nearly zero, strongly rejecting the model in a classical sense. The Bayesian F-test also provides strong evidence (not shown) against the null hypothesis. The probability of the null curve looks very similar to the red line in Figure 6, quickly declining to an extended zero-probability range. Interestingly, in this case the average absolute alpha is 1.93% per annum, which is much higher than the 1.42%/1.45% under the HXZ model (with/without UMD and HML^m). Yet, in our model comparison analysis, the six-factor model was strongly preferred to the HXZ model. This is an example of the sort of conflict discussed in Barillas and Shanken (2015), who argue that model comparison must be based on excluded-factor restrictions, whereas test-asset metrics can be misleading. Indeed, we have verified that the ratio of the absolute-test probability for the HXZ model (incorporating UMD and HML^m) to that for the six-factor model is very small, except for Sharpe multiples close to one.

Now we turn to the results for the 25 portfolios formed on sorts by book-to-market and investment. For the HXZ model, the GRS statistic for the test-asset restrictions is 1.53 with a p-value of 0.05. The average absolute alpha is 1.44% per annum. Adding the excluded factors UMD and HML^m increases the GRS statistic to 1.94. The p-value is now much smaller, 0.004, but the average absolute alphas are only slightly higher at 1.47%. The Bayesian results are plotted in Figure 7. The probability based solely on test assets (blue line) is substantial and never below 20%. As indicated in the figure, the probability exceeds one-half for Sharpe ratios a bit greater than 1.2 (σ_α around 1.8%) and beyond. With HML^m and UMD considered (red line), however, the probability for the null is close to zero for all but the tightest priors. Thus, challenging the HXZ model to price excluded factors reduces the probability of the null substantially for a wide range of priors.

[Figure 7]

Next, we use the same 25 book-to-market and investment portfolios to evaluate the six-factor model. The GRS statistic is 2.89 with p-value nearly zero. More interestingly, the average absolute alpha is 2.88% in this case, which is double the value under the HXZ model. As with the size/momentum test portfolios, focusing on this test-asset metric would incorrectly give the impression that the six-factor model is inferior to the HXZ model. Figure 8 plots the Bayesian F-test results. The probability for the six-factor model here is greater than that in Figure 7 for the HXZ model when the excluded factors are incorporated (red line), but lower otherwise (blue line). Again, this shows that the absolute test results are in line with the model comparison analysis, both favoring the six-factor model, provided that the HXZ model is asked to price the value and momentum factors as well as the test assets.

[Figure 8]

We conclude this section with some additional observations about the Bayesian analysis. First, the probability for the models in Figures 7 and 8 rebounds from zero in each case and becomes substantial, approaching one (apparent for the blue line in Figure 7) as the Sharpe multiple and prior standard deviation for alpha get large. 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 may be even further from the values of alpha envisioned under the alternative when the Sharpe multiple 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).

The differing classical and Bayesian views about model validity that emerge in Figure 7 also deserve further comment. The p-value of 5% in the evaluation based solely on test assets would typically be interpreted as evidence against the null. However, the posterior probability (blue line) for the null is substantially higher than 5% for all priors, exceeding 50% for some more behavioral priors. This finding is consistent with Lindley's paradox: in sufficiently large samples, the posterior probability corresponding to a fixed p-value will be close to one, even if the p-value is small.[26] This divergence reflects a fundamental difference between posterior probabilities and p-values. Whereas the former reflects likelihoods under both the null and alternatives, the latter is a tail probability under the null that makes no reference to the distribution under the alternative. Nonetheless, p-values are often treated in practice *as if* they are posterior probabilities. The findings in Figure 7 serve as a reminder that this can lead to less than sensible conclusions and highlights what some perceive as an advantage of the Bayesian

[26] Intuitively, an alpha estimate with a t-statistic of two, say, will be close to zero when the sample is very large. The likelihood for $\alpha = 0$ will be quite high in this case, whereas the likelihood for alternative values that have substantial prior probability but are further from zero, will be much lower. As a result, the Bayes factor (ratio of marginal likelihoods) will strongly favor the null.

approach.[27]

Results for an Approximate Model

We have seen that for priors with moderate to fairly large Sharpe multiples, the evidence in Figure 8 favors an alternative statistical model over the restricted pricing model {Mkt SMB IA ROE UMD HML^{m}}, the "winner" in our model comparison contest. But perhaps the pricing model, nonetheless, provides 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*.[28] We address these issues in the Bayesian framework by modifying the null prior for a model to accommodate relatively small average deviations from the exact specification. The modified BF formula was given earlier in Proposition 1.

The black dotted line in Figure 9 (test assets 25 size/momentum portfolios) indicates extremely strong evidence against the exact null, $\alpha = 0$, as discussed earlier. The blue line in the figure shows the results of an analysis in which the prior under the *null* now assumes that $\sigma_{\alpha 0}$ = 1% (annualized). This allows for deviations from the exact version of {Mkt SMB ROE IA UMD HML^{m}} that have expected value of about 0.8% in magnitude. Such deviations give rise to a higher Sharpe ratio under the approximate null hypothesis, about 10% larger than that for the exact null. Thus, whereas the starting point previously was at a Sharpe ratio multiple of 1.0, the blue line now starts at a ratio just over 1.1. Not surprisingly, the posterior probabilities in Figure 9 for the less restrictive null model are higher than the probabilities obtained earlier for the exact model. There is no longer a "zero probability range" for the approximate null, but the model probabilities are still less than 0.5 over what we would consider the more relevant range of prior Sharpe multiples. Thus, the less restrictive alternative is still favored.

[27] Sample size is automatically incorporated in the BF. While it is sometimes recognized that the significance level in a classical test should be adjusted to reflect sample size, this can be difficult to operationalize and is generally ignored.

[28] In the case of exact models, BFs still provide an indication of the "relative success" of the models at predicting the data, e.g., Kass and Raftery (1995), or the "comparative support" the data provide for the models, e.g., Berger and Pericchi (1996).

To further explore the fit of the six-factor model, we increase $\sigma_{\alpha 0}$ to 1.5% (expected alpha about 1.2%), which corresponds to a Sharpe multiple just over 1.2. The probability for this level of approximation, shown by the black dashed curve, is now greater than 0.5 over most of the prior range. This sort of sensitivity analysis provides a computationally simple and conceptually appealing Bayesian complement to the descriptive statistics employed by Fama and French (2015a) to evaluate "goodness of fit" for a misspecified model. An advantage of this extension of the Bayesian framework over the conventional F test is that it allows more subtle and informative inferences to be obtained in situations where the sample size is large and models are routinely rejected at conventional levels as above or, e.g., in Fama and French (2015b).

6. Conclusion

We have derived 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 test that permits an analysis of Bayesian model comparison, i.e., the computation of model probabilities for the collection of all possible pricing models that are based on subsets of the given factors.

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) 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 have analyzed 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. The factors examined in Assness and Frazzini (2013), Hou, Xue and

Zhang (2015a,b) and Fama and French (2015a,b) have been studied in our preliminary empirical exploration, 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.

References

Ang and Chen, 2007, CAPM over the long run: 1926-2001, *Journal of Empirical Finance*, 14, 1-40.

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 and Jay Shanken, 2015, Which Alpha?, Working paper, Emory University.

Barillas, Francisco and Jay Shanken, 2015, Which Alpha?, Working paper, Emory University.

Berger J.O. and Pericchi L.R. 1996, The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*, 91, 433, p. 109-122.

Black, Fisher, Michael C. Jensen and Myron Scholes, 1972, The Capital Asset Pricing Model: Some Empirical Tests, in *Studies in the Theory of Capital markets*. Michael C. Jensen, ed. New York: Praeger, pp. 79-121

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.

De Moor, Lieven, Geert Dhaene and Piet Sercu, 2015, On Comparing Zero-alpha Tests across Multifactor Asset Pricing Models, *Journal of Banking & Finance*.

Dybvig, Phillip H, 1983, An explicit bound on individual assets' deviations form APT pricing in a finite economy, *Journal of Financial Economics* 12, 483-196.

Fama, Eugene F., 1996, Multifactor Portfolio Efficiency and Multifactor Asset Pricing, *Journal of Financial and Quantitative Analysis* 31, 441-65.

Fama, Eugene F., 1998, Determining the Number of Priced State Variables in the ICAPM, *Journal of Financial and Quantitative Analysis* 33, 217-31.

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, 2006, The Value Premium and the CAPM, *Journal of Finance* 61, 2163-2185.

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

Fama, Eugene F., And Kenneth R. French, 2015b, Dissecting anomalies with a five-factor model, Manuscript, Booth School of business, University of Chicago.

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

Gibbons, M., Ross, S., Shanken, J., 1989, A test of the efficiency of a given portfolio. *Econometrica* 57, 1121-1152.

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

Hou, Kewei, Xue Chen, and Zhang, Lu, 2015a, Digesting Anomalies: An Investment Approach, *Review of Financial Studies* 28, 650-705.

Hou, Kewei, Xue Chen, and Zhang, Lu, 2015b, A Comparison of New Factor Models, Working paper.

Jeffreys, H. 1961, *Theory of Probability*, 3rd ed. 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, D. and R. 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, 141-68.

MacKinlay, A. Craig, 1995 Multifactor models do not explain deviations form the Capital Asset pricing Models, *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: Second Edition, Wiley Series in Probability and Statistics.

Sharpe, F. William, 1964, Capital asset prices: a theory of market equilibrium under conditions of risk, *Journal of Finance,* 19, 425-442.

Shanken, Jay, 1987a, Multivariate proxies and asset pricing relations: Living with the Roll critique, *Journal of Financial Economics,* 18, 91-110.

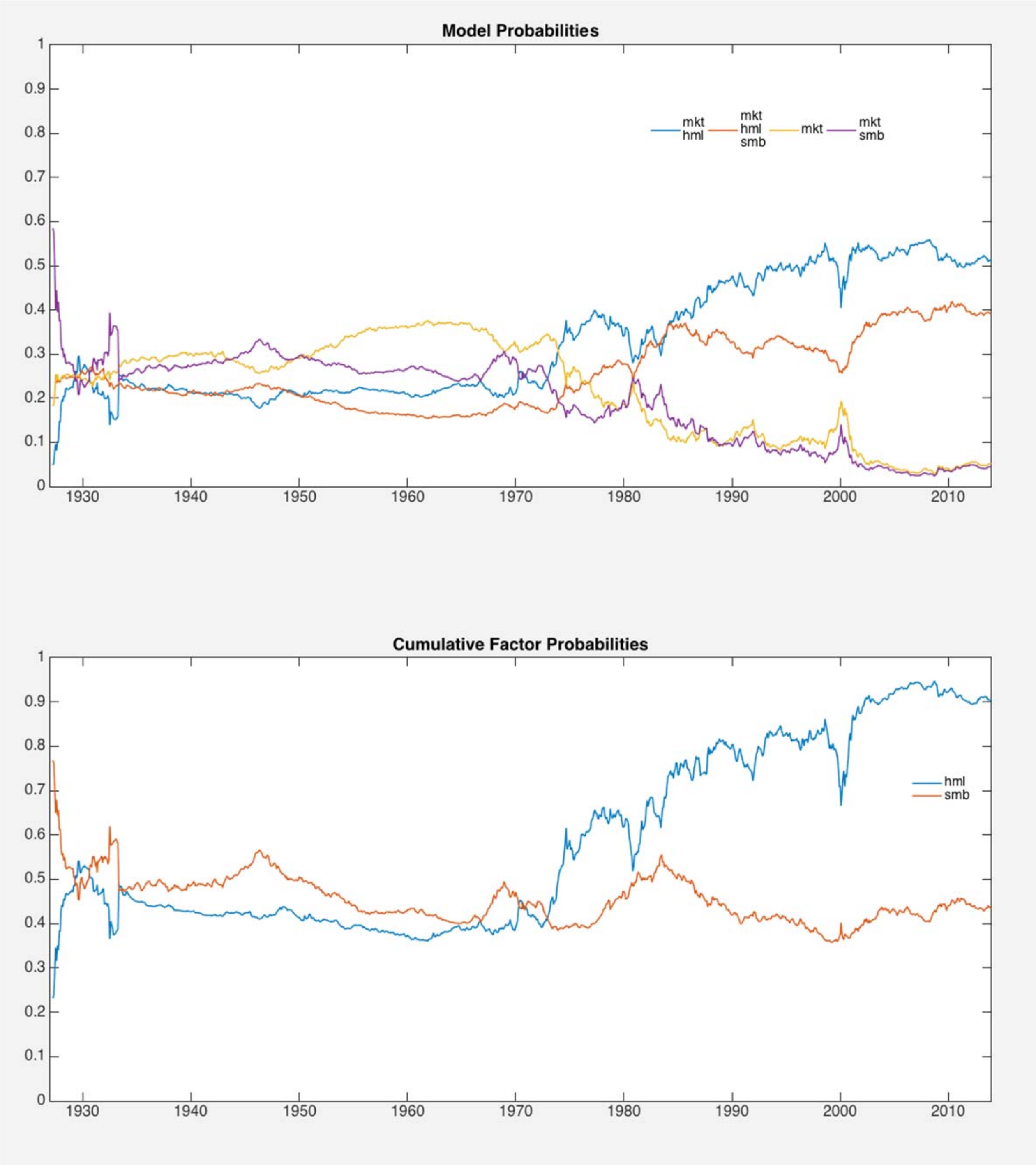
Shanken, Jay, 1987b, A Bayesian approach to testing portfolio efficiency, *Journal of Financial Economics,* 19, 195-216.

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, J. L. and F. 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, pp. 307-333.


Model Probabilities

Cumulative Factor Probabilities

Figure 1: Models based on the FF3 factors, sample 1927-2013, prior Sh_{max} = 1.27 x Sh(Mkt).

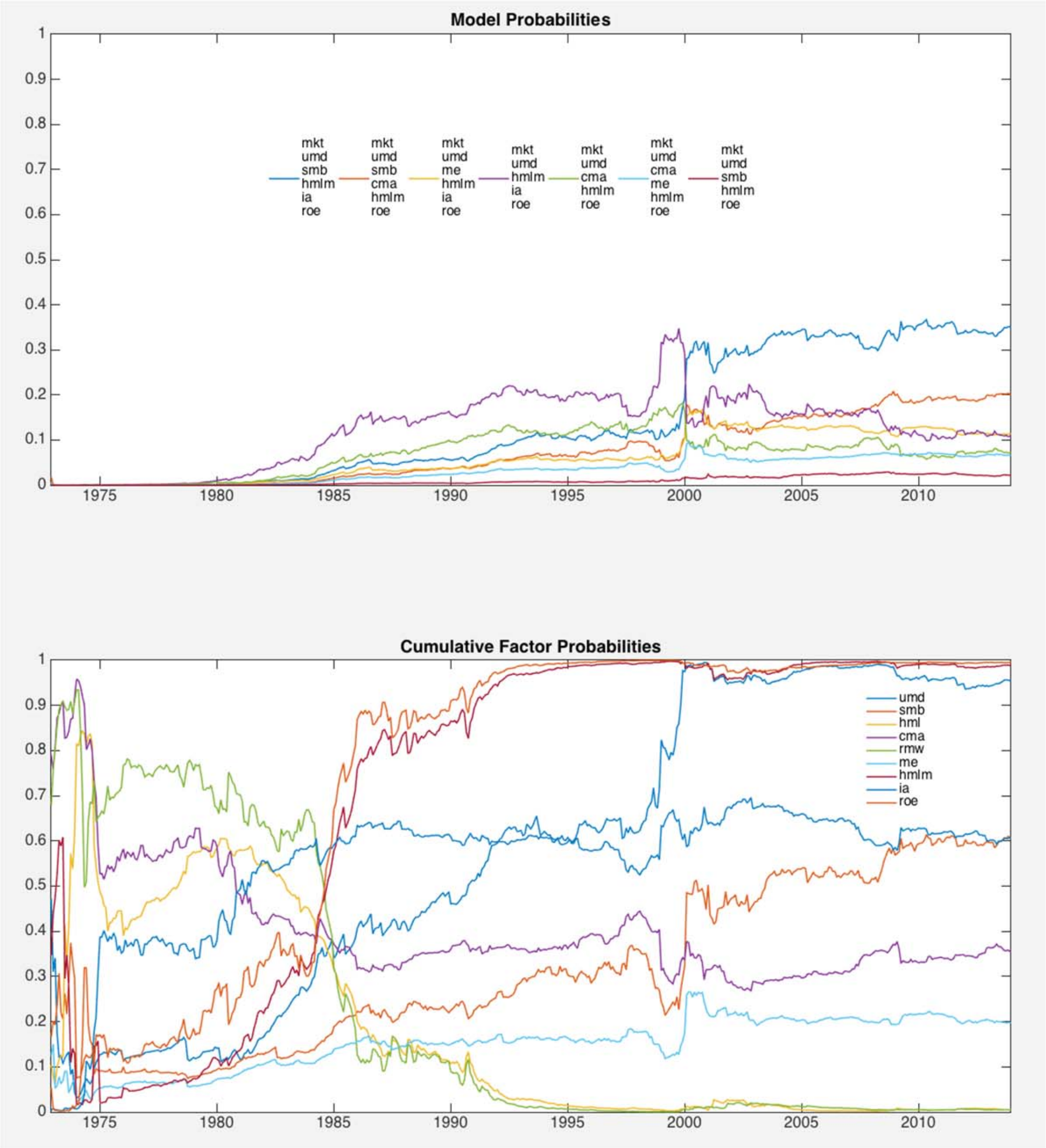


Figure 2: Models based on 10 prominent factors, sample 1972-2013, prior Sh_{max} = 1.5 x Sh(Mkt).

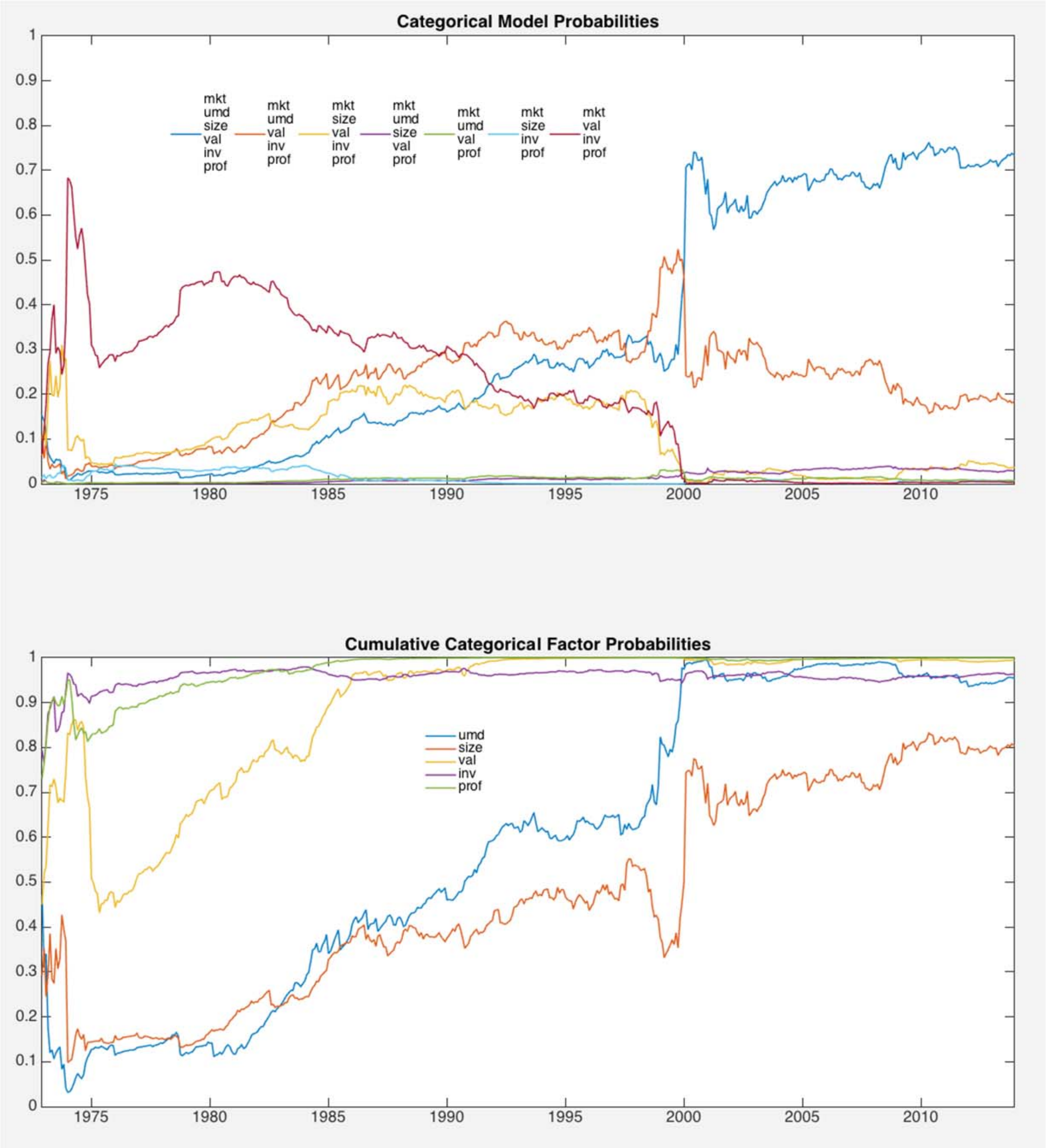


Figure 3: Categorical models based on 10 factors (including 2 versions of size, value, investment and profitability factors), sample 1972-2013, prior Sh_{max} = 1.5 x Sh(Mkt).

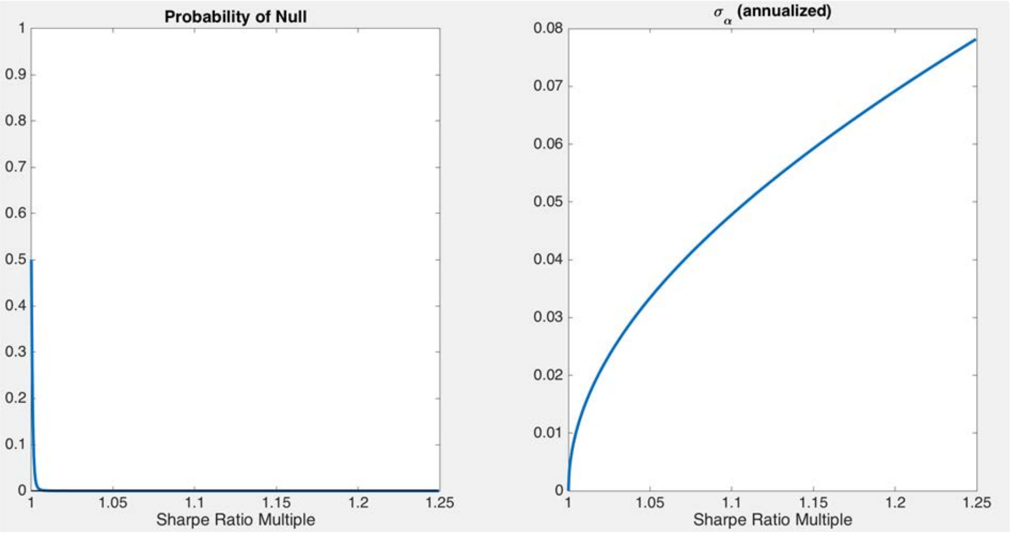


Figure 4: HML^m is not redundant in relation to the other factors {Mkt SMB ROE IA UMD} in the top model. Bayesian intercept test for HML^m. Sample 1972-2013. Horizontal axis: prior Sharpe ratio for the alternative as a multiple of the ratio under the null hypothesis (latter is 0.44 or 3.9 times the Market Sharpe ratio).

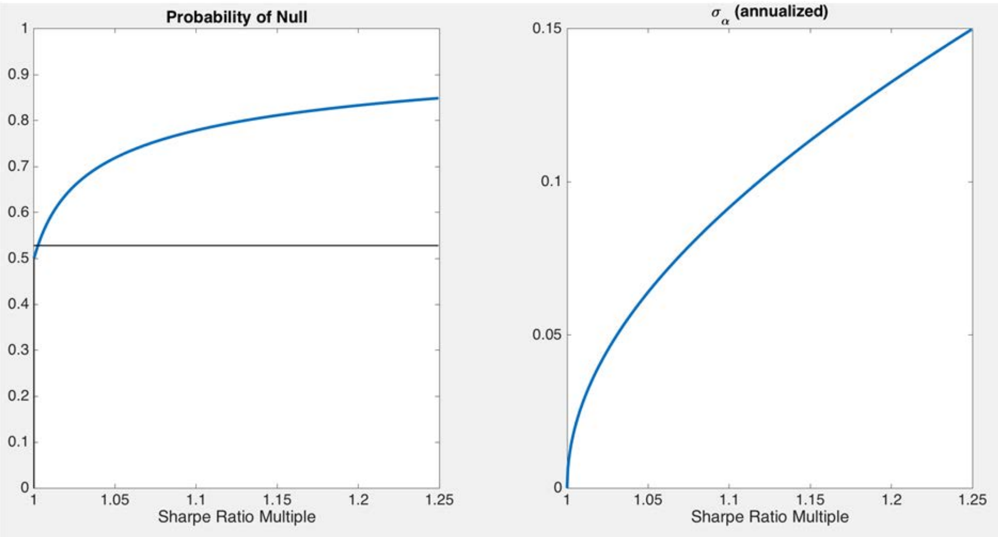


Figure 5: Bayesian intercept test for UMD on the model {Mkt SMB ROE IA}. Sample 1972-2013. Horizontal axis: prior Sharpe ratio for the alternative as a multiple of the ratio under the null hypothesis (latter is 0.44 or 3.9 times the Market Sharpe ratio).

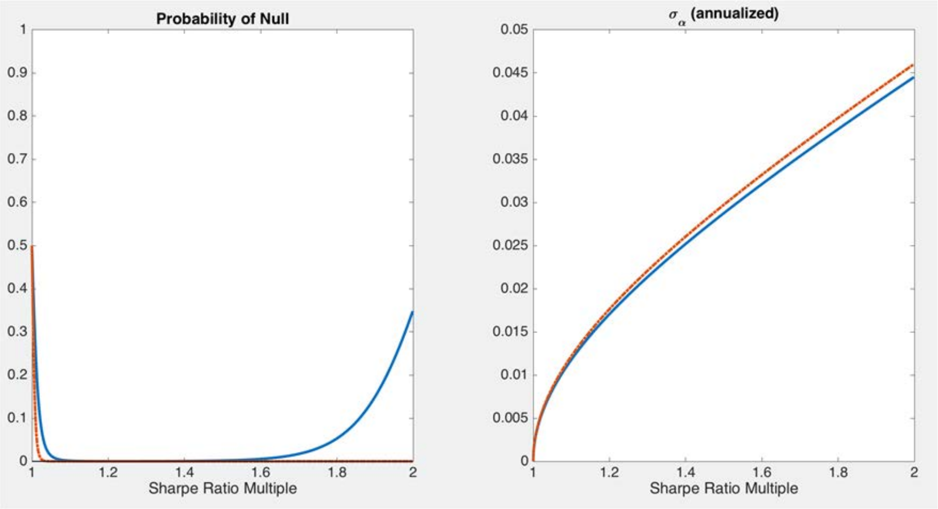


Figure 6: Sample 1972-2013, Model = {Mkt ME IA ROE}. Test assets = 25 size-momentum portfolios (blue line) plus UMD, HML^m (red line). Sample 1972-2013. Horizontal axis: prior Sharpe ratio for the alternative as a multiple of the ratio under the null hypothesis (latter is 0.43 or 3.8 times the Market Sharpe ratio).

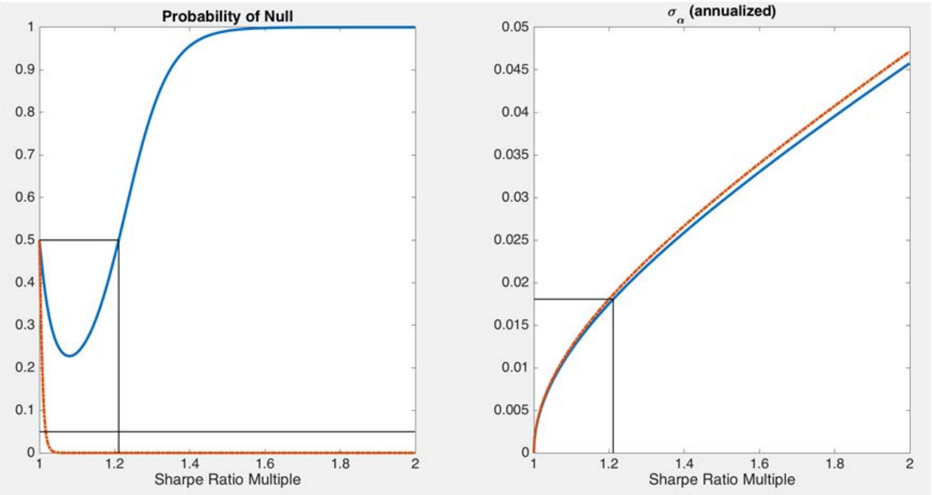


Figure 7: Sample 1972-2013, Model = {Mkt ME IA ROE}. Test assets = 25 Book-to-market/investment portfolios (blue line) plus UMD, HML^m (red line). Horizontal line shows the conventional p-value. Horizontal axis: prior Sharpe ratio for the alternative as a multiple of the ratio under the null hypothesis (latter is 0.43 or 3.8 times the Market Sharpe ratio).

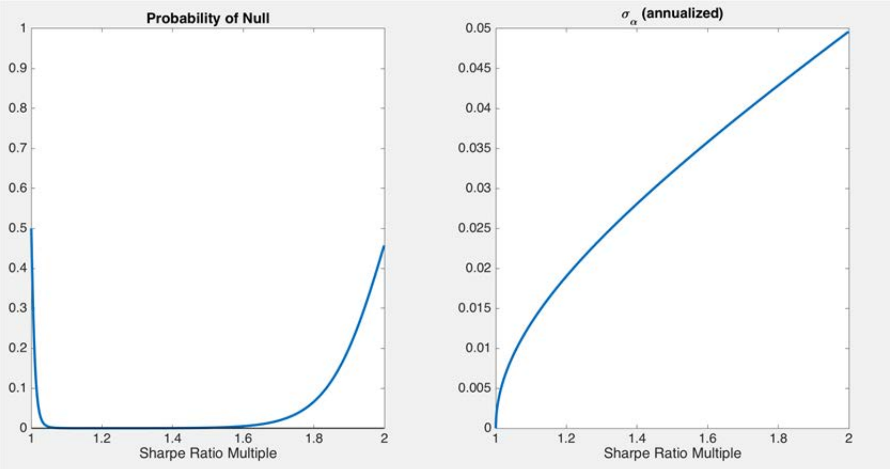


Figure 8: Sample 1972-2013, Model = {Mkt SMB IA ROE UMD HML^{m}}. Test assets: 25 Book-to-market/investment portfolios. Horizontal axis: prior Sharpe ratio for the alternative as a multiple of the ratio under the null hypothesis (latter is 0.50 or 4.8 times the Market Sharpe ratio).

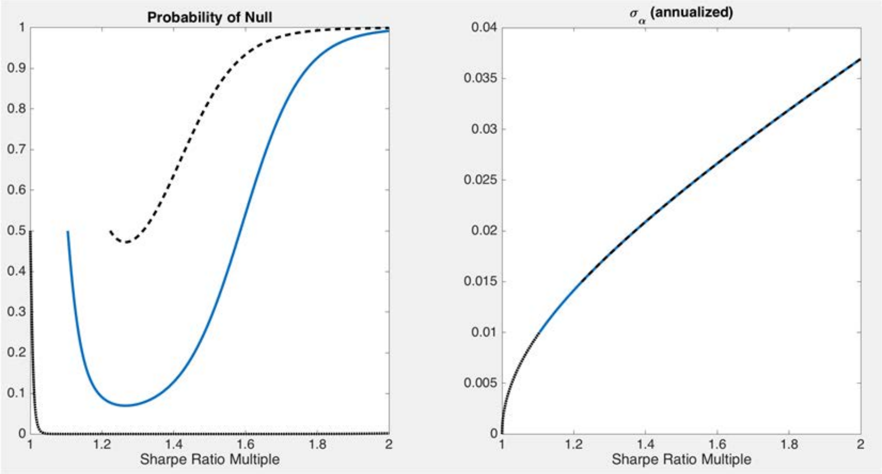


Figure 9: Sample 1972-2013, Model = {Mkt SMB IA ROE UMD HML^{m}}, Test assets = 25 size-momentum portfolios. Exact model (black dotted). $\sigma_{\alpha0}$ = 1% (blue line) or $\sigma_{\alpha0}$ =1.5% (black dashed) under the approximate null hypothesis. Horizontal axis: prior Sharpe ratio for the alternative as a multiple of the ratio under the exact null hypothesis (latter is 0.51 or 4.5 times the Market Sharpe ratio).

Table 1

Fama-French factors - Posterior Model Probabilities for different Prior Sharpe Multiples

Data Sample: Jan 1927 to Dec 2013
Market Sharpe Ratio = 0.115
3-factor Sharpe Ratio = 0.142 or 1.23*Market Sharpe Ratio

Model/Prior Multiple[*]	**1.13**	**1.27**	**1.58**	**2.24**
Mkt HML	44.9	51.3	58.6	65.3
Mkt HML SMB	42.2	39.1	32.1	23.1
Mkt	6.6	5.2	5.7	8.1
Mkt SMB	6.3	4.5	3.6	3.4

[*] Multiple of Mkt Sharpe ratio under 3-factor alternative.

Table 2

10 factors - Posterior Model Probabilities for different Prior Sharpe Multiples

Data Sample: Jan 1972 to Dec 2013
Market Sharpe Ratio = 0.113
6-factor (best model) Sharpe Ratio = 0.51 or 4.5*Market Sharpe Ratio

Model/Prior Multiple[*]	**1.25**	**1.5**	**2**	**3**
Mkt SMB ROE IA HML^m UMD	21.9	35.1	42.6	45.3
Mkt SMB ROE CMA HML^m UMD	12.6	20.1	24.2	25.7
Mkt ME ROE IA HML^m UMD	9.1	11.4	10.6	9.2
Mkt ROE IA HML^m UMD	14.6	10.9	6.6	5.0
Mkt ROE CMA HML^m UMD	9.0	7.2	4.7	3.7
Mkt ME ROE CMA HML^m UMD	0.2	6.6	6.1	5.3
Mkt SMB ROE IA HML^m	3.8	1.6	0.6	2.8

[*] Multiple of Mkt Sharpe ratio under 6-factor alternative.

Table 3: Categorical Models - Posterior Probabilities for different Prior Sharpe Multiples

Data Sample: Jan 1972 to Dec 2013
Market Sharpe Ratio = 0.113
6-factor (best model) Sharpe Ratio = 0.51 or 4.5*Market Sharpe Ratio

Model/Prior Multiple*	**1.25**	**1.5**	**2**	**3**
Mkt SIZE PROF INV VAL MOM	50.8	73.4	83.6	85.6
Mkt PROF INV VAL MOM	25.2	18.3	11.3	8.8
Mkt SIZE PROF INV VAL	11.6	3.6	1.1	5.0
Mkt SIZE PROF VAL MOM	2.0	2.9	3.5	4.6
Mkt PROF VAL MOM	1.0	0.7	0.5	0.5
Mkt SIZE PROF INV	3.4	0.5	0.1	0.0
Mkt PROF INV VAL	3.3	0.3	0.2	0.0

Table 4: Relative Probabilities for each Categorical Factor in the 10-factor Analysis

Factor/Prior Multiple[*]	**1.25**	**1.5**	**2**	**3**
ROE	96.0	99.5	99.9	100
IA	63.6	63.1	63.1	63.2
SMB	70.1	75.4	80.0	83.0
HML^m	93.9	99.5	99.9	100

[*] Multiple of Mkt Sharpe ratio under 6-factor alternative.

The remaining PROF, INV, SIZE and VAL probability goes to RMW, CMA, ME and HML, respectively.

Table 5

Intercepts for each Factor in the Highest-Probability Model on the other Five-factors

This table presents annualized alphas from regressions of each factor on the other factors in the model {Mkt SMB ROE IA UMD HML^m}. Sample period Jan 1972 to Dec 2013.

Factor	SMB	ROE	IA	UMD	HML^m
Alpha (t-statistic)	5.09 (3.14)	6.97 (6.08)	1.20 (1.50)	6.60 (3.96)	6.07 (5.26)