Factor Model Selection Using the ICAPM*

Paul Glasserman† Jeremias Huber‡ Harry Mamaysky§

January 4, 2026

Abstract

Using market data together with household-level consumption data, we introduce a quantitative measure of the ICAPM sign-consistency property of Maio and Santa-Clara (2012) for factor models. We find that more consistent factor models have less persistent alphas, and more stable betas and out-of-sample mean squared errors. We propose a novel statistical test for the sign of the consistency-stability relationship across many factor models and over time. Our methodology allows for the efficient identification of the historically most ICAPM-consistent factor models and factors from a large candidate pool. Our results suggest that historically consistent models are likely to be stable in the future.

JEL classification: G12, G50

Keywords: ICAPM, consumption-based asset pricing, factor models

Online Appendix: `https://sites.google.com/view/hmamaysky`

*We are grateful to Paulo Maio for his detailed comments about an earlier draft of this paper. We also thank Geert Bekaert for his comments and suggestions.

†Columbia Business School, pg20@columbia.edu.

‡International Finance Corporation, jhuber@ifc.org.

§Columbia Business School, hm2646@columbia.edu.

1 Introduction

In this paper, we introduce a robust statistical test for evaluating factor model consistency with the Merton (1973) ICAPM. Our work builds on the insight of Maio and Santa-Clara (2012) (MSC) that factors which positively predict market returns should have positive cross-sectional risk prices. MSC propose a binary test of model fit: Either a model fits their forecasting and risk price sign prediction across all model factors or it doesn't. Our view is that, for several reasons, this evaluation should instead be a matter of degree. First, we argue that the MSC sign property is a desirable feature of a model, but not, strictly speaking, a restriction imposed by the ICAPM. Second, we note that a factor's consistency with the sign property may depend on which other factors are included in a model, the choice of forecasting horizons, and the choice of test assets used to estimate risk prices. These considerations lead us to formulate measures of the degree of consistency with the MSC sign property. Our main findings show that these measures provide valuable information about a model's performance.

Our method evaluates the degree of ICAPM consistency by averaging—across all of a model's factors—the product of each factor's t-statistic in a forecasting regression with the t-statistic associated with that factor's price of risk from a Fama and MacBeth (1973) analysis. The product of t-statistics will be large and positive when the two coefficients are significant and have the same sign; it will be large and negative when the two coefficients are significant and have opposite signs; and it will be close to zero if either or both coefficients fail to be significant. We apply this method using both market data, as in Maio and Santa-Clara (2012), and household-level consumption data, which is better motivated theoretically, and we show that both consistency measures are "useful."

We measure a factor model's usefulness for investors along three dimensions. First, we ask whether alphas relative to a factor model are persistent. In a factor model that correctly prices all assets, alphas should reflect temporary deviations from the model's risk-based return forecast, and should thus not be persistent. Second, we ask whether betas relative to a factor model are persistent. In a factor model that properly captures assets' risk loadings, betas should be stable over time. Finally, we ask whether using betas relative to a factor model produces a good hedge to assets' systematic risk exposures. Again, a properly specified factor model should yield a good hedging tool, which we measure via a factor model's ratio of in-sample to out-of-sample mean squared errors (MSE ratio).

We refer to models that do well across these three dimensions as stable. Using the original eight MSC factor models and all three- and four-factor models which can be

formed from the market in combination with the Jensen et al. (2023) 13 factor themes, we show that more ICAPM consistent models are more stable.

To measure model consistency, we first run a set of forecasting regressions. These regressions test the ability of a model's state variables to forecast future stock market index returns, index return variance, consumption growth, and the variance of cross-sectional consumption growth. We then run a Fama and MacBeth (1973) analysis to estimate each factor's risk premium relative to a specific set of test assets. These analyses are run for every year of our sample using expanding windows. We then multiply the t-statistics from the forecasting regressions with the t-statistics from the price of risk analysis. Large positive values for these products indicate good alignment between the signs of the estimates, as required for the Maio and Santa-Clara (2012) ICAPM-consistency property. Negative values indicate an inconsistent factor model.[1] We refer to the average of t-statistic products as the model consistency index, or MCI.

Then, for year t, we evaluate the stability of factor models' alphas, betas, hedging quality (measured via MSE ratios), and risk prices between years t and $t+1$. These are obtained relative to the same set of test assets used in the MCI calculation. For all factor models in a given set, we run a regression of a factor model's alpha, beta, hedging, and price-of-risk stability on their MCIs. In year t, the sign of the coefficient from this cross-sectional regression determines the relationship between ICAPM consistency and model stability. Over time, the signs of these cross-sectional regression coefficients indicate whether factor model MCI and factor model stability have a positive, negative, or negligible relationship.

This analysis reveals that the alpha coefficients (i.e., from all the rolling cross-sectional regressions of alpha stability on model consistency) are generally negative, that the beta and MSE ratio coefficients are generally positive, and that the price of risk coefficients are on average close to zero. Therefore, more consistent models have less persistent alphas, more stable betas and better hedging properties, but model consistency seems largely unrelated to future factor risk prices.

We propose a novel statistical test of the above findings. The test sums a sign indicator (+1 for positive and significant, -1 for negative and significant, and 0 otherwise) for each of the rolling cross-sectional coefficients for a given outcome variable, e.g., for alpha stability versus MCIs. We then create a simulated distribution for this sum under the null of no relationship between consistency and stability, while allowing for correlation across the rolling cross-sectional regression coefficients. The test confirms that our results for alphas,

[1]The sign test is reversed when considering forecasts of return and consumption growth variance.

betas, and MSE ratios are highly statistically significant, while the price-of-risk results are not.

Our methodology allows investors to track which factor models have performed well over time, which factors have been frequently present in the most ICAPM-consistent models, and which factor models are presently the most ICAPM-consistent. Of the factor models analyzed in Maio and Santa-Clara (2012), the historically most consistent ones have been the Fama and French (1992) and Carhart (1997) models. Of the 13 factor themes in the Jensen et al. (2023) data, the ones that are most frequently present in consistent factor models over time are low leverage, profit growth, and profitability. When looking at the most consistent factor models using the most recent data available in our study, the most prevalent factor themes are, again, profit growth and profitability. Prior work, reviewed in Novy-Marx and Medhat (2025), shows the significance of profitability for the cross section of returns. Our results complement prior work by arriving at these factors through a very different approach to factor selection.

Based on the historical relationship between model consistency and future stability, our analysis leads to the normative implication that investors using factor models in practice will be well served by selecting models which are ICAPM consistent. Our methodology further allows for the efficient identification of the small number of ICAPM-consistent factor models and factors from a large pool of candidates.

1.1 Literature Review

Our work builds on Maio and Santa-Clara (2012), who test the consistency of factor models with respect to the ICAPM by examining sign consistency between cross-sectional risk premia of factors and predictive forecasting coefficients of their respective state variables. Lin (2022) applies their method to show that economic uncertainty does not satisfy the MSC restriction, while Cooper and Maio (2018) show that asset growth and investment factors are consistent with ICAPM restrictions. Work by Boons (2016) and Barroso et al. (2021) is supportive of the ICAPM as a testing framework for factor models and finds that time-variation of risk premia is driven by changes in the predictive relationship between state variables and future macroeconomic outcomes, such as consumption growth, for a similar set of state variables as in Maio and Santa-Clara (2012). Chabi-Yo et al. (2025) use the ICAPM directly to construct a risk factor model, while Gonçalves (2021) finds a downward-sloping equity term structure at long maturities to be consistent with the ICAPM. Cooper et al. (2021), on the other hand, suggest testing factor models with regard to their consistency relative to the arbitrage pricing theory of Ross (1976).

The distance metric $\|\mathrm{E}[MR] - 1\|$ of Hansen and Jagannathan (1997) can be understood as a measure of how consistent a pricing model (given by the SDF M) is with the no-arbitrage requirement that discounted gross returns should equal 1. Examples of such tests of asset pricing models include Kan and Robotti (2008) and Li et al. (2010).[2] Another example of consistency tests is the GRS statistic, proposed in Gibbons et al. (1989), which evaluates whether the alphas of a set of test assets are jointly zero relative to a factor model, a condition that should be satisfied by the "true" factor model. Related tests have been proposed by Barillas and Shanken (2017), who advocate evaluating factor models by their ability to produce small pricing errors, in the GRS sense, for competing factor models, and Harvey and Liu (2021) who discuss important implentation issues in conducting alpha-based tests. Lastly, Feng et al. (2020) propose a test of individual factors based on their marginal contribution to a high-dimensional SDF. While all these tests rely on cross-sectional model restrictions, our proposed model consistency measure combines cross-sectional (price of risk) with time-series (forecasting coefficient) information.

Several recent papers evaluate out-of-sample Sharpe ratios of SDF portfolios, e.g., Didisheim et al. (2025), Bryzgalova et al. (2022), Kozak et al. (2018, 2020). Jacquier et al. (2025) and Kan et al. (2024) propose refinements to testing factor models based on out-of-sample Sharpe ratios relative to in-sample optimal Sharpe ratios. One problem with these tests is that optimal Sharpe ratio portfolios are often not implementable in practice due to extreme positions, often in illiquid assets.

Another measure of model usefulness examines the out-of-sample R^2 of a factor model, either in its cross-sectional (see, e.g., Cooper et al. 2021) or in its predictive variant (see, e.g., Kelly et al. 2019). Kan et al. (2013) derive a test for the difference in in-sample in R^2s between factor models as a way of comparing their usefulness. Other work examines factor models' root mean squared pricing errors, mean absolute pricing errors, or unexplained idiosyncratic variation (see, e.g., Bryzgalova et al. 2022 and Lettau and Pelger 2020). Relying on in-sample fit to predict out-of sample properties assumes a high degree of persistence that may not be economically motivated. Our model consistency measure, on the other hand, is theoretically motivated and is associated with lower out-of-sample MSEs. For a review of factor model testing procedures we refer the reader to Giglio et al. (2022) and Kelly et al. (2019).

Notably, our approach can be used to evaluate the ICAPM consistency of machine-learning (ML) based pricing models, as well as of traditional factor models. If the stochas-

[2]Didisheim et al. (2025) propose using consistency measures, e.g., the $\mathrm{E}[MR] = 1$ condition, to evaluate pricing models ex-post.

tic discount factor can be written as

$$M_{t+1} = a_t - b^\top R_{t+1} \tag{1}$$

for some K-dimensional set of factor returns R_{t+1}, we can compute the MCI of the proposed set of factors. As Didisheim et al. (2025) point out, many neural-network based pricing models can be expressed as in (1) with R_{t+1} the return of a set of "characteristic-managed portfolio(s)." Because our procedures relies only on the time-series of factor returns and not on the underlying sorting variable (see Section 4.2), it can be applied with no adjustment to evaluate the consistency of ML-based asset pricing models.

2 Is There an ICAPM Sign Restriction?

Maio and Santa-Clara (2012) put forward restrictions on the time-series and cross-sectional properties of a factor model that should hold for the model to be consistent with the ICAPM of Merton (1973). The main restriction they propose is that if a state variable positively (negatively) forecasts investment opportunities, then its innovations should earn a positive (negative) price of risk in the cross section.

The intuition underlying this condition works as follows. If a state variable positively forecasts returns, then assets that negatively covary with the state variable hedge reinvestment risk; investor hedging demand bids up their prices, so these assets should carry a negative risk premium in the cross section. Similarly, assets that positively covary with the state variable should earn a positive risk premium. This reasoning links the sign of the forecasting relation with the sign of the risk price. A closely related point is made in Merton (1973), bottom of p. 876. The argument of MSC is economically appealing, and we will rely on this insight in proposing measures of model consistency. However, we question whether the restriction proposed by MSC is, strictly speaking, implied by the ICAPM and how it should be applied in practice. We argue that consistency with the MSC sign property is best viewed as a matter of degree, rather than as a test to pass or fail.

Theoretical argument. Following Cochrane (2009), MSC approximate the continuous-time ICAPM pricing equation in discrete time as

$$\mathrm{E}_t\left(R_{i,t+1} - R_{f,t+1}\right) = \gamma \,\mathrm{Cov}_t\left(R_{i,t+1}, R_{m,t+1}\right) + \gamma_z \,\mathrm{Cov}_t\left(R_{i,t+1}, \Delta z_{t+1}\right), \tag{2}$$

where $R_{i,t+1}$ denotes the return on an asset, $R_{f,t+1}$ the risk-free rate (known at time t),

$R_{m,t+1}$ the return on the market, and Δz_{t+1} the innovation or change in an ICAPM state variable, all measured from time t to $t+1$. Here, γ is the risk price associated with the market, and γ_z is the risk price associated with the state variable.

MSC argue that the ICAPM restricts the sign of γ_z as follows. Suppose the state variable is positively correlated with future aggregate returns, so

$$\mathrm{Cov}_t\left(R_{m,t+2}, \Delta z_{t+1}\right) > 0; \tag{3}$$

suppose asset i is positively correlated with the state variable, so

$$\mathrm{Cov}_t\left(R_{i,t+1}, \Delta z_{t+1}\right) > 0. \tag{4}$$

MSC assert that (3) and (4) imply

$$\mathrm{Cov}_t\left(R_{i,t+1}, R_{m,t+2}\right) > 0. \tag{5}$$

They then argue that under (5), asset i does not provide a hedge for reinvestment risk and must therefore earn a positive risk premium in (2), meaning that

$$\gamma_z \,\mathrm{Cov}_t\left(R_{i,t+1}, \Delta z_{t+1}\right) > 0. \tag{6}$$

But then (4) implies $\gamma_z > 0$. Through this argument, the sign of the forecasting relation in (3) determines the sign of γ_z.

The problem with this argument is that (3) and (4) do not imply (5): two random variables can be positively correlated with each other and yet have correlations of opposite signs with a third random variable.[3] It is therefore possible to maintain the assumptions in (3)–(4) while leaving the sign of γ_z undetermined.[4] [5]

A modified argument avoids this gap and establishes the sign-consistency property in the special case of a *traded* factor. Suppose that (3) holds for a traded factor z_t. As $R_{i,t+1}$, take the return on z_t itself. In this case, (5) holds because it restates (3). Given (5), the original economic argument of Maio and Santa-Clara (2012) implies (6). We can now

[3]For example, suppose $\mathrm{E}[Y] = \mathrm{E}[Z] = \mathrm{E}[YZ] = 0$ and $\mathrm{E}[Y^2] > \mathrm{E}[Z^2] > 0$. Take $X_1 = Y + Z$ and $X_2 = Y - Z$. Then $\mathrm{E}[X_1 X_2] = \mathrm{E}[Y^2] - \mathrm{E}[Z^2] > 0$ and $\mathrm{E}[X_1 Z] > 0$, but $\mathrm{E}[X_2 Z] < 0$.

[4]This issue is not an artifact of the discrete-time approximation (2); a similar argument can be formulated in continuous time.

[5]Maio and Santa-Clara (2012) also propose a sign restriction for state variables that positively forecast market *volatility*, replacing $R_{m,t+2}$ with $R^2_{m,t+2}$ in (3). In this case, an asset satisfying (4) would hedge increased reinvestment volatility and should therefore carry a negative risk premium. Thus, Maio and Santa-Clara (2012) propose that $\mathrm{Cov}_t\left(R^2_{m,t+2}, \Delta z_{t+1}\right) > 0$ implies $\gamma_z < 0$.

correctly conclude that $\gamma_z > 0$ because $\text{Cov}_t\left(R_{i,t+1}, \Delta z_{t+1}\right) > 0$ when $R_{i,t+1}$ is the return on the factor. In our empirical analysis, we work with traded factors, except in revisiting specific models from Maio and Santa-Clara (2012).[6]

Multiple state variables. The theoretical difficulty with the original argument of Maio and Santa-Clara (2012) becomes more acute with multiple state variables, in which case Δz_{t+1} and γ_z are vectors. We may suppose that (3) holds for every component of Δz_{t+1}; if a state variable was negatively correlated with future returns, we could change its sign.

Consider the simplest case — two uncorrelated state variables. Suppose $R_{i,t+1}$ loads positively on one state variable, as in (4), and negatively on the other. As both state variables positively forecast future market returns, it is unclear if the asset should earn a positive or negative risk premium. We therefore cannot determine the sign in (6) (which is now a dot product of two vectors), much less the signs of the components of γ_z.

Further complications arise with correlated state variables. One can orthogonalize the state variables, but the transformed variables then depend on which variables are included in the first place. This points to a further issue: the sign consistency property proposed by Maio and Santa-Clara (2012) cannot be thought of as a property of an individual state variable — it relies on the set of state variables specified by a model. Moreover, the model must claim to have identified all ICAPM state variables, or at least that any other state variables are orthogonal to those included. Otherwise, omitted state variables could change the signs of the risk prices on included state variables.

Consumption versus market returns. The Maio and Santa-Clara (2012) sign condition asserts that the positive covariance between the state variable and future market returns in (3) implies a positive risk price for the state variable. In Merton (1973) and Campbell (1993), the risk price depends fundamentally on the relationship between the state variable and consumption, not the market; market returns enter as a proxy for changes in aggregate wealth. We will therefore investigate sign consistency with respect to both the market and consumption. In addition, following Maio and Santa-Clara (2012), we explore a version of (3) where we replace the market return with the future variance of market returns. Finally, we investigate a version of (3) where the market return is replaced with the cross-sectional variance of consumption growth. This is motivated by the incomplete risk sharing model of Constantinides and Duffie (1996), which suggests that the covariance of returns with the variance of consumption growth commands a negative risk premium.

Choice of horizon. Merton (1973) is formulated in continuous time. The precise counter-

[6]In Online Appendix Section A1 we also examine the empirical frequency with which (3) and (4) are associated with (5), and find some empirical support for this claim.

part to the sign of the forecasting relation in (3) is the sign of the derivative of optimal consumption with respect to the state variable; see equation (18) in Merton (1973). An empirical implementation requires a discrete-time formulation, as in (2). Like MSC, we consider horizons of 1, 12, and 60 months. We note, however, that longer horizons take us further from the theoretical ICAPM and thus further from what could be considered restrictions imposed by the ICAPM.

Sign consistency as a matter of degree. In light of the foregoing considerations, we do not see the sign consistency property proposed by MSC as a restriction imposed by the ICAPM. We do, however, see it as an economically appealing property — a property that may indicate a "better" model. Our goal is therefore not to accept or reject models, as MSC do, but rather to investigate ways in which greater alignment with the sign consistency property leads to better model performance. To this end, we treat sign consistency as a matter of degree, rather than a binary outcome. We develop a way to measure the degree of sign consistency exhibited by a model, and we document benefits to model performance associated with higher degrees of sign consistency.

Carrying this out requires a systemic approach to measuring consistency. MSC reject some models based on the signs of estimated risk prices that are not statistically significant; we prefer to incorporate the degree of significance in our measure of consistency. MSC estimate forecasting regressions at multiple horizons but use only the longest horizon to determine the signs of forecasting coefficients; we prefer to combine results over multiple horizons, again to measure the extent of consistency. We also consider aggregate measures that combine market-based and consumption-based forecasts, and forecasts of volatilities as well as levels. Several models are rejected in MSC based on a separate consideration of the magnitude of the market risk prices γ in (2); we focus on the sign-consistency property and do not consider the magnitude of γ in our consistency measures. Furthermore, we measure consistency for multifactor models and not for isolated factors. We do this by including all of a model's factors in running forecasting regressions and estimating risk prices, which is consistent with MSC (see their footnote 11).

3 Data

We use consumption data from the Consumer Expenditure Survey (CES) collected by the U.S. Bureau of Labor Statistics. Following Brav et al. (2002) and Malloy et al. (2009), our consumption definition is constrained to non-durable goods and services for stockholder households, and we follow their outlined procedure as closely as possible. The

resulting time series contains monthly observations of stockholder consumption growth defined relative to three months prior, which reflects the interval between household survey interviews. The consumption data covers the time period from March 1984 to December 2020, though no data are available for March, April, and May 1986 as well as March, April, and May 1996. Figure 1 shows quarter-over-quarter stockholder consumption growth and the cross-sectional variance of consumption growth in each month. More details on series construction can be found in Section A2 of the Online Appendix.

We use the factor and test asset data used in Maio and Santa-Clara (2012), though our sample is longer, running from July 1963 to December 2023.[7] The non-traded MSC state variables include the term spread ($TERM$), defined as the yield spread between ten-year and one-year Treasury bonds, and the default spread (DEF), calculated as the difference in yield between BAA and AAA corporate bonds. These are both obtained from the St. Louis Fed's FRED website. We download data used to construct the dividend yield (DY) and the price-to-earnings ratios (PE) from Robert Shiller's website.[8] We use the value spread (VS) measure suggested by Campbell and Vuolteenaho (2004) which calculates the difference between the log of book-to-market ratios of small-value and small-growth portfolios. We also use the one-month Treasury bill as the risk-free rate (RF). These data are obtained from Ken French's website. In addition, we construct the Cochrane and Piazzesi (2005) bond market risk factor, which is used as a state variable in Koijen et al. (2017).

The second set of MSC factors consists of long-short portfolios, including the excess return of the CRSP value-weighted U.S. market index, the returns on the size (SMB), value (HML), and momentum (UMD) factors.[9] We obtain the returns of the liquidity factor (L) used in the factor model of Pástor and Stambaugh (2003) from Ľubŏs Pástor's website. As a proxy for the market return variance we use $SVAR$, the cumulative sum of squared daily returns of the S&P 500 from Amit Goyal's website; see Welch and Goyal (2007). We follow MSC and use the 25 size and book-to-market sorted portfolios as one set of test assets, and the 25 size and momentum sorted portfolios as the other set of test

[7]The factor models considered in Maio and Santa-Clara (2012) are Carhart (1997), Fama and French (1992), Fama and French (1993), Campbell and Vuolteenaho (2004), Hahn and Lee (2006), Koijen et al. (2017), Petkova (2006), and Pástor and Stambaugh (2003). We refer to these as C, FF3, FF5, CV, HL, KLVN, P, and PS, respectively. The complete set of factors underlying these models are listed in Table 1. For the specific factors in each model, see Section A3 of the Online Appendix.

[8]DY is computed using the sum of annual dividends, while PE uses the moving average over ten years for earnings of the S&P 500 index, both normalized by the level of the S&P 500 and in logarithmic form.

[9]In Table 1, SMB^* and HML^* refer to transformations introduced by Maio and Santa-Clara (2012), and $CUMD$ and CL are cumulative 60-month returns on UMD and L. These transformations are explained in Section 4.2.

assets. When testing the MSC factor models, we use monthly returns.

Additionally, we use the 13 factor theme portfolios in Jensen et al. (2023) which we refer to as JKP factors and which are based on the following themes: accruals, debt issuance, investment, low leverage, low risk, momentum, profit growth, profitability, quality, seasonality, short-term reversal, size and value. Table 1, Panel B, describes those 13 factor themes. These factors are obtained from the `https://jkpfactors.com/` website and are based on 153 characteristics and factors which the authors then clustered into factor themes. Unlike the MSC data, which are not uniformly available on a daily frequency, the JKP data contains both monthly and daily factor returns. When testing factor models using the JKP data, we use daily returns as this increases the precision of our estimates. Repeating our analysis using monthly JKP returns yields very similar results.

4 Model Consistency Indexes

As discussed in Section 2, we want to quantify the degree to which a model is consistent with the MSC sign property. We introduce a model consistency index for this purpose.

Our model consistency calculation involves three steps. First, we estimate the price of risk for every factor in a factor model. Then we estimate the factors' forecasting coefficients for market or consumption outcomes at different horizons. Finally, we measure the degree to which the signs of the risk prices and forecasting coefficients are consistent. We view sign consistency as a property of a model, not an isolated state variable. We therefore estimate forecasting coefficients and risk prices through multivariate regressions with all of a model's factors included. The signs of these parameters may depend on which factors are in the model.

These calculations are done annually using expanding windows, with a minimum window length of seven full years across all factors and factor-derived forecasting state variables.[10] For estimates using only market data, we have 49 yearly observations going from 1975 to 2023. Estimates that use consumption data are based on 31 yearly observations from 1990 to 2020. We are careful to use no forward-looking information in this analysis, so all year t metrics would have been known to investors as of the end of year t.

[10]Due to shorter availability of liquidity factor in Pástor and Stambaugh (2003) and the 60-month cumulation period to compute the associated state variable, June 1968 is the first date for which we observe all relevant variables.

4.1 Risk Prices

Consider a factor model ϕ, e.g., the Fama and French (1992) three-factor model, and a set of test assets $\mathcal{A}$. The covariance risk prices for a state variable z is the coefficient γ_z in a multivariate extension of (2),

$$\mathrm{E}_t\left(R_{i,t+1} - R_{f,t+1}\right) = \gamma \,\mathrm{Cov}_t\left(R_{i,t+1}, R_{m,t+1}\right) + \sum_{z \in \phi \backslash m} \gamma_z \,\mathrm{Cov}_t\left(R_{i,t+1}, \Delta z_{t+1}\right), \qquad (7)$$

where $\phi \backslash m$ is the set of factors in model ϕ, excluding the market factor, and $i \in \mathcal{A}$. We estimate these risk prices using GMM (Hansen 1982) with a modified Fama and MacBeth (1973) methodology described in Maio and Santa-Clara (2012), p. 595.[11] This produces a vector of risk prices $\boldsymbol{\gamma}^{\phi,\mathcal{A}}$ for the factor model and test asset combination in each model estimation window. To be consistent with Maio and Santa-Clara (2012), we use two groups of test assets: the 25 Fama-French size-value sorted portfolios ($\mathcal{A}_{s,v}$) and their 25 size-momentum sorted portfolios ($\mathcal{A}_{s,m}$). More details about this estimation process are available in Section A4 of the Online Appendix.

We consider three model sets in our analysis. The first consists of the eight factor models tested in Maio and Santa-Clara (2012), which are listed in footnote 7, and which we describe in more detail in Section A3 of the Online Appendix. Some of the MSC factors consist of long-short stock portfolios, while others, e.g., $TERM$ and DEF, are constructed from other market and macro variables, as explained in Section 3. In particular, following Maio and Santa-Clara (2012), we do not use factor replicating portfolios for the non-traded state variables. Our other factor models consist of the overall U.S. stock market, plus subsets of the 13 thematic factors from Jensen et al. (2023), which are summarized in Panel B of Table 1. The JKP3 model set includes the overall market, and all two-factor combinations of the 13 thematic factors, for a total of 78 (13 choose 2) models. The JKP4 model set includes the market and all three-factor combinations, for a total of 286 models. We refer to a model set, e.g., the eight MSC models, as Φ and to a specific model in this model set, e.g., the Fama-French three-factor model, as $\phi \in \Phi$.

[11] In our baseline model we do not include factor returns as test assets, which is common in the literature (see footnote 18 in Raponi et al. 2019). However, based on the suggestion in Lewellen et al. 2010, we also re-estimated the GMM after imposing the additional moment conditions that average factor returns should equal the factor risk premia. This did not change our results.

4.2 Forecasting Estimates

For each model ϕ (e.g., one of the eight MSC factor models) we run forecasting regressions using monthly data

$$Y^{\tau}_{t-o,t+h} = c^{\tau}_{0h} + \boldsymbol{c}^{\tau}_{h}\boldsymbol{Z}^{\phi}_{t} + u^{\tau}_{t,t+h}, \tag{8}$$

where $\boldsymbol{Z}^{\phi}_{t}$ is a vector of forecasting variables associated with model ϕ (explained below), and Y^{τ} is one of four outcome variables: CRSP value weighted market returns (M), realized S&P 500 return variance (SVAR), consumption growth (C), or the cross-sectional variance of consumption growth (CVAR). We refer to $\tau \in \{\text{M, SVAR, C, CVAR}\}$ as the test type. These outcome variables are summarized in Table 2. Note that regression (8) does not depend on a specific set $\mathcal{A}$ of test assets, but the risk price analysis in (7) does.

For market returns and market variance, both of which are used by Maio and Santa-Clara (2012), $o = 0$ in (8), and the forecasting horizon is one of $h \in \{1, 12, 60\}$ months. For example, for market returns $Y_{t,t+12}$ is the return, including dividends, from the market close at the end of month t to the market close at the end of month $t + 12$. The market return variance is calculated over the same time period.

For consumption growth and cross-sectional consumption variance, $o = 3$ months and $h \in \{0, 2, 11\}$ months. Consumption data are reported monthly for a set of three rolling cohorts, each reporting on a different quarterly cycle. For consumption growth, $Y_{t-3,t}$ is the quarter-over-quarter growth rate of the single cohort reporting in month t. $Y_{t-3,t+2}$ includes the sum of the quarter-over-quarter consumption growth of cohorts reporting in months $\{t, t+1, t+2\}$. And $Y_{t-3,t+11}$ includes the sum of quarter-over-quarter consumption growth of cohorts reporting in months $\{t, t+1, t+2, \ldots, t+11\}$. The timing for cross-sectional consumption growth variance is identical. For C and CVAR and the $\{o = 3, h = 0\}$ specification, the regression in (8) is contemporaneous.

We follow Maio and Santa-Clara (2012) in defining the forecasting variables $\boldsymbol{Z}^{\phi}_{t}$ from factors. For the state variables $TERM$, DEF, DY, RF, VS, and CP, we use the level of the variable itself as the forecasting variable. For SMB and HML, we apply the transformations in equations (22) and (23) in Maio and Santa-Clara (2012) to define forecasting variables SMB^{*} and HMB^{*}.[12] For UMD and L and all JKP factors, the forecasting variable is the rolling 60-month cumulative factor return.

[12] With these transformations, SMB approximates changes in SMB^{*}, and HML approximates changes in HML^{*}, giving SMB^{*} and HML^{*} an interpretation as state variables.

4.3 Sign Consistency

We seek to quantify the degree to which a factor model's risk prices have signs that are consistent with their forecasting coefficients. To motivate our consistency index, we start with a simple case. Suppose we estimate $\boldsymbol{\gamma}_z^{\phi,\mathcal{A}}$, the risk price for factor z in model ϕ estimated from assets $\mathcal{A}$, and suppose we also estimate $\boldsymbol{c}_{h,z}^M$, the market-return forecasting coefficients for the same factor z over horizon h. We want a simple way to quantify the degree to which these parameters exhibit sign-consistency, based on our estimates. We want a measure that is

- large and positive when the two estimates have the same sign and are significant;
- large (in magnitude) and negative when the estimates have opposite signs and are significant;
- close to zero if either estimate fails to be significant.

These properties are satisfied qualitatively by taking as our measure the product of the t-statistics $t(\boldsymbol{\gamma}_z^{\phi,\mathcal{A}})$ and $t(\boldsymbol{c}_{h,z}^M)$ of the two estimates. This product is our core measure of sign consistency.

In the case of forecasting coefficients $\boldsymbol{c}_{h,z}^\tau$, with $\tau \in \{\text{SVAR, CVAR}\}$, the ICAPM predicts opposite signs for risk prices and forecasting coefficients, so we flip the sign of the product of t-statistics in measuring consistency. This preserves the property that larger positive values of our measure indicate greater sign-consistency.

We now extend this measure from an individual factor z to a factor model ϕ by averaging over pairs of coefficients. For a factor model ϕ, set of test assets $\mathcal{A}$, and test type τ, the model consistency index (MCI) is defined as follows:

$$\begin{aligned} MCI(\phi,\tau,\mathcal{A}) &= s(\tau) \times \frac{1}{(|\phi|-1)\times|H|} \sum_{z\in\phi\setminus m} t(\boldsymbol{\gamma}_z^{\phi,\mathcal{A}}) \times \Big(\sum_{h\in H} t(\boldsymbol{c}_{h,z}^\tau)\Big), \quad \text{where} \\ s(\tau) &= \begin{cases} +1 & \text{for } \tau \in \{\text{M, C}\} \\ -1 & \text{for } \tau \in \{\text{SVAR, CVAR}\} \end{cases}, \end{aligned} \tag{9}$$

where $t(\boldsymbol{\gamma}_Z^{\phi,\mathcal{A}})$ is the t-statistic associated with the factor z in the price-of-risk estimation for factor model ϕ and test asset set $\mathcal{A}$ (from equation 7), $t(\boldsymbol{c}_{h,z}^\tau)$ is the t-statistic associated with z in the horizon h forecasting regression of type τ (from equation 8), and H is the set of horizons. The t-statistics for factor risk prices use GMM standard errors as described in Section A4 of the Online Appendix, and t-statistics from (8) use Newey and West (1987)

standard errors with h lags. The market factor is excluded from (9) based on the a priori assumption that any valid factor model should contain the market return.

The final model consistency index for factor model ϕ and test type τ averages across the two sets of tests assets

$$MCI(\phi, \tau) = \frac{MCI(\tau, \phi, \mathcal{A}_{s,v}) + MCI(\tau, \phi, \mathcal{A}_{s,m})}{2}, \tag{10}$$

where $\mathcal{A}_{s,v}$ are the 25 size-value sorted Fama-French portfolios and $\mathcal{A}_{s,m}$ are the 25 size-momentum sorted Fama-French portfolios.

Discussion. The intuition for (9) follows from the analysis of Section 2. A factor that positively forecasts future market returns or contemporaneous or future consumption growth is risky and should thus command a positive risk premium. Therefore, we expect the t-statistics from the price-of-risk regression and the different horizon forecasting regressions to have the same sign. Furthermore, higher prices of risk should be associated with greater predictability of future returns, resulting in a higher value of the index. On the other hand, positive prices of risk associated with very negative return or consumption forecasts are highly inconsistent with the model and produce lower values of the index. If either the risk price or forecasting coefficient of a factor is close to zero, then the ICAPM doesn't restrict the sign of the other effect, and that factor's contribution to the index will be close to zero. The product of t-statistics is a parsimonious representation of this logic. In contrast, factors that positively forecast future return or consumption volatility are hedges and should command negative risk premia. For $\tau \in \{SVAR, CVAR\}$, consistency requires that the t-statistics for risk prices and forecasting coefficients have opposite signs. That is why $s(\tau) = -1$ in these cases. With this change of sign, the product of t-statistics argument goes through as before.[13]

Since the MCI score is an average of products of pairs of t-statistics, the square root of the MCI has the same units as a t-statistic and thus its magnitude has a similar interpretation. MCI scores higher than 1.96^2 suggest strong agreement between risk premia and forecasting coefficients. On the other hand, negative MCI scores are strongly indicative of violation of ICAPM sign consistency. We make formal connections between the magnitude of MCIs and out-of-sample measures of model stability in Section 5.

Sample MCI calculation. Table 3 gives an example of the model consistency index calculation for the Fama and French (1992) three-factor model, i.e., ϕ = FF3, with test

[13]We experimented with a version of (9) that ran each t-statistic through a tanh function to reduce the impact of outliers, but found that this did not lead to better performance on the tests in Section 5.

assets $\mathcal{A}_{s,v}$ (25 size-value portfolios) using data through December of 1985. In this model, $\phi\backslash m$ contains the SMB and HML factors, whose risk prices are positive with t-statistics of 1.51 and 4.31, respectively (Panel A). We restrict attention to test type $\tau = M$, so equation (8) uses the market return as the dependent variable. We evaluate the forecasting coefficients for SMB and HML (transformed as explained Section 4.2) across three forecasting horizons. Panel B shows that $\{t(c^M_{h,SMB})\}_{h\in\{1,12,60\}} = \{-0.14, 0.79, 1.22\}$ and $\{t(c^M_{h,HML})\}_{h\in\{1,12,60\}} = \{1.99, 1.82, 18.51\}$. The MCI is calculated as in (9), by multiplying the t-statistics of the factor risk prices with the t-statistics of their respective forecasting coefficients, and then averaging across the $(|\phi| - 1) \times |H| = 6$ pairs. We find that $MCI(FF3, M, \mathcal{A}_{s,v}) = 16.5$, which indicates a high degree of ICAPM consistency.

MCI correlations. For each test type τ, we use (9) to calculate $MCI(\phi, \tau, \mathcal{A})$ for the eight MSC models and each of the test asset sets $\mathcal{A}_{s,m}$ and $\mathcal{A}_{s,v}$. We calculate these MCIs using data up to and including 2020, the last year for which data on all four MCIs are available. Each test type $\tau \in \{M, SVAR, C, CVAR\}$ is represented by its MCI observation vector (for the 16 model-test assets combinations). We then calculate the correlation matrix for these four observation vectors. We then apply the same procedure replacing the eight MSC models with the 286 JKP4 models.[14]

Table 4 shows the correlations of these four vectors, calculated separately for the MSC and JKP4 models. Table A2 of the Online Appendix shows correlations of the four MCI measures for the JKP3 models. With the exception of {M,CVAR}, all MCI pairs have positive correlations for the MSC, JKP3, and JKP4 sets of factor models, suggesting some agreement among the different consistency measures.

4.4 Comparison with Maio and Santa-Clara (2012)

An advantage of the MCI measure in (10) is that it systematizes the appealing economic logic of Maio and Santa-Clara (2012). We next compare the qualitative conclusions in Table 1 of their paper to our MCI measures evaluated using the same 1963-2008 data sample.[15] Our Table 5 reproduces the Maio and Santa-Clara (2012) model validity assessments in the $MSC(M)$ and $MSC(SVAR)$ columns. We show our $MCI(M, \phi, \mathcal{A})$ measure in the $MCI(M)$ column and our $MCI(SVAR, \phi, \mathcal{A})$ measure in the $MCI(CVAR)$ column. Here we are not aggregating the MCI measures for the $\mathcal{A}_{s,v}$ and $\mathcal{A}_{s,m}$ test sets, but showing them separately for comparison with Maio and Santa-Clara (2012).

[14]For the each test type, this yields a vector of length $8 \times 2 = 16$ for MSC and length $286 \times 2 = 572$ for JKP4 models.

[15]Table A1 of the Online Appendix explains how Maio and Santa-Clara (2012) draw their conclusions.

The four models with the highest model consistency scores for the $\tau = M$ (market return) test type are FF3, C, and PS with the size-value test assets and the C model with the size-momentum test assets (see footnote 7 for a list of the MSC models and their abbreviations). All four cases were flagged by Maio and Santa-Clara (2012) as consistent with their market return ICAPM test. In addition, all ten model and test asset pairs that receive negative MCI scores using (9) are also rejected by MSC.

Some notable differences between our methodology and MSC's assessments include the CV and FF5 model tests using $\tau = M$ and the size-value test assets, as well as the model C tests using $\tau = SVAR$ for both sets of test assets. Here MSC reject ICAPM consistency, often based on a single instance of sign mismatch, including cases in which the sign is not statistically significant. Our measures in (9) weigh the overall degree of consistency and account for the degree of statistical significance of the estimated parameters.

The effect of accounting for statistical significance is illustrated in Table 5 by the KLVN model with the size-value test assets. This case passes the MSC(M) sign-consistency test, but we give it a relatively low MCI(M) score of 0.935 because the estimated risk prices are not significantly different from zero. A similar contrast can be seen for the FF5 model with the size-value test assets. The model passes the MSC(SVAR) test but gets a weak MCI(SVAR) score of 0.535.

Overall, our systematized version of the MSC ICAPM consistency idea generally leads to similar conclusions to the ones they reached on their original sample, though with some differences.[16] One appealing aspect of our methodology is that it is straightforward to run over different time periods because a subjective assessment of sign consistency is not required.

The last two columns of Table 5 show that the MCI measures for different models and test types change when the time period is extended to include our full sample. Time variation in ICAPM consistency suggests that the ability of models to provide effective guidance to researchers and practitioners may also vary over time. We explore this possibility in the next section.

5 Evaluating Factor Models

ICAPM model consistency is of practical interest to the extent that more consistent factor models perform better, in some sense, than less consistent ones. We propose measuring

[16]Section A5 of the Online Appendix has a detailed description of the results in in Maio and Santa-Clara (2012).

model performance via time variation in factor model alphas, betas, risk prices, and goodness-of-fit metrics. It would be desirable for the following relationships to hold in high MCI factor models:

i) Alphas should be zero on average. Non-zero alphas should therefore indicate temporary mispricing, which should not be persistent;

ii) Betas of assets to factors should be stable.

To the extent that risk prices (the gammas from equation 7) reflect time-varying forecasting properties of factors, as suggested by Barroso et al. (2021), if the latter change, then risk prices should change as well.

iii) We therefore do not expect gamma stability to be associated with more ICAPM-consistent models.

Finally, we expect that

iv) Historical betas should do a good job of explaining future realized returns relative to future factor returns;

If these properties hold, then high MCI models can be said to be better specified than low MCI models.

5.1 Model Stability Measures

To evaluate properties (i)–(iv), we will calculate measures of model stability and examine if greater stability is associated with higher MCIs. To evaluate model stability, we compare model parameters and performance over time.

For each model ϕ, set of test assets $\mathcal{A}$, and year t, we run a standard factor regression using either daily (for JKP models) or monthly (for MSC models) data to estimate factor model alphas and betas in year t and then in year $t+1$

$$R_{is} - R_{fs} = \alpha_i + \beta_{mi}(R_{ms} - R_{fs}) + \boldsymbol{\beta_i}^\top \boldsymbol{\phi}_s^{\setminus m} + \epsilon_{is}, \tag{11}$$

where R_{is} is the return for test asset $i \in \mathcal{A}$ in period s, R_{fs} is the risk-free rate (known at time $s-1$), R_{ms} is the market return, and $\boldsymbol{\phi}_s^{\setminus m}$ is the vector of non-market factor returns from model ϕ. We also estimate gammas (risk prices) in years t and $t+1$. We refer to year t as the *training window* and year $t+1$ as the *evaluation window*. We use the factor model residuals in year t to calculate the training-window mean-squared error (MSE); we

then use the training-window alphas and betas to calculate the evaluation-window MSE. (We will also consider versions of these calculations with longer training windows.) In each period, each $\{\phi, \mathcal{A}\}$ pair produces $|\mathcal{A}|$ alphas and MSEs (one per test asset), $|\phi| - 1$ gammas (one per factor excluding the market factor), and $|\mathcal{A}| \times (|\phi| - 1)$ betas (one per test asset-factor combination). We evaluate model stability by comparing estimates in the training and evaluation windows.

We calculate the alpha stability (or closeness) using cosine similarity

$$c(\boldsymbol{\alpha}^-, \boldsymbol{\alpha}^+; \phi, \mathcal{A}) \equiv \frac{\boldsymbol{\alpha}^- \cdot \boldsymbol{\alpha}^+}{\|\boldsymbol{\alpha}^-\| \|\boldsymbol{\alpha}^+\|}, \tag{12}$$

where $\boldsymbol{\alpha}^-$ represents the vector of training-window alphas and $\boldsymbol{\alpha}^+$ represents the vector of evaluation-window alphas. Applying (12) to the training- and evaluation-window non-market betas yields a measure of beta stability $c(\boldsymbol{\beta}^-, \boldsymbol{\beta}^+; \phi, \mathcal{A})$, and applying (12) to the gammas yields a measure of gamma stability $c(\boldsymbol{\gamma}^-, \boldsymbol{\gamma}^+; \phi, \mathcal{A})$. For the beta comparisons, we stack the beta vectors across all test assets, and the $\boldsymbol{\gamma}$s are vectors of risk prices for each non-market factor. We use cosine similarity rather than correlation because the average level of evaluation-window outcomes also matters, and not only variation around the average.

Evaluation-window MSEs will likely be considerably higher than training window MSEs because we use training window alphas and betas in the evaluation window. To account for this relationship, we measure MSE stability (in percent) using either

$$s_{mr}(\boldsymbol{\mu}^-, \boldsymbol{\mu}^+; \phi, \mathcal{A}) = \frac{100}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \frac{\mu_i^-}{\mu_i^+} \qquad \text{or} \qquad s_{rm}(\boldsymbol{\mu}^-, \boldsymbol{\mu}^+; \phi, \mathcal{A}) = 100 \times \frac{\sum_{i \in \mathcal{A}} \mu_i^-}{\sum_{i \in \mathcal{A}} \mu_i^+}. \tag{13}$$

The s_{mr} measure is the mean of ratios of training window (μ_a^-) to evaluation window (μ_a^+) MSEs, and s_{rm} is the ratio of the means. We discuss s_{mr} here and consider s_{rm} in the Online Appendix. In both cases, the higher the MSE ratio, the greater is the ability of training-window alphas and betas to explain test asset returns in the evaluation window. Models which show lower out-of-sample mean-squared errors are associated with greater hedging efficacy.

An Example

Figure 2 shows alpha, beta, gamma, and MSE stability measures for the year 1977 (so the evaluation window is 1978). Each panel is a scatter plot of one of the stability measures

from (12) and s_{mr} from (13), shown on the y-axis. On the x-axis is each model ϕ's MCI from (9) evaluated using the entire sample until 1977 for test type $\tau = M$ and test asset set $\mathcal{A}$, i.e., $MCI(\phi, M, \mathcal{A})$. Because there are eight MSC models and two sets of test assets (size-value and size-momentum portfolios), each scatter plot has 16 points.

The upper left panel in the figure shows that higher MCI factor models are associated with less persistent alphas, as model mispricing tends to be less persistent for higher MCI models. The upper right panel shows that more consistent models exhibit more stable betas. The lower left panel shows that gammas, or risk prices, tend to be more stable for the more consistent models, though the relationship between gamma stability and MCIs is not statistically significant in this example. Finally, the lower right panel shows that more consistent models are associated with higher mean MSE ratios in (13), suggesting more consistent models yield betas which are more useful for ex-post portfolio hedges. The Figure 2 slopes relating the four model stability measures to model consistency are:

	Alpha	Beta	Gamma	MSE Ratios
Coefficient	-0.1	0.05	0.01	2.44
T-stat	-3.08	2.12	0.41	2.27

Figure A4 of the Online Appendix shows the same results but using rank correlations in (12) (instead of cosine similarity) and the s_{rm} measure from (13) instead of s_{mr}. The results are similar.

5.2 Model Stability Over Time

The left part of Figure 3 repeats this analysis for each year of the sample, using expanding windows, with the first one starting in 1975. The four subpanels of the left part of the figure plot the time series of the four slopes calculated in Figure 2.

For each year t we run a cross-sectional regression of each one of the model stability measures $\psi(\phi, \mathcal{A}) \in \{c(\boldsymbol{\alpha}^-, \boldsymbol{\alpha}^+; \phi, \mathcal{A}), c(\boldsymbol{\beta}^-, \boldsymbol{\beta}^+; \phi, \mathcal{A}), c(\boldsymbol{\gamma}^-, \boldsymbol{\gamma}^+; \phi, \mathcal{A}), s_{mr}(\boldsymbol{\mu}^-, \boldsymbol{\mu}^+; \phi, \mathcal{A})\}$, from (12) and (13), on $MCI(\phi, \tau, \mathcal{A})$, with $\tau = M$,

$$\psi(\phi, \mathcal{A}) = a_{\tau,t} + b_{\tau,t} \times MCI(\phi, \tau, \mathcal{A}) + \epsilon(\phi, \mathcal{A}). \tag{14}$$

With the MSC model set and two sets of test assets, we have 16 $\{\phi, \mathcal{A}\}$ combinations and thus 16 observations in each regression. Figure 3 plots $b_{M,t}$ over time for each of the four possible values of ψ. Negative and significant coefficients are marked with a red x, and positive and significant coefficients marked with a green dot. Significance is measured at the 5% level using White standard errors in each cross-sectional regression.

In the upper-left subpanel of the figure, the MCI-alpha stability relationship is positive and significant four times, but it is negative and significant 10 times. This indicates a tendency for higher MCI models to produce less persistent alphas; an ideal model would eliminate alphas, so less persistence in alphas is preferable. The MCI-beta stability relationship is positive and significant 19 times, and is never negative and significant. This is a strong sense in which higher MCI models are more stable over time. The evidence for MCI-gamma stability is mixed, with six positive and significant and three negative and significant coefficients. Finally, the evidence is supportive of a positive MCI-mean MSE ratio relationship, with five positive and significant coefficients and two negative and significant coefficient. For models with higher MCIs, the evaluation-window pricing errors calculated from training-window model parameters tend to be smaller, relative to training-window pricing errors.

The right part of Figure 3 repeats this analysis for the JKP4 set of 286 models. Each cross-sectional regression (14) now uses 286×2 observations for all possible $\{\phi, \mathcal{A}\}$ pairs, rather than the 16 observations for the MSC models. The larger sample size leads to more precise slope estimates and thus a larger number of significant observations. The results are consistent with the MSC analysis. Higher MCI models are associated with less alpha persistence, more stable betas and gammas, and lower evaluation-window relative MSEs.[17] However, for reasons that we discuss later, our analysis in Section 6 finds that the MCI-gamma relationship is generally the weakest of the four.

5.3 Model Consistency for Other ICAPM Test Types

We extend the analysis of Figure 3 to all test types by running (14) for all $\tau \in \{M, SVAR, C, CVAR\}$. Figure 4 summarizes the results over time. The left side of the figure shows the results for the eight MSC factor models and the right side shows the results for the 286 JKP4 factor models. Each subpanel of the figure counts the number of significantly positive and negative coefficients in the rolling cross-sectional relationships between the MCIs and the four model stability measures. Because we include the market- and consumption-based tests in each subpanel, the time period of analysis is restricted to the 31 years for which both sets of tests are available, and thus differs from the time period of Figure 3.

The top left chart in each panel shows the number of times that the MCI-alpha slope coefficients are positive and significant or negative and significant in all the rolling re-

[17] Figure A5 of the Online Appendix repeats this analysis using rank correlation in (12) and s_{rm} from (13). Figure A6 repeats the analysis with cosine similarity using the JKP3 models. All results are similar.

gressions. For both sets of factor models and the four test types, the MCI-alpha stability relationship is most often negative when it's significant, meaning that high MCI models are systematically associated with less alpha persistence. The results are more pronounced for the JKP4 models, because, with a larger cross-section, the slope coefficient in (14) is better measured.

Similarly, the top-right chart in each panel shows that the MCI-beta relationship is strongly positive. This is especially true for the JKP4 models, where the larger cross-section allows for more precise slope estimates. The lower-left charts examine price-of-risk stability as a function of model MCIs. For the market-based tests, labeled "M", higher model consistency is associated with more stable risk prices. However, for the other three tests, the evidence is mixed, without a clear relationship between MCIs and gamma stability. The lower-right figures demonstrate that higher MSE ratios are generally associated with higher MCIs, though this is not true for $\tau = SVAR$.[18]

Figure 5 shows the average R-squareds of the rolling regressions of Figure 3 for the market-type tests, as well as summary results for the SVAR, consumption, and CVAR test types. Again, the analysis is restricted to the 31 years in which each test-type can be performed. For the MSC models, there isn't a pronounced pattern in R-squareds, with the exception that the beta-MCI R-squareds are highest for the $\tau = M$ test types. For the JKP4 models, a clear pattern emerges: The consumption- and CVAR-type tests generally have higher R-squareds than their market- and SVAR-counterparts. The association of future outcomes with MCIs is highest for the consumption-based tests. As discussed in Section 2, the ICAPM sign-consistency property applies more directly to consumption and extends to the market only as a proxy, so this empirical pattern is consistent with our expectation based on theory.[19]

Robustness. Figure A11 of the Online Appendix reports very similar results using only the one-month (market) or contemporaneous (consumption) horizons for the outcome regression in (8). This is closest in spirit to the continuous-time formulation of the ICAPM. Figure A12 of the Online Appendix shows this analysis is unchanged if we expand the set of test assets to include 25 size-investment and 25 size-profitability sorted portfolios.

[18]Figure A7 of the Online Appendix shows the results when using rank correlations in (12) and s_{rm} from (13). Figure A8 shows the cosine similarity and s_{mr} results for the JKP3 models. All results are similar.

[19]Figure A9 of the Online Appendix repeats this analysis using rank correlation in (12) and s_{rm} from (13) Figure A10 of the Online Appendix repeats the cosine similarity analysis, s_{mr} analysis for the JKP3 models. All results are similar.

6 Statistical Tests of MCI-Stability Relationship

Using one-year training and one-year evaluation windows, Figures 2, 3, and 4 show that, for all four test types, there is evidence that model ICAPM consistency is associated with future factor model stability. Table 6 extends this analysis to also consider one-, two-, and three-year training windows, as well as the JKP three-factor models.[20] We use monthly data for the analysis involving MSC factors and daily data for the JKP models. We also develop a formal statistical test for whether MCIs are associated with out-of-sample factor model stability across the four measures of Section 5.1.

Each cell of the table corresponds to a different model set, training window, and test statistic combination (e.g., MSC model set, three-year training window, and alpha cosine similarity). For each test type τ, we run a set of regressions from (14), one for each stability measure, in annual expanding windows starting in 1975. Every time the cross-sectional regression coefficient $b_{\tau,t}$ in (14) comes up positive and significant, a +1 is added to the cell count, and every time the coefficient is negative and significant and -1 is added to each cell:

$$\xi(\Phi, T, \psi, \tau, t) = \begin{cases} +1 & \text{if } b_{\tau,t} \text{ is positive and significant,} \\ -1 & \text{if } b_{\tau,t} \text{ is negative and significant,} \\ 0 & \text{otherwise;} \end{cases}$$

T is the length of the model training window in years, Φ is the model set (MSC, JKP3, or JKP4), and ψ is the stability measure (alpha cosine similatity, beta cosine similarity, etc.), τ is the test type, and t is the end year of the expanding window. Significance in (14) is determined at the 5% level using White standard errors.

We then aggregate the ξs across the four different test types τ and years t

$$j(\Phi, T, \psi) = \sum_{\tau} \sum_{t} \xi(\Phi, T, \psi, \tau, t). \tag{15}$$

Because there are 48 years of data for the $\tau \in \{M, SVAR\}$ tests and 31 years for the $\tau \in \{C, CVAR\}$ tests, there are a total of 158 ξ observations for each j statistic. The minimum and maximum values of j are therefore -158 and $+158$, respectively. Each cell in Table 6 shows its respective value for $j(\Phi, T, \psi)$.

[20] In Table 6, the number after "pre" is the number of years in the training window. The "post" number is the length of the evaluation window and is always 1. Table 6 uses all available market observations, so its "pre 1, post 1" values are not identical to the net counts calculated in Figure 4, which are based on time periods for which both market and consumption data are available.

6.1 Inference for j-statistics

To obtain p-values for the j-statistic in (15) under the null of no MCI-to-model-stability relationship, we draw 158 multinomial random variables—each proxying for one of the ξs—which have a 2.5% of having value -1, a 2.5% probability of having value +1, and a 95% probability of having the value zero. Summing across 158 such draws gives one observation of j under the null. The challenge is that each such draw is not independent, e.g., four of the draws correspond to the cross-sectional regressions of beta cosine similarity on MCIs for each of the four τs (test types) in the year 1998. To allow for the tests to be correlated, we sample the multinomial random variables by first sampling 158 multivariate normal variables with a single pairwise correlation parameter of ρ. Each normal variable is discretized to $\{-1, 0, +1\}$ using the 5% cutoffs of a standard normal distribution.

For each value of ρ, we repeat this 158-variable draw 2,000,000 times. For each simulation run, we calculate the mean and median of all pairwise correlations of the discretized variables and the mean of all pairwise correlations of the raw (i.e., non-discretized) variables. We then average these three means across all 2,000,000 simulation runs. Figure A15 of the Online Appendix plots these three simulation averages for each value of ρ. The simulation average of the mean pairwise correlations of the raw variables forms the expected 45-degree line relative to the simulation ρ parameter (blue line). The simulation average of the median pairwise correlations of the discretized variables, the green line, shows a convex relationship between correlations of the raw variables and correlations of the discretized variables. The discretized correlations are uniformly lower than the raw ones, and approach the raw correlation curve only as ρ approaches 1.[21]

For each ρ and each simulation run, we then sum the 158 discretized random variables to form a single realization of the j-statistic from (15) under the null of no relationship between MCIs and model stability. Across the 2,000,000 simulation runs for $\theta \in \{0, 5, 10, \ldots, 80\}$, we calculate the fraction of times that $j \leq -\theta$ or $j \geq \theta$, i.e.,

$$p(\theta, \rho) \equiv \frac{|\{i : j_i \leq -\theta \text{ or } j_i \geq \theta\}|}{2,000,000}.$$

This represents the two-sided p-value for the j-statistic under the null of no MCI and stability relationship. By definition, $p(0, \rho) = 1$. When $\rho = 0$, even $p(5, 0)$ is quite low, because the distribution of js across all draws is so tightly centered around 0. As the correlation increases, the j-statistics become more dispersed. In the limit, as $\rho \to 1$,

[21] The averages across simulations of the mean and median pairwise correlation of the discretized draws are virtually identical.

the j-statistics can take on only three values $\{-158, 0, 158\}$ with probabilies 2.5%, 95%, and 2.5% respectively. Table A3 of the Online Appendix shows, for every ρ, the average of the mean pairwise discretized correlations from Figure A15, as well as $p(\theta, \rho)$ for all thresholds.

The j-statistic from (15) in each cell in Table 6 contains the sum of $\xi(\cdot, \tau, t)$ across the four test types and across all years for which the cross-sectional regression in (14) is run. Because ρ is a latent parameter, to obtain a p-value for each j-statistic, we estimate $\hat{\rho}_{disc}(\Phi, T, \psi)$, the discretized version of the correlation, in the data and use our simulation results (from Table A3) to find the mapping from $\hat{\rho}_{disc}$ to ρ. We calculate $\hat{\rho}_{disc}(\Phi, T, \psi)$ as the average of the six cross-sectional correlations among $\xi(\cdot, \tau, \cdot)$ for $\tau \in \{M, SVAR, C, CVAR\}$, as well as the four time-series correlations between one-year lags of the ξs corresponding to the same test type, i.e., $\xi(\cdot, \tau, t)$ with $\xi(\cdot, \tau, t-1)$, and the four time-series correlations at two-year lags.[22] Panel B of Table A4 shows $\hat{\rho}_{disc}(\Phi, T, \psi)$ across factor model sets, training window horizons, and model stability measures. For each j-value in Table 6, we find its associated $\hat{\rho}_{disc}$ and then look up the interpolated p-value from Table A3.[23]

6.2 Results

The left column of Table 6 shows the j-statistic results for the alpha stability measure in (12). All j-statistics for alpha stability are non-positive, suggesting that test asset alphas relative to more ICAPM-consistent factor models are less persistent. Because alphas relative to the "correct" factor model should be zero, less alpha persistence indicates a more properly specified model. Interestingly, the MSC factor model j-statistics are smaller in magnitude than those associated with the JKP3 model set, but the MSC j-statistics have lower p-values. This is because the realized correlations $\hat{\rho}_{disc}(MSC, T, \alpha)$ are lower than $\hat{\rho}_{disc}(JKP3, T, \alpha)$, which leads to lower p-values for a given value of j (see Table A4).

The results for beta stability (second column) are all very positive and highly statistically significant. Higher MCIs predict more stable betas. Those interested in using

[22]The cross-sectional correlations are restricted to the 31 years for which we have data for all four test types, while the time-series correlations use 31 and 48 years respectively. Panel A of Table A4 of the Online Appendix shows that the means of correlations in these three categories exhibit a good deal of cross-sectional variation. The overall mean equals the average of the three category-level means weighted by $\{6, 4, 4\}$.

[23]Negative values of $\hat{\rho}_{disc}(\Phi, T, \psi)$ are treated as $\hat{\rho}_{disc}(\Phi, T, \psi) = 0$ thus leading to a more conservative (higher) p-value estimate for the corresponding $j(\Phi, T, \psi)$.

factor models as a risk management tool would prefer betas to be stable over time, and thus should select factor models with higher MCI scores. Related to beta stability, out-of-sample mean squared errors (third column), measured via s_{mr} from (13), are lower for more ICAPM-consistent models (since higher s_{mr}s are associated with lower out-of-sample mean-squared errors, relative to training-window errors) and the result is highly statistically significant.

The MCI-stability results for the price-of-risk measures—the γs from equation (7)—are shown in the fourth column of Table 6. We find little evidence of a relationship between price-of-risk stability and model consistency. There are two potential reasons. First, it could be that gamma similarity is less precisely measured since it is derived from fewer underlying observations: alphas, betas and MSEs can be observed on the test asset level, but gammas are observed on the factor level only (see the discussion in Section 5.1). Second, the only requirement that sign-consistency places on risk prices is that they should have the proper relationship with forecasting regression coefficients in (9). The ICAPM does not suggest that risk prices of even the "correct" model must remain constant over time if the forecasting relationships change. And our empirical results identify no such tendency.[24]

Table A5 of the Online Appendix repeats the above analysis using the rank correlation versions of the stability measures from Section 5.1, as well as the s_{rm} from (13). The results are broadly consistent with those in Table 6.

7 Selecting Factor Models

The top panel of Figure 6 shows the MCI from (10) averaged across the four test types

$$MCI(\phi) = \frac{1}{4} \sum_{\tau \in \{M, SVAR, C, CVAR\}} MCI(\phi, \tau) \tag{16}$$

for the eight MSC models (see footnote 7), for each year from 1990 through 2020. The most consistent models, based on their average MCIs over time, are $FF3$, the Fama and French (1992) three-factor model, and C, the Carhart (1997) model, which adds momentum to the Fama-French three-factor model. The least consistent of the MSC models are CV, HL, and P. Interestingly, the most consistent models—$FF3$ and C—contain only tradable equity portfolios $\{SMB, HML, UMD\}$, in addition to the market.

[24]This is consistent with the finding in Barroso et al. (2021) risk premia can switch signs over time, driven by the directionality of how state variables forecast future consumption growth.

But the least consistent models contain non-tradable factors, many of which are related to bond markets: $\{TERM, DEF\}$ in the case of HL, $\{TERM, PE, VS\}$ for CV, and $\{TERM, DEF, DY, RF\}$ for P.

The bottom panel of Figure 6 shows the output of (16), but with τ restricted to $\{M, SVAR\}$. Because these tests do not use consumption data, they contain 15 years of additional data at the start of the sample and three additional years at the end. The results are qualitatively consistent with those shown in the top panel of the figure, except that the P and PS models fare better when covariance with consumption and consumption volatility is not considered.

Because the JKP4 model set has 286 models, it is not practical to plot all of their consistency scores in one heatmap. Instead, in each year we select the top 10 highest MCI models from JKP4 using (16). Figure 7 then tabulates the number of times each of the 13 major theme factors from Jensen et al. (2023) appears in the top ten most consistent models. For example, in the last evalution of the MCI test statistic, the profit growth theme factor appeared in seven of the top ten models, the profitability theme factor appeared in five of the ten top models, while the accruals, seasonality and short term reversal theme factors appeared zero times. The most heavily represented themes across all time periods were profit growth and profitability, with low leverage in third place.

The bottom panel of Figure 7 shows the same analysis but restricting (16) to $\tau \in \{M, SVAR\}$ test types. Because consumption data is not required, the bottom panel covers a longer data sample. Similarly to the top panel, the most frequent themes in the most consistent JKP4 models are low leverage, profit growth, and profitability. The themes that are least present in the top ten most consistent models are accruals, debt issuance, momentum, seasonality, and short term reversals.

Robustness. Figures A13 and A14 of the Online Appendix show these results remain robust to a shorter forecasting horizon in (8) and to the inclusion 25 size-profitability and 25 size-investment sorted portfolios as additional test assets.

7.1 Recent Consistent Themes

The takeaways for practitioners who build factor models to gauge mispricing and for risk management is to focus on factors falling into the themes we identify as most consistent and to avoid factors falling into themes we classify as being less consistent.

To gain greater insights into the which factors are represented in the most recent

consistent models, Table 7 shows a decomposition of the last column in Figure 7. The top panel in the table shows the top ten models for the MCI measure in (16) which uses all four ICAPM test types. These tests are performed in December 2020, the last year in which we have consumption data. The tenth best model contains the market factor, and in addition it contains the investment, profit growth, and size theme factors. The top model contains the market, and in addition the profit growth, profitability, and value theme factors. Adding up across the rows of the table produces the number of times each of the 13 Jensen et al. (2023) theme factors appears in the top ten models.

The bottom panel of the table shows the same analysis, but for a version of (16) which includes only the market and SVAR test types. These tests are from December 2023 because we have more recent non-consumption data. Similar to the top panel, the two most prominent theme factors are profit growth and profitability, with investment, low leverage, size, value, and quality all in second place by prominence in both panels.

Table A6 of the Online Appendix provides more granularity into the top model identified in Table 7. It decomposes the $MCI(\phi)$ statistic in (16) into the eight $MCI(\phi, \tau, \mathcal{A})$s (i.e., the four test types times the two sets of test assets) from (9), and shows the top JKP4 model for each such test type-test asset pair. Again, profit growth and profitability feature as the two most prominent theme factors. Prior work has established the importance of profitability measures in explaining the cross section of stock returns, including Novy-Marx and Medhat (2025), Kogan et al. (2023), Bouchaud et al. (2019), and references therein. Our findings complement earlier work by arriving at these measures through a different path, focusing on sign consistency and model stability.

The bottom panel of Table 7 shows the MCIs of top ten MCI models, as well as several other measures of model fit: the R^2s from time-series (first stage) or cross-sectional (second stage, equation 7) Fama-MacBeth regressions, as well as the GRS test statistics for whether first-stage alphas are jointly zero. All of these are derived from regressions of test portfolios on factor models.[25] For each test statistic, Table 7 also shows the fraction of all 286 JKP4 models with worse fits, i.e., smaller R^2s and larger GRS statistics. High MCI models tend to have reasonably high R^2s and low GRS test statistics, though with a good deal of variation. We conclude that our MCI measure produces a model ranking which is distinct from the rankings produced by time-series and cross-sectional R^2s and GRS statistics respectively.

[25]While the GRS test and cross-sectional R^2's are computed at the same level of aggregation as our MCI (by model and test asset set), the time-series R^2's are averaged across all test assets (25 per test asset set) for each model.

8 Conclusion

Using the idea from Maio and Santa-Clara (2012) that state variables that forecast positive market returns should have positive risk premia, we introduce a new measure of ICAPM consistency for factor models. We show that consumption- and market-based versions of the measure lead to similar conclusions. Our consistency measure is associated with less alpha persistence, more beta persistence, and greater hedging efficacy, which we collectively refer to as model stability. Our approach can be used to identify factor models which have historically been more ICAPM consistent, as well as factors which are frequently present in high consistency models. Combining these insights, the models identified by our approach as being ICAPM consistent are likely to be stable in the future.

Unlike other factor model testing approaches, which typically evaluate whether a given factor model satisfies a theoretically predicted set of moment conditions, our approach is normative. Market practitioners will be well served to use more ICAPM consistent factor models in their work.

References

Barillas, F. and Shanken, J. (2017). Which Alpha? *The Review of Financial Studies*, 30(4):1316–1338.

Barroso, P., Boons, M., and Karehnke, P. (2021). Time-varying state variable risk premia in the ICAPM. *Journal of Financial Economics*, 139(2):428–451.

Boons, M. (2016). State variables, macroeconomic activity, and the cross section of individual stocks. *Journal of Financial Economics*, 116(3):489–511.

Bouchaud, J., Krüger, P., Landier, A., and Thesmar, D. (2019). Sticky expectations and the profitability anomaly. *Journal of Finance*, 74(2):639–674.

Brav, A., Constantinides, G., and Geczy, C. (2002). Asset pricing with heterogeneous consumers and limited participation: Empirical evidence. *Journal of Political Economy*, 110:793–824.

Bryzgalova, S., Huang, J., and Julliard, C. (2022). Bayesian solutions for the factor zoo: We just ran two quadrillion models. *Journal of Finance*, 78(1):487–557.

Campbell, J. and Vuolteenaho, T. (2004). Bad beta, good beta. *American Economic Review*, 94:1249–1275.

Campbell, J. Y. (1993). Intertemporal asset pricing without consumption data. *The American Economic Review*, 83(3):487–512.

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

Chabi-Yo, F., Gonçalves, A. S., and Loudis, J. A. (2025). An intertemporal risk factor model. *Management Science*, 71(8):6518–6544.

Cochrane, J. and Piazzesi, M. (2005). Bond risk premia. *American Economic Review*, 95:138–160.

Cochrane, J. H. (2009). *Asset pricing: Revised edition*. Princeton university press.

Constantinides, G. M. and Duffie, D. (1996). Asset pricing with heterogeneous consumers. *Journal of Political Economy*, 104(2):219–240.

Cooper, I., Ma, L., Maio, P., and Philip, D. (2021). Multifactor models and their consistency with the APT. *The Review of Asset Pricing Studies*, 11(2):402–444.

Cooper, I. and Maio, P. (2018). Asset growth, profitability, and investment opportunities. *Management Science*, 65(9):3988–4010.

Didisheim, A., Ke, S., Kelly, B., and Malamud, S. (2025). APT or "AIPT"? The surprising dominance of large factor models. working paper.

Fama, E. and French, K. (1992). The cross-section of expected stock returns. *Journal of Finance*, 47:427–465.

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

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3):607–636.

Feng, G., Giglio, S., and Xiu, D. (2020). Taming the factor zoo: A test of new factors. *Journal of Finance*, 75(3):1327–1370.

Gibbons, M. R., Ross, S. A., and Shanken, J. (1989). A test of the efficiency of a given portfolio. *Econometrica*, 57(5):1121–1152.

Giglio, S., Kelly, B., and Xiu, D. (2022). Factor Models, Machine Learning, and Asset Pricing. *Annual Review of Financial Economics*, 14(1):337–368.

Gonçalves, A. S. (2021). Reinvestment risk and the equity term structure. *Journal of Finance*, 76(5):2153–2197.

Hahn, J. and Lee, H. (2006). Yield spreads as alternative risk factors for size and book-to-market. *Journal of Financial and Quantitative Analysis*, 41:245–269.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054.

Hansen, L. P. and Jagannathan, R. (1997). Assessing specification errors in stochastic discount factor models. *The Journal of Finance*, 52(2):557–590.

Harvey, C. R. and Liu, Y. (2021). Lucky factors. *Journal of Financial Economics*, 141(2):413–435.

Jacquier, A., Muhle-Karbe, J., and Mulligan, J. (2025). In-sample and out-of-sample sharpe ratios for linear predictive models. *Quantitative Finance*.

Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023). Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518.

Kan, R. and Robotti, C. (2008). Model comparison using the hansen-jagannathan distance. *The Review of Financial Studies*, 22(9):3449–3490.

Kan, R., Robotti, C., and Shanken, J. (2013). Characteristics are covariances: A unified model of risk and return. *Journal of Finance*, 68(6):2617–2649.

Kan, R., Wang, X., and Zheng, X. (2024). In-sample and out-of-sample sharpe ratios of multi-factor asset pricing models. *Journal of Financial Economics*, 155:103837.

Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524.

Kogan, L., Li, J., and Zhang, H. H. (2023). Operating hedge and gross profitability premium. *The Journal of Finance*, 78(6):3387–3422.

Koijen, R., Lustig, H., and Van Nieuwerburgh, S. (2017). The cross-section and time series of stock and bond returns. *Journal of Monetary Economics*, 88:50–69.

Kozak, S., Nagel, S., and Santosh, S. (2018). Interpreting factor models. *The Journal of Finance*, 73(3):1183–1223.

Kozak, S., Nagel, S., and Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, 135(2):271–292.

Lettau, M. and Pelger, M. (2020). Factors that fit the time series and cross-section of stock returns. *The Review of Financial Studies*, 33(5):2274–2325.

Lewellen, J., Nagel, S., and Shanken, J. (2010). A skeptical appraisal of asset-pricing tests. *Journal of Financial Economics*, 96:175–194.

Li, H., Xu, Y., and Zhang, X. (2010). Evaluating asset pricing models using the second hansen-jagannathan distance. *Journal of Financial Economics*, 97(2):279–301.

Lin, Q. (2022). Is economic uncertainty a valid intertemporal capm state variable? *Review of Asset Pricing Studies*, 12(4):999–1040.

Maio, P. and Santa-Clara, P. (2012). Multifactor models and their consistency with the ICAPM. *Journal of Financial Economics*, 106(2):586–613.

Malloy, C. J., Moskowitz, T. J., and Vissing-Jørgensen, A. (2009). Long-run stockholder consumption risk and asset returns. *The Journal of Finance*, 64(6):2427–2479.

Merton, R. (1973). An intertemporal capital asset pricing model. *Econometrica*, 41:867–887.

Newey, W. and West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–08.

Novy-Marx, R. and Medhat, M. (2025). Profitability retrospective: What have we learned? Working Paper 33601, National Bureau of Economic Research.

Pástor, L. and Stambaugh, R. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy*, 111:642–685.

Petkova, R. (2006). Do the fama–french factors proxy for innovations in predictive variables? *Journal of Finance*, 61:581–612.

Raponi, V., Robotti, C., and Zaffaroni, P. (2019). Testing beta-pricing models using large cross-sections. *The Review of Financial Studies*, 33(6):2796–2842.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–360.

Welch, I. and Goyal, A. (2007). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.

Stockholder Consumption Growth and Cross-sectional Variance

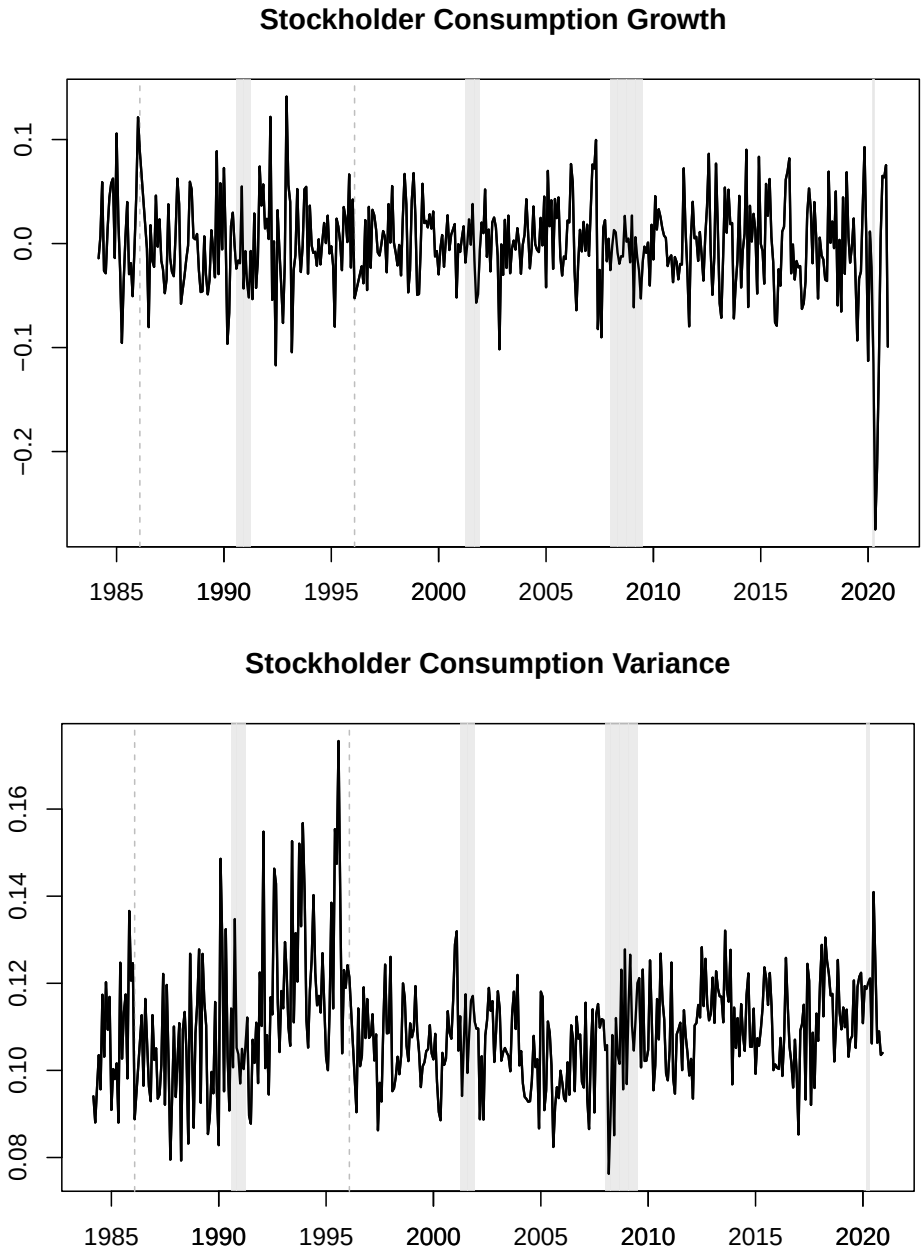


Figure 1: This figure shows the stockholder consumption growth series and the cross-sectional variance in stockholder consumption growth computed from the Bureau of Labor Statistics Consumer Expenditure Survey (CES). Each monthly observation refers to consumption growth relative to three months prior or the variance thereof. The sample period covered ranges from March 1984 to December 2020. NBER dated recessions are marked in gray bars and dotted vertical lines indicate times in which the survey experienced a break. Breaks correspond to the years 1986 and 1996 where we are missing observations for March, April and May.

Scatter plots of cosine similarity one year into the future versus market-based MCI in 1977

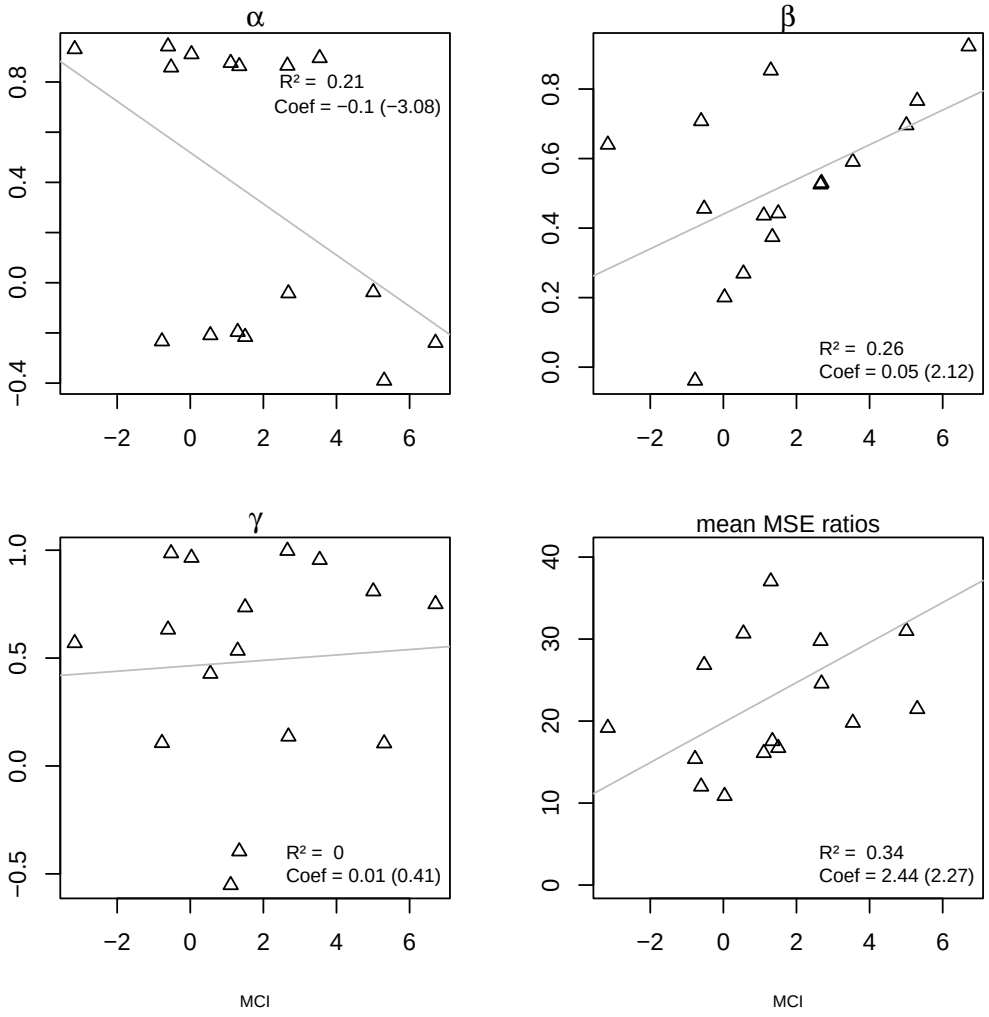


Figure 2: These scatter plots show outcome variables plotted against the market-based model consistency index (MCI) for the Maio and Santa-Clara (2012) models. Future outcomes are compared to their respective values one year prior, i.e. the last year used to calculate the MCI. Each plot has 16 data points corresponding to the outcomes and MCIs of eight factor models and two sets of test assets. The outcome measures the cosine similarity of regression alphas, betas, Fama-MacBeth gammas, and the ratio of mean-square errors using monthly data. For each coefficient estimate we report the corresponding t-statistic based on White standard errors in parentheses.

Time series plots of regression coefficients of the market-based MCI predicting cosine similarity one year into the future

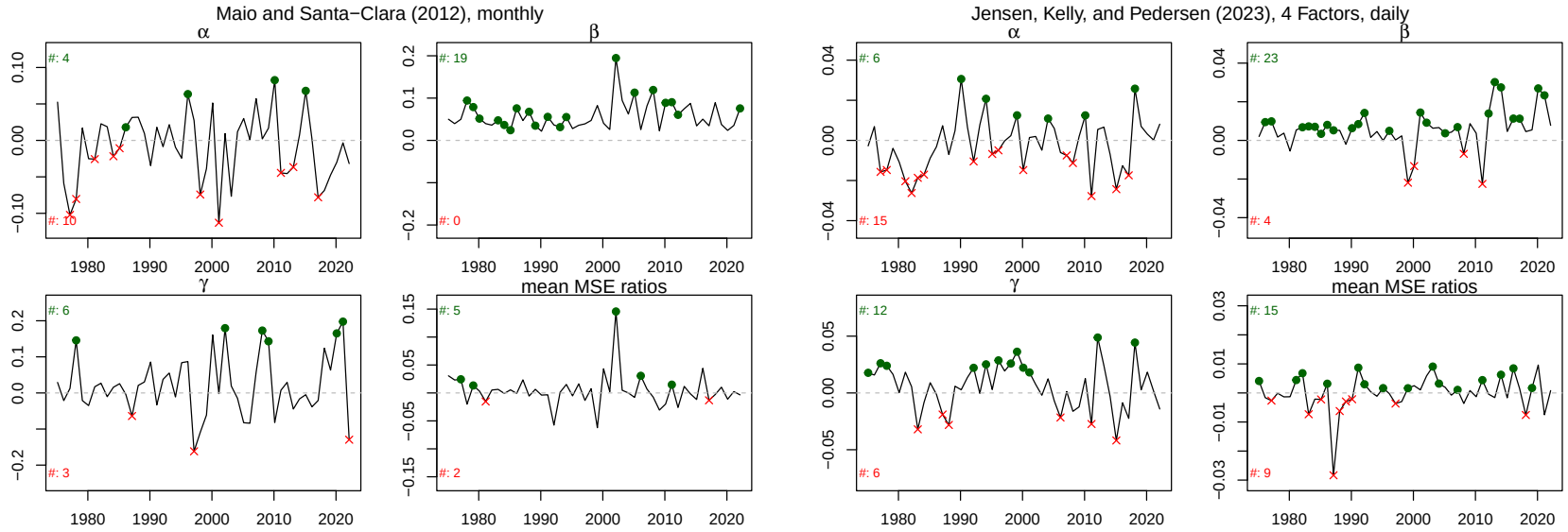


Figure 3: These time series plots show the regression coefficient on the market-based model consistency index (MCI) forecasting outcomes one year into the future for MSC (2012) and JKP4 models for each year. Future outcomes are compared to their respective values one year prior, i.e. the last year used to calculate the MCI. Positive significant values are indicated in green above the dotted line and negative significant ones in red below. The results use White standard errors. The outcome measures the cosine similarity of regression alphas, betas, Fama-MacBeth gammas, and the ratio of mean-square errors using monthly data for MSC (2012) models and daily data for JKP4 models. For MSC (2012) models we are limited using monthly data as some factors were not available on a daily basis.

Counts of positive and negative significant coefficients in rolling regressions predicting cosine similarity one year into the future using MCI

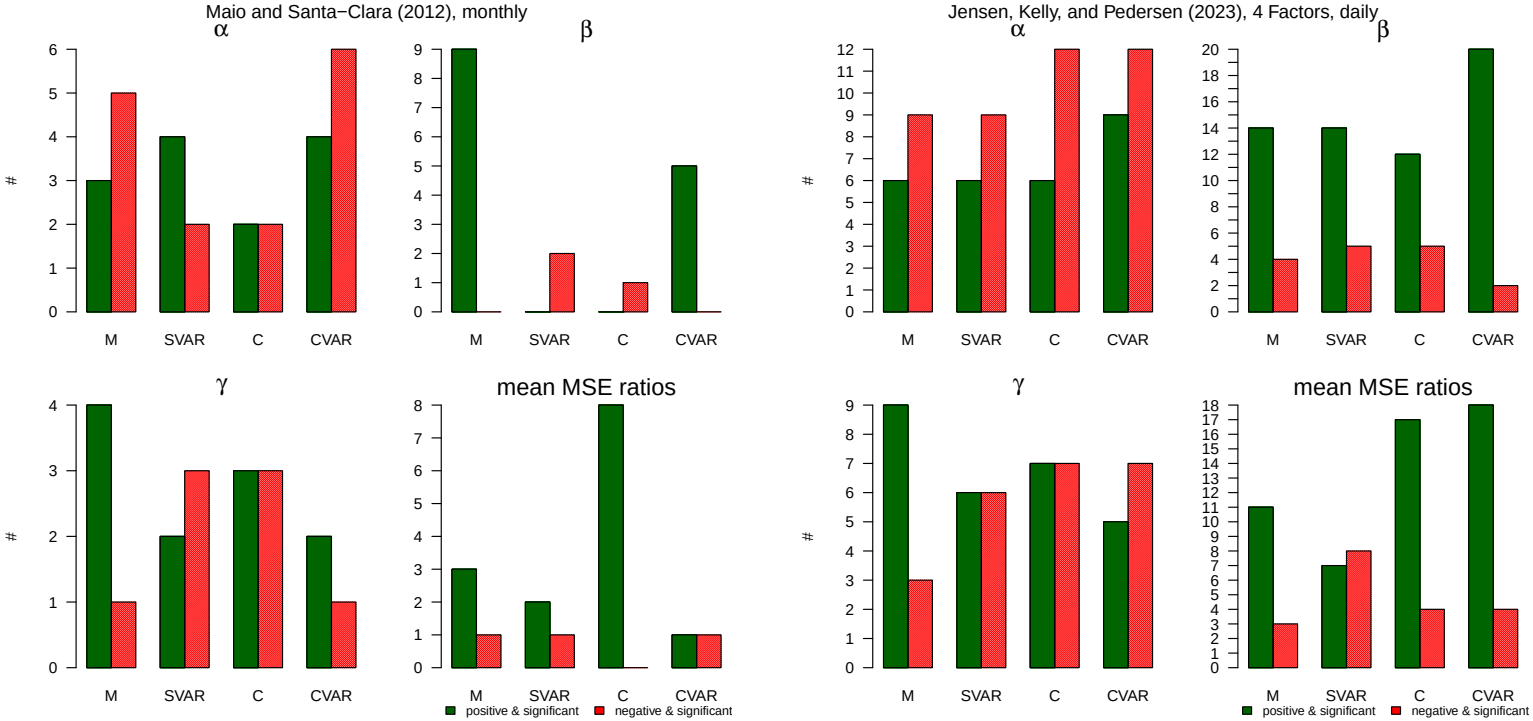


Figure 4: This figure reports the number of positive and negative regression coefficients predicting outcomes one year into the future for MSC (2012) and JKP4 models. We only count significant coefficients using White standard errors. We consider four variants of model consistency, two related to the market's return and two related to stockholder consumption. Future outcomes are compared to their respective values one year prior, i.e. the last year used to calculate model consistency. The outcome measures the cosine similarity of regression alphas, betas, Fama-MacBeth gammas, and the ratio of mean-square errors using monthly data for MSC (2012) models and daily data for JKP4 models. The time period is adjusted for the market (and SVAR) based consistency to allow for comparison with the consumption (and CVAR) based consistency index and thus ranges from 1990 to 2020. For MSC (2012) models we are limited using monthly data as some factors were not available on a daily basis.

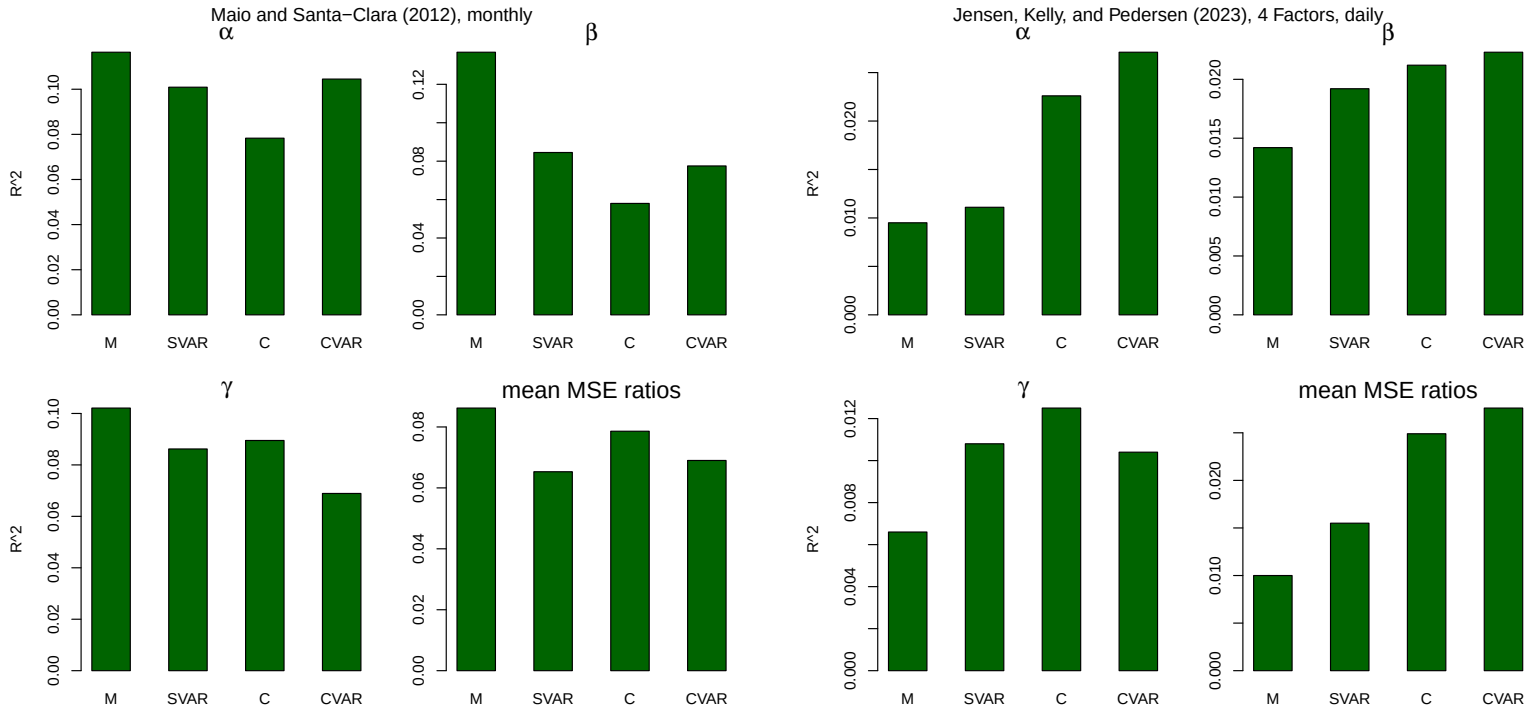


Figure 5: The bar plots report the average R^2s of the regressions of one-year stability measures on MCIs of the MSC (2012) and JKP4 models using the four ICAPM test types $\tau \in \{M, SVAR, C, CVAR\}$. Future outcomes are compared their respective values one year prior, i.e. the last year used to calculate model consistency. The outcome measures the cosine similarity of regression alphas, betas, Fama-MacBeth gammas, and the ratio of mean-square errors using monthly data for MSC (2012) models and daily data for JKP4 models. Also here, the time period is adjusted for the market (and SVAR) based consistency to allow for comparison with the consumption (and CVAR) based consistency index and thus ranges from 1990 to 2020. For MSC (2012) models we are limited using monthly data as some factors were not available on a daily basis.

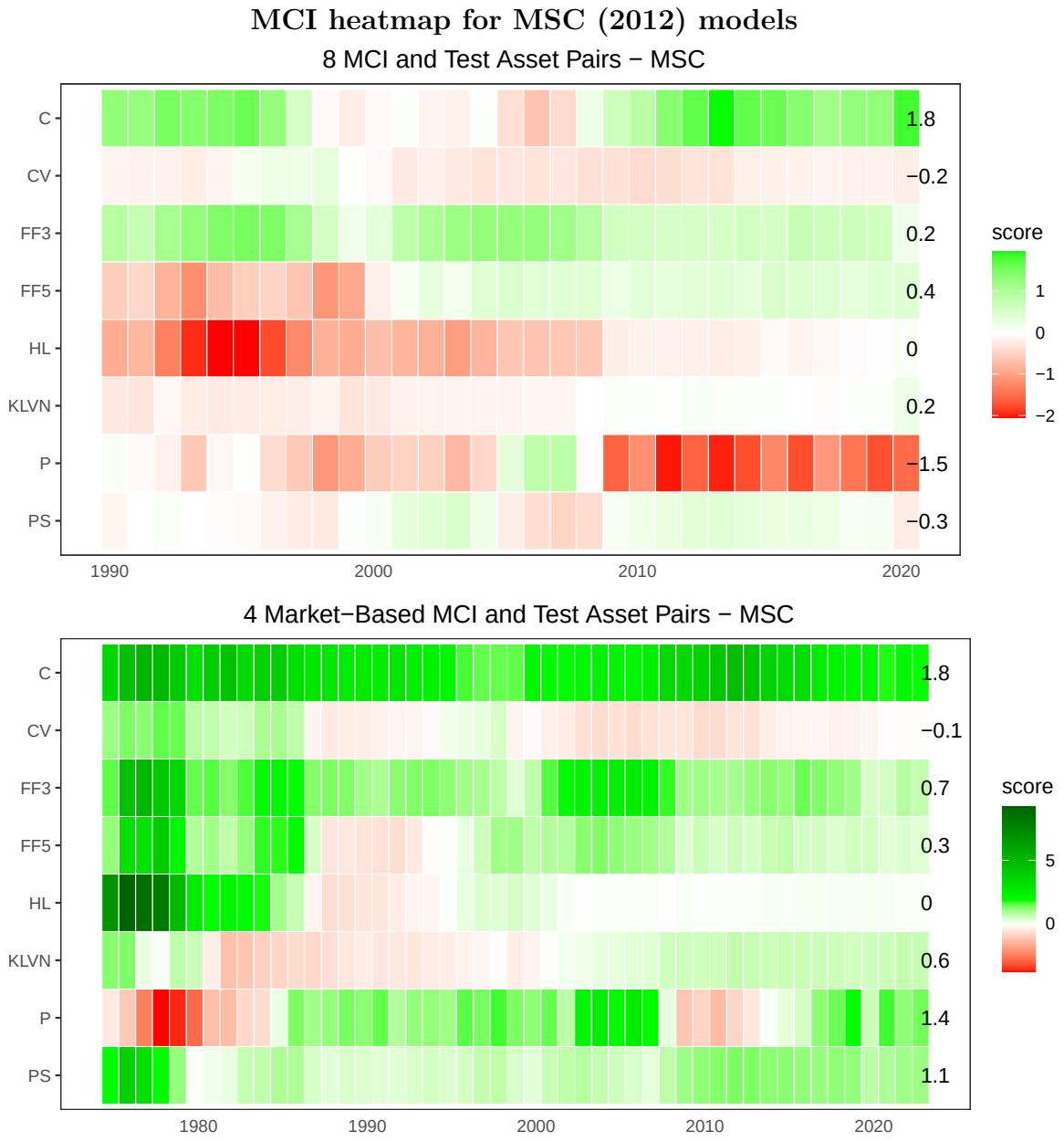


Figure 6: The heatmaps above show the MCI for each year and MSC (2012) model. The first subfigure computes the MCI over its four variants and both sets of test assets (eight pairs) by averaging across those. The second subfigure computes the MCI over its two market-based variants and both sets of test assets again by averaging. Due to the shorter time period for which we have available consumption data, the second subplot contains more yearly observations. MCIs are computed using all available past data up to the end of a given year while requiring 7 full years of past data for all factors and factor-derived state variables. The column at the end indicates the final MCI value.

Heatmap counting how often a given factor theme is included among the top 10 most consistent JKP4 models

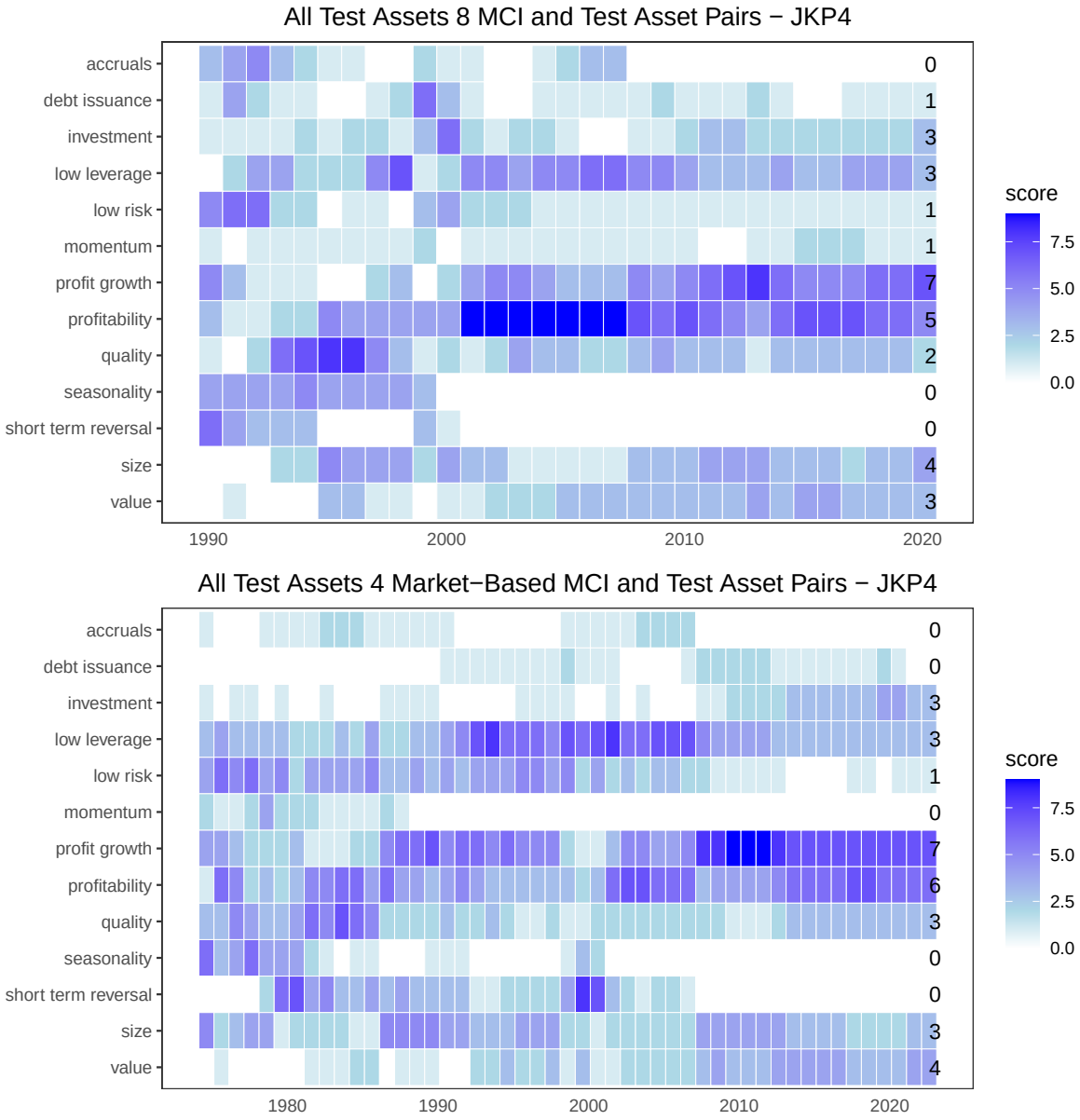


Figure 7: The heatmaps above present the count for any given year of how often a given factor theme occurs in the top 10 most consistent models among JKP4 models. Those most consistent models are computed across either all 8 MCI and Test Asset Pairs or across 4 Market-Based MCI and Test Asset Pairs. Due to the shorter time period for which we have available consumption data, the second subplot contains more yearly observations. MCIs are computed using all available past data and up to the end of a given year while requiring 7 full years of past data for all factors and factor-derived state variables. The last column indicates the most recent count computed for each combined MCI (calculated over the full sample period).

Table 1: Data Summary

This table summarizes the factor data used in this paper. The factors we use either stem from the set of factors used in Maio and Santa-Clara (2012) or the 13 factor themes in Jensen, Kelly and Pedersen (2023). Panel A summarizes the former and Panel B the latter.

Panel A : Maio and Santa-Clara (2012) Factor Data (monthly levels)

	MSC (2012) sample: 06/1963 - 12/2008					Full sample: 06/1963 - 12/2023				
Statistic	N	Mean	St. Dev.	Min	Max	N	Mean	St. Dev.	Min	Max
TERM	546	0.008	0.011	-0.031	0.033	726	0.010	0.012	-0.031	0.034
DEF	546	0.010	0.004	0.003	0.034	726	0.010	0.004	0.003	0.034
DY	546	-3.549	0.417	-4.502	-2.775	726	-3.651	0.411	-4.502	-2.775
RF	546	0.005	0.002	0.000	0.014	726	0.004	0.003	0.000	0.014
PE	546	2.789	0.408	1.915	4.077	726	2.874	0.432	1.915	4.818
VS	546	1.569	0.153	1.283	2.128	726	1.624	0.185	1.283	2.206
CP	546	0.010	0.016	-0.052	0.069	726	0.008	0.014	-0.050	0.057
SMB^*	546	-0.138	0.215	-0.796	0.223	726	-0.182	0.281	-1.316	0.223
HML^*	546	-2.846	1.061	-6.007	-0.992	726	-3.352	1.498	-8.557	-0.992
CUMD	546	0.512	0.198	-0.135	1.090	726	0.389	0.292	-0.498	1.090
L	546	-0.003	0.057	-0.392	0.277	726	-0.0001	0.056	-0.392	0.277
CL	487	-0.193	0.575	-1.610	0.667	667	-0.018	0.645	-1.610	1.320

Panel B: Jensen, Kelly and Pedersen (2023) Factor Themes (% per month)

	June 1963 - December 2023					
Statistic	N	Mean	St. Dev.	Min	Max	L/S
accruals	726	0.248	1.160	−3.880	4.710	−
debt issuance	726	0.221	0.705	−2.640	6.640	−
investment	726	0.275	2.120	−10.300	14.600	−
low leverage	726	0.009	2.960	−18.400	28.700	+
low risk	726	0.141	3.750	−26.700	22.100	+
momentum	726	0.360	3.190	−22.200	23.200	+
profit growth	726	0.171	1.090	−6.750	5.330	+
profitability	726	0.227	2.110	−14.800	16.400	+
quality	726	0.286	1.550	−6.820	6.140	+
seasonality	726	0.129	0.561	−2.330	2.580	+
short term reversal	726	0.143	1.200	−8.360	11.100	−
size	726	0.109	2.280	−9.230	14.100	−
value	726	0.341	3.240	−28.100	21.100	+

+/− indicates a long/short position in the characteristic

Table 2: Summary of Test Types

This table summarizes the outcome variables used in (8) which determine test types $\tau \in \{M, MVAR, C, CVAR\}$.

τ	Explanation of test type
M	Cumulative return of the CRSP value-weighted index from month t to $t+h$. These are computed on a daily or monthly basis.
SVAR	Return variance of the S&P 500 index from month t to $t+h$. These are computed using daily data.
C	Growth of stockholder consumption from month $t-3$ to $t+h$ as explained in Section A2 of the Online Appendix.
CVAR	Variance of cross-sectional consumption growth of shareholders from month $t-3$ to $t+h$.

Table 3: Example MCI Calculation

This table presents an example for computing one variant of the MCI, the market based MCI for test asset set A, for the year 1985 and the Fama and French (1992) three-factor model. Test asset set A refers to the 25 portfolios sorted on size and book-to-market ratio available on Kenneth French's data library. The MCI is computed from the Fama-McBeth risk prices of all non-market factors in Panel A as well as the forecasting coefficients of the same factors with respect to the market returns cumulated over $h = 1, 12, 60$ months into the future in Panel B. The equation below the two panels computes the MCI from (9) with $\phi = FF3$, test type $\tau = M$ (market), and test asset set $\mathcal{A}_{s,v}$.

Panel A: Risk Prices

Model	Test Assets	Factor	Risk Price	t-statistic
FF3	A	SMB	3.45	1.51
FF3	A	HML	10.13	4.31

Panel B: Forecasting Regression Coefficients

Model	Factor	Horizon	Coefficient	t-statistic
FF3	SMB	1	-0.003	-0.14
FF3	SMB	12	0.15	0.79
FF3	SMB	60	0.16	1.22
FF3	HML	1	0.01	1.99
FF3	HML	12	0.06	1.82
FF3	HML	60	0.28	18.51

MCI Calculation

$MCI(FF3, M, \mathcal{A}_{s,v})$=[1.51(-0.14+0.79+1.22)+4.31(1.99+1.82+18.51)]/6=16.5

Table 4: Correlation Between MCIs

This table shows the cross-sectional correlation among MCIs in 2020 paired by set of test assets both for the models in Maio and Santa-Clara (2012) as well as the JKP4 models explained in Section 4.1. We chose 2020 because it is the last year for which we observe all four MCIs, which are calculated using all information until that year. Each test type is associated with an 8×2 MCI vector in the left panel and 286×2 MCI vector in the right panel.

	MSC (2012) Models				JKP4 Models			
	M	SVAR	C	CVAR	M	SVAR	C	CVAR
M		0.295	0.350	−0.029		0.589	0.477	−0.139
SVAR			0.562	0.269			0.464	0.100
C				0.677				0.251
CVAR								

Table 5: Model Consistency Index (MCI) versus MSC (2012)

This table shows for each model and test asset pair examined in Maio and Santa-Clara (2012) if the authors concluded that the model was consistent with the ICAPM (✓) or if they found evidence of inconsistency (✗). The column M-MSC indicates that this conclusion was derived by the authors based on the arguments regarding (future) market returns, while M-SVAR refers to the an equivalent conclusion based on (future) market variance. We contrast the findings in Maio and Santa-Clara's (2012) with our market (M) and market variance (SVAR) based MCIs both over the same sample period as well as over the full sample at our disposal.

		MSC sample period: June 1963 - December 2008				June 1963 - December 2023	
Model	Test Assets	$MCI(M)$	$MSC(M)$	$MCI(SVAR)$	$MSC(SVAR)$	$MCI(M)$	$MCI(SVAR)$
HL	$\mathcal{A}_{s,v}$	0.681	✗	−0.025	✗	2.180	−1.090
P	$\mathcal{A}_{s,v}$	0.067	✗	1.040	✗	2.490	−1.500
CV	$\mathcal{A}_{s,v}$	2.440	✗	0.250	✗	3.430	1.230
KLVN	$\mathcal{A}_{s,v}$	0.935	✓	−0.496	✗	1.690	−1.740
FF3	$\mathcal{A}_{s,v}$	4.640	✓	6.220	✓	2.280	4.410
C	$\mathcal{A}_{s,v}$	4.230	✓	3.810	✗	0.965	3.010
PS	$\mathcal{A}_{s,v}$	2.460	✓	1.360	✗	2.110	1.930
FF5	$\mathcal{A}_{s,v}$	2.280	✗	0.535	✓	2.810	−1.580
HL	$\mathcal{A}_{s,m}$	−2.530	✗	1.620	✗	−2.880	1.750
P	$\mathcal{A}_{s,m}$	−1.600	✗	1.430	✗	2.690	1.880
CV	$\mathcal{A}_{s,m}$	−4.050	✗	−0.942	✗	−3.660	−1.430
KLVN	$\mathcal{A}_{s,m}$	1.250	✗	0.474	✗	0.320	2.220
FF3	$\mathcal{A}_{s,m}$	−0.203	✗	−3.800	✗	0.120	−4.200
C	$\mathcal{A}_{s,m}$	3.760	✓	2.330	✗	0.588	2.690
PS	$\mathcal{A}_{s,m}$	−0.065	✗	−0.809	✗	0.284	−0.084
FF5	$\mathcal{A}_{s,m}$	0.021	✗	0.561	✗	−0.193	0.311

Table 6: Net Number of Positive minus Negative Forecasting Significant Coefficients

Each cell shows the net number of significant positive and negative coefficients from (15) for alpha, beta, MSE, and gamma stability across different horizons and factor model sets for the 4 ICAPM test types run in annual expanding windows as explained in Section 6. P-values are based on the simulation study explained in Section 6.1.The analysis for MSC factors uses monthly data while analysis of the JKP factors uses daily data.

	α cosine sim	β cosine sim	mean MSE ratios	γ cosine sim
pre 3 post 1 - MSC	2 (0.79)	24 (0)	12 (0)	9 (0.02)
pre 2 post 1 - MSC	-6 (0.09)	22 (0)	5 (0.48)	7 (0.06)
pre 1 post 1 - MSC	-6 (0.09)	19 (0)	11 (0)	3 (0.46)
pre 3 post 1- JKP3	-11 (0.2)	45 (0.01)	-2 (0.77)	-1 (0.89)
pre 2 post 1- JKP3	-12 (0.2)	41 (0)	23 (0.02)	-2 (0.79)
pre 1 post 1- JKP3	-13 (0.18)	48 (0.01)	41 (0.03)	6 (0.42)
pre 3 post 1- JKP4	-23 (0.04)	74 (0.01)	16 (0.02)	1 (0.89)
pre 2 post 1- JKP4	-26 (0)	75 (0.01)	23 (0.06)	1 (0.82)
pre 1 post 1- JKP4	-23 (0)	64 (0.01)	37 (0.03)	5 (0.11)

Table 7: JKP4 - Themes in Top 10 Most Consistent Models

This table shows the themes included in the top 10 most consistent four-factor models constructed using three themes from the 13 factor themes in Jensen et al. (2023), as well as the market. The top 10 most consistent models refer to those factor model for which we observe the highest MCI averaged across all 8 MCI and test asset pairs using data up to December 2020 in Panel A, or all 4 market-based MCI and test asset pairs using data up to December 2023 in Panel B. Checkmarks (✓) indicate that a given theme is part of the factor model selected as one of the ten most consistent factor models according to the combined MCI. Columns respond to the ranking of the models by MCI of the model consisting of the market and the three checkmarked themes. In the bottom panel, we report–at the same level of aggregation–the associated average R^2s for cross-sectional and time-series regressions and the GRS test statistic, as explained in Section 7.1. The fraction of the 286 JKP4 models with lower R^2s or higher GRS statistics. Finally, the last column counts the occurrence of a given theme in the top 10 most consistent models.

Panel A: 8 MCI and Test Asset Pairs

Theme/Rank	10	9	8	7	6	5	4	3	2	1	total
accruals											0
debt issuance									✓		1
investment	✓	✓					✓				3
low leverage				✓				✓	✓		3
low risk						✓					1
momentum				✓							1
profit growth	✓		✓			✓	✓	✓	✓	✓	7
profitability		✓		✓	✓		✓			✓	5
quality		✓			✓						2
seasonality											0
short term reversal											0
size	✓		✓			✓		✓			4
value			✓		✓					✓	3
MCI	3.63	3.67	3.83	3.89	3.95	4.03	4.10	4.16	4.26	5.04	
CS-R^2	0.968 (0.81)	0.977 (0.97)	0.968 (0.80)	0.967 (0.78)	0.967 (0.78)	0.967 (0.77)	0.964 (0.70)	0.965 (0.74)	0.976 (0.94)	0.966 (0.75)	
TS-R^2	0.86 (0.93)	0.80 (0.50)	0.87 (0.98)	0.82 (0.72)	0.83 (0.74)	0.86 (0.90)	0.81 (0.59)	0.87 (0.98)	0.79 (0.37)	0.82 (0.70)	
GRS	3.26 (0.83)	2.75 (0.94)	3.76 (0.65)	4.51 (0.35)	3.71 (0.67)	3.52 (0.73)	3.34 (0.81)	4.70 (0.26)	3.03 (0.89)	3.70 (0.68)	

Panel B: 4 Market-Based MCI and Test Asset Pairs

Theme/Rank	10	9	8	7	6	5	4	3	2	1	total
accruals											0
debt issuance											0
investment			✓		✓				✓		3
low leverage				✓		✓				✓	3
low risk	✓										1
momentum											0
profit growth	✓	✓	✓	✓	✓	✓	✓				7
profitability					✓	✓	✓	✓	✓	✓	6
quality								✓	✓	✓	3
seasonality											0
short term reversal											0
size		✓	✓	✓							3
value	✓	✓					✓	✓			4
MCI	3.14	3.15	3.25	3.32	3.93	4.37	4.54	4.58	4.61	5.36	
CS-R^2	0.962 (0.63)	0.968 (0.80)	0.968 (0.81)	0.965 (0.74)	0.964 (0.70)	0.961 (0.59)	0.966 (0.75)	0.967 (0.78)	0.977 (0.97)	0.971 (0.86)	
TS-R^2	0.83 (0.75)	0.87 (0.98)	0.86 (0.93)	0.87 (0.98)	0.81 (0.59)	0.81 (0.65)	0.82 (0.70)	0.83 (0.74)	0.80 (0.50)	0.81 (0.66)	
GRS	3.77 (0.65)	3.76 (0.65)	3.26 (0.83)	4.70 (0.26)	3.34 (0.81)	4.86 (0.21)	3.70 (0.68)	3.71 (0.67)	2.75 (0.94)	4.50 (0.35)	