It Takes Two to Tango: Economic Theory and Model Uncertainty for Equity Premium Prediction*

Daniele Bianchi† Alexandre Rubesam‡ Andrea Tamoni§

First draft: July 1, 2022 This version: January 26, 2024

Abstract

We assess whether data-driven statistical methods and, in particular, forecast combination strategies can provide additional information about expected market returns beyond that of theoretically motivated predictors. The results indicate that averaging forecasts from the theoretically motivated predictors and combination strategies enhances prediction accuracy relative to using each forecasting approach individually. Our findings demonstrate that flexible statistical methods could be used to *boost* economic theory rather than *dilute* its importance for equity premium predictability. Yet, forecast combination approaches can extract additional information and no theoretical predictor in isolation is likely to be *the* expected return on the market.

Keywords: Equity premium; Return predictability; Forecast combination; Encompassing forecasts; Model complexity; Shrinkage.

JEL codes: C53, C22, G11, G12.

*The authors thank conference participants at the 2023 FMA Annual meeting (Chicago) and the 2023 QuantMinds International (London), as well as seminar participants at the Eli Broad College of Business (Michigan State), the Finance Hub seminar (São Paulo/Online), and EDHEC Business School (Lille).

†School of Economics and Finance, Queen Mary University of London, Mile End, London E1 4NS. E-mail: d.bianchi@qmul.ac.uk. Web: whitesphd.com.

‡Department of Finance, IESEG School of Management, France. E-mail: a.rubesam@ieseg.fr. Web: https://www.rubesam.com/.

§Rutgers Business School. Address: Department of Finance, 1 Washington Pl, Newark, NJ 07102. E-mail: andrea.tamoni.research@gmail.com. Web: https://www.andreatamoni.com.

1 Introduction

In an important contribution, Welch and Goyal (2008) document that several well-known candidate predictors display little ability to forecast the equity premium out-of-sample when taken in isolation. Since then, a host of papers have proposed alternative economic mechanisms leading to new variables designed to predict future excess market returns (e.g., Polk et al., 2006; Ferreira and Santa-Clara, 2011; Martin, 2017; Campbell, 2018). These approaches impose economically-motivated model restrictions on either the predictors to be used, the value of the regression coefficients, or both, producing parsimonious models with good forecasting performance. Importantly, these restrictions are firmly anchored on theoretical foundations, such as the present-value identity or the capital asset pricing model (CAPM).[1] At the same time, a rapidly expanding literature has introduced flexible statistical techniques designed to handle the challenging problem of forecasting returns with a wealth of predictors (Elliott et al., 2013; Rossi, 2018; Kelly and Xiu, 2023).

Perhaps surprisingly, statistical models are rarely compared with theoretically-motivated models, and vice versa. Such a comparison is crucial to gauge whether a theoretical predictor truly represents *the* expected market return. It also quantifies the information potentially overlooked when dismissing a specific theoretically-backed predictor. In this paper, we fill this gap by comparing a handful of simple theoretically-motivated models against data-driven methods that do not prioritize any specific economic theory. We devote particular attention to forecast combination methods that make indistinct use of all available predictors.

Our contribution is twofold. First, from a methodological perspective, we extend the *complete subset regressions* (CSR) proposed by Elliott et al. (2013) to account for the un-

[1] In the remainder of this paper, we refer to models that impose economically motivated restrictions as "theoretically motivated" models for brevity. We define precisely the meaning of this term in Section 2.2.

certainty about the size of the forecasting model.[2] Specifically, we introduce a probabilistic approach, based on the *model confidence set* (MCS) of Hansen et al. (2011), to select multiple models of different sizes that produce a statistically equivalent out-of-sample predictive performances. Our novel forecast combination approach outperforms other statistical methods and serves as a benchmark for theoretically-motivated return predictive regressions. Second, we provide an intuitive yet effective procedure to couple the strict parsimonious structure imposed by theoretically motivated models with the richness provided by forecast combination methods. Our approach takes an agnostic perspective as to which predictive framework is preferable while providing new insights into the role of a given economic theory for equity premium predictability. In particular, we show that, although no theoretically motivated model subsumes the forecast from CSRs, a simple average of the two forecasts is generally superior to each forecast taken individually.

In our empirical analysis, we consider an extensive set of 31 predictors for the equity premium. This is about twice the number of predictors considered by Welch and Goyal (2008) and comparable to Goyal et al. (2021). In addition to fundamental variables – such as the term and the default spreads – and technical indicators based on past prices and volume (Neely et al., 2014), we include several theoretically motivated predictors (see, e.g., Polk et al., 2006; Ferreira and Santa-Clara, 2011; Martin, 2017; Campbell, 2018). We also evaluate, within a model combination strategy, the gains from imposing economic restrictions such as positive forecast constraints (see, e.g., Campbell and Thompson, 2008). In addition to univariate models prescribed by theory, we consider popular penalised regression methods and conventional forecast combination techniques.

Several results stand out. First, we document that theoretically motivated models per-

[2]The size of a model is given by the number of included regressors (out of the many candidate predictors).

form well in terms of out-of-sample R^2 (henceforth, R^2_{OOS}). In particular, the "sum-of-parts" (SOP) proposed by Ferreira and Santa-Clara (2011) – which is based on a present-value identity – and the cross-sectional premium (CSP) proposed by Polk et al. (2006) – which combines cross-sectional CAPM restrictions into a Gordon growth dividend model – attain positive R^2_{OOS} of 1.46% and 1.42%, respectively, when forecasting the excess market returns one-month ahead. These R^2_{OOS} are higher than what is attained by penalized regression methods such as the lasso and elastic net. Nevertheless, we also find that CSR generally performs on par with simple theoretically motivated models, if not better in some instances. Notably, the choice of the dimension(s) of the models to combine and the imposition of economic constraints play an important role in obtaining such competitive performance, which requires qualification.

Implementing CSR requires selecting the dimension(s) of the models to combine. Given a total of p predictors, one can combine all models that include only $K = 1$ predictor, all the models that include $K = 2$ predictors, and so on. The simplest possibility is to choose a given K, e.g. $K = 1$, and average over all fixed-size models (e.g., Rapach et al., 2009). Alternatively, one can acknowledge K to be a hyper-parameter whose "optimal" value(s) is inherently uncertain. In this regard, Elliott et al. (2013) show that there exists a bias-variance trade-off for different values of K and propose a recursive procedure to choose a single value of K based on the cumulative out-of-sample mean squared error (MSE).

We build upon this intuition and propose a procedure to select *all values of* K that produce a statistically equivalent performance. Specifically, we apply the model confidence set approach of Hansen et al. (2011) to the CSR forecasts for all values of K. This yields a set of models that includes all the values of K for which we cannot reject the hypothesis of equal predictive ability. We then average the forecasts for all values of K in this set. Our

approach outperforms other forecast combination methods, including the simple combination proposed by Rapach et al. (2009) and a CSR that aggregates across all values of K.

This forecast combination approach – which, to the best of our knowledge, is new to the literature – performs remarkably well both in terms of R^2_{OOS} and economic utility gains. In particular, the CSR that averages *only* over the models with equivalent predictive ability delivers an R^2_{OOS} of about 1.87% before imposing economic constraints. In addition, the same model combination approach delivers an economic gain of 4.53% relative to the historical mean benchmark, which represents an increase of almost 1.5% relative to the Ferreira and Santa-Clara (2011) model, the best among the theoretically motivated models.

Importantly, the R^2_{OOS} of our preferred CSR raises to 2.23% after imposing economic restrictions on the sign of the equity premium. This is the largest value attained among all models considered. The CSR approach with sign restriction also delivers the highest utility gains, a fact that extends to the model averaging context the existing evidence on the value of economic restrictions on individual regressions (see, e.g., Pettenuzzo et al., 2014).

Overall, our findings underscore that, although theoretically motivated models often claim to be *the* expected market return, one cannot rule out a priori the key role of other variables in understanding the dynamics of the US equity premium. That is, model uncertainty is pervasive, regardless of how strong the argument might be in favour of a given economic theory of returns predictability. Yet, by the same token, combining a large set of models is unlikely to be a panacea: if one keeps adding noisy predictors to the model, averaging will not generate any predictive gain. Simply put, economic theory still matters.

Motivated by these results, we implement a series of forecast encompassing tests whereby we impose a dogmatic view on a given theoretical predictor and test the extent of the

additional information on market expected returns captured by our novel CSR approach. The results show that no individual theoretically motivated model incorporates all the relevant information on expected returns; that is, the forecasts implied by such models do not subsume the predictions obtained from our forecast combination approaches. This supports the view that no theoretical predictor is likely to be *the* expected return on the market. However, these results do not exclude that pooling forecasts from theoretically motivated predictors and data-driven methods may lead to more accurate predictions of the equity premium.

We test out-of-sample this hybrid forecasting approach and show that a simple average of the forecasts from theoretically motivated predictive regressions and CSRs substantially outperforms both predictions taken separately. For example, the baseline forecast using the Polk et al. (2006) model achieves an R^2_{OOS} of 0.67% and economic gains of 0.94% relative to the historical mean benchmark. A naive average of this baseline model with the prediction from our preferred CSR produces an R^2_{OOS} of 2.25% and an economic gain of 4% annually. Results using other theoretically motivated models also reveal similar statistical and economic improvements. Importantly, we are not only improving upon the theoretical predictors but also relative to the CSRs itself. For instance, combining the SOP predictor with our CSR implementation attains an R^2_{OOS} of 2.4% compared with the 1.8% of the same CSR approach, which does not exploit the information in the SOP.

We also show that the naive approach of combining theoretical forecasts with statistical ones by fixing the relative weight to 0.5 is superior to other combination approaches. The reason is that the optimal weight is extremely unstable when estimated in real time. Finally, we investigate what would be the optimal weight ex-post, and find a value of about 0.5 over the full sample from 1953 to 2021. This value increases to 0.7 over the more recent sample from 1996 to 2021. Although this evidence suggests that, over this period, theoretical

predictors are getting closer to be the expected return of the market, we find that this change is mostly attributable to a deterioration of the statistical benchmark over this recent period.

We conclude that theoretically motivated models are informative but far from being the expected return on the market. At the same time, averaging forecasts from all predictors, disregarding their possibly different economic underpinning, is unlikely to be the solution. Therefore, we recommend to decouple the parsimonious economic structure of theoretically motivated models from the agnostic view offered by data-driven methods such as complete subset regressions, and then to combine the two separate forecasts to model expected returns.

1.1 Related literature

A vast literature examines the ability of different fundamentals, sentiment, and technical factors to predict the aggregate excess market return.[3] The linear regression forecasting framework is a cornerstone in this literature, and the set of monthly candidate predictors of Welch and Goyal (2008) is widely used.

Early on, several studies have recognised that adding all predictors in a single regression results in poor out-of-sample predictive performance. Since then, the literature has uncovered several techniques that enhance the out-of-sample performance of predictive regressions.[4] The first consists of imposing economically motivated restrictions on the regression forecasts

[3] Another main strand in the literature pertains to the predictability in the cross-section of returns, where a host of characteristics (the so-called *factor zoo*) that purportedly predict returns has been identified, see e.g., McLean and Pontiff (2016); Green et al. (2017); Hou et al. (2020).

[4] We focus mostly on the literature that relies on standard predictive linear regression models. Another strand of the literature focuses on model instability. It considers regime switches, structural breaks, and time-varying coefficients; see, e.g., Paye and Timmermann (2006); Guidolin and Timmermann (2007); Lettau and Van Nieuwerburgh (2008); Pástor and Stambaugh (2009); Henkel et al. (2011); Dangl and Halling (2012). Cederburg et al. (2023) provides an economic framework to investigate the importance of accounting for time-varying volatility when forecasting the equity premium.

(Campbell and Thompson, 2008; Pettenuzzo et al., 2014; Li and Tsiakas, 2017), or relying on theoretical foundations to either preselect or design relevant predictors (e.g., Tsiakas et al., 2020; Kelly and Pruitt, 2013; Huang et al., 2015; Dong et al., 2022), or to restrict the form of the predictive model (Ferreira and Santa-Clara, 2011; Campbell, 2018; Martin, 2017).

A second conventional modelling tool is to regularise the model parameter estimates, i.e., penalised regressions such as lasso or ridge (Li and Tsiakas, 2017; Dong et al., 2022). Opposite to this, there are model averaging techniques. For instance, Avramov (2002) and Cremers (2002) show that Bayesian model averaging outperforms model selection approaches that attempt to choose a single best model. Similarly, Rapach et al. (2009) show that a parsimonious model that combines (i.e., averages) forecasts from univariate predictive models strongly outperforms all individual models.[5] Elliott et al. (2013) propose a *complete subset regression* approach – which we leverage in our paper – that combines forecasts from all possible linear regression models with a fixed number of predictors K. They show that combining a small set of predictors (up to 6) outperforms the historical mean and other competing approaches, such as shrinkage and Bayesian model averaging, with performance peaking at $K = 2$.[6] More recently, Giannone et al. (2021) proposed a Bayesian approach with a prior that allows for sparsity and shrinkage. They show that dense models and Bayesian model averaging produce the best results in the Welch and Goyal (2008) data set.[7]

[5] In the context of bond return predictability, Lin et al. (2018) propose using a linear combination of the combined forecast from univariate regressions and the historical average return.

[6] This empirical result aligns with those reported by Tsiakas et al. (2020), who show that several low-dimensional models perform well with 2 or 3 predictors. Tsiakas et al. (2020) explain the good performance of low-dimensional models based on the realization that some predictors work better during expansionary periods (e.g., the dividend yield), while others during recessions (e.g., the long-term return on government bonds and the term spread).

[7] We note, however, that comparison of their results with those from the literature is hampered by the fact that they work with annual observations only, and do not report metrics such as the R^2_{OOS}. In unreported analysis, we have done extensive tests with similar Bayesian models, which did not outperform the simpler approaches we consider in our paper.

2 Forecasting the equity premium

Denote with p the total number of available predictors, $x_{i,t}$, $i = 1, \ldots, p$, with K the dimension of the model ($K = 1$ being univariate, $K = 2$ bivariate, etc.), and with T the number of observations used to estimate a model. r_{t+1} is the excess return on the market. Next, we discuss the statistical and theoretically motivated approaches to forecasting.

2.1 Forecasting methods

It is common to assume that excess returns are linear in the set of predictors:

$$r_{t+1} = \alpha + \boldsymbol{\beta}' x_t + \varepsilon_{t+1}, \tag{1}$$

with x_t a p-dimensional set of predictors (see, e.g., Campbell and Thompson, 2008; Welch and Goyal, 2008; Rapach et al., 2009). Given a sample $\{r_{\tau+1}, x_\tau\}, \tau = 1, \ldots, T-1$, the set of parameters $(\alpha, \boldsymbol{\beta})$ is estimated by minimising the mean squared error. We refer to the model that includes all p predictors as the "kitchen sink" regression model (OLS KS).

Throughout the empirical analysis, we build upon the linear framework (1) either by adding different layers of complexity in the loss function or by taking a prior view on the composition of x_t. Next, we describe each departure from this conventional model.

Penalised regressions. The first class of models we implement is penalised linear regressions. In its general form, a penalised regression entails adding a penalty term on top of the

mean squared error $\mathcal{L}_{OLS}(\alpha, \boldsymbol{\beta}) = \frac{1}{T}\sum_{\tau=1}^{T-1} \varepsilon_{\tau+1}^2$,

$$\mathcal{L}(\alpha, \boldsymbol{\beta}; \cdot) = \underbrace{\mathcal{L}_{OLS}(\alpha, \boldsymbol{\beta})}_{\text{Loss Function}} + \underbrace{\lambda\phi(\boldsymbol{\beta}; \cdot)}_{\text{Penalty Term}} . \tag{2}$$

Depending on the functional form of the penalty term, the regression coefficients can be shrunk towards zero (the ridge regression of Hsiang, 1975), exactly set to zero (the lasso of Tibshirani, 1996), or a combination of the two (the elastic net of Zou and Hastie, 2005):

$$\phi(\boldsymbol{\beta}; \cdot) = \begin{cases} \sum_{j=1}^{p} \beta_j^2 & \text{ridge} \\ \sum_{j=1}^{p} |\beta_j| & \text{lasso} \\ (1-\delta)\sum_{i=1}^{p} \beta_i^2 + \delta \sum_{i=1}^{p} |\beta_i| & \text{elastic net} \end{cases}$$

The hyper-parameter $\lambda \geq 0$ governs the degree of shrinkage, and δ is a parameter for blending the L_1 and L_2 components in the penalty term. While the ridge is a dense model in which all predictors enter the model space, the lasso produces a sparse model which selects the variables that are deemed relevant for forecasting. A drawback of the lasso is that the L_1 penalty term will select, in a somewhat arbitrary way, one predictor from a group of highly correlated predictors (Zhao and Yu, 2006). The elastic net mitigates this tendency by adding an L_2 component in the penalty. In the empirical analysis, we set $\delta = 0.5$ following the recommendation of Hastie and Qian (2016).

Bayesian shrinkage. The use of a prior distribution centered at zero on the regression parameters $\boldsymbol{\beta}$ offers a way of regularizing the parameter estimates in a Bayesian context. Several hierarchical shrinkage priors have been proposed in the statistical literature, with

perhaps the simplest being a standard Normal prior on each individual slope coefficient $\beta_j \sim \mathcal{N}(0, \tau^2)$.[8] In this paper, we consider a popular hierarchical shrinkage prior, namely the "horseshoe" prior proposed by Carvalho et al. (2010): $\beta_i \sim \mathcal{N}(0, \sigma^2\lambda_i^2\tau^2)$, $\lambda_i \sim C^+(0,1)$, $\tau \sim C^+(0,1)$, and $p(\sigma^2) \sim 1/\sigma^2$, where $C^+(0,1)$ denotes the standard half-Cauchy distribution.

The horseshoe is a global-local shrinkage prior (Polson and Scott, 2011; Bhattacharya et al., 2016) that maintains aggressive shrinkage of unimportant coefficients without affecting the largest ones. The prior combines a local shrinkage parameter for each coefficient (λ_i) with a global shrinkage parameter (τ), thus providing more versatility in detecting sparse signals compared to other shrinkage approaches. Furthermore, this prior has the added benefit of being fully data-driven, i.e., it requires minimal tuning of hyper-parameters.[9]

Partial least squares. A benchmark data compression methodology used in empirical asset pricing is the so-called partial least squares (PLS) (see, e.g., Kelly and Pruitt, 2013, 2015). With PLS, the common components of the predictors are derived by conditioning on the joint distribution of the target variable and the regressors. We use the three-pass procedure proposed by Kelly and Pruitt (2015), a special case of PLS, using realized subsequent market returns as the only proxy.[10]

[8] Given the value of the variance in the prior distribution τ^2, the posterior mean for $\boldsymbol{\beta}$ is the solution to

$$\arg\min_{\boldsymbol{\beta}} \mathcal{L}_{OLS}(\alpha, \boldsymbol{\beta}) + \frac{\sigma^2}{\tau^2}\sum_{j=1}^{p}\beta_j^2 \tag{3}$$

which is similar to the standard ridge regression outlined above.

[9] We rely on the Matlab implementation of the method provided by Bhattacharya et al. (2016). We thank the authors for making their code available at https://github.com/antik015/Fast-Sampling-of-Gaussian-Posteriors.

[10] We thank Seth Pruitt for making his code available at https://sethpruitt.net/research/downloads/.

Forecast combination. The fourth class of competing forecasting strategies is based on the idea that, unless the correct forecasting model can be identified *ex-ante*, one can improve the prediction of excess market returns by averaging the forecasts from different models (e.g., Bates and Granger, 1969; Clemen, 1989; Timmermann, 2004; Rapach et al., 2009).

Let $\widehat{r}_{i,t+1}$ denote the forecast from the predictive model $i \in \{1, \ldots, n\}$. A combined forecast can be constructed as the weighted average:

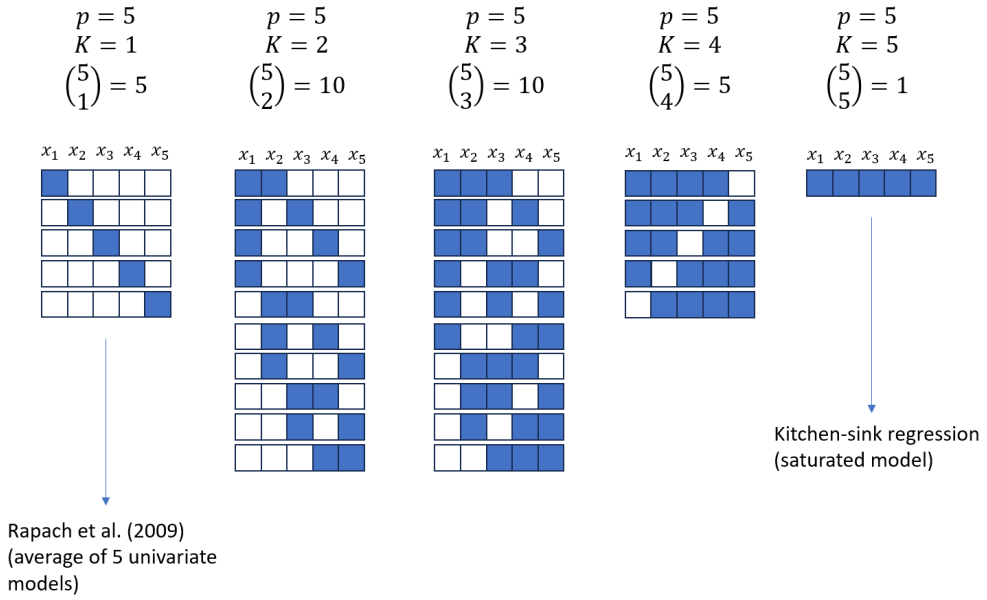
$$\widehat{r}^{\mathcal{C}}_{t+1} = \sum_{i=1}^{n} \omega_i \widehat{r}_{i,t+1}, \tag{4}$$

where $\widehat{r}^{\mathcal{C}}_{t+1}$ is the final combination forecast, $\omega_{i,t}$ is the weight associated to each individual out-of-sample prediction $\widehat{r}_{i,t+1}$ at time t. The simplest and often most effective forecasting combination method is the simple mean of forecasts whereby $\omega_i = 1/n$.

When forecasting the equity premium, Rapach et al. (2009) found that an equal weight average of simple univariate predictive regressions $\widehat{r}_{i,t+1} = \widehat{\alpha} + \widehat{\beta} x_{i,t}$ produce forecasts that outperform a naive rolling mean. This simple combination of univariate forecasts represents a special case of the complete subset regression (CSR) method proposed by Elliott et al. (2013). For a fixed number of predictors $K \leq p$, the complete subset of models is the collection of the $n_{p,K} = \binom{p}{K}$ possible regression models that include K out of the total number of potential predictors p. When $K = 1$, we have $n_{p,1} = K$ such that $\widehat{r}_{i,t+1} = \widehat{\alpha} + \widehat{\beta} x_{i,t}$ for $i = 1, \ldots, K$. Instead, for $K = p$ we have $n_{p,p} = 1$, meaning there is only one model that includes all available predictors, and the model reduces to the kitchen sink regression. Figure 1 illustrates the sets of models in CSR with $p = 5$ for each value of K.

Depending on the size of K and p, the number of total models to be considered can become computationally prohibitive. For instance, in our empirical application, we consider

Figure 1: CSR models with $p = 5$



Visual representation of all complete subset regression models with a total of $p = 5$ predictors.

a total of $p = 31$ predictors. Already with $K = 12$ one needs to evaluate $n_{12,31} = 141,120,525$ models. Elliott et al. (2013) address this issue and show that a uniform sampling of models with relatively small draws works well. In our implementation, we consider $K \in \{1, \ldots, 31\}$. For each given K, if the number of possible models is less than 5,000, we estimate all possible models. Otherwise, we estimate 5,000 models randomly selected (without replacement) from the set of all possible models. This is similar to the subsampling approach proposed by De Nard et al. (2022) in the cross-sectional asset pricing context.

The above description of the CSR method makes it clear that K is a hyperparameter to be determined. Elliott et al. (2013) propose a recursive procedure that selects, at each point in the forecasting process, the single value of K with the lowest cumulative out-of-sample MSE. In this paper, we follow an approach similar to that proposed by Elliott et al. (2013),

but we also propose an alternative procedure, which is new to the literature.

In particular, we propose to automatically select multiple values of K that produce comparable predictive performance. This is based on two insights: (1) combining CSR forecasts obtained with multiple values of K will further reduce variance and improve the aggregate forecast; and (2) choosing multiple values of K with similar predictive performance mitigates the risk of failing to select the single optimal K, which may happen due to sample variation. Our proposed approach to select an appropriate set of values for K is based on the model confidence set procedure of Hansen et al. (2011) applied to the CSR forecasts. Next, we describe this approach in detail.

Complete subset regression meets model confidence sets. To select an appropriate set of values for K, we apply the model confidence set procedure of Hansen et al. (2011) to the CSR forecasts for all values of K in each validation sample. This yields a model confidence set (MCS) that includes all the values of K for which we cannot reject the hypothesis of equal predictive ability. We then average the forecasts for all values of K in this set. We now provide a description of the MCS procedure.

Given a set $\mathcal{M}$ of models, the MCS approach is based on a sequence of pairwise significance tests, in which models found to be significantly inferior are eliminated. When no further elimination is possible, the remaining models form a model confidence set, interpreted to contain the best model(s) with some confidence. More formally, assume there are m models in $\mathcal{M}$, and let $L_i = L(r_{t+1}, \widehat{r}_{i,t+1})$ denote a loss function for a generic forecasting model i at time t, where $\widehat{r}_{i,t+1}$ is the forecast of the time-t market return obtained with model i. The relative performance between two competing forecasting models i and j is defined as $d_{ij} = L_i - L_j$. The hypotheses tested in the MCS procedure are of the form

$H_{0,\mathcal{M}} : E(d_{ij}) = 0, \quad \forall i, j \in \mathcal{M}$. The test statistic has a nonstandard distribution, and p-values are obtained using bootstrap techniques.[11]

Given a significance level α, we denote by $\widehat{\mathcal{M}}_{1-\alpha\%}$ the model confidence set. For example, the $\widehat{\mathcal{M}}_{75\%}$ is interpreted to contain the best model with 75% confidence. We apply this procedure using the MSE as our metric. In particular, we use the series of squared errors in the validation sample to run the MCS procedure and select the best models. The forecast is the average of the forecasts from the complete subset regressions for all values of K that belong to $\widehat{\mathcal{M}}_{1-\alpha\%}$ at each point in time. In total, we consider four different versions of model combination using CSR:[12]

- CSR ($K = 1$): this is the simple combination of univariate forecasts in equation (4), i.e., the combination used in Rapach et al. (2009).
- CSR (average all K): we form a combination of the forecasts from the complete subset regressions for all values of K, i.e., the aggregate forecast from this approach is essentially a combination of combinations.
- CSR (optimal K): this approach selects, at each point in the forecasting exercise, the single value of K with the lowest validation sample MSE.
- CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$): this approach averages the forecasts from the complete subset regressions for all values of K that are part of the model confidence set using the MSE criterion.

[11] Based on the recommendation in the corrigendum to Hansen et al. (2011), we use the TR statistic to test the hypothesis.

[12] Besides the statistical (MSE) criterion, we also try economic (quadratic utility) criteria to select the optimal value for K or the model confidence set. Since the analysis based on MSE delivers almost the same conclusions as that using the utility criterion, in the main text we focus on CSR (optimal K) and CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) and relegate to Appendix C.3 the results for utility-based CSR.

2.2 Theoretically motivated model restrictions

We now describe in more detail the types of models to which we refer in this paper as "theoretically motivated". In essence, these are models that can still be written in the form of a predictive regression as in (1), but that impose constraints, derived from economic theory, on the set of predictors to be included in the model, the value of their corresponding coefficients, or both. It is important to highlight that many candidate predictors of the equity risk premium proposed in the literature can be said to be, at least partly, motivated by theory. However, the models we consider in this section go further in that they specify exactly which predictors to use and, in most cases, the value of the coefficients in the predictive regression. We test the predictors in these theoretically motivated models both in isolation and in conjunction with forecast combination methods

We select four models that fit into this category and which we describe next. They each satisfy three conditions. Firstly, the candidate predictor(s) is theoretically grounded in present value relations, the no-arbitrage principle, equilibrium conditions such as the CAPM, or a combination thereof. Second, the proposed predictor(s) is observable or, alternatively, the original paper is explicit about how to proxy for it. Thirdly, the proposed predictor(s) is linked to the next month's return, rather than to a stream of future returns at an arbitrary long horizon. Moreover, if the (accounting or arbitrage) identity explicitly specifies the predictor's loading, eliminating the need for an extra regression to derive the expected market return, we enforce this additional restriction.

Drifting steady-state valuation model. Campbell (2018) shows that when the log dividend-price ratio follows a random walk, and the log dividend growth rate g_t is condition-

ally normal and homoskedastic, then

$$E_t\left[r_{t+1}\right] \approx \underbrace{\frac{D_t}{P_t}\left(1+G_{t+1}\right)+\exp\left(E_t\left[g_{t+1}\right]\right)+\frac{1}{2}\mathrm{Var}_t\left(r_{t+1}\right)-E_t\left[1+r_{f,t+1}\right]}_{x_t} \tag{5}$$

where $E_t\left[g_{t+1}\right]$ and $\mathrm{Var}_t\left(r_{t+1}\right)$ denote market participants' conditional expectation of future log dividend growth and the conditional variance of log returns and $E_t\left[1+r_{f,t+1}\right]$ is the conditional expectation of the real risk-free rate. x_t is a version of the dividend yield adjusted for dividend growth and the real interest rate.

We construct an estimate of the adjusted dividend yield using the historical sample means of real dividend growth and the real risk-free rate, and the historical sample variance of log stock returns up to date t.[13]. In principle, we could run a classic predictive univariate regression $\widehat{r}_{t+1} = \widehat{\alpha} + \widehat{\beta} x_t$. However, imposing the theoretical restriction $\widehat{r}_{t+1} = x_t$ works better in practice, so we follow this approach.

The "sum-of-parts" (SOP). Ferreira and Santa-Clara (2011) write the log stock returns as the sum of the (log) growth in the price-earnings ratio, the growth in earnings, and the dividend–price ratio:

$$\log\left(1+R_{t+1}\right) = gm_{t+1} + ge_{t+1} + dp_{t+1},$$

where R_{t+1} is the market return, $gm_{t+1} = \log\frac{P_{t+1}/E_{t+1}}{P_t/E_t}$, $ge_{t+1} = \log\frac{E_{t+1}}{E_t}$, and $dp_{t+1} = \log\left(1+\frac{D_{t+1}}{P_{t+1}}\right)$. The authors propose to forecast separately the components of the stock market return:

$$\log\widehat{\left(1+R_{t+1}\right)} = \widehat{gm}_{t+1|t} + \widehat{ge}_{t+1|t} + \widehat{dp}_{t+1|t} \tag{6}$$

[13]The term $\frac{D_t}{P_t}\left(1+G_{t+1}\right)$ is scaled by 12 since the numerator is the 12-month trailing dividend.

The SOP method assumes no multiples growth ($\widehat{gm}_{t+1|t} = 0$) and that the dividend–price ratio follows a random walk so that $\widehat{dp}_{t+1|t} = dp_t$; and it proxies for $\widehat{ge}_{t+1|t}$ with the 20-year moving average of the log growth in earnings per share up to time t, denoted by $\bar{g}_t$. In all, we use $x_t = \bar{g}_t + dp_t$ as the forecast of $\log\widehat{(1+R_{t+1})}$.[14]

Cross-sectional equity premium (CSP). Polk et al. (2006) combine the Gordon (1962) stock-valuation model with the Sharpe-Lintner CAPM to obtain a prediction for expected return:

$$\frac{D_{i,t}}{P_{i,t-1}} \approx \beta_i E_{t-1}\left[r_t\right] - E\left[g_i - r_t^f\right]$$

where $E\left[g_i - r_t^f\right]$ is expected dividend growth minus the interest rate, and betas and the risk-free rate are assumed to be constant.

This expression leads to a cross-sectional measure of the equity premium. Specifically, regress the cross-section of dividend yields on betas and expected dividend growth:

$$\frac{D_{i,t}}{P_{i,t-1}} \approx \lambda_{0,t-1} + \lambda_{1,t-1}\beta_i + \lambda_{2,t-1}E\left[g_i\right]$$

and then use $\lambda_{1,t-1}$ as a predictor in a linear regression to forecast the next period's equity premium. Kelly and Pruitt (2013) propose a related procedure to extract market expectations of future returns from cross-sectional present value relations. We follow Polk et al. (2006) and use the association between valuation rank and beta as our measure of the cross-sectional beta premium (and do not control for expected growth). As discussed in Polk et al. (2006), ranks are a transformation of the underlying multiples robust to outliers.

[14] The variable x_t is a proxy for expected log return, whereas we are forecasting excess returns (without log). One could run a predictive regression and use $\widehat{r}_{t+1} = \widehat{\alpha} + \widehat{\beta}x_t$ as the forecast, but we find that this model underperforms the historical average. We therefore use a fixed coefficient model.

Simple VIX (SVIX). Martin (2017) provides a lower bound to the expected excess returns on the market:

$$E_t\left[r_{t+1}\right] \geq R_{f,t} \times \text{SVIX}_{t+1}^2$$

where the SVIX index measures the risk-neutral variance, $\text{SVIX}_{t+1}^2 = \text{var}_t^*\left(\frac{R_{t+1}}{R_{f,t}}\right)$, and it is calculated at time t based on the prices of options that mature next month (at time $t+1$). In this paper, we use the squared SVIX index as a proxy for the equity premium, assuming the bound is tight.[15]

A natural question to ask is why we did not include, e.g., the dividend-price ratio in our list of theoretically-motivated predictors. The reason is that all of the predictors mentioned above convey information about *next period* market excess returns. On the other hand, the dividend-price ratio (without any additional assumptions) would inform not only about the next period but also about long-run returns and future dividend growth.

2.3 Out-of-sample forecasting performance

We compare the forecasts obtained from each methodology to the historical average excess stock returns. In particular, we calculate the out-of-sample predictive R^2, as suggested by Campbell and Thompson (2008). The R_{OOS}^2 is akin to the in-sample R^2 and is calculated as

$$R_{OOS}^2 = 1 - \frac{\sum_{t=0}^{T-1}\left(r_{t+1} - \widehat{r}_{t+1}\right)^2}{\sum_{t=0}^{T-1}\left(r_{t+1} - \overline{r}_{t+1}\right)^2}$$

where $\widehat{r}_{t+1}$ is the forecast value from a given approach from the start date of the estimation sample through date t and $\overline{r}_{t+1}$ is the historical sample average of the excess return estimated

[15] In future versions, we plan to replace the SVIX with the exact equity premium proxy computed by Tetlock (2023).

from the start date through t. Here T is the size of the out-of-sample forecast evaluation period. A positive value for R^2_{OOS} means the predictive regression has a lower average mean-squared prediction error than a "no-predictability" benchmark.

The R^2_{OOS} is a measure of the statistical performance of a forecasting model. However, it is possible for a return forecasting model to have a negative R^2_{OOS} and still generate significant profits when used in the context of a dynamic trading strategy (Kelly et al., 2022). Consider, for example, a model that accurately predicts the direction of the market return one period ahead but gets the scale wrong. Although such a model has a negative R^2_{OOS}, it is certainly useful to develop a profitable trading strategy. Therefore, we also assess the economic value of a given forecasting model relative to the historical average benchmark by comparing certain equivalent returns for a mean-variance investor who dynamically allocates his capital between the stock market and the risk-free asset, using forecasts from the model and the recursive sample mean (see, e.g., Welch and Goyal, 2008; Goyal et al., 2021; Dangl and Halling, 2012 and Pettenuzzo et al., 2014; Tsiakas et al., 2020; Dong et al., 2022).

Specifically, at month t, the investor decides on an equity allocation w_t by solving the utility maximization problem

$$\max_{w_t} E_t\left[U(r_{p,t+1})\right] = \widehat{r}_{p,t+1} - \frac{\gamma}{2}\widehat{\sigma}^2_{p,t+1}, \tag{7}$$

where $\widehat{r}_{p,t+1} = w_t\widehat{r}_{t+1} + (1-w_t)r_f$ and $\widehat{\sigma}^2_{p,t+1} = w_t^2\widehat{\sigma}^2_{t+1}$ are forecasts of the investor's portfolio return and variance at time $t+1$, based on information up to time t, and γ is the coefficient of relative risk aversion. The solution to problem (7) is $w_t = \frac{1}{\gamma}\frac{\widehat{r}_{t+1}}{\widehat{\sigma}^2_{t+1}}$. Given a time series of out-of-sample portfolio returns based on a specific forecasting model, we calculate the certain equivalent return as $CER_p = \bar{r}_p - \frac{\gamma}{2}\widehat{\sigma}^2_p$, where $\bar{r}_p$ and $\widehat{\sigma}^2_p$ are the mean and variance of the

portfolio return over the out-of-sample period. Similarly, we calculate the certain equivalent return of the portfolio obtained using the benchmark sample average forecasts, CER_{hist}. The spread $\Delta CER = CER_p - CER_{hist}$ can be interpreted as the fee a risk-averse investor is willing to pay to access the strategy implied by a given forecasting model.

2.4 Sample splitting and parameters tuning

Our sample period is from November 1928 to December 2021. Each forecasting model is estimated using a rolling-window approach with a training window of 240 months and a validation window of 60 months. Therefore, the out-of-sample period started in November 1953. Forecast errors in each validation window are used to select the hyper-parameters, such as the shrinkage parameter for penalised regressions, the optimal number of predictors K, and to define the model confidence set for our complete subset regression model. Upon choosing the hyper-parameters, each model is retrained using the combined training and validation samples so that forecasts for the following month – the test sample – are made using, for e.g., the optimal K or the optimal set of models. Models without any hyper-parameters – such as the linear model with all predictors included horseshoe shrinkage and the theoretically motivated regressions – are estimated directly with the combined training and validation samples.

Our use of a rolling window is needed to satisfy the assumption of the bootstrap procedure to compute the MCS (see Section 6.1 in Hansen et al., 2011). However, we also investigate the performance of all forecasting models under an expanding window approach (c.f., Section 3.3), which may shed light on whether differences in performance are potentially related to model instability, structural breaks, or changes in regime (e.g., Paye and Timmermann, 2006;

Lettau and Van Nieuwerburgh, 2008; Henkel et al., 2011; Dangl and Halling, 2012).

In the expanding window approach, we start with the same training and validation periods as in the rolling window approach, but the windows are expanded each month, with the validation period kept at 20% of the combined training and validation periods. Using this approach, we obtain forecasts for the same out-of-sample period as with the rolling window approach. However, there are two disadvantages to using an expanding window. First, as already discussed, the assumptions of the bootstrap needed for the MCS procedure are not met and the selection of statistically equivalent models is likely not reliable. Thus, we do not report MCS p-values. Second, we can only include predictors for which we have complete data over the entire sample period.

For the model confidence set, we follow Hansen et al. (2011) and choose a 75% interval ($\widehat{\mathcal{M}}_{75\%}$). When comparing models in terms of their ability to forecast the market return, we consider a quadratic loss function. Finally, to compute the CER, we use a risk aversion coefficient of 5 and (out-of-sample) estimates of the market variance using a rolling window of 60 months. Also, to keep the optimal portfolio weights w_t within a reasonable range that could be implemented in practice, we impose the restriction that $-1 \leq w_t \leq 2$ when dealing with unconstrained forecasts.

3 Empirical results

We use 31 predictors in our study, which we compile from various sources. Table A.1 in the Appendix provides a summary. We start with 11 predictors from Welch and Goyal (2008): the dividend yield (DY), the earnings/price ratio (EP), the market volatility ($RVOL$)

calculated following Mele (2007), the book/market ratio (BM), the net equity expansion ($NTIS$), the Treasury bill rate (TBL), the long-term return of government bonds (LTR), the term spread (TMS), the default yield spread (DFY), the default return spread (DFR), and inflation ($INFL$).[16]

Using the same set of data, we also calculate the predictor in equation (5) from Campbell (2018), which we denote by DP^{Drift}, as well as the two components of the SOP model, namely, the moving average of the log-growth in earnings and the logarithm of one plus the dividend yield (see (6) for more details). Using data from CRSP and Compustat, we calculate the cross-sectional risk premium (CSP) following Polk et al. (2006). We obtain data on short interest (SI) from Dave Rapach's website.[17] In addition, we calculate the $SVIX^2$ as in Martin (2017) using data from WRDS/OptionMetrics, and we follow Dong et al. (2022) and construct the average return from 172 long-short portfolios ($\overline{r}_{LS}$) obtained from Chen and Zimmermann (2022).[18]

Finally, we follow Neely et al. (2014) and construct a group of return predictors which consist of various technical analysis indicators, such as (1) moving average crossovers of different lengths $MA(n_{short}, n_{long})$ – which equal one if a moving average of n_{short} months of the S&P 500 index is above a moving average of n_{long} months, for $n_{short} \in \{1, 2, 3\}$ and $n_{long} \in \{9, 12\}$ –, (2) time-series momentum indicators $TSMOM(n)$ – which equals one if the S&P 500 index is greater than its value n months ago, for $n \in \{9, 12\}$ –, and (3) the

[16] We thank Amit Goyal for making the data available at https://sites.google.com/view/agoyal145. We do not include the dividend/price ratio because it has a near-perfect correlation with the dividend yield. Likewise, we exclude the payout ratio (DE) since is a linear combination of the dividend/price ratio and the earnings/price ratio. Finally, we also exclude the long-term yield (LTY), which is a linear combination of the Treasury bill rate (TBL) and the term spread (TMS).

[17] We thank Dave Rapach for making this available on his website at https://sites.google.com/slu.edu/daverapach.

[18] The returns on the anomaly portfolios are collected from the Open Source Asset Pricing website https://www.openassetpricing.com/ used in Chen and Zimmermann (2022).

signed-volume moving average crossover $MA^{VOL}(n_{short}, n_{long})$ – which is constructed using the signed volume of the S&P 500 index, with the same values for n_{short} and n_{long}.

3.1 Equity premium forecasts

We start by investigating the predictive ability of theoretically motivated models and several statistical models designed to accommodate a large set of predictors. Table 1 shows the results. The columns report the R^2_{OOS}, the mean-squared error (MSE) and the relative certainty equivalent return (ΔCER). The last three columns report the same metrics when a positive constraint is imposed on the conditional forecasts, i.e., in this case we use $\widehat{r}^{POS}_{t+1} = \max(0, \widehat{r}_{t+1})$, where $\widehat{r}_{t+1}$ is the forecast value from a given approach (see, e.g., (Campbell and Thompson, 2008)).

Several results stand out. First, theoretically motivated models perform well in terms of R^2_{OOS}; this is particularly the case for CSP (Polk et al., 2006) and SOP (Ferreira and Santa-Clara, 2011) ($R^2_{OOS} = 1.420\%$ and 1.465%, respectively) and, to a lesser extent, for the Campbell (2018) drifting steady-state valuation model ($R^2_{OOS} = 0.40\%$). The SOP attains the largest CER, almost double that of CSP and DP^{Drift}. Consistent with the intuition that theoretically motivated models already impose meaningful economic restrictions at the outset, we observe that adding a positive forecast constraint does not significantly affect the performance of theoretical predictors as testified by a similar R^2_{OOS} and ΔCER.

The comparable R^2_{OOS} for CSP and SOP hides some important facts. First, SOP emerges as the best theoretically motivated predictor if one adopts ΔCER as the performance metric: the SOP commands a fee of about 3.31% whereas an investor would pay 1.43% to access the CSP forecasts. Second, CSP and SOP carry some *non-redundant* information about expected

returns. In Figure 2, we plot the difference between the cumulative squared prediction errors of each of the theoretically motivated models and the historical average benchmark. In these plots, an increasing pattern indicates that the corresponding predictive model outperforms the benchmark. The graphs show that the CSP conveys substantial information at the beginning of the sample (until 1965), whereas the performance of SOP is particularly good post-2000 and before 2010.[19] Furthermore, the unconditional correlation between the CSP and SOP forecasts is only 0.44. This is *prima facie* evidence that neither of these predictors is likely to be *the* expected return on the market in isolation.

Turning to Panel B, we observe that Bayesian shrinkage and partial least squares fail to deliver a positive R^2_{OOS}. Among penalised regression models, the performance of the ridge model stands out, although the $R^2_{OOS} = 0.748$ is only half of that attained by CSP and SOP. When a positive forecast constraint is imposed, the performance of the horseshoe prior improves substantially and is now on par with that of ridge regression. We interpret this evidence as an indication in favour of "dense" models for forecasting stock returns.

One needs to turn to complete subset regressions (CSR) to obtain R^2_{OOS} larger than those observed for theoretical predictors. Importantly, an agnostic average across all possible model sizes (CSR average all K) still does not perform well, with a negative $R^2_{OOS} = -0.53$.[20] The simple combination forecast of Rapach et al. (2009), i.e., CSR with $K = 1$, obtains a higher $R^2_{OOS} = 0.56$, although it still does not ourperform CSP or SOP. Our proposed extensions of the baseline CSR (using a single optimal model size K or the average of models in the 75% model confidence set, $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) deliver the largest R^2_{OOS} - of more than 1.80% - and the largest ΔCER. A positive forecast constraint further improves our proposed CSR methods.

[19] Figure C.2 in the Appendix shows similar plots for all preditive models.

[20] Note that the CSR employs all predictors except those labelled as theoretically motivated and analyzed in Panel A.

Figure 3 provides some insights into what drives the performance of different combination methods.[21] In particular, we compare the kitchen sink regression (OLS KS), the simple combination by Rapach et al. (2009) and the two preferred CSR forecasts, namely the CSR that dynamically selects the optimal model size K and the one that averages over models in the 75% model confidence set. Compared to OLS KS, CSRs reduce the forecast variance and, as such, do well in terms of R^2_{OOS}. However, the simple combination shrinks forecasts excessively, making the forecasts very close to the historical average benchmark. Between the optimal model size K and averaging models in the MCS, the latter has a lower variance, although sometimes it produces extreme forecasts. Given that the CSR with optimal K and the CSR with average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$ emerge as the best performing models (among the statistical ones considered in the paper), we focus most of our discussion on these two models.

The evidence in Panels A and B suggests that there is fundamental information in theoretically motivated models and arguably in *other* predictors, once properly combined. Naturally, the question arises: Can one use information from both sets to enhance equity premium prediction? Panel C makes an attempt to answer this question by considering the same statistical models of Panel B, where the set of predictors is now enlarged to include the theoretical ones listed in Panel A. Quite surprisingly, the performance of CSR methods (and of all others) deteriorates substantially in terms of R^2_{OOS}.[22]

[21] In Appendix B.1, we discuss the bias-variance tradeoff as the model dimension K varies and show that the proposed methods to dynamically select optimal value(s) of K outperform CSR models with static K. Furthermore, in Appendix B.2, we discuss the time-variation in the model dimension and investigate its economic determinants.

[22] The fact that increasing the number of predictors alone does not necessarily improve the predictions is also seen in other contexts. For example, Bianchi et al. (2021) shows that a neural network with macroeconomic variables and forward rates performs worse than a more parsimonious combination of two separate networks. A possible explanation for this degradation in performance is the near multicollinearity introduced by adding correlated predictors. For example, in our case, the dividend yield is a component in DP^{Drift} and SOP.

In the next section, we show that it is possible to achieve better statistical and economic performance by combining theoretical models together with other predictors. Importantly, our approach also provides a test of whether a given theoretically motivated model is the expected return on *the* market, hence subsuming other predictors.

3.2 Testing the expected return on the market

Full-sample estimates. Table 2 provides preliminary evidence that there is information in other predictors that is not subsumed by theoretically motivated variables. As a result, theoretical predictors proposed in the literature may not be *the* expected return on the market. In particular, we build upon Harvey et al. (1998) and run the following regression:

$$r_{t+1} = \delta \widehat{r}^{Theo}_{t+1} + (1-\delta)\widehat{r}^{CSR}_{t+1} + \varepsilon_{t+1} \qquad \text{with} \qquad 0 \leq \delta \leq 1 \tag{8}$$

where $\widehat{r}^{Theo}_{t+1}$ and $\widehat{r}^{CSR}_{t+1}$ represent the out-of-sample forecast of the equity premium from a given theoretically motivated predictor and a CSR model, respectively. When $\delta = 1$ the theoretical predictor encompasses the $\widehat{r}^{CSR}_{t+1}$ forecast. On the contrary, when $\delta = 0$, it is the prediction from the CSR that encompasses the expected returns implied by $\widehat{r}^{Theo}_{t+1}$. In the case that both $\widehat{r}^{Theo}_{t+1}$ and $\widehat{r}^{CSR}_{t+1}$ contain valuable information on the equity premium, δ should be between zero and one (see, e.g., Granger and Ramanathan, 1984). Thus, we can use (8) to evaluate whether theoretical and statistical forecast specifications can be fruitfully combined to produce a superior characterization of the expected return on the market.

The theoretical models used to compute $\widehat{r}^{Theo}_{t+1}$ are the CSP, DP^{Drift} and SOP (see Panel A, B, and C, respectively). To proxy for $\widehat{r}^{CSR}_{t+1}$ we consider the CSR with $K = 1$, with optimal K, and with average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$ (from left to right). The table shows that the loading δ is around

0.4 when we pit the theoretical predictor against the CSR that accounts for optimal model size or CSR that averages across models of different sizes. This value of δ is far from the benchmark value of one. Interestingly, when using the CSR ($K = 1$) or CSR (Average all K) as a data-driven approach, the estimates $\widehat{\delta}$ are closer to one, which would be consistent with $\widehat{r}_{t+1}^{Theo}$ being *the* expected return on the market.[23] This analysis shows that it is important to benchmark theoretical models against combination methods that account for model size uncertainty, like our preferred CSR approach with average $\widehat{\mathcal{M}}_{75\%}^{\text{MSE}}$. These methods do indeed provide additional information on the expected returns, which is not captured by competing model averaging methods. More generally, combining theoretical models with other macro or trend variables shows promise since the R_{OOS}^2 for the full sample of forecasts is as high as 2.7%, and the CER is as large as 4.7%. In fact, these R_{OOS}^2 are even larger than those resulting from applying a positive forecast constraint to CSR (optimal K) and CSR (average $\widehat{\mathcal{M}}_{75\%}^{\text{MSE}}$) (2.23 and 2.15, respectively, as shown in Table 1). This highlights the effectiveness of integrating theoretical predictors with statistical model averaging forecasts to capture the impact of sensible economic constraints on the statistical model.

Real-Time Expected Returns Estimates. Table 2 reported the δ estimated from Eq. (8) using the full set of out-of-sample forecasts $\widehat{r}_{t+1}^{Theo}$ and $\widehat{r}_{t+1}^{CSR}$. We now instead estimate δ from the perspective of a real-time investor. The purpose of this exercise is twofold. First, we want to understand whether the $\delta = 0.4$ estimated in Table 2 masks any meaningful time variation. For example, it could be the case that, in some periods, $\widehat{r}_{t+1}^{Theo}$ could indeed

[23] The optimal δs are smaller for CSR (average all K) than for CSR ($K = 1$). This may be surprising at first, since CSR (average all K) attains a negative R_{OOS}^2 in Table 1. However, observe that CSR (average all K) delivers a positive $R_{OOS}^2 = 1.616$ once we impose the positive constraint. Thus, when we combine CSR (average all K) with theoretical forecasts, we boost the performance in two ways: First, because the theoretical forecasts contain quite different information; and second, because the combined forecasts likely turn positive.

be considered the expected return on the market by a real-time investor. Second, we aim to investigate whether the good performance observed in Table 2 for a model that combines theoretical models with the (combined) forecasts from other variables also holds when implemented in real-time.

Figure 4 shows the real-time δ estimated from Eq. (8) at each month via constrained least squares using either a rolling window of five-years of out-of-sample forecasts (blue line) or an expanding window starting with 60 months of out-of-sample forecasts (red line). Shaded areas indicate NBER recessions. From left to right, the different columns report the estimated loading $\hat{\delta}$ when using CSP, DP^{Drift} or SOP to compute the forecast $\widehat{r}_{t+1}^{Theo}$. Two interesting facts emerge. First, there is instability in the δ when it is estimated on a rolling window (blue line). The null hypothesis that $\widehat{r}_{t+1}^{Theo}$ encompasses the expected market return implied by CSR can be arguably rejected more strongly outside recessions. This is more evident for $\widehat{r}_{t+1}^{CSR}$ proxied by CSR ($K = 1$), which is a simple combination of univariate forecasts. Second, the weights resulting from combining the theoretically motivate model with CSR (optimal K) are similar to those obtained from employing CSR (average $\widehat{\mathcal{M}}_{75\%}^{\text{MSE}}$). The rank correlation between the δ_t obtained with either specification is 0.86 (average across the CSP, DP^{Drift}, and SOP models). This is consistent with the comparable performance in terms of R^2_{OOS} of theoretically-motivated predictive models, in particular CSP and SOP, outlined in Table 1 and Table 2.

Implications for out-of-sample forecasting. Table 3 reports the results from the combined forecasts $\widehat{r}_{t+1} = \delta_t \widehat{r}_{t+1}^{Theo} + (1 - \delta_t)\widehat{r}_{t+1}^{CSR}$ using the real-time estimates of Figure 4; i.e., δ_t is estimated from $t - 59$ to t to avoid look-ahead bias or double usage of the target equity

premium r_{t+1}.[24]

This approach is similar to the iterated mean combination (IMC) proposed by Lin et al. (2018) with two key differences. First, we combine the forecast combination method with a given theoretically motivated model rather than with the running mean of market returns. We do so because our objective is not to provide yet another shrinkage strategy towards the unconditional average of returns but rather to test the encompassing property of theoretically justified predictors, i.e., their ability to capture the expected return on the market out-of-sample. Second, we restrict $\delta_t \in [0, 1]$ so that it can be interpreted as the relative weight assigned to, a priori, complementary and potentially not mutually exclusive proxies for the expected return on the market.

Panel A of Table 3 reports results for the three best CSR models ($K = 1$, optimal K, and average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$), which are similar to those in Table 1 over this shorter period. The first row of Panels B, C and D ("Baseline") reports the performance of the theoretically motivated models. When compared to Panel A of Table 1 we observe degradation in R^2_{OOS} for CSP (from 1.42% to 0.67%). On the other hand, the performance of DP^{Drift} and SOP appears to be more stable.

Our main interest lies in the performance of our IMC implementation. The results consistently show that the IMC forecast that combines the baseline with the CSR (optimal K) or the CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) outperforms (both in terms of statistical and economic metrics) the theoretically motivated models, but fails to improve upon the respective CSR models in Panel A that exclude them. One potential reason for such dismal performance is the large instability in the rolling window estimates of δ_t (c.f., the blue line in Figure 4).

[24] Because of the additional estimation step for δ_t, we lose five years of data relative to Table 1.

To address the instability of δ_t, we follow a naive model averaging (NMA) approach, which fixes $\delta = 0.5$.[25] This value is close to the full sample estimates reported in Table 2. The results in Table 3 show that fixing the δ_t to 0.5 improves the performance across models, with R^2_{OOS} above 2.2% for any of the theoretically motivated models combined with the CSR (optimal K) or the CSR that averages across the model confidence set. The best model is the naive method that combines SOP and CSR with optimal model size K; this model yields the highest $R^2_{OOS} = 2.46\%$ and has a $\Delta CER = 4.66\%$, which is second only to that of the CSR with optimal K in Panel A ($\Delta CER = 4.70\%$). Imposing the positivity constraint on the resulting combined forecasts improves R^2_{OOS} and ΔCER even further.[26] Our interpretation of these results is that, regardless of the approach taken, economic theory is an important consideration when forming forecasts of the equity risk premium. Our dynamic CSR forecasts perform well on their own but can be improved by imposing the positivity constraint (which is rooted in economic theory, as the equity premium should be positive) or through a simple combination with forecasts from the theoretically motivated models of the equity risk premium. In the case of SOP, both seem to help.

As highlighted in Figure 4, a feasible alternative to stabilize the estimates of δ_t is to adopt an expanding rather than a rolling window approach: the expanding window estimates (red line) are indeed less volatile.[27] Table 4 reports the performance when using such δ_t estimated over an expanding window.[28] The results confirm that more stable values of δ lead to larger

[25] This approach is similar in spirit to the one proposed by Chen et al. (2022). Whereas Chen et al. (2022) shrink forecasts from univariate predictive regressions towards the sample average of market excess returns, we focus on a combination of theoretically motivated models and statistical models with many predictors.

[26] We impose the positivity constraint on the combined forecast, i.e., the constrained forecasts are calculated as $\max(\widehat{r}_{t+1}, 0) = \max(\delta_t \widehat{r}^{Theo}_{t+1} + (1-\delta_t)\widehat{r}^{CSR}_{t+1}, 0)$.

[27] Note that $\widehat{r}^{Theo}_{t+1}$ and $\widehat{r}^{CSR}_{t+1}$ are still obtained from a rolling sample. We do so to isolate the effect of instability in δ and allow a comparison between Tables 3 and 4.

[28] The results in Panel A and those for the naive model averaging (NMA) that fixes $\delta = 0.5$ are, of course, identical between Tables 3 and 4, and reported only for reader's convenience.

R^2_{OOS} and ΔCER compared to when δ_t is estimated with a rolling window. Nevertheless, the NMA approach that fixes δ outperforms an expanding window combination of theoretical and CSR forecasts. This holds for all the theoretically motivated models and the three versions of CSR.

A natural question is whether fixing $\delta_t = 0.5$ represents an optimal compromise, or one could further improve the out-of-sample forecasting performance by using alternative fixed values for δ. We formally investigate this question by showing the out-of-sample R^2_{OOS} across different values of $\delta = 0, 0.1, \ldots, 1$, and for all theoretical predictors and CSR methods. Figure 5 shows the results. The leftmost point on the graph is the R^2_{OOS} that would be obtained by the given CSR method. The rightmost point is the R^2_{OOS} obtained by the the theoretically motivated model in isolation.[29] The graph shows that, for our preferred versions of CSR, $\delta = 0.5$ is nearly optimal for CSP and SOP, whereas a lower loading is needed on DP^{Drift}. Indeed, the naive investor setting $\delta = 0.5$ exhibits an R^2_{OOS} (from Table 3) lower by 0.38%, 0.12%, and 0.08% compared to the maximum reported in Figure 5 for the CSP, DP^{Drift}, and SOP models, respectively.[30]

Overall, we can conclude that none of the theoretically motivated models analyzed seems to be *the* expected return on the market. Nevertheless, theoretically motivated models carry meaningful information on the equity premium so that, when combined together with other variables, the out-of-sample predictability of stock market returns substantially increases.

[29] Since δ_t is fixed, there is no need for an initial 5-year burn-in period to estimate it. Therefore, Figure 5 uses the entire OOS period from Nov 1953 to Dec 2021, and the left and rightmost points math the values reported in Table 1.

[30] Specifically, for CSR (optimal K) and CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$), the R^2_{OOS} peaks at $\delta = 0.449, 0.306, 0.449$ in the context of CSP, DP^{Drift}, and SOP. These values are very close to those reported in Table 2. The difference arises due to the discretization of the deltas to generate the figure.

Properties of Real-Time Expected Returns Estimates. Figure 6 depicts the time series of expected returns derived from the CSR (optimal K) in the top left panel, as well as the time series resulting from employing a specific theoretically motivated model in isolation or in combination with the CSR (optimal K) (represented by the blue and dashed red lines, respectively, in top right and bottom panels). Initially, we note that the volatility of expected return forecasts for individual theoretically motivated models is very low, ranging from only 0.32% for DP^{drift} to 1.44% for CSP (annualized), whereas the volatility of realized returns is close to 15%. The highest volatility is for the CSR (optimal K), at 2.78%. The NMA approach delivers expected returns that have a volatility ranging from 1.39% (for DP^{drift}) to 1.66% (for CSP). The average annualized forecasts of the market risk premium for the NMA that combines CSR (optimal K) with CSP, DP^{drift}, and SOP, respectively, are 6.30%, 5.35%, and 5.74%, compared to the average excess return of 7.07%. Finally, we note that, although the correlations between the forecasts of the CSP, DP^{drift}, and SOP models are low (between -0.31 and 0.25), the correlations between the NMA forecasts are high, between 0.88 and 0.95, due to the common CSR component which has much higher volatility.

3.3 Individual forecasts based on an expanding window

All the results discussed so far are obtained using a rolling window approach to produce the forecasts for both $\widehat{r}_{t+1}^{Theo}$ and $\widehat{r}_{t+1}^{CSR}$. This allows us to incorporate additional predictors for which data is only available more recently. Nevertheless, an expanding window approach has often been used as an alternative to rolling-window implementations. For example, Rapach et al. (2009), Elliott et al. (2013), and Neely et al. (2014) rely on an expanding window approach whereas Li and Tsiakas (2017) and Tsiakas et al. (2020) use a rolling window

approach.[31] To compare our results with previous studies, Tables C.6 and C.7 reproduce the evidence in Tables 1 and 3 using the expanding window approach.

Table C.7 reports the results using all predictors for which data is available throughout the sample period. For consistency, the historical average used in calculating the R^2_{OOS} in this case is also based on an expanding window. The first notable difference with respect to our previous results is a widespread decrease in R^2_{OOS} and ΔCER for statistical methods (see Panel A). For example, without the positive constraint on the prediction, the CSR with optimal K displays an $R^2_{OOS} = 0.41$ relative to the 1.85 reported in Table 3. Furthermore, the CSR (optimal K) underperforms the CSR ($K = 1$). However, once we impose the positive forecast constraint, the CSR (optimal K) beats the simple combination in terms of R^2_{OOS} of 0.69% while producing the highest ΔCER (which equals 2.39%). The CSR that relies on the MCS shows a negative R^2_{OOS}, which is not surprising, given that underlying assumptions for the combination procedure are unlikely to hold.[32]

Another interesting result is the deterioration in the performance of the theoretically motivated models based on expanding window regressions relative to the results obtained with a rolling window. The CSP (Polk et al., 2006) and DP^{Drift} (Campbell, 2018) now produce negative R^2_{OOS} (see row "Baseline" in Panels B and C). These results suggest a substantial model instability, which can be more accurately attenuated with a rolling window forecasting approach. More importantly, we continue to find that the naive combination – with $\delta = 0.5$ – of a given theoretically motivated predictor with the CSR (optimal K)

[31] See Cooper and Gulen (2006) for a discussion of the merits of the two approaches.

[32] The implementation of the MCS procedure relies on bootstrapping the differences in loss between forecasts from each pair of models, which are assumed to be stationary. This assumption is unlikely to hold when the parameters are estimated based on an expanding window (see Hansen et al., 2011, Section 6.1). Because of this, we are careful in interpreting the results obtained with the complete subset regression methods that rely on the MCS procedure to select different values of K.

delivers the best performance. For instance, for CSP, we observe an $R^2_{OOS} = 0.74$ and a ΔCER= 2.21%. Similarly, we see an $R^2_{OOS} = 0.82$ and a $\Delta CER = 2.53\%$ for SOP. Overall, we confirm that theoretical predictors do not produce encompassing expected returns with respect to CSR and that combining theoretical and statistical forecasts through a simple weighted average enhances out-of-sample predictive accuracy compared to using each forecast in isolation.

3.4 Evidence over the post-1996 sample

The purpose of this section is twofold. First, we want to investigate the robustness of our conclusions over an alternative sample period. At the same time, the SVIX proposed by Martin (2017) becomes available over this shorter sample. Thus, we can test whether this important variable – which has been widely used in the literature to measure the expected market return – subsumes information by other predictors.[33] Panel A in Table 5 shows the performance of each theoretically motivated predictor over the sample from 1996 to 2021. Compared to Table 1, the R^2_{OOS} halves from 1.42% to 0.70% for CSP, it decreases from 1.46% to 0.42% for SOP, and it turns negative for DP^{Drift}. The decline in statistical performance is also visible in Figure 2, where the cumulative mean squared error flattens post-90s for SOP and DP^{Drift}, while registering a notable reduction around the 2000 recession for SOP, and a severe decline in 2000 and post-2010 for DP^{Drift}. Despite the large decrease in statistical performance, the SOP yields the best economic gain with a CER=2.33%. Over this sample period, the SVIX achieves the largest R^2_{OOS} among theoretical predictors and the second largest economic gain (ΔCER= 1.61%).

[33] Tables 5, 6, and 7 in this subsection are the counterparts of Tables 1, 2 and 3, respectively.

Panel B shows that the worsening in performance over this sample period is not unique to theoretically motivated models, but it also afflicts statistical models. In particular, we see that the R^2_{OOS} is in negative territory also for the CSRs. Imposing the positive forecast constraint restores a positive R^2_{OOS} only for our proposed methods, namely the CSRs with optimal K and the CSR averaging over models in the $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$ set. Panel C shows that adding predictors from the theoretically motivated models to the set of variables used in Panel B does not help the statistical models. On the contrary, we now observe that both the CSR (optimal K) and the CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) display negative R^2_{OOS}, even after imposing the positive forecast constraint.

Overall, the results in Table 5 raise the possibility that, over this sample, the theoretical predictors are getting closer to be the expected return on the market. This conjecture is supported to some degree by the results in Table 6. The weight on CSP increases from about 0.44 to 0.77, whereas that on SOP goes from 0.44 to 0.66 (CSR optimal K). The SVIX displays a similar weight of about 0.66. As we have already observed in Table 2, it is important to use a forecast combination model that accounts for either optimal model size or optimal set of models; otherwise, one would estimate δs at least as large as 0.85 for CSP and SOP, and 0.77 for SVIX. Importantly, the full-sample estimate of δ yields sizable R^2s: focusing on the forecast from CSR (optimal K), the R^2_{OOS} are 0.81, 0.34, 0.70 and 1.39% for CSP, DP^{Drift}, SOP and SVIX, respectively. These values are all greater than those obtained by each theoretically motivated predictor taken in isolation (0.70, -0.69, 0.42, and 0.92%; c.f. Table 5, Panel A). This evidence prompts an investigation of the out-of-sample performance of the IMC and NMA.

Table 7 shows that CSP yields an $R^2_{OOS} = 1.06\%$ and virtually zero economic gain. However, when CSP is naively combined (i.e., $\delta = 0.5$) with CSR that averages over the

$\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$ set, these figures become 1.11% and 3.49%. The gain for SOP is also large: when combined with CSR (average over $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$), the R^2_{OOS} raises from 0.53% to 0.82% and the CER increases from 2.5% to 3.8%. The SVIX is no exception; alone it yields an $R^2_{OOS} = 0.61\%$ and a $\Delta CER = 1.04\%$.[34] These values increase substantially for the NMA approach that uses the CSR averaging over $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$: the R^2_{OOS} is now = 1.51% and the CER= 3.67%. It is also interesting to compare the model that estimates δs (IMC) rather than restricting it to 0.5 as in NMA. Panels B to E of Table 7 reveal that, in general, IMC performs poorly relative to NMA. This is attributed to the volatility in estimated δ, particularly evident when using a rolling window (refer to Figure 7). Indeed, for the combination of CSP with CSR (optimal K), we observe a more stable δ and, hence, similar performance between IMC and NMA ($R^2_{OOS} = 0.93\%$ vs $R^2_{OOS} = 1.04\%$).

Given the excellent performance obtained by fixing $\delta = 0.5$, one may still wonder if this is the optimal weight to combine theoretically motivated models and statistical models. Figure 8 answers this question by recomputing the out-of-sample R^2_{OOS} for different values of δ. For CSP, SOP and SVIX, we find a value larger than 0.5. In particular, the maximum R^2_{OOS} are achieved for δ equal to 0.75, 0.65, and 0.68 when these predictors are combined with the CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$). These values align closely with the full sample estimates reported in Table 6, confirming that, over this sample, the forecasts from theoretically motivated models are getting closer to the expected return on the market than those of statistical models. At the same time, the figure quantifies the "gain" in terms of R^2_{OOS} for an investor that knows

[34] The performance of theoretically motivated predictors is not stable. We already observed this point when comparing Panel A in Table 1 to Table 3. Similarly, when comparing Panel A in Table 5 (sample: 1996 to 2021) to Table 7 (sample: 2001 to 2021) we see that the R^2_{OOS} increases from 0.70 to 1.01 for CSP but decreases from 0.92 to 0.61 for the SVIX. By imposing a tight structure on the role of theoretical predictors, the evidence suggests that the NMA approach could mitigate the instability of theoretical predictors over time.

the true optimal delta, relative to a naive investor that simply adopts $\delta = 0.5$. In particular, knowledge of the optimal delta would improve the R^2_{OOS} by 0.15%, 0.08% and 0.11% for CSP, SOP and SVIX, respectively.

Overall, combining theoretically motivated models with forecast combination methods, particularly our preferred variants of CSR, proves effective in enhancing the statistical performance of each individual approach in isolation. A question may arise regarding whether the improvement brought by theoretically motivated models is confined to specific data points, such as recessions. To address this, Figure C.3 illustrates the cumulative squared error of the CSR model (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) relative to that of the theoretically motivated model. An increase in the plot indicates instances where the theoretically motivated model outperforms (i.e., delivers lower mean squared error than) the CSR model. Several observations emerge. Firstly, there is an upward trend post-90s, implying a consistent contribution of the theoretically motivated model to overall performance, not limited to specific data points. At the same time, fluctuations around this trend highlight instances where our CSR approach outperforms the identity-based predictors. Second, the 2008 crisis is a distinctive period where our CSR significantly outperforms theoretically motivated models such as CSP, DP^{Drift}, and SOP. This is supported by the negative spike in cumulative squared error shown in Figure C.3, underscoring the relevance of statistical models even during crises. Additionally, this evidence aligns with our analysis of the weight assigned to the theoretically motivated model (δ) relative to CSR. Consistent with the general upward trend post-90s observed in Figure C.3, we estimate larger δs in Table 6 compared to Table 2 . Furthermore, in the last row of Figure 7, we observe that the δs from the combination approach with an expanding window decrease around the 2008 crisis, supporting the fact that CSR performs better than the theoretically motivated model during this period.

In sum, our main conclusions continue to hold over this more recent - and shorter - sample period. First, the simple, naive combination of theoretically motivated model with macroeconomic and trend-based variables improves the performance of individual theoretically motivated model over purely statistical models. Second, although over this sample the theoretically motivated models perform better than the statistical models, and seem to be closer to the expected return on the market, there is substantial information contained in other variables that provide statistical and economic gains to a real-time investor.

3.5 Optimal Portfolios

The certain equivalents reported thus far consolidate into a single number the economic value of using forecasts to construct optimal market timing portfolios. From a practical point of view, the characteristics of these portfolios are relevant to investors, due to factors such as transaction costs, use of leverage etc. Therefore, in Tables C.2 through C.5 we report statistics of the portfolios associated with different forecasts.[35] These tables report annualized average returns, standard deviations, and corresponding Sharpe ratios (SR), as well as the average allocation to the market portfolio ($\overline{w}$), its standard deviation ($\sigma(w)$), and the average monthly portfolio turnover.

Tables C.2 reports results using raw forecasts, i.e., without imposing the positivity constraint, over the longer sample period from Nov 1958 to Dec 2021. For comparison purposes, Panel A reports results for a static buy and hold strategy that is always 100% invested in the stock market. This strategy produces an annual excess return of approximately 7%, with a

[35] All optimal portfolios are obtained with $\gamma = 5$. Since forecasts from different models have different variances, it is not surprising to find that the optimal portfolios vary significantly in terms of their annualized volatility.

volatility of close to 15%, yielding a Sharpe ratio of 0.48. Using the historical average return as a forecast, an investor would have outperformed the market by 164 basis points, while slightly reducing the volatility, increasing the Sharpe ratio to 0.61. This would entail a low turnover of 3.35% per month, which is not surprising, as the historical average is quite stable over time. In Panel B, we report results for standalone CSR models. The best performer in terms of Sharpe ratio is CSR (Optimal K), which delivers an annualized return of 14.16%, with volatility of 15.08%, for a Sharpe ratio of 0.93. To achieve this, however, the investor would have to turnover on average 41% of their portfolio every month.

Panels C, D, and E of Table C.2 report results for theoretically motivated models and different combinations with CSR models. Among the baseline models, the best results are produced by using forecasts from the DP^{drift} model (SR=1.02) and the SOP model (SR = 1.00). The DP^{drift}, in particular, achieves the highest SR while having a very low average turnover of only 1.44%. One interesting aspect is that, since forecasts from the theoretically motivated models have lower volatility, the corresponding optimal portfolios have lower allocations to the market than the portfolios constructed using CSR forecasts. For example, the average allocation to the market for the CSR (Optimal K) portfolio is approximately 70%, whereas the average allocation is 40% for DP^{drift} and 51% for SOP.

Turning to the optimal portfolio constructed using combinations of identity-based and CSR models, the first result that stands out is that portfolios obtained with the simple average (NMA) always outperform those that use optimization to select the value of δ. Among the portfolios obtained using NMA forecasts, those that use CSR (Optimal K) forecasts appear to deliver better results. These portfolios all achieve Sharpe ratios that are close to the highest Sharpe ratio of either model. For instance, the portfolio constructed using the average of DP^{drift} (respectively, SOP) and CSR (Optimal K) forecasts achieves a SR of 0.98

(respectively, 1.00). As expected, the average turnover of these NMA portfolios is higher than that of the portfolios constructed using DP^{drift} and SOP forecasts directly, but lower compared to portfolios constructed with CSR (Optimal K). Imposing the positivity constraint does not change the overall conclusions and the corresponding numbers are reported in Table C.3.

Table C.4 reports results for the more recent period from January 2001 to December 2021, for which we also include results using the $SVIX^2$ model. The buy and hold strategy produces a SR of 0.57 over this period, which is sightly higher compared to the longer sample period. However, for this period the portfolio constructed using the historical average return underperforms the market, producing a SR of 0.39. Among the CSR models, the best performance is now obtained using the CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) forecasts, which achieves a SR of 0.72. In contrast, the portfolios constructed with the forecasts from identity-based models all underperform the buy and hold strategy. The portfolio constructed using the CSP forecast, the best among the baseline models in Panels D through F, achieves a SR of 0.55. The NMA portfolios achieve slightly lower SRs than those of the portfolios constructed using the corresponding CSR forecasts, but with lower turnover. The results with the positivity constraint are, once again, similar, and are reported in Table C.5.

We can draw several conclusions from these analyses. First, it is possible to outperform the market using equity risk premium forecasts from both statistical as well as theoretically motivated models. Second, no single model appears to produce the best results consistently, and the relationship with the statistical performance of the models is not straightforward. For example, over the sample period from Nov 1958 to December 2021, CSR produces somewhat higher R^2_{OOS} compared to the identity-based models (see Table 4), but this does not seem to translate to an advantage in terms of risk-adjusted returns. The opposite seems to happen

in the more recent period in Table C.4. Finally, we note that market-timing portfolios, particularly constructed using CSR forecasts, have relatively high turnover, which suggests that an analysis of transaction costs is warranted. Although we do not propose a detailed analysis of transaction costs in this paper, we note that the results are not likely to disappear under reasonable transaction costs. For example, in Table C.2, the average turnover of the portfolio constructed using a NMA combination of SOP and CSR (optimal K) was 26.794%. Assuming a cost of 50bps per transaction (see., e.g. Marquering and Verbeek, 2004; Tsiakas et al., 2020), the total annual cost of this optimal portfolio would be approximately 161bp. The SR would be reduced from 1.00 to 0.87, still comparing favorably to the buy and hold strategy (SR = 0.48). If an investor implements such strategies using futures contracts, which have much lower transaction costs, the effect will be even less relevant.

4 Conclusions

In this paper, we investigate the extent to which theoretically motivated models can capture *the* expected return on the market. We start by documenting that theoretically motivated predictors perform well in terms of economic and statistical performance. We then turn to a variety of statistical approaches that can handle a large number of predictors and show that combination methods can compete with the theoretical predictive models. However, not all combination methods work equally well. A simple combination of univariate models or an agnostic combination of all possible model sizes is suboptimal. In other words, the model size matters. We propose to either select the dimension delivering the lowest (in an out-of-sample sense) MSE, or to combine models of different dimensions that are deemed statistically indistinguishable in terms of MSE. We show that these variants of complete

subset regressions yield out-of-sample R^2_{OOS} of similar magnitude to theoretically motivated models and larger economic gains as measured by the certainty equivalent.

Given this evidence, we turn to the key questions of this paper: (1) whether theoretically motivated models subsume information in a large set of candidate predictors of the equity premium, and (2), if not, whether there is a benefit in combining forecasts from the two approaches. Since complete subset regressions (CSRs), in particular those that account for optimal model size, emerge as the best statistical method among those we considered, we propose to combine a given theoretically motivated model with the forecasts obtained from a combination method, i.e., we use an iterated combination method. We also restrict the sum of weights on the theoretical and CSR forecasts to be one, so that the expected return on the market is given by the theoretically motivated model under consideration at one extreme, and by the CSR forecasts at the other. We show that the weight on theoretically motivated models is generally less than one, suggesting that it is unlikely that a theoretically motivated model in isolation is *the* expected return on the market. However, the optimal weight is extremely unstable. Indeed, a naive approach that fixes that weight to a given value yields superior performance relative to other iterated combination approaches. Quantitatively, the fixed weight with the best performance is about 0.5 over the full sample from 1953 to 2021, but rises to 0.7 when we consider the more recent sample from 1996 to 2021.

We conclude that theoretically motivated models are informative but far from being the expected return on the market. At the same time, averaging all variables disregarding their possibly different economic underpinning is unlikely to be the solution. Therefore, our recommendation to financial economists is to decouple the parsimonious economic structure of theoretical models from the agnostic view offered by data-driven methods such as complete subset regressions, and then to combine the two separate forecasts to model expected returns.

This supports a rather intuitive, although often unappreciated view: complex statistical methods could be used to *boost* theoretically motivated predictors rather than dilute the importance of economic theory for equity premium prediction. Yet, each predictor in isolation is unlikely to be *the* expected return on the market.

References

Avramov, Doron (2002) "Stock return predictability and model uncertainty," *Journal of Financial Economics*, 64 (3), 423–458.

Baker, Malcolm and Jeffrey Wurgler (2006) "Investor sentiment and the cross-section of stock returns," *The journal of Finance*, 61 (4), 1645–1680.

Bates, J. M. and C. W. J. Granger (1969) "The combination of forecasts," *Operational Research Quarterly*, 20, 451–468, 10.2307/3008764.

Bhattacharya, Anirban, Antik Chakraborty, and Bani K Mallick (2016) "Fast sampling with Gaussian scale mixture priors in high-dimensional regression," *Biometrika*, asw042.

Bianchi, Daniele, Matthias Büchner, and Andrea Tamoni (2021) "Bond risk premiums with machine learning," *The Review of Financial Studies*, 34 (2), 1046–1089.

Campbell, John Y. (2018) *Financial Decisions and Markets: A Course in Asset Pricing*: Princeton University Press.

Campbell, John Y. and Samuel B. Thompson (2008) "Predicting Excess Stock Returns out of Sample: Can Anything Beat the Historical Average?" *The Review of Financial Studies*, 21 (4), 1509–1531.

Carvalho, Carlos M, Nicholas G Polson, and James G Scott (2010) "The horseshoe estimator for sparse signals," *Biometrika*, 97 (2), 465–480.

Cederburg, Scott, Travis L Johnson, and Michael S O'Doherty (2023) "On the economic significance of stock return predictability," *Review of Finance*, 27 (2), 619–657.

Chen, Andrew Y. and Tom Zimmermann (2022) "Open Source Cross-Sectional Asset Pricing," *Critical Finance Review*, 27 (2), 207–264.

Chen, Huafeng (Jason), Liang Jiang, and Weiwei Liu (2022) "Predicting Returns Out of Sample: A Naïve Model Averaging Approach," *The Review of Asset Pricing Studies*, 13 (3), 579–614.

Chen, Huafeng, Liang Jiang, and Weiwei Liu (2022) "Predicting Returns Out of Sample: A Naïve Model Averaging Approach," *The Review of Asset Pricing Studies*, raac021.

Clemen, Robert T (1989) "Combining forecasts: A review and annotated bibliography," *International Journal of Forecasting*, 5 (4), 559–583.

Cooper, Michael and Huseyin Gulen (2006) "Is time-series-based predictability evident in real time?" *The Journal of Business*, 79 (3), 1263–1292.

Cremers, KJ Martijn (2002) "Stock return predictability: A Bayesian model selection perspective," *The Review of Financial Studies*, 15 (4), 1223–1249.

Dangl, Thomas and Michael Halling (2012) "Predictive regressions with time-varying coefficients," *Journal of Financial Economics*, 106 (1), 157–181.

De Nard, Gianluca, Simon Hediger, and Markus Leippold (2022) "Subsampled factor models for asset pricing: The rise of Vasa," *Journal of Forecasting*, 41 (6), 1217–1247.

Dong, Xi, Yan Li, David E Rapach, and Guofu Zhou (2022) "Anomalies and the expected market return," *The Journal of Finance*, 77 (1), 639–681.

Elliott, Graham, Antonio Gargano, and Allan Timmermann (2013) "Complete subset regressions," *Journal of Econometrics*, 177 (2), 357–373, Dynamic Econometric Modeling and Forecasting.

Ferreira, Miguel A and Pedro Santa-Clara (2011) "Forecasting stock market returns: The sum of the parts is more than the whole," *Journal of Financial Economics*, 100 (3), 514–537.

Giannone, Domenico, Michele Lenza, and Giorgio E. Primiceri (2021) "Economic Predictions With Big Data: The Illusion of Sparsity," *Econometrica*, 89 (5), 2409–2437.

Gordon, Myron J (1962) "The savings investment and valuation of a corporation," *The Review of Economics and Statistics*, 37–51.

Goyal, Amit, Ivo Welch, and Athanasse Zafirov (2021) "A Comprehensive Look at the Empirical Performance ofEquity Premium Prediction II," *Available at SSRN 3929119*.

Granger, Clive WJ and Ramu Ramanathan (1984) "Improved methods of combining forecasts," *Journal of forecasting*, 3 (2), 197–204.

Green, Jeremiah, John R. M. Hand, and X. Frank Zhang (2017) "The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns," *The Review of Financial Studies*, 30 (12), 4389–4436.

Guidolin, Massimo and Allan Timmermann (2007) "Asset allocation under multivariate regime switching," *Journal of Economic Dynamics and Control*, 31 (11), 3503–3544.

Hansen, Peter R, Asger Lunde, and James M Nason (2011) "The model confidence set," *Econometrica*, 79 (2), 453–497.

Harvey, David I, Stephen J Leybourne, and Paul Newbold (1998) "Tests for forecast encompassing," *Journal of Business & Economic Statistics*, 16 (2), 254–259.

Hastie, Trevor and Junyang Qian (2016) "Glmnet vignette," *Working paper, Stanford University*, –.

Henkel, Sam James, J Spencer Martin, and Federico Nardari (2011) "Time-varying short-horizon predictability," *Journal of financial economics*, 99 (3), 560–580.

Hou, Kewei, Chen Xue, and Lu Zhang (2020) "Replicating anomalies," *The Review of Financial Studies*, 33 (5), 2019–2133.

Hsiang, TC (1975) "A Bayesian view on ridge regression," *Journal of the Royal Statistical Society: Series D*, 24 (4), 267–268.

Huang, Dashan, Fuwei Jiang, Jun Tu, and Guofu Zhou (2015) "Investor sentiment aligned: A powerful predictor of stock returns," *The Review of Financial Studies*, 28 (3), 791–837.

Kelly, Bryan and Seth Pruitt (2013) "Market Expectations in the Cross-Section of Present Values," *The Journal of Finance*, 68 (5), 1721–1756.

——— (2015) "The three-pass regression filter: A new approach to forecasting using many predictors," *Journal of Econometrics*, 186 (2), 294–316, High Dimensional Problems in Econometrics.

Kelly, Bryan T, Semyon Malamud, and Kangying Zhou (2022) "The Virtue of Complexity in Return Prediction."

Kelly, Bryan T and Dacheng Xiu (2023) "Financial machine learning," *Available at SSRN*.

Lettau, Martin and Stijn Van Nieuwerburgh (2008) "Reconciling the return predictability evidence," *The Review of Financial Studies*, 21 (4), 1607–1652.

Li, Jiahan and Ilias Tsiakas (2017) "Equity premium prediction: The role of economic and statistical constraints," *Journal of financial markets*, 36, 56–75.

Lin, Hai, Chunchi Wu, and Guofu Zhou (2018) "Forecasting corporate bond returns with a large set of predictors: An iterated combination approach," *Management Science*, 64 (9), 4218–4238.

Marquering, Wessel and Marno Verbeek (2004) "The economic value of predicting stock index returns and volatility," *Journal of Financial and Quantitative Analysis*, 39 (2), 407–429.

Martin, Ian (2017) "What is the Expected Return on the Market?" *The Quarterly Journal of Economics*, 132 (1), 367–433.

McLean, R David and Jeffrey Pontiff (2016) "Does academic research destroy stock return predictability?" *The Journal of Finance*, 71 (1), 5–32.

Mele, Antonio (2007) "Asymmetric stock market volatility and the cyclical behavior of expected returns," *Journal of financial economics*, 86 (2), 446–478.

Neely, Christopher J., David E. Rapach, Jun Tu, and Guofu Zhou (2014) "Forecasting the Equity Risk Premium: The Role of Technical Indicators," *Management Science*, 60 (7), 1772–1791.

Pástor, L'uboš and Robert F Stambaugh (2009) "Predictive systems: Living with imperfect predictors," *The Journal of Finance*, 64 (4), 1583–1628.

Paye, Bradley S and Allan Timmermann (2006) "Instability of return prediction models," *Journal of Empirical Finance*, 13 (3), 274–315.

Pettenuzzo, Davide, Allan Timmermann, and Rossen Valkanov (2014) "Forecasting stock returns under economic constraints," *Journal of Financial Economics*, 114 (3), 517–553.

Polk, Christopher, Samuel Thompson, and Tuomo Vuolteenaho (2006) "Cross-sectional forecasts of the equity premium," *Journal of Financial Economics*, 81 (1), 101–141.

Polson, Nicholas G. and James G. Scott (2011) "Shrink globally, act locally: sparse Bayesian regularization and prediction," in *Bayesian statistics 9*, 501–538: Oxford Univ. Press, Oxford.

Rapach, David E., Jack K. Strauss, and Guofu Zhou (2009) "Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy," *The Review of Financial Studies*, 23 (2), 821–862.

Rossi, Alberto (2018) "Predicting Stock Market Returns with Machine Learning," *Working paper, Maryland University*, –.

Tetlock, Paul C. (2023) "The Implied Equity Premium," *Working paper, Columbia University*, –.

Tibshirani, Robert (1996) "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58 (1), 267–288.

Timmermann, A. (2004) "Forecast combinations," in Elliott, G., C. W. J. Granger, and A. Timmermann eds. *Handbook of Economic Forecasting*, 1, Chap. 4, 135–196: North Holland.

Tsiakas, Ilias, Jiahan Li, and Haibin Zhang (2020) "Equity premium prediction and the state of the economy," *Journal of Empirical Finance*, 58, 75–95.

Welch, Ivo and Amit Goyal (2008) "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," *The Review of Financial Studies*, 21 (4), 1455–1508.

Zhao, Peng and Bin Yu (2006) "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7 (90), 2541–2563, http://jmlr.org/papers/v7/zhao06a.html.

Zou, Hui and Trevor Hastie (2005) "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (2), 301–320.

Table 1: Equity premium forecasts (Nov 1953 to Dec 2021)

	No constraint			Positive forecast constraint		
	R^2_{OOS}	MSE	Δ CER	R^2_{OOS}	MSE	Δ CER
Panel A: theoretically-motivated models						
CSP (Polk et al., 2006)	1.42	1.775	1.433	1.436	1.774	1.46
DP^{drift} (Campbell, 2018)	0.404	1.793	1.695	0.404	1.793	1.695
SOP (Ferreira and Santa-Clara, 2011)	1.465	1.774	3.143	1.537	1.773	3.233
Panel B: models with all predictors except those in Panel A						
OLS KS	−13.435	2.042	0.154	−3.843	1.869	2.893
Ridge	0.748	1.787	3.099	1.072	1.781	3.044
Lasso	0.242	1.796	2.412	0.814	1.786	2.501
Elastic Net	0.096	1.798	2.341	0.615	1.789	2.362
Horseshoe	−0.276	1.805	1.913	1.078	1.781	2.711
PLS	−5.216	1.894	−3.58	−2.327	1.842	−1.676
CSR(K=1)	0.561	1.79	1.535	0.612	1.789	1.607
CSR(optimal K)	1.874	1.766	4.534	2.234	1.76	4.461
CSR(average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	1.829	1.767	4.135	2.148	1.762	4.407
CSR(average all K)	−0.525	1.81	2.501	1.616	1.771	3.849
Panel C: models with all predictors						
OLS KS	−15.328	2.076	2.246	−6.176	1.911	3.154
Ridge	−0.137	1.803	3.933	0.741	1.787	3.668
Lasso	−1.072	1.82	2.227	−0.205	1.804	2.31
Elastic Net	−0.761	1.814	2.423	−0.009	1.8	2.388
Horseshoe	0.028	1.8	2.241	1.064	1.781	2.771
PLS	−5.265	1.895	−3.338	−2.369	1.843	−1.528
CSR (K=1)	0.651	1.788	1.642	0.713	1.787	1.733
CSR (optimal K)	0.57	1.79	4.38	0.975	1.783	4.117
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	0.935	1.783	4.469	2.08	1.763	4.707
CSR (average all K)	−0.387	1.807	3.114	1.945	1.765	4.079

Table 2: Full-sample estimation of optimal combinations (Nov 1958 to Dec 2021)

	CSP (Polk et al., 2006)			
	CSR ($K=1$)	CSR (Optimal K)	CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	CSR (Average all K)
$\hat{\delta}$	0.791	0.443	0.449	0.638
R^2	1.485	2.656	2.645	2.348
ΔCER	1.885	4.358	4.23	3.668
	DP^{drift} (Campbell, 2018)			
	CSR ($K=1$)	CSR (Optimal K)	CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	CSR (Average all K)
$\hat{\delta}$	0.425	0.312	0.319	0.572
R^2	0.749	2.255	2.227	1.595
ΔCER	2.01	4.331	4.005	2.736
	SOP (Ferreira and Santa-Clara, 2011)			
	CSR ($K=1$)	CSR (Optimal K)	CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	CSR (Average all K)
$\hat{\delta}$	1	0.442	0.446	0.665
R^2	1.465	2.567	2.505	2.141
ΔCER	3.143	4.725	4.393	3.817

The table reports estimates of the optimal combination of theoretical models with complete subset regression (CSR) models. The models are of the form $\hat{r}_{t+1} = \delta \hat{r}^{Theo}_{t+1} + (1-\delta)\hat{r}^{CSR}_{t+1}$, where $\hat{r}^{Theo}_{t+1}$ is the forecast from a theoretical model of the equity risk premium, $\hat{r}^{CSR}_{t+1}$ is the forecast from a CSR model, and $0 \leq \delta \leq 1$. The theoretical models considered are the CSP model of Polk et al. (2006), the DP^{drift} model of Campbell (2018), and the SOP model of Ferreira and Santa-Clara (2011). CSR ($K=1$) indicates the simple combination of univariate predictive regressions. CSR (optimal K) indicates a combination where the optimal K in the CSR procedure is chosen based on the lowest mean squared error in each validation sample. CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) indicates a combination using the average of models in the 75% model confidence set, using the Hansen et al. (2011) procedure.

Table 3: Iterated mean combinations (rolling 60-month δ) and enhanced theoretical models (Nov 1958 to Dec 2021)

	No constraint			Positive forecast constraint		
	R^2_{OOS}	MSE	Δ CER	R^2_{OOS}	MSE	Δ CER
Panel A: CSR						
CSR ($K = 1$)	0.491	1.816	1.515	0.544	1.815	1.592
CSR (optimal K)	1.852	1.791	4.701	2.234	1.784	4.626
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	1.817	1.792	4.310	2.158	1.785	4.605
Panel B: CSP (Polk et al., 2006)						
Baseline	0.672	1.812	0.940	0.688	1.812	0.968
IMC (Baseline, CSR ($K = 1$))	0.299	1.819	1.327	0.376	1.818	1.456
IMC (Baseline, CSR (optimal K))	1.533	1.797	4.073	2.006	1.788	4.201
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	1.548	1.796	3.731	1.870	1.791	3.933
NMA (Baseline, CSR ($K = 1$))	0.902	1.808	1.918	0.870	1.809	1.872
NMA (Baseline, CSR (optimal K))	2.257	1.784	4.023	2.173	1.785	3.929
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	2.259	1.784	3.975	2.115	1.786	3.783
Panel C: DP^{drift} (Campbell, 2018)						
Baseline	0.564	1.814	2.040	0.564	1.814	2.040
IMC (Baseline, CSR ($K = 1$))	0.417	1.817	1.398	0.430	1.817	1.418
IMC (Baseline, CSR (optimal K))	0.893	1.808	3.385	1.417	1.799	3.627
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	0.934	1.808	2.706	1.415	1.799	3.167
NMA (Baseline, CSR ($K = 1$))	0.775	1.811	2.176	0.775	1.811	2.176
NMA (Baseline, CSR (optimal K))	2.217	1.784	4.179	2.255	1.784	4.280
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	2.207	1.784	4.035	2.212	1.784	4.110
Panel D: SOP (Ferreira and Santa-Clara, 2011)						
Baseline	1.205	1.803	2.901	1.281	1.801	2.998
IMC (Baseline, CSR ($K = 1$))	−8.666	1.983	−1.948	−0.708	1.838	1.197
IMC (Baseline, CSR (optimal K))	1.384	1.799	3.895	1.93	1.790	4.151
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	1.305	1.801	3.54	1.863	1.791	3.959
NMA (Baseline, CSR ($K = 1$))	1.021	1.806	2.474	1.073	1.805	2.546
NMA (Baseline, CSR (optimal K))	2.463	1.780	4.665	2.575	1.778	4.726
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	2.408	1.781	4.406	2.512	1.779	4.504

Table 4: Iterated mean combinations (expanding δ) and enhanced theoretical models (Nov 1958 to Dec 2021)

	No constraint			Positive forecast constraint		
	R^2_{OOS}	MSE	Δ CER	R^2_{OOS}	MSE	Δ CER
Panel A: CSR						
CSR ($K = 1$)	0.491	1.816	1.515	0.544	1.815	1.592
CSR (optimal K)	1.852	1.791	4.701	2.234	1.784	4.626
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	1.817	1.792	4.310	2.158	1.785	4.605
Panel B: CSP (Polk et al., 2006)						
Baseline	0.636	1.813	1.336	0.67	1.813	1.402
IMC (Baseline, CSR ($K = 1$))	0.636	1.813	1.336	0.67	1.813	1.402
IMC (Baseline, CSR (optimal K))	1.711	1.794	3.657	1.902	1.79	3.752
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	1.688	1.794	3.503	1.833	1.791	3.668
NMA (Baseline, CSR ($K = 1$))	0.902	1.808	1.918	0.870	1.809	1.872
NMA (Baseline, CSR (optimal K))	2.257	1.784	4.023	2.173	1.785	3.929
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	2.259	1.784	3.975	2.115	1.786	3.783
Panel C: DP^{drift} (Campbell, 2018)						
Baseline	0.564	1.814	2.040	0.564	1.814	2.04
IMC (Baseline, CSR ($K = 1$))	0.529	1.815	1.794	0.529	1.815	1.794
IMC (Baseline, CSR (optimal K))	1.96	1.789	4.627	2.31	1.783	4.647
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	2.008	1.788	4.125	2.193	1.785	4.277
NMA (Baseline, CSR ($K = 1$))	0.775	1.811	2.176	0.775	1.811	2.176
NMA (Baseline, CSR (optimal K))	2.217	1.784	4.179	2.255	1.784	4.280
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	2.207	1.784	4.035	2.212	1.784	4.110
Panel D: SOP (Ferreira and Santa-Clara, 2011)						
Baseline	1.205	1.803	2.901	1.281	1.801	2.998
IMC (Baseline, CSR ($K = 1$))	−2.886	1.877	−1.432	0.44	1.817	1.516
IMC (Baseline, CSR (optimal K))	1.993	1.788	4.241	2.284	1.783	4.446
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	1.886	1.790	3.865	2.192	1.785	4.263
NMA (Baseline, CSR ($K = 1$))	1.021	1.806	2.474	1.073	1.805	2.546
NMA (Baseline, CSR (optimal K))	2.463	1.780	4.665	2.575	1.778	4.726
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	2.408	1.781	4.406	2.512	1.779	4.504

Table 5: Equity premium forecasts (Jan 1996 to Dec 2021)

	No constraint			Positive forecast constraint		
	R^2_{OOS}	MSE	Δ CER	R^2_{OOS}	MSE	Δ CER
Panel A: theoretically-motivated models						
CSP (Polk et al., 2006)	0.703	1.917	−0.273	0.784	1.915	−0.205
DP^{drift} (Campbell, 2018)	−0.692	1.944	1.373	−0.692	1.944	1.373
SOP (Ferreira and Santa-Clara, 2011)	0.418	1.922	2.329	0.698	1.917	2.701
$SVIX^2$ (Martin, 2017)	0.922	1.912	1.613	0.922	1.912	1.613
Panel B: models with all predictors except those in Panel B						
OLS KS	−16.419	2.247	1.895	−7.879	2.082	3.803
Ridge	−2.05	1.97	0.715	−1.157	1.953	1.313
Lasso	−3.101	1.99	−0.455	−1.398	1.957	0.237
Elastic Net	−3.241	1.993	−0.5	−1.569	1.96	0.165
Horseshoe	−3.066	1.989	−0.095	−0.76	1.945	0.616
PLS	−5.076	2.028	−2.428	−3.784	2.003	−0.789
CSR (K=1)	−0.177	1.934	0.964	−0.05	1.931	1.156
CSR (optimal K)	−0.311	1.936	3.169	0.144	1.927	3.353
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	−0.606	1.942	3.358	0.254	1.925	4.005
CSR (average all K)	−3.241	1.993	2.413	−0.278	1.936	3.759
Panel C: models with all predictors						
OLS KS	−16.941	2.257	3.487	−10.423	2.131	3.589
Ridge	−2.831	1.985	3.183	−0.603	1.942	3.369
Lasso	−3.789	2.003	0.167	−1.854	1.966	0.745
Elastic Net	−3.594	2	0.06	−1.44	1.958	0.687
Horseshoe	−3.147	1.991	−0.166	−0.685	1.943	0.636
PLS	−5.798	2.042	−2.371	−3.775	2.003	−0.842
CSR (K=1)	−0.137	1.933	0.97	0.016	1.93	1.207
CSR (optimal K)	−1.533	1.96	4.33	−0.608	1.942	4.224
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	−3.102	1.99	3.539	−0.192	1.934	4.173
CSR (average all K)	−2.78	1.984	2.939	0.39	1.923	3.974

Table 6: Full-sample estimation of optimal combinations (Jan 1996 to Dec 2021)

	CSP (Polk et al., 2006)			
	CSR ($K=1$)	CSR (Optimal K)	CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	CSR (Average all K)
$\hat{\delta}$	1	0.766	0.747	0.853
R^2	0.703	0.808	0.874	0.825
ΔCER	-0.273	1.537	1.552	1.018

	DP^{drift} (Campbell, 2018)			
	CSR ($K=1$)	CSR (Optimal K)	CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	CSR (Average all K)
$\hat{\delta}$	0.275	0.442	0.489	0.681
R^2	-0.091	0.336	0.364	0.025
ΔCER	1.515	3.037	2.815	2.246

	SOP (Ferreira and Santa-Clara, 2011)			
	CSR ($K=1$)	CSR (Optimal K)	CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	CSR (Average all K)
$\hat{\delta}$	1	0.656	0.681	0.839
R^2	0.418	0.695	0.706	0.558
ΔCER	2.329	3.246	3.204	2.822

	$SVIX^2$ (Martin, 2017)			
	CSR ($K=1$)	CSR (Optimal K)	CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	CSR (Average all K)
$\hat{\delta}$	0.858	0.656	0.662	0.767
R^2	0.953	1.39	1.462	1.346
ΔCER	1.908	3.626	3.545	3.287

The table reports estimates of the optimal combination of theoretical models with complete subset regression (CSR) models. The models are of the form $\hat{r}_{t+1} = \delta \hat{r}^{Theo}_{t+1} + (1-\delta)\hat{r}^{CSR}_{t+1}$, where $\hat{r}^{Theo}_{t+1}$ is the forecast from a theoretical model of the equity risk premium, $\hat{r}^{CSR}_{t+1}$ is the forecast from a CSR model, and $0 \leq \delta \leq 1$. The theoretical models considered are the CSP model of Polk et al. (2006), the DP^{drift} model of Campbell (2018), and the SOP model of Ferreira and Santa-Clara (2011). CSR ($K=1$) indicates the simple combination of univariate predictive regressions. CSR (optimal K) indicates a combination where the optimal K in the CSR procedure is chosen based on the lowest mean squared error in each validation sample. CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) indicates a combination using the average of models in the 75% model confidence set, using the Hansen et al. (2011) procedure.

Table 7: Iterated mean combinations (rolling 60-month δ) and enhanced theoretical models - $SVIX^2$ period (Jan 2001 to Dec 2021)

	No constraint			Positive forecast constraint		
	R^2_{OOS}	MSE	Δ CER	R^2_{OOS}	MSE	Δ CER
Panel A: CSR						
CSR ($K = 1$)	−0.192	1.835	1.494	−0.026	1.832	1.73
CSR (optimal K)	−0.205	1.835	4.406	0.389	1.825	4.64
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	−0.563	1.842	4.67	0.56	1.821	5.474
Panel B: CSP (Polk et al., 2006)						
Baseline	1.059	1.812	−0.001	1.165	1.81	0.083
IMC (Baseline, CSR ($K = 1$))	0.237	1.827	1.566	0.509	1.822	1.887
IMC (Baseline, CSR (optimal K))	0.929	1.815	3.904	1.703	1.8	4.384
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	0.707	1.819	3.998	1.729	1.8	4.506
NMA (Baseline, CSR ($K = 1$))	0.618	1.82	1.18	0.618	1.82	1.18
NMA (Baseline, CSR (optimal K))	1.044	1.813	3.449	1.265	1.808	3.719
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	1.108	1.811	3.494	1.413	1.806	3.748
Panel C: DP^{drift} (Campbell, 2018)						
Baseline	−0.682	1.844	1.685	−0.682	1.844	1.685
IMC (Baseline, CSR ($K = 1$))	−0.701	1.844	0.849	−0.658	1.844	0.911
IMC (Baseline, CSR (optimal K))	−0.898	1.848	3.009	−0.117	1.834	3.36
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	−0.908	1.848	2.655	−0.149	1.834	3.046
NMA (Baseline, CSR ($K = 1$))	−0.157	1.835	1.991	−0.157	1.835	1.991
NMA (Baseline, CSR (optimal K))	0.54	1.822	3.433	0.723	1.818	3.719
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	0.604	1.821	3.382	0.927	1.815	3.752
Panel D: SOP (Ferreira and Santa-Clara, 2011)						
Baseline	0.529	1.822	2.516	0.895	1.815	2.975
IMC (Baseline, CSR ($K = 1$))	−8.447	1.986	0.101	1.281	1.808	4.724
IMC (Baseline, CSR (optimal K))	0.104	1.83	3.436	0.989	1.814	3.983
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	0.006	1.832	3.773	1.161	1.81	4.385
NMA (Baseline, CSR ($K = 1$))	0.272	1.827	2.125	0.455	1.823	2.375
NMA (Baseline, CSR (optimal K))	0.836	1.816	4.07	1.37	1.807	4.413
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	0.818	1.817	3.81	1.542	1.803	4.365

Table 7: *(Continued)*

	No constraint			Positive forecast constraint		
	R^2_{OOS}	MSE	Δ CER	R^2_{OOS}	MSE	Δ CER
Panel D: $SVIX^2$ (Martin, 2017)						
Baseline	0.608	1.82	1.036	0.608	1.82	1.036
IMC (Baseline, CSR ($K = 1$))	-7.162	1.963	-0.007	2.517	1.786	3.09
IMC (Baseline, CSR (optimal K))	0.471	1.823	3.796	1.15	1.811	4.012
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	0.277	1.827	3.378	1.215	1.809	3.64
NMA (Baseline, CSR ($K = 1$))	0.66	1.82	1.874	0.66	1.82	1.874
NMA (Baseline, CSR (optimal K))	1.442	1.805	3.805	1.322	1.807	3.616
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	1.514	1.804	3.666	1.521	1.804	3.678

The table reports the out-of-sample r-squared (R^2_{OOS}), the out-of-sample mean squared error (MSE), and the change in the certain equivalent return (Δ CER) for portfolios constructed based on each forecast, relative to portfolios constructed using the historical average benchmark. Panel A reports results for complete subset regressions (CSR). The remaining panels report results for different baseline theoretically-motivated model (i.e. identical to those in Table 1), iterated mean combinations (IMC) and naive model averages (NMA) of the baseline model and different versions of CSR. The IMC models are of the form $\hat{r}_{t+1} = \delta \hat{r}^{Theo}_{t+1} + (1-\delta)\hat{r}^{CSR}_{t+1}$, where $\hat{r}^{Theo}_{t+1}$ is the forecast from a theoretical model of the equity risk premium, $\hat{r}^{CSR}_{t+1}$ is the forecast from a CSR model, and $0 \leq \delta \leq 1$ is estimated via constrained least squares using a rolling window of 60 months. The NMA models fix $\delta = 0.5$.

Table 8: Iterated mean combinations (expanding δ) and enhanced theoretical models - $SVIX^2$ period (Jan 2001 to Dec 2021)

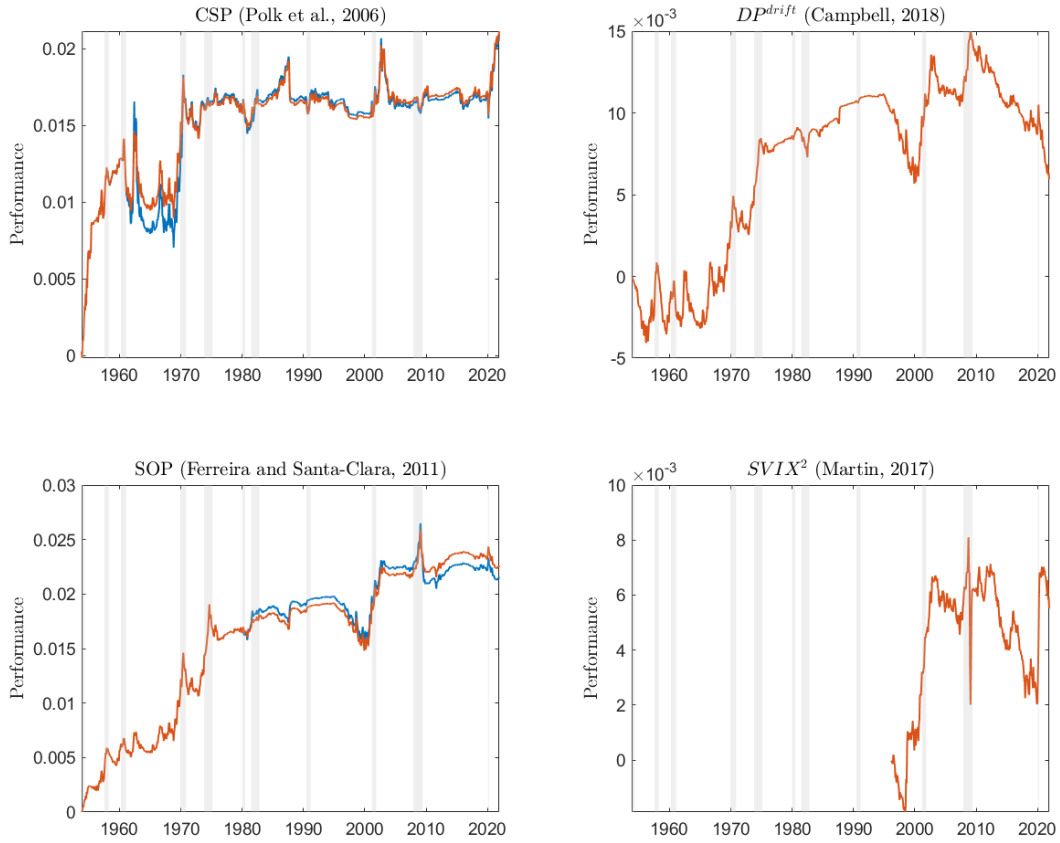
	No constraint			Positive forecast constraint		
	R^2_{OOS}	MSE	Δ CER	R^2_{OOS}	MSE	Δ CER
Panel A: CSR						
CSR ($K=1$)	−0.192	1.835	1.494	−0.026	1.832	1.73
CSR (optimal K)	−0.205	1.835	4.406	0.389	1.825	4.64
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	−0.563	1.842	4.67	0.56	1.821	5.474
Panel B: CSP (Polk et al., 2006)						
Baseline	1.059	1.812	−0.001	1.165	1.81	0.083
IMC (Baseline, CSR ($K=1$))	0.446	1.823	1.701	0.717	1.818	2.022
IMC (Baseline, CSR (optimal K))	0.676	1.819	4.096	1.436	1.805	4.618
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	0.406	1.824	4.021	1.574	1.803	4.513
NMA (Baseline, CSR ($K=1$))	0.618	1.82	1.18	0.618	1.82	1.18
NMA (Baseline, CSR (optimal K))	1.044	1.813	3.449	1.265	1.808	3.719
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	1.108	1.811	3.494	1.413	1.806	3.748
Panel C: DP^{drift} (Campbell, 2018)						
Baseline	−0.682	1.844	1.685	−0.682	1.844	1.685
IMC (Baseline, CSR ($K=1$))	−0.499	1.841	1.546	−0.499	1.841	1.546
IMC (Baseline, CSR (optimal K))	−0.42	1.839	3.292	0.396	1.824	3.403
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	−0.441	1.84	3.351	0.51	1.822	3.634
NMA (Baseline, CSR ($K=1$))	−0.157	1.835	1.991	−0.157	1.835	1.991
NMA (Baseline, CSR (optimal K))	0.54	1.822	3.433	0.723	1.818	3.719
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	0.604	1.821	3.382	0.927	1.815	3.752
Panel D: SOP (Ferreira and Santa-Clara, 2011)						
Baseline	0.529	1.822	2.516	0.895	1.815	2.975
IMC (Baseline, CSR ($K=1$))	0.412	1.824	2.732	0.402	1.824	2.635
IMC (Baseline, CSR (optimal K))	−0.119	1.834	3.324	1.086	1.812	3.859
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	−0.064	1.833	3.854	1.233	1.809	4.326
NMA (Baseline, CSR ($K=1$))	0.272	1.827	2.125	0.455	1.823	2.375
NMA (Baseline, CSR (optimal K))	0.836	1.816	4.07	1.37	1.807	4.413
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	0.818	1.817	3.81	1.542	1.803	4.365

Table 8: *(Continued)*

	No constraint			Positive forecast constraint		
	R^2_{OOS}	MSE	Δ CER	R^2_{OOS}	MSE	Δ CER
Panel D: $SVIX^2$ (Martin, 2017)						
Baseline	0.608	1.82	1.036	0.608	1.82	1.036
IMC (Baseline, CSR ($K = 1$))	-1.852	1.866	-0.069	0.117	1.829	0.79
IMC (Baseline, CSR (optimal K))	-0.02	1.832	2.177	1.182	1.81	2.935
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	0.199	1.828	2.76	1.291	1.808	3.315
NMA (Baseline, CSR ($K = 1$))	0.66	1.82	1.874	0.66	1.82	1.874
NMA (Baseline, CSR (optimal K))	1.442	1.805	3.805	1.322	1.807	3.616
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$))	1.514	1.804	3.666	1.521	1.804	3.678

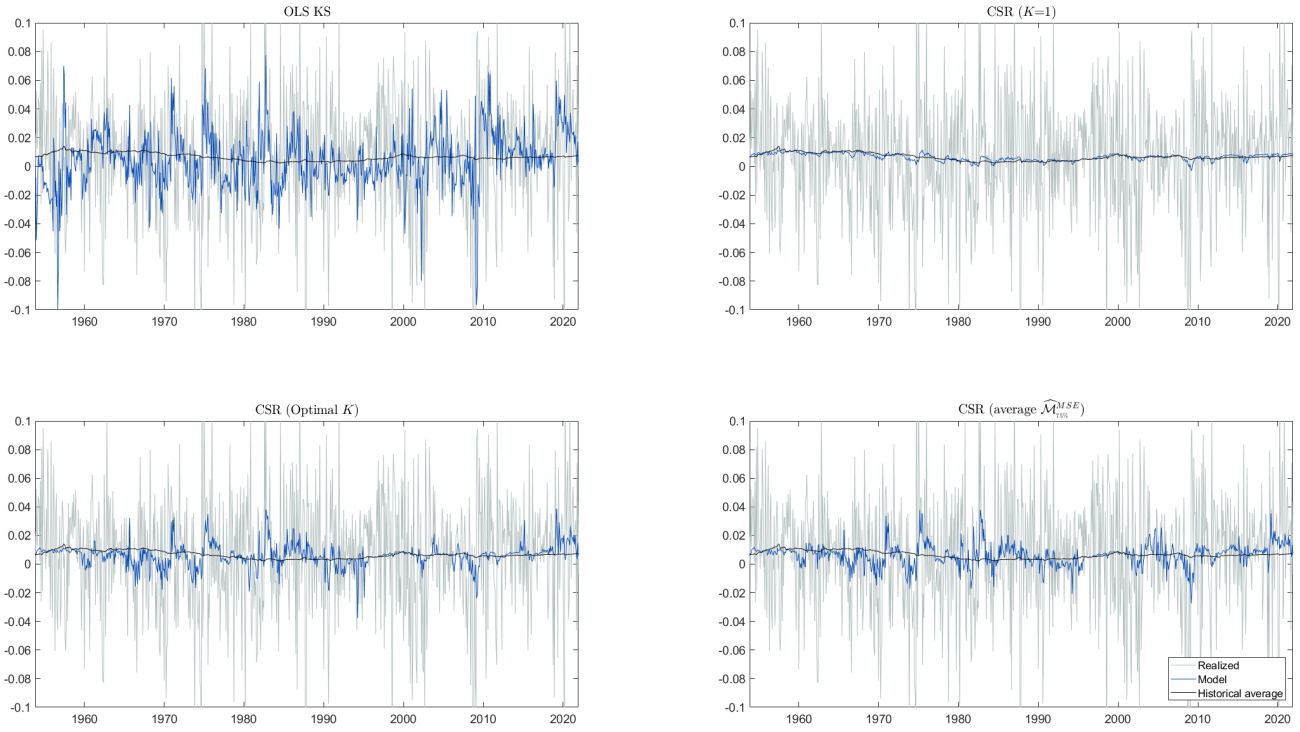
The table reports the out-of-sample r-squared (R^2_{OOS}), the out-of-sample mean squared error (MSE), and the change in the certain equivalent return (Δ CER) for portfolios constructed based on each forecast, relative to portfolios constructed using the historical average benchmark. Panel A reports results for complete subset regressions (CSR). The remaining panels report results for different baseline theoretically-motivated model (i.e. identical to those in Table 1), iterated mean combinations (IMC) and naive model averages (NMA) of the baseline model and different versions of CSR. The IMC models are of the form $\hat{r}_{t+1} = \delta \hat{r}^{Theo}_{t+1} + (1-\delta)\hat{r}^{CSR}_{t+1}$, where $\hat{r}^{Theo}_{t+1}$ is the forecast from a theoretical model of the equity risk premium, $\hat{r}^{CSR}_{t+1}$ is the forecast from a CSR model, and $0 \leq \delta \leq 1$ is estimated via constrained least squares using an expanding window months. The NMA models fix $\delta = 0.5$.

Figure 2: Performance of equity risk premium forecast models - Theoretical Models



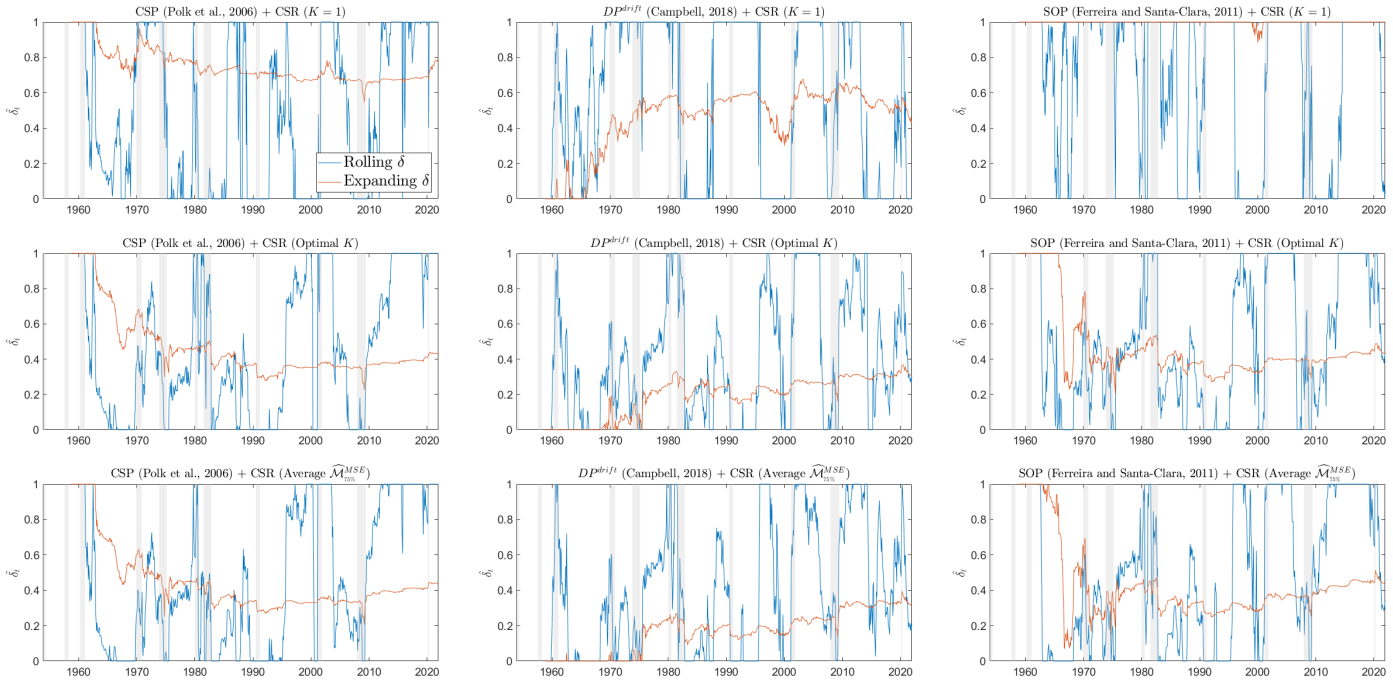
The figure displays the out-of-sample performance of identity-based models to forecast the equity risk premium. Each panel refers to a specific identity-based predictor. Each graph shows the cumulative squared error of the historical mean forecast minus the cumulative squared error of the alternative. The red line shows results when the positivity constraint is enforced, while the blue line shows results from the raw forecasts. Shaded areas indicate recessions.

Figure 3: Realized returns and forecasts from OLS model and complete subset regressions



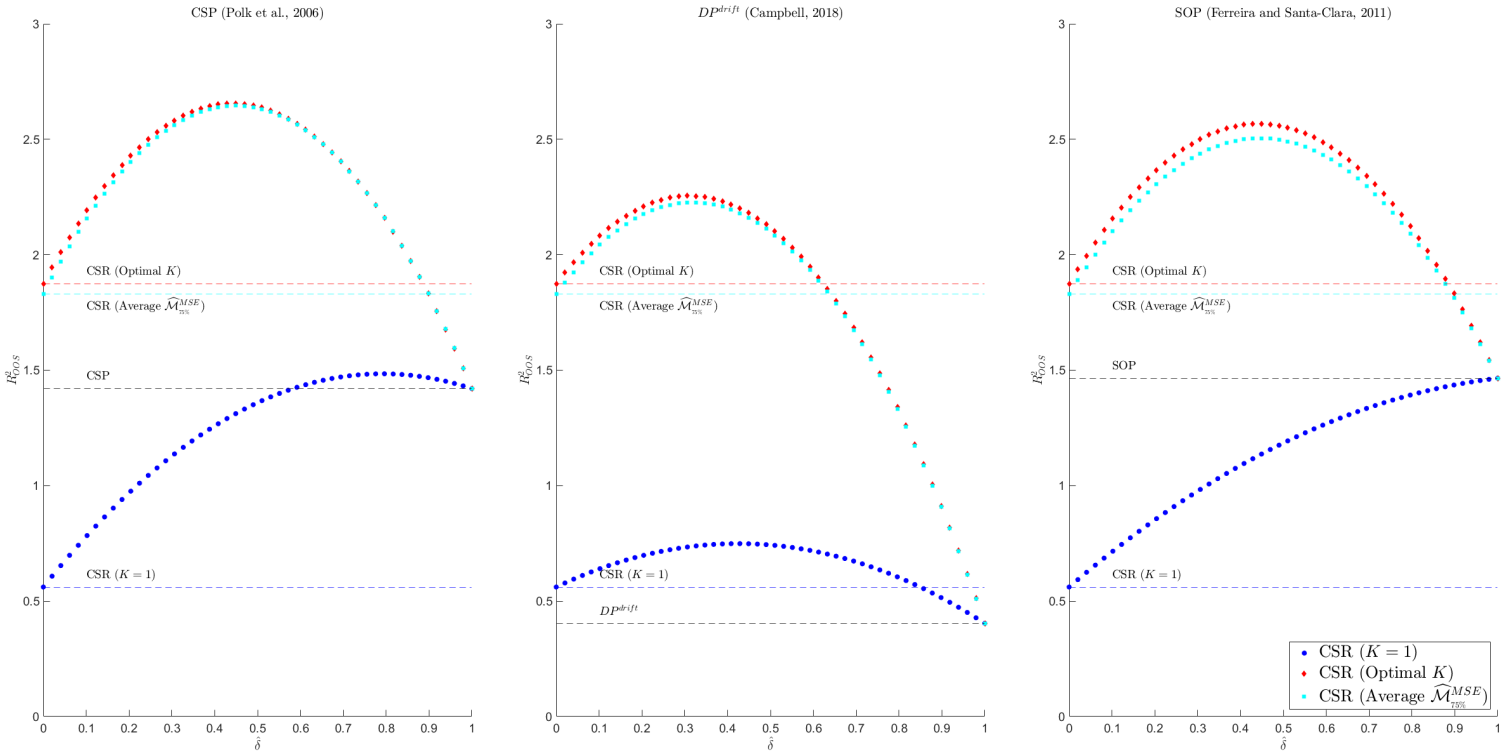
The solid gray line represents the realized excess market return. The solid black line represents the forecast using the historical average benchmark. The solid blue line represents the forecasts from the model shown in the title of each graph. Models are estimates using all available predictors, excluding those in theoretical models.

Figure 4: Optimal combination weight (δ)



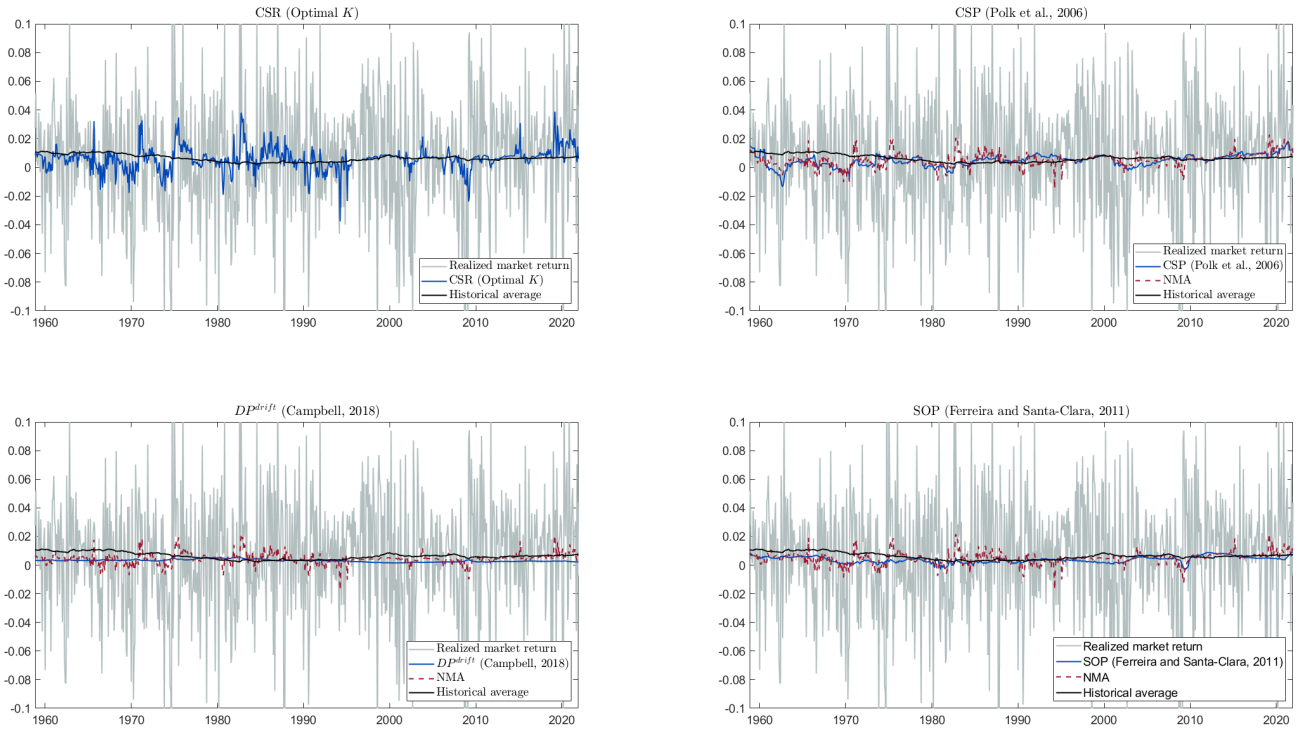
Each graph shows estimates of the optimal weight in a combination of theoretical models with complete subset regression (CSR) models. The models are of the form $\hat{r}_{t+1} = \delta \hat{r}^{Theo}_{t+1} + (1-\delta)\hat{r}^{CSR}_{t+1}$, where $\hat{r}^{Theo}_{t+1}$ is the forecast from a theoretical model of the equity risk premium, $\hat{r}^{CSR}_{t+1}$ is the forecast from a CSR model, and $0 \leq \delta \leq 1$. Optimal values of δ in each combination are obtained by constrained least square using either a rolling window of 60 months (blue line) or an expanding window (red line) approach. The theoretical models considered are the CSP model of Polk et al. (2006), the DP^{drift} model of Campbell (2018), and the SOP model of Ferreira and Santa-Clara (2011). CSR ($K = 1$) indicates the simple combination of univariate predictive regressions. CSR (lowest MSE) indicates a combination where the optimal K in the CSR procedure is chosen based on the lowest mean squared error in each validation sample. CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) indicates a combination using the average of models in the 75% model confidence set, using the Hansen et al. (2011) procedure. Shaded areas indicate NBER recessions.

Figure 5: R^2_{OOS} of combinations of theoretical models and and complete subset regressions



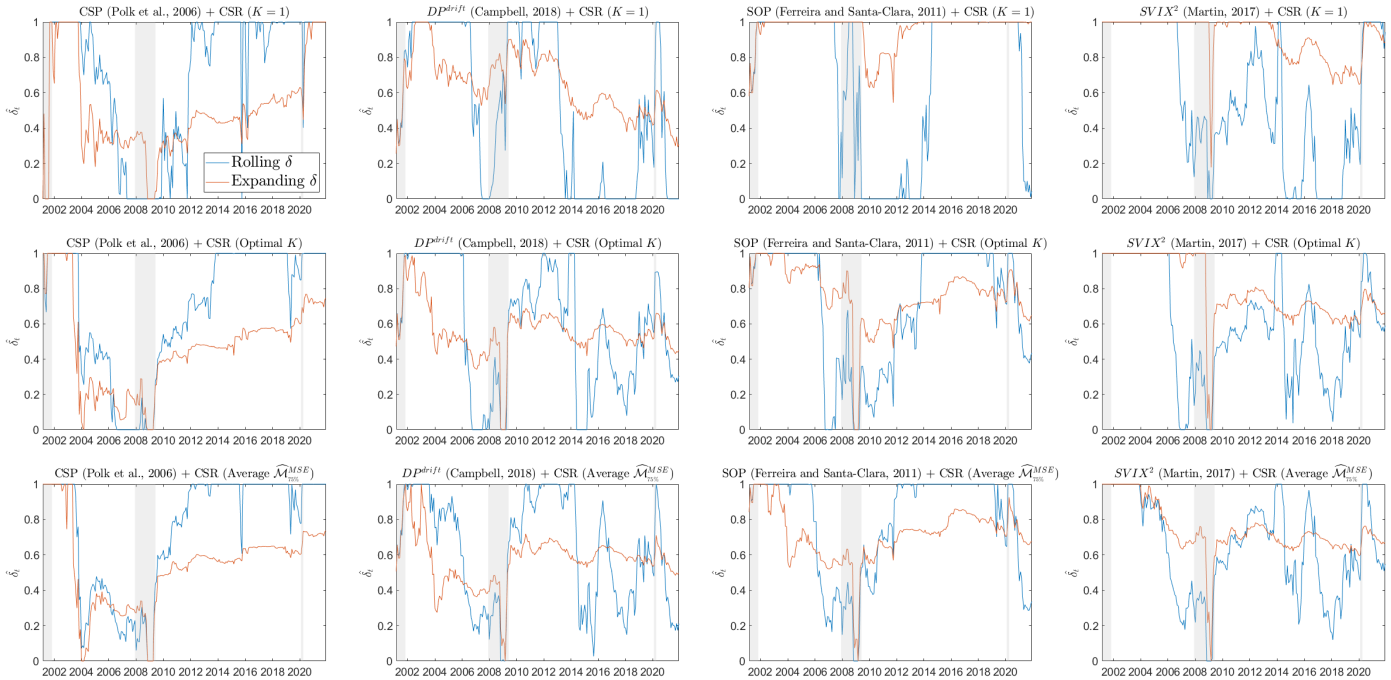
Each panel shows the out-of-sample R^2_{OOS} of models of the form $\hat{r}_{t+1} = \delta\hat{r}^{Theo}_{t+1} + (1-\delta)\hat{r}^{CSR}_{t+1}$, where $\hat{r}^{Theo}_{t+1}$ is the forecast from a theoretical model of the equity risk premium, $\hat{r}^{CSR}_{t+1}$ is the forecast from a CSR model, and $0 \leq \delta \leq 1$. The theoretical models considered are the CSP model of Polk et al. (2006), the DP^{drift} model of Campbell (2018), and the SOP model of Ferreira and Santa-Clara (2011). In each panel, dashed horizontal lines represent the R^2_{OOS} of the standalone model indicated. Blue circles, red diamonds, and magenta squares represent combinations of the theoretical model and each version of CSR for values of $1-\delta$ shown on the x axis. CSR (lowest MSE) indicates a combination where the optimal K in the CSR procedure is chosen based on the lowest mean squared error in each validation sample. CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) indicates a combination using the average of models in the 75% model confidence set, using the Hansen et al. (2011) procedure.

Figure 6: Forecasts of theoretical models and naive combinations with complete subset regressions



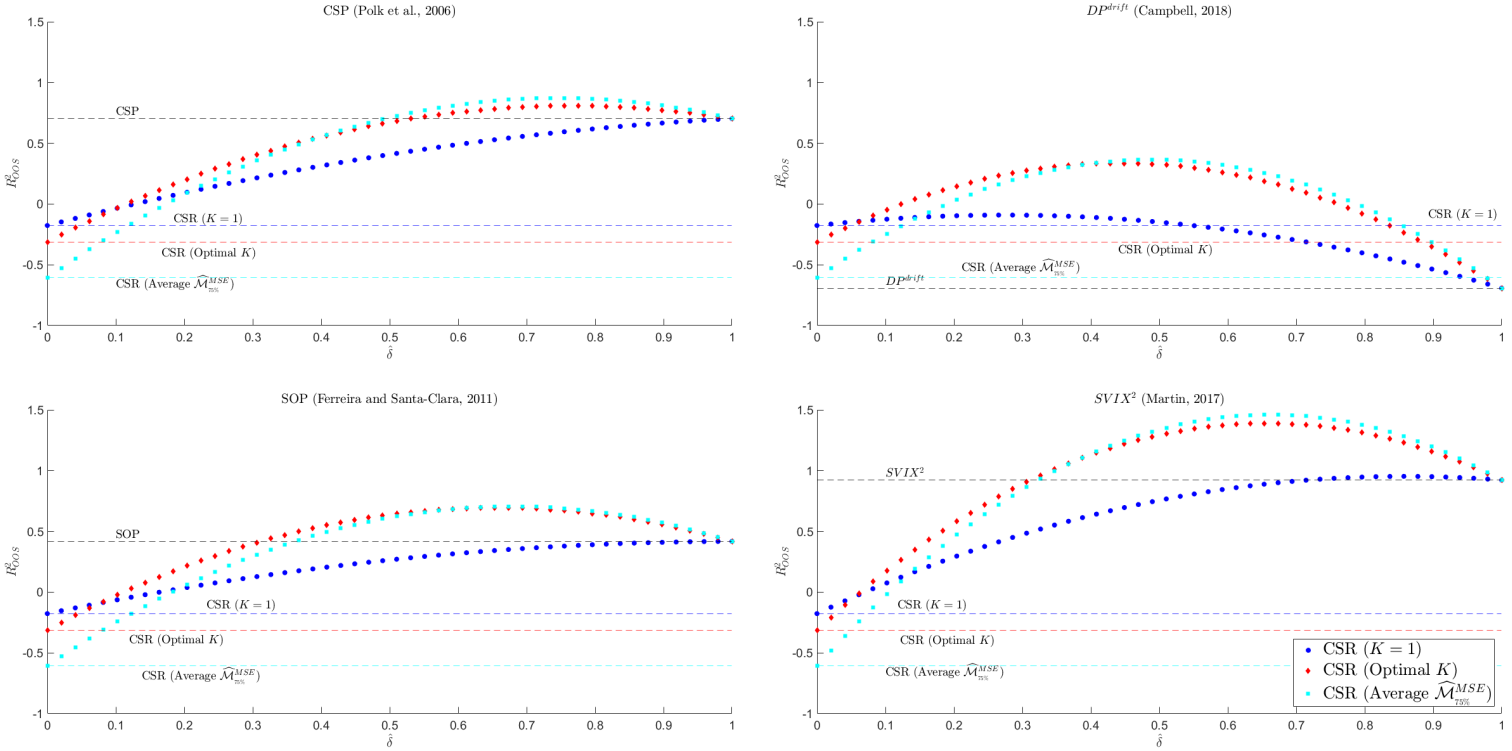
The solid gray line represents the realized excess market return. The solid black line represents the forecast using the historical average benchmark. The solid blue line represents the forecasts from the model shown in the title of each graph. The dashed red line indicates the naive model combination of the theoretical model and the Complete Subset Regression (Lowest MSE) model. Shaded areas indicate NBER recessions.

Figure 7: Optimal combination weight (δ) - $SVIX^2$ period (Jan 2001 - Dec 2021)



Each graph shows estimates of the optimal weight in a combination of theoretical models with complete subset regression (CSR) models. The models are of the form $\hat{r}_{t+1} = \delta \hat{r}^{Theo}_{t+1} + (1-\delta)\hat{r}^{CSR}_{t+1}$, where $\hat{r}^{Theo}_{t+1}$ is the forecast from a theoretical model of the equity risk premium, $\hat{r}^{CSR}_{t+1}$ is the forecast from a CSR model, and $0 \leq \delta \leq 1$. Optimal values of δ in each combination are obtained by constrained least square using either a rolling window of 60 months (blue line) or an expanding window (red line) approach. The theoretical models considered are the CSP model of Polk et al. (2006), the DP^{drift} model of Campbell (2018), and the SOP model of Ferreira and Santa-Clara (2011). CSR ($K = 1$) indicates the simple combination of univariate predictive regressions. CSR (lowest MSE) indicates a combination where the optimal K in the CSR procedure is chosen based on the lowest mean squared error in each validation sample. CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) indicates a combination using the average of models in the 75% model confidence set, using the Hansen et al. (2011) procedure. Shaded areas indicate NBER recessions.

Figure 8: R^2_{OOS} of combinations of theoretical models and and complete subset regressions - $SVIX^2$ period (Jan 1996 - Dec 2021)



Each panel shows the out-of-sample R^2_{OOS} of models of the form $\hat{r}_{t+1} = \delta)\hat{r}^{Theo}_{t+1} + (1-\delta)\hat{r}^{CSR}_{t+1}$, where $\hat{r}^{Theo}_{t+1}$ is the forecast from a theoretical model of the equity risk premium, $\hat{r}^{CSR}_{t+1}$ is the forecast from a CSR model, and $0 \leq \delta \leq 1$. The theoretical models considered are the CSP model of Polk et al. (2006), the DP^{drift} model of Campbell (2018), the SOP model of Ferreira and Santa-Clara (2011), and the $SVIX^2$ model of Martin (2017). In each panel, dashed horizontal lines represent the R^2_{OOS} of the standalone model indicated. Blue circles, red diamonds, and magenta squares represent combinations of the theoretical model and each version of CSR for values of $1-\delta$ shown on the x axis. CSR (lowest MSE) indicates a combination where the optimal K in the CSR procedure is chosen based on the lowest mean squared error in each validation sample. CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) indicates a combination using the average of models in the 75% model confidence set, using the Hansen et al. (2011) procedure.

Online Appendix

A Description of data

Table A.1: Summary of predictors used in the study

Variable	Description	Source of data
DY	Dividend yield	
EP	Earnings/Price	
$RVOL$	Equity risk premium volatility (Mele 2007, JFE)	
BM	Book/Market ratio	
$NTIS$	Net equity expansion	Amit Goyal's website
TBL	Treasury bill rate	
LTR	Long-term rate of return for government bonds	
TMS	Term Spread	
DFY	Default Yield Spread	
DFR	Default Return Spread	
$INFL$	Inflation	
CSP	Cross-sectional Premium (Polk et al, 2006)	Self-calculated (CRSP/Compustat)
SI	Short interest	Dave Rapach's website
DP^{drift}	Adjusted dividend/price ratio (Campbell, 2018)	Self-calculated using Goyal's data
SOP	Sum of moving average of growth in earnings ($\widehat{ge}$) and log of Dividend/Price ratio ($\widehat{dp}$) (Ferreira and Santa-Clara, 2011)	Self-calculated using Goyal's data
$SVIX^2$	Bound on equity premium (Martin, 2017)	Self-calculated from CRSP/Compustat data
$\overline{r}_{LS}$	Average of long-short anomalies portfolios (inspired by Dong et al, 2022)	Self-calculated with data from OSAP
$MA(1,9)$	Moving average cross-over $L_2$1 month/9 months)	Self-calculated using Goyal's data
$MA(1,12)$	Moving average cross-over (1 month, 12 months)	Self-calculated using Goyal's data
$MA(2,9)$	Moving average cross-over (2 months, 9 months)	Self-calculated using Goyal's data
$MA(2,12)$	Moving average cross-over (2 months, 12 months)	Self-calculated using Goyal's data
$MA(3,9)$	Moving average cross-over (3 months, 9 months)	Self-calculated using Goyal's data
$MA(3,12)$	Moving average cross-over (3 months, 12 months)	Self-calculated using Goyal's data
$TSMOM(9)$	Time-series momentum indicator (9 months)	Self-calculated using Goyal's data
$TSMOM(12)$	Time-series momentum indicator (12 months)	Self-calculated using Goyal's data
$VOL(1,9)$	Volume-based indicator (1 month, 9 months)	Self-calculated (CRSP)
$VOL(1,12)$	Volume-based indicator (1 month, 12 months)	Self-calculated (CRSP)
$VOL(2,9)$	Volume-based indicator (2 months, 9 months)	Self-calculated (CRSP)
$VOL(2,12)$	Volume-based indicator (2 months, 12 months)	Self-calculated (CRSP)
$VOL(3,9)$	Volume-based indicator (3 months, 9 months)	Self-calculated (CRSP)
$VOL(3,12)$	Volume-based indicator (3 months, 12 months)	Self-calculated (CRSP)

B Complete Subset Regressions: Additional Analysis

B.1 Properties of the forecast combination

In the main manuscript, we have emphasized the good performance – both from a statistical and economic perspectives – generated by forecast combination strategies. We now provide more insights on the key properties of these forecasts. Given the discussion in Section 3.1 and the established ranking of the models, our discussion will centre on the forecast properties of CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) and of CSR (optimal K) based on a rolling window estimation.

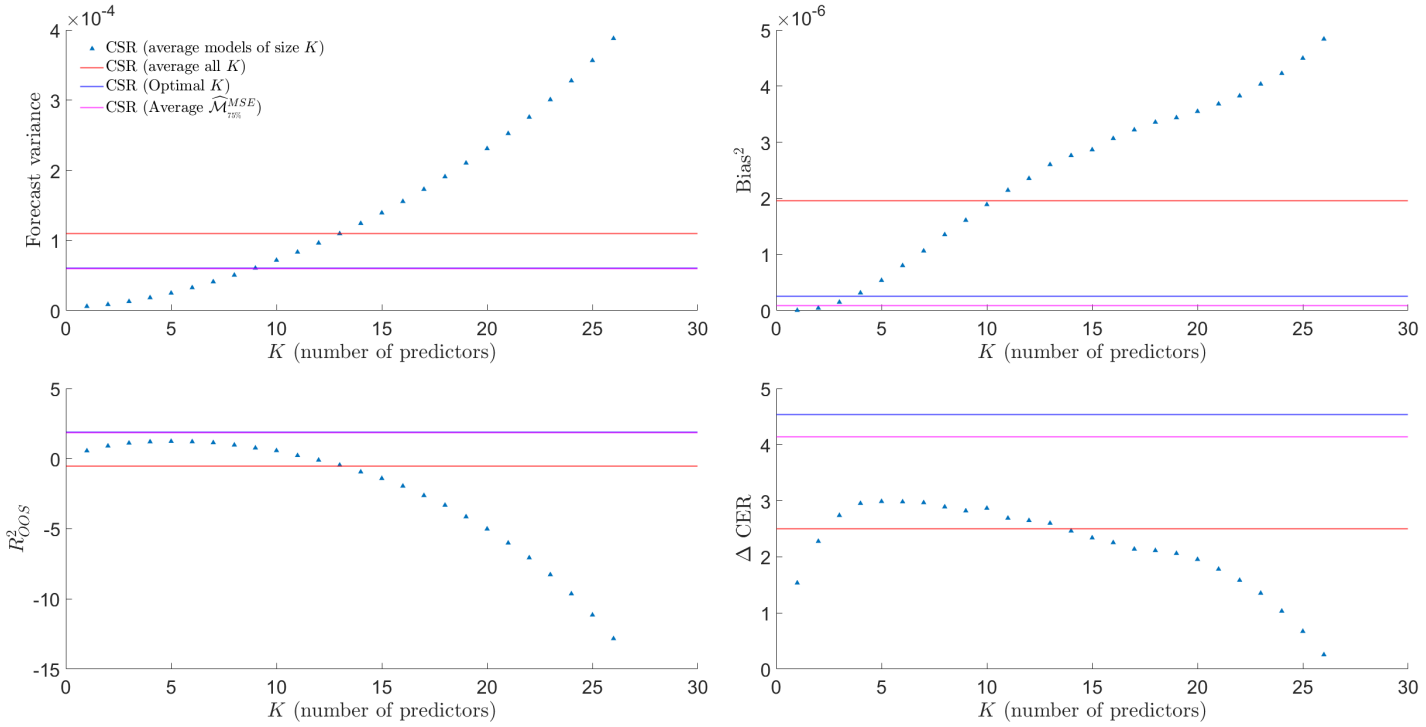
Figure B.1 reports the variance (upper left), squared bias (upper right), R^2_{oos} (bottom right), and the Δ CER of complete subset regressions forecasts for different values of K. In each graph, the solid triangles represent the corresponding value for the complete subset regression model that averages models with the corresponding number of predictors, shown in the x axis. The red line indicates the value attained by the forecast that average across all values of K, while the blue and magenta lines represent the value for the forecasts from the complete subset regressions (lowest MSE) and complete subset regression (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$). The upper left graph shows that the forecast variance increases with the value of K, as we move closer to the kitchen-sink regression, similar to the results reported by Elliott et al. (2013).[36] We observe that the complete subset regressions (lowest MSE) and complete subset regression (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) reduce forecast variance significantly (blue and magenta lines) relative to the forecast that averages across all K (red line).

The upper right graph in Figure B.1 shows also an increasing pattern for the square bias as a function of K. The forecasts from the CSR (optimal K) and CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) achieve a squared bias which is close to the minimum across all values of K. The bottom left graph in Figure B.1 shows the R^2_{oos} of complete subset regressions for each value of K. The R^2_{oos} remains positive up to $K = 12$, and the R^2_{oos} of the complete subset regression (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$), shown by the magenta line, is slightly superior to that of any individual K, which further confirms that the method is able to select appropriate sets of values for K over time.

[36] Note that, although we have 31 predictors in total, we only report results for complete subset regressions with up to 29 predictors in this graph. The reason is that the data for two predictors (short interest, SI, and $SVIX^2$) becomes available later than other variables, and therefore we only have forecasts for $K = 30$ and $K = 31$ once those predictors have a complete history in the combined training and validation sample.

Finally, the bottom right graph in Figure B.1 reports the ΔCER for the different complete subset regression models. Although the complete subset regression beats the historical average for all values of K in terms of certain equivalent returns, we notice substantial gains from our two preferred CSR models, which achieve a higher economic performance compared to CSRs with any specific value for K.

Figure B.1: Out-of-sample statistical and economic performance



Solid triangles represent results obtained using a complete subset regression model in which forecasts from all models with a given number of predictors (K) is averaged. The red line labeled "Average" represents a model that averages forecasts from all models which up to 15 predictors. The blue line labeled "CSR (optimal K)" represents a model in which a single value of K in the complete subset regression model is chosen at each iteration to minimize the validation mean squared error. The magenta line labeled "CSR (Average $\widehat{\mathcal{M}}_{75\%}$)" corresponds to dynamic selection of a set of values of K using the model confidence set approach of Hansen et al. (2011). The top left, top right, bottom left, and bottom right panels report forecast variances, squared biases, out-of-sample r-squared, and the difference between the certain equivalent return of portfolios formed using each model, relative to the historical mean.

B.2 The dynamics of model uncertainty

Given the good performance of the CSRs, one may wonder how volatile are the optimal model dimension K and the size of the model confidence set over which we average models. The short answer, for both, is: a lot. Figure B.2 shows the optimal model dimension over time, when K is chosen according to the lowest MSE in each validation sample. In several circumstances it is optimal to average over models that use most or even all of the predictors, e.g. in the early 1960s, mid 1990s, in the period 2003-2007, and finally in the last part of our sample. We also observe persistent sequences of months where the combination of univariate models ($K = 1$) is optimal, e.g., in the late 1990s and between 2010 and 2017 (with minor interruptions).[37]

Figure B.2 shows the dimension of the models in the MCS (chosen using the MSE) over time. At each month, a blue cross for a specific value of K indicates that the complete subset regression with K predictors was included in the model confidence set, and therefore included in the forecast. The red line indicates the number of different complete subset regressions over which we are averaging. For example, in the late 1990s, only univariate models enter the MCS (K=1 and the crosses overlap with the red line). The graph also shows that, occasionally, the complete subset regressions is optimal. This case is denoted by a red line flat at 29 to 31 (depending on availability of predictors, as discussed previously).

Although using the full set of predictors is sometimes optimal, it is clear that an intermediate value – i.e., K lower than the total number of predictors – is often common, for e.g., in the period 1970-1980 and again in the mid 1980s. Furthermore, it is interesting to note that low dimensional models – denoted by a white area in the top part of the graph – were optimal in the early part of our sample, in the late 1970s and 1990s, and late 2000s.

[37] The relatively high level of variation in the optimal K that we document contrasts with the mostly stable (and low) values of K obtained by Elliott et al. (2013, , Fig. 9). The differences, however, are reconciled when we compare the method to select the hyperparameter K. Elliott et al. (2013) use the *cumulative* out-of-sample performance of the complete subset regressions for each value of K to select the optimal K at each month, starting with a period of five years of out-of-sample forecasts, whereas we rely on a rolling validation window with a length of 60 months to select the optimal K. Figure C.4 in the Appendix shows our implementation of both methods using an expanding window for consistency with their results. Using the Elliott et al. (2013) approach produces, as expected, lower and more stable values for the optimal K over time. However, we note that the performance of this approach is not superior to that of the approach we propose using rolling windows.

Overall, the graph suggests that although it is clearly important to consider the dimension of the model K as an uncertain parameter, i.e., averaging unconditionally over all possible model sizes is unlikely to be optimal for equity premium forecasting.

Next, we investigate the relationship between model dimension and macroeconomic conditions by regressing the optimal K in the top panel of Figure B.2 and the median K in the bottom panel of the same figure onto three variables: market realised volatility (RV), defined as the square root of sum of squares of daily returns on each month; the Baker and Wurgler (2006) sentiment index (BW); and the NBER recession indicator (REC).[38] It is important to highlight that the optimal number of predictors using either the lowest MSE or median model dimension in the MCS at month t are determined using the validation window, which consists of the previous 5 years. Because of this, it is appropriate to include lags of the explanatory variables. However, including too many lags would shorten the sample and would likely result in multi-collinearity issues.

To keep the model parsimonious and to use as much of the data as possible, we adopt the following approach. For each explanatory variable, we calculate lags of 1 to 24 months. We then include in the regression the contemporaneous value, an average of the lags from 1 to 12 months, and an average of the lags from months 13 to 24:

$$\begin{aligned} MD_t = \beta_0 &+ \beta_1 RV_t + \beta_2 \overline{RV}_{t-12:t-1} + \beta_3 \overline{RV}_{t-24:t-13} \\ &+ \beta_4 BW_t + \beta_5 \overline{BW}_{t-12:t-1} + \beta_6 \overline{BW}_{t-24:t-13} \\ &+ \beta_7 REC_t + \beta_8 \overline{REC}_{t-12:t-1} + \beta_9 \overline{REC}_{t-24:t-13} + \varepsilon_t, \end{aligned}$$

where MD_t denotes the model dimension, which is either the optimal K obtained using the MSE, or the median model dimension from the models in the MCS. We report results for different models including each explanatory variable (as well as its average lags, as explained above) one at a time, and the three variables (and their average lags) simultaneously. We also report the total impact of each variable, calculated as the sum of the corresponding estimated

[38] The realised volatility is the square root of the $SVAR$ variable obtained from Amit Goyal's webpage at https://sites.google.com/view/agoyal145?pli=1. The Baker and Wurgler (2006) sentiment index was downloaded from Jeffrey Wurgler's website at https://pages.stern.nyu.edu/~jwurgler/. The NBER recession indicator was downloaded from the St. Louis Federal Reserve website at https://fred.stlouisfed.org/series/USRECM).

regression coefficients. For example, the total impact of RV is calculated as $\widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3$.[39]

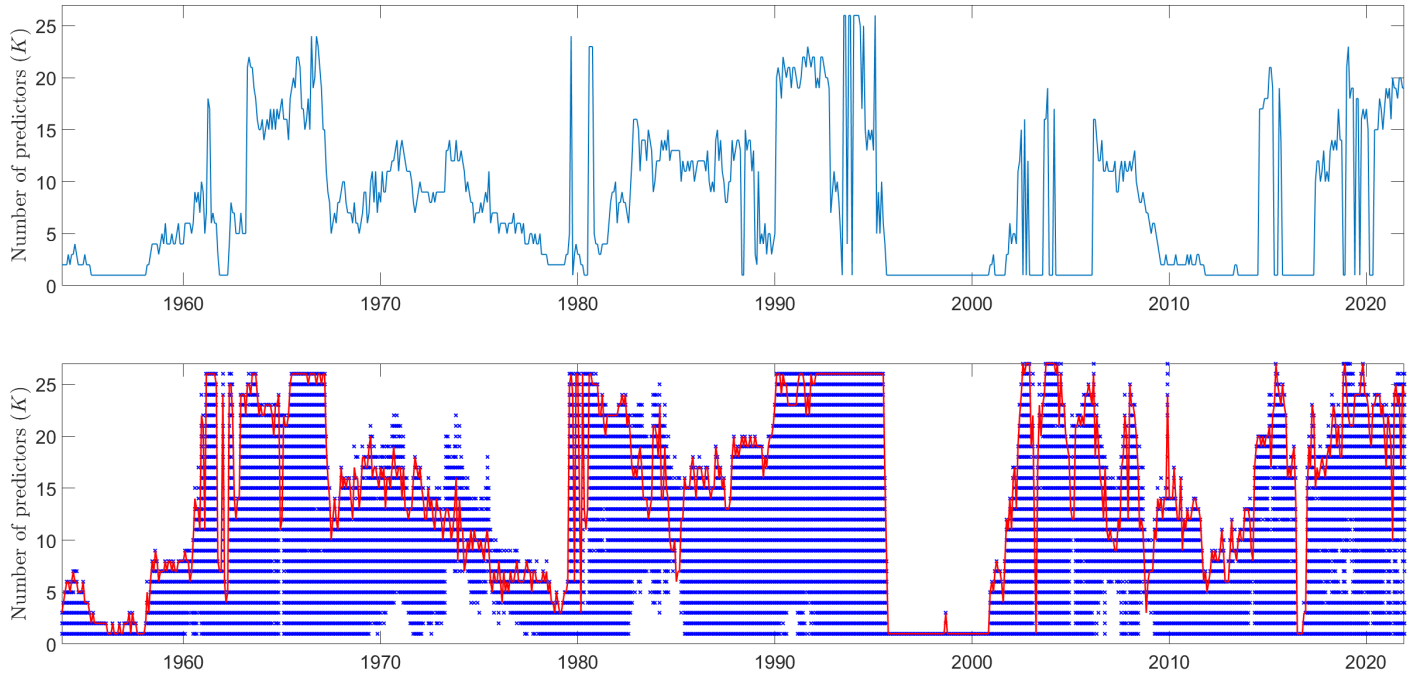
Table B.1 shows the results. Columns (1) through (4) report results when model dimensionality is proxied with the optimal K using the MSE criterion. Column (1) reports results when only the current and lagged values of RV_t are used. The results suggest that an increase in realised volatility (contemporaneous or lagged) is associated with a decrease in model dimensionality, i.e., during turbulent periods, models of lower dimension are typically selected. The total effect of RV is negative and significant at the 5% level.

When only the Baker and Wurgler (2006) sentiment index (BW) is considered, the results in column (2) suggest that neither the contemporaneous, nor the lagged effects are significant; however, the total effect os BW is positive and significant at the 5% level. This suggests that periods of high sentiment are associated with models of higher dimension. Column (3) reports results using only the recession indicator and its lags. Overall, there does not seem to be much association between recessions and the model dimension.

We find similar results when all regressors are used simultaneously in column (4), with RV and BW being negatively and positively associated with model dimensionality. Interestingly, when controlling for volatility and sentiment, the overall impact of REC becomes positive, although it is not statistically significant. Columns (5) through (8) of Table B.1 report results using as response variable the median model dimension in the MCS procedure. When all predictors are used simultaneously in column (8), similar conclusions are reached in terms of the signs of the overall impacts of the regressors in comparison with column (4), although the realised volatility is only statistically significant at the 10% level, while recessions now appear to be positive associated with higher median model dimension.

[39]The corresponding standard errors are calculated using the covariance matrix of errors corrected for heteroscedasticity and autocorrelation using the Newey-West procedure.

Figure B.2: Optimal model dimension in complete subset regression



The top graph shows the optimal number of predictors selected in the complete subset regression (Lowest MSE) approach. At each point in time during the forecasting exercise, all complete subset regressio models are estimated in the training window. Forecasts are made for the validation set, and the optimal value of K is the one with the lowest validation mean squared error. The bottom graph shows which sets of K are included in the CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) approach over time. At each point in time during the forecasting exercise, the model confidence set (MCS) procedure of Hansen et al. (2011) is run to identify a set of models for which the null of equal predictive ability is not rejected. The values of K that are part of the model confidence set at each point are shown as a blue "x". The red line shows how many values of K are part of the MCS at each point in time.

Table B.1: Determinants of model dimension

	Optimal K (lowest MSE in validation)				Median K - models in MCS			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	8.086***	8.086***	6.951***	6.951***	9.223***	9.223***	8.593***	8.180***
RV_t	−0.293			−0.269	−0.153			−0.359
$\overline{RV}_{t-12:t-1}$	−0.240			−1.159**	−0.356			−1.087
$\overline{RV}_{t-24:t-13}$	−1.284***			−1.471**	−0.197			−0.192
BW_t		0.480		1.290		−0.458		−0.048
$\overline{BW}_{t-12:t-1}$		0.367		−0.755		1.042		0.427
$\overline{BW}_{t-24:t-13}$		0.585		1.021		−0.036		0.279
REC_t			−0.589	−0.750			1.312	1.471
$\overline{REC}_{t-12:t-1}$			3.304	5.568**			1.646	3.352**
$\overline{REC}_{t-24:t-13}$			0.218	3.178*			1.484	2.530*
Total RV	−1.817***			−2.898***	−0.706*			−1.638***
Total BW		1.433***		1.556***		0.548		0.658
Total REC			2.932	7.996***			4.441***	7.352***
R^2_{Adj}	0.045	0.030	0.011	0.128	0.012	0.019	0.043	0.127

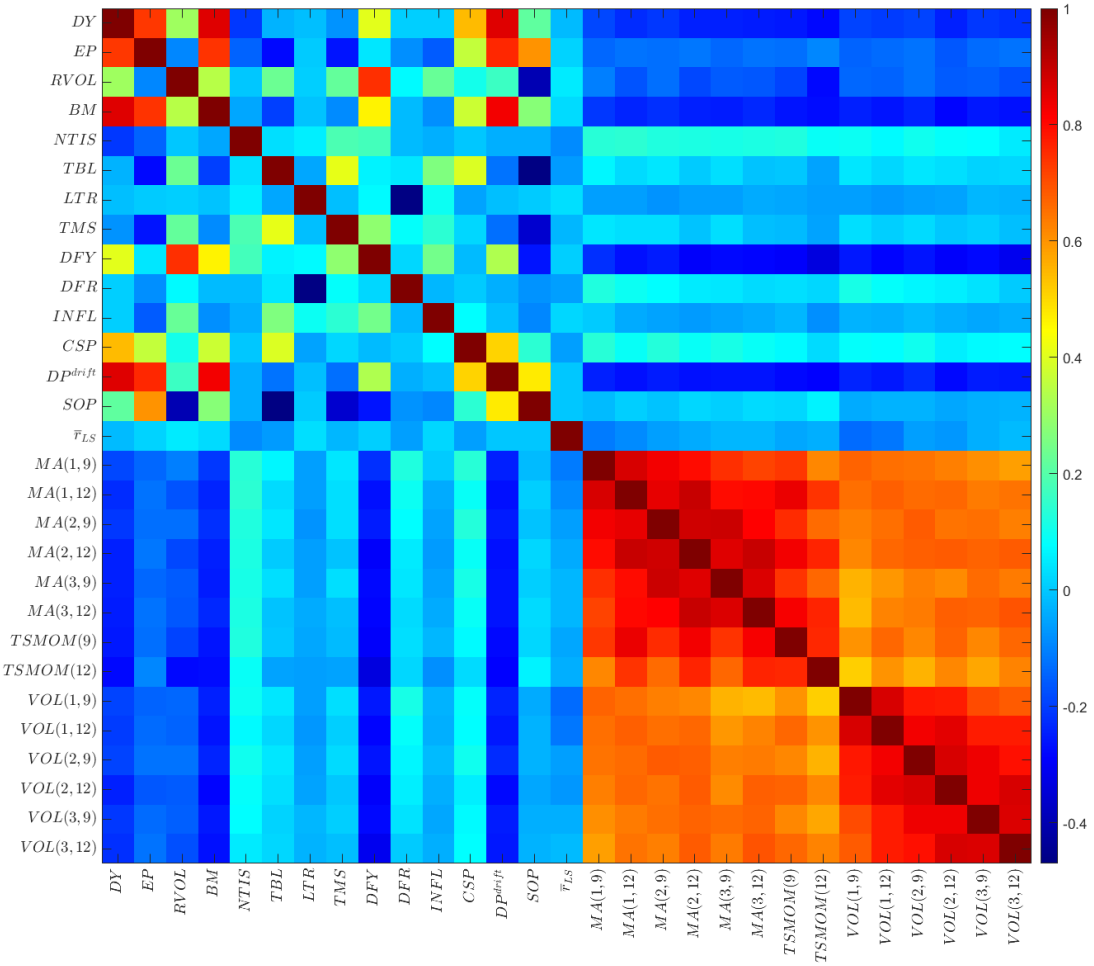
The table reports results from regression models of the form

$$\begin{aligned} MD_t = \beta_0 &+ \beta_1 RV_t + \beta_2 \overline{RV}_{t-12:t-1} + \beta_3 \overline{RV}_{t-24:t-13} \\ &+ \beta_4 BW_t + \beta_5 \overline{BW}_{t-12:t-1} + \beta_6 \overline{BW}_{t-24:t-13} \\ &+ \beta_7 REC_t + \beta_8 \overline{REC}_{t-12:t-1} + \beta_9 \overline{REC}_{t-24:t-13} + \varepsilon_t. \end{aligned}$$

MD_t denotes the model dimension, either the optimal K from the complete subset regression that minimizes the MSE in each validation sample (see Figure B.2) or the median model dimension of the models in the MCS using the MSE criterion (the median of the selected Ks in Figure B.2). RV_t, BW_t, and REC_t denote contemporaneous values of realized volatility, the Baker and Wurgler (2006) investor sentiment index, and a recession dummy. The remaining variables are averages of past values of the corresponding predictors. For example, $\overline{RV}_{t-12:t-1}$ is the average of the lags of RV from time $t-12$ to time $t-1$. The values reported under "Total RV", "Total BW", and "Total REC" are the sums of the corresponding coefficients. For example, "Total RV" in column1.000is the sum of the coefficients on RV_t, $\overline{RV}_{t-12:t-1}$, and $\overline{RV}_{t-24:t-13}$. *, **, and *** denotes statistical significance at the 10%, 5%, and 1% levels, respectively. p-values are calculated using heteroscedasticity and autocorrelation consistent standard errors using the Newey-West procedure.

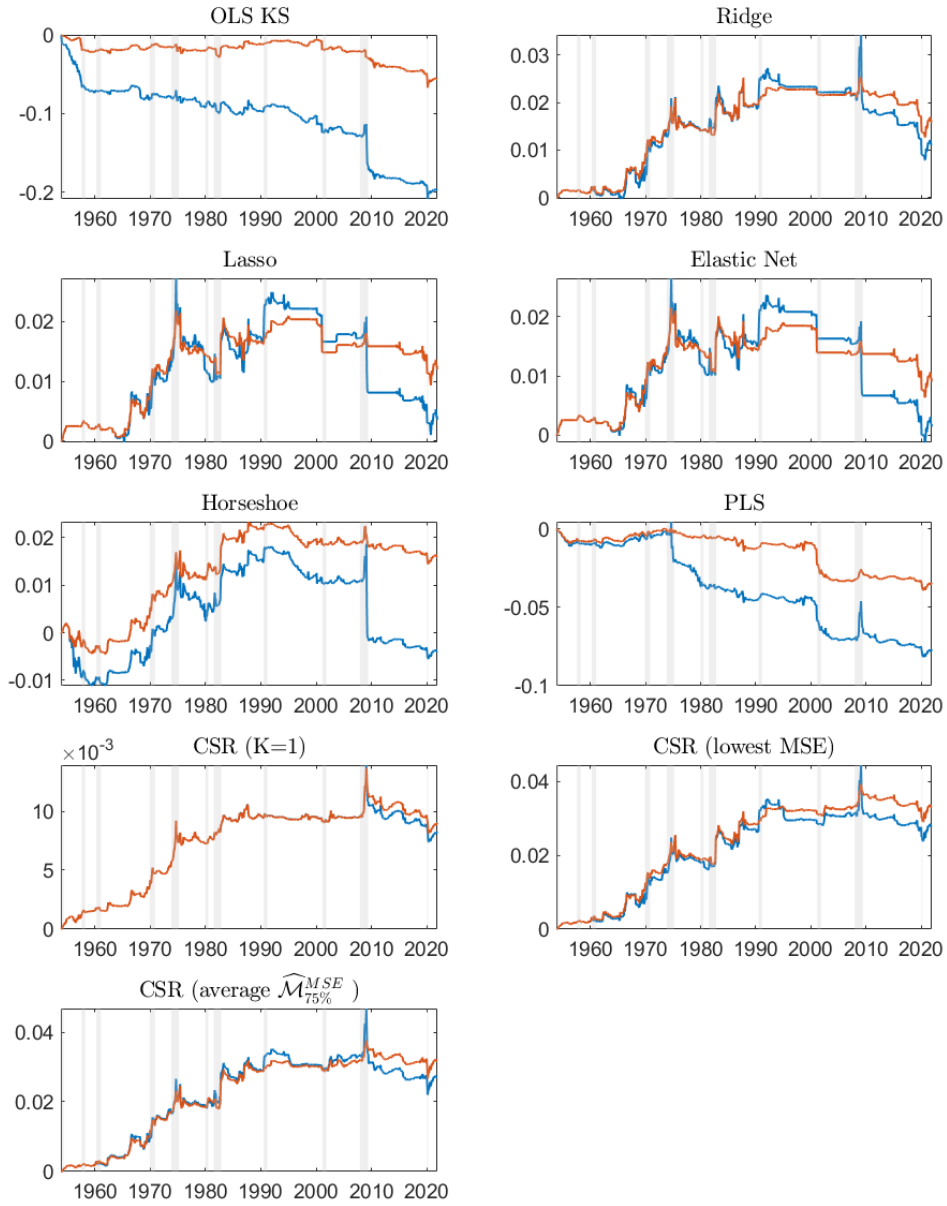
C Additional tables and graphs

Figure C.1: Correlations among predictors



The figure shows a correlation heatmap of the main predictors used in this study that are available throughout the entire sample period from November 1928 to Decembe 2021.

Figure C.2: Performance of equity risk premium forecast models - All Models



The figure displays the out-of-sample performance of statistical models to forecast the equity risk premium. Each graph shows the cumulative squared error of the historical mean forecast minus the cumulative squared error of the alternative. The red line shows results when the positivity constraint is enforced, while the blue line shows results from the raw forecasts. Shaded areas indicate recessions.

Figure C.3: Performance of equity risk premium forecast models - Identity-based models vs CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)

The figure displays the out-of-sample performance of identity-based models to forecast the equity risk premium, relative to the complete subset regression model that averages over dimensions in model confidence set. Each graph shows the cumulative squared error of the CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) forecast minus the cumulative squared error of the alternative. The red line shows results when the positivity constraint is enforced, while the blue line shows results from the raw forecasts. Shaded areas indicate recessions.

Table C.1: Equity premium forecast with different sets of predictors (Nov 1953 to Dec 2021)

	R^2_{OOS} - No constraint			R^2_{OOS} - Positive forecast constraint		
	ALL	MACRO	TECH	ALL	MACRO	TECH
OLS KS	−13.435	−6.781	−3.697	−3.843	−2.375	−1.361
Ridge	0.748	0.316	−1.056	1.072	1.294	−0.351
Lasso	0.242	−1.026	−1.076	0.814	0.389	−0.281
Elastic Net	0.096	−0.888	−1.022	0.615	0.513	−0.25
Horseshoe	−0.276	−0.499	−0.502	1.078	1.256	−0.336
PLS	−5.216	−3.304	−2.794	−2.327	−0.814	−1.309
CSR (K=1)	0.561	0.912	0.046	0.612	0.984	0.219
CSR (optimal K)	1.874	0.691	−0.867	2.234	1.846	−0.251
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	1.829	1.439	−0.557	2.148	2.107	−0.069
CSR (average all K)	−0.525	0.65	−0.467	1.616	1.707	0.221

The table reports the out-of-sample r-squared (R^2_{OOS}) of different forecasting models for three sets of predictors. The columns "ALL", "MACRO", and "TECH" refers to models estimated using all available predictors, only macroeconomic predictors, and only predictors based on technical indicators, respectively. OLS KS is the "kitchen sink" regression model using all predictors. PLS is the partial least squares method. CSR denotes complete subset regressions using different approaches to select the optimal model dimension(s). CSP is the cross-sectional premium predictor of Polk et al. (2006). SOP is the "sum-of-parts" model of Ferreira and Santa-Clara (2011).

Table C.2: Asset allocation based on equity premium forecasts (Nov 1958 to Dec 2021)

	No constraint					
	Ave.	Std.	SR	$\overline{w}$	$\sigma(w)$	Turnover
Panel A: Buy and Hold and historical average						
Buy and Hold	7.068	14.741	0.477	100	0	0
Historical average	8.711	14.212	0.61	88.226	54.529	3.354
Panel B: Baseline CSR models						
CSR ($K = 1$)	9.613	13.322	0.718	86.577	53.185	8.52
CSR (Optimal K)	14.048	15.081	0.929	70.839	79.666	41.093
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	14.163	15.738	0.897	73.261	82.749	45.187
Panel C: CSP (Polk et al., 2006)						
Baseline	8.896	13.108	0.675	62.492	63.883	8.457
IMC (Baseline, CSR ($K = 1$))	9.238	13.038	0.705	76.994	61.278	11.481
IMC (Baseline, CSR (optimal K))	12.82	14.264	0.896	64.983	77.383	33.61
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	12.455	14.232	0.872	66.518	76.642	34.296
NMA (Baseline, CSR ($K = 1$))	9.247	12.112	0.76	74.512	49.054	7.744
NMA (Baseline, CSR (optimal K))	11.951	13.064	0.911	66.486	64.456	26.56
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	12.042	13.275	0.904	67.975	64.254	28.277
Panel D: DP^{drift} Campbell (2018)						
Baseline	6.793	6.609	1.021	40.584	17.021	1.435
IMC (Baseline, CSR ($K = 1$))	8.358	11.487	0.724	69.869	50.063	7.181
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	10.947	12.491	0.873	58.69	67.089	34.088
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	10.444	12.77	0.814	58.812	67.344	35.205
NMA (Baseline, CSR ($K = 1$))	8.182	9.686	0.84	63.9	33.276	4.549
NMA (Baseline, CSR (optimal K))	10.871	11.01	0.983	56.879	52.087	26.89
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	10.947	11.403	0.956	58.328	52.71	28.617
Panel E: SOP (Ferreira and Santa-Clara, 2011)						
Baseline	8.245	8.205	1	50.937	35.906	5.836
IMC (Baseline, CSR ($K = 1$))	6.895	14.397	0.476	13.993	96.752	21.711
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	11.155	11.998	0.926	56.907	61.954	30.767
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	10.995	12.319	0.889	57.767	62.412	30.794
NMA (Baseline, CSR ($K = 1$))	8.908	10.533	0.842	69.076	42.328	5.725
NMA (Baseline, CSR (optimal K))	11.74	11.686	1.001	61.861	57.634	26.794
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	11.762	12.157	0.964	63.452	59.032	28.453

The table reports statistics for optimal portfolios based on forecasts from different models. The columns show a portfolio's annual average return (Ave.), standard deviation of returns (Std.), Sharpe ratio (SR), the average and standard deviation of the allocation to the market portfolio ($\overline{w}$ and $\sigma(w)$, respectively), and the average portfolio turnover.

Table C.3: Asset allocation based on equity premium forecasts - positive forecast constraint (Nov 1958 to Dec 2021)

	positive forecast constraint					
	Ave.	Std.	SR	$\overline{w}$	$\sigma(w)$	Turnover
Panel A: Buy and Hold and historical average						
Buy and Hold	7.068	14.741	0.477	100	0	0
Historical average	8.711	14.212	0.61	88.226	54.529	3.354
Panel B: Baseline CSR models						
CSR ($K = 1$)	9.679	13.306	0.724	86.699	52.954	8.392
CSR (Optimal K)	13.315	14.181	0.936	80.476	64.08	30.701
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	13.736	14.792	0.926	83.243	67.111	34.099
Panel C: CSP (Polk et al., 2006)						
Baseline	8.712	12.779	0.678	66.875	56.938	6.985
IMC (Baseline, CSR ($K = 1$))	9.245	12.851	0.716	78.592	58.264	10.404
IMC (Baseline, CSR (optimal K))	12.294	13.314	0.92	73.688	64	24.595
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	12.455	14.232	0.872	66.518	76.642	34.296
NMA (Baseline, CSR ($K = 1$))	9.196	12.104	0.756	74.898	48.353	7.435
NMA (Baseline, CSR (optimal K))	11.592	12.651	0.913	71.112	56.56	21.276
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	11.557	12.826	0.898	72.543	56.443	22.844
Panel D: DP^{drift} Campbell (2018)						
Baseline	6.793	6.609	1.021	40.584	17.021	1.435
IMC (Baseline, CSR ($K = 1$))	8.376	11.485	0.726	69.891	50.03	7.137
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	10.652	11.599	0.915	65.547	55.695	25.899
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	10.347	11.863	0.869	65.442	56.532	27.417
NMA (Baseline, CSR ($K = 1$))	8.182	9.686	0.84	63.9	33.276	4.549
NMA (Baseline, CSR (optimal K))	10.703	10.51	1.014	60.522	45.379	21.908
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	10.734	10.887	0.982	61.785	46.508	24.016
Panel E: SOP (Ferreira and Santa-Clara, 2011)						
Baseline	8.325	8.163	1.015	51.453	34.991	5.44
IMC (Baseline, CSR ($K = 1$))	8.337	11.797	0.703	47.272	66.41	12.73
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	10.871	11.061	0.979	64.135	49.059	22.557
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	10.823	11.319	0.952	64.809	49.925	23.062
NMA (Baseline, CSR ($K = 1$))	8.972	10.517	0.849	69.205	42.089	5.623
NMA (Baseline, CSR (optimal K))	11.476	11.115	1.029	66.133	50.032	21.669
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	11.5	11.55	0.992	67.702	51.574	23.533

The table reports statistics for optimal portfolios based on forecasts from different models. The columns show a portfolio's annual average return (Ave.), standard deviation of returns (Std.), Sharpe ratio (SR), the average and standard deviation of the allocation to the market portfolio ($\overline{w}$ and $\sigma(w)$, respectively), and the average portfolio turnover.

Table C.4: Asset allocation based on equity premium forecasts - $SVIX^2$ period (Jan 2001 to Dec 2021)

	No constraint					
	Ave.	Std.	SR	$\overline{w}$	$\sigma(w)$	Turnover
Panel A: Buy and Hold and historical average						
Buy and Hold	8.409	14.838	0.566	100	0	0
Historical average	5.467	14.129	0.386	85.423	45.435	3.868
Panel B: Baseline CSR models						
CSR ($K = 1$)	6.127	12.895	0.474	86.24	52.829	9.609
CSR (Optimal K)	10.654	15.194	0.7	85.607	76.137	26.963
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	11.831	16.353	0.723	91.136	83.733	40.559
Panel C: CSP (Polk et al., 2006)						
Baseline	7.706	17.007	0.452	90.965	67.456	6.866
IMC (Baseline, CSR ($K = 1$))	7.409	14.652	0.505	88.313	67.425	10.556
IMC (Baseline, CSR (optimal K))	10.745	15.956	0.673	83.377	78.046	18.081
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	10.738	15.829	0.678	86.96	75.859	22.656
NMA (Baseline, CSR ($K = 1$))	6.939	14.537	0.476	88.855	57.417	7.395
NMA (Baseline, CSR (optimal K))	9.62	15.093	0.637	88.435	69.763	18.298
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	10.127	15.694	0.644	90.856	71.12	24.791
Panel D: DP^{drift} Campbell (2018)						
Baseline	3.108	6.154	0.503	33.519	19.093	1.357
IMC (Baseline, CSR ($K = 1$))	4.489	11.25	0.398	68.599	58.723	8.474
IMC (Baseline, CSR (optimal K)	7.248	12.27	0.59	62.662	70.097	22.122
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	6.677	11.91	0.56	60.639	67.242	25.959
NMA (Baseline, CSR ($K = 1$))	4.621	9.284	0.496	60.367	35.465	5.224
NMA (Baseline, CSR (optimal K))	7.342	11.718	0.625	63.314	56.534	20.975
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	7.798	12.553	0.62	65.68	59.217	27.75
Panel E: SOP (Ferreira and Santa-Clara, 2011)						
Baseline	5.645	10.301	0.547	65.984	36.854	4.207
IMC (Baseline, CSR ($K = 1$))	5.52	14.061	0.392	−6.51	91.334	24.709
IMC (Baseline, CSR (optimal K)	7.453	11.901	0.625	66.95	53.269	16.572
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	7.704	11.755	0.654	66.192	51.276	16.483
NMA (Baseline, CSR ($K = 1$))	5.89	11.471	0.512	76.6	43.779	6.159
NMA (Baseline, CSR (optimal K))	9.001	13.349	0.673	78.964	61.135	20.192
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	9.312	14.179	0.656	81.619	64.245	27.075
Panel F: $SVIX^2$ (Martin, 2017)						
Baseline	5.223	12.185	0.428	40.401	37.985	13.239
IMC (Baseline, CSR ($K = 1$))	5.574	14.29	0.389	−15.842	79.815	27.341
IMC (Baseline, CSR (optimal K)	8.595	13.151	0.653	63.475	65.634	25.195
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	8.151	13.11	0.621	61.77	63.579	28.613
NMA (Baseline, CSR ($K = 1$))	5.467	11.166	0.489	64.406	33.182	9.187
NMA (Baseline, CSR (optimal K))	8.103	12.365	0.654	67.188	51.258	23.828
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	8.541	13.265	0.643	69.426	54.427	30.681

Table C.5: Asset allocation based on equity premium forecasts - positive forecast constraint - $SVIX^2$ period (Jan 2001 to Dec 2021)

	Positive forecast constraint					
	Ave.	Std.	SR	$\overline{w}$	$\sigma(w)$	Turnover
Panel A: Buy and Hold and historical average						
Buy and Hold	8.409	14.838	0.566	100	0	0
Historical average	5.467	14.129	0.386	85.423	45.435	3.868
Panel B: Baseline CSR models						
CSR ($K = 1$)	6.333	12.848	0.492	86.6	52.137	9.24
CSR (Optimal K)	10.058	14.059	0.714	91.007	66.154	21.709
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	11.71	15.179	0.771	98.89	70.043	32.321
Panel C: CSP (Polk et al., 2006)						
Baseline	7.783	16.999	0.457	91.295	66.985	6.613
IMC (Baseline, CSR ($K = 1$))	7.692	14.601	0.526	89.002	66.409	9.934
IMC (Baseline, CSR (optimal K))	10.44	14.939	0.698	88.196	69.71	13.72
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	10.484	14.835	0.706	91.36	67.932	18.032
NMA (Baseline, CSR ($K = 1$))	6.939	14.537	0.476	88.855	57.417	7.395
NMA (Baseline, CSR (optimal K))	9.548	14.633	0.652	90.302	66.421	16.081
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	9.997	15.196	0.657	93.611	66.241	21.571
Panel D: DP^{drift} Campbell (2018)						
Baseline	3.108	6.154	0.503	33.519	19.093	1.357
IMC (Baseline, CSR ($K = 1$))	4.548	11.245	0.403	68.655	58.65	8.361
IMC (Baseline, CSR (optimal K)	6.838	10.958	0.623	66.566	63.673	18.224
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	6.319	10.578	0.596	64.654	60.717	21.65
NMA (Baseline, CSR ($K = 1$))	4.621	9.284	0.496	60.367	35.465	5.224
NMA (Baseline, CSR (optimal K))	7.011	10.614	0.659	66.049	51.431	17.767
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	7.492	11.426	0.655	69.478	52.251	23.352
Panel E: SOP (Ferreira and Santa-Clara, 2011)						
Baseline	6.072	10.239	0.592	66.682	35.324	3.949
IMC (Baseline, CSR ($K = 1$))	7.823	10.244	0.762	36.016	56.001	13.228
IMC (Baseline, CSR (optimal K)	7.231	10.53	0.685	70.888	44.13	13.001
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	7.528	10.328	0.728	70.143	41.792	13.24
NMA (Baseline, CSR ($K = 1$))	6.118	11.432	0.534	76.901	43.167	5.922
NMA (Baseline, CSR (optimal K))	8.686	12.325	0.704	81.741	55.478	17.586
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$	9.143	13.119	0.696	85.333	56.819	23.451
Panel F: $SVIX^2$ (Martin, 2017)						
Baseline	5.223	12.185	0.428	40.401	37.985	13.239
IMC (Baseline, CSR ($K = 1$))	6.711	11.216	0.597	27.41	45.315	15.08
IMC (Baseline, CSR (optimal K)	8.076	11.98	0.673	67.065	59.325	21.709
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	7.675	11.932	0.642	65.645	56.736	24.351
NMA (Baseline, CSR ($K = 1$))	5.467	11.166	0.489	64.406	33.182	9.187
NMA (Baseline, CSR (optimal K))	7.539	11.743	0.641	68.66	48.377	21.37
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$	8.12	12.595	0.644	71.805	49.673	26.601

C.1 Results using an expanding window

Table C.6: Equity premium forecast using all predictors (Nov 1953 to Dec 2021) - Expanding window approach

	No constraint		Positive forecast constraint	
	R^2_{OOS}	Δ CER	R^2_{OOS}	Δ CER
Panel A: models with all predictors				
OLS KS	−9.952	−2.528	−2.82	0.498
Ridge	0.54	1.906	0.796	2.076
Lasso	−0.86	−0.857	−0.399	−0.395
Elastic Net	−0.975	−0.906	−0.455	−0.407
Horseshoe	−58.589	−5.95	−15.169	−0.382
PLS	−4.251	−6.309	−3.216	−4.643
CSR (K=1)	0.635	1.699	0.635	1.699
CSR (optimal K)	0.472	2.041	0.728	2.229
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	−0.13	1.549	0.455	1.894
CSR (average all K)	−0.719	0.91	0.243	1.752
Panel B: theoretically-motivated models				
CSP (Polk et al., 2006)	0.165	0.505	0.686	1.447
Campbell (fixed)	−0.293	0.787	−0.293	0.787
SOP (Ferreira and Santa-Clara, 2011)	0.776	2.235	0.848	2.325

The table reports the out-of-sample r-squared (R^2_{OOS}) and the change in the certain equivalent return (Δ CER) for portfolios constructed based on each forecast, relative to portfolios constructed using the historical average benchmark. Models are estimated using an expanding window approach. The initial training is from November 1928 to October 1948 (240 months), the initial validation window is from November 1948 to October 1953 (60 months), and the first out-of-sample forecast is for November 1953. Thereafter, windows are expanded by one month, with the validation period kept at 20% of the combined training and validation periods.

Table C.7: Iterated mean combinations (Expanding window and δ) (Nov 1958 to Dec 2021)

	No constraint			Positive forecast constraint		
	R^2_{OOS}	MSE	Δ CER	R^2_{OOS}	MSE	Δ CER
Panel A: CSR						
CSR ($K = 1$)	0.607	1.802	1.761	0.607	1.802	1.761
CSR (optimal K)	0.412	1.806	2.189	0.685	1.801	2.391
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	−0.207	1.817	1.693	0.415	1.805	2.064
Panel B: CSP (Polk et al., 2006)						
Baseline	−0.594	1.824	0.041	−0.082	1.815	0.989
IMC (Baseline, CSR ($K = 1$))	0.251	1.808	1.512	0.441	1.805	1.714
IMC (Baseline, CSR (optimal K))	0.28	1.808	1.783	0.527	1.803	2.018
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	−0.058	1.814	1.382	0.331	1.807	1.76
NMA (Baseline, CSR ($K = 1$))	0.684	1.801	1.996	0.589	1.802	1.901
NMA (Baseline, CSR (optimal K))	0.741	1.8	2.21	0.7	1.8	2.179
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	0.435	1.805	1.751	0.533	1.803	1.996
Panel C: DP^{drift} (Campbell, 2018), fixed						
Baseline	−0.079	1.814	1.2	−0.079	1.814	1.2
IMC (Baseline, CSR ($K = 1$))	0.535	1.803	1.673	0.535	1.803	1.673
IMC (Baseline, CSR (optimal K))	0.286	1.808	1.949	0.52	1.804	2.174
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	−0.13	1.815	1.487	0.261	1.808	1.795
NMA (Baseline, CSR ($K = 1$))	0.445	1.805	1.781	0.445	1.805	1.781
NMA (Baseline, CSR (optimal K))	0.504	1.804	2.127	0.557	1.803	2.181
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	0.244	1.809	1.799	0.384	1.806	1.988

.7-1!

Table C.3: *(Continued)*

	No constraint			Positive forecast constraint		
	R^2_{OOS}	MSE	Δ CER	R^2_{OOS}	MSE	Δ CER
Panel D: SOP (Ferreira and Santa-Clara, 2011), fixed						
Baseline	0.566	1.803	2.061	0.643	1.801	2.157
IMC (Baseline, CSR ($K = 1$))	−1.312	1.837	−0.166	−0.195	1.817	0.786
IMC (Baseline, CSR (optimal K))	0.565	1.803	2.136	0.632	1.802	2.216
IMC (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	0.422	1.805	1.935	0.555	1.803	2.125
NMA (Baseline, CSR ($K = 1$))	0.747	1.799	2.184	0.749	1.799	2.188
NMA (Baseline, CSR (optimal K))	0.819	1.798	2.529	0.889	1.797	2.603
NMA (Baseline, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	0.556	1.803	2.287	0.737	1.8	2.478

The table reports the out-of-sample r-squared (R^2_{OOS}), the out-of-sample mean squared error (MSE), and the change in the certain equivalent return (Δ CER) for portfolios constructed based on each forecast, relative to portfolios constructed using the historical average benchmark. Panel A reports results for complete subset regressions (CSR). The remaining panels report results for different baseline theoretically-motivated model (i.e. identical to those in Table 1), iterated mean combinations (IMC) and naive model averages (NMA) of the baseline model and different versions of CSR. The IMC models are of the form $\hat{r}_{t+1} = \delta \hat{r}^{Theo}_{t+1} + (1-\delta)\hat{r}^{CSR}_{t+1}$, where $\hat{r}^{Theo}_{t+1}$ is the forecast from a theoretical model of the equity risk premium, $\hat{r}^{CSR}_{t+1}$ is the forecast from a CSR model, and $0 \leq \delta \leq 1$ is estimated via constrained least squares using an expanding window months. The NMA models fix $\delta = 0.5$.

C.2 Results of models à la Lin et al. (2018) and Chen et al. (2022)

Table C.4: Combinations of historical average and complete subset regressions (Nov 1958 to Dec 2021)

	No constraint			Positive forecast constraint		
	R^2_{OOS}	MSE	Δ CER	R^2_{OOS}	MSE	Δ CER
Panel A: CSR						
CSR ($K = 1$)	0.491	1.816	1.515	0.544	1.815	1.592
CSR (optimal K)	1.852	1.791	4.701	2.234	1.784	4.626
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	1.817	1.792	4.31	2.158	1.785	4.605
Panel B: Combinations of historical average and CSR, rolling δ						
IMC ($\overline{r}_{t+1}$, CSR ($K = 1$))	0.47	1.816	1.517	0.524	1.815	1.595
IMC ($\overline{r}_{t+1}$, CSR (optimal K))	1.287	1.801	3.639	1.75	1.793	3.865
IMC ($\overline{r}_{t+1}$, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	1.523	1.797	3.685	1.918	1.79	3.961
Panel C: Combinations of historical average and CSR, expanding δ						
IMC ($\overline{r}_{t+1}$, CSR ($K = 1$))	0.491	1.816	1.515	0.544	1.815	1.592
IMC ($\overline{r}_{t+1}$, CSR (optimal K))	1.807	1.792	4.483	2.27	1.783	4.58
IMC ($\overline{r}_{t+1}$, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	1.883	1.79	4.196	2.176	1.785	4.432
Panel C: Combinations of historical average and CSR, Naive Model Averaging						
NMA ($\overline{r}_{t+1}$, CSR ($K = 1$))	0.281	1.82	0.851	0.281	1.82	0.851
NMA ($\overline{r}_{t+1}$, CSR (optimal K))	1.877	1.79	3.503	1.928	1.79	3.558
NMA ($\overline{r}_{t+1}$, CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	1.82	1.792	3.32	1.85	1.791	3.355
Uni.OLS + NMA ((Chen et al., 2022)	0.351	1.818	0.789	0.392	1.818	0.846

The table reports the out-of-sample r-squared (R^2_{OOS}), the out-of-sample mean squared error (MSE), and the change in the certain equivalent return (Δ CER) for portfolios constructed based on each forecast, relative to portfolios constructed using the historical average benchmark. Panel A reports results for complete subset regressions (CSR). The remaining panels report results for different baseline theoretically-motivated model (i.e. identical to those in Table 1), iterated mean combinations (IMC) and naive model averages (NMA) of the baseline model and different versions of CSR. The IMC models are of the form $\hat{r}_{t+1} = \delta\overline{r}_{t+1} + (1-\delta)\hat{r}^{CSR}_{t+1}$, where $\overline{r}_{t+1}$ is historical average benchmark forecast of the equity risk premium using data until time t, $\hat{r}^{CSR}_{t+1}$ is the forecast from a CSR model, and $0 \leq \delta \leq 1$ is estimated via constrained least squares using an expanding window months. The NMA models fix $\delta = 0.5$.

C.3 Using utility as a criterion to select model dimension

The results for CSR models in the main text relied on the statistical performance (i.e., the MSE) in the validation window to select the optimal value of K or, to a set of values of K that were part of the model confidence set. As discussed in Section 2.1, we also explored corresponding versions of CSR that use utility as a criterion to select values of K. These approaches are referred to as CSR (highest utility) and CSR (average $\widehat{\mathcal{M}}^{\text{U}}_{75\%}$). Table C.5 reports results using these alternative methodologies. The left part of the table reports results for unconstrained forecasts. In terms of R^2_{OOS}, the utility-based CSR models fare worse compared to other CSR approaches: both the CSR (highest utility) and the CSR (average $\widehat{\mathcal{M}}^{\text{U}}_{75\%}$) models deliver negative R^2_{OOS} (-0.81% and -0.01%, respectively). On the other hand, the CSR (highest utility) achieves a ΔCER of 4.45%, the highest value among all the unconstrained models. As discussed previously, this apparent contradiction between statistical and economic performance is indicative of the model's ability to predict market movements better than the historical average, while potentially not being accurate in terms of scale or the variance of the forecasts.

When the positive forecast constraint is imposed, both models deliver positive R^2_{OOS}, although not as high as other versions of CSR. For example, the CSR (average $\widehat{\mathcal{M}}^{\text{U}}_{75\%}$) approach achieves an R^2_{OOS} of 1.31%, which is lower than the R^2_{OOS} of the CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$), 2.03%. Both utility-based CSR models continue to deliver values of ΔCER which are only inferior to that of the CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$). Because of this, our preference continues to be for the latter approach, which seems to offer the best mix of statistical and economic performance.

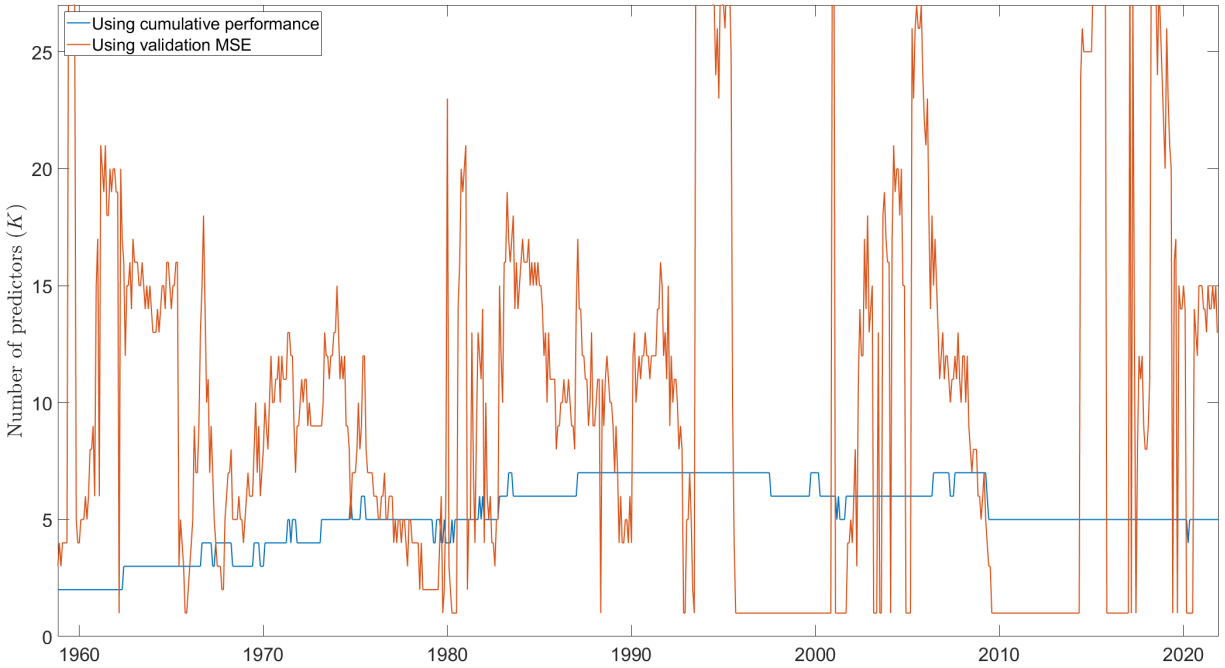
Table C.5: Complete subset regressions using utility criteria vs MSE criterion

	No constraint			Positive forecast constraint		
	R^2_{OOS}	MSE	Δ CER	R^2_{OOS}	MSE	Δ CER
CSR (optimal K)	1.874	1.766	4.534	2.234	1.76	4.461
CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$)	1.829	1.767	4.135	2.148	1.762	4.407
CSR (highest utility)	−0.761	1.814	4.191	0.652	1.788	4.379
CSR (average $\widehat{\mathcal{M}}^{\text{U}}_{75\%}$)	0.909	1.784	3.475	1.707	1.769	4.034

The table reports the out-of-sample r-squared (R^2_{OOS}), the out-of-sample mean squared error (MSE) and corresponding MCS p-values using the Hansen et al. (2011) methodology, the change in the certain equivalent return (Δ CER) for portfolios constructed based on each forecast, relative to portfolios constructed using the historical average benchmark, and the corresponding MCS p-values using a utility-based loss function. CSR denotes complete subset regressions using different approaches to select the optimal model dimension(s). The results based on MSE are the same as in Table 1, and reported here for reader convenience and to facilitate the comparison. CSR (highest utility) uses the certainty equivalent to select the model size. CSR (average $\widehat{\mathcal{M}}^{\text{U}}_{75\%}$) uses the certainty equivalent to determine the models in the confidence set.

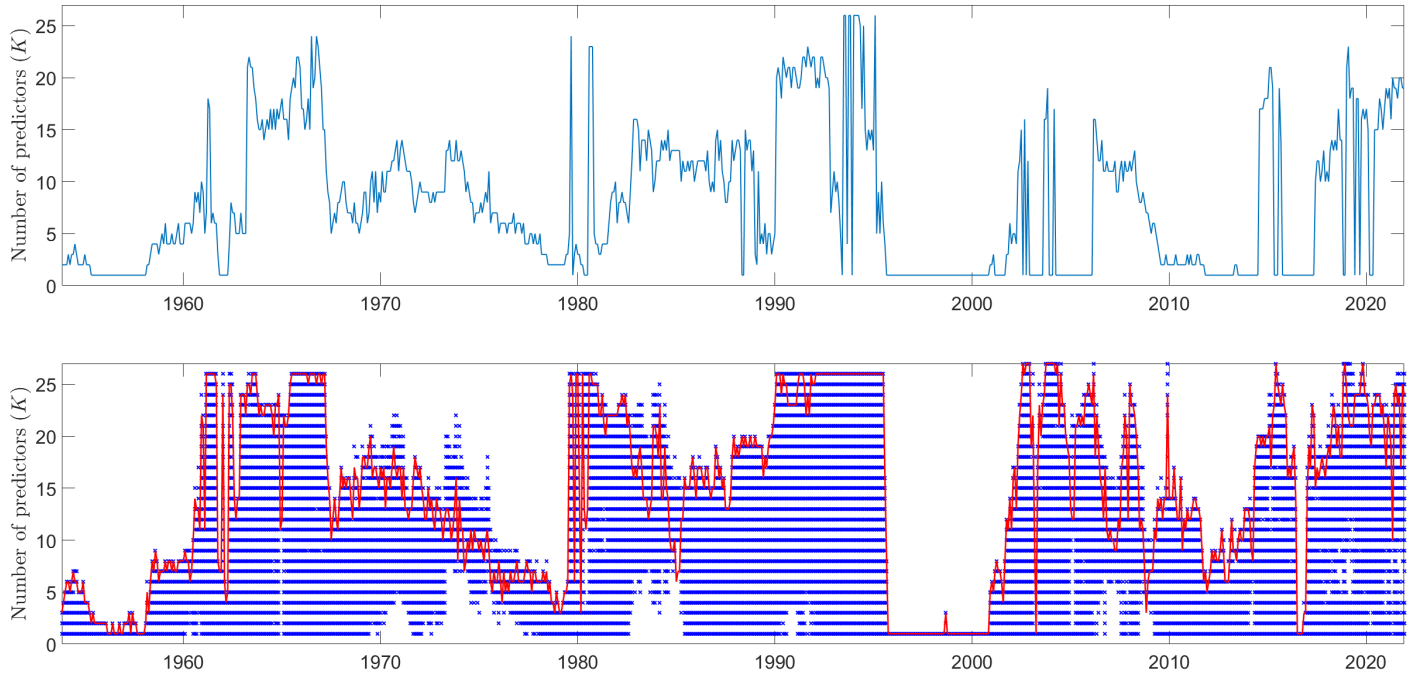
C.4 The dynamics of model uncertainty: Additional Results

Figure C.4: Optimal number of predictors in complete subset regression - comparison with Elliott et al. (2013)



The figure shows the optimal number of predictors selected in the complete subset regression under an expanding window approach using either the validation MSE or the approach in Elliott et al. (2013), which relies on the cumulative out-of-sample performance. At each point in time during the forecasting exercise, all complete subsets from $K = 1$ to $K = 29$ are estimated in the training window. The optimal K using the validation MSE selects, at each month, the value of K with the lowest validation mean squared error. The optimal K in the Elliott et al. (2013) approach relies on the cumulative performance of the complete subset regressions for each value of K, using information up to, but not including, the current month. A starting period of five years of monthly out-of-sample forecasts is used initially, and thereafter the performance is calculate by accumulating the out-of-sample periods.

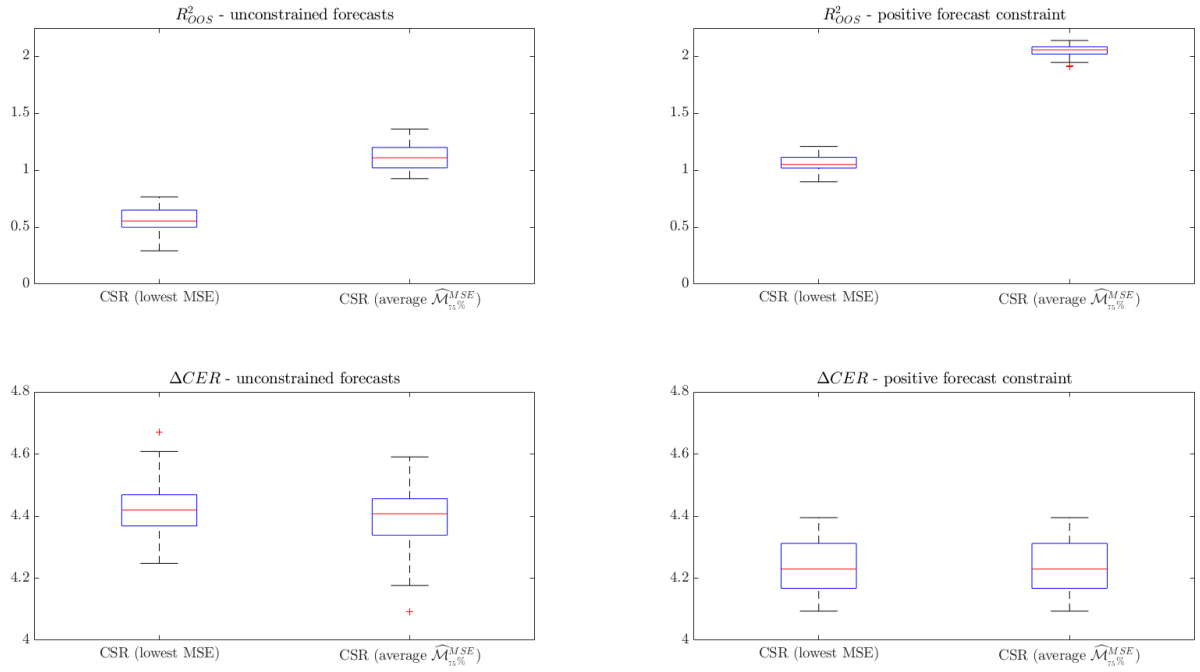
Figure C.5: Optimal number of predictors in complete subset regression - expanding window approach



The top graph shows the optimal number of predictors selected in the complete subset regression (Lowest MSE) model using an expanding window approach. At each point in time during the forecasting exercise, all complete subset regression models are estimated in the training window. Forecasts are made for the validation set, and the optimal value of K is the one with the lowest validation mean squared error. The bottom graph shows which sets of K are included in the CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) approach over time. At each point in time during the forecasting exercise, the model confidence set (MCS) procedure of Hansen et al. (2011) is run to identify a set of models for which the null of equal predictive ability is not rejected. The values of K that are part of the model confidence set at each point are shown as a blue "x". The red line shows how many values of K are part of the MCS at each point in time.

C.5 Results using 20 different replications

Figure C.6: R^2_{OOS} and ΔCER of complete subset regressions in 20 different replications



The figure shows boxplots of the out-of-sample r-squared (R^2_{OOS}, top charts) and the change in the certain equivalent return (Δ CER, bottom charts) relative to portfolios constructed using the historical average benchmark for forecasts obtained using two complete subset regression approaches in 20 replications. The CSR (optimal K) indicates the approach where the optimal value of K is chosen based on the lowest mean squared error in each validation sample. The CSR (average $\widehat{\mathcal{M}}^{\text{MSE}}_{75\%}$) indicates the complete subset regression that averages forecasts of models in the 75% model confidence set.