AP 2014–01

Chair for Applied Stochastics and Risk Management

Helmut Schmidt Universität
Universität der Bundeswehr Hamburg

Faculty of Economics and Social Sciences
Department of Mathematics/Statistics

Working Paper

Forecasting Equity Premia using Bayesian Dynamic Model Averaging

Joscha Beckmann and Rainer Schüssler

May 08, 2014

Working Paper AP 2014–01
May 08, 2014

Forecasting Equity Premia using Bayesian Dynamic Model Averaging

Joscha Beckmann

University of Duisburg-Essen
Campus Essen
Faculty of Economics
Chair for Macroeconomics
UniversitätsstraSSe 12, D-45117 Essen, Germany

E-mail: joscha.beckmann@uni-due.de

Rainer Schüssler

Helmut Schmidt University
Faculty of Economics and Social Sciences
Department of Mathematics/Statistics
Chair for Applied Stochastics and Risk Management
Holstenhofweg 85, D-22043 Hamburg, Germany

URL: www.hsu-hh.de/stochastik
Phone: +49 (0)40 6541-2861
E-mail: schuessr@hsu-hh.de

Working Paper
Please use only the latest version of the manuscript. Distribution is unlimited.

Supervised by: Prof. Dr. Gabriel Frahm
Chair for Applied Stochastics and
Risk Management

URL: www.hsu-hh.de/stochastik

Forecasting Equity Premia using Bayesian Dynamic Model Averaging*

Joscha Beckmann[a] and Rainer Schüssler[b†]

[a]University of Duisburg-Essen and Kiel Institute for the World Economy

[b]Helmut-Schmidt University, Hamburg, and CQE, Münster

April 2014

Abstract

We introduce a Bayesian version of Dynamic Model Averaging for predicting aggregate stock returns. Our approach simultaneously accounts for (i) parameter instability, (ii) time-varying volatility, (iii) model uncertainty and (iv) time-varying model weights. We analyze the predictability of S&P 500 returns and assess which components of forecast models pay off in terms of statistical accuracy and economic value. We document that statistical and economic evaluation metrics can be in sharp contrast. While stochastic volatility emerges as important in terms of density forecast accuracy and economic gains, return prediction models that use economic covariates are helpful only within very limited periods.

JEL: C11, G11

Keywords: Asset allocation; Density forecasting; Model averaging

*We benefited greatly from discussions with Miguel Belmonte, Roberto Casarin, Gabriel Frahm, Gary Koop, Dimitris Korobilis, Francesco Ravazzolo, Christopher Sims and Mark Trede. We also thank conference and seminar participants at: the 4th European Seminar on Bayesian Econometrics (ESOBE), Norges Bank, the 67th European Meeting of the Econometric Society, Gothenburg, the 88th Western Economic Association, Seattle, and the Research Seminar at Strathclyde University 2012, Glasgow, for constructive comments.

†Corresponding author. Tel.: +49 40 6541 2861.

E-mail addresses: joscha.beckmann@uni-due.de (J. Beckmann) and schuessr@hsu-hh.de (R. Schüssler).

1 Introduction

This paper introduces a Bayesian version of Dynamic Model Averaging for predicting aggregate stock returns, accounting for many sources of uncertainty. As the data generating process of the equity premium is expected to be complex and evolving over time, we introduce a highly flexible econometric technique. Precisely, our approach handles (i) parameter instability, (ii) time-varying volatility, (iii) model uncertainty and (iv) time-varying model weights.

Out-of-sample predictability of equity premia is at the core of financial economics and has been the subject of numerous studies. However, empirical evidence of predictability is still controversial. Particularly, it is open to debate whether point forecasts generated by models with economic covariates are consistently superior relative to simple benchmark models such as the prevailing historical mean. Some authors, such as Bossaerts and Hillion (1999) and Welch and Goyal (2008), remain sceptical regarding conditional predictability while other studies report results in favour of conditional predictability. Examples include Ang and Bekaert (2002), Campbell and Thompson (2008), Rapach, Strauss, and Zhou (2010), Ferreira and Santa-Clara (2011), Dangl and Halling (2012) and Neely, Rapach, Tu, and Zhou (2010).[1]

Many caveats have been identified which complicate out-of-sample forecasts. Instability in the relation between stock returns and predictor variables over time is deemed as one of them; see, e.g., Pesaran and Timmermann (1995), Xia (2001), Paye and Timmermann (2006), Welch and Goyal (2008) and Dangl and Halling

[1] A recent survey is provided by Rapach and Zhou (2012).

(2012).[2] Further, predictive regressions for equity premia tend to be too volatile, an issue which has been mitigated by shrinking forecasts toward the historical mean by combining forecasting models; see, e.g., Rapach, Strauss, and Zhou (2010) and Elliott, Gargano, and Timmermann (2013).[3] Another issue is related to the specification uncertainty for models, that is, which combination of predictor variables most accurately summarizes the impact of predictors on equity premia. This source of uncertainty has been commonly addressed in Bayesian Model Averaging (BMA) frameworks; see, e.g., Avramov (2002), Cremers (2002) and Dangl and Halling (2012).

A further unsettled issue is the relationship between statistical and economic metrics of forecast evaluation. Studies documenting predictability agree that even small improvement in statistical accuracy relative to the historical mean can result in sizeable utility gains. Focussing on density rather than point forecasts, Johannes, Korteweg, and Polson (2013) document empirical evidence that simultaneously accomodating for an ensemble of features (such as time-varying expected returns, time-varying volatility and accounting for estimation error) is necessary for out-of-sample portfolio benefits. They conclude that "there is no single "silver bullet" generating out-of-sample gains." This finding suggests the need for a flexible methodology to analyze equity premium predictability, at least in terms of economic utility.

[2] Using a Bayesian framework and accounting for many dimensions of uncertainty, Pettenuzzo and Timmermann (2011) show that structural breaks in the relation between stock returns and predictor variables can crucially influence optimal portfolio allocation.

[3] An alternative shrinkage device in order to limit estimation error of parameters is imposing economic constraints on the coefficients; see Campbell and Thompson (2008) and Pettenuzzo, Timmermann, and Valkanov (2013) for a more involved approach.

We specify a large set of time-varying parameter (TVP) models which differ with regard to included explanatory variables as well as to the specifed dynamics for the evolution of coefficients and volatility. All these choices represent dimensions of model uncertainty, which we address in a Bayesian version of Dynamic Model Averaging. The combination approach generates an aggregate predictive density at each point in time. Despite the multitude of channels entertained in order to enhance model flexibility, our framework retains transparency. That is, the evolution of coefficients for individual models can be tracked over time, as well as individual model weights can be monitored over time. This feature is helpful in disentangling the numerous effects at work for generating the overall forecast.

Our econometric setup introduces a Bayesian version of Dynamic Model Averaging. We label our approach *Bayesian Dynamic Model Averaging (BDMA)*. Raftery, Kárný, and Ettler (2010) developed a strategy to combine models that allows not only parameters but also entire forecast models to change over time. The approach parsimoniously models uncertainty about both coefficients and models using discount factors[4] and has been successfully applied in economic applications; see, e.g. the study about forecasting inflation by Koop and Korobilis (2012). DMA with constant discount factors assumes that certain values are appropriate during all periods. This is not a realistic assumption and certain values are more likely to be only "locally appropriate" (West and Harrison, 1997). We rigorously treat the uncertainty about the involved discount factors within a data-adaptive Bayesian setting. This involves marginalizing out uncertainties about coefficients, the vari-

[4]Discounting/Forgetting approaches are well established in the state space literature; see West and Harrison (1997).

ance and model weights. The key advantage of *BDMA* over DMA is that our setup *allows* parameters, volatility and models to change (gradually or even abruptly) over time rather than *impose* them to change. This added flexibility generates many simpler model configurations as special cases of the most flexible model configuration.[5] With *BDMA* allowing for weighting the recent forecast performance of models more heavily than the forecast performance in the more distant past, model weights may vary over time. In contrast to classical BMA, *(B)DMA* does not invoke the assumption that one of the models captures the true data generating process and considers all of the specified models as potentially misspecified.[6]

Based on the proposed econometric technique, we contribute to the literature on equity premium forecasting by carefully dissecting the relative role of uncertainty regarding regressors, evolution of coefficients, observational variance and the model weights in terms of economic and statistical forecasting gains. For each configuration, we evaluate density forecasts with respect to statistical accuracy and economic value. In our real-time asset allocation exercise, an investor maximizes expected utility using the density forecasts. Analyzing a large set of models (and, hence, different modelling assumptions) enables us to identify general patterns that emerge in order to create superior forecasts.

We consider monthly US equity premia from $1927:01$ to $2012:12$ along with a standard set of twelve explanatory variables and evaluate the forecasts in terms

[5] Among these simpler models are, e.g., equally weighted univariate constant regression models (as proposed by Rapach, Strauss, and Zhou (2010)), classical BMA of time-varying parameter models (as advanced by Dangl and Halling (2012)), linear regression models with constant mean and constant variance (as used by Welch and Goyal (2008)) or the simple historical mean.

[6] This desirable feature is also adopted in other model combination schemes recently proposed; see Hoogerheide, Kleijn, Ravazzolo, Van Dijk, and Verbeek (2010), Geweke and Amisano (2011) and Billio, Casarin, Ravazzolo, and van Dijk (2013).

of statistical and economic criteria. Our results adduce empirical evidence that shrinking and combining forecasts can result in more precise point forecasts relative to the prevailing historical mean. We find that model specifications allowing for stochastic volatility improve density forecast accuracy and increase economic gains, as measured by certainty equivalent returns (CERs) and the Sharpe ratio. Our methodology does not identify any of the included covariates as particularly important for predicting equity premia over a considerably long period of time. With respect to the degree of instability of coefficients, stable and gradually evolving behavior is favoured rather than abruptly changing coefficients. We document disagreement between statistical and economic metrics of forecast performance, that is, point prediction accuracy and economic gains. With density forecasts and economic criteria being more in agreement, exploiting the entire return distribution for asset allocation rather than focus on point predictions pays off. Most importantly, however, while utility gains are generally higher for more flexible methods when evaluated over the entire evaluation period, this result is not robust. The identified gains are largely driven by exceptional and short periods of time, particularly the time period around the Oil Shock ($1973 - 1975$) and the Subprime Crisis (2008/09).

The rest of the paper is organized as follows. Section 2 introduces the predictive regressions. Section 3 lays out the *Bayesian Dynamic Model Averaging* approach. Section 4 reports the empirical results for our analysis of equity premium predictability. Section 5 concludes. Some additional analytical results are shown in greater detail in the Appendix.

2 Time-varying Parameter Models

The model universe in our analysis consists of linear TVP models. Thus the building blocks of the multimodel forecast are of the same type. The specified TVP models differ with regard to the included explanatory variables and the values that control the evolution of (possibly) time-varying coefficients and (possibly) time-varying observational volatility. For ease of presentation, we drop model indices and show the structure of a typical dynamic linear model for $t = 1, ..., T$, consisting of an observation equation (1) and a system equation (2),

$$y_t = F_t^{'}\theta_t + v_t, \qquad v_t \sim N(0, V_t) \tag{1}$$

$$\theta_t = \theta_{t-1} + w_t, \qquad w_t \sim N(0, V_t W_t^*). \tag{2}$$

The TVP model allows for a time-varying linear relationship between the univariate (scalar) variable y_t (in our case: the equity premium) and the vector of the explanatory variables F_t, observed at time $t-1$. $F_t = [1, X_{t-1}]$ is an $m \times 1$ vector of predictors for equity premia, θ_t is an $m \times 1$ vector of coefficients (states). We adopt a strict out-of-sample approach. That is, for predicting y_t, only information at or before time $t-1$ is used. To state precisely on which information set beliefs about parameters are formed, let denote $I_t = [y_t, y_{t-1,}..., y_1, F_t, F_{t-1}, ..., F_1, \text{Priors}_{t=0}]$. This information set contains all realized values of the variable of interest, all realizations of the considered predictive variables as well as the priors for the system coefficients (θ_0) and the observational variance (V_0). As the system equation (2) indicates, the evolution of the system coefficients is assumed to follow a random

walk, with coefficients being exposed to random shocks w_t.[7]

Adopting a (conditionally) normally distributed prior for the system coefficients and an inverse-gamma distributed prior for the observational variance results in a fully conjugate Bayesian analysis, ensuring that prior and posterior distribution come from the same family of distributions. The conjugate specification at some arbitrary time t can be expressed as

$$V_t|I_t \sim IG\left[\frac{n_t}{2}, \frac{n_t S_t}{2}\right], \tag{3}$$

$$\theta_t|I_t \sim t_{n_t}\left[m_t, S_t C_t^*\right], \tag{4}$$

$$\theta_t|I_t, V_t \sim N\left[m_t, V_t C_t^*\right]. \tag{5}$$

S_t is a point estimate for the observational variance V_t. n_t denotes the degrees of freedom for the (unconditionally on V_t) t-distributed coefficients. The point estimate for the coefficient vector is m_t with scale matrix $C_t = S_t C_t^*$. The forecast of y_t (i.e., the predictive density) is obtained by integrating out the uncertainty in the states θ_t and the volatility V_t, rendering a t-distributed predictive density. In A.1, we will describe in detail, how, at some arbitrary time t, beliefs are formed for the variable of interest and how new observations lead to an update for the estimated system coefficients, their associated scale matrix and for the estimate of the observational variance.

[7] All variances and covariances in the dynamic linear model are scaled by the unknown observational variance V_t. Unscaled (co-)variances are indcated by asterisks, e.g., in the case of the system variance, $W_t = V_t W_t^*$. For this aspect as wells as for the description of TVP models in general, our adopted notation is based on West and Harrison (1997). For a discussion about the random walk assumption in TVP models for coefficients in the context of equity premia, see Dangl and Halling (2012).

We adopt a discount factor approach for modelling the unknown sequences for V_t and W_t. For the latter, consider the transition from the posterior time-$t-1$ estimate for the scale matrix of coefficients (C_{t-1}) to the time-t prior for the scale matrix of coefficients (R_t),

$$R_t = C_{t-1} + W_t. \tag{6}$$

To accomodate the additional uncertainty involved in the estimate for the coefficients proceeding from time $t-1$ to time t, C_{t-1} is inflated by the system variance W_t. Instead of estimating W_t, our adopted discount approach involves replacing W_t by

$$W_t = \frac{1-\delta}{\delta} C_{t-1}, 0 < \delta \leq 1, \tag{7}$$

and, hence,

$$R_t = \frac{1}{\delta} C_{t-1}. \tag{8}$$

δ is a discount factor providing that observations s periods in the past have weight δ^s. This implies an age-weighted estimation with an effective window size of $(1-\delta)^{-1}$; see Hannan, McDougall, and Poskitt (1989). For $\delta = 1$, the case of constant parameters is included, $\delta < 1$ explicitly allows for variability in the system coefficients. Values of δ near 1 are consistent with gradual parameter evolution, whereas low values of δ allow for abrupt parameter changes. In our empirical application, we will consider a grid of values for $\delta \in \{\delta_1, ..., \delta_d\}$ to allow for different degrees of parameter instability. Notice, however, that δ is fixed

within each individual model. The data support for different degrees of parameter instability is hence displayed at the level of the multimodel forecast, reflecting the data support across models with different values of δ at each point in time.

In a similar fashion as for W_t, we adopt a discount approach for the evolution of the observational variance, V_t. Since the assumption of a constant observational variance is unappealing in the context of financial applications, our econometric technique allows for stochastic volatility. Imposing a decay factor β, $0 < \beta \leq 1$, the degree of adaptiveness to new data is controlled. Updating the (inverse-gamma) posterior distribution of V_t involves updating the degrees of freedom

$$n_t = \beta n_{t-1} + 1 \tag{9}$$

and the point estimate

$$S_t = S_{t-1} + \frac{S_{t-1}}{n_t}\left(\frac{e_t^2}{Q_t} - 1\right), \tag{10}$$

see (3). e_t denotes the prediction error of a model and Q_t the scale associated with the t-distributed forecast y_t, see (30) in A.1. Note from (9) that, for $\beta = 1$, $n_t \to \infty$ for increasing t. It is readily seen from (10) that this results in $S_t = S$, and, hence, the case of constant variance is recovered for $\beta = 1$. For $\beta < 1$, n_t converges to the constant, limiting degrees of freedom, $n_t \to (1-\beta)^{-1}$, implying a limit to the accuracy with which the variance at any time is estimated. (10) shows, that if the prediction error e_t of a model coincides with its expectation Q_t (i.e., $e_t^2 = Q_t$), $S_t = S_{t-1}$. Prediction errors above the expected error lead to an increase in the estimated observational variance and vice versa.

In the case of stochastic volatility ($\beta < 1$), the estimate for the observational variance is updated according to new data, discounting past information to reflect changes in volatility, with the updated posterior distribution being more heavily weighted on the new observation than in the case of constant variance. The representation

$$S_t = (1-\beta)\sum_{s=0}^{t-1}\beta^s\left(\frac{e_{t-s}^2 S_{t-s-1}}{Q_{t-s}}\right) \tag{11}$$

for the point estimate S_t has the form of an exponentially weighted moving average of the standardised forecast errors. Thus, the estimate of the variance continues to adapt to new data, while older data are further discounted as time progresses. We consider a grid of values $\beta \in \{\beta_1, ..., \beta_b\}, 0 < \beta \leq 1$. b indicates the discrete number of grid points considered. Just as δ, β is fixed within each individual model.

We denote a typical model in our universe by $M_j, j = 1, ..., J$. Each model is defined by its set of considered regressors, the presumed variability in the coefficients (governed by the discount factor δ) and the dynamics of the volatility (characterized by the decay factor β). With a set of K explanatory variables (in addition to the intercept[8]), b grid points for β and d grid points for δ, $J = 2^K \cdot b \cdot d$ models are available at each point in time. Empirical evidence for particular model configurations (i.e., for certain values of δ, β and subsets of explanatory variables from the K candidates) is uncovered at each point in time through their data support (i.e., the attached model weight for particular model configurations). In the next step we will address the issue of combining the individual models.

[8] All models in our universe include an intercept.

3 Bayesian Dynamic Model Averaging

The large set of models at disposal raises the issue of how to optimally combine them. We propose a flexible weighting scheme which nests BMA and equal model weighting as special cases. The approach draws on insights from DMA proposed by Raftery, Kárný, and Ettler (2010). DMA allows for exponential discounting in the weight dynamics according to the past forecast performance of the individual models, thus allowing recent data to be emphasized.[9] DMA involves specifying a discount factor to control down-weighting of older data. We generalize Raftery's implementation of DMA by addressing the uncertainty about the discount factor, calculating it in a data-adaptive fashion.

Let denote $p\left(M_i|I_{t-1}\right)$ the updated model weight for model i at time $t-1$. $\mathcal{P}\left(M_i|I_{t-1}\right)$ indicates the *prediction* weight for model i at time $t-1$ (or stated differently: the prior weight for time t). α is a discount factor, $0 \leq \alpha \leq 1$, and shrinks the posterior model weights toward equal weights,

$$\mathcal{P}\left(M_i|I_{t-1}\right) = \frac{p\left(M_i|I_{t-1}\right)^{\alpha}}{\sum_{j=1}^{J} p\left(M_j|I_{t-1}\right)^{\alpha}}. \tag{12}$$

Updating model weights is accomplished by using Bayes' rule,

$$p\left(M_i|I_t\right) = \frac{p\left(y_t|M_i, I_{t-1}\right)\mathcal{P}\left(M_i|I_{t-1}\right)}{\sum_{j=1}^{J} p\left(y_t|M_j, I_{t-1}\right)\mathcal{P}\left(M_j|I_{t-1}\right)}. \tag{13}$$

[9] Emphasizing recent data when combining models is also well known in the literature about point forecasting; see, e.g., Stock and Watson (2004).

Obviously, for $\alpha = 0$, all models are equally weighted,[10] while for $\alpha = 1$, there is no discounting and, hence, BMA is recovered as a special case. The connection between predictive and marginal likelihoods (and, thus, between DMA and classical BMA) is shown in A.2). BMA attaches equal weights to all data from $s = 1, ..., t$ and, as t gets larger, posterior model probabilities will typically change only slightly as new data points are added. Allowing for $\alpha < 1$ increases flexibility as model weights may change more rapidly.

Using Raftery's version of DMA with a discount factor α, the *predictive* weight attached to model i is

$$\begin{aligned} \mathcal{P}\left(M_i|I_{t-1}\right) &\propto \left[\mathcal{P}\left(M_i|I_{t-2}\right)p\left(y_{t-2}|M_{i,}I_{t-2}\right)\right]^{\alpha} \\ &= \prod_{s=1}^{t-1} p\left(y_s|M_i, I_{t-s-1}\right)^{\alpha^s}. \end{aligned} \tag{14}$$

Thus, model i will be attached more weight if it has provided accurate forecasts in terms of predictive likelihoods in the (recent) past compared to its peers. The discount factor α controls the exponential discounting of likelihoods according to their recency.

As, however, a certain value of α might only be locally appropriate, we let α evolve over time and integrate out the associated uncertainty. Initializing the process of model combinations involves specifying priors on model weights, $p\left(M_i|I_0\right)$, $\forall i = 1, ..., J$. To obtain predictive weights, we use an equation similar

[10] It is well-known in the forecasting literature that equal model weighting is a tough benchmark; see, e.g., Geweke and Amisano (2012).

to (12), but in contrast to (12), we sum over the discrete set of considered grid points for α.

$$\mathcal{P}\left(M_i|I_{t-1}\right) = \sum_{v=1}^{a} \frac{\overbrace{p\left(M_i|I_{t-1}\right)^{\alpha_v}}^{:=\mathcal{P}(M_i|I_{t-1},\alpha_v)}}{\sum_{j=1}^{J} p\left(M_j|I_{t-1}\right)^{\alpha_v}} \cdot p\left(\alpha_v|I_{t-1}\right). \tag{15}$$

$p\left(M_i|I_{t-1}\right)$ refers to the time $t-1$ posterior model weights. We consider values on the grid $\alpha_v \in \{\alpha_1, \alpha_2, .., \alpha_a\}$, where $0 \leq \alpha_v \leq 1$ and a denotes the number of grid points. The updating step for model weights is accomplished by

$$p\left(M_i|I_t\right) = \sum_{v=1}^{a} \frac{p\left(y_t|M_i, I_{t-1}\right)\mathcal{P}\left(M_i|I_{t-1}, \alpha_v\right)}{\sum_{j=1}^{J} p\left(y_t|M_j, I_{t-1}\right)\mathcal{P}\left(M_j|I_{t-1}, \alpha_v\right)} \cdot p\left(\alpha_v|I_t\right), \tag{16}$$

where the predictive likelihood of model i,

$$p\left(y_t|M_i, I_{t-1}\right) \sim \frac{1}{\sqrt{Q_{t,i}}} t_{\beta n_{t-1},i}\left(\frac{y_t - \widehat{y}_{t,i}}{\sqrt{Q_{t,i}}}\right), \tag{17}$$

is used to a assess the forecast performance for model i and is obtained by evaluating the predictive density at the actual value y_t. $\widehat{y}_{t,i}$, $Q_{t,i}$ and $\beta n_{t-1,i}$ denote the point estimate, the scale and the degrees of freedom of the predictive density for a particular model i, respectively. High values of the predictive likelihoods are associated with good forecast performance.

The time$-t$ posterior of a particular grid point for the discount factor α is obtained as

$$p\left(\alpha_z|I_t\right) = \frac{\sum_{j=1}^{J} p\left(y_t|M_j, I_{t-1}\right) \mathcal{P}\left(M_j|I_{t-1}, \alpha_z\right) p\left(\alpha_z|I_{t-1}\right)}{\sum_{v=1}^{a}\sum_{j=1}^{J} p\left(y_t|M_j, I_{t-1}\right) \mathcal{P}\left(M_j|I_{t-1}, \alpha_v\right) p\left(\alpha_v|I_{t-1}\right)}, \forall z = 1, .., a, \quad (18)$$

where $\sum_{j=1}^{J} p\left(y_t|M_j, I_{t-1}\right) \mathcal{P}\left(M_j|I_{t-1}, \alpha_z\right)$ is the predictive likelihood of the multi-model involving all J considered models with weights governed by the particular value α_z.

There are at least two motivating aspects for the use of likelihood discounting. First, it is reasonable to think that more recent data will provide more relevant information for predicting, since recent data are in many situations more likely to occur in a similar (economic) environment. Second, the discounting approach with its provided shrinkage toward equal weights can prevent attaching the entire weight to one particaular model, as it is (asymptotically) the case for standard BMA which cumulates the unweighted likelihoods. In a stable environment, high values for α are expected to be supported by the data, while in unstable periods low values for α are likely to be favored, reflecting the need for changes in model weights. When focussing on a particular variable (or combination of variables), that is, set aside specification uncertainty, the combination of (possibly) time-varying coefficients ($\delta < 1$) and (possibly) time-varying model weights ($\alpha < 1$) amounts to a version of averaging across estimation windows as analyzed by Pesaran and Pick (2011).

Point forecasts for the overall forecast model are obtained as

$$\widehat{y}_t|I_{t-1} = \sum_{j=1}^{J} \left(F_{t,j} m_{t-1,j}\right) \mathcal{P}\left(M_j|I_{t-1}\right). \tag{19}$$

Note that *BDMA* represents a shrinkage device for (slope) coefficients. Models which do not include a subset of particular regressors implicitly set the associated coefficients to zero, thereby shrinking those coefficients in the overall forecast model toward zero. In our setup, the model which considers all predictors to be unnecessary, is nested. If the entire weight is attached to this particular model, the overall forecast model collapses to the historical mean. Having layed out the econometric setup, we next turn to our empirical analysis.

4 Empirical Analysis

4.1 Data

We use the proposed methodology to forecast (simple) excess returns of the S&P 500 index covering the period from 1927 : 01 to 2012 : 12 at the monthly horizon.[11] We draw on a standard set of explanatory variables, previously employed in the study by Welch and Goyal (2008).[12] For the sake of brevity, we include only a list of the predictive variables here and refer to Welch and Goyal (2008) for a detailed

[11] Our choice for this time period is driven by data availability. Choosing such a long period, we want to mitigate concerns of sample selection bias for our empirical results.

[12] The dataset is provided by Amit Goyal: (`http://www.hec.unil.ch/agoyal/`). Of course, further variables could be added. For example, Boudoukh, Michaely, Richardson, and Roberts (2007) propose alternative measures of payout yield rather than the classical dividend yield, while Neely, Rapach, Tu, and Zhou (2010) employ technical indicators. If, then, the set of condsidered explanatory variables (say, $k > 15$) becomes very large, it is no longer possible to evaluate all models. However, in this case, we could employ Markov Chain Monte Carlo Composition (Madigan, York, and Allard, 1995; Raftery, Madigan, and Hoeting, 1997) or stochastic search algorithms (e.g. the stochastic shotgun search algorithm proposed by Hans, Dobra, and West (2007)) to explore the model space and calculate model probabilities.

discussion of the dataset and the data sources.

- Log dividend yield (*dy*): difference between the log of dividends on the S&P 500 index and the log of one-month-lagged prices.
- Earnings-to-price ratio (*ep*): difference between the log of earnings and the log of stock prices.
- Dividend-payout ratio (*dpayr*): difference between the log of dividends and the log of earnings.
- Stock variance (*svar*): sum of squared daily returns.
- Book-to-market ratio (*bmr*): book to market ratio value for the Dow Jones Industrial Average.
- Net issuing activity (*ntis*): ratio of twelve-month moving sums of net issues by NYSE listed stocks to the total market capitalization of NYSE stocks.
- T-bill rate (*tbl*): interest rate on a three-month Treasury bill (secondary market).
- Long-term yield (*lty*): long-term government bond yield.
- Long-term return (*ltr*): return on long-term government bonds.
- Default return spread (*dfr*): long-term corporate bond return minus the long-term government bond return.
- Default yield spread (*dfy*): difference between BAA- and AAA-rated corporate bond yields.

- Inflation (*inf*): Consumer Price Index (all urban consumers) from the Bureau of Labor Statistics, lagged by one additional month.

Taking into account these listed predictor variables, our set of potential regressors comprises $K = 12$ variables. We set aside a period for initializing the estimation and report results for the evaluation period from $1947:01$ to $2012:12$.

4.2 Prior Choices

4.2.1 TVP Models

To initialize the sequential prediction and updating of the TVP models, we have to choose a (normally/inverse-gamma) prior distribution for the coefficients and the observational variance, that is $V_0|I_0 \sim IG\left[\frac{n_0}{2}, \frac{n_0 S_0}{2}\right]$ and $\theta_0|I_0, V_0 \sim N\left[m_0, C_0\right]$. We use the empirical variance of the index returns from the "burn-in" period from $1927:01$ to $1946:12$ to determine S_0 and choose $n_0 = 5$ to to express our initial uncertainty about the observational variance. We set $m_0^{(j)} = 0_{n_j \times 1}$, $C_0^{(j)} = g \cdot I_{n_j}$ with $g = 10^{13}$ for all models $j = 1, ..., J$, where n_j denotes the number of variables in model $j = 1, ..., J$. Thus we center the initial values for the system coefficients around zero, surrounded by a high degree of uncertainty. This diffuse prior thus allows for data patterns to be quickly adapted at the beginning of the estimation.

Specifying the range for the grid of values of δ and β, we define the range

[13] Alternative values for g such as $g = 1$ or $g = 100$ do not affect our results as the effect of the prior variance quickly disappears. Hence, the reported numbers in our application are robust with respect to the prior specification. If we, however, choose a very low value for g and thus a very intense shrinkage for the coefficients toward zero, we would prevent the models from learning. In a multimodel setting with small models already nested in the setup, such tight priors would be pointless.

which is covered when summing over the degree of variability in θ and V. We choose $\delta \in \{0.95; 0.99; 1\}$ and $\beta \in \{0.80; 0.90; 1\}$.[14] Choosing the upper bounds is motivated by the purpose to recover the special case of constant parameters and constant observational variance. $\delta = 0.99$ allows for gradual evolution of the coefficients, while $\delta = 0.95$ models abrupt changes in coefficients. In the latter case, the evolution would be highly unstable. In our view, setting the lower bound to 0.95 presents a compromise between allowing for sharp changes in the evolution of the coefficients and limiting the possibility for extremely erraneous behaviour of the coefficients. $\beta \in \{0.8; 0.9; 1\}$ covers a broad range from high variation in volatility ($\beta = 0.8$) to constant volatility ($\beta = 1$).

4.2.2 Model Combination

We initially assign equal weights to each possible model configuration, that is, $p(M_j|I_0) = \frac{1}{b \cdot d \cdot 2^K}, \forall j = 1, ..., J$. Thus, initially, all models are equally likely. On the level of model combination, we have to choose the range of α. We set $\alpha \in \{0; 0.80; 0.90; 0.95; 0.99; 1\}$ and, hence, cover the range from classical BMA ($\alpha = 1$) to equal weights ($\alpha = 0$). We assign equal initial weights for each considered grid point for α, i.e., $p(\alpha_z|I_0) = \frac{1}{a}, \forall z = 1, ..., a$.[15]

[14] Choosing beta-distributed priors for β and δ could be regarded as a more natural choice. However, in this case, we would have to give up conjugacy of our analysis, substantially increasing the computational burden for estimating the models.

[15] As a robustness check, we initially favor equal weighting, that is $\alpha = 0$. We assign $p(\alpha = 0|I_0) = 0.8$ and distribute the remaining weight equally among the remaining values for α. In our analysis, the impact of the prior rapidly decays with data support for $\alpha = 0$ being very similar to the standard case for initially equal weighted values for α. This pattern holds true if we favor other values for α at the beginning.

4.3 Forecast Models

We present results for a multitude of forecast models, nested as special cases in our preferred and most general implementation of the proposed method (*BDMA*). The benchmark models arise by imposing (one or more) constraints on the regressors (k) and the discount factors (δ, β, α) governing the evolution of coefficients, the observational variance and the model weights. This analysis enables us to empirically assess the relative importance of the different dimensions of uncertainty. To this end, we focus on model configurations which have either been explicitly proposed in previous studies or contribute to disentangling various effects (e.g., model size, importance of particular predictors, constant vs time-varying coefficients/variance, dynamics of combination weights). The set of models comprise the following configurations (the mnemonics label the models in the tables) and is summarized in Table 1:

- *BDMA:* Forecasts using *Bayesian Dynamic Model Averaging* without restrictions. $\delta \in \{0.95; 0.99; 1\}$, $\beta \in \{0.8; 0.9; 1\}$ and $\alpha \in \{0; 0.8; 0.9; 0.95; 0.99; 1\}$.

- *BDMS*: Forecasts using *Bayesian Dynamic Model Selection*. This involves assigning the entire model weight to the model with the currently highest predictive weight at each point in time.

- *CC-CV-BMA*: Forecasting with the restrictions of constant coefficients (CC), constant observational variance (CV) and standard Bayesian Model Averaging (BMA). Technically, this involves setting $\delta = 1$, $\beta = 1$, $\alpha = 1$.

- *CV*: Forecasts with the restriction of constant observational variance ($\beta = 1$).

- *CC*: Forecasts with the restriction of constant coefficients ($\delta = 1$).

- *CV-BMA*: Forecasts with constant observational variance and using standard BMA, $\beta = 1$ and $\alpha = 1$. This econometric setup is employed in the study of Dangl and Halling (2012).

- *CC-EW*: Forecasts with constant coefficients and equal model weights, $\delta = 1$ and $\alpha = 0$.

- *EW*: Forecasts with equally weighted models, $\alpha = 0$.

- *BMA*: Forecasts using standard BMA ($\alpha = 1$).

- *Kitchen-Sink Models*[16]: Forecasts with all predictors included ($k = K$). *Kitchen-Sink-CC-CV Models* further assume both constant coefficients and constant variance.

- *Large Models*: Forecasts with at least 9 predictors ($k \geq 9$). *Large Models CC-CV* further assume both constant coefficients and constant variance.

- *Medium Models*: Forecasts restricted to models with 5 to 8 predictors ($5 \leq k \leq 8$). *Medium Models CC-CV* further assume both constant coefficients and constant variance.

- *Small Models*: Forecasts with a maximum of 4 predictors ($k \leq 4$). *Medium Models CC-CV* further assume both constant coefficients and constant variance.

[16] This model specification is known as "kitchen-sink" regression because it throws "everything but the kitchen sink" into the regression.

- *Univariate Models*: Forecasts with only one predictor ($k = 1$). The configurations *Univariate Models-CC-CV* further assume both constant coefficients and constant variance, while the configurations *Univariate Models-CC-CV-EW* in addition assign equal weights. This econometric setup is used in the study of Rapach, Strauss, and Zhou (2010).

- *Historical Mean-CC*: Forecasts without additional regressors ($k = 0$ and $\delta = 1$). Point forecasts are in this case identical to the unconditional prevailing mean. β is not restricted, allowing for time-varying volatility.[17]

4.4 Model Characteristics

To assess which configurations are supported by the data, we present some model characteristics for our preferred and most general aggregate forecast model, the *BDMA* configuration. Figure 1 presents point predictions along with credibility intervals of the equity premium for the entire evaluation period. The shrinkage provided by the combination of models keeps the point forecasts relatively stable. The width of the credibility intervals substantially varies over time, a manifestation of the data support for stochastic volatility, as documented in Figure 2. While the data support for $\beta = 1$, corresponding to the constant volatility model, rapidly converges to zero, $\beta = 0.9$ is favored by the data most of the time, accompanied by occasional spikes for $\beta = 0.80$ (e.g., in the Subprime Crisis).[18] Altogether, a

[17] The historical mean is the most common benchmark for evaluating forecasting models for the equity premium. We allow for stochastic volatility in this configuration. Most studies also incorporate time-varying volatility, however, in ad-hoc approaches such as rolling windows; see, e.g., Campbell and Thompson (2008) and Neely, Rapach, Tu, and Zhou (2010).

[18] In these periods, particularly the error variances of the homoscedastic model versions are too small. As a consequence, point forecasts of the homoscedastic model versions are far in the

Table 1: Forecast Models.
The table summarizes the forecast models with their imposed restrictions on regressors, coefficients, variance and the model weighting scheme. $(-)$ indicates that no restrictions are imposed.

Model configuration	Regressors	Coefficients	Variance	Model weights
BDMA	$-$	$-$	$-$	$-$
BDMS	$-$	$-$	$-$	*single model*
CC-CV-BMA	$-$	$\delta = 1$	$\beta = 1$	$\alpha = 1$
CV	$-$	$-$	$\beta = 1$	$-$
CV-BMA	$-$	$-$	$\beta = 1$	$\alpha = 1$
CC	$-$	$\delta = 1$	$-$	$-$
CC-EW	$-$	$\delta = 1$	$-$	$\alpha = 0$
EW	$-$	$-$	$-$	$\alpha = 0$
BMA	$-$	$-$	$-$	$\alpha = 1$
Kitchen-Sink	$k = 12$	$-$	$-$	$-$
Kitchen-Sink-CC-CV	$k = 12$	$\delta = 1$	$\beta = 1$	$-$
Large Models	$k \geq 9$	$-$	$-$	$-$
Large Models CC-CV	$k \geq 9$	$\delta = 1$	$\beta = 1$	$-$
Medium Models	$5 \leq k \leq 8$	$-$	$-$	$-$
Medium Models-CC-CV	$5 \leq k \leq 8$	$\delta = 1$	$\beta = 1$	$-$
Small Models	$k \leq 4$	$-$	$-$	$-$
Small Models-CC-CV	$k \leq 4$	$\delta = 1$	$\beta = 1$	$-$
Univariate Models	$k = 1$	$-$	$-$	$-$
Univariate Models-CC-CV	$k = 1$	$\delta = 1$	$\beta = 1$	$-$
Univariate Models-EW	$k = 1$	$-$	$-$	$\alpha = 0$
Univariate Models-CC-CV-EW	$k = 1$	$\delta = 1$	$\beta = 1$	$\alpha = 0$
Historical Mean-CC	$k = 0$	$\delta = 1$	$-$	$-$

high degree of data adaptiveness for the variance is selected for the overall model.

With respect to the empirical evidence for time-varying coefficients, Figure 3 documents that the data support for abruptly changing coefficients ($\delta = 0.95$) quickly converges to zero. Gradually evolving coefficients (corresponding to $\delta = 0.99$) and constant coefficients ($\delta = 1$) both receive data support, with $\delta = 0.99$ being favored until the mid-90s, when this pattern is reversed. During the Subprime Crisis, the importance of time-varying coefficients increases again.

With respect to model combinations, Figure 4 shows empirical evidence for the favored degree of likelihood discounting over time. Though entire forecasting models are allowed to rapidly change over time (corresponding to low values for α), $\alpha = 0.99$ and $\alpha = 1$ are attached the highest weights. Data support for other values of α converges to zero. While $\alpha = 0.99$ increasingly gains support until the mid-90s, $\alpha = 1$ dominates afterwards until the Subprime Crisis. The pattern of the values for $\alpha = 1$ and $\alpha = 0.99$ are to a certain extent developing in line with the $\delta = 1$ and $\delta = 0.99$. Between the mid-90s and the Subprime Crisis, small models with stable coefficients are favored, along with BMA weighting. Inspecting Figure 1, this pattern might be traced back to a learning process during an increasingly stable environment.

Figure 5, presenting the inclusion probabilities for each regressor, shows that none of the considered predictors emerges as particularly important over the entire evaluation period. Generally, the inclusion probabilities for the predictors mildly

tails of the predictive distributions, resulting in small predictive likelihoods.

fluctuate around 0.5 until the mid-90s. Thereafter, most variable lose data support, while the predictor *net issuing activity (ntis)* gains importance.

Altogether, the figures for the data support of the model configurations indicate that the support for certain configurations are able to change rapidly. Figures 2, 3 and 4 demonstrate that the support for the discount factors for β, δ and α may quickly adapt to new data. Also, inclusion probabilities for particular regressors abruptly change in some cases (see Figure 5). Why, then, no predictor emerges as important for a considerably long period of time? Is it, because there are no useful predictors for a longer time, or is it because the econometric setup is unable to identify them? From the outset of our model configuration (at least the most flexible version), we know, that if the correct model is nested, it would ultimatively be detected.[19] Of course, this is an asymptotic result and we do not know how long it would take to detect it.[20] However, it is unrealistic that any of the nested models is literally true. Hence, if we consider all individual forecast models to be misspecified, the task of the flexible model averaging process is to detect locally appropriate models, that is locally suitable approximations to the data generating process, and rapidly increase the weights corresponding

[19] BMA is nested in our approach. If the correct model is included among the considered model universe, it will eventually be attached the entire weight, since the marginal likelihood of the correct model will increasingly dominate the marginal likelihood of all remaining models. If the correct model was indeed among the specified configurations, α would converge to 1, as it, for this case, would be of no use to discount likelihoods and the weighting scheme would collapse to classical BMA.

[20] Raftery, Kárný, and Ettler (2010) conduct a simulation study in order to assess whether DMA is able to track both changing parameters and models for the case of the right model being included in the considered model universe. They demonstrate that DMA quickly adapts to (even abrupt) parameter changes and changes of the entire model. Since our Bayesian version of DMA even increases model flexibility, our econometric technique is supposed to have power to rapidly detect changes in parameters and models.

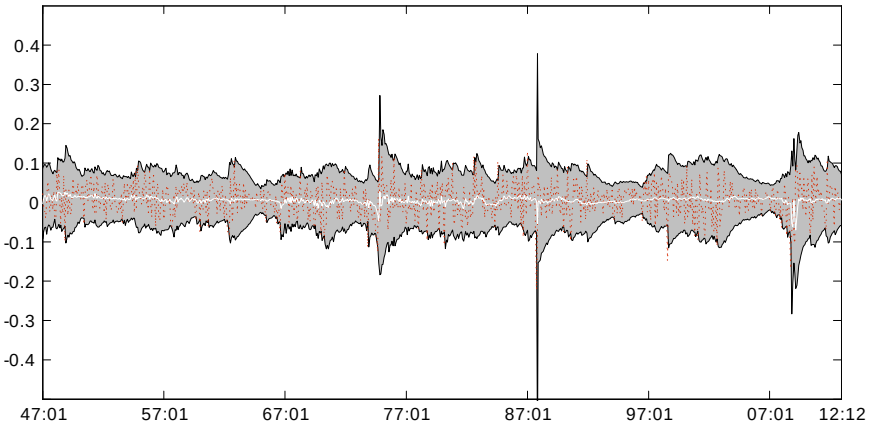

Figure 1: Predictive densities for the S&P 500 returns. The figure presents the $(5\%, 95\%)$ credibility intervals for the predictive returns (grey shaded area). Point forecasts are indicated by the solid white line, realized returns are displayed by the dotted red line.

to the appropriate configurations. Although our approach offers a great amount of flexibility for adapting to changes in the data generating process, we do not identify any regressors to be important for a considerably long period of time. This points to a scenario in which either no regressors are relevant over a considerably long period of time, or the marginal impact of regressors changes in an erratic, unpredictable fashion.

4.5 Forecast Evaluation

4.5.1 Statistical Evaluation

As a measure of density forecast accuracy, we assess our models in terms of predictive log likelihoods (SumPL), involving the entire predictive distribution. We also report the out-of-sample R^2 (Campbell and Thompson, 2008), denoted by R^2_{OOS}. The R^2_{OOS} measures the proportional reduction in mean squared prediction error

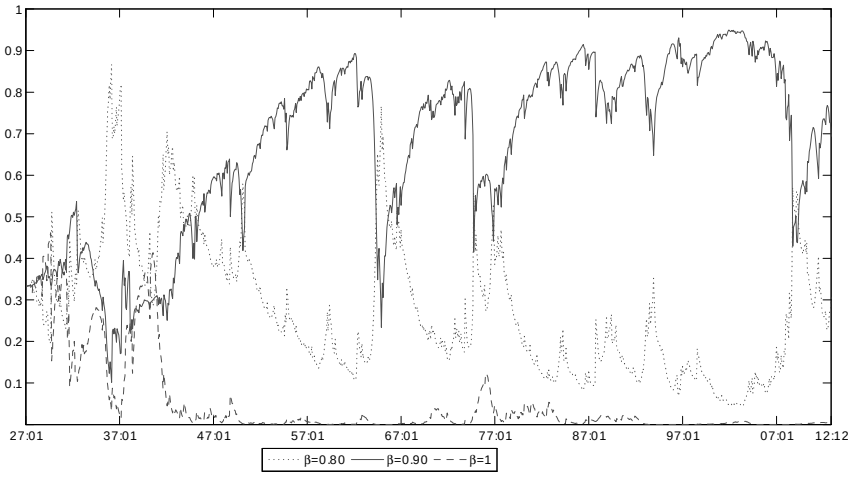

Figure 2: Data support for different values of β over time. The figure shows the inclusion probabilities for the considered grid points.

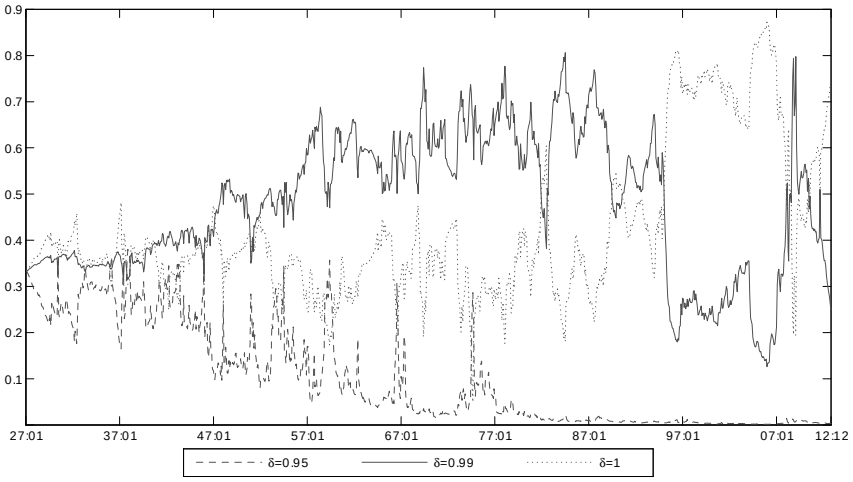

Figure 3: Data support for different values of δ over time. The figure shows the inclusion probabilities for the considered grid points.

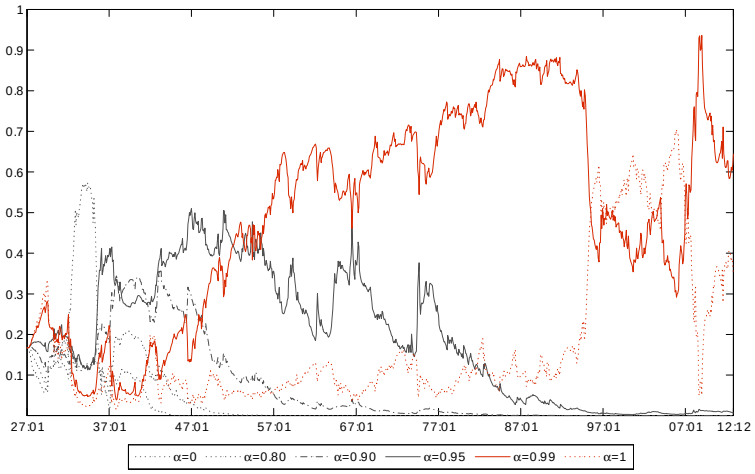

Figure 4: Data support for different values of α over time. The figure shows the inclusion probabilities for the considered grid points.

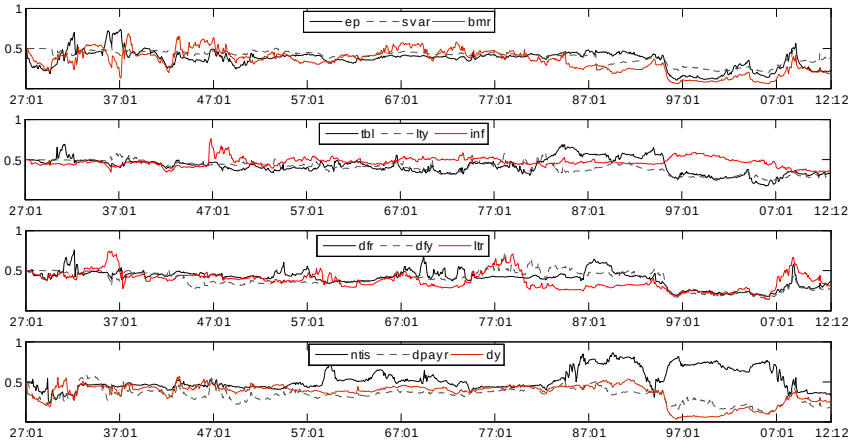

Figure 5: Inclusion probabilities for the twelve predictors over time.

(MSPE) for an arbitrary forecast model i relative to the historical average.[21] When assessing statistical significance, we have to take into account that all entertained model configurations nest the historical mean. We therefore employ the Clark and West (2007) MSPE-adjusted statistic. This involves testing the null hypothesis that the historical average MSPE is greater than the model's forecast MSPE against the one-sided (upper tail) alternative that the historical average MSPE is greater than the model's forecast MSPE ($H_0 : R^2_{OOS} \leq 0$ vs $H_1 : R^2_{OOS} > 0$). Interestingly, for some R^2_{OOS} statistics in Table 2, we reject $R^2_{OOS} \leq 0$ in favor of $R^2_{OOS} > 0$, even though R^2_{OOS} is negative.[22] The best R^2_{OOS} is achieved for equally weighted univariate models with constant coefficients.[23] This nested configuration is the econometric setup used by Rapach, Strauss, and Zhou (2010). It is striking that this model configurations and the *Univariate Models-EW* configuration do better in terms of R^2_{OOS} during recessions than during expansions, while this is not the case for the remaining models. Generally, we observe a clear pattern with small models displaying better R^2_{OOS} than large models.

In terms of predictive likelihoods, the implication from Table 2 is a clear one: models allowing for variation in volatility substantially ouperform constant volatility models in terms of density forecast accuracy. This result is in strong agreement with Figure 2, documenting essentially no data support for $\beta = 1$. Other features, such as time-varying coefficients or the number of regressors in the models, seem to

[21] $R^2_{OOS} = 1 - \left(\frac{MSPE_i}{MSPE_{hist}}\right)$. If $R^2_{OOS} > 0$, the forecast model i is more accurate than the historical average in terms of the MSPE.

[22] Rejecting H_0 despite a negative test statistic is possible, since the MSPE-adjusted statistic proposed by Clark and West (2007) is a test of population-level predictability.

[23] The MSPE of this configuration is less than the MSPE of any forecast with only one regressor and even less than any of the MSPEs of any of the $b \cdot d \cdot 2^K = 3 \cdot 3 \cdot 2^{12} = 36,864$ individual forecast models. Results are available upon request.

play, if any, a minor role. In this respect, also the forecast weighting scheme does not appear to be important. Equal weighted models do worse than models with unrestricted weights due to the imposed inclusion of constant volatility models.

Figure 6 documents the evolution of cumulated differences in the sum of predictive log likelihoods between the *BDMA* and the *Historical Mean-CC* configuration. Including regressors has not provided value in terms of increased density forecast accuracy, evaluated over the entire period from 1947 : 01 to 2012 : 12. We also employ this kind of graphical device (see Figure 7) to assess the evolution of the cumulated squared errors between the *BDMA* and the *Historical Mean-CC* configuration. It is apparent that point forecast accuracy is higher for the *BDMA* configuration during the time of the Oil Shock ($1973 - 1975$) and the Subprime Crisis (2008/09). However, immediately after those periods forecast performance deteriorates again. This points to the difficulty for models of returning to other "regimes" of the economy. Altogether, the provided flexibility for the *BDMA* does not pay off in terms of point and density forecasting accuracy. It is interesting in this context to recall the statement of Welch and Goyal (2008) about the influence of the Oil Shock on results for point prediction accuracy: "If we exclude the Oil Shock, most models [`note`: models with economic covariates] perform even worse - many were statistically significant in the past only because of the stellar model performance during these contiguous unusual years. One can only imagine whether our profession would have been equally comfortable rationalizing away these years "as unusual" if they had been the main negative and not the main positive influence."

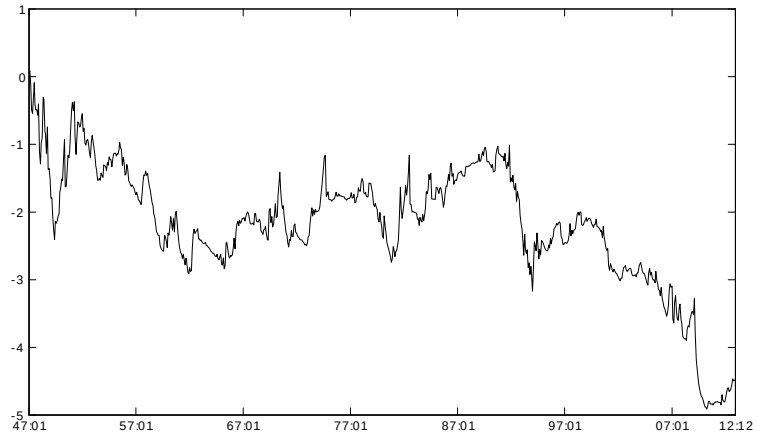

Figure 6: Evolution of cumulated differences in the sum of predictive log likelihoods between the *BDMA* and the *Historical Mean-CC* configuration. Positive (negative) values indicate better (worse) cumulated performance up to the considered point in time for the *BDMA* configuration relative to the *Historical Mean-CC* configuration.

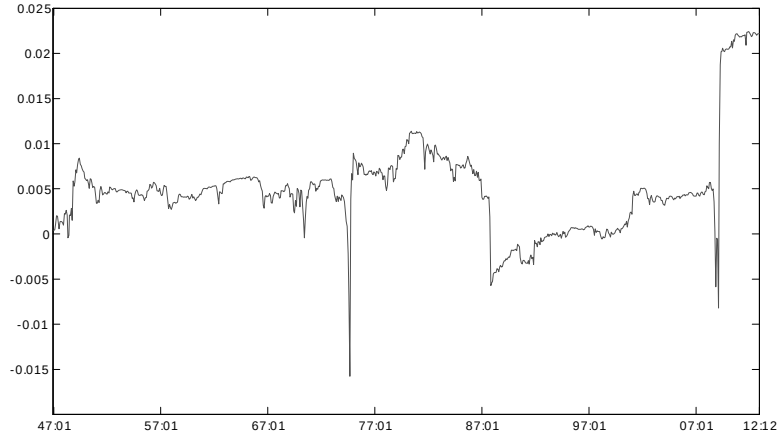

Figure 7: Evolution of cumulated differences in the sum of squared errors between *BDMA* and the *Historical Mean-CC* configuration. Negative (positive) values indicate better (worse) cumulated performance up to the considered point in time for the *BDMA* configuration relative to the *Historical Mean-CC* configuration.

Table 2: Statistical evaluation.

Sum(PL) indicates the sum of predictive log likelihoods. R^2_{OOS} measures the percentage reduction in mean squared prediction error (MSPE) based on the forecast of the respective model relative to the historical average benchmark forecast. Statistical significance is assessed by the Clark and West (2007) test. *,**, *** indicate significance at the 10%, 5% and 1% level, respectively, that the historical average MSPE is less or equal to the respective predictive model's MSPE against the alternative that the historical average MSPE is greater than the predictive model's MSPE. R^2_{OOS} statistics are computed for the entire $1947:01-2012:12$ forecast evaluation period and separately for NBER-dated expansions (exp.) and recessions (rec.).

Model configuration	Sum(PL)	R^2_{OOS}	R^2_{OOS}(exp.)	R^2_{OOS} (rec.)
BDMA	1402	-1.58	-0.02^{*}	-5.20
BDMS	1393	-3.30	-2.90	-4.23
CC-CV-BMA	1263	-5.98	-5.95	-6.03
CV	1274	-2.46	-0.82^{*}	-6.26
CV-BMA	1282	-3.07	-1.62^{*}	-6.42
CC	1402	-1.39	-0.70	-3.00
CC-EW	1365	-3.15	-2.02	-5.76
EW	1366	-2.54^{*}	-0.96^{*}	-6.22
BMA	1404	-0.67	-0.61	-0.81
Kitchen-Sink	1384	-10.11^{*}	-7.40^{*}	-16.40
Kitchen-Sink-CC-CV	1252	-16.58	-15.30	-19.54
Large Models	1394	-3.95	-1.42^{*}	-9.81
Large Models-CC-CV	1261	-7.45	-5.86	-11.14
Medium Models	1400	-1.92	-0.12^{**}	-6.11
Medium Models-CC-CV	1266	-2.54	-1.56	-4.82
Small Models	1404	-1.07	-0.13^{*}	-3.25
Small Models-CC-CV	1268	0.15^{*}	0.28^{**}	-0.14
Univariate Models	1405	-0.91	-0.51	-1.84
Univariate Models-CC-CV	1267	0.37^{**}	0.47^{**}	0.16
Univariate Models-EW	1378	0.40	-0.25	1.91
Univariate Models-CC-CV-EW	1267	0.64^{***}	0.54^{**}	0.88^{**}
Historical Mean-CC	1406	0.00	0.00	0.00

4.5.2 Economic Evaluation

To evaluate the economic value of our proposed forecast method, we analyze them within a real-time portfolio allocation. We consider an investor who allocates wealth between the S&P 500 index and one-month T-Bills. At the end of each month $t-1$, the investor chooses the fraction ϕ_t to be held in the stock index for the period $(t-1, t]$, based on the overall density forecast of stock index returns in $t-1$. We limit ϕ_t in the interval $[0; 1.5]$ and assume an investor with power utility.[24] At the end of each period, the investor maximizes the power utility function

$$u(W_t) = \frac{W_t^{1-\gamma}}{1-\gamma}, \quad \gamma > 1, \tag{20}$$

where γ is the coefficient of relative risk aversion and W_t denotes the wealth at time t, which is equal to

$$W_t = W_{t-1}\left[(1-\phi_t)(1+r_{f,t-1}) + \phi_t\left(1 + r_{f,t-1} + \widetilde{y}_t - 2c\left|\phi_t - \phi_{t-1}\right|\right)\right]. \tag{21}$$

$r_{f,t-1}$ the one-step ahead risk-free rate rate from $t-1$ to t and $\widetilde{y}_t$ the one-step ahead forecast of the equity premium made at time $t-1$. c indicates a fixed percentage of transaction costs on each traded dollar.[25] Setting $W_{t-1} = 1$, the

[24] Assuming power utility, accomodating for higher moments might be useful. Our aggregate model is a mixture Student-t distribution and, hence, able to reflect higher moments.

[25] The multiplication by 2 is due to the fact that the investor rebalances investments in both stocks and bonds. The considered strategies in our empirical analysis vary with respect to turnover. To compare results in a fair setting, we include transaction costs.

investor's optimization problem can be expressed as

$$\max_{\phi_t \in [0;1.5]} \mathbb{E}_{t-1} \left[\frac{\left((1-\phi_t)(1+r_{f,t-1}) + \phi_t \left(1 + r_{f,t-1} + \widetilde{y}_t - 2c\left|\phi_t - \phi_{t-1}\right|\right)\right)^{1-\gamma}}{1-\gamma} \right]. \quad (22)$$

The expectation depends on the predictive density for the equity premium. That is, the investor faces the problem

$$\max_{\phi_t \in [0;1.5]} \int u(W_t)\, p(\widetilde{y}_t | I_{t-1})\, d\widetilde{y}_t. \quad (23)$$

To approximate the integral in (23), we generate a large number N[26] of random numbers from the (mixture Student-t) predictive density[27] and employ numerical optimization to find

$$\max_{\phi_t \in [0;1.5]} \frac{1}{N} \sum_{n=1}^{N} \left[\frac{\left((1-\phi_t)(1+r_{f,t-1}) + \phi_t \left(1 + r_{f,t-1} + \widetilde{y}_t^{(n)} - 2c\left|\phi_t - \phi_{t-1}\right|\right)\right)^{1-\gamma}}{1-\gamma} \right]. \quad (24)$$

Table 3 reports two measures to assess the economic value of the forecast

[26] We set $N = 100,000$.

[27] Using t-distributed predictive densities of returns, expected utility can be infinite. We therefore monitor if draws from the monthly return distribution are smaller than -100% or greater than $+100\%$. In these cases, we would restrict the forecasts to be within the range $[-100\%; +100\%]$. However, in our simulations the bounds are never hit.

Table 3: Economic evaluation.

This table reports portfolio performance measures for an investor with power utility and risk aversion coefficient $\gamma = 3$. The investor allocates monthly between equities and risk-free bills using one of the different forecasting models for the equity risk premium. Δ is the annualized CER. Δ statistics are also reported separately for NBER-dated expansions (exp.) and recessions (rec.). We report the annualized Sharpe ratio (SR) for each configuration as a further measure of economic value of the forecast models. All numbers are reported after deducting proportional transaction costs of 50 basis points ($c = 0.0005$) per transaction. All numbers are in percent.

Model configuration	Δ(overall)	Δ(exp.)	Δ(rec.)	SR
BDMA	9.33	10.31	4.52	57.53
BDMS	6.53	8.24	−1.88	39.13
CC-CV-BMA	5.00	6.68	−3.23	21.06
CV	8.09	8.81	4.56	52.58
CV-BMA	8.14	9.02	3.83	52.02
CC	8.52	9.94	1.56	50.92
CC-EW	6.56	8.12	−1.05	36.43
EW	8.47	9.34	4.16	55.67
BMA	8.47	9.82	1.85	50.97
Kitchen-Sink	9.19	10.32	3.65	55.83
Kitchen-Sink-CC-CV	6.28	7.45	0.58	33.85
Large Models	9.65	10.36	6.16	60.05
Large Models-CC-CV	6.79	7.99	0.89	39.11
Medium Models	9.59	10.45	5.45	60.00
Medium Models-CC-CV	6.81	7.91	1.41	42.57
Small Models	9.13	9.99	4.89	57.00
Small Models-CC-CV	7.00	7.74	3.36	53.54
Univariate Models	8.52	9.59	3.25	51.72
Univariate Models-CC-CV	6.76	8.03	0.51	44.08
Univariate Models-EW	7.89	8.55	4.66	51.13
Univariate Models-CC-CV-EW	6.91	7.97	1.71	47.58
Historical Mean-CC	8.02	10.28	−3.02	47.04

Table 4: Economic evaluation for univariate forecasts.
This table displays the economic value in terms of CERs for each regressor. Four variants are considered: no restrictions (−), constant coefficients (CC), constant volatility (CV), both constant coefficients and constant volatility (CC-CV).

	CER			
Variable	-	CC	CV	CC-CV
dy	7.75	6.84	6.53	6.50
ep	7.39	7.83	6.96	7.31
dpayr	8.18	8.06	7.18	7.03
svar	7.88	7.46	6.70	6.42
bmr	6.72	6.47	6.25	5.47
ntis	8.51	7.53	7.53	6.94
tbl	8.78	8.80	6.81	6.59
lty	8.40	8.63	6.48	6.50
ltr	8.17	7.54	7.01	6.38
dfr	8.43	8.06	6.94	6.79
dfy	7.89	6.76	7.10	6.34
inf	7.99	8.26	6.81	6.79

models: the annualized CER[28] for power utility with $\gamma = 3$[29] and the annualized Sharpe ratio (SR). Further, we also separately report results for the CERs for National Bureau of Economic Research (NBER)-dated expansions and recessions.

A key question is how the imposed restrictions on regressors, coefficients, volatility and model weights of various model configurations affect economic value

[28] For power utility, the annualized CER (in percent) is calculated as

$$CER = \left\{ \left[(1-\gamma) E^{-1} \sum_{t=B+1}^{B+E} u(W_t) \right]^{\frac{1}{1-\gamma}} - 1 \right\} \cdot 1200.$$

B denotes the "burn-in" sample and E indicates the evaluation period. $E^{-1} \sum_{t=B+1}^{B+E} u(W_t)$ denotes the mean realized utility.

[29] We experimented with lower and higher risk aversion coefficients. Our findings are qualitatively unaffected. However, as expected, differences in CERs and Sharpe ratios are smaller (higher) for higher (lower) risk aversion coefficients.

in terms of CERs and Sharpe ratios. Table 3 documents that, in general, restrictions negatively affect performance: model configurations with imposed constant coefficients and constant variance fare worse than competitor configurations allowing for time-varying coefficients and time-varying volatility. Particularly, if various restrictions are imposed simultaneaously, the performance measures deteriorate. This pattern is striking for the *CC-CV-BMA* model with an annualized CER of exactly 500 basis points and an annualized Sharpe ratio of roughly 21%. In contrast to point prediction accuracy in terms of R^2_{OOS}, shrinkage does generally not provide value with respect to CERs and Sharpe ratios. Even the Kitchen-Sink model with all 12 regressors fares comparatively well, given its R^2_{OOS} of approximately -10%. This is due to the inlusion of stochastic volatility. Even if point forecasts are inaccurate, the high predictive volatility for next month's index return (including estimation errors for the coefficients) leads to reductions in the share of the risky asset. The *Large Models* and *Medium Models* configurations do even slightly better in terms of our economic evaluation criteria than the unrestricted configuration *BDMA*, while the *Small Models* and *Univariate Models* configurations perform worse than *BDMA*. The favourite model configuration in terms of R^2_{OOS}, *Univariate Models-CC-CV-EW*, underperforms the *Historical Mean-CC* configuration in terms of economic gains. Altogether, this shows that there can be substantial disagreement between statistical and economic metrics of utility.[30] The agreement between density forecast accuracy (in terms of predictive log likelihoods) and economic measures is stronger, while the economic evaluation

[30]It would also be of great interest to formally assess the statistical significance of the CERs. However, formal tests for this issue have not been fully developed. See McCracken and Valente (2012) for initial results.

criteria CER and Sharpe ratio strongly agree. This finding is completely in line with the results of Cenesizoglu and Timmermann (2012) who document a weak relationship between point forecast accuracy and economic evaluation criteria, but a stronger agreement between predictive density forecasts and economic value.

The *CV-BMA* configuration corresponds to the econometric setup employed by Dangl and Halling (2012). They document that allowing for time-varying coefficients, along with model combinations using BMA, provides sizeable utility gains as well as substantial and significant improvement in Clark and West (2007) MSPE-adjusted statistics. However, in contrast to their study, we do not employ the cross-sectional beta premium (*csp*)[31] as a regressor to predict equity premia.[32] Our results for the *CV-BMA* configuration are clearly inferior to those reported by Dangl and Halling (2012), both in terms of utility gains as well as for MSPE-adjusted statistics. However, if we include the *csp* variable into our set of regressors and also adopt the (slightly) different study design employed by Dangl and Halling (2012), we are able to reproduce their results. Differences in results are clearly due to the *csp* variable and not to differences in the study design.

The *BDMS* approach only uses a single (potentially different) model to forecast at each point in time, and, hence, allows for fast switching of models. However,

[31] The cross-sectional beta premium (*csp*) quantifies the relative valuation of high- and low-beta stocks according to Polk, Thompson, and Vuolteenaho (2006).

[32] Amit Goyal states: "The *csp* data in the original paper was incorrect. It was an auxiliary series from the Polk+ paper [`note`: Polk, Thompson, and Vuolteenaho (2006)]. We could not replicate their primary *csp* data and thus their results. This may well be our problem, not their's." The supplementary file with updated results and corrections (including this statement) is available at: `http://www.hec.unil.ch/agoyal/docs/PaperTables2009.pdf.` The variable *csp* is included in the dataset provided by Amit Goyal and also used by Dangl and Halling (2012). The dataset is available at: (`http://www.hec.unil.ch/agoyal/`).

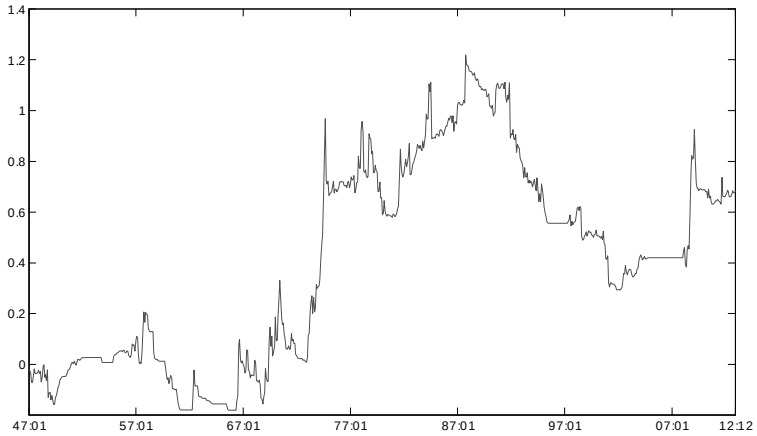

Figure 8: Evolution of cumulated differences in realized utility between *BDMA* and the *Historical Mean-CC* configuration. Positive (negative) values indicate better (worse) cumulated performance up to the considered point in time for the *BDMA* configuration relative to the *Historical Mean-CC* configuration.

both in terms of economic and statistical performance, this configuration does not fare well. This is to be expected, given the broad empirical support for slow changing model weights (see Figure 4).

Table 4 presents results for univariate regressions (for each considered regressor) in terms of CERs. We show differences between results when no further restrictions are imposed, for constant coefficient, constant variance and for both constant coefficients and constant variance. The latter configuration is employed in the study by Welch and Goyal (2008). A clear indication from Table 4 is that imposing constant volatility is detrimental to utility gains. On the other hand, the effects of imposing constant coefficients in univariate models is ambiguous.

When assessing economic value separately for expansions and recessions, it is apparent that forecast models which overall fare better relative to the historical

mean forecast achieve improvements during recessions. This pattern is well-known in the literature on equity premium prediction; see, e.g., Rapach, Strauss, and Zhou (2010), Henkel, Martin, and Nardari (2011) and Dangl and Halling (2012). Further, the observation of increased predictability during recessions is supported by asset pricing theory (Campbell and Cochrane, 1999).[33]

Figure 8 shows the evolution of cumulated differences in realized utility between the *BDMA* and the *Historical Mean-CC* configuration (for an investor with power utility and $\gamma = 3$). Though the cumulated difference is positive for most of the time, this graphical device sheds light on the way how the difference between the two models has evolved over time. The more flexible model fared better only during two periods, namely, around the Oil Shock $1973 - 75$ and the Subprime Crisis 2008/09. This finding is helpful in two respects: firstly, since those periods happened to be recessions, it clarifies why the model has done better than the simpler model during recessions. Secondly, it reveals that the outperformance of the *BDMA* model is driven by small subperiods, particularly the period around the Oil Shock.

5 Conclusion

This article has introduced a Bayesian version of Dynamic Model Averaging. The setup allows for rigorous modelling of uncertainties. We specify a large set of individual models, differing with respect to included regressors, stability of coefficients and volatility dynamics. Individual forecasts are monitored, tracking model per-

[33] In the habit formation model, Campbell and Cochrane (1999) argue that risk aversion increases during economic downturns, thereby generating equity premium predictability.

formance over time in terms of predictive likelihoods. Flexible model combination schemes shift weights according to the (recent) forecast performance of the models. The aggregate model sequentially generates predictive densities, accounting for many sources of uncertainty.

We forecast monthly US equity premia and evaluate the forecasts in terms of statistical and economic criteria. Our results add empirical evidence that shrinking and combining forecasts can result in more precise point forecasts relative to the prevailing historical mean. We find that model specifications allowing for stochastic volatility improve density forecast accuracy and increase economic gains, as measured by CERs and the Sharpe ratio. Our proposed methodology does not identify any of the included covariates to be particularly important for predicting equity premia over a considerably long period of time. With respect to the degree of instability of coefficients, stable and gradually evolving behavior is favored rather than abruptly changing coefficients. There is only a low degree of likelihood discounting, being consistent with slow changing model weights. We document disagreement between statistical and economic metrics of forecast performance, that is, point prediction accuracy and economic gains. With density forecasts and economic criteria being more in agreement, exploiting the entire return distribution for asset allocation rather than focus on point predictions appears to pay off. Most importantly, however, while utility gains are generally higher for more flexible models when evaluated over the entire evaluation period, this result is not robust. The identified gains are largely driven by exceptional and short periods of time, particularly the time period around the Oil Shock ($1973 - 1975$) and the Subprime Crisis (2008/09).

The implications of our results reconcile seemingly contradictory views in the literature: Welch and Goyal (2008) take a sceptical view with respect to enhancing point prediction accuracy for equity premia by means of economic covariates. Just as we, they ascribe positive findings to the time period around the Oil Shock. At the same time, they argue that the forecast models would not have been helpful for asset allocation relative to the simple historical benchmark. Subsequent studies such as Rapach, Strauss, and Zhou (2010), Ferreira and Santa-Clara (2011), Rapach, Strauss, and Zhou (2010) and Johannes, Korteweg, and Polson (2013) adopted more elaborated econometric methods. With these various techniques, empirical evidence in favour of predictability and economic utility gains (of a similar magnitude as in our study for the *BDMA* configuration) is adduced. However, these studies do not investigate the impact of the Oil Shock.

Our approach directly nests some of the econometric settings of those studies (Welch and Goyal, 2008; Rapach, Strauss, and Zhou, 2010; Dangl and Halling, 2012) and is similar to the setup employed by Johannes, Korteweg, and Polson (2013). By construction, our approach avoids "cherry-picking". Against this background, we reinforce the statement advanced by Welch and Goyal (2008): "Although it is possible to search for, to occasionally stumble upon, and then to defend some seemingly statistically significant models, we interpret our results to suggest that a healthy scepticism is appropriate when it comes to predicting the equity premium, [...]. The models do not seem robust." Further, our results point to the superiority of stochastic volatility models in terms of density forecast accuracy and economic gains. Therefore we strenghten the conclusion drawn by Cenesizoglu and Timmermann (2012) that "[...] the debate on return predictabil-

ity has focused too narrowly on statistical measures of forecast precision such as root mean squared forecast errors [...]."

References

Ang, A., and G. Bekaert (2002): "International asset allocation with regime shifts," *Review of Financial studies*, 15(4), 1137–1187.

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

Billio, M., R. Casarin, F. Ravazzolo, and H. K. van Dijk (2013): "Time-varying combinations of predictive densities using nonlinear filtering," *Journal of Econometrics*, 177(2), 213 – 232, Dynamic Econometric Modeling and Forecasting.

Bossaerts, P., and P. Hillion (1999): "Implementing statistical criteria to select return forecasting models: what do we learn?," *Review of Financial Studies*, 12(2), 405–428.

Boudoukh, J., R. Michaely, M. Richardson, and M. R. Roberts (2007): "On the importance of measuring payout yield: Implications for empirical asset pricing," *The Journal of Finance*, 62(2), 877–915.

Campbell, J. Y., and J. H. Cochrane (1999): "By Force of Habit: A Consumption Based Explanation of Aggregate Stock Market Behavior," *Journal of Political Economy*, 107(2), 205–251.

Campbell, J. Y., and S. B. Thompson (2008): "Predicting excess stock returns out of sample: Can anything beat the historical average?," *Review of Financial Studies*, 21(4), 1509–1531.

Cenesizoglu, T., and A. Timmermann (2012): "Do return prediction models add economic value?," *Journal of Banking & Finance*, 36(11), 2974 – 2987.

Clark, T. E., and K. D. West (2007): "Approximately normal tests for equal predictive accuracy in nested models," *Journal of Econometrics*, 138(1), 291–311.

Cremers, K. J. M. (2002): "Stock Return Predictability: A Bayesian Model Selection Perspective," *Review of Financial Studies*, 15(4), 1223–1249.

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

Dawid, A. P. (1984): "Present position and potential developments: Some personal views: Statistical theory: The prequential approach," *Journal of the Royal Statistical Society. Series A (General)*, pp. 278–292.

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

Ferreira, M. A., and P. 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.

Geweke, J., and G. Amisano (2011): "Optimal prediction pools," *Journal of Econometrics*, 164(1), 130–141.

——— (2012): "Prediction with misspecified models," *The American Economic Review*, 102(3), 482–486.

Geweke, J., and C. Whiteman (2006): "Bayesian forecasting," *Handbook of economic forecasting*, 1, 3–80.

Hannan, E. J., A. McDougall, and D. Poskitt (1989): "Recursive estimation of autoregressions," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 217–233.

Hans, C., A. Dobra, and M. West (2007): "Shotgun stochastic search for large p regression," *Journal of the American Statistical Association*, 102(478), 507–516.

Henkel, S. J., J. S. Martin, and F. Nardari (2011): "Time-varying short-horizon predictability," *Journal of Financial Economics*, 99(3), 560–580.

Hoogerheide, L., R. Kleijn, F. Ravazzolo, H. K. Van Dijk, and M. Verbeek (2010): "Forecast accuracy and economic gains from Bayesian model averaging using time-varying weights," *Journal of Forecasting*, 29(1-2), 251–269.

Johannes, M., A. Korteweg, and N. Polson (2013): "Sequential Learning, Predictability, and Optimal Portfolio Returns," *The Journal of Finance*, pp. n/a–n/a.

Koop, G., and D. Korobilis (2012): "Forecasting inflation using dynamic model averaging," *International Economic Review*, 53(3), 867–886.

MADIGAN, D., J. YORK, AND D. ALLARD (1995): "Bayesian graphical models for discrete data," *International Statistical Review/Revue Internationale de Statistique*, pp. 215–232.

MCCRACKEN, M., AND G. VALENTE (2012): "Testing the economic value of asset return predictability," *FRB of St. Louis Working Paper No.*

NEELY, C. J., D. E. RAPACH, J. TU, AND G. ZHOU (2010): "Forecasting the equity risk premium: the role of technical indicators," *Federal Reserve Bank of St. Louis Working Paper Series*.

PAYE, B. S., AND A. TIMMERMANN (2006): "Instability of return prediction models," *Journal of Empirical Finance*, 13(3), 274–315.

PESARAN, M. H., AND A. PICK (2011): "Forecast Combination Across Estimation Windows," *Journal of Business & Economic Statistics*, 29(2), 307–318.

PESARAN, M. H., AND A. TIMMERMANN (1995): "Predictability of stock returns: Robustness and economic significance," *The Journal of Finance*, 50(4), 1201–1228.

PETTENUZZO, D., AND A. TIMMERMANN (2011): "Predictability of stock returns and asset allocation under structural breaks," *Journal of Econometrics*, 164(1), 60–78.

PETTENUZZO, D., A. TIMMERMANN, AND R. VALKANOV (2013): "Forecasting Stock Returns under Economic Constraints," Working Papers 57, Brandeis University, Department of Economics and International Businesss School.

Polk, C., S. Thompson, and T. Vuolteenaho (2006): "Cross-sectional forecasts of the equity premium," *Journal of Financial Economics*, 81(1), 101–141.

Prado, R., and M. West (2010): *Time Series: Modeling, Computation, and Interface*, Chapman and Hall/CRC Texts in Statistical Science Series. Chapman & Hall/CRC.

Raftery, A. E., M. Kárný, and P. Ettler (2010): "Online Prediction Under Model Uncertainty via Dynamic Model Averaging: Application to a Cold Rolling Mill.," *Technometrics : a journal of statistics for the physical, chemical, and engineering sciences*, 52(1), 52–66.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997): "Bayesian model averaging for linear regression models," *Journal of the American Statistical Association*, 92(437), 179–191.

Rapach, D., and G. Zhou (2012): "Forecasting stock returns," *Handbook of Economic Forecasting*, 2.

Rapach, D. E., J. K. Strauss, and G. Zhou (2010): "Out-of-sample equity premium prediction: Combination forecasts and links to the real economy," *Review of Financial Studies*, 23(2), 821–862.

Stock, J. H., and M. W. Watson (2004): "Combination forecasts of output growth in a seven-country data set," *Journal of Forecasting*, 23(6), 405–430.

Welch, I., and A. Goyal (2008): "A comprehensive look at the empirical performance of equity premium prediction," *Review of Financial Studies*, 21(4), 1455–1508.

West, M., and J. Harrison (1997): *Bayesian forecasting and dynamic models*. Springer, 2nd edn.

Xia, Y. (2001): "Learning about predictability: The effects of parameter uncertainty on dynamic asset allocation," *The Journal of Finance*, 56(1), 205–246.

A Appendix

A.1 Structure of Dynamic Linear Models

Building on the specification of the dynamic linear model in equations (1) and (2), we describe the sequential updating of the the beliefs about system coefficients, the scale matrix of the coefficients and the observational variance. Suppose, at some arbitrary time $t-1$, we have already observed y_{t-1}. Hence, we are able to form a posterior belief about the values of the unobservable coefficients $\theta_{t-1}|I_{t-1}$ and of the observational variance $V_{t-1}|I_{t-1}$. These posteriors are normally/inverse-gamma distributed

$$V_{t-1}|I_{t-1} \sim IG\left[\frac{n_{t-1}}{2}, \frac{n_{t-1}S_{t-1}}{2}\right], \tag{25}$$

$$\theta_{t-1}|I_{t-1}, V_{t-1} \sim N\left[m_{t-1}, V_{t-1}C^*_{t-1}\right]. \tag{26}$$

After integrating out the uncertainty in the observational variance, the posteriors of the coefficients are t-distributed as

$$\theta_{t-1}|I_{t-1} \sim t_{n_{t-1}}\left[m_{t-1}, S_{t-1}C^*_{t-1}\right]. \tag{27}$$

The prior distribution of the time-varying regression coefficients, $\theta_t|I_{t-1}$ accomodates for the system coefficients being exposed to shocks, increasing the system variance by W_t,

$$\theta_t|I_{t-1} \sim t_{\beta n_{t-1}}\left[m_{t-1}, S_{t-1}C^*_{t-1} + S_{t-1}W^*_t\right]. \tag{28}$$

(6), (7) and (8) in the main text show the discount approach for specifying W_t.

The prior for the observational variance is

$$V_t|I_{t-1} \sim IG\left[\beta\frac{n_{t-1}}{2}, \beta\frac{n_{t-1}S_{t-1}}{2}\right]. \tag{29}$$

Notice the difference between the posterior for the observational variance in (25) and the prior for the observational variance in (29). The modelling approach for the evolution of the observational variance assumes that the observational variance is subject to some random disturbance over the time interval $(t-1, t]$. The discount factor $\beta \in \{\beta_1, ..., \beta_b\}$, $\beta \in (0; 1]$ models a decay of information between the time points and retains the marginal inverse gamma form of the prior and posterior distribution, ensuring conjugacy. Based on the time $t-1$ posterior (25), deriving $V_t|I_{t-1}$ involves a random-walk like stochastic beta/inverse-gamma evolution for the sequence of observational variances, resulting in the time-t prior distribution (29). It has the same location as (25), that is, $\mathbb{E}_{t-1}(V_t) = \mathbb{E}_{t-1}(V_{t-1}) = S_{t-1}$ but increased dispersion through the discounting of the degrees of freedom (see (9) in the main text).[34]

The predictive density of y_t is obtained by integrating the conditional density of y_t over the range of θ_t and V_t. Let $\vartheta(y; \mu, \sigma^2)$ denote the density of a normal distribution evaluated at y and $IG(V; a, b)$ the density of an $IG(a, b)$ distributed

[34] The variance discounting approach induces robustness and protection against potential biases in estimation of the state vector and can also protect against aspects of model misspecification; see Prado and West (2010), page 132.

variable evaluated at V. We obtain the predictive density as

$$\begin{aligned} p\left(y_t|I_{t-1}\right) &= \int_0^\infty \left[\int \vartheta\left(\widetilde{y}_t; F_t^{'}\theta_t, V_t\right) \vartheta\left(\theta_t; m_{t-1}^{'}, V_t\left(C_{t-1}^{*}+W_t^{*}\right)\right) d\theta_t\right] \\ &\quad \times IG\left(\widetilde{V}_t; \beta\frac{n_{t-1}}{2}, \beta\frac{S_{t-1}n_{t-1}}{2}\right) dV_t \\ &= \int_0^\infty \vartheta\left(\widetilde{y}_t; F_t^{'}m_{t-1}, V_t\left[1+F_t^{'}\left(C_{t-1}^{*}+W_t^{*}\right)F_t\right]\right) \\ &\quad \times IG\left(\widetilde{V}_t; \beta\frac{n_{t-1}}{2}, \beta\frac{S_{t-1}n_{t-1}}{2}\right) dV_t. \end{aligned}$$

The predictive density

$$p\left(y_t|I_{t-1}\right) = t_{\beta n_{t-1}}\left(\widetilde{y}_t; F_t^{'}m_{t-1}, \underbrace{S_{t-1}\cdot\underbrace{\left[1+F_t^{'}\left(\underbrace{C_{t-1}^{*}+W_t^{*}}_{:=R_t^{*}}\right)F_t\right]}_{:=Q_t^{*}}}_{:=Q_t}\right) \tag{30}$$

is Student-t distributed with location $F_t^{'}m_{t-1,}$ scale Q_t and βn_{t-1} degrees of freedom, evaluated at $\widetilde{y}_t$. R_t denotes the prior variance of the coefficient vector θ_t. S_{t-1} represents the estimate for the observational variance. With all inputs for the predictive density determined, the prediction step is finished and we continue to outline the update step.

After the y_t has materialized, the priors about the system coefficients and the observational variance are updated based on the prediction error

$$e_t = y_t - \widehat{y}_t. \tag{31}$$

Combining the time−t prior (29) for the observational variance[35]

$$p\left(V_t|I_{t-1}\right) \propto V_t^{\frac{-\beta n_{t-1}}{2}} \exp\left(-\frac{\beta n_{t-1} S_{t-1}}{2V_t}\right) \tag{32}$$

for $V_t > 0$ with the (conditionally) normal likelihood
$y_t|I_{t-1}, V_t \sim N\left(F_t' m_{t-1,} V_t \frac{Q_t}{S_{t-1}}\right)$,

$$p\left(y_t|V_t, I_{t-1}\right) \propto V_t^{\frac{1}{2}} \exp\left(\frac{-e_t^2 S_{t-1}}{2V_t Q_t}\right), \tag{33}$$

we obtain the inverse-gamma distributed posterior for the observational variance

$$p\left(V_t|I_t\right) \propto V_t^{\frac{-\beta n_t}{2}} \exp\left(-\frac{n_t S_t}{2V_t}\right), \tag{34}$$

with the updated point estimate S_t and the updated degrees of freedom n_t (see (9) and (10) in the main text).

The $m \times 1$ adaptive coefficient vector

[35] The variance discounting approach underlies a multiplicative model for generating $V_t|I_{t-1}$ from $V_{t-1}|I_{t-1}$. Suppose γ_t to be a beta distributed random variable, independent of V_{t-1}, with density $p\left(\gamma_t|I_{t-1}\right) \sim Beta\left[\beta\frac{n_{t-1}}{2}, (1-\beta)\frac{n_{t-1}}{2}\right]$, and $\mathbb{E}_{t-1}\left(\gamma_t\right) = \beta$ for $0 < \gamma_t < 1$. Given V_{t-1}, set $V_t = \gamma_t V_{t-1}/\beta$. The resulting distribution of V_t is the time-t prior (29). The evolution $V_t = \gamma_t V_{t-1}/\beta$ formally models stochastic variation in the observational variance sequence. The variance discounting arises from a stochastic evolution in which the V_t sequence changes as a result of independent random "shocks" γ_t/β. The variance discounting approach is documented in detail in West and Harrison (1997), (page 360 et seq.), and Prado and West (2010), (page 132 et seq.).

$$A_t = \frac{R_t F_t}{Q_t} \tag{35}$$

relates the precision of the estimated coefficients to the variance, and hence, the information content of the current observation. A_t determines the degree to which the updated values for estimates of the coefficients react to new observations. Updating for point estimates of the system coefficients and the associated estimate of the scale matrix is completed by

$$m_t = m_{t-1} + A_t e_t, \tag{36}$$

$$C_t = \frac{S_t}{S_{t-1}} \left(R_t - A_t A_t' Q_t \right). \tag{37}$$

A.2 Connection between Marginal and Predictive Likelihoods

To show the connection between DMA and classical (static) BMA, we exploit the link between marginal and predictive likelihoods (Dawid, 1984; Geweke and Whiteman, 2006; Raftery, Kárný, and Ettler, 2010). The marginal likelihood in period t for model j is expressed by the product of past predictive likelihoods for $s = 1, ..., t$.

$$p\left(y^t | M_j\right) = \prod_{s=1}^{t} p\left(y_s | M_j, I_{s-1}\right), \tag{38}$$

where $y^t = \{y_{1,\ldots,}y_t\}$.

In the BMA framework, posterior model probabilities are expressed using Bayes factors for pairwise comparisons (of individual models or model combinations). The Bayes factor for model j against model k is defined by the ratio of marginal likelihoods, $\mathcal{B}_{jk} = \frac{p\left(y^t|M_j\right)}{p(y^t|M_k)}$. It follows from (38) that the ratio of Bayes factors can be expressed as

$$\mathcal{B}_{jk} = \prod_{s=1}^{t} \mathcal{B}_{jk,s} = \frac{\prod_{s=1}^{t} p\left(y_s|I_{s-1}, M_j\right)}{\prod_{s=1}^{t} p\left(y_s|I_{s-1}, M_k\right)}, \tag{39}$$

where $\mathcal{B}_{jk,s}$ is the sample-specific Bayes factor for sample s.

The posterior model probabilities are then (in the case of equal prior probabilities)

$$\frac{p\left(M_j|I_t\right)}{p\left(M_k|I_t\right)} = \prod_{s=1}^{t} \left(\mathcal{B}_{jk,s}\right)^{\alpha^{t-s}}.$$

For $\alpha < 1$, the posterior model probabilities are equal to the exponentially age-weighted product of sample-specific Bayes factors. For $\alpha = 1$, we obtain the usual Bayes factors for arbitrary models and predictive likelihoods correspond to marginal likelihoods (at least, if we consider the entire data set from $s = 1$).