Center for Quantitative Economics

Forecasting Exchange Rates under Model and Parameter Uncertainty

Joscha Beckmann [†] and Rainer Schüssler [‡]

32/2014

[†] University of Duisburg-Essen and Kiel Institute for the World Economy, Germany

[‡] Helmut-Schmidt University Hamburg and Department of Economics, University of Münster, Germany

wissen.leben
WWU Münster

Forecasting Exchange Rates under Model and Parameter Uncertainty[*]

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

August 2014

Abstract

We introduce a forecasting method that closely matches the econometric properties required by the theory on exchange rate prediction. Our approach formally models (i) when (and if) explanatory variables enter or leave a regression model, (ii) the degree of parameter instability, (iii) the (potentially) rapidly changing relevance of regressors, and (iv) the appropriate shrinkage intensity over time. We consider (short-term) forecasting of six major US dollar exchange rates using a standard set of macro fundamentals. Our results indicate the importance of shrinkage and flexible model selection/averaging criteria to avoid poor forecasting results.

JEL: F31, F37, G17

Keywords: Exchange rates forecasting; Time-varying parameter models; Shrinkage; Model selection/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 8th Rimini Bayesian Econometrics Workshop of RCEA, the 89th Western Economic Association, Denver, the 12th INFINITI conference on International Finance, Prato, the Research Seminar at the Westphalian University 2014, Münster, 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

Forecasting problems in economics and finance are in many cases complicated because potential predictive power of the considered regressors appears to be undermined by overfitting and instabilities, resulting in poor out-of-sample forecasting performance.[1] The forecasting literature has addressed those issues focusing on parsimonious models that limit the effect of parameter estimation error through various shrinkage or regularization techniques. Furthermore, forecast combinations have turned out to be useful to stabilize forecasts, since they are robust to structural breaks and model misspecification; see, e.g., Rapach, Strauss, and Zhou (2010).

Particularly, exchange rate forecasting is known as very tough. Although economic fundamentals are considered to contain information with regard to future exchange rate movements, the forecasting performance of exchange rate models has turned out to be frequently inferior to a naive random walk benchmark, a finding that dates back to the seminal study by Meese and Rogoff (1983). Given the lack of success in predicting exchange rates by macro fundamentals, exchange rates are considered as largely disconnected from economic fundamentals. This phenomenon constitutes the "exchange rate disconnect" puzzle (Engel, Mark, and West, 2008).[2] The prevailing view is that Meese and Rogoff's finding has not been convincingly overturned until today. Providing a comprehensive survey study, Rossi (2013) finds that the forecasting ability crucially depends on the choice of

[1] See Rossi, Elliott, and Timmermann (2012) for a recent study on forecasting a very broad set of financial and economic variables under model instability.

[2] The "exchange rate disconnect" puzzle particularly refers to short-term forecasting with horizons of up to one year.

predictors, the forecast horizon, the sample period, the type of forecasting models, and forecast evaluation method. Despite some encouraging result for certain predictors such as Taylor-Rule based forecasts (Molodtsova and Papell, 2009), no predictor or model seems to provide systematically superior forecasts compared to a random walk. Rossi (2013) concludes that predictability only appears occasionally for some countries and short periods of time.

Sarno and Valente (2009) consider forecasting exchange rates using a predictive procedure that allows the relationship between exchange rates and fundamentals to evolve in a very flexible fashion. They conclude that the poor out-of-sample forecasting ability of exchange rate models may be caused by poor in-sample model selection criteria rather than by the lack of information embedded in the fundamentals and that the difficulty in selecting the best predictive model is largely due to frequent shifts in the fundamentals. This finding fuels the search for a model selection/averaging procedure that is able to keep up with frequent model changes. Recent rational expectations models ascribe the instablity between exchange rates and macro fundamentals to imperfect knowledge. Facing incomplete and heterogeneous information, investors in the foreign exchange market attach excessive weight to an observed fundamental - the "scapegoat" variable - during some period (Bacchetta and Van Wincoop, 2004; Bacchetta and Van Wincoop, 2006; Bacchetta and Van Wincoop, 2013).[3] Markiewicz (2012) proposes a learning theory in which

[3] They rationalize exchange rate movements by a shift in an unobserved fundamental (e.g. liquidity trades). Searching for an explanation for the exchange rate change, investors in the foreign exchange rate market may attribute such a movement to an observed macro fundamental. The concerned macro fundamental becomes the "scapegoat" and feeds back to investors' trading strategies, resulting in time-varying weights for the fundamentals. For survey evidence that agents in the foreign exchange rate market frequently change the weight they ascribe to fundamentals, see Cheung and Chinn (2001) and Fratzscher, Sarno, and Zinna (2012).

forecasts based on the selected macro variable feeds back into the actual exchange rate dynamics. The theoretical argument behind those rational expectation models is that investors focus excessively on a time-varying subset of fundamentals that changes over time. This gives rise to the need for an economertric forecasting technique that is able to handle rapid shifts in parameters and allows the relevant subset of economic fundamentals to change over time. That is, an appropriate econometric model should be able to accomodate both parameter instability and model uncertainty. Furthermore, the specified model universe ought to be general enough to comprise all possible models of exchange rate behavior considered plausible by the researcher as well as to allow also for the possibility that none of the regressors is indeed useful for forecasting and, in this case, the model should collapse to a simple random walk specification. Our approach allows a researcher to include a multitude of different model specifications, while (s)he may rely on the mechanism of the method to automatically eliminate potentially unnecessary model features (such as regressors or time-varying coefficients) and, hence, ensures parsimony.[4]

Recent empirical studies on exchange rate prediction employ shrinkage techniques and flexible model averaging or selection criteria: Wright (2008) and Corte, Sarno, and Tsiakas (2009) use Bayesian Model Averaging, Li, Tsiakas, and Wang (2014) use the elastic net as a shrinkage technique and report encouraging results.

[4] An alternative to choose appropriate model specifications in an automated fashion would be sequential hypothesis testing. However, there are at least two problematic issues that arise with such a strategy: (i) Pre-testing problems, (ii) hypothesis tests are designed for constant parameter models. Sequential hypothesis testing analyzes whether a restriction holds globally. However, in time-varying parameter models, restrictions should be allowed to hold locally, that is, at some points in time but not at others. Hypothesis testing cannot properly address this issue.

Berge (2014) uses the gradient boosting method as a shrinkage device. Kouwenberg, Markiewicz, Verhoeks, and Zwinkels (2013) employ a backward elimination rule as model selection criterion that intends to capture the (potentially rapidly changing) set of relevant fundamentals which most accurately predicts exchange rates.

The outlined theoretical and empirical literature on exchange rate forecasting suggests a variety of desired characteristics with respect to a prediction procedure. Our approach is meant to closely match those demands. Against this background, we design a statistical approach that formally models (i) when (and if) explanatory variables enter or leave a regression model, (ii) the degree of parameter instability, (iii) the (potentially) rapidly changing relevance of regressors, and (iv) the appropriate shrinkage intensity over time. We use our proposed method to dissect the different effects that influence forecasting performance for exchange rates. Particularly, we focus on the following key questions: Which set of macro fundamentals, if any, is relevant for forecasting at each point in time? Are time-varying coefficients helpful? Is it worthwile to consider flexible model averaging/selection criteria? How intensively are forecasts shrunk towards zero, that is, how strong is the data support for the random walk model? Are the flexible models able to outperform the random walk benchmark?

With respect to the methodological contribution, our work falls into the domain of shrinkage in time-varying parameter models. We extend the complete subset regression approach advanced by Elliott, Gargano, and Timmermann (2013) as a shrinkage technique for constant linear regression models to a setting that allows for time-varying coefficients, and include flexible model weighting schemes both within and across subsets. There are only few studies in the econometric literature that allow for changing complexity in time-varying parameter (TVP) models:[5] These methods include the time-varying dimension model (Chan, Koop, Leon-Gonzalez, and Strachan, 2012), the Dynamic Model Averaging approach (Koop and Korobilis, 2012) and the normal-gamma autoregressive (NGAR) process prior approach (Kalli and Griffin, 2014).[6] Our suggested approach has some appealing properties: It allows for time-varying and predictor-specific shrinkage intensity, it is transparent and computationally efficient and avoids arbitrary choices to be made by the researcher.

The rest of the paper is organized as follows. Section 2 describes the predictive regressors based on standard empirical exchange rate models, Section 3 introduces the employed model specifications and the econometric methodology underlying our forecasting strategies. In Section 4, we run a Monte Carlo simulation to analyze the behavior of our proposed forecasting method. In Section 5, we report our empirical results. Section 6 concludes.

[5] From a theoretical perspective, one could argue that parameter shrinkage should be sufficient to induce parsimony into TVP models and there was no further need for modelling explicit model change (i.e., where the model dimension can be reduced or expanded over time by setting time-varying coefficients to zero). The argument is that coefficients are allowed to be estimated zero when they are temporarily unnecessary and thus the dimension of the model should (at least approximately) change over time. In this case, model uncertainty would automatically be addressed by modelling parameter instability. However, in practice, such over-parameterized TVP models can lead to poor forecast performance (see the forecasting results in our paper for the "kitchen-sink" models or those in, e.g., Koop and Korobilis (2012) or Chan, Koop, Leon-Gonzalez, and Strachan (2012)).

[6] The approach by Groen, Paap, and Ravazzolo (2013) involves a latent variable that indicates whether a regressor is included in or excluded from the model. The binary decision is irreversible. Hence, the approach is limited in the sense that the relevance of a variable is measured globally, rather than being allowed to fluctuate over time. Similarly, the hierarchical shrinkage prior approach (Belmonte, Koop, and Korobilis, 2014) effectively results in a variable selection method since it shrinks some of the regression coefficients extremely close to zero for the whole time series.

2 The Menu of Fundamentals

Our considered set of variables for predicting end-of-month (log) exchange rate returns comprises four regressors that are based on standard models. The predictive variable in period t, $x_{i,t}$, is defined by the regressors $i = 1, ..., K = 5$. In addition to an intercept ($x_{1,t}$), the regressors are:

2.1 Uncovered Interest Parity

The regressor *UIP* is based on the uncovered interest parity condition as follows:

$$x_{2,t} = i_t - i_t^*. \tag{1}$$

i_t is the domestic one-month nominal interest rate, i_t^* is the foreign one-month nominal interest rate (proxied by Eurodeposit interest rates). The interest rate differential ($i_t - i_t^*$) is identical to the forward premium ($f_t - s_t$) since the literature agrees that the covered interest parity condition holds (Akram, Rime, and Sarno, 2008). f_t denotes the log of the one-month forward exchange rate at time t (i.e., the rate agreed at time t for an exchange of currencies at $t + 1$). s_t denotes the log of the exchange rate.[7]

2.2 Purchasing Power Parity

The regressor *PPP* is based on the purchasing power parity condition:

$$x_{3,t} = p_t - p_t^* - s_t. \tag{2}$$

[7] All data series are obtained from `Datastream`.

p_t denotes the log of the domestic price level, p_t^* the log of the foreign price level.[8]

2.3 Asymmetric Taylor Rule

The regressor *AsyTaylor* is based on the (asymmetric) Taylor (1993) rule as follows:

$$x_{4,t} = 1.5\left(\pi_t - \pi_t^*\right) + 0.1\left(g_t - g_t^*\right) + 0.1\left(s_t + p_t^* - p_t\right). \tag{3}$$

π_t is the domestic inflation rate, π_t^* is the foreign inflation rate, g_t the domestic output gap and g_t^* the foreign output gap. We measure the output gap as the (percent) deviation of real output from an estimate of its potential level calculated using the Hodrick and Prescott (1997) filter.[9] Fixing the parameters to $(1.5, 0.1, 0.1)$, we follow a standard choice in the literature (Molodtsova and Papell, 2009).

2.4 Monetary Fundamentals

The regressor *Monetary* employs monetary fundamentals as follows:

$$x_{5,t} = \left(m_t - m_t^*\right) - \left(ip_t - ip_t^*\right) - s_t. \tag{4}$$

m_t denotes the log of the domestic money supply and m_t^* the log of the foreign money supply.[10] $ip_t^{(*)}$ is the log of the domestic (foreign) industrial production.

[8] Prices are approximated by consumer price indices.

[9] We set the smoothing parameter to $14,400$ as in Molodtsova and Papell (2009).

[10] We use the aggregates M0 and M1 as proxies for money supply.

3 Model Specifications

We define a range of model specifications, starting with a simple constant linear regression model in Section 3.1. Then, we will extend the model step by step. In order to limit the effect of parameter estimation error, we employ the complete subset regression approach (Elliott, Gargano, and Timmermann, 2013) in Section 3.2. With K potential regressors, there are 2^K model combinations if each regressor is either included in or excluded from the model. The complete subset regression involves running predictive regressions for all model configurations that keep the number of predictors fixed. A subset k comprises $\binom{K}{k}$ models of which each of them includes k predictors. Elliott, Gargano, and Timmermann (2013) suggest assigning equal weights to all models within a subset k to provide an aggregate (point) forecast of the subset. Hence, K aggregate subset forecasts are available at each point in time. The authors propose to select the forecast of the subset which would have given the best forecasting performance up to the given point in time as the overall forecast. To increase flexibility, we expand their setup in the following directions: By rewriting the regression model into a state space representation (Section 3.3), we allow for time-varying coefficients. This way, we obtain density forecasts for each model and exploit them to introduce flexible weighting schemes within the subsets (Section 3.4). To combine the (density) forecasts across subsets, we use optimal prediction pools (Geweke and Amisano, 2011), in Section 3.5. We analyze how the shrinkage effect of our model comes into play in Section 3.6.

3.1 Kitchen-Sink Regression

We start with a simple linear regression model with constant parameters:

$$\Delta s_t = \sum_{i}^{K} \alpha + \beta_i x_{i,t} + \varepsilon_{t+1}, \quad \varepsilon_{t+1} = N\left(0, \sigma_\varepsilon^2\right), \tag{5}$$

where Δs_t denotes the difference in the log exchange rate between period t and $t-1$. This model specification is sometimes referred to as "kitchen sink" regression because it throws "everything but the kitchen sink" into the regression. With many possible regressors and only a small sample size (small n, large T), economic forecasting models that include all considered regressors are in many cases plagued by parameter estimation error, resulting in a poor forecasts in terms of mean squared prediction errors. For the case of constant linear regression models, many techniques have been advanced to alleviate the concern of overfitting .[11] We employ complete subset regressions as a shrinkage technique for two reasons. First, the technique outperforms many other shrinkage techniques both in a Monte Carlo experiment and in an equity index forecasting exercise; see Elliott, Gargano, and Timmermann (2013). Second, the method is extendable to the case of time-varying coefficients and flexible model combination schemes.

[11] Those include, among others, bagging (Breiman, 1996), the elastic net (Zou and Hastie, 2005), lasso (Tibshirani, 1996) or Bayesian Model Averaging (Raftery, Madigan, and Hoeting, 1997). All shrinkage methods have one common characteristic: They aim at improving the variance-bias tradeoff to enhance out-of-sample forecasting results. To illustrate this argument, consider a simple linear regression model $y = X\beta + \varepsilon$ with $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = \sigma_\varepsilon^2$. The mean-squared error (MSE) of β can be decomposed as folows. $MSE\left(\widehat{\beta}\right) = \mathbb{E}\left[\left(\widehat{\beta} - \beta\right)\right]^2 = Bias\left(\widehat{\beta}\right)^2 + \mathbb{V}\left(\widehat{\beta}\right)$, where $Bias\left(\widehat{\beta}\right) = \mathbb{E}\left(\widehat{\beta}\right) - \beta$. While the OLS estimator is unbiased, the shrinkage estimator is usually biased. However, its variance is in many cases lower than that of the OLS estimator (in an extreme case, for a random walk forecast without drift, it is 0). As shown by Tibshirani (1996), the MSE for the model forecasts is directly linked to the MSE of the estimator: $MSE\left(y - \widehat{y}\right)^2 = \mathbb{E}\left(y - \widehat{y}\right)^2 = MSE\left(\widehat{\beta}\right) + \sigma_\varepsilon^2$. Hence, the forecasting accuracy can be improved by reducing $MSE\left(\widehat{\beta}\right)$.

3.2 Complete Subset Regressions

For a given set of potential predictor variables, the forecasts from all possible linear regression models that keep the number of predictors fixed, are combined. A complete subset is defined by the set of models that include $k \leq K$ regressors .With K possible predictors, there are K unique univariate models and $n_{k,K} = \binom{K}{k}$ different k-variate models for $k \leq K$. With K regresors in the full model and k regressors chosen for each of the 'short' models, there will be $\binom{K}{k}$ subset regressions to average over within each complete subset, where each regressor in subset k is included a total of $n_{k-1,K-1} = \binom{K-1}{k-1}$ times. To get an insight into how the method provides shrinkage, we outline the mechanism in A.1, closely following the setup of Elliott, Gargano, and Timmermann (2013) and refer to their work for a more detailed presentation.

3.3 State Space Representation

We introduce state space representation for a dynamic linear regression model to accomodate time-varying coefficients. The specified TVP models differ with regard to the included explanatory variables (with 2^K possible combinations) and the values that control the evolution of (possibly) time-varying coefficients. 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 (6) and a system equation (7),

$$
\begin{aligned}
y_t &= F_t^{'}\theta_t + v_t, \qquad v_t \sim N(0, V_t) && (6) \\
\theta_t &= \theta_{t-1} + w_t, \qquad w_t \sim N(0, V_t W_t^*). && (7)
\end{aligned}
$$

The TVP model allows for a time-varying linear relationship between the univariate (scalar) variable y_t (in our case: log exchange rate returns Δs_t) and the vector of the explanatory variables F_t, observed at time $t-1$.[12] $F_t = [1, X_{t-1}]$ is a $r \times 1$ vector of predictors for exchange rates, where $r \leq K$. θ_t is an $r \times 1$ 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 (7) indicates, the evolution of the system coefficients is assumed to follow a random walk, with coefficients being exposed to random shocks w_t.[13]

Adopting a (conditionally) normally distributed prior for the system coefficients and an inverse-gamma distributed prior for the observational variance results in a

[12] We will also consider direct 12-month ahead forecasts in our empirical work. However, to keep our notation simple, we condition on the information set in period $t-1$.

[13] 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 notation is based on West and Harrison (1997). In empirical macroeconomics, there is a widespread consensus to model time-varying parameters as random walks; see, e.g., Cogley and Sargent (2005) or Primiceri (2005).

fully conjugate Bayesian analysis, ensuring that both prior and posterior come from the same family of distributions. The conjugate specification at some arbitrary time t can be expressed as

$$\begin{aligned} V_t|I_t &\sim IG\left[\frac{n_t}{2}, \frac{n_t S_t}{2}\right], \quad (8) \\ \theta_t|I_t &\sim t_{n_t}\left[m_t, S_t C_t^*\right], \quad (9) \\ \theta_t|I_t, V_t &\sim N\left[m_t, V_t C_t^*\right]. \quad (10) \end{aligned}$$

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. To initialize the sequential prediction and updating process, we have to specify m_0, C_0 and S_0.[14] 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.2, 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 and their associated scale matrix.

We adopt a discount factor approach for modelling the unknown sequence for W_t. Consider the transition from the posterior time $t-1$ estimate for the scale

[14] In our empirical work, we use the empirical variance of the log exchange rate returns from the "burn-in" period to determine S_0 and choose $n_0 = 5$ to to express our initial uncertainty about the observational variance. For a model with k regressors, we set $m_0 = 0_{k\times 1}$, $C_0 = g \cdot I_k$ with $g = 10$. Thus we center the initial values for the system coefficients around zero, surrounded by a high degree of uncertainty. This diffuse prior allows for data patterns to be quickly adapted at the beginning of the estimation. The results are qualitatively unaffected by alternative choices for g, n_0 and S_0.

matrix of coefficients (C_{t-1}) to the time t prior for the scale matrix of coefficients (R_t),

$$R_t = C_{t-1} + W_t. \tag{11}$$

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, the discount approach involves replacing W_t by

$$W_t = \frac{1-\delta}{\delta} C_{t-1}, 0 < \delta \leq 1, \tag{12}$$

and, hence,

$$R_t = \frac{1}{\delta} C_{t-1}. \tag{13}$$

δ is a discount factor providing that observations s periods in the past have weight δ^s, implying an age-weighted estimation with an effective window size of $(1-\delta)^{-1}$; see Hannan, McDougall, and Poskitt (1989).[15] For $\delta = 1$, the case of constant parameters is included,[16] $\delta < 1$ 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 work, we will consider a grid of values for $\delta \in \{\delta_1, ..., \delta_d\}$ to allow for different degrees of parameter instability. Concretely, we will consider a grid cov-

[15] The discounting/forgetting approach is well established in the state space literature, see West and Harrison (1997)

[16] In this case, the diagonal elements in R_t and C_t will converge to 0 as t increases, ruling out any uncertainty about the value of the cofficients. To see this, consider that in Equation (35) both R_t and $A_t A_t^{'} Q_t$ are positive (semi-)definite and $S_t = S$ for increasing t, since $n_t \rightarrow \infty$; see Equations (31) and (32).

ering $\delta \in \{0.96; 0.97; 0.98; 0.99; 1\}$.[17] Notice, however, that δ is fixed within each individual model. The data support for different degrees of parameter instability is displayed at the level of the multimodel forecast, reflecting the data support across models with different values of δ at each point in time.

3.4 Flexible Model Averaging and Selection

There are $d \cdot \binom{K}{k}$ individual models in a typical subset k. The large set of models at disposal raises the issue of an appropriate model averaging or selection scheme. Elliott, Gargano, and Timmermann (2013) propose to assign equal weights to the models, argumenting that this simple weighting scheme has turned out to be difficult to be beaten by more flexible weighting schemes. However, in our considered model universe that includes also time-varying parameter models, simply averaging the models can lead to poor forecasting results if a large part of the model pool is inappropriate. Suppose, for example, constant coefficient models are appropriate (over the entire period or at a certain point in time). Then, with $\delta \in \{0.96; 0.97; 0.98; 0.99; 1\}$, constant parameter models make up only $\frac{1}{5}$ of the model pool in each subset. With equal weighting, there would be no mechanism to control for this issue. For this reason, we search for model averaging/selection procedures that choose (temporarily) appropriate models in a data-adaptive fashion. We will introduce two methods, Bayesian Dynamic Model Averaging (BDMA) and Bayesian Dynamic Model Selection (BDMS). The BDMA approach nests classical BMA and equal weighting as special cases. With forecasting densities of the

[17] The boundaries are set based on the following considerations. As we want to allow for time-varying coefficients rather than impose them, constant coefficients ought to be included in the model. The lower boundary is set to 0.96 since we also want to include the possibility of very unstable coefficients.

models provided by the estimation of the state space models, our combination procedures exploits the models' log predictive likelihoods for model combination rather than measures of point forecasting accuracy.

3.4.1 Bayesian Dynamic Model Averaging

Our BDMA approach draws on insights from Dynamic Model Averaging (DMA) proposed by Raftery, Kárný, and Ettler (2010).[18] DMA employs exponential discounting in the weight dynamics according to the past forecast performance of the individual models, thus allowing recent data to be emphasized.[19] 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 put another way: the prior weight for time t). α is a discount factor, $0 \leq \alpha \leq 1$, and shrinks the posterior model weights towards equal weights,

$$\mathcal{P}\left(M_i|I_{t-1}\right) = \frac{p\left(M_i|I_{t-1}\right)^{\alpha}}{\sum\limits_{j=1}^{J} p\left(M_j|I_{t-1}\right)^{\alpha}}. \tag{14}$$

Updating model weights is accomplished by using Bayes' rule,

[18] See Koop and Korobilis (2012) for an application to inflation forecasting.

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

$$p\left(M_i|I_t\right) = \frac{p\left(y_t|M_i, I_{t-1}\right)\mathcal{P}\left(M_i|I_{t-1}\right)}{\sum\limits_{j=1}^{J} p\left(y_t|M_j, I_{t-1}\right)\mathcal{P}\left(M_j|I_{t-1}\right)}, \tag{15}$$

where the predictive likelihood of model i,

$$p\left(y_t|M_i, I_{t-1}\right) \sim \frac{1}{\sqrt{Q_{i,t}}} t_{n_{t-1},i}\left(\frac{y_t - \widehat{y}_{i,t}}{\sqrt{Q_{i,t}}}\right), \tag{16}$$

is used to a assess the forecasting 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 $n_{t-1,i}$ 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. Obviously, for $\alpha = 0$, all models are equally weighted, while for $\alpha = 1$, there is no discounting and, hence, BMA is recovered as a special case.[20] 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

$$\begin{aligned} \mathcal{P}\left(M_i|I_{t-1}\right) &\propto \left[\mathcal{P}\left(M_i|I_{t-2}\right) p\left(y_{t-2}|M_i, I_{t-2}\right)\right]^{\alpha} \\ &= \prod_{s=1}^{t-1} p\left(y_{t-s}|M_i, I_{t-s-1}\right)^{\alpha^s}. \end{aligned} \tag{17}$$

[20] In this case, the predictive likelihoods are identical to the marginal likelihoods; see Raftery, Kárný, and Ettler (2010).

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 (sum) out the associated uncertainty. Initializing the process of model combinations involves specifying priors on model weights, $p\left(M_i|I_0\right)$, $\forall i = 1, ..., J$.[21] To obtain predictive weights, we use an equation similar to (14), but in contrast to (14), we sum over the discrete set of considered grid points for α.

$$\mathcal{P}\left(M_i|I_{t-1}\right) = \sum_{v=1}^{a} \frac{\overbrace{p\left(M_i|I_{t-1}\right)^{\alpha_v}}^{:=\mathcal{P}(M_i|I_{t-1},\alpha_v)}}{\sum\limits_{j=1}^{J} p\left(M_j|I_{t-1}\right)^{\alpha_v}} \cdot p\left(\alpha_v|I_{t-1}\right). \tag{18}$$

$p\left(M_i|I_{t-1}\right)$ 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.[22] We will consider the grid $\alpha \in \{0; 0.80; 0.90; 0.95; 0.99; 1\}$. The updating step for model weights is accomplished by

$$p\left(M_i|I_t\right) = \sum_{v=1}^{a} \frac{p\left(y_t|M_i, I_{t-1}\right)\mathcal{P}\left(M_i|I_{t-1}, \alpha_v\right)}{\sum\limits_{j=1}^{J} p\left(y_t|M_j, I_{t-1}\right)\mathcal{P}\left(M_j|I_{t-1}, \alpha_v\right)} \cdot p\left(\alpha_v|I_t\right). \tag{19}$$

[21] In our empirical work, we initially assign equal weights to each model configuration, that is $p\left(M_i|I_0\right) = \frac{1}{d \cdot \binom{K}{k}}, \forall i = 1, ..., J$.

[22] In our empirical work, we assign equal weights to each considered grid point for α, i.e. $p\left(\alpha_z|I_0\right) = \frac{1}{a}, z = 1, ...a$.

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

$$p\left(\alpha_z|I_t\right) = \frac{\sum_{j=1}^{J} p\left(y_t|M_j, I_{t-1}\right) \mathcal{P}\left(M_j|I_{t-1}, \alpha_z\right) p\left(\alpha_z|I_{t-1}\right)}{\sum_{v=1}^{a}\sum_{j=1}^{J} p\left(y_t|M_j, I_{t-1}\right) \mathcal{P}\left(M_j|I_{t-1}, \alpha_v\right) p\left(\alpha_v|I_{t-1}\right)}, \forall z = 1, .., a, \quad (20)$$

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 calm 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 in Pesaran and Timmermann

(2007), Pesaran and Pick (2011) and Pesaran, Pick, and Pranovich (2013).[23]

3.4.2 Bayesian Dynamic Model Selection

In contrast to BDMA, BDMS chooses a single model within each subset to provide the subset forecast rather than average over individual models. At each point in time, the model with the (currently) highest log predictive score is used as the subset forecast, the remaining models are (temporarily) removed.

3.5 Optimal Prediction Pools

In the previous section, we have addressed the issue of how to combine or select models within a subset k. In the following section, we focus on combining the aggregate density forecasts of the K subsets. To this end, we employ optimal prediction pools (Geweke and Amisano, 2011). The method combines models so as to maximize the log predictive score.and has several attractive theoretical properties such as it does not assume that the true model is included in the specified model set.

Let $p\left(y_t|I_{t-1},\mathcal{M}_k\right)$ denote the (combined) density forecast of subset $k \leq K$ for period t. In period $t-1$, the aggregate predictive density for period t using the linear prediction pool is

$$p\left(y|I_{t-1}\right)=\sum_{k=0}^{K} w_{k,t-1} p\left(y_t|I_{t-1},\mathcal{M}_k\right). \tag{21}$$

[23] Averaging information across window sizes has turned out as successful in many instances; see Rossi (2013).

The weight vector $w_{t-1} = (w_{0,t-1}, ..., w_{K,t-1})$ satsisfies $\sum_{k=0}^{K} w_{k,t} = 1$ and $w_{k,t} \geq 0$, $\forall k \leq K$. The optimal weight vector $w^{*}_{k,t-1}$ is to be chosen such as to maximize the log predictive score up to period $t-1$:

$$f\left(w_{t-1}|I_{t-1}\right) = \sum_{s=1}^{t-1} \ln \sum_{k=0}^{K} w_{k,t-1} p\left(y_s|I_{s-1}, \mathcal{M}_k\right). \tag{22}$$

3.6 The Role of Shrinkage

Our forecasting method can be considered as a hierarchical combination of TVP models. To explore how the shrinkage works, we decompose the point forecast of (an arbitrary) period t as

$$\widehat{y_t}|I_{t-1} = \sum_{k=0}^{K} w^{*}_{k,t-1} \circ \left(\widehat{\Upsilon}^{agg}_{k,t-1}\right)' \cdot \digamma_t, \tag{23}$$

where

$$\widehat{\Upsilon}^{agg}_{k,t-1} = \sum_{j=1}^{d\cdot\binom{K}{k}} \widehat{\Upsilon}_{j,t-1} \mathcal{P}\left(M_j|I_{t-1}\right). \tag{24}$$

$\digamma_t$ denotes the full $K \times 1$ vector of regressors in period t. Following the notation in Section 3.3, $\digamma_t = [1, X_{t-1}]$ and $k = K$. $\widehat{\Upsilon}_{j,t-1}$ denotes the updated full $K \times 1$ coefficient vector in period $t-1$ for model j. For example, if the individual model j includes the regressors $1, 3$ and 5, but not the regressor 2 and 4, the estimated coefficient $m_{j,t-1}$ (updated in period $t-1$) is an $r \times 1$ vector with $r = 3$. In the full coefficient vector $\widehat{\Upsilon}_{j,t-1}$, all entries for which the associated regressor is excluded from the model, are filled up with zeros. In the particular example case, we have $\widehat{\Upsilon}'_{j,t-1} = \left(m^{(1)}_{j,t-1}, 0, m^{(3)}_{j,t-1}, 0, m^{(5)}_{j,t-1}\right)$. $m^{(i)}_{j,t-1}$ denotes the estimated coefficient associated with the i-th regressor of $\digamma_t$. $\widehat{\Upsilon}^{agg}_{k,t-1}$ denominates the aggregate coefficient vector for subset k. $\circ$ indicates the Hadamard product.

Equations 23 and 24 illustrate that the final model forecast can be decomposed into a linear combination of the estimated model coefficients and the regressors. This renders the method transparent. The shrinkage intensity may evolve over time. Suppose, for example, none of the regressors is important at a certain point around period t. In this case, the weight $w_{0,t-1}$ attached to the random walk forecast is expected to be high. In the extreme case, $w_{0,t-1} = 1$ and, hence, the final model forecast for period t is 0. If, however, around another period, some regressors become important, the weight attached to the random walk forecast is expected to decrease. It is worth to note that by pooling the aggregate subset density forecasts, models with different complexity have equal chances to turn out as important. The subset $k = 0$ contains only $\binom{K}{0} = 1$ model, i.e., the random walk forecast, while the subset $k = 3$ comprises $5 \cdot \binom{5}{3} = 50$ models for $d = 5$ and $K = 5$. However, each subset k provides only one aggregate forecast density $p\left(y_s | I_{s-1}, \mathcal{M}_k\right)$ in period s.[24] The OLS estimate is, of course, also recovered as a special case. In this case, $w_{K,t-1} = 1$, $\mathcal{P}\left(M_j | I_{t-1}\right) = 1$ for the model which includes all K regressors and assumes constant coefficients.

4 Monte Carlo Simulation

We consider a small Monte Carlo simulation to assess our method's ability to recover the generating data mechanism. Particularly, we are interested if (a flex-

[24] In contrast, in a BMA (DMA) weighting scheme over the entire model pool, the random walk model would be assigned only $\frac{1}{10}$ of the weight of the subset with 3 regressors (for $K = 5$ and $d = 5$), if all the models have equal marginal (predictive) likelihoods. Hence, BMA (DMA) automatically disadvantages very sparse models. Addressing this issue by assigning high a priori weights to very sparse models is overturned after few periods in the updating process.

ible version of) our approach manages to rapidly adjust to gradually or abruptly changing coefficients. For the setup of the Monte Carlo study, we assume a TVP model of the form:

$$Y_{t+1} = \sum_{i=1}^{4} \theta_{i,t} x_{i,t} + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim N\left(0, \sigma_{\varepsilon}^{2}\right).$$

We set $\sigma_{\varepsilon}^{2} = 0.01$, $t = 1, ..., 500$, and consider the following structure for the evolution of the coefficients:

$$\begin{aligned}
\theta_{1,t} &= \begin{cases} -0.2 & , 120 < t < 300 \\ 0.6 & , otherwise \end{cases}, \\
\theta_{2,t} &= \begin{cases} 0 & , t < 120 \\ -5 \times 10^{-4} \cdot t & , otherwise \end{cases}, \\
\theta_{3,t} &= \begin{cases} 0.8 - 0.2 \cdot t/120 & , t < 120 \\ 0.4 + 0.2 \cdot t/120 & , 120 < t < 300 \\ 0 & , otherwise \end{cases}, \\
\theta_{4,t} &= \begin{cases} 0.8 & , t > 300 \\ 0 & , otherwise \end{cases}.
\end{aligned}$$

Figure 1 presents the results for three different model settings. In the most restrictive setting, we set $\delta = 1$, $\alpha = 1$, and hence, do not allow for time-varying coefficients. The graph clearly indicates that constant coefficients fail in picking up the changes in the coefficients. The setting that allows for time-varying coefficients and combines the models via BMA within the subsets, $\delta \in$

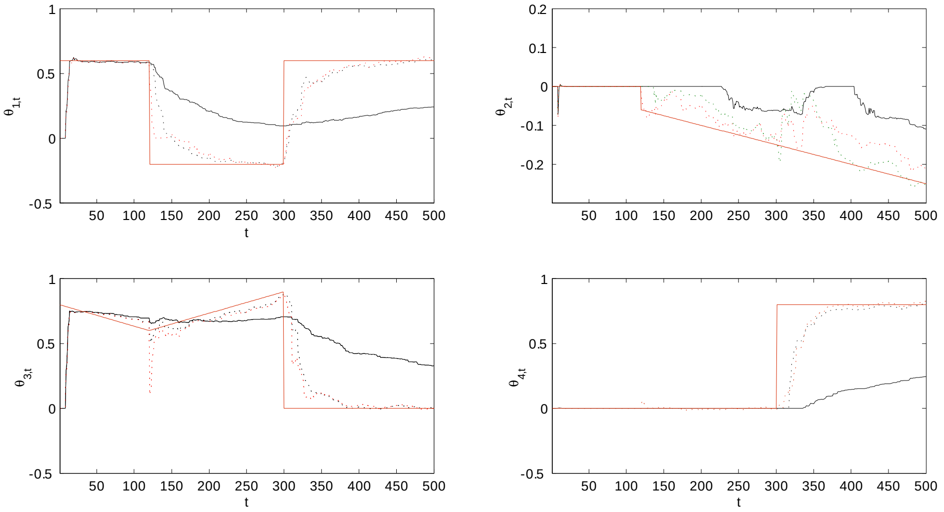

Figure 1: Evolution of coefficients. The figure presents the evolution of the coefficients in the simulation experiment. The solid red line indicates the true evolution of the coefficient. The solid black line shows the evolution of the estimated coefficients under the most restrictive setting ($\delta = 1$, $\alpha = 1$). The dotted black line indicates the behavior of the estimated coefficients under Bayesian Model Averaging of TVP models, $\delta \in \{0.96; 0.97; 0.98; 0.99; 1\}$ and $\alpha = 1$. The dotted red line displays the evolution of the estimated coefficients under the most flexible model setting with $\delta \in \{0.96; 0.97; 0.98; 0.99; 1\}$ and $\alpha \in \{0; 0.80; 0.90; 0.95; 0.99; 1\}$.

$\{0.96; 0.97; 0.98; 0.99; 1\}$, $\alpha = 1$, does considerably better in tracking both gradual as well as abrupt coefficient changes. We observe only marginal performance improvement for our most flexible setting with time-varying coefficients and BDMA weights, $\delta \in \{0.96; 0.97; 0.98; 0.99; 1\}$, $\alpha \in \{0; 0.80; 0.90; 0.95; 0.99; 1\}$. For some cases, this version is slightly less sluggish in adapting to parameter changes. However, this effect becomes more pronounced if the variance of the data generating process is increased. We next turn to our empirical work, reporting and discussing the forecasting results.

5 Empirical Analysis

Our forecasting exercise comprises the following currencies: the British pound (GBP), Japanese yen (JPY), German mark/euro (DEM), Canadian dollar (CAD), Swiss franc (SWF) and the Australian dollar (AUD). The (monthly) data cover the period from 1975 : 03 to 2013 : 06. We report forecasting results after a "burn-in" period from 1985 : 04 to 2013 : 06 for a range of model specifications.

Table 1 shows the results for point prediction accuracy in terms of the out-of-sample R^2 (R$^2_{OOS}$) proposed by Campbell and Thompson (2008) for one-month ahead forecasts.[25] Table 2 shows the forecasting results for direct twelve-month

[25] The out-of sample R^2 is calculated as

$$R^2_{OOS} = 1 - \frac{\sum\limits_{t=\tau+1}^{T-1} (y_{t+1} - \widehat{y}_{t+1}|I_t)^2}{\sum\limits_{t=\tau+1}^{T-1} \left(y_{t+1} - \widehat{bm}_{t+1}|I_t\right)^2},$$

where τ denotes the "burn-in" sample, $\widehat{y}_{t+1}|I_t$ refers to the point forecast of the respective model configuration, and $\widehat{bm}_{t+1}|I_t$ to the point forecast of the benchmark model. As we use the ahead forecasts. Our set of model configurations is divided into two main groups, the constant parameter models ($\delta = 1$) and the TVP models ($\delta \in \{0.96; 0.97; 0.98; 0.99; 1\}$). The kitchen sink specification considers only the subset that includes all K regressors. That is, $w_{K,t} = 1$, $\forall t$. The *Subset-Regression-EW* specification employs the complete subset regressions approach with equal weighting of the models within the subsets (as suggested by Elliott, Gargano, and Timmermann (2013)). However, to combine the models across subsets, optimal prediction pools are employed.[26] The *Subset-Regression-BDMA* and *Subset-Regression-BDMS* model configurations (outlined in Section 3.4) allow for flexible weighting schemes within the subsets. We will comment on the results, supported by some graphical devices, in the context of the key questions we have raised at the beginning. All graphical devices are based on our baseline results, the one-month ahead forecasting configuration.

Which set of macro fundamentals (if any) is relevant for forecasting at each point in time? Figure 2 shows the inclusion probabilities for the regressors over time. The inclusion probabilities are simply calculated as summing over the predictive model probabilities that include a particular regressor i, that is $\sum_{k=0}^{K} \sum_{j=1}^{d \cdot \binom{K}{k}} \mathcal{P}\left(M_j | I_{t-1}\right) \cdot \mathbb{I}_{\{i \in M_j\}}$ for period t. The regressor *AsyTaylor* receives relatively constant support over time for the GBP and JPY. For the DEM, the *PPP* regressor gains some importance after the Subprime Crisis, while the other regres-

random walk without drift as the benchmark model, $\widehat{bm}_{t+1}|I_t$ is always 0. The random walk without drift is known as the toughest bennchmark in the exchange rate forecasting literature; see Rossi (2013). If we use the random walk with drift as benchmark model, i.e., the prevailing unconditional mean, our results are revolved around: for this case, our flexible model configurations regularly and (to a large extent significantly) outperform the the benchmark in terms of R^2_{OOS}. Results are omitted but are available upon request.

[26] We use optimal prediction pools to provide comparability to the other model configurations. However, if we recursively select the hyperparameter k (that is, choose the subset k which would have given the best forecasting performance in terms of the MSE), our results are generally slightly worse. Results are available upon request.

Table 1: Prediction accuracy of one-step ahead forecasts.
R^2_{OOS} measures the percentage reduction in mean squared prediction error (MSPE) based on the forecast of the respective model relative to the random walk benchmark forecast. Statistical significance is assessed by the Clark and West (2007) test. a,b,c indicate significance at the 10%, 5% and 1% level, respectively, that the random walk MSPE is less or equal to the respective predictive model's MSPE against the alternative that the random walk MSPE is greater than the predictive model's MSPE. R^2_{OOS} statistics are computed for the $1985:04-2013:06$ forecast evaluation period.

Model configuration	$\mathrm{R}^2_{OOS}\%$					
	GBP	*JPY*	*DEM*	*CAD*	*SWF*	*AUD*
Constant Parameter Models						
Kitchen Sink	-2.55	-1.47	-4.08	-0.81	-1.09	-3.02
Subset Regressions - EW	-0.13	0.62	0.00	-0.01	0.35	-0.04
Subset Regressions - BDMA	-0.17	0.38^b	0.00	0.00	0.48^a	-0.03
Subset Regressions - BDMS	-0.14	0.73^b	0.00	-0.04	0.88^b	0.00
TVP Models						
Kitchen Sink	-6.02	-0.27	-4.17	-9.36	-8.84	-10.01
Subset Regressions - EW	-0.23	0.51	0.04^a	-0.06	0.09	-0.04
Subset Regressions - BDMA	-0.07	0.48	0.06^b	-0.23	-0.05	-0.13
Subset Regressions - BDMS	0.20	0.51	0.15	-0.56	-0.43	-0.09

sors are essentially removed from the aggregate model. Also, for the AUD none of the regressors is included. For the CAD, the *UIP* and *Monetary* regressors gain importance after the Subprime Crisis. For the remaining exchange rates, the *Monetary* regressor turns out as unneccessary for short-term forecasting.[27] We observe an interesting pattern for the SWF. The *UIP* and *PPP* regressors display rapidly changing inclusion probabilities that seem to move in opposite directions from the middle of the sample. This pattern illustrates the flexibility embedded in our approach, allowing the weights attached to fundamentals to change abruptly if required by the data. Bottom line, none of the regressor seems to be important for forecasting across all countries.

[27] This finding is in line with Engel, Mark, and West (2008).

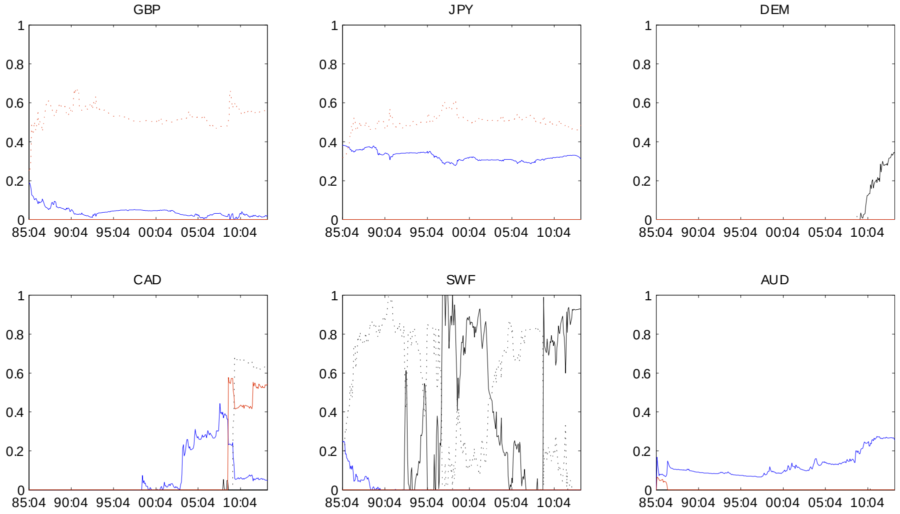

Figure 2: Evolution of inclusion probabilities for the regressors of the *TVP-Subset Regressions-BDMA* model configuration. The dotted red (black) line indicates the evolution for the inclusion probability of the *AsyTaylor* (*UIP*) regressor. The solid black (red) line shows the evolution of the inclusion probability of the *PPP* (*Monetary*) regressor. The solid blue line tracks the inclusion of the intercept.

Table 2: Prediction accuracy of twelve-step ahead forecasts.
R^2_{OOS} measures the percentage reduction in mean squared prediction error (MSPE) based on the forecast of the respective model relative to the random walk benchmark forecast. Statistical significance is assessed by the Clark and West (2007) test. a,b,c indicate significance at the 10%, 5% and 1% level, respectively, that the random walk MSPE is less or equal to the respective predictive model's MSPE against the alternative that the random walk MSPE is greater than the predictive model's MSPE. R^2_{OOS} statistics are computed for the 1985 : 04 − 2013 : 06 forecast evaluation period.

Model configuration	R^2_{OOS}%					
	GBP	*JPY*	*DEM*	*CAD*	*SWF*	*AUD*
Constant Parameter Models						
Kitchen Sink	−108.92	−14.87	−83.71	−30.60	−38.69	−68.93
Subset Regressions - EW	0.00	−0.60	0.00	−1.02	0.00	0.00
Subset Regressions - BDMA	0.00	−0.15	0.00	−0.04	0.00	0.00
Subset Regressions - BDMS	0.00	−0.04	0.00	0.00	0.34	0.00
TVP Models						
Kitchen Sink	−107.95	−13.80	−83.68	−30.67	−37.00	−55.73
Subset Regressions - EW	0.00	−0.62	0.00	−0.92	0.00	0.00
Subset Regressions - BDMA	0.00	−0.26	0.00	−0.02	0.00	0.00
Subset Regressions - BDMS	0.00	0.83	0.00	0.00	0.00	0.00

Are time-varying coefficients helpful? For the *Kitchen Sink* model configurations, the increased flexibility of the TVP has a negative effect on forecasting performance for the one-step ahead forecasts (except for the JPY), while the effect on the twelve-step ahead forecasts is ambiguous. For the configurations that use shrinkage and model averaging/selection, results are mixed and no clear pattern arises whether the TVP models outperform their constant counterparts.

Is it worthwile to consider flexible model averaging/selection criteria? The BDMA and BDMS model averaging/selection schemes provide, is sum, slightly better results than the equal-weighting benchmark.

How intensively are the forecasts shrunk towards zero? Figure 3 displays the results for the *Kitchen Sink* approach without shrinkage, Figure 4 corresponds to the final results after shrinkage based on the flexible *TVP-Subset Regression-BDMA* configuration. A comparison readily shows that the coefficients display significantly less variation after shrinkage has been introduced. However, abrupt changes in coefficients can still be observed, in particular around the end of the sample after the emergence of the Subprime Crisis. Figure 5 shows the weights assigned to the random walk forecast over time. If the weight equals one, the final forecast of the model is 0 and, hence, the shrinkage intensity is maximal. Overall, the shrinkage intensity is high (except for the JPY, it is always above 50%). For the DEM, the shrinkage intensity is maximal until near the end of the sample. Figure 5 is intimately related to Figure 2. The inclusion probabilities of all macro fundamentals are close to zero for most of the time. Near the end of the sample, the *PPP* regressor gains importance and is assigned an increasing weight at the expense of the random walk model. The shrinkage intensity for the CAD also changes over time, while it is roughly constant for the remaining countries. An interesting piece of evidence is that the previous shown flexibility of our model pays off in terms of forecasting performance for the JPY and SWF since some flexible model configurations do comparatively well for both countries. This suggests that our approach is able to detect (temporarily) relevant information embedded in the macro fundamentals. Figure 6 shows the MSE of the aggregate subset forecasts as a function of the number of predictors that are included. For all six currencies, the MSE increases when more than three predictors are included. This highlights the issue of overfitting and the superiority of parsimonious models. As our model combination/selection exploits the prediction densities rather the conditional means, Figure 7 is intended to provide information to whether the criteria MSE and log predictive likelihoods favor a similar degree of model complexity. The measures

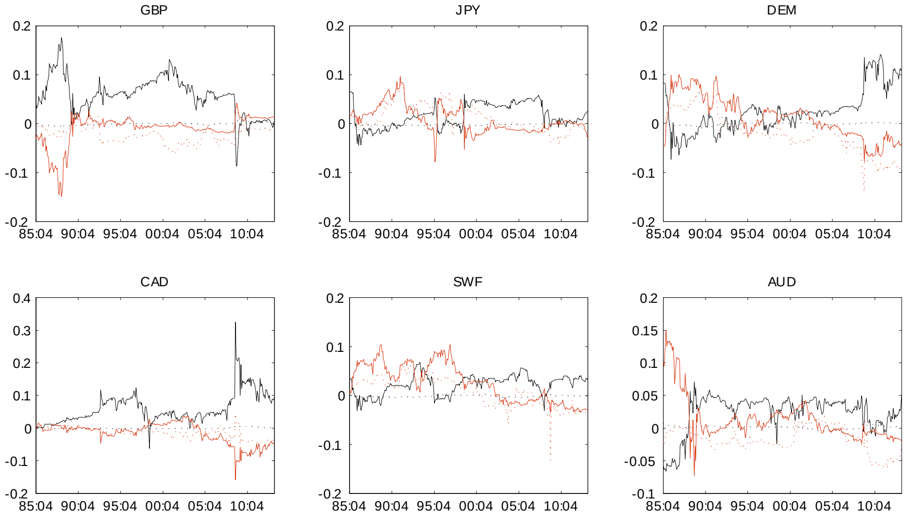

Figure 3: Evolution of coefficients of the *TVP - Kitchen Sink* model configuration. The dotted red (black) line indicates the evolution of the coefficient associated with the *AsyTaylor (UIP)* regressor. The solid black (red) line shows the evolution of the *PPP (Monetary)* regressor.

largely agree, suggesting that our combination scheme is appropriate, although we focus on point forecast accuracy.[28]

Are the flexible models able to outperform the random walk benchmark? The results are mixed for the one-step ahead forecasts. For the twelve-month horizon, none of the model configurations significantly outperforms the random walk forecast. Instead, forecasts of the flexible models are intensively shrunk towards zero. Bottom line, the considered model configurations cannot consistently outperform the random walk benchmark.

As a robustness check, the set of potential regressors has been extended by including several other predictors, namely changes in cumulated trade balances,

[28] We have already mentioned that the R^2_{OOS} generally decreases if we use the recursive subset selection strategy proposed by Elliott, Gargano, and Timmermann (2013) that is based on forecasting accuracy measured by the MSE.

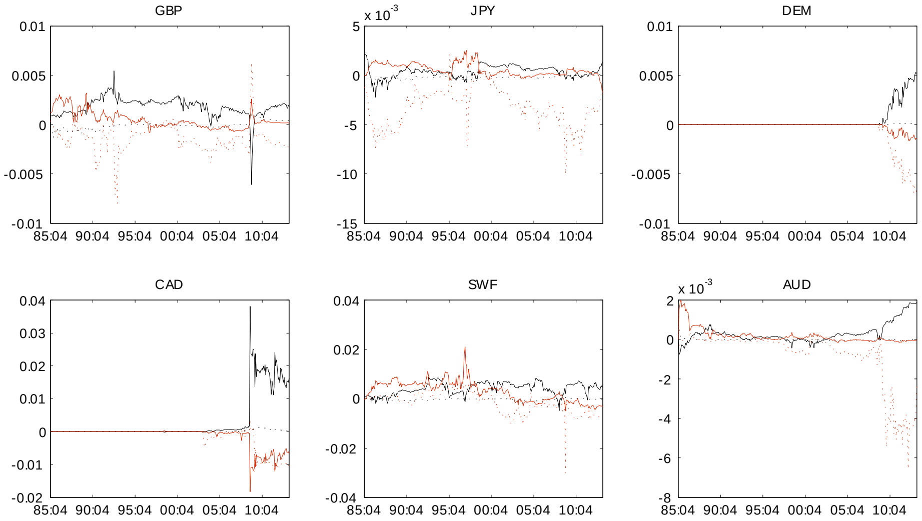

Figure 4: Evolution of coefficients of the *TVP-Subset Regressions-BDMA* model version. The dotted red (black) line indicates the evolution of the coefficient associated with the *AsyTaylor (UIP)* regressor. The solid black (red) line shows the evolution of the *PPP (Monetary)* regressor.

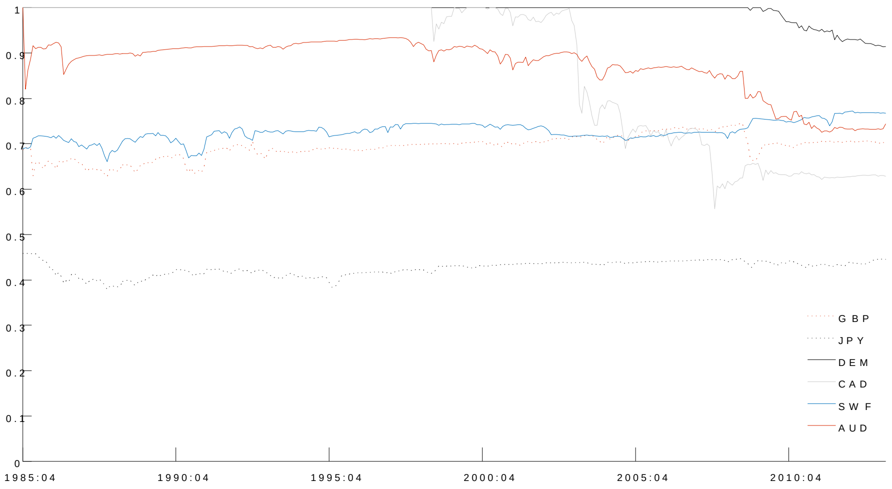

Figure 5: Shrinkage to the random walk. The figure shows the weights attached to the subset with zero regressors (i.e., the random walk) for the considered exchange rates over time.

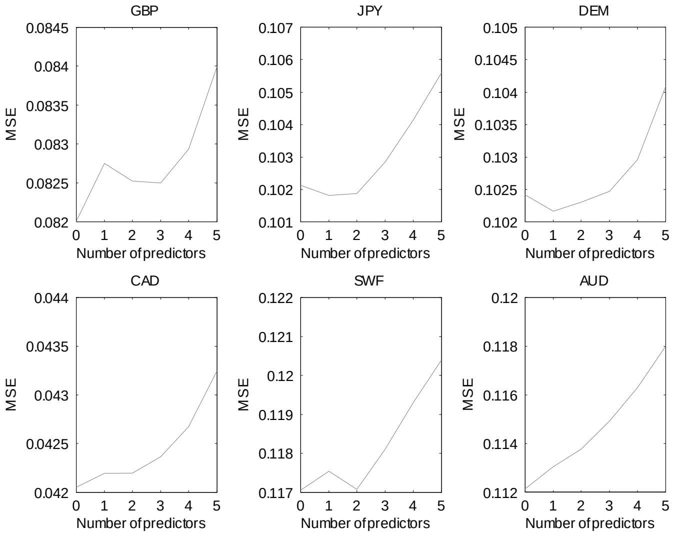

Figure 6: Out-of-sample forecast performance (MSE). The mean squared errors of the *TVP-Subset Regressions-BDMA* model configurations are shown as a function of the number of predictors.

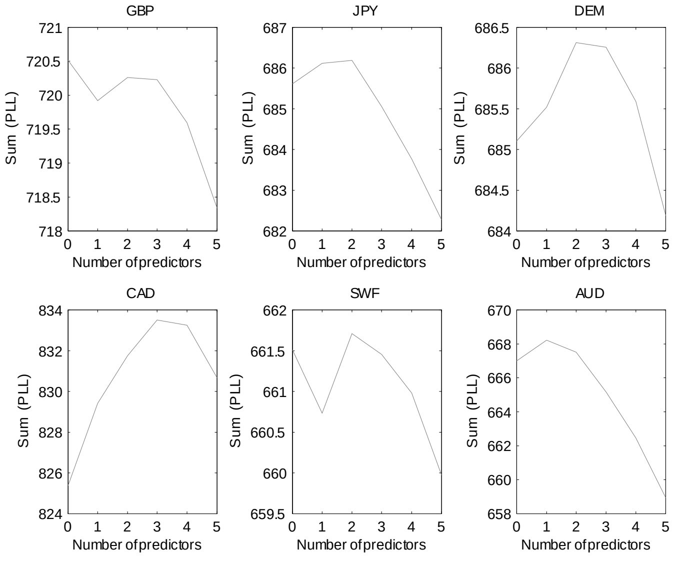

Figure 7: Out-of-sample forecast performance (Predictive Log Likelihoods). The sums of predictive log likelihoods of the *TVP-subset regressions-BDMA* model configurations are shown as a function of the number of predictors.

stock price returns and commodity index returns. We have also analyzed whether including the considered fundamentals directly as regressors rather than pre-estimate models (as outlined in Section 2) changes the results. Furthermore, we have experimented with different sample periods. To conserve space, the results are not displayed since our our main findings are qualitatively unaffected in all cases. The results are available upon request.

6 Conclusion

We have introduced a method for shrinkage in TVP models that is tailored to the econometric demands for exchange rate forecasting models. The method has a range of desirable properties, particularly it allows for time-varying and predictor-specific shrinkage intensity. Our empirical results amass evidence for the prevailing view that (short-term) forecasting of exchange rates with macro fundamentals cannot systematically beat a naive random walk benchmark in terms of point prediction accuracy (measured by the root mean squared error). The naive random walk model is frequently selected as the most appropriate model. However, the empirical findings suggest that changing relevance of nformation embedded in macro fundamentals is detected in the flexible model versions.

The emphasis of our analysis has dealt with designing a flexible econometric forecasting technique. Against this background, we are confident to have provided a suitable setup for conditional predictability, as far as it genuinely exists. While our results are robust to alternative predictors, they do not rule out predictability for other settings, such as real-time (rather than revised) macro fundamentals. As our method is of general purpose (though inspired by the demands for predicting exchange rates), it is suited for predicting other variables in macroeconomics and finance.

Our results affirm the importance for shrinkage and flexible model averaging/selection criteria to avoid poor forecasting performance. The researcher using the approach benefits from the automated shrinkage procedure. (S)he avoids the risk of misspecification associated with a simpler model (such as the random walk) from the outset and, simultaneously, circumvents the caveat of overfitting that is associated with using unrestricted TVP models. Thus the suggested method provides a "failsafe mechanism" against inappropriate model choices.

References

AKRAM, Q. F., D. RIME, AND L. SARNO (2008): "Arbitrage in the foreign exchange market: Turning on the microscope," *Journal of International Economics*, 76(2), 237–253.

BACCHETTA, P., AND E. VAN WINCOOP (2004): "A scapegoat model of exchange rate fluctuations," Discussion paper, National Bureau of Economic Research.

——— (2006): "Can Information Heterogeneity Explain the Exchange Rate Determination Puzzle?," *American Economic Review*, 96(3), 552–576.

——— (2013): "On the unstable relationship between exchange rates and macroeconomic fundamentals," *Journal of International Economics*, 91(1), 18–26.

BELMONTE, M. A., G. KOOP, AND D. KOROBILIS (2014): "Hierarchical Shrinkage in Time-Varying Parameter Models," *Journal of Forecasting*, 33(1), 80–94.

BERGE, T. J. (2014): "Forecasting disconnected exchange rates," *Journal of Applied Econometrics*, 29(5), 713–735.

BREIMAN, L. (1996): "Bagging predictors," *Machine learning*, 24(2), 123–140.

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.

CHAN, J. C., G. KOOP, R. LEON-GONZALEZ, AND R. W. STRACHAN (2012): "Time varying dimension models," *Journal of Business & Economic Statistics*, 30(3), 358–367.

Cheung, Y.-W., and M. D. Chinn (2001): "Currency traders and exchange rate dynamics: a survey of the US market," *Journal of International Money and Finance*, 20(4), 439–471.

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.

Cogley, T., and T. J. Sargent (2005): "Drifts and volatilities: monetary policies and outcomes in the post WWII US," *Review of Economic dynamics*, 8(2), 262–302.

Corte, P. D., L. Sarno, and I. Tsiakas (2009): "An Economic Evaluation of Empirical Exchange Rate Models," *Review of Financial Studies*, 22(9), 3491–3530.

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

Engel, C., N. C. Mark, and K. D. West (2008): "Exchange Rate Models Are Not As Bad As You Think," in *NBER Macroeconomics Annual 2007, Volume 22*, NBER Chapters, pp. 381–441. National Bureau of Economic Research, Inc.

Fratzscher, M., L. Sarno, and G. Zinna (2012): "The scapegoat theory of exchange rates: the first tests," Working Paper Series 1418, European Central Bank.

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

Groen, J. J., R. Paap, and F. Ravazzolo (2013): "Real-time inflation forecasting in a changing world," *Journal of Business & Economic Statistics*, 31(1), 29–44.

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.

Hodrick, R. J., and E. C. Prescott (1997): "Postwar US business cycles: an empirical investigation," *Journal of Money, credit, and Banking*, pp. 1–16.

Kalli, M., and J. E. Griffin (2014): "Time-varying sparsity in dynamic regression models," *Journal of Econometrics*, 178(2), 779–793.

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

Kouwenberg, R., A. Markiewicz, R. Verhoeks, and R. C. Zwinkels (2013): "Model Uncertainty and Exchange Rate Forecasting," *Available at SSRN 2291394*.

Li, J., I. Tsiakas, and W. Wang (2014): "Predicting Exchange Rates Out of Sample: Can Economic Fundamentals Beat the Random Walk?," *Journal of Financial Econometrics*.

Markiewicz, A. (2012): "Model uncertainty and exchange rate volatility," *International Economic Review*, 53(3), 815–844.

Meese, R. A., and K. Rogoff (1983): "Empirical exchange rate models of the seventies : Do they fit out of sample?," *Journal of International Economics*, 14(1-2), 3–24.

Molodtsova, T., and D. H. Papell (2009): "Out-of-sample exchange rate predictability with Taylor rule fundamentals," *Journal of International Economics*, 77(2), 167–180.

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

Pesaran, M. H., A. Pick, and M. Pranovich (2013): "Optimal forecasts in the presence of structural breaks," *Journal of Econometrics*, 177(2), 134–152.

Pesaran, M. H., and A. Timmermann (2007): "Selection of estimation window in the presence of breaks," *Journal of Econometrics*, 137(1), 134–161.

Primiceri, G. E. (2005): "Time varying structural vector autoregressions and monetary policy," *The Review of Economic Studies*, 72(3), 821–852.

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. 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.

Rossi, B. (2013): "Exchange rate predictability," *Journal of Economic Literature*, 51(4), 1063–1119.

Rossi, B., G. Elliott, and A. Timmermann (2012): "Advances in forecasting under instability," *by G. Elliott, and A. Timmermann. Elsevier-North Holland.*

Sarno, L., and G. Valente (2009): "Exchange Rates and Fundamentals: Footloose or Evolving Relationship?," *Journal of the European Economic Association*, 7(4), 786–830.

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.

Taylor, J. B. (1993): "Discretion versus policy rules in practice," in *Carnegie-Rochester conference series on public policy*, vol. 39, pp. 195–214. Elsevier.

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

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

Wright, J. H. (2008): "Bayesian Model Averaging and exchange rate forecasts," *Journal of Econometrics*, 146(2), 329–341.

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

A Appendix

A.1 The Shrinkage Mechanism in Complete Subset Regressions

Suppose we are interested in predicting the univariate (scalar) variable y_{T+1} in a simple linear regression model with k predictors $x_T \in \mathbb{R}^k$, and a history of data, $\{y_{t+1}, x_t\}_{t=0}^{T-1}$. Let $\mathbb{E}(x_t' x_t) = \Sigma_X$ for all t, and without loss of generality, assume that $\mathbb{E}(x_t) = 0$ for all t. To focus on regressions that include only a subset of the predictors, define β to be a $K \times 1$ vector with coefficients in the rows representing included regressors and zeros in the rows of excluded variables. $y = (y_1, ..., y_T)$ is a $T \times 1$ vector and $X = (x_0, x_1 ..., x_{T-1})'$ stacks the x observations into a $T \times K$ matrix. Denote the generalized inverse of a matrix A by A^-. Let S_i be a $K \times K$ matrix with zeros everywhere except for ones in the diagonal cells corresponding to included variables, such that if the (j, j) element of S_i is one, the jth regressor is included, while if this element is zero, the jth regressor is excluded. Sums over i are sums over all permutations of S_i. The subset regression estimators can be represented as a weighted average of the components of the full regression OLS estimator, $\widehat{\beta}_{OLS}$. Elliott, Gargano, and Timmermann (2013) show that, for a large sample size and under general conditions, the estimator for the complete subset regression, $\widehat{\beta}_{k,K}$, can be written as

$$\widehat{\beta}_{k,K} = \Lambda_{k,K} \widehat{\beta}_{OLS} + o_p\left(1\right),$$

where

$$\Lambda_{k,K} = \frac{1}{n_{k,K}} = \sum_{i=1}^{n_{k,K}} \left(S_i' \Sigma_X S_i \right)^- \left(S_i' \Sigma_X \right).$$

To gain insight into how the method works as a shrinkage estimator, we will first focus on the special case when the covariates are orthonormal. In this case, $\widehat{\beta}_{k,K} = \lambda_{k,K} \widehat{\beta}_{OLS}$, where $\lambda_{k,K} = 1 - \left(\frac{n_{k,K-1}}{n_{k,K}} \right)$ is a scalar. To see this, note that for this special case, $\widehat{\beta}_{OLS} = X'y$, while each of the subset regression estimates can be written $\widehat{\beta}_i = S_i X' y$. The complete subset regression estimator is then given by

$$\begin{aligned}
\widehat{\beta}_{k,K} &= \frac{1}{n_{k,K}} \sum_{i=1}^{n_{k,K}} \widehat{\beta}_i \\
&= \frac{1}{n_{k,K}} \sum_{i=1}^{n_{k,K}} S_i X' y \\
&= \left(\frac{1}{n_{k,K}} \sum_{i=1}^{n_{k,K}} S_i \right) \widehat{\beta}_{OLS}.
\end{aligned}$$

The result follows by noting that the elements of $\sum_{i=1}^{n_{k,K}} S_i$ are zero for the off-diagonal terms, and equal the number of times the regressor is included in the subset regressions for the diagonal terms, that is $\binom{n_{K-1}}{n_{k-1}}$ times. In turn, the diagonal terms equal $n_{k,K}$ minus the number of times a regressor is excluded, which gives the result, noting that the solution is the same for each diagonal. The smaller k relative to K, the greater the amount of shrinkage.

For the general case, where regressors are correlated, the subset regression coefficients are not simple regressor-by-regressor shrinkages of the OLS estimates, and will depend on the full covariance matrix of all regressors. Specifically, $\Lambda_{k,K}$

is not diagonal and each element of $\widehat{\beta}$ is approximately a weighted sum of all of the elements in $\widehat{\beta}_{OLS}$. The weights depend not only on $\{k, K\}$ but on all elements in Σ_X, denoted Σ_{ij}. For example, if $K = 3$ and $k = 1$, we have

$$\Lambda_{1,3} = \frac{1}{3} \begin{pmatrix} 1 & \frac{\Sigma_{12}}{\Sigma_{11}} & \frac{\Sigma_{13}}{\Sigma_{11}} \\ \frac{\Sigma_{12}}{\Sigma_{22}} & 1 & \frac{\Sigma_{23}}{\Sigma_{22}} \\ \frac{\Sigma_{13}}{\Sigma_{33}} & \frac{\Sigma_{23}}{\Sigma_{33}} & 1 \end{pmatrix}.$$

Each row of $\Lambda_{1,3}$ is the result of including a particular subset regression in the average. For example, the first row gives the first element of $\widehat{\beta}_{1,3}$ as a weighted sum of the OLS regressors $\widehat{\beta}_{OLS}$. Apart from the multiplication with $\frac{1}{3}$, its own coefficient is given a relative weight of one while the remaining coefficients are those we expect from accomodating the omitted variable bias. The effect of dividing by $n_{1,3} = 3$ is to shrink all coefficients, including its own coefficient, towards zero. For $k > 1$, each regressor gets included more often in the regressions. This increases the their effect on $\Lambda_{k,K}$ through a higher inclusion frequency, but decreases their effect through the omitted variable bias. Since the direct effect is larger than the omitted variable bias, an increased k generally reduces the amount of shrinkage. Of course, in the limit as $k = K$, there is no shrinkage and the method is identical to OLS.

A.2 The Structure of Dynamic Linear Models

Based on the specification of the dynamic linear model in equations (6) and (7), 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

$$\begin{aligned} V_{t-1}|I_{t-1} &\sim IG\left[\frac{n_{t-1}}{2}, \frac{n_{t-1}S_{t-1}}{2}\right], && (25) \\ \theta_{t-1}|I_{t-1}, V_{t-1} &\sim N\left[m_{t-1}, V_{t-1}C^*_{t-1}\right]. && (26) \end{aligned}$$

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,

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

Equations (11), (12) and (13) in the main text show the discount approach for specifying W_t.

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 $\vartheta\left(y;\mu,\sigma^2\right)$ denote the density of a normal distribution evaluated at y and $IG\left(V;a,b\right)$ the density of an $IG\left(a,b\right)$ distributed variable evaluated at V. We obtain the predictive density as

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

The predictive density

$$p\left(y_t|I_{t-1}\right) = t_{n_{t-1}}\left(\widetilde{y}_t;F_t^{'}m_{t-1,}\underbrace{S_{t-1}\cdot\underbrace{\left[1+F_t^{'}\left(\underbrace{C_{t-1}^*+W_t^*}_{:=R_t^*}\right)F_t\right]}_{:=Q_t^*}}_{:=Q_t}\right) \tag{29}$$

is Student-t distributed with location $F_t^{'}m_{t-1,}$ scale Q_t and n_{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

$$e_t = y_t - \widehat{y}_t, \tag{30}$$

playing a key role in signal conditioning learning. Updating the degrees of freedom is accomplished by

$$n_t = n_{t-1} + 1 \tag{31}$$

and the point estimate for the observational variance is updated as

$$S_t = S_{t-1} + \frac{S_{t-1}}{n_t}\left(\frac{e_t^2}{Q_t} - 1\right). \tag{32}$$

Q_t denotes the scale associated with the t-distributed forecast y_t, see (29) in A.2. Equation (32) 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.

The $r \times 1$ adaptive coefficient vector[29]

[29] Rewriting the adaptive vector as $A_t = \frac{S_{t-1}\left(C^*_{t-1}+W^*_{t-1}\right)F_t}{S_{t-1}\left[1+F_t'\underbrace{\left(C^*_{t-1}+W^*_t\right)}_{R^*_t}F_t\right]} = \frac{R_tF_t}{Q_t}$ shows that the adaptiveness to new observations does not depend on S_{t-1}.

$$A_t = \frac{R_t F_t}{Q_t} \tag{33}$$

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

$$m_t = m_{t-1} + A_t e_t, \tag{34}$$

$$C_t = \frac{S_t}{S_{t-1}} \left(R_t - A_t A_t' Q_t \right). \tag{35}$$