ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

A loss discounting framework for model averaging and selection in time series models☆

Dawid Bernaciak *, Jim E. Griffin

Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

ARTICLE INFO

Keywords:
Bayesian model synthesis
Density forecasting
Forecast combination
Forecast averaging
Multilevel discounting

ABSTRACT

We introduce a loss discounting framework for model and forecast combination, which generalises and combines Bayesian model synthesis and generalized Bayes methodologies. We use a loss function to score the performance of different models and introduce a multilevel discounting scheme that allows for a flexible specification of the dynamics of the model weights. This novel and simple model combination approach can be easily applied to large-scale model averaging/selection, handle unusual features such as sudden regime changes and be tailored to different forecasting problems. We compare our method to established and state-of-the-art methods for several macroeconomic forecasting examples. The proposed method offers an attractive, computationally efficient alternative to the benchmark methodologies and often outperforms more complex techniques.

1. Introduction

Recent developments in econometric modelling and machine learning techniques, combined with increasingly easy access to vast computational resources and data, have led to a proliferation of forecasting models yielding either point forecasts or full forecast density functions. This trend has been met with a renewed interest in tools that can effectively use these different forecasts, such as model selection, or forecast combination, pooling, or synthesis, *e.g.*, Stock and Watson (2004), Hendry and Clements (2004), Hall and Mitchell (2007), Raftery, Kárnỳ, and Ettler (2010), Geweke and Amisano (2011) Waggoner and Zha (2012), Koop and Korobilis (2012), Billio, Casarin, Ravazzolo, and Van Dijk (2013), Del Negro, Hasegawa, and Schorfheide (2016), Yao et al. (2018), McAlinn and West (2019), Diebold, Shin, and Zhang (2022), Li, Kang, and Li (2023) to mention just a few. Wang, Hyndman, Li, and Kang (2023) provide an excellent review of work in this area.

Combining forecasts from different models, rather than using a forecast from a single model, is intuitively appealing and justified by improved empirical performance (see *e.g.* Bates & Granger, 1969; Stock & Watson, 2004). Hendry and Clements (2004) suggested that combining point forecasts provides insurance against poor performance by individual models which are misspecified, poorly estimated or non-stationary.

In density forecasting, the superiority of a combination over single models is less clear. Bayesian model averaging (BMA) (Leamer, 1978) is a simple and coherent approach to weight forecasts in a combination, but may not be optimal under logarithmic scoring when the set of models to be combined is misspecified (Diebold, 1991). Since sets of models usually do not include the true data-generating mechanism, this result has driven substantial literature proposing alternatives to BMA. Hall and Mitchell (2007) proposed a logarithmic scoring rule for a time-invariant linear pool with weights on the simplex, which leads to a forecast density combination that minimises Kullback–Leibler divergence to the true but unknown density. This idea has been developed for Bayesian

☆ The results presented in this paper were reproduced by the Editor-in-Chief on 20 March 2023.

* Corresponding author.
E-mail address: dawid.bernaciak@ucl.ac.uk (D. Bernaciak).

https://doi.org/10.1016/j.ijforecast.2024.03.001

Please cite this article as: D. Bernaciak and J.E. Griffin, A loss discounting framework for model averaging and selection in time series models. International Journal of Forecasting (2024), https://doi.org/10.1016/j.ijforecast.2024.03.001.

estimation (Geweke & Amisano, 2011), Markov switching weights (Waggoner & Zha, 2012) and dynamic linear pools (Billio et al., 2013; Del Negro et al., 2016). These approaches often lead to better forecasting performance but at the cost of increased computational expense. A computationally cheaper alternative directly adjusts the model weights from BMA to allow time-variation (Raftery, Gneiting, Balabdaoui, & Polakowski, 2005) leading to dynamic model averaging (DMA) (Raftery et al., 2010), which uses an exponential discounting of Bayes factors with a discount/forgetting/decay[1] to achieve time-varying model weights. Performance can be sensitive to the discount factor, and Koop and Korobilis (2012) suggested using logarithmic score maximisation to find an optimal discount factor for DMA. Beckmann, Koop, Korobilis, and Schüssler (2020) applied this idea to model selection and developed the dynamic model learning (DML) method with an application to foreign exchange forecasting. Outside the formal Bayesian framework, Diebold et al. (2022) suggested a simple average of the forecasts from a team of N (or less) forecasters chosen using the average logarithmic scores in the previous rw-periods. This can be seen as a localised and simplified version of Hall and Mitchell (2007).

Recently, McAlinn and West (2019) and McAlinn, Aastveit, Nakajima, and West (2020) proposed a broad theoretical framework called Bayesian Predictive Synthesis (BPS), which includes the majority of proposed Bayesian techniques as special cases. They propose a novel forecast combination method using latent factor regression, cast as a Bayesian seemingly unrelated regression (SUR) and showed better performance than the BMA benchmark and an optimal linear pool. However, the approach can be computationally demanding with a large pool of models. Tallman and West (2023) use entropic tilting to expand the BPS framework to more general aims than forecast accuracy (such as return maximisation in portfolio allocation).

This paper describes our loss discounting framework (LDF), which extends DMA and DML to the general loss function (in a similar spirit to Tallman & West, 2023), and more general discounting dynamics. Acomputationally efficient time-varying discounting scheme is constructed through a sequence of pools of meta-models, which starts with the initial pool. Meta-models at one layer are constructed by combining meta-models at the previous layer using a DMA/DML type rule with different discount factors. We show that LDF can outperform other benchmark methods and is more robust to hyperparameter choice than DMA/DML in simulated data and foreign exchange forecasting using econometric fundamentals and a large pool of models. We also show how tailoring the approach to constructing long-short foreign exchange portfolios can lead to economic gains. A second example illustrates the limitations of our methodology in US inflation forecasting.

The paper is organised as follows. Section 2 presents some background leading into a description of the proposed methodology in Section 3. In Section 4, the performance of the LDF approach is examined in a simulated example and applications to foreign exchange and US inflation forecasting. We discuss our approach and set out directions for further research in Section 5.

2. Background

It is common in Bayesian analysis (Bernardo & Smith, 2009; Yao et al., 2018, and references therein) to distinguish three types of model pools $\mathcal{M} = \{M_1, M_2, \ldots, M_K\}$: $\mathcal{M}$-closed – the true data generating process is described by one of the models in $\mathcal{M}$ but is unknown to researchers; $\mathcal{M}$-complete – the model for the true data generating process exists but is not in $\mathcal{M}$, which is viewed as a set of useful approximating models; $\mathcal{M}$-open – the model for the true data generating process is not in $\mathcal{M}$ and the true model cannot be constructed either in principle or due to a lack of resources, expertise etc.[2] Model selection based on BMA only converges to the true model in the $\mathcal{M}$-closed case (see *e.g.* Diebold, 1991) and can perform poorly otherwise.

There are several reasons to believe that econometric problems are outside the $\mathcal{M}$-closed setting. Firstly, real-world forecasting applications often involve complex systems, and the model pool will only include approximations at best. One might argue that econometric modellers have an inherent belief that the models they propose provide a reasonable approximation to the data-generating process even if certain process features escape the capabilities of the supplied methodologies. Secondly, in many applications, the data-generating process is not constant in time (Del Negro et al., 2016) and may involve regime changes and considerable model uncertainty. For example, in the foreign exchange context, Bacchetta and Van Wincoop (2004) proposed the scapegoat theory, suggesting that investors display a rational confusion about the true source of exchange rate fluctuations. If an unobservable or unknown factor affects an exchange rate movement, investors may attribute this movement to some other observable macroeconomic fundamental variable. This induces regimes where different market observables might be more or less important.

These concerns motivate a model averaging framework that is suitable for $\mathcal{M}$-complete (or even $\mathcal{M}$-open) situations and incorporates time-varying model weights. We use $\pi_{t|s,k}$ to represent the weight of model k at time t using information to time s and use the forecast combination density

$$p(y_t|y_s) = \sum_{k=1}^{K} \pi_{t|s,k}\, p_k(y_t|y_s) \tag{2.1}$$

where $p_k(y_t|y_s)$ represents the forecast density of model k at time t using information $y_1, \ldots, y_s$, which we call the predictive likelihood. DMA (Raftery et al., 2010), assumes that $s = t - 1$ and updates $\pi_{t+1|t,k}$ using the observation

[1] The terms discount/forgetting/decay factor are used interchangeably in this paper.

[2] Clarke et al. (2013) give, a slightly unusual, example of works of William Shakespeare as an $\mathcal{M}$-open problem. The works (data) have a true data-generating process (William Shakespeare), but one can argue that it makes no sense to model the mechanism by which the data was generated.

at time t, y_t, and a forgetting factor, denoted by α, by the recursion

$$\pi_{t|t,k} = \frac{\pi_{t|t-1,k}\, p_j(y_t|y_{t-1})}{\sum_{l=1}^{K} \pi_{t|t-1,l}\, p_l(y_t|y_{t-1})}, \tag{2.2}$$

$$\pi_{t+1|t,k} = \frac{\pi_{t|t,k}^{\alpha} + c}{\sum_{l=1}^{K} \pi_{t|t,l}^{\alpha} + c}, \tag{2.3}$$

where c is a small positive number introduced to avoid model probability being brought to machine zero by aberrant observations.[3] The log-sum-exp trick is an alternative way of handling this numerical instability, which would, at least in part, eliminate the need for the constant c. We leave the role of this parameter to further research.

The recursions in (2.2) and (2.3) amount to a closed form algorithm to update the probability that model k is the best predictive model given information up to time t, for forecasting at time t. A model receives a higher weight if it performed well in the recent past, and the discount factor α controls the importance one attaches to the recent past. For example, if $\alpha = 0.7$, the forecast performance 12 periods before the last one receives approximately 2% of the importance of the most recent observation. However, if $\alpha = 0.9$, this importance is as high as 31%. Therefore, lower values of α lead to large changes in the model weights. In particular, $\alpha \to 0$ would lead to equal model weights, and $\alpha = 1$ recovers the standard BMA.

DMA has been shown to perform well in econometric applications whilst avoiding the computational burden of calculating large-scale Markov Chain Monte Carlo (MCMC) or sequential Monte Carlo associated with methods such as Waggoner and Zha (2012). Del Negro et al. (2016) showed that DMA performed comparably to their novel dynamic prediction pooling method in forecasting inflation and output growth. It was subsequently expanded and successfully used in econometric applications by Koop and Korobilis (2012, 2013) and Beckmann et al. (2020). In the first two papers, the authors compare DMA for a few possible values of discount factors α, whereas, in the latest paper, the authors follow the recommendation of Raftery et al. (2010) to estimate the forgetting factor online in the context of Bayesian model selection. We find that estimating the forgetting factor is key to the performance of DMA, as we will show in our simulation study and empirical examples. Our LDF provides a general approach by combining multiple layers of discounting with time-varying discount factors to provide better performance and robustness for the hyperparameter choice.

[3] Yusupova, Pavlidis, and Pavlidis (2019) note that this constant is only present in the original work by Raftery et al. (2010) and then in the implementation by Koop and Korobilis (2012) but then subsequently dropped in further works, software packages and citations. They also notice that this constant has a non-trivial and often critical effect on the dynamics of weight changes. We comment on this aspect in Appendix C.

3. Methodology

Our proposed loss discounting framework (LDF) provides a method of updating time-varying model weights using flexible discounting of a general measure of model performance. The flexible discounting is achieved by defining layers of meta-models using the simple discount scheme in (2.2) and (2.3). The approach can be used for dynamic model averaging and dynamic model selection. For example, we can define a pool of forecast combination densities by applying (2.1) with different discount factor values. We refer to the elements of this pool as meta-models. We can subsequently find the best meta-model average (or best meta-model) by again applying exponential discounting to the past performance of the meta-models. This leads to an approach with two layers, but clearly, we could continue the process by defining a pool of meta-models at one layer by applying the forecast combination in (2.1) to a pool of meta-models at the previous layer.

The method has two key features. The first key feature of the model averaging (selection) we develop is the ability to shrink the pool of the relevant models (show greater certainty across time in a single model) in times of low volatility and to encompass more models (display greater variation in model selection) when the volatility of the system is high. The second key feature is using a generalised measure of model performance, which enables users to define the scores/losses directly connected with their final goal. As we show in the empirical study, aligning model scores to the final purpose leads to better performance.

3.1. Loss discounting framework

We first describe how a score can generalize DMA and then describe our discounting scheme using meta-models. The score or loss (we will use these terms interchangeably) is defined for the prediction of an observation with predictive distribution p and observed value y and denoted $S(p, y)$. This measures the quality of the predictive distribution if the corresponding observed value is y. For a set of K models, we assume that the (one-step ahead) predictive distribution for model k at time t is $p_{k,t} = p_k(y_t|y_{t-1})$ we define the log-discounted predictive likelihood for the kth model at time t using discount factor α to be

$$\mathrm{LDPL}_{t,k}(\alpha) = \sum_{i=1}^{t-1} \alpha^{i-1} S\left(p_{k,t_i}, y_{t-i}\right).$$

We define a model-averaged predictive density

$$\sum_{k=1}^{K} w_{t|t-1,k}(\alpha)\, p_k(y_t|y_{t-1})$$

where

$$\left(w_{t|t-1,1}(\alpha), \ldots, w_{t|t-1,K}(\alpha)\right) = \mathrm{softmax}\left(\mathrm{LDPL}_{t,1}(\alpha), \ldots, \mathrm{LDPL}_{t,K}(\alpha)\right).$$

This generalizes the use of the logarithmic scoring in DMA. The use of scoring rules for Bayesian updating for

parameters was pioneered by Bissiri, Holmes, and Walker (2016) (rather than inference about models in forecast combination) and is justified in a $\mathcal{M}$-open or misspecified setting. Loaiza-Maya, Martin, and Frazier (2021) extend this approach to econometric forecasting. They consider equally weighted sums (*i.e.* $\alpha = 1$ for layer 2). Miller and Dunson (2019) justify using a powered version of the likelihood of misspecified models.

Each meta-model is defined using a recipe for model or meta-model averaging/selection. We consider a specific type of such recipe, which is based on exponential discounting of the scores with different discount factors from a set of possible values $\mathcal{S}_\alpha = \{\alpha_1, \ldots, \alpha_M\}$. To lighten the notation, we write $w(m)$ and $\text{LDPL}(m)$ to denote weights and log-discounted predictive likelihoods evaluated at α_m.

In the first layer, each model in the model pool is scored and the ith meta-model is defined by applying either DMA or DML with discounting α_i and the weights defined above.

Then, to construct the second layer, the meta-models in the first layer are scored, and the ith meta-model is again defined by applying either DMA or DML with discounting α_i to these scores. This iterative process can be easily extended to an arbitrary number of layers. We highlight two parallels between the methods used in LDF for time series models and concepts in Bayesian modelling. The first parallel is between the layers of meta-models in LDF and using hyperpriors in the Bayesian hierarchical models; similarly to deciding on the setup of hyperpriors in the hierarchical models, LDF allows for varying depth and type of meta-model layers appropriate for the use case in question. We also draw an analogy between the model selection versus the maximum a posteriori probability (MAP) estimate of the quantity and model weights in model averaging versus full posterior distribution.

To provide a full description of the approach, we will write the forecast densities of the K models as $p_1^{(0)}(y_t|y_{t-1}), \ldots, p_K^{(0)}(y_t|y_{t-1})$ to make notation consistent. At every other layer, we define predictive meta-models, an average of (meta-)models in the previous layers. At the first layer, we directly use the forecast combination in (2.1)

$$p_m^{(1)}(y_t|y_{t-1}) = \sum_{k=1}^{K} w_{t|t-1,k}^{(1)}(m)\, p_k^{(0)}(y_t|y_{t-1})$$

and, for $n \geqslant 2$, we apply (2.1) to the M meta-models specified at the previous layer,

$$p_m^{(n)}(y_t|y_{t-1}) = \sum_{k=1}^{M} w_{t|t-1,k}^{(n)}(m)\, p_k^{(n-1)}(y_t|y_{t-1}).$$

To define the weights $w_{t|t-1,k}^{(n)}$, we extend the log-discounted predictive likelihood for the kth (meta-)model at the nth layer at time t using discount factor α_m to be

$$\text{LDPL}_{t,k}^{(n)}(m) = \sum_{i=1}^{t-1} \alpha_m^{i-1} S\left(p_k^{(n)}, y_{t-i}\right). \tag{3.1}$$

The weights in layer n are constructed using either softmax[4] (to give a form of (meta-)model averaging) or argmax (to give a form of (meta-)model selection). We use the notation L_n to represent this operation in the nth layer, which can take the value of s (softmax) or a (argmax). If $L_n = s$,

$$\left(w_{t|t-1,1}^{(n)}(m), \ldots, w_{t|t-1,K}^{(n)}(m)\right) = \text{softmax}\left(\text{LDPL}_{t,1}^{(n-1)}(m), \ldots, \text{LDPL}_{t,K}^{(n-1)}(m)\right)$$

if $n = 1$, or

$$\left(w_{t|t-1,1}^{(n)}(m), \ldots, w_{t|t-1,M}^{(n)}(m)\right) = \text{softmax}\left(\text{LDPL}_{t,1}^{(n-1)}(m), \ldots, \text{LDPL}_{t,M}^{(n-1)}(m)\right)$$

if $n \geqslant 2$.

If $L_n = a$,

$$w_{t|t-1,k}^{(n)}(m) = \begin{cases} 1 & k = k^\star(m) \\ 0 & k \neq k^\star(m) \end{cases}$$

where

$$k^\star(m) = \text{argmax}\left(\text{LDPL}_{t,1}^{(r-1)}(m), \ldots, \text{LDPL}_{t,K}^{(r-1)}(m)\right)$$

if $n = 1$ or, if $n \geqslant 2$,

$$k^\star(m) = \text{argmax}\left(\text{LDPL}_{t,1}^{(r-1)}(m), \ldots, \text{LDPL}_{t,M}^{(r-1)}(m)\right).$$

The N-layer LDF with score S and with choice L_n (equal to s or a) at layer n will be written $\text{LDF}^N_{L_1 L_2 \ldots L_N}(S)$.

The scheme only needs a single discount factor to be chosen in the final meta-model layer. An expert might set this parameter or calculate it on a calibration sample if the data sample is sufficiently large to permit a robust estimation. LDF refers to the discount factor in the final meta-model layer as α.

As well as defining a model combination at each layer, $\text{LDF}^N_{L_1 L_2 \ldots L_N}(S)$ also leads to a discount model averaging of the initial model set for any N since

$$p_m^{(N)}(y_t|y_{t-1}) = \sum_{k_N=1}^{M} w_{t|t-1,k_N}^{(N)}(m)\, p_{k_N}^{(N-1)}(y_t|y_{t-1}) \tag{3.2}$$

$$= \sum_{k_1=1}^{K} \left[\sum_{k_2=1}^{M} \cdots \sum_{k_N=1}^{M} w_{t|t-1,k_N}^{(N)}(m) \times \prod_{p=1}^{N-1} w_{t|t-1,k_p}^{(p)}(k_{p+1}) \right] p_{k_1}^{(0)}(y_t|y_{t-1}). \tag{3.3}$$

Given this setup, the models and meta-models are either averaged by using the softmax function or selected by using the argmax function applied to the log-discounted predictive likelihood.

3.2. Special cases

3.2.1. Dynamic model averaging

The updates of the dynamic model averaging weights in (2.3) correspond to passing $\text{LDPL}_{t,1}^{(0)}, \ldots, \text{LDPL}_{t,K}^{(0)}$ with the logarithmic scoring function through the softmax function. In DMA, we only have one level of discounting where $p_k(y_t|y_{t-1})$ are the different forecaster densities.

[4] $\text{softmax}(a_1, \ldots, a_J) = \left(\frac{\exp\{a_1\}}{\sum_{j=1}^{J}\exp\{a_j\}}, \ldots, \frac{\exp\{a_J\}}{\sum_{j=1}^{J}\exp\{a_j\}}\right)$

Therefore, we could denote DMA as LDF^1_{S} where the superscript indicates a single level of loss discounting and the s subscript indicates the use of the softmax function.

3.2.2. Dynamic model learning

Dynamic model learning (DML) (Beckmann et al., 2020) provides a way to choose a single discount factor for model selection optimally. In DML, the logarithmic scoring functions $S\left(p_k^{(0)}, y_{t-i}\right)$ are passed through an argmax function to select the best model. We could refer to DML as $\text{LDF}^2_{\text{a,a}}$ with the second layer of meta-models prior restricted to a single point on the grid, namely $\mathcal{S}_\alpha = \{1\}$ for $n = 2$.

A similar idea for model averaging, using the softmax function for selecting an ensemble of parameters α, was developed in Zhao, Xie, and West (2016).

3.2.3. Two-layer model averaging/selection within loss discounting framework

The loss discounting framework allows us to describe more general setups for discounting in forecast combinations, such as these models with two or more meta-model levels. In this paper, we focus on LDF with two layers of meta-models such as $\text{LDF}^2_{\text{S,a}}$, $\text{LDF}^2_{\text{a,S}}$, $\text{LDF}^2_{\text{a,a}}$ and $\text{LDF}^2_{\text{S,S}}$, as well as, the limiting cases such as $\text{LDF}^\infty_{\text{S}\cdots\text{S}}$. In contrast to DMA and DML, having two (with $\alpha \neq 1$) or more layers of meta-models makes the discount factors in the other layers time-dependent, which, as we show in the following sections, leads to an improved performance of model averaging and selection.

In terms of computation time, our proposed algorithm is very fast as it relies on simple addition and multiplication. This is an advantage over more sophisticated forecast combination methods when the time series is long and/or we would like to incorporate a large (usually greater than 10) number of forecasters.

As mentioned before, $\text{LDF}^2_{\text{a,a}}$ is a generalised version of DML presented in Beckmann et al. (2020) where implicitly the authors suggest $\alpha = 1$, i.e., all past performances of the forgetting factors are equally weighted. In the limit $\alpha \to 0$, we would choose the discount factor α, which performed best in the latest run, disregarding any other history. The $\text{LDF}^2_{\text{a,S}}$ specification is a hybrid between model selection and model averaging. The first layer performs the model selection for each discount factor, and the second layer performs the model averaging for the discount factors. Therefore, we select a single model for each discount factor, but then we take a mixture of discount factors, which results in a mixture of models.

3.3. Properties of LDF as $N \to \infty$

Considering the impact of additional layers in an LDF model is natural. If we use either the softmax or the argmax at all layers, the weights for each model converge as $N \to \infty$, so adding more layers has a diminishing effect on the sequence of predictive distributions. Intuitively, we have a diminishing impact on the final result for the softmax functions as we take weighted averages of the weighted averages of the models. For the argmax/model selection, the LDF approach settles on a single model for any discount factor in the final layer. The detailed and rigorous proofs are provided in the technical Appendix A. We demonstrate in the empirical sections that the sequence converges to a predictive distribution, often the best or nearly best-performing setup of the LDF framework.

3.4. Comments

Low variation in LDPL across time leads to model weight concentration on fewer models, and higher variation in LDPL leads to the opposite; the model weights are more evenly spread. This is because in the presence of high variation in LDPL, the lower discount factors will be preferred, and hence, the faster forgetting will accommodate the regime switching.

If one believes that the data generating process (DGP) is present in the pool, LDF will not perform as well as BMA, which will asymptotically converge to the suitable model quicker than LDF. Conversely, if the DGP is not among the models in the pool, LDF adapts by adjusting the models' weights over time to approximate the DGP.

Following the argument in Del Negro et al. (2016) to interpret DMA in terms of a Markov switching model, our extension allows a time-varying transition matrix, *i.e.*, $Q_t = (q(t)_{kl})$. The gradual forgetting of the performance of the discount factor α allows for a change of optimal discount factor when the underlying changes in the transition matrix are required. However, we also show that our two-layer model specification outperforms the standard DMA model even when the transition matrix is non-time-varying. This point will be further illustrated in Appendix B.

4. Examples

Our methodology best suits data with multiple regime switches with a potentially time-varying transition matrix. As such, it is beneficial for modelling data such as inflation levels, interest or foreign exchange rates. We illustrate our model using a simulated example and two real data examples. The supplementary materials for our examples are given in Appendix B, Appendix C, Appendix D and Appendix E.

We compare examples of our LDF to several popular model-averaging methodologies. The approaches used are

- Multi-layer LDF - 2 hyperparameters, i.e., α, c;
- BMA - 0 hyperparameters;
- DMA - 2 hyperparameters, i.e., α, c
- BPS (McAlinn & West, 2019) - 5 hyperparameters, i.e., β discount factor for state evolution matrices, δ discount factor for residual volatility, n_0 prior number of degrees of freedom, s_0 prior on BPS observation variance, R_0 prior covariance matrix of BPS coefficients;
- best N-average (Diebold et al., 2022) - 2 hyperparameters, i.e., N number of models, rolling window length rw.
- DeCo (Billio et al., 2013) - 5 hyperparameters (defaults and online estimation options are available).

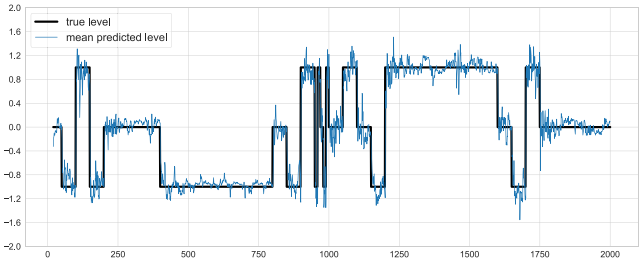


Fig. 1. Simulation – True data generating process mean and mean predicted level according to $\text{LDF}^2_{\text{S,S}}$.

We evaluate the performance of the models by calculating the out-of-sample mean log predictive score (MLS)

$$\text{MLS} = \frac{1}{T-s}\sum_{t=s+1}^{T} \log p(y_t|y_1,\ldots,y_{t-1}),$$

and log predictive density ratios (LPDR) for a chosen reference model m^*

$$\text{LPDR}(\tau) = \sum_{t=s+1}^{\tau} \log\{p_{m^*}(y_t|y_1,\ldots,y_{t-1})/p_{LDF}(y_t|y_1,\ldots,y_{t-1})\},$$

where $y_1,\ldots,y_s$ are the observations for a calibration period and T is the total number of observations and p_{LDF} correspond to the selected LDF model.

4.1. Simulation study

The data generating process (DGP) of Diebold et al. (2022) is

$$y_t = \mu_t + x_t + \sigma_y\epsilon_t, \quad \epsilon_t \sim N(0,1), \tag{4.1}$$

$$x_t = \phi_x x_{t-1} + \sigma_x v_t, \quad v_t \sim N(0,1), \tag{4.2}$$

where y_t is the variable to be forecast, x_t is the long-run component of y_t, μ_t is the time-varying level (in Diebold et al. (2022) set to 0). We can interpret μ_t as a piecewise-constant deterministic signal with a finite state space that accounts for regime switches. The error terms are all i.i.d and uncorrelated. It is assumed that the data generating process is known to each forecaster apart from the level component μ_t. Each individual forecaster k models x_t with noise and applies different level η_k to y_t:

$$z_{kt} = x_t + \sigma_{tk}\nu_{kt}, \quad \nu_{kt} \sim N(0,1), \tag{4.3}$$

$$\tilde{y}_{kt} = \eta_k + z_{kt} + \sigma_y\epsilon_t, \quad \epsilon_t \sim N(0,1). \tag{4.4}$$

Notice that the individual forecasters' levels do not vary over time. This emulates a situation where forecasters can access different sets of information and/or models that might guide a different level of choice. It emulates an $\mathcal{M}$-complete or even $\mathcal{M}$-open setting where no forecaster is correct at all times.

In contrast to Diebold et al. (2022), we allow the variable y_t to have multiple regime switches. The settings are as follows: $\phi_x = 0.9$, $\sigma_x = 0.3$, $\sigma_y = 0.3$, $\sigma_{tk} = 0.1\,\forall k$, $K = 20, T = 2001$, $\eta_k = -2 + 0.2105(k-1)$, $k = 1,\ldots,K$ and finally:

$$\mu_t = \begin{cases} 0, & \text{for } t \in [0,49] \bigcup [200,399] \bigcup [800,849] \bigcup [970,979] \\ & \bigcup [1000,1049] \bigcup [1600,1650] \bigcup [1700,2001] \\ 1, & \text{for } t \in [100,150] \bigcup [900,949] \bigcup [960,969] \bigcup [990,999] \\ & \bigcup [1050,1099] \bigcup [1200,1599] \bigcup [1700,1749] \\ -1, & \text{otherwise.} \end{cases}$$

More examples are discussed in Appendix B, where we draw the levels from a Markov switching model ten times. For LDF we set $S_\alpha = \{1.0, 0.99, 0.95, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.001\}$ and $c = 10^{-20}$ similarly to Koop and Korobilis (2012).

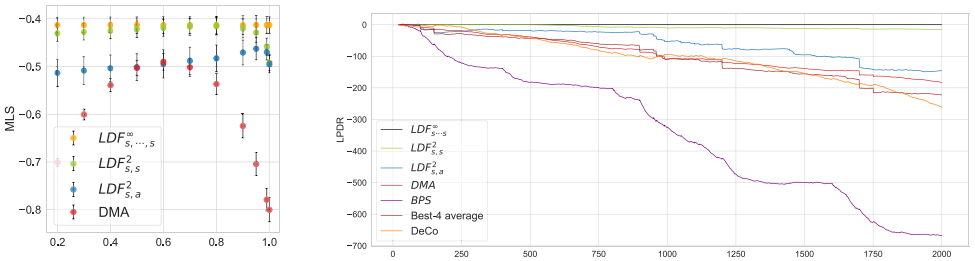
In Fig. 1, we present how the synthesised agent forecast level of $\text{LDF}^2_{\text{S,S}}$ adjusts to the mean levels implied by the DGP. We can see that the model is very reactive to the mean predicted level, following the true DGP mean closely with only a small time lag.

All results[5] are based on ten runs, where the levels were fixed, but the random numbers regenerated. The standard Bayesian model averaging (MLS = −4.34) fared poorly since it quickly converged to the wrong model. DeCo (MLS = −0.57) was adapted to output 39 quantiles from which we calculated the log scores[6] and it did not cope well with abrupt level changes in our numerical example, overestimating the variance and leading to poorer scores. BPS (MLS = −0.73) with normal agent predictive densities[7] performed better than BMA but struggled to adjust quickly to the regime changes, resulting in low log scores at the change points. The N-average method performed better (we chose the rolling window of five observations that performed best), with an MLS of −0.52

[5] Except BPS for which we performed only one run due to computational cost.

[6] 39 quantiles in increments of 0.025. We used the default setting in the DeCo package with $\Sigma = 0.09$ (matching our DGP) and with learning and parameter estimation. The quantiles indicated that the normal distribution could well approximate the output.

[7] We used the original set of parameters (adjusted $\beta = 0.95$ and $\delta = 0.95$ to get better results as proposed by the authors of the paper but adjusted the prior variance to match the σ_y parameter. The model was run for 5000 MCMC paths with 3000 paths burnin period. All other runs of BPS (which achieved worse results) are detailed in Appendix B).



(a) MLS versus values of α. (b) LPDR of the competing models.

Fig. 2. Simulation – (a) The MLS versus values of α for LDF and α for DMA in the x-axis. The error bars correspond to the standard deviation of MLS over ten runs. (b) LPDR of the competing models with a calibration period of 250.

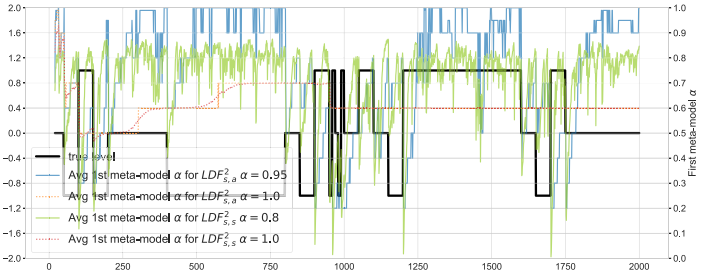


Fig. 3. Simulation – Comparison of the average α parameters in the first meta-model layer for $\text{LDF}^2_{\text{S,a}}$ model with $\alpha = 0.95$ versus $\alpha = 1$ as well as $\text{LDF}^2_{\text{S,S}}$ with $\alpha = 0.8$. We observe more dynamic adaptation of the discount parameter in the first meta-model layer when the final $\alpha < 1$.

for $N = 3$ and $N = 4$, than BMA and BPS and similarly to the standard DMA method of Raftery et al. (2010).

Crucially, we note that DMA's performance varies significantly depending on the hyperparameter choice, whereas the performance of multilayered LDF methods does not. This is clearly illustrated in Fig. 2(a). One could adopt various strategies to try to find the hyperparameters. The most basic one would be based on tuning the hyperparameter on the calibration period and keeping the parameter constant after that. In this case, for example, if we set the calibration period to 250, the methods choose discounts as: DMA 0.5 (MLS = −0.50); $\text{LDF}^2_{\text{S,S}}$ 0.6 (MLS = −0.42); $\text{LDF}^2_{\text{S,a}}$ 0.7 (MLS = −0.49). For comparison the stable state $\text{LDF}^{\infty}_{\text{S,...,S}}$ achieves MLS = −0.41. The non-LDF model averaging models, namely BPS, DeCO and best N-average, were tuned to achieve the best performance to the entire sample a posteriori in contrast to LDF models where we select a single configuration based on the initial sample of 250 observations. Another strategy could be based on selecting the best discount factor at each time step (online) based on the expanding window, potentially exponentially weighted - this boils down to an LDF approach with an additional argmax layer. In this case, DMA simply becomes $\text{LDF}^2_{\text{S,a}}$ and, as shown, can lead to better results. Even better results and more robustness can be achieved using $\text{LDF}^2_{\text{S,S}}$ where a mix of discount factors is being used.

In Fig. 2(b) we present the LPDR for the tested models against $\text{LDF}^{\infty}_{\text{S,...,S}}$. The LDF models (including DMA) generally performed better; however, the results suggest that the 2-layer LDF, which can weight multiple discount factors model, is more robust to abrupt level changes than the other methods.

Fig. 3 show how the average parameter α in the first meta-model layer dynamically changes using $\text{LDF}^2_{\text{S,a}}$ with $\alpha = 0.95$ and $\text{LDF}^2_{\text{S,S}}$ with $\alpha = 0.8$. It is close to 1 in periods of stability and closer to 0 during abrupt changes. In comparison, for $\alpha = 1$, the average parameter α in the first meta-model layer is stable, oscillating around 0.6. As mentioned before, this variation in parameter α might be beneficial since the lower the α parameter, the more models that will be considered, and the final outcome might show more uncertainty. Additionally, a lower parameter α facilitates the ability to quickly re-weight the models to adapt to the new regime. In times of stability, it might be better to narrow down the meaningful forecasts to a smaller group by increasing the parameter α. This illustrates how the two-layer model provides useful flexibility in the discount factors in the first meta-model layer. Another observation from Fig. 3 concerns the average values of discount parameters α in the first layer across time. For $\text{LDF}^2_{\text{S,a}}$ the average α in the first payer for $\alpha = 0.95$ in the second layer is 0.75 and for $\text{LDF}^2_{\text{S,S}}$ with $\alpha = 0.8$ in the last layer it is 0.71. The average α for both LDF models with $\alpha = 0.1$ in the last layer is 0.61.

4.2. Foreign exchange forecasts

We consider exchange rate forecasting (see Rossi, 2013, for a comprehensive review). The random walk is a typical benchmark as it corresponds to the claim that the exchange rates are unpredictable, but Rossi (2013) argues that economic variables can have time-varying predictive power. Beckmann et al. (2020) consider exploiting this predictive ability using DML with a pool of Time-Varying Parameter Bayesian Vector Autoregressive (TVP-BVAR)

models with different subsets of economic fundamentals. We closely follow their setup. Appendix D.1 and Appendix D.2 describe the model.[8]

We use a set of G10 currencies: Australian dollar (AUD), Canadian dollar (CAD), euro (EUR), Japanese yen (JPY), New Zealand dollar (NZD), Norwegian krone (NOK), Swedish krona (SEK), Swiss franc (CHF), pound (GBP) and US dollar (USD). All currencies are expressed in terms of the amount of dollars per unit of a foreign currency, i.e. the domestic price of a foreign currency. The data is monthly[9] and runs from November 1989 to July 2020. This is a more up-to-date data set than the one used in other studies, but similar in length.[10] We use the macroeconomic fundamentals:

- Uncovered Interest Rate Parity (UIP) which postulates that, given the spot rate S_t, the expected rate of appreciation (or depreciation) is approximately:
$$\frac{\mathbb{E}(S_{t+h} - S_t)}{S_t} = i_t - i_t^*, \tag{4.5}$$
where i_t is the domestic and i_t^* is the foreign interest rate corresponding to the time horizon h of the return.[11]
- Long-short interest rate difference - the difference between the ten-year benchmark government yield and one month deposit rate.
- Stock growth - monthly return on the main stock index of each of the G10 currencies/countries.
- Gold price - monthly change in the gold price.

The data is standardised based on the mean and standard deviation calibrated to an initial training period of 10 years.

We consider using all possible models as our pool (which consists of 2048 models, including all possible subsets of the fundamentals). Comparison of the N-average method (Diebold et al., 2022), DeCo (Billio et al., 2013) and BPS (McAlinn & West, 2019) with the competing methods is not available for this pool due to computational cost, and so we consider a small pool (which consists of the 32 models based on UIP only and time-constant parameters). An exhaustive list of model parameter settings is outlined in Appendix Appendix D.2.

We compare the performance of the competing model averaging techniques using the logarithmic score and economic evaluation using Sharpe ratios of a long-short currency portfolio. We find that LDF provides superior performance according to the logarithmic score and demonstrate how these differences in scores manifest themselves in an economic evaluation.

4.2.1. Small model pool - analysis of scores

Fig. 4 compares the logarithmic score for DMA and some specifications of LDF. LDF provides better performance for an optimal choice of the hyperparameter and is more robust to the choice of the hyperparameters than DMA.[12] Interestingly, for model selection, we note that the proposed two-layer LDF specification $\text{LDF}^2_{\text{a,a}}$ (as well as $\text{LDF}^{\infty}_{\text{a,...,a}}$) methodology improved upon the DML method (Beckmann et al., 2020), which as we recall is $\text{LDF}^2_{\text{a,a}}$ with $\alpha = 1$. The best scores in model averaging/selection were achieved for $\text{LDF}^2_{\text{S,S}}$/$\text{LDF}^2_{\text{a,S}}$ specification with $\alpha = 0.9$.

The average value of discount parameters α in the first meta-model layer across time, for $\text{LDF}^2_{\text{S,S}}$ with $\alpha = 0.9$ is 0.77 which was very similar for $\alpha = 1$. However, the variability of α in the first meta-model layer for $\alpha < 1$ was much larger, i.e. α being closer to 0 during times of increased volatility and closer to 1 during calmer times (same observation of either pool of models).

Fig. 5 shows the LDPR for the competing methods on an expanding window with the hyperparameters of LDF calibrated using the first ten years of data and the competing models calibrated in the sample. The LDPRs show considerable time variation with the sudden drops in performance of $\text{LDF}^2_{\text{S,a}}$ and Best-4 average models correspond to the period of big FX volatility increases as measured by Barclays G10 FX index.

In comparison to other methods, the $\text{LDF}^2_{\text{S,S}}$ method with $\alpha = 0.9$ performs best (MLS = 22.16), followed by other two layer LDF specifications and the 4-model average (MLS = 22.10). The BPS method (MLS = 21.60) did not perform well here. Similarly, the DeCo (MLS = 18.31) method using multivariate normal approximation[13] In terms of model performance out of sample, $\text{LDF}^2_{\text{S,S}}$ calibrated only on the initial ten years of data (to select α) - $\alpha = 0.8$ (MLS = 22.15) - still outperforms the other non-LDF models that were calibrated in-sample. The detailed results are presented in Table D.3 in Appendix D. The stable state LDF models performed similarly to the two layer specification, $\text{LDF}^{\infty}_{\text{S,...,S}}$ achieves MLS = 22.13 and $\text{LDF}^{\infty}_{\text{a,...,a}}$ scores MLS = 22.07.

[8] Following Koop and Korobilis (2013), we adopt an Exponentially Weighted Moving Average (EWMA) estimator for the measurement covariance matrix to avoid the need for the posterior simulation for multivariate stochastic volatility. This is different than Beckmann et al. (2020) who use the approximation derived by Triantafyllopoulos (2011).

[9] If month-end data was not available, it was substituted with the beginning of the month data or monthly average. These substitutions were unavoidable for some of the data in the 1980s. See Appendix Appendix D.7 for more details.

[10] Kouwenberg, Markiewicz, Verhoeks, and Zwinkels (2017), consider the data beginning from 1973. However, data samples from the 1970s and 1980s can vary between data providers, and the available quotes are of lower quality than the newer data. For example, due to the illiquidity of financial instruments. The details concerning the data sources and any proxies used are presented in the data Appendix Appendix D.7.

[11] In this context, we use the one month deposit rates. Theoretically, one should use the one month rates from the appropriate cross-currency curves. However, we assume that the difference between the deposit rates in the two countries provides a good proxy for the interest rate differential.

[12] In Appendix D.3 we show that with a dense grid of allowable values for α the points in Fig. 4 become smooth curves.

[13] For DeCo, we checked that the marginal distributions are well described by the normal distribution. We then output the covariance matrix from the DeCo source code to complete the multivariate normal approximation.

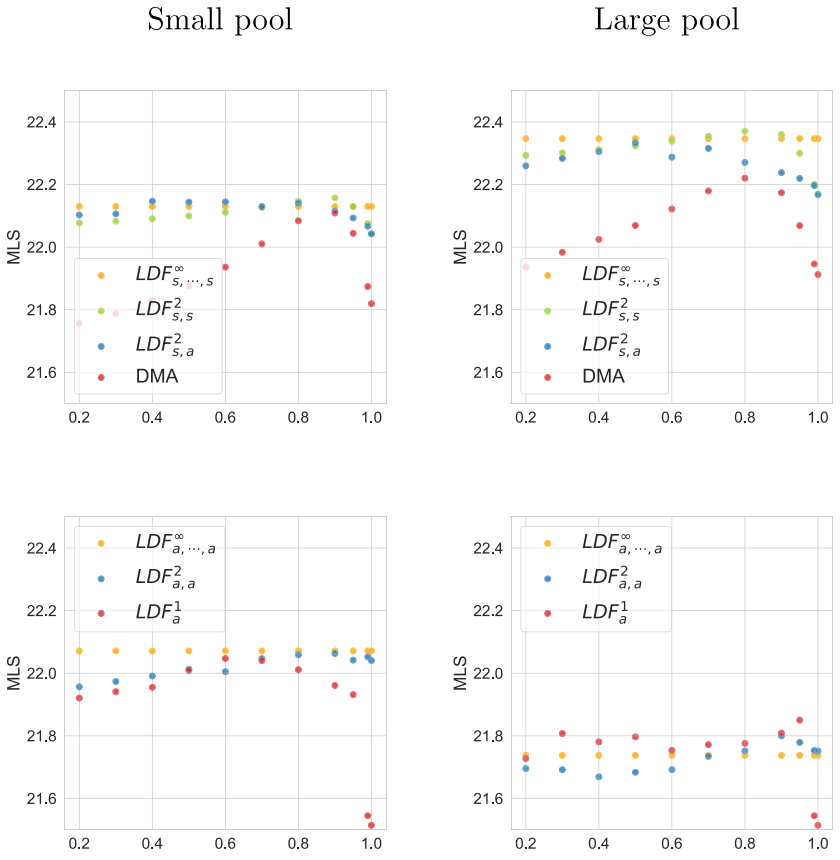


Fig. 4. FX – MLS versus values of α for LDF and α for DMA in the x-axis for the small and large model pool. The upper plots show the cases of model averaging, whereas the lower plots show model selection.

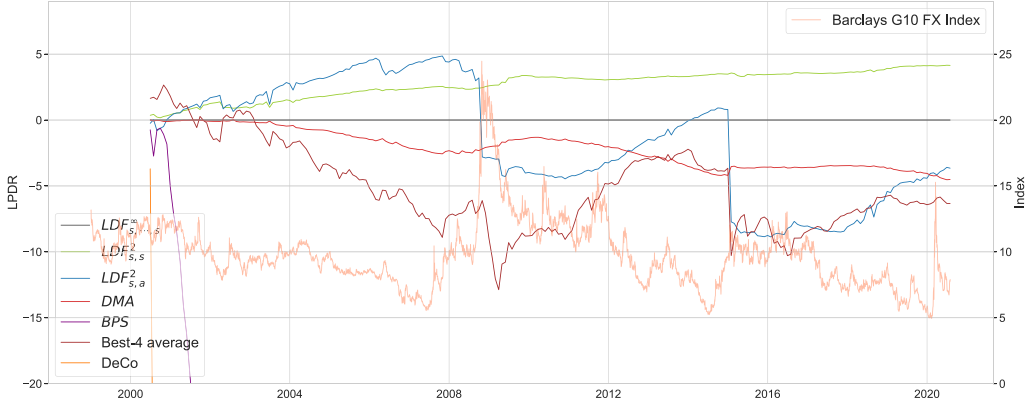


Fig. 5. FX – LPDR for model averaging in the small model pool. $LDF^{2}_{S,S}$ provides the best performance robust to increases in the FX volatility.

4.2.2. Large model pool - analysis of scores

We can only consider the LDF methods for the large pool of models due to the run times of the other methods. Additional meta-model layers have a similar impact as in the small model pool but with a less pronounced effect on model selection and a greater effect on model averaging Fig. 4. Again, the best model averaging scores were achieved for $LDF^{2}_{S,S}$ specification with $\alpha = 0.8$ (MLS = 22.37), and the stable state LDF models performed similarly to the two-layer specification, MLS = 22.35 for $LDF^{\infty}_{S,\ldots,S}$. For model selection, the multi-layer specification of LDF introduces the robustness to hyperparameters but does not necessarily outperform the single-layer LDF in terms of log scores. Interestingly, in the larger pool, the EWMA random walk[14] (RW) (decay factor 0.97) model

[14] I.e. we estimate the volatility of the random walk based on the exponentially weighted moving average.

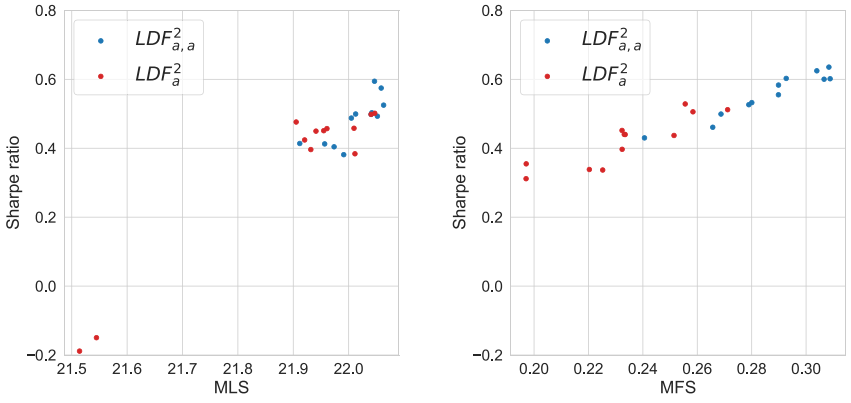


Fig. 6. FX – mean score values versus achieved Sharpe ratios. The log scores were used in the left-hand side plot; the focused scores were used in the right-hand side plot.

was not the best model of all models considered (MLS = 21.77), but it performed almost on par with the *a posteriori* best model (MLS = 21.78). This indicates that finding a single model that outperforms the random walk, even from a big pool of models, is hard.

4.2.3. Economic evaluation of model selection

We consider economic evaluation by constructing a portfolio of long and short currency positions targeting 10% annual volatility with 8bps transaction costs. We measure the performance by looking at the cumulative wealth over time, as well as the Sharpe ratio, which captures the risk-adjusted performance, applied to the smaller model pool of 32 models. To target the Sharpe ratio, we define the score as the portfolio returns divided by the portfolio standard deviation based on a rolling twelve-month window.

We concentrate on LDF configurations that select a single model at a time, that is $\text{LDF}^2_{\text{a,a}}$ and LDF^1_{a}. Portfolios are constructed by maximizing the returns subject to a fixed risk per model, as in Beckmann et al. (2020). Model averaging cannot be used as the correlation between the investment strategies would inevitably change the target risk level of the portfolio. An alternative approach to portfolio construction was presented in Tallman and West (2023), who use the multivariate focused prediction score in the context of model averaging where each model aims to minimise the risk subject to a fixed return target.

$\text{LDF}^2_{\text{a,a}}$ only narrowly outperforms LDF^1_{a} on the log score (Fig. 4) but has a higher Sharpe ratio (Fig. 6). This aligns with the observations in Beckmann et al. (2020), noting that small differences in the log scores can translate to noteworthy economic differences. The right panel of Fig. 6 shows the mean focused scores (MFS) where there are apparent differences (unlike the log scores) with the double discounting version of LDF achieving better scores leading to higher Sharpe ratios and higher final wealth as seen in Fig. 7. The double discounting of $\text{LDF}^2_{\text{a,a}}$ allows the discount factor to drop in times of higher volatility, such as during the great financial crisis, the Chinese crash or the Brexit referendum. For $\alpha = 0.7$ in the second layer, the time average value of α in the first layer is 0.71 and with $\alpha = 1$ in the first layer it is 0.80. This is in contrast to DML ($\text{LDF}^2_{\text{a,a}}$ $\alpha = 1$) and LDF^1_{a} specifications where in the former the discount factor settles at 0.9 and does not move and in the latter it is just fixed to a predetermined constant value.

Fig. 8 shows the portfolio composition through time. We note that the weights display stability when the portfolio value experiences periods of growth, and the sudden weight changes correspond to periods of growth plateauing. The weights generally follow the carry trade strategy, which is well documented in the literature, see Della Corte and Tsiakas (2012) and references therein.

4.3. US inflation forecasts

The final study considers an example of McAlinn and West (2019), which involves forecasting the quarterly US inflation rate between 1961/Q1 and 2014/Q4. Here, the inflation rate corresponds to the annual percentage change in a chain-weighted GDP price index. There are four competing models: M_1 includes one period lagged inflation rate, M_2 includes period one, two and three lagged inflation interest and unemployment rates, M_3 includes period one, two and three lagged inflation rate only, and M_4 includes period one lagged inflation interest and unemployment rates. All four models provide Student-t distributed forecasts with around 20 degrees of freedom.

The distinguishing features of this example are the small number of models and the existence of time periods when none of the models or model combinations lying on simplex provide an accurate mean forecast. In this example, we will see the limitation of the LDF and other simplex-based methodologies, which are unable to correct for forecasting biases if bias-corrected models are not explicitly available in the pool.

The BPS method (MLS = 0.06) dominates all other methodologies since it allows model combinations that do not adhere to simplex. There were six dates in the evaluation period where the mean of the BPS synthesised model was greater than the maximum of the underlying models. The ability to go beyond simplex proved to be one of the key factors in the superior performance.

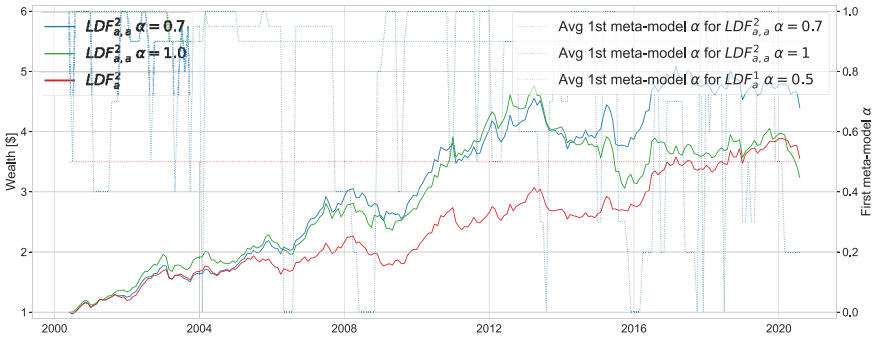


Fig. 7. FX – Money through time and discount factor α through time for $\text{LDF}^2_{\text{a,a}}$ with $\alpha = 0.7$ and $\alpha = 1$, and LDF^1_{a} with $\alpha = 0.5$.

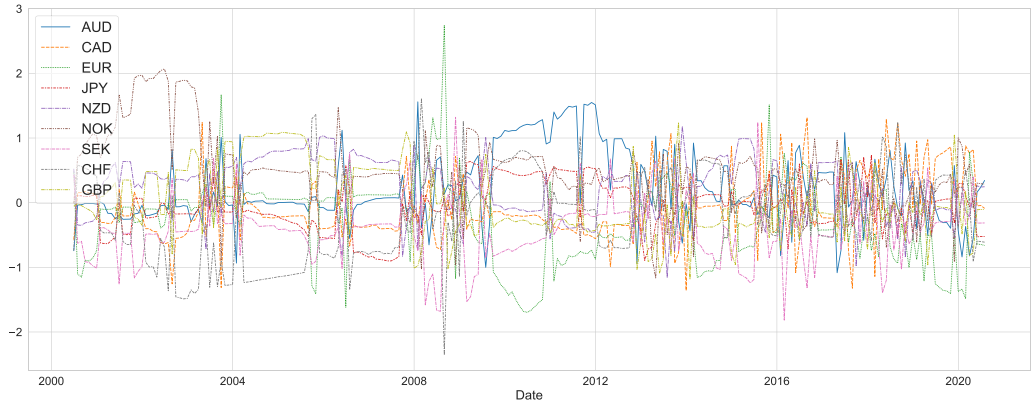


Fig. 8. FX – Portfolio composition through time for $\text{LDF}^2_{\text{a,a}}$ with $\alpha = 0.7$. We can see that long stretches of stable composition correspond to the growth periods, and the periods of sudden portfolio changes correspond to the periods of money growth plateauing.

The next most effective method was the N-model average of Diebold et al. (2022), which for $N = 2$ and $N = 3$ models had an MLS equal to -0.01 and provided better performance than the best single model (M_2, MLS = -0.02). For $N = 2$, out of the 100 evaluation points, the algorithm selected the pair (M_0, M_1) 35 times, the pair (M_2, M_3) 49 times and the pair (M_1, M_3) 16 times. On the other hand, both 2-level LDF model averaging and DMA methods did not work well in this example, but they improved upon picking just a single model. The poor performance of 2-level LDF and DMA could mostly be attributed to the highly dynamic nature of these methods which sometimes attached too much weight to a single model that would score poorly (see Fig. 9).

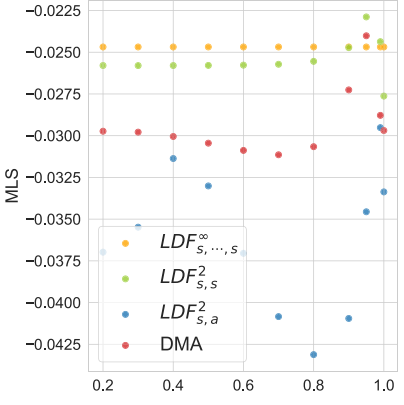


Fig. 9. US inflation – The MLS versus values of α for LDF and α for DML in the x-axis.

5. Discussion

This paper contributes to the model averaging and selection literature by introducing a loss discounting framework that encompasses dynamic model averaging, first presented by Raftery et al. (2010), generalises dynamic model learning (Beckmann et al., 2020) and introduces additional model averaging or selection specifications. The framework allows for general dynamics for model weights and works well with focused scores for goal-oriented decision-making. The methodology offers extra flexibility, which can lead to better forecast scores and yield results that are less sensitive to the choice of hyperparameters. This is particularly important in a more realistic online forecasting setting where the selection of the globally optimal hyperparameters is often unattainable. It also empowers users to choose the model specification in terms of the number of levels of discounting layers suitable for the problem at hand.

We show that our proposed methodology performs well in both the simulation study and the empirical examples based on the exchange rate forecasts, where we show the superiority of our approach both for model averaging as well as model selection, where for the latter, we also demonstrate how the differences in the scores translate to noteworthy economic gains. We find that the LDF can be a good choice when: the number of forecasters is fairly large, and sophisticated methods become burdensome; if we want to have only a small number of hyperparameters to calibrate; we suspect that we are in the $\mathcal{M}$-complete/open setting, and different models might be optimal at different times, but there is no consistent bias to be eliminated across all models; if we believe that scoring forecasters on the joint predictive density or joint utility basis is reasonable.

The LDF is not the panacea for model synthesis, and the performance of different model synthesis methods depends on the problem (as seen in the empirical studies). However, LDF can often achieve competitive performance with low computational overhead using flexible dynamics and general model scores in an easy-to-implement and compute framework.

There are multiple open avenues to explore. Many current forecast combination methods described in the literature assume that the pool of forecasters does not change over time (see *e.g.* Diebold et al., 2022; McAlinn & West, 2019; Raftery et al., 2010). This is a substantial limitation in some situations, for example, if a pool of experts provides the forecasts.

Let us first consider the situation of a new agent being added to the existing pool of forecasters. The existing forecasters already have a track record of forecasts and corresponding scores. A new forecaster could be included with an initial weight. This could be fairly easily achieved in the LDF by considering a few initial scores. It is unclear what this weight should be, especially in more formal methodologies that relax the simplex restriction like McAlinn and West (2019). Similarly, forecasters may drop out completely, or for some quarters, before providing new forecasts. Again, it is generally hard to know how to weight these forecasters. The LDF provides a rationale, and we should be using an estimate of that forecaster's score when a forecast is made. This time series prediction problem can be approached using standard methods.

We noted in the empirical sections that the best-performing discount factor in the second layer is larger than the average across-time discount factor in the first layer. We showed that as one keeps adding more and more layers of meta-models, the weights converge to an equilibrium, i.e., adding more layers does not change the scores any more. Any choice of the discount factor in the final layer leads to the same score and discount factors in all other layers.

It would also be of interest to consider the case when the models to be combined themselves allow for sharp breaks (Gerlach, Carter, & Kohn, 2000; Huber, Kastner, & Feldkircher, 2019). Intuitively, more flexible models will lead to weights with fewer fluctuations if the models can represent the true DGP (for example, if one model is correctly specified, then LDF should be able to replicate BMA roughly). We believe that using out-of-sample log predictive scores to calculate weights in LDF will avoid the problems of overfitting found using in-sample estimation methods. Therefore, we believe that LDF can take advantage of more flexible models and robustifies against the use of overly simple models.

As mentioned, we use joint predictive log-likelihood as a statistical measure of out-of-sample forecasting performance in most examples. It indicates how likely the realisation of the modelled variable was conditional on the model parameters. The logarithmic scoring rule is strictly proper, but it severely penalises low probability events, and hence it is sensitive to tail or extreme cases, see Gneiting and Raftery (2007). A different proper scoring rule could be used when needed, or if a decision is to be made based on the outcomes of model averaging/selection, then a focused score (or utility) aligned with the final goal can be used, as demonstrated in one of our examples.

Furthermore, since the scoring function is often based on the joint forecast probability density function, our methodology is not best suited to take strength from forecasters who might be good at forecasting one or more variables but not the others. This is partially because our methodology does not consider any dependency structure between expert models, and the weighting is solely performance-based. An extension introducing a way to take the agent inter-dependencies into consideration would be of considerable interest.

More broadly, the exponential discounting recipe could be generalised and expanded by any forecast of the scores which could involve more parameters.

6. Data and code availability

The data and the code in Python, R and MATLAB reproducing the results in this paper are freely availabe on https://github.com/dbernaciak/ldf.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2024.03.001.

References

Bacchetta, P., & Van Wincoop, E. (2004). A scapegoat model of exchange-rate fluctuations. *American Economic Review*, *94*(2), 114–118.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*, 451–468.

Beckmann, J., Koop, G., Korobilis, D., & Schüssler, R. A. (2020). Exchange rate predictability and dynamic Bayesian learning. *Journal of Applied Econometrics*, *35*, 410–421.

Bernardo, J. M., & Smith, A. F. (2009). *Bayesian theory*: *vol. 405*, John Wiley & Sons.
Billio, M., Casarin, R., Ravazzolo, F., & Van Dijk, H. K. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, *177*, 213–232.
Bissiri, P. G., Holmes, C. C., & Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B.*, *78*, 1103–1130.
Clarke, J. L., Clarke, B., Yu, C.-W., et al. (2013). Prediction in $\mathcal{M}$-complete problems with limited sample size. *Bayesian Analysis*, *8*, 647–690.
Del Negro, M., Hasegawa, R. B., & Schorfheide, F. (2016). Dynamic prediction pools: An investigation of financial frictions and forecasting performance. *Journal of Econometrics*, *192*, 391–405.
Della Corte, P., & Tsiakas, I. (2012). Statistical and economic methods for evaluating exchange rate predictability. In J. James, I. W. Marsh, & L. Sarno (Eds.), *Handbook of exchange rates* (pp. 221–263). John Wiley & Sons, Ltd.
Diebold, F. X. (1991). A note on Bayesian forecast combination procedures. In P. Hackl, & A. H. Westlund (Eds.), *Economic structural change* (pp. 225–232). Springer.
Diebold, F. X., Shin, M., & Zhang, B. (2022). On the aggregation of probability assessments: Regularized mixtures of predictive densities for Eurozone inflation and real interest rates. *Journal of Econometrics*, Article 105321.
Gerlach, R., Carter, C., & Kohn, R. (2000). Efficient Bayesian inference for dynamic mixture models. *Journal of the American Statistical Association*, *95*(451), 819–828.
Geweke, J., & Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, *164*, 130–141.
Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359–378.
Hall, S. G., & Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, *23*, 1–13.
Hendry, D. F., & Clements, M. P. (2004). Pooling of forecasts. *The Econometrics Journal*, *7*, 1–31.
Huber, F., Kastner, G., & Feldkircher, M. (2019). Should I stay or should I go? A latent threshold approach to large-scale mixture innovation models. *Journal of Applied Econometrics*, *34*(5), 621–640.
Koop, G., & Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, *53*, 867–886.
Koop, G., & Korobilis, D. (2013). Large time-varying parameter VARs. *Journal of Econometrics*, *177*, 185–198.
Kouwenberg, R., Markiewicz, A., Verhoeks, R., & Zwinkels, R. C. J. (2017). Model uncertainty and exchange rate forecasting. *Journal of Financial and Quantitative Analysis*, *52*, 341–363.
Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*: *vol. 53*, Wiley New York.
Li, L., Kang, Y., & Li, F. (2023). Bayesian forecast combination using time-varying features. *International Journal of Forecasting*, *39*(3), 1287–1302.
Loaiza-Maya, R., Martin, G. M., & Frazier, D. T. (2021). Focused Bayesian prediction. *Journal of Applied Econometrics*, *36*, 517–543.
McAlinn, K., Aastveit, K. A., Nakajima, J., & West, M. (2020). Multivariate Bayesian predictive synthesis in macroeconomic forecasting. *Journal of the American Statistical Association*, *115*, 1092–1110.
McAlinn, K., & West, M. (2019). Dynamic Bayesian predictive synthesis in time series forecasting. *Journal of Econometrics*, *210*, 155–169.
Miller, J. W., & Dunson, D. B. (2019). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, *114*, 1113–1125.
Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*, 1155–1174.
Raftery, A. E., Kárnỳ, M., & Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, *52*, 52–66.
Rossi, B. (2013). Exchange rate predictability. *Journal of Economic Literature*, *51*, 1063–1119.
Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*, 405–430.
Tallman, E., & West, M. (2023). Bayesian predictive decision synthesis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkad109.
Triantafyllopoulos, K. (2011). Time-varying vector autoregressive models with stochastic volatility. *Journal of Applied Statistics*, *38*, 369–382.
Waggoner, D. F., & Zha, T. (2012). Confronting model misspecification in macroeconomics. *Journal of Econometrics*, *171*, 167–184.
Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting*, *39*(4), 1518–1547.
Yao, Y., Vehtari, A., Simpson, D., Gelman, A., et al. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*, 917–1007.
Yusupova, A., Pavlidis, N. G., & Pavlidis, E. G. (2019). Adaptive dynamic model averaging with an application to house price forecasting. arXiv:1912.04661.
Zhao, Z. Y., Xie, M., & West, M. (2016). Dynamic dependence networks: Financial time series forecasting and portfolio decisions. *Applied Stochastic Models in Business and Industry*, *32*(3), 311–332.