Bregman Model Averaging for Forecast Combination

Yi-Ting Chen[†][1], Chu-An Liu[‡][2], and Jiun-Hua Su[§][2]

[1]Department of Finance, Center for Research in Econometric Theory and Applications, National Taiwan University
[2]Institute of Economics, Academia Sinica

January 31, 2024

Abstract

We propose a unified model averaging (MA) approach and establish its asymptotic optimality for a wide class of forecasting targets. The asymptotic optimality is achieved by minimizing an asymptotic risk based on the expected *Bregman divergence* of a combined-forecast sequence from a forecasting-target sequence under the local(-to-zero) asymptotics. This approach is flexibly applicable to generate MA methods in different forecasting contexts, including, but not limited to, univariate or multivariate mean forecasts, volatility forecasts, probabilistic forecasts and density forecasts. We also conduct Monte Carlo simulations (empirical applications) to show that compared to related existing methods, the MA methods generated by this approach perform reasonably well in finite samples (real data).

Keywords: Bregman divergence, forecast combination, loss function, model averaging

JEL Classification: C18, C32, C53

[†]Yi-Ting Chen gratefully acknowledges the research support by the National Science and Technology Council, Taiwan (NSTC-111-2410-H-002-085).

[‡]Chu-An Liu gratefully acknowledges the research support by the Center for Research in Econometric Theory and Applications, Taiwan (Grant No. 113L8601).

[§]Address correspondence to Jiun-Hua Su, 128 Academia Road, Section 2, Nankang, Taipei, 115 Taiwan; E-mail address: jhsu@econ.sinica.edu.tw.

1 Introduction

Forecast combination is undoubtedly one of the most essential concepts in economic forecasts; see, e.g., Bates and Granger (1969), Granger and Ramanathan (1984), Granger (1989), and Diebold and Lopez (1996) for earlier methods and Elliott and Timmermann (2008, 2016b), Petropoulos et al. (2022), and Wang et al. (2023) for reviews. More recently, researchers have proposed different frequentist model averaging (MA) methods for establishing the asymptotic optimality of a linear combination of the individual forecasts generated from a set of estimated candidate models. Some of these MA methods assume that the candidate models have fixed specification biases for a non-parametric data generating process (DGP), while others adopt the local(-to-zero) asymptotics setting by assuming that the candidate models are locally misspecified for a parametric DGP; see, e.g., Hansen (2008), Hansen (2010), and Cheng and Hansen (2015) for the former and Wan et al. (2014), Liu and Kuo (2016), and Chen and Liu (2023) for the latter.

Conventionally, researchers develop forecast-combination methods in the mean-forecast context where the forecasting target is a univariate random variable, the candidate models are linear predictive regressions, the parameters are estimated by the least squares (LS) method, the loss function of forecast error is quadratic, and the asymptotic optimality is defined by minimizing an asymptotic risk based on the mean squared forecast error (MSFE). Although this context is standard in the literature, it is too restrictive in practice. Indeed, the loss function is not necessarily quadratic from the viewpoint of practitioners; see, e.g., Timmermann (2006), Lee (2007), Patton and Timmermann (2007a, 2007b), and Barendse and Patton (2022). Researchers may also be interested in other types of forecasting targets. Examples include, but are not limited to, the multivariate vector autoregressive (VAR) forecasts considered by Ang and Piazzesi (2003) and Bauer et al. (2012) for macroeconomic applications, the volatility forecasts considered by Andersen et al. (2003), Hansen and Lunde (2005), Patton (2011), and Wang et al. (2016) for financial applications, the probabilistic forecasts considered by Estrella and Mishkin (1998), Kauppi and Saikkonen (2008), and Erdogan et al. (2015) for recession prediction, and the density forecasts considered by Diebold et al. (1998) and Berkowitz (2001), among many others, for risk management.

Importantly, since the existing frequentist MA methods are context-specific by construction, the asymptotic optimality of a specific forecast-combination method may not hold when the loss function, the forecasting target, the model specification *or* the estimation method is beyond its designed context. Although there are certain studies that explore the problem of forecast combination beyond the univariate mean-forecast context, such as the VAR forecast combination considered by Liao et al. (2019) and Liao and Tsay (2020), the density forecast combination considered by Hall and Mitchell (2007), Geweke and Amisano (2011), and Pauwels and Vasnev (2016), among others, it still lacks a unified MA approach that is flexibly applicable to establish asymptotically optimal combined forecasts in various forecasting contexts, to the best of our knowledge.

In this paper, we fill the gap by proposing such a unified MA approach for forecast combination based on the local-asymptotics setting. We adopt this setting because it is useful for dealing with the trade-off between specification bias and estimation uncertainty of the forecasts generated from the estimated candidate models. This setting is also popular in the MA literature and the forecasting literature; see, e.g., Hjort and Claeskens (2003), Claeskens and Hjort (2008), Hansen (2014), and Liu (2015) for the former and Elliott et al. (2013), Wan et al. (2014), Liu and Kuo (2016), Hirano and Wright (2017), and Chen and Liu (2023) for the latter. The proposed approach is established in a generalized forecasting context where the forecasting target, the candidate models and the estimation method are not pre-specified in theoretical analysis. Correspondingly, we measure the divation of a combined forecast from a forecasting target using the Bregman divergence (BD), which was proposed by Bregman (1967) for measuring the deviation of one point h_1 from another point h_2 in an inner-product space. Given a smooth and strictly convex 'loss-generating function' $\psi(\cdot)$, the BD is defined as the remainder of the first-order Taylor expansion of $\psi(h_1)$ around h_2. The BD has been applied to estimation (Zhang et al., 2009; Zhang et al., 2010), optimization (Zhang, 2003), principal component analysis (Collins et al., 2001), clustering analysis (Banerjee et al., 2005b; Banerjee et al., 2007) and forecast evaluation and comparison (Gneiting and Raftery, 2007; Gneiting, 2011; Laurent et al., 2013; Patton, 2020) in different research fields. We apply the BD to model averaging for forecast combination.

We build our unified approach on the BD for the following reasons. First, the strict convexity of $\psi(\cdot)$ allows us to interpret the BD as a generalized loss function of forecast error; see also Laurent et al. (2013) and Patton (2020) for related discussions. Second, the BD encompasses a number of important divergence measures that are suitable for different forecasting contexts, such as the squared ℓ^2-norm for univariate or multivariate mean forecasts, the 'QLIKE loss' for volatility forecasts, and the Kullback-Leibler (KL) divergence for probabilistic forecasts or density forecasts; see also Banerjee et al. (2005b) and Gneiting and Raftery (2007) for more examples. Third, given an information set, the conditional mean of a forecasting target is the optimal forecast that minimizes the Bregman risk, that is the expected BD, of an arbitrary forecast from the forecasting target; see Banerjee et al. (2005a). This property is essential for establishing the asymptotic optimality of our approach.

Following Chen and Liu (2023), we adopt the setting of combining forecast sequences, which is more general than the setting of combining single forecasts. We show that, in the generalized forecasting context, the Bregman risk of a combined-forecast sequence consists of two components that are, respectively, irreducible and reducible by the choice of combination weights. Importantly, the reducible component can be normalized as a mean squared error between the combined-forecast sequence and the conditional-mean sequence (of the forecasting target), *weighted by* the Hessian matrix of $\psi(\cdot)$, in large samples under the local-asymptotics setting. Accordingly, we define an asymptotic risk of the combined-forecast sequence, establish the Bregman MA approach by minimizing this asymptotic risk, and facilitate the use of the Bregman MA approach by providing a consistent estimator of the asymptotic risk.

Our approach, like several existing MA methods, involves balancing the specification biases and the estimation uncertainty of the candidate models in an optimal way; it stands apart from other MA methods in two ways. First, unlike most of related existing methods focusing on the setting of combining single forecasts, our approach adopts the setting of combining forecast sequences. As explained by Chen and Liu (2023), the latter setting is more general than the former setting, and has important implications on the asymptotic risk of combined forecasts. In particular, the in-sample estimation scheme and the asymptotic ratio of the forecast-sequence length relative to the in-sample size are irrelevant to combining single forecasts but essential for combining forecast sequences. In the case of combining forecast sequences, the asymptotic risk under the recursive scheme is not greater than that under the fixed scheme (or the rolling scheme); moreover, such a risk shrinks and eventually attains a lower bound, which is a squared-bias component, as the asymptotic ratio goes to infinity. Chen and Liu (2023) showed these results in the conventional context of univariate mean forecasts. We show that these results also hold for a generalized forecasting context.

Second, more importantly, our approach is flexibly applicable to different forecasting contexts with different forecasting targets, models, estimation methods or loss functions. This flexibility is essential not only for theoretical unification but also for empirical applications. We further utilize our approach to generate new MA methods in the contexts of univariate and multivariate mean forecasts, volatility forecasts, probabilistic forecasts and density forecasts. Considering these forecasting contexts is essential for macroeconomic and financial applications, as mentioned previously. Our approach includes the MA method of Chen and Liu (2023) as a particular example in the conventional context of univariate mean forecasts. In other forecasting contexts, our approach provides the theoretical foundation of combining forecast sequences and thereby generates useful alternatives to related existing methods.

In this study, we also conduct Monte Carlo simulations (empirical applications) to illustrate that compared to existing MA methods for VAR forecast combinations and density forecast combinations, the MA methods generated by our unified approach perform reasonably well in finite samples (real data).

The remainder of this paper is organized as follows. In Section 2, we introduce our approach. In Section 3, we apply our approach to establish new MA methods in different forecasting contexts, and compare the resulting methods with related existing methods. Section 4 includes Monte Carlo simulations. Section 5 includes empirical applications. We conclude this paper in Section 6. All proofs, additional theoretical results, further simulation evidence, and empirical findings are collected in the online appendix.

Throughout this paper, we maintain the assumption that all random variables are defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and use the following notation. For a generic n-vector-valued function $h = (h_1, \dots, h_n)^\top$ with an argument $\phi = (\phi_1, \dots, \phi_m)^\top \in \mathbb{R}^m$, we write $\nabla_\phi h$ for an $m \times n$ matrix with its (i, j) element equal to $\partial h_j / \partial \phi_i$; specifically,

$$\nabla_\phi h = \begin{bmatrix} \frac{\partial h_1}{\partial \phi_1} & \cdots & \frac{\partial h_n}{\partial \phi_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_1}{\partial \phi_m} & \cdots & \frac{\partial h_n}{\partial \phi_m} \end{bmatrix}.$$

For any positive integer $\tau \geq 2$, let $\nabla_\phi^\tau h \coloneqq \nabla_\phi \operatorname{vec}(\nabla_\phi^{(\tau-1)} h)$ and $\nabla_\phi^1 h \coloneqq \nabla_\phi h$, where vec denotes the vec operator. We suppress the subscript ϕ when there is little risk of confusion. Given a generic $m \times n$ matrix A, we denote its Frobenius norm by $\|A\| \coloneqq (\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2)^{1/2}$, where a_{ij} is the (i, j) element of A. Using the Frobenius norm throughout this paper entails no loss of generality because of the norm equivalence over a finite-dimensional vector space. We also write $\operatorname{tr}(A)$ for the sum of the main diagonal elements of a squared matrix A.

2 The Proposed Approach

2.1 A Generalized Forecasting Framework

Suppose that at *each* point in time $t = T, \dots, T + P - 1$, we are required to forecast an $n \times 1$ target f_{t+1}, which is realized at $t + 1$, after observing a $d \times 1$ vector X_t of predictors. To fulfill this requirement, we attempt to use a parametric model $g : \mathcal{X} \times \Theta \subseteq \mathbb{R}^d \times \mathbb{R}^{\bar{k}} \to \mathbb{R}^n$ to provide a forecast $g_t(\theta) \coloneqq g(X_t, \theta)$ of the target f_{t+1} for some parameter θ.[1] This framework accommodates several forecasting contexts of interest in the literature on economic forecast; see Section 3. Despite the different forecasting objects of interest, we aim to establish our MA approach for forecast combination in a unified manner. To this end, we consider the loss incurred by using a generic forecast h of f_{t+1} at time t to be the BD (of h relative to f_{t+1}):

$$D_\psi : (f_{t+1}, h) \in \Psi \times \mathcal{H} \mapsto D_\psi(f_{t+1}, h) \coloneqq \psi(f_{t+1}) - \psi(h) - \langle \nabla \psi(h), f_{t+1} - h \rangle,$$

[1] Rigorously speaking, the mapping g should be treated as a skeleton, which is a noise-free counterpart of a model, as defined in Tong (1990, p. 96). In this paper, we use the two terms, model and skeleton, interchangeably when no confusion may arise.

where the loss-generating function ψ is strictly convex on $\Psi \subseteq \mathbb{R}^n$ and differentiable on $\mathcal{H} \subseteq \Psi$,[2] $\nabla\psi$ is the gradient of ψ (that is, the $n \times 1$ vector of derivatives of ψ) and $\langle\cdot,\cdot\rangle$ denotes the usual inner product on $\mathbb{R}^n \times \mathbb{R}^n$. Intuitively, it is sensible to use the BD for evaluating the forecasting performance because $D_\psi(f_{t+1}, h)$ is the remainder of the first-order Taylor expansion of $\psi(f_{t+1})$ around h. In addition, the strict convexity of ψ implies that $D_\psi(f_{t+1}, h) \geq 0$, where the equality holds if and only if $f_{t+1} = h$. More importantly, as shown in Theorem 1 of Banerjee et al. (2005a), for every $\mathcal{F}_t$-measurable forecast $h_{t+1|t}$ of f_{t+1},

$$\mathbb{E}[D_\psi(f_{t+1}, h_{t+1|t})] \geq \mathbb{E}[D_\psi(f_{t+1}, f^*_{t+1|t})],$$

where $f^*_{t+1|t} \coloneqq \mathbb{E}[f_{t+1}|\mathcal{F}_t]$ is the conditional mean of the forecasting target, and $\mathcal{F}_t$ is the σ-field generated by X_t. Phrased differently, the *Bregman risk* $\mathbb{E}[D_\psi(f_{t+1}, h_{t+1|t})]$ is minimized at $h_{t+1|t} = f^*_{t+1|t}$; accordingly, this property provides a guideline for the forecast optimality under the decision-theoretical framework.

Since the class of $\mathcal{F}_t$-measurable functions is too large, it is impractical, if not impossible, to effectively search for $f^*_{t+1|t}$. Instead, taking into account a variety of specification searches based on economic theories or statistical evidence, for example the interpretive search and simplification search as indicated in Leamer (1978), we consider J candidate models of the form $\{g_t(\theta_{(j)}) : \theta_{(j)} \in \Theta_{(j)}\}$ with $\Theta_{(j)} \coloneqq \{S_{(j)}\theta : \theta \in \Theta\}$, where $S_{(j)}$ is a $\bar{k} \times \bar{k}$ identity matrix with $\bar{k} - k_j$ row(s) replaced with the $1 \times \bar{k}$ zero vector. Assuming without loss of generality that the $\bar{k} \times 1$ zero vector is in the set Θ, we have the property that $\Theta_{(j)} \subseteq \Theta$ for every j. We require the number J of candidate models to be finite but allow these models to be pairwise non-nested.

Rather than selecting one out of these J models, we propose an MA approach to combine the forecast sequences generated by the candidate models.[3] Suppose that at each $t = T, \dots, T+P-1$, an estimator $\hat{\theta}_{(j),t}$ of parameter $\theta_{(j)}$ is available for all candidate models, we thus have a sequence $\{\hat{f}_{t+1|t}(\boldsymbol{w})\}_{t=T}^{T+P-1}$ of combined one-step-ahead forecasts, where $\hat{f}_{t+1|t}(\boldsymbol{w}) \coloneqq \sum_{j=1}^J w_j g_t(\hat{\theta}_{(j),t})$ for some $\boldsymbol{w} = (w_1, \dots, w_J)^\top$. We are interested in finding a weight vector $\boldsymbol{w} \in \mathbb{W} \coloneqq \{(w_1, \dots, w_J)^\top \in [0,1]^J : \sum_{j=1}^J w_j = 1\}$ such that the *averaged Bregman risk*

$$\mathcal{B}(\boldsymbol{w}; T, P) \coloneqq \frac{1}{P} \sum_{t=T}^{T+P-1} \mathbb{E}[D_\psi(f_{t+1}, \hat{f}_{t+1|t}(\boldsymbol{w}))]$$

[2] In the literature on convex optimization, $\mathcal{H}$ is considered to be the collection $\mathrm{dom}(\partial\psi)$ of points at which ψ is subdifferentiable; see Definition 9.2 of Beck (2017) for example. In this paper, $\mathcal{H}$ is allowed to be a proper subset of $\mathrm{dom}(\partial\psi)$ because forecasting models are user-specified.

[3] As argued in Clemen (1989), "using a combination of forecasts amounts to an admission that the forecaster is unable to build a properly specified model."

is as small as possible. In the case where $P \geq 2$, this is in pursuit of the optimality of $\boldsymbol{w}$ for a *sequence* $\{\hat{f}_{t+1|t}(\boldsymbol{w})\}_{t=T}^{T+P-1}$ of forecasts, not for a single forecast $\hat{f}_{T+1|T}(\boldsymbol{w})$, in the sense of minimizing the averaged Bregman risk $\mathcal{B}(\boldsymbol{w}; T, P)$. The setting of combining forecast sequences follows Chen and Liu (2023), and is more general than the conventional setting of $P = 1$ in combining single forecasts.

Admittedly, such an averaged risk in general depends on the estimation scheme of $\theta_{(j)}$'s, which will be addressed in the subsequent section. Before delving further into the effect of estimation scheme of $\theta_{(j)}$'s, we pause to study the structure of $\mathcal{B}(\boldsymbol{w}; T, P)$, which is essential for establishing the proposed approach. For this purpose, we make the following two high-level assumptions.

H1 For every j and $t \geq T$,

$$\limsup_{T\to\infty} \mathbb{E}\left[\left\|\sqrt{T}(\hat{g}_{j,t+1|t} - f^*_{t+1|t})\right\|^3\right] < \infty,$$

where $\hat{g}_{j,t+1|t}$ is the jth candidate forecast of f_{t+1}.

H2 The loss-generating function ψ is twice differentiable and its associated Hessian matrix is Lipschitz continuous on $\mathcal{H}$.

Under Assumption H1, the deviation of the combined forecast $\hat{f}_{t+1|t}(\boldsymbol{w})$ from $f^*_{t+1|t}$, the optimal forecast that minimizes the Bregman risk, shrinks at the rate of $T^{-1/2}$; specifically, $\hat{f}_{t+1|t}(\boldsymbol{w}) = f^*_{t+1|t} + \mathrm{O_p}\left(T^{-1/2}\right)$. This property is ensured by the local-asymptotics setting that imposes the 'local misspecification' and smoothness assumptions, as indicated in Assumptions P1-P5 later. Inspection of the proof in Proposition 1 below reveals that the requirement of finite third moment of $\left\|\sqrt{T}(\hat{g}_{j,t+1|t} - f^*_{t+1|t})\right\|$ can be weakened to finite second moment whenever the norm of the Hessian matrix $\nabla^2\psi$ is uniformly bounded by a constant with probability one. Assumption H2 is technical and typically met in our forecasting context, as shown in the concrete examples in the next section.

Equipped with these high-level assumptions, we are now in the position to investigate the averaged Bregman risk $\mathcal{B}(\boldsymbol{w}; T, P)$ in the following proposition.

Proposition 1. *For every* $\boldsymbol{w} \in \mathbb{W}$*,*

$$\mathcal{B}(\boldsymbol{w}; T, P) = \frac{1}{P}\sum_{t=T}^{T+P-1} \mathbb{E}[D_\psi(f_{t+1}, f^*_{t+1|t})] + \frac{1}{P}\sum_{t=T}^{T+P-1} \mathbb{E}[D_\psi(f^*_{t+1|t}, \hat{f}_{t+1|t}(\boldsymbol{w}))].$$

Suppose that Assumptions H1 and H2 hold. For every $\boldsymbol{w} \in \mathbb{W}$*,*

$$\frac{1}{P}\sum_{t=T}^{T+P-1} \mathbb{E}[D_\psi(f^*_{t+1|t}, \hat{f}_{t+1|t}(\boldsymbol{w}))]$$

$$= \frac{1}{2P} \sum_{t=T}^{T+P-1} \mathbb{E}\left[(\hat{f}_{t+1|t}(\boldsymbol{w}) - f^*_{t+1|t})^\top \nabla^2 \psi(f^*_{t+1|t})(\hat{f}_{t+1|t}(\boldsymbol{w}) - f^*_{t+1|t})\right] + \mathrm{o}\left(T^{-1}\right)$$

as $T \to \infty$. If, in addition, $\limsup_{t\to\infty} \mathbb{E}\left[\left\|\nabla^2 \psi(f^*_{t+1|t})\right\|^3\right] < \infty$, then

$$\frac{1}{P} \sum_{t=T}^{T+P-1} \mathbb{E}[D_\psi(f^*_{t+1|t}, \hat{f}_{t+1|t}(\boldsymbol{w}))] = \mathrm{O}\left(T^{-1}\right).$$

Proposition 1 indicates that the averaged Bregman risk $\mathcal{B}(\boldsymbol{w}; T, P)$ is the sum of an irreducible Bregman risk $P^{-1}\sum_{t=T}^{T+P-1} \mathbb{E}[D_\psi(f_{t+1}, f^*_{t+1|t})]$ and a reducible Bregman risk $P^{-1}\sum_{t=T}^{T+P-1} \mathbb{E}[D_\psi(f^*_{t+1|t}, \hat{f}_{t+1|t}(\boldsymbol{w}))]$. It also shows that the difference between the reducible Bregman risk and the mean squared error between the combined-forecast sequence $\{\hat{f}_{t+1|t}(\boldsymbol{w})\}_{t=T}^{T+P-1}$ and the conditional-mean sequence $\{f^*_{t+1|t}\}_{t=T}^{T+P-1}$, *weighted by* the Hessian matrix $\nabla^2\psi(f^*_{t+1|t})$, is of order $\mathrm{o}\left(T^{-1}\right)$. Moreover, the reducible Bregman risk is of order $\mathrm{O}\left(T^{-1}\right)$ provided that every element of $\nabla^2\psi(f^*_{t+1|t})$ has finite third moment for all t sufficiently large. Since this result holds for all weights whereas the irreducible Bregman risk is of order $\mathrm{O}\left(1\right)$, it could be difficult for a theoretically optimal forecast combination to outperform the equal-weight scheme in terms of the averaged Bregman risk. This corresponds to a phenomenon known as the 'forecast combination puzzle' in literature.

2.2 Asymptotic Risk of Combined Forecasts

Recognizing that the irreducible Bregman risk is invariant to the combination weights and the reducible Bregman risk is of order $\mathrm{O}\left(T^{-1}\right)$, we consider the asymptotic risk of a combined-forecast sequence $\{\hat{f}_{t+1|t}(\boldsymbol{w})\}_{t=T}^{T-P+1}$ to be an asymptotic normalization of $\mathcal{B}(\boldsymbol{w}; T, P)$:

$$\begin{aligned}\mathcal{R}(\boldsymbol{w}) &\coloneqq \lim_{T\to\infty} \frac{2T}{P} \sum_{t=T}^{T+P-1} \mathbb{E}[D_\psi(f^*_{t+1|t}, \hat{f}_{t+1|t}(\boldsymbol{w}))] \\ &= \lim_{T\to\infty} \frac{T}{P} \sum_{t=T}^{T+P-1} \mathbb{E}\left[(\hat{f}_{t+1|t}(\boldsymbol{w}) - f^*_{t+1|t})^\top \nabla^2\psi(f^*_{t+1|t})(\hat{f}_{t+1|t}(\boldsymbol{w}) - f^*_{t+1|t})\right].\end{aligned}$$

To proceed with our theoretical analysis on $\mathcal{R}(\boldsymbol{w})$, we introduce notation that facilitates the presentation of the specification about candidate forecasts. For every j, let $\Theta_j \coloneqq \left\{S_j\theta_{(j)} \in \mathbb{R}^{k_j} : \theta_{(j)} \in \Theta_{(j)} \subseteq \mathbb{R}^{\bar{k}}\right\}$, where S_j is the $k_j \times \bar{k}$ submatrix of $S_{(j)}$ obtained by crossing out the rows equal to the $1\times\bar{k}$ zero vector. For every $\theta_j \in \Theta_j$, let $g_{j,t}(\theta_j) \coloneqq g_t(S_j^\top \theta_j)$. We write $\hat{g}_{j,t+1|t}$ for the jth candidate forecast obtained by the jth candidate model $g_{j,t}$; specifically, $\hat{g}_{j,t+1|t} \coloneqq g_{j,t}(\hat{\theta}_{j,t}) = g_t(\hat{\theta}_{(j),t})$, where $\hat{\theta}_{(j),t} = S_j^\top \hat{\theta}_{j,t} \in \Theta_{(j)}$ and $\hat{\theta}_{j,t}$ is an estimator of a (pseudo-true) parameter $\theta_j^* \in \Theta_j$. For example, we may consider the 'minimum BD' (MBD) estimator that minimizes the in-sample BD in the fixed, rolling, or recursive scheme:

$$\hat{\theta}_{j,t} \coloneqq \begin{cases} \underset{\theta_j \in \Theta_j}{\arg\min} \dfrac{1}{T-1} \displaystyle\sum_{s=1}^{T-1} D_\psi(f_{s+1}, g_{j,s}(\theta_j)), & \text{fixed;} \\ \underset{\theta_j \in \Theta_j}{\arg\min} \dfrac{1}{T-1} \displaystyle\sum_{s=t+1-T}^{t-1} D_\psi(f_{s+1}, g_{j,s}(\theta_j)), & \text{rolling;} \\ \underset{\theta_j \in \Theta_j}{\arg\min} \dfrac{1}{t-1} \displaystyle\sum_{s=1}^{t-1} D_\psi(f_{s+1}, g_{j,s}(\theta_j)), & \text{recursive.} \end{cases} \tag{1}$$

In this case, the corresponding population parameter is $\theta_j^* \coloneqq \arg\min_{\theta_j \in \Theta_j} \mathbb{E}[D_\psi(f_{t+1}, g_{j,t}(\theta_j))]$.[4] Note that Zhang et al. (2009) also considered the MBD estimator, but they did not consider the forecasting problem and the estimation schemes. It is also important to note that adopting the MBD estimator is not necessary in our theoretical analysis. Our approach allows the estimation problem and the forecasting problem to have different loss functions. This is useful for maintaining the flexibility of our approach in applications. Instead of focusing on a specific estimator, we impose primitive assumptions on a working estimator $\hat{\theta}_{j,t}$ and its corresponding population parameter θ_j^* as follows.

P1 (i) The following convergence in distribution holds: As $t \to \infty$,

$$\begin{bmatrix} \sqrt{t}(\hat{\theta}_{1,t} - \theta_1^*) \\ \vdots \\ \sqrt{t}(\hat{\theta}_{J,t} - \theta_J^*) \end{bmatrix} \xrightarrow{\text{d}} \begin{bmatrix} Z_1 \\ \vdots \\ Z_J \end{bmatrix},$$

where each Z_j has mean zero and finite second moment.

(ii) For every j and k, the sequence $\left\{ t(\hat{\theta}_{j,t} - \theta_j^*)(\hat{\theta}_{k,t} - \theta_k^*)^\top \right\}$ is asymptotically uniformly integrable and independent of X_t.

(iii) $\limsup_{t\to\infty} \mathbb{E}\left[\left\| \sqrt{t}(\hat{\theta}_{j,t} - \theta_j^*) \right\|^8 \right] < \infty$.

P2 For each j, the candidate model $g_{j,t}$ is twice continuously differentiable and satisfies

$$\limsup_{t\to\infty} \mathbb{E}\left[\left\| \nabla_{\theta_j} g_{j,t}(\theta_j^*) \right\|^8 \right] < \infty \quad \text{and} \quad \limsup_{t\to\infty} \mathbb{E}\left[\sup_{\theta_j \in \Theta_j} \left\| \nabla_{\theta_j}^2 g_{j,t}(\theta_j) \right\|^8 \right] < \infty.$$

P3 The unrestricted model g_t is twice continuously differentiable and satisfies

$$\limsup_{t\to\infty} \mathbb{E}\left[\left\| \nabla_\theta g_t(\theta^*) \right\|^8 \right] < \infty \quad \text{and} \quad \limsup_{t\to\infty} \mathbb{E}\left[\sup_{\theta \in \Theta} \left\| \nabla_\theta^2 g_t(\theta) \right\|^8 \right] < \infty.$$

[4] As indicated by Hansen (2022, p. 792), it is not easy to give general conditions under which the pseudo-true parameter is identified in the context of a nonlinear regression model. Throughout this paper, we simply assume the pseudo-true parameter θ_j^* is identified.

P4 For each j, $\left\|\theta^*_{(j)} - \theta^*\right\| = \mathrm{O}\left(T^{-1/2}\right)$ as $T \to \infty$, where $\theta^*_{(j)} := S_j^\top \theta^*_j \in \Theta_{(j)}$, and $\theta^* \in \Theta$ is defined by the condition $f^*_{t+1|t} = g_t(\theta^*)$.

P5 The loss-generating function ψ is twice differentiable and satisfies

$$\limsup_{t\to\infty} \mathbb{E}\left[\left\|\nabla^2 \psi(f^*_{t+1|t})\right\|^4\right] < \infty.$$

Assumption P1(i) requires that the working estimator $\hat{\theta}_{j,t}$ is $\sqrt{t}$-consistent. For example, we show in Theorem A.1 of the online appendix that under some regularity conditions, the recursive estimator in (1) is $\sqrt{t}$-consistent and asymptotically normal. To accommodate the fixed and rolling estimators in (1), which are $\sqrt{T}$-consistent and asymptotically normal as shown in the online appendix, we can consider an alternative assumption:

P1' (i) The following convergence in distribution holds: As $t \geq T \to \infty$,

$$\begin{bmatrix} \sqrt{T}(\hat{\theta}_{1,t} - \theta^*_1) \\ \vdots \\ \sqrt{T}(\hat{\theta}_{J,t} - \theta^*_J) \end{bmatrix} \xrightarrow{\mathrm{d}} \begin{bmatrix} Z_1 \\ \vdots \\ Z_J \end{bmatrix},$$

where each Z_j has mean zero and finite second moment.

(ii) For every j and k, the sequence $\left\{T(\hat{\theta}_{j,t} - \theta^*_j)(\hat{\theta}_{k,t} - \theta^*_k)^\top\right\}$ is asymptotically uniformly integrable and independent of X_t.

(iii) $\limsup_{T\to\infty} \mathbb{E}\left[\left\|\sqrt{T}(\hat{\theta}_{j,t} - \theta^*_j)\right\|^8\right] < \infty$.

Assumption P1' differs from Assumption P1 only in the normalization; thus, we focus on Assumption P1 in the following discussion. Assumption P1(ii) is a technical requirement. On the one hand, the asymptotic uniform integrability could be established by the large deviation inequalities in the context of maximum likelihood (ML) estimation or by the maximal inequalities in the context of M-estimation, as claimed in Nishiyama (2010, Theorem 1) and Negri and Nishiyama (2017). On the other hand, the argument of asymptotic independence is sensible under Assumption P1(i) because this assumption implies that $t(\hat{\theta}_{k,t} - \theta^*_k)(\hat{\theta}_{j,t} - \theta^*_j)^\top$ converges in distribution to $Z_k Z_j^\top$, as $t \to \infty$, which is independent of X_t. Under Assumption P1(iii), the eighth moment of $\left\|\sqrt{t}(\hat{\theta}_{j,t} - \theta^*_j)\right\|$ is finite for all t sufficiently large. This assumption is satisfied if $\sqrt{t}(\hat{\theta}_{j,t} - \theta^*_j)$ is asymptotically normal (a typical case under Assumption P1(i)) and the sequence $\left\{t^4(\hat{\theta}_{j,t} - \theta^*_j)^8\right\}$ is asymptotically uniformly integrable because these two conditions ensure the moment convergence, $\lim_{t\to\infty} \mathbb{E}\left[\left\|\sqrt{t}(\hat{\theta}_{j,t} - \theta^*_j)\right\|^8\right] = \mathbb{E}\left[\|Z_j\|^8\right] < \infty$, as indicated in Theorem 2.20 of van der Vaart (1998). Assumption P2 requires smoothness of the candidate model $g_{j,t}$ and moment restrictions on its derivatives. These moment restrictions can be weakened to

$$\limsup_{t\to\infty} \mathbb{E}\left[\left\|\nabla_{\theta_j} g_{j,t}(\theta_j^*)\right\|^4\right] < \infty \quad\text{and}\quad \limsup_{t\to\infty} \mathbb{E}\left[\sup_{\theta_j\in\Theta_j}\left\|\nabla^2_{\theta_j} g_{j,t}(\theta_j)\right\|^4\right] < \infty$$

if the second derivatives of ψ are uniformly bounded by some constant; for example, in Chen and Liu (2023), $\psi(f) = f^2/2$ for all $f \in \mathbb{R}$ and thus $\nabla^2\psi(f^*_{t+1|t}) = 1$ for all t. Assumption P3 imposes a similar requirement on the unrestricted model g_t. Assumption P4 requires the unrestricted model to be correctly specified and the restricted models to be locally misspecified. Assumption P5 requires that the Hessian matrix $\nabla^2\psi$ evaluated at $f^*_{t+1|t}$ admits moments of at least order 4. Such a moment condition is generally satisfied in our forecasting context if mild regularity assumptions on $f^*_{t+1|t}$ are imposed.

In order to elaborate on the effect of an estimation scheme, we need some notation. Let $\boldsymbol{\vartheta}_T$ be the $\bar{k} \times J$ matrix with its j-th column equal to $\sqrt{T}(S_j^\top \theta_j^* - \theta^*)$. Also, let $\kappa_0 := 1$ and $\kappa_\lambda := \lambda^{-1}\log\{1+\lambda\}$ for $\lambda > 0$. For every $\lambda \geq 0$, let $\boldsymbol{\Omega}_\lambda := \kappa_\lambda \boldsymbol{A} + \boldsymbol{B}$, where $\boldsymbol{A}$ is the $J \times J$ matrix with its (j,k) element equal to

$$\boldsymbol{A}_{(j,k)} = \operatorname{tr}\left(\mathbb{E}\left[\nabla_{\theta_j} g_{j,t}(\theta_j^*)\nabla^2\psi(f^*_{t+1|t})[\nabla_{\theta_k} g_{k,t}(\theta_k^*)]^\top\right]\mathbb{E}\left[Z_k Z_j^\top\right]\right)$$

and $\boldsymbol{B}$ is the $J \times J$ matrix satisfying

$$\boldsymbol{B} = \lim_{T\to\infty} \boldsymbol{\vartheta}_T^\top \mathbb{E}\left[\nabla_\theta g_t(\theta^*)\nabla^2\psi(f^*_{t+1|t})\left[\nabla_\theta g_t(\theta^*)\right]^\top\right]\boldsymbol{\vartheta}_T.$$

As pointed out in Section A.1 of the online appendix, if $\hat{\theta}_{k,T_i}$ and $\hat{\theta}_{j,T_i}$ are the recursive MBD estimators in (1), then the covariance matrix $\mathbb{E}\left[Z_k Z_j^\top\right]$ is of the sandwich form; in particular, $\mathbb{E}\left[Z_k Z_j^\top\right] = \boldsymbol{M}_{kk}^{-1}\boldsymbol{W}_{kj}\boldsymbol{M}_{jj}^{-1}$ for some matrices $\boldsymbol{M}_{kk}$, $\boldsymbol{W}_{kj}$, and $\boldsymbol{M}_{jj}$.

We are now ready to explore $\mathcal{R}(\boldsymbol{w})$ in the following theorem, which nests Chen and Liu's (2023) Theorem 1 as a special case.[5]

Theorem 1. *Suppose that the sequence $\{X_t\}$ of predictors is stationary. If Assumptions H1-H2 and P1-P5 hold, then*

$$\mathcal{R}(\boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{\Omega}_{\bar{\lambda}} \boldsymbol{w},$$

where $\bar{\lambda} := \lim_{T\to\infty} P/T \in [0,\infty)$. Replacing Assumption P1 with Assumption P1' and maintaining the other assumptions, we obtain

$$\mathcal{R}(\boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{\Omega}_0 \boldsymbol{w}.$$

[5] The matrix $\boldsymbol{A}$ can be written compactly and this compact form is used to show that Theorem 1 generalizes Chen and Liu's (2023) Theorem 1. Details are given in Section A.2 of the online appendix.

Theorem 1 provides a clear guidance on the choice of an estimation scheme for forecast combination. The asymptotic risk $\mathcal{R}(\boldsymbol{w})$ of a combined-forecast sequence $\{\hat{f}_{t+1|t}(\boldsymbol{w})\}_{t=T}^{T-P+1}$ can be decomposed into two terms: one term $\kappa_{\bar{\lambda}}\boldsymbol{w}^\top \boldsymbol{A}\boldsymbol{w}$ reflects the estimation uncertainty; the other term $\boldsymbol{w}^\top \boldsymbol{B}\boldsymbol{w}$ reflects the local model misspecification. If the asymptotic P/T ratio $\bar{\lambda}$ is greater than 0, then the recursive scheme, in comparison to the fixed and rolling schemes, yields a smaller asymptotic risk $\mathcal{R}(\boldsymbol{w})$ because the term $\kappa_{\bar{\lambda}}\boldsymbol{w}^\top \boldsymbol{A}\boldsymbol{w}$ reflecting the estimation uncertainty decreases with $\bar{\lambda}$. This decrease is in line with our intuitive understanding because the recursive scheme, compared with the fixed and rolling schemes, uses more observations in the estimation of parameters. In contrast, the term $\boldsymbol{w}^\top \boldsymbol{B}\boldsymbol{w}$ reflecting the local model misspecification is invariant to $\bar{\lambda}$ and thus commonly shared by three estimation schemes. In the case of $\bar{\lambda} = 0$, for example $P = 1$ in the context of combining the single forecasts, the recursive scheme has no aforementioned competitive edge over the fixed and rolling schemes in terms of the asymptotic risk $\mathcal{R}(\boldsymbol{w})$. These results are consistent with the implication of Chen and Liu's (2023) Theorem 1 and Corollary 1 established in the univariate mean-forecast context.

2.3 Bregman Model Averaging

Importantly, Theorem 1 suggests that the asymptotically optimal combination-weight vector, accounting for the trade-off between the estimation uncertainty and local model misspecification of the candidate forecasts, should be

$$\boldsymbol{w}_{\bar{\lambda}}^* \coloneqq \underset{\boldsymbol{w}\in\mathbb{W}}{\arg\min}\, \boldsymbol{w}^\top \boldsymbol{\Omega}_{\bar{\lambda}}\boldsymbol{w}.$$

Adopting $\boldsymbol{w}_{\bar{\lambda}}^*$, among all weight vectors in $\mathbb{W}$, and thereby constructing the combined-forecast sequence $\{\hat{f}_{t+1|t}(\boldsymbol{w}_{\bar{\lambda}}^*)\}_{t=T}^{T+P-1}$, we would obtain the smallest averaged Bregman risk provided that T is large. Note that in the recursive scheme, $\boldsymbol{w}_{\bar{\lambda}}^*$ is dependent on the asymptotic P/T ratio $\bar{\lambda}$, whereas in the fixed and rolling schemes, $\boldsymbol{w}_{\bar{\lambda}}^* = \boldsymbol{w}_0^*$ is free of $\bar{\lambda}$. We refer to this asymptotically optimal forecast-combination approach as the Bregman MA approach.

To facilitate the Bregman MA approach, we consider an empirical context where the asymptotically optimal combination-weight vector $\boldsymbol{w}_{\bar{\lambda}}^*$ is estimated in the following way. Suppose that, at time $T_i \coloneqq T + i - 1$ for some $i = 1, \ldots, R$, we estimate θ^* and θ_j^* by their corresponding estimators $\hat{\theta}_{T_i}$ and $\hat{\theta}_{j,T_i}$, which are calculated using the sample $\{(f_t, X_{t-1})\}_{t=2}^{T_i}$. This sample is also used to calculate an estimator $\hat{\mathbb{E}}_i[Z_k Z_j^\top]$ of $\mathbb{E}\left[Z_k Z_j^\top\right]$, the covariance matrix between $\hat{\theta}_{k,T_i}$ and $\hat{\theta}_{j,T_i}$.[6] The estimation-uncertainty matrix $\boldsymbol{A}$ is estimated by $\hat{\boldsymbol{A}}_i$

[6] To facilitate the unified exposition in this section, we defer concrete estimators $\hat{\theta}_{T_i}$ and $\hat{\theta}_{j,T_i}$ to the next section where forecasting targets and candidate models are spelt out.

with its (j,k) element defined as

$$\hat{\boldsymbol{A}}_{i,(j,k)} \coloneqq \frac{1}{T_i-1}\sum_{s=1}^{T_i-1} \operatorname{tr}\left(\nabla_{\theta_j} g_{j,s}(\hat{\theta}_{j,T_i})\nabla^2\psi(g_s(\hat{\theta}_{T_i}))\left[\nabla_{\theta_k} g_{k,s}(\hat{\theta}_{k,T_i})\right]^\top \hat{\mathbb{E}}_i[Z_k Z_j^\top]\right).$$

The squared-bias matrix $\boldsymbol{B}$ is estimated by

$$\hat{\boldsymbol{B}}_i \coloneqq \hat{\boldsymbol{\vartheta}}_{T_i}^\top \left[\frac{1}{T_i-1}\sum_{s=1}^{T_i-1} \nabla_\theta g_s(\hat{\theta}_{T_i})\nabla^2\psi(g_s(\hat{\theta}_{T_i}))\left[\nabla_\theta g_s(\hat{\theta}_{T_i})\right]^\top\right]\hat{\boldsymbol{\vartheta}}_{T_i},$$

where $\hat{\boldsymbol{\vartheta}}_{T_i}$ is the $\bar{k}\times J$ matrix with its j-th column equal to $\sqrt{T}(S_j^\top\hat{\theta}_{j,T_i}-\hat{\theta}_{T_i})$. For every $\lambda\geq 0$, we can solve the quadratic programming problem

$$\hat{\boldsymbol{w}}_{i,\lambda} \coloneqq \underset{\boldsymbol{w}\in\mathbb{W}}{\arg\min}\, \boldsymbol{w}^\top\hat{\boldsymbol{\Omega}}_{i,\lambda}\boldsymbol{w},$$

where $\hat{\boldsymbol{\Omega}}_{i,\lambda}\coloneqq\kappa_\lambda\hat{\boldsymbol{A}}_i+\hat{\boldsymbol{B}}_i$. Taking $\lambda=\hat{\lambda}\coloneqq P/T$, we call $\hat{\boldsymbol{w}}_{i,\hat{\lambda}}$ the 'plug-in' estimator of $\boldsymbol{w}_{\bar{\lambda}}^*$; similarly, replacing $\hat{\lambda}$ with 0, we have the 'fixed' estimator $\hat{\boldsymbol{w}}_{i,0}$. It is worth noting that the potential time variation of $\{\hat{\boldsymbol{w}}_{i,\hat{\lambda}}\}_{i=1}^R$ and $\{\hat{\boldsymbol{w}}_{i,0}\}_{i=1}^R$ is attributed to the nature of recursive estimation. The asymptotically optimal combination-weight vector $\boldsymbol{w}_{\bar{\lambda}}^*$ is, however, time-invariant in our unified framework.

We now discuss the asymptotic properties of $\hat{\boldsymbol{w}}_{i,\hat{\lambda}}$. It is expected that $\hat{\theta}_{T_i}$, $\hat{\theta}_{j,T_i}$, and $\hat{\mathbb{E}}_i[Z_kZ_j^\top]$ are consistent for θ^*, θ_j^*, and $\mathbb{E}\left[Z_kZ_j^\top\right]$, respectively. Thus, $\hat{\boldsymbol{A}}_i$ is consistent for the estimation-uncertainty matrix $\boldsymbol{A}$. By construction, $\hat{\lambda}$ is consistent for $\bar{\lambda}$, and hence $\kappa_{\hat{\lambda}}$ is consistent for $\kappa_{\bar{\lambda}}$. Therefore, the asymptotic property of $\hat{\boldsymbol{w}}_{i,\hat{\lambda}}$ is essentially determined by that of $\hat{\boldsymbol{B}}_i$. To explore the asymptotic property of $\hat{\boldsymbol{B}}_i$, let $\vartheta_{j,T}$ and $\hat{\vartheta}_{j,T_i}$ denote the j-th column of $\boldsymbol{\vartheta}_T$ and $\hat{\boldsymbol{\vartheta}}_{T_i}$, respectively. For every $i=1,\ldots,R$,

$$\begin{aligned}\hat{\vartheta}_{j,T_i}-\vartheta_{j,T} &= \sqrt{T}(S_j^\top\hat{\theta}_{j,T_i}-\hat{\theta}_{T_i})-\sqrt{T}(S_j^\top\theta_j^*-\theta^*)\\ &= \sqrt{\frac{T}{T_i}}\left[S_j^\top\sqrt{T_i}(\hat{\theta}_{j,T_i}-\theta_j^*)-\sqrt{T_i}(\hat{\theta}_{T_i}-\theta^*)\right].\end{aligned}$$

If $\sqrt{T_R}(\hat{\theta}_{j,T_R}-\theta_j^*)=\mathrm{O_p}\left(1\right)$ and $\sqrt{T_R}(\hat{\theta}_{T_R}-\theta^*)=\mathrm{O_p}\left(1\right)$, then

$$\begin{aligned}\hat{\vartheta}_{j,T_R}-\vartheta_{j,T} &= \sqrt{\frac{T}{T+R-1}}\left[S_j^\top\sqrt{T_R}(\hat{\theta}_{j,T_R}-\theta_j^*)-\sqrt{T_R}(\hat{\theta}_{T_R}-\theta^*)\right]\\ &= \sqrt{\frac{T}{T+R-1}}\,\mathrm{O_p}\left(1\right).\end{aligned}$$

The presumption is guaranteed whenever $\hat{\theta}_{j,T}$ and $\hat{\theta}_T$ are $\sqrt{T}$-consistent. Let $r_R\coloneqq T/(T+R-1)$. Following Theorem A.1 of the online appendix, it is expected that as $T\to\infty$, $\hat{\boldsymbol{\vartheta}}_{T_R}-\boldsymbol{\vartheta}_T-r_R^{1/2}\boldsymbol{Z}=\mathrm{o_p}\left(1\right)$ for some $\bar{k}\times J$ matrix $\boldsymbol{Z}$ with every column being normally distributed. Therefore, as $T \to \infty$, every column of $\hat{\boldsymbol{\vartheta}}_{T_R} - \boldsymbol{\vartheta}_T$ should be asymptotically normal provided that r_R does not converge to zero; hence, $\hat{\boldsymbol{B}}_R$ is asymptotically random and inconsistent for $\boldsymbol{B}$. However, if $r_R \to 0$ as $T \to \infty$, then $\hat{\boldsymbol{\vartheta}}_{T_R} - \boldsymbol{\vartheta}_T$ tends to zero in probability and hence $\hat{\boldsymbol{B}}_R$ is consistent for $\boldsymbol{B}$. Following a similar argument to the proof of Theorem 3 of Chen and Liu (2023), it is expected that the plug-in estimator $\hat{\boldsymbol{w}}_{R,\hat{\lambda}}$ is asymptotically random in the case where r_R does not converge to zero as $T \to \infty$; in contrast, it is consistent for $\boldsymbol{w}^*_{\hat{\lambda}}$ whenever $r_R \to 0$ as $T \to \infty$.

3 Applications and Comparison

In this section, we show that the Bregman MA approach is flexibly applicable to different forecasting contexts provided that the forecasting target f_{t+1}, the loss-generating function ψ, the candidate models $\{g_{j,t}\}_{j=1}^J$ and the estimators $\{\hat{\theta}_{j,t}\}_{j=1}^J$ are properly specified and that the strict convexity of ψ on Ψ and the Lipschitz continuity of $\nabla^2\psi$ on $\mathcal{H}$ are verified. Investigating these properties of ψ is important because the asymptotically optimal combination-weight vector $\boldsymbol{w}^*_{\hat{\lambda}}$ depends on the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, both of which in turn depend on the Hessian matrix $\nabla^2\psi$. We apply our approach to generate new MA methods for mean forecast combination in Section 3.1, volatility forecast combination in Section 3.2, probabilistic forecast combination in Section 3.3 and density forecast combination in Section 3.4. In Section 3.5, we further compare our MA methods with related existing MA methods in these four forecasting contexts from methodological viewpoints.

3.1 Mean Forecast Combination

In the mean-forecast context, f_{t+1} is a random vector in $\mathbb{R}^n$, which is realized at time $t+1$, and $f^*_{t+1|t} = \mathbb{E}[f_{t+1}|\mathcal{F}_t]$. Given some $n \times n$ positive definite matrix V, we consider the loss-generating function

$$\psi_m : h \in \mathcal{H} = \Psi = \mathbb{R}^n \mapsto h^\top V^{-1} h.$$

This function is $2\|V\|^{-1}$-strongly convex on $\mathbb{R}^n$,[7] and has the Hessian matrix $\nabla^2\psi_m(h) = 2V^{-1}$ for all $h \in \mathbb{R}^n$. The corresponding BD of h relative to f_{t+1}:

$$D_{\psi_m}(f_{t+1}, h) = (f_{t+1} - h)^\top V^{-1}(f_{t+1} - h)$$

[7] Let $\mu > 0$ be a constant. We say a continuously differentiable function $\psi : \Psi \subseteq \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex on a convex set $\Psi' \subseteq \Psi$ if for any $h_1, h_2 \in \Psi'$,

$$\psi(h_2) \geq \psi(h_1) + \langle \nabla\psi(h_1), h_2 - h_1 \rangle + \frac{\mu}{2}\|h_2 - h_1\|^2.$$

If the function ψ is μ-strongly convex on Ψ', then it is strictly convex on Ψ'. The reader is referred to Boyd and Vandenberghe (2004) and Beck (2017) for further properties of the strong convexity.

is the squared ℓ^2-norm of forecast error $f_{t+1} - h$ weighted by the matrix V^{-1}. In the case where V is the identity matrix, the Bregman risk reduces to the well-known MSFE $\mathbb{E}\left[D_{\psi_m}(f_{t+1}, h)\right] = \mathbb{E}\left[\|f_{t+1} - h\|^2\right]$, which is adopted in "nearly all the work on forecasting with VARs," as argued in Elliott and Timmermann (2016a, p. 186).

In this forecasting context, VAR models serve as representative candidate models (i.e., $\{g_{j,t}\}_{j=1}^{J}$), and the recursive MBD estimator in (1) with $\psi = \psi_m$ and $t = T_i$ simplifies to the LS estimator:

$$\hat{\theta}_{j,T_i} = \underset{\theta_j \in \Theta_j}{\arg\min} \sum_{s=1}^{T_i - 1} \|f_{s+1} - g_{j,s}(\theta_j)\|^2;$$

meanwwhile,

$$\hat{\theta}_{T_i} = \underset{\theta \in \Theta}{\arg\min} \sum_{s=1}^{T_i - 1} \|f_{s+1} - g_s(\theta)\|^2.$$

In the online appendix, we derive the formulae of $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ for this forecasting context. Accordingly, our approach generates an MA method for combining mean-forecast sequences. In the recursive scheme, this method generates a combined mean-forecast sequence $\{\hat{f}_{t+1|t}(\boldsymbol{w})\}_{t=T_i}^{T_i+P-1}$, in which $\hat{f}_{t+1|t}(\boldsymbol{w}) = \sum_{j=1}^{J} w_j g_{j,t}(\hat{\theta}_{j,t})$ and $\boldsymbol{w}$ is set to be the plug-in estimator of $\boldsymbol{w}^*_{\hat{\lambda}}$:

$$\hat{\boldsymbol{w}}_{i,\hat{\lambda}} = \underset{\boldsymbol{w} \in \mathbb{W}}{\arg\min}\, \boldsymbol{w}^\top \hat{\boldsymbol{\Omega}}_{i,\hat{\lambda}} \boldsymbol{w},$$

where $\hat{\boldsymbol{\Omega}}_{i,\hat{\lambda}} = \kappa_{\hat{\lambda}} \hat{\boldsymbol{A}}_i + \hat{\boldsymbol{B}}_i$, $\hat{\boldsymbol{A}}_i$ is the $J \times J$ matrix with its (j,k) element equal to

$$\hat{\boldsymbol{A}}_{i,(j,k)} = \frac{2}{T_i - 1} \sum_{s=1}^{T_i - 1} \operatorname{tr}\left(\left[\nabla_{\theta_j} g_{j,s}(\hat{\theta}_{j,T_i})\right]^\top \hat{\boldsymbol{M}}_{i,jj}^{-1} \hat{\boldsymbol{W}}_{i,jk} \hat{\boldsymbol{M}}_{i,kk}^{-1} \nabla_{\theta_k} g_{k,s}(\hat{\theta}_{k,T_i})\right),$$

$$\hat{\boldsymbol{B}}_i = \hat{\boldsymbol{\vartheta}}_{T_i}^\top \left[\frac{2}{T_i - 1} \sum_{s=1}^{T_i - 1} \nabla_\theta g_s(\hat{\theta}_{T_i}) \left[\nabla_\theta g_s(\hat{\theta}_{T_i})\right]^\top\right] \hat{\boldsymbol{\vartheta}}_{T_i},$$

$$\hat{\boldsymbol{M}}_{i,jj} = \frac{2}{T_i - 1} \sum_{s=1}^{T_i - 1} \left[\nabla_{\theta_j} g_{j,s}(\hat{\theta}_{j,T_i})\right] \left[\nabla_{\theta_j} g_{j,s}(\hat{\theta}_{j,T_i})\right]^\top$$

and

$$\hat{\boldsymbol{W}}_{i,jk} = \frac{4}{T_i - 1} \sum_{s=1}^{T_i - 1} \left[\nabla_{\theta_j} g_{j,s}(\hat{\theta}_{j,T_i})[f_{s+1} - g_s(\hat{\theta}_{T_i})]\right] \left[\nabla_{\theta_k} g_{k,s}(\hat{\theta}_{k,T_i})[f_{s+1} - g_s(\hat{\theta}_{T_i})]\right]^\top,$$

for $j, k = 1, 2, \ldots, J$.

3.2 Volatility Forecast Combination

In the volatility-forecast context, $f_{t+1} = \text{vech}(\Sigma_{t+1})$ and $f^*_{t+1|t} = \text{vech}(\Sigma^*_{t+1|t})$, in which vech is the usual half-vectorization operator, Σ_{t+1} is an $\tilde{n} \times \tilde{n}$ symmetric and positive definite matrix that serves as a volatility proxy at time $t+1$, and $\Sigma^*_{t+1|t} \coloneqq \mathbb{E}[\Sigma_{t+1}|\mathcal{F}_t]$. Given the collection $\mathcal{S}^{++}_{\tilde{n}}$ of all $\tilde{n} \times \tilde{n}$ symmetric and positive definite matrices, we consider the loss-generating function

$$\psi_v : h \in \Psi = \mathcal{L}_{\text{ower}} \circ \mathcal{S}^{++}_{\tilde{n}} \mapsto -\log\{\det(\text{vech}^{-1}(h))\},$$

where $\mathcal{L}_{\text{ower}} \circ \mathcal{S}^{++}_{\tilde{n}} \coloneqq \{\text{vech}(H) \in \mathbb{R}^{\tilde{n}(\tilde{n}+1)/2} : H \in \mathcal{S}^{++}_{\tilde{n}}\}$. This function is well-defined because the determinant of a positive definite matrix $\text{vech}^{-1}(h)$ is always positive.[8] To explore the curvature of ψ_v on $\mathcal{H} = \mathcal{L}_{\text{ower}} \circ \mathcal{M}_{\tilde{n}} \coloneqq \{\text{vech}(H) \in \mathbb{R}^{\tilde{n}(\tilde{n}+1)/2} : H \in \mathcal{M}_{\tilde{n}} \subseteq \mathcal{S}^{++}_{\tilde{n}}\}$ with $\mathcal{M}_{\tilde{n}}$ being a collection of some $\tilde{n} \times \tilde{n}$ symmetric and positive definite matrices, we define $\lambda_{\max}(\mathcal{M}_{\tilde{n}}) \coloneqq \sup\{\lambda_{\max}(H) : H \in \mathcal{M}_{\tilde{n}}\}$ and $\lambda_{\min}(\mathcal{M}_{\tilde{n}}) \coloneqq \inf\{\lambda_{\min}(H) : H \in \mathcal{M}_{\tilde{n}}\}$, where $\lambda_{\max}(H)$ and $\lambda_{\min}(H)$ are the largest and smallest eigenvalue of H, respectively. We write $G_{\tilde{n}}$ for the $\tilde{n}^2 \times \tilde{n}(\tilde{n}+1)/2$ *duplication matrix*, which makes the connection between vec and vech operators by the defining property: $\text{vec}(M) = G_{\tilde{n}}\, \text{vech}(M)$ for every $\tilde{n} \times \tilde{n}$ symmetric matrix M. See Harville (2008) for a more detailed treatment of the duplication matrix. In the online appendix, we prove the following result.

Proposition 2. *Let $\tilde{n} \leq 3$ be a positive integer and $\otimes$ denote the Kronecker multiplication. The Hessian matrix associated with the function ψ_v is*

$$\nabla^2 \psi_v(h) = G^\top_{\tilde{n}} \left([\text{vech}^{-1}(h)]^{-1} \otimes [\text{vech}^{-1}(h)]^{-1}\right) G_{\tilde{n}}$$

for any $h \in \mathcal{L}_{ower} \circ \mathcal{S}^{++}_{\tilde{n}}$. Furthermore, for any $H_1, H_2 \in \mathcal{M}_{\tilde{n}}$,

$$(i) \quad D_{\psi_v}(\text{vech}(H_1), \text{vech}(H_2)) \geq \frac{\|\text{vech}(H_1) - \text{vech}(H_2)\|^2}{2[\lambda_{max}(\mathcal{M}_{\tilde{n}})]^2};$$

$$(ii) \quad \left\|\nabla^2\psi_v(\text{vech}(H_1)) - \nabla^2\psi_v(\text{vech}(H_2))\right\| \leq \frac{4\tilde{n}^{9/2}\lambda_{max}(\mathcal{M}_{\tilde{n}})}{[\lambda_{min}(\mathcal{M}_{\tilde{n}})]^4}\|\text{vech}(H_1) - \text{vech}(H_2)\|.$$

This proposition suggests that if $\tilde{n} \leq 3$, then the function ψ_v is strictly convex on $\mathcal{L}_{\text{ower}} \circ \mathcal{S}^{++}_{\tilde{n}}$ and $[\lambda_{\max}(\mathcal{M}_{\tilde{n}})]^{-2}$-strongly convex on $\mathcal{L}_{\text{ower}} \circ \mathcal{M}_{\tilde{n}}$ provided that $\lambda_{\max}(\mathcal{M}_{\tilde{n}}) < \infty$. If in addition $\lambda_{\min}(\mathcal{M}_{\tilde{n}}) > 0$, then the Hessian matrix $\nabla^2\psi_v$ is Lipschitz continuous on $\mathcal{L}_{\text{ower}} \circ \mathcal{M}_{\tilde{n}}$. In other words, Proposition 2 indicates the conditions on $\mathcal{M}_{\tilde{n}}$ in the context of volatility

[8] Since the operator vech : $\mathcal{S}^{++}_{\tilde{n}} \to \mathcal{L}_{\text{ower}} \circ \mathcal{S}^{++}_{\tilde{n}} \subseteq \mathbb{R}^{\tilde{n}(\tilde{n}+1)/2}$ is a bijective mapping, its inverse mapping $\text{vech}^{-1} : \mathcal{L}_{\text{ower}} \circ \mathcal{S}^{++}_{\tilde{n}} \to \mathcal{S}^{++}_{\tilde{n}}$ is well-defined.

forecasts such that the high-level assumptions in Section 2 are ensured.[9] In Section A.3 of the online appendix, we show that the corresponding BD (of $\text{vech}(H)$ relative to $\text{vech}(\Sigma^*_{t+1|t})$) is

$$D_{\psi_v}(\text{vech}(\Sigma^*_{t+1|t}), \text{vech}(H)) = \log\{\det(H)\} + \text{tr}(H^{-1}\Sigma^*_{t+1|t}) - \log\{\det(\Sigma^*_{t+1|t})\} - \tilde{n}.$$

Replacing the latent matrix $\Sigma^*_{t+1|t}$ with its proxy Σ_{t+1}, we obtain the multivariate QLIKE loss function considered by Patton (2011, Eq. (24)) and Patton and Sheppard (2009, Eq. (50)):

$$\text{QLIKE}(\Sigma_{t+1}, H) \coloneqq \log\{\det(H)\} + \text{tr}(H^{-1}\Sigma_{t+1}) - \log\{\det(\Sigma_{t+1})\} - \tilde{n}.$$

The resulting Bregman risk, as a criterion of forecasting performance, is robust in the sense that the ranking of competing volatility forecasts is invariant to the replacement of $\Sigma^*_{t+1|t}$ with Σ_{t+1}, as indicated in Proposition 2 of Laurent et al. (2013).

In financial applications, it is common to set $\Sigma_{t+1} = u_{t+1}u_{t+1}^\top$, where u_{t+1} denotes an $\tilde{n} \times 1$ column vector that is observable at $t+1$ and satisfies the restriction $\mathbb{E}[u_{t+1}|\mathcal{F}_t] = 0$.[10] In this context, multivariate GARCH models, such as the BEKK model in Engle and Kroner (1995), the dynamic conditional correlation model in Engle (2002), and the varying correlation model in Tse and Tsui (2002), serve as representative candidate models (i.e., $\{g_{j,t}\}_{j=1}^J$). The recursive MBD estimator in (1) with $\psi = \psi_v$ and $t = T_i$ corresponds to the Gaussian quasi-ML (QML) estimator:

$$\hat{\theta}_{j,T_i} = \underset{\theta_j \in \Theta_j}{\arg\max} \sum_{s=1}^{T_i-1} \left\{ -\frac{\tilde{n}}{2}\log\{2\pi\} - \frac{1}{2}\log\{\det(g_{j,s}(\theta_j))\} - \frac{1}{2}u_{s+1}^\top \left[g_{j,s}(\theta_j)\right]^{-1} u_{s+1} \right\};$$

meanwhile,

$$\hat{\theta}_{T_i} = \underset{\theta \in \Theta}{\arg\max} \sum_{s=1}^{T_i-1} \left\{ -\frac{\tilde{n}}{2}\log\{2\pi\} - \frac{1}{2}\log\{\det(g_s(\theta))\} - \frac{1}{2}u_{s+1}^\top \left[g_s(\theta)\right]^{-1} u_{s+1} \right\}.$$

For every j and t, let $g_t^{\text{vec}} \coloneqq \text{vec}(g_t)$ and $g_{j,s}^{\text{vec}} \coloneqq \text{vec}(g_{j,t})$. In the online appendix, we further derive the formulae of $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ for this forecasting context. Using the formulae, our approach generates an MA method for combining volatility-forecast sequences. In the recursive scheme, this method generates a combined volatility-forecast sequence $\{\hat{f}_{t+1|t}(\boldsymbol{w})\}_{t=T_i}^{T_i+P-1}$, in which $\hat{f}_{t+1|t}(\boldsymbol{w}) = \sum_{j=1}^J w_j g_{j,t}(\hat{\theta}_{j,t})$ and $\boldsymbol{w}$ is set to be the plug-in estimator of $\boldsymbol{w}^*_{\hat{\lambda}}$:

$$\hat{\boldsymbol{w}}_{i,\hat{\lambda}} = \underset{\boldsymbol{w} \in \mathbb{W}}{\arg\min}\, \boldsymbol{w}^\top \hat{\boldsymbol{\Omega}}_{i,\hat{\lambda}} \boldsymbol{w},$$

[9] We conjecture that Proposition 2 also holds for every positive integer $\tilde{n} > 3$. To the best of our knowledge, Proposition 2 is the first attempt in literature to quantify the convexity of the function ψ_v in the cases where $\tilde{n} > 1$.

[10] We note in passing that the matrix $u_{t+1}u_{t+1}^\top$ is not positive definite (but positive semidefinite) and thus $\det(u_{t+1}u_{t+1}^\top) = 0$. It would be innocuous to the evaluation of forecasting performances if we remove terms irrelevant to the choice of a forecasting model from the QLIKE loss function and consider the revised QLIKE function $\widetilde{\text{QLIKE}}(u_{t+1}u_{t+1}^\top, H) \coloneqq \log\{\det(H)\} + u_{t+1}^\top H^{-1} u_{t+1}$.

where $\hat{\boldsymbol{\Omega}}_{i,\hat{\lambda}} = \kappa_{\hat{\lambda}} \hat{\boldsymbol{A}}_i + \hat{\boldsymbol{B}}_i$, $\hat{\boldsymbol{A}}_i$ is the $J \times J$ matrix with its (j,k) element satisfying

$$\begin{aligned}&(T_i - 1)\hat{\boldsymbol{A}}_{i,(j,k)} \\ &= \sum_{s=1}^{T_i-1} \operatorname{tr}\left(\nabla_{\theta_j} g_{j,s}^{\text{vec}}(\hat{\theta}_{j,T_i}) \left[[g_s(\hat{\theta}_{T_i})]^{-1} \otimes [g_s(\hat{\theta}_{T_i})]^{-1} \right] \left[\nabla_{\theta_k} g_{k,s}^{\text{vec}}(\hat{\theta}_{k,T_i}) \right]^\top \hat{\boldsymbol{M}}_{i,kk}^{-1} \hat{\boldsymbol{W}}_{i,kj} \hat{\boldsymbol{M}}_{i,jj}^{-1} \right),\end{aligned}$$

$$\hat{\boldsymbol{B}}_i = \hat{\boldsymbol{\vartheta}}_{T_i}^\top \left[\frac{1}{T_i - 1} \sum_{s=1}^{T_i-1} \nabla_\theta g_s^{\text{vec}}(\hat{\theta}_{T_i}) \left[[g_s(\hat{\theta}_{T_i})]^{-1} \otimes [g_s(\hat{\theta}_{T_i})]^{-1} \right] \left[\nabla_\theta g_s^{\text{vec}}(\hat{\theta}_{T_i}) \right]^\top \right] \hat{\boldsymbol{\vartheta}}_{T_i}.$$

$$(T_i - 1)\hat{\boldsymbol{M}}_{i,jj} = \sum_{s=1}^{T_i-1} \nabla_{\theta_j} g_{j,s}^{\text{vec}}(\hat{\theta}_{j,T_i}) \left[[g_{j,s}(\hat{\theta}_{j,T_i})]^{-1} \otimes [g_{j,s}(\hat{\theta}_{j,T_i})]^{-1} \right] \left[\nabla_{\theta_j} g_{j,s}^{\text{vec}}(\hat{\theta}_{j,T_i}) \right]^\top$$

and

$$\begin{aligned}&(T_i - 1)\hat{\boldsymbol{W}}_{i,kj} \\ &= \sum_{s=1}^{T_i-1} \left[\nabla_{\theta_k} g_{k,s}^{\text{vec}}(\hat{\theta}_{k,T_i}) \left[[g_{k,s}(\hat{\theta}_{k,T_i})]^{-1} \otimes [g_{k,s}(\hat{\theta}_{k,T_i})]^{-1} \right] \operatorname{vec}\left(\Sigma_{s+1} - g_s(\hat{\theta}_{T_i}) \right) \right] \\ &\qquad \cdot \left[\nabla_{\theta_j} g_{j,s}^{\text{vec}}(\hat{\theta}_{j,T_i}) \left[[g_{j,s}(\hat{\theta}_{j,T_i})]^{-1} \otimes [g_{j,s}(\hat{\theta}_{j,T_i})]^{-1} \right] \operatorname{vec}\left(\Sigma_{s+1} - g_s(\hat{\theta}_{T_i}) \right) \right]^\top,\end{aligned}$$

for $j, k = 1, 2, \ldots, J$.

3.3 Probabilistic Forecast Combination

In the probabilistic-forecast context, $f_{t+1} = (f_{t+1,1}, \ldots, f_{t+1,n})^\top \in \mathbb{R}^n$ is one-hot encoded; specifically, $\sum_{\ell=1}^n f_{t+1,\ell} = 1$ and every $f_{t+1,\ell}$ is either zero or one. Correspondingly, the conditional mean $f_{t+1|t}^* = \mathbb{E}[f_{t+1}|\mathcal{F}_t]$ represents the probability parameter of a multinomial distribution conditional on $\mathcal{F}_t$. We consider the loss-generating function

$$\psi_p : h = (h_1, \cdots, h_n)^\top \in \Delta_0 \mapsto \sum_{\ell=1}^n h_\ell \log\{h_\ell\},$$

where $\Delta_0 := \left\{ (p_1, \cdots, p_n)^\top \in \mathbb{R}^n : \min\{p_1, \ldots, p_n\} \geq 0 \text{ and } \sum_{\ell=1}^n p_\ell = 1 \right\}$ is an $(n-1)$-dimensional simplex. We agree on the convention $0 \log\{0\} := 0$. In this application, we take $\Psi = \Delta_0$ and $\mathcal{H} = \Delta_{p_0} := \left\{ (p_1, \cdots, p_n)^\top \in \mathbb{R}^n : \min\{p_1, \ldots, p_n\} \geq p_0 \text{ and } \sum_{\ell=1}^n p_\ell = 1 \right\}$ with some constant $p_0 \in (0, 1/n)$. The function ψ_p is strictly convex on Δ_0; furthermore, it is twice continuously differentiable on Δ_{p_0} and its Hessian matrix $\nabla^2 \psi_p(h)$ is a diagonal matrix with the (ℓ, ℓ) element equal to the inverse of the ℓth element of h for all $h \in \Delta_{p_0}$. It is thus straightforward to show that on Δ_{p_0}, ψ_p is 1-strongly convex and $\nabla^2 \psi_p$ is Lipschitz continuous.

The corresponding BD of $h = (h_1, \dots, h_n)^\top$ relative to $f^*_{t+1|t} = (f^*_{t+1|t,1}, \dots, f^*_{t+1|t,n})^\top$ is the KL divergence of h from $f^*_{t+1|t}$ minus a term irrelevant to the choice of h; to be specific,

$$\begin{aligned} D_{\psi_p}(f^*_{t+1|t}, h) &= -\sum_{\ell=1}^{n} f^*_{t+1|t,\ell} \log\{h_\ell\} \\ &= \underbrace{\sum_{\ell=1}^{n} f^*_{t+1|t,\ell} \log\left\{\frac{f^*_{t+1|t,\ell}}{h_\ell}\right\}}_{:=\mathrm{KL}(f^*_{t+1|t},\, h)} - \sum_{\ell=1}^{n} f^*_{t+1|t,\ell} \log\{f^*_{t+1|t,\ell}\}. \end{aligned}$$

If h is an $\mathcal{F}_t$-measurable parametric function, this BD can also be viewed as the negative expected quasi-log-likelihood conditional on $\mathcal{F}_t$; particularly,

$$D_{\psi_p}(f^*_{t+1|t}, h) = -\,\mathbb{E}\left[\sum_{\ell=1}^{n} f_{t+1,\ell} \log\{h_\ell\} \middle| \mathcal{F}_t\right].$$

In this forecasting context, multinomial logit models, or multinomial probit models, serve as representative candidate models (i.e., $\{g_{j,t}\}_{j=1}^{J}$), and the recursive MBD estimator in (1) with $\psi = \psi_p$ and $t = T_i$ reduces to the QML estimator:

$$\hat{\theta}_{j,T_i} = \underset{\theta_j \in \Theta_j}{\arg\max} \sum_{s=1}^{T_i-1} \sum_{\ell=1}^{n} f_{s+1,\ell} \log\{g_{j,s,\ell}(\theta_j)\},$$

where $g_{j,s,\ell}$ denotes the ℓth element of $g_{j,s}$; meanwhile,

$$\hat{\theta}_{T_i} = \underset{\theta \in \Theta}{\arg\max} \sum_{s=1}^{T_i-1} \sum_{\ell=1}^{n} f_{s+1,\ell} \log\{g_{s,\ell}(\theta)\},$$

with $g_{s,\ell}$ being the ℓth element of g_s. In the online appendix, we further derive the formulae of $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ for this forecasting context. Given the formulae, our approach generates an MA method for combining probabilistic-forecast sequences. In the recursive scheme, this method generates a combined probabilistic-forecast sequence $\{\hat{f}_{t+1|t}(\boldsymbol{w})\}_{t=T_i}^{T_i+P-1}$, in which $\hat{f}_{t+1|t}(\boldsymbol{w}) = \sum_{j=1}^{J} w_j g_{j,t}(\hat{\theta}_{j,t})$ and $\boldsymbol{w}$ is set to be the plug-in estimator of $\boldsymbol{w}^*_{\hat{\lambda}}$:

$$\hat{\boldsymbol{w}}_{i,\hat{\lambda}} = \underset{\boldsymbol{w} \in \mathbb{W}}{\arg\min}\, \boldsymbol{w}^\top \hat{\boldsymbol{\Omega}}_{i,\hat{\lambda}} \boldsymbol{w},$$

where $\hat{\boldsymbol{\Omega}}_{i,\hat{\lambda}} = \kappa_{\hat{\lambda}} \hat{\boldsymbol{A}}_i + \hat{\boldsymbol{B}}_i$, $\hat{\boldsymbol{A}}_i$ is the $J \times J$ matrix with its (j,k) element equal to

$$\hat{\boldsymbol{A}}_{i,(j,k)} = \frac{1}{T_i - 1} \sum_{s=1}^{T_i-1} \sum_{\ell=1}^{n} \left[g_{s,\ell}(\hat{\theta}_{T_i})\right]^{-1} \left[\nabla_{\theta_j} g_{j,s,\ell}(\hat{\theta}_{j,T_i})\right]^\top \hat{\boldsymbol{M}}^{-1}_{i,jj} \hat{\boldsymbol{W}}_{i,jk} \hat{\boldsymbol{M}}^{-1}_{i,kk} \nabla_{\theta_k} g_{k,s,\ell}(\hat{\theta}_{k,T_i}),$$

$$\hat{\boldsymbol{B}}_i = \hat{\boldsymbol{\vartheta}}^\top_{T_i} \left[\frac{1}{T_i - 1} \sum_{s=1}^{T_i-1} \sum_{\ell=1}^{n} \left[g_{s,\ell}(\hat{\theta}_{T_i})\right]^{-1} \nabla_\theta g_{s,\ell}(\hat{\theta}_{T_i}) \left[\nabla_\theta g_{s,\ell}(\hat{\theta}_{T_i})\right]^\top\right] \hat{\boldsymbol{\vartheta}}_{T_i},$$

$$(T_i-1)\hat{\boldsymbol{M}}_{i,jj} = \sum_{s=1}^{T_i-1}\sum_{\ell=1}^{n}\left[g_{j,s,\ell}(\hat{\theta}_{j,T_i})\right]^{-1}\nabla_{\theta_j} g_{j,s,\ell}(\hat{\theta}_{j,T_i})\left[\nabla_{\theta_j} g_{j,s,\ell}(\hat{\theta}_{j,T_i})\right]^\top$$

and

$$\begin{aligned}&(T_i-1)\hat{\boldsymbol{W}}_{i,jk}\\ =&\sum_{s=1}^{T_i-1}\left[\sum_{\ell=1}^{n}\frac{[f_{s+1,\ell}-g_{s,\ell}(\hat{\theta}_{T_i})]\nabla_{\theta_j} g_{j,s,\ell}(\hat{\theta}_{j,T_i})}{g_{j,s,\ell}(\hat{\theta}_{j,T_i})}\right]\left[\sum_{\ell=1}^{n}\frac{[f_{s+1,\ell}-g_{s,\ell}(\hat{\theta}_{T_i})]\nabla_{\theta_k} g_{k,s,\ell}(\hat{\theta}_{k,T_i})}{g_{k,s,\ell}(\hat{\theta}_{k,T_i})}\right]^\top,\end{aligned}$$

for $j,k=1,2,\dots,J$.

3.4 Density Forecast Combination

In the density-forecast context, we are interested in forecasting the probability density function of y_{t+1}, which is a continuous random variable realized at time $t+1$, using X_t. Our approach is applicable to this context by applying the MA method proposed in Section 3.3 to the implied probabilistic-forecasting setting of density forecast in the limiting case where n goes to infinity.

To define the implied probabilistic-forecast setting, we consider a sequence $\{z_0, z_1, \dots, z_{\tilde{n}+1}\}$ such that $-M = z_0 < z_1 < \cdots < z_{\tilde{n}} < z_{\tilde{n}+1} = M$, for some large $M > 0$, and

$$z_i - z_{i-1} = \begin{cases} 2\Delta_{\tilde{n}}, & \text{if } i = 2,\dots,\tilde{n};\\ \Delta_{\tilde{n}}, & \text{otherwise,}\end{cases}$$

with $\Delta_{\tilde{n}} \coloneqq M/\tilde{n}$, and partition $\mathbb{R}$ into $\tilde{n}+2$ intervals:

$$\mathcal{I}_\ell = \begin{cases}(-\infty, -M], & \text{if } \ell = 1;\\ (z_{\ell-1}-\Delta_{\tilde{n}}, z_{\ell-1}+\Delta_{\tilde{n}}], & \text{if } 2 \le \ell \le \tilde{n}+1;\\ (M, \infty), & \text{if } \ell = \tilde{n}+2.\end{cases}$$

Let $n = \tilde{n}+2$ and $f_{t+1} = (f_{t+1,1},\dots,f_{t+1,n})^\top$ be the $n\times 1$ vector with its ℓth element $f_{t+1,\ell}$ equal to one if $y_{t+1}\in\mathcal{I}_\ell$, and zero otherwise. Correspondingly, $f^*_{t+1|t} = \mathbb{E}[f_{t+1}|\mathcal{F}_t]$ is the probability parameter of the multinomial distribution of f_{t+1} conditional on $\mathcal{F}_t = \sigma(X_t)$, the sigma field generated by X_t. Specifically, we have $f^*_{t+1|t} = (f^*_{t+1|t,1},\dots,f^*_{t+1|t,n})^\top$ with the ℓth element equal to

$$f^*_{t+1|t,\ell} = \int_{\mathcal{I}_\ell}\varphi_{t+1|t}(y)\,\mathrm{d}y,$$

where $\varphi_{t+1|t}$ denotes the probability density function of y_{t+1} conditional on X_t. Let Θ be a subset of the $\bar{k}$-dimensional Euclidean space and $h_{t+1|t}(\cdot\,;\theta)$ denote a generic probability density function conditional on X_t for every $\theta \in \Theta$. For every $\theta \in \Theta$, we let $h_{t+1|t}^{(n)}(\theta)$ be an $n \times 1$ vector with the ℓth element:

$$h_{t+1|t,\ell}^{(n)}(\theta) := \int_{\mathcal{I}_\ell} h_{t+1|t}(y;\theta)\,\mathrm{d}y,$$

for $\ell = 1, 2, \ldots, n$. We can apply the loss-generating function ψ_p and the BD $D_{\psi_p}(f_{t+1|t}^*, h)$ to the aforementioned f_{t+1} and the probabilistic model $h = h_{t+1|t}^{(n)}(\theta)$ for every n in this implied probabilistic-forecast setting.

As discussed in the probabilistic-forecast context, we consider the recursive MBD estimator in (1) with $\psi = \psi_p$ and $t = T_i$, which reduces to the QML estimator based on the probabilistic model $h_{t+1|t}^{(n)}(\theta)$:

$$\hat{\theta}_{T_i}^{(h,n)} := \underset{\theta\in\Theta}{\arg\max} \sum_{s=1}^{T_i-1} \sum_{\ell=1}^{n} f_{s+1,\ell} \log\{h_{s+1|s,\ell}^{(n)}(\theta)\}.$$

In comparison, the conditional probability density model $h_{t+1|t}(\cdot\,;\theta)$ has the original QML estimator:

$$\hat{\theta}_{T_i}^{(h)} := \underset{\theta\in\Theta}{\arg\max} \sum_{s=1}^{T_i-1} \log\{h_{s+1|s}(y_{s+1};\theta)\}. \tag{2}$$

We denote their population counterparts by

$$\theta^{(h,n)} := \underset{\theta\in\Theta}{\arg\max}\, \mathbb{E}\left[\sum_{\ell=1}^{n} f_{t+1,\ell} \log\left\{h_{t+1|t,\ell}^{(n)}(\theta)\right\}\right]$$

and

$$\theta^{(h)} := \underset{\theta\in\Theta}{\arg\max}\, \mathbb{E}\left[\log\left\{h_{t+1|t}(y_{t+1};\theta)\right\}\right]. \tag{3}$$

Proposition 3 below states the main result of the implied probabilistic-forecasting setting in the limiting case where n goes to infinity: under regularity conditions, the difference between the implied QML estimator $\hat{\theta}_{T_i}^{(h,n)}$ and the original QML estimator $\hat{\theta}_{T_i}^{(h)}$ is asymptotically negligible.

Proposition 3. *Suppose that Conditions (A)-(F) listed in Section A.4 of the online appendix hold. If for every n, $\hat{\theta}_{T_i}^{(h,n)}$ converges in probability to $\theta^{(h,n)}$ and $\hat{\theta}_{T_i}^{(h)}$ converges in probability to $\theta^{(h)}$ as $T_i \to \infty$, then for every $\eta > 0$, there is a positive integer $\bar{n}$ such that for all $n \geq \bar{n}$,*

$$\lim_{T_i\to\infty} \mathbb{P}\left(\left\|\hat{\theta}_{T_i}^{(h,n)} - \hat{\theta}_{T_i}^{(h)}\right\| > \eta\right) = 0.$$

In this proposition, the consistency of $\hat{\theta}_{T_i}^{(h,n)}$ for every n and that of $\hat{\theta}_{T_i}^{(h)}$ are required and should be valid under mild primitive assumptions; moreover, Conditions (A)-(F) are imposed to ensure that as n tends to infinity, the expected KL divergence $\mathbb{E}\left[\mathrm{KL}\left(f^*_{t+1|t}, h^{(n)}_{t+1|t}(\theta)\right)\right]$ converges uniformly on Θ to the expected KL divergence $\mathbb{E}\left[\mathrm{KL}\left(\varphi_{t+1|t}(\cdot), h_{t+1|t}(\cdot\,;\theta)\right)\right]$.

The conditional density function $\varphi_{t+1|t}$ could be specified in different manners. In financial applications, univariate GARCH-type models with different choices of conditional distribution specification serve as representative examples. Taking specification searches into consideration, we address the density forecast combination by applying the proposed MA approach to the implied probabilistic-forecast setting in the limiting case. For every $\theta \in \Theta \subseteq \mathbb{R}^{\bar{k}}$, let $g_t(\cdot\,;\theta)$ be the unrestricted model of $\varphi_{t+1|t}$. For every $j = 1, \ldots, J$, let Θ_j be the subset of $\mathbb{R}^{k_j}$, as defined in the beginning of Section 2.2. For every $\theta_j \in \Theta_j$, we write $g_{j,t}(\cdot\,;\theta_j) \coloneqq g_t(\cdot\,;S_j^\top \theta_j)$ for the candidate models of $\varphi_{t+1|t}$. We let $\hat{\theta}_{T_i}$ be the estimator in (2) and $\theta^\star$ the population parameter in (3) by setting $h_{t+1|t}(\cdot\,;\theta) = g_t(\cdot\,;\theta)$. Similarly, we write $\hat{\theta}_{j,T_i}$ for the estimator in (2) and $\theta_j^\star$ for the population parameter in (3) with $h_{t+1|t}(\cdot\,;\theta)$ replaced by $g_{j,t}(\cdot\,;\theta_j)$ and maximum being taken over Θ_j. Under suitable regularity conditions, the implied probabilistic-forecast setting should have the limits:

$$\lim_{n\to\infty} \boldsymbol{A}_{(j,k)} = \lim_{n\to\infty} \mathbb{E}\left[\frac{g_{j,t}(y_{t+1};\theta_j^\star) g_{k,t}(y_{t+1};\theta_k^\star)}{[g_t(y_{t+1};\theta^\star)]^2}\right.$$
$$\left.\cdot \left[\nabla_{\theta_j} \log\left\{g_{j,t}(y_{t+1};\theta_j^\star)\right\}\right]^\top \boldsymbol{M}_{jj}^{-1} \boldsymbol{W}_{jk} \boldsymbol{M}_{kk}^{-1} \nabla_{\theta_k} \log\left\{g_{k,t}(y_{t+1};\theta_k^\star)\right\}\right],$$

$$\lim_{n\to\infty} \boldsymbol{B} = \lim_{T\to\infty} \boldsymbol{\vartheta}_T^\top \mathbb{E}\left[\nabla_\theta \log\left\{g_t(y_{t+1};\theta^\star)\right\} \left[\nabla_\theta \log\left\{g_t(y_{t+1};\theta^\star)\right\}\right]^\top\right] \boldsymbol{\vartheta}_T,$$

$$\lim_{n\to\infty} \boldsymbol{M}_{jj} = \mathbb{E}\left[\frac{g_{j,t}(y_{t+1};\theta_j^\star)}{g_t(y_{t+1};\theta^\star)} \nabla_{\theta_j} \log\left\{g_{j,t}(y_{t+1};\theta_j^\star)\right\} \left[\nabla_{\theta_j} \log\left\{g_{j,t}(y_{t+1};\theta_j^\star)\right\}\right]^\top\right]$$

and

$$\lim_{n\to\infty} \boldsymbol{W}_{jk} = \mathbb{E}\left[\nabla_{\theta_j} \log\left\{g_{j,t}(y_{t+1};\theta_j^\star)\right\} \left(\nabla_{\theta_k} \log\left\{g_{k,t}(y_{t+1};\theta_k^\star)\right\}\right)^\top\right].$$

Thus, our approach generates an MA method for density forecast combination. In the recursive scheme, this method generates a combined density-forecast sequence $\left\{\sum_{j=1}^J w_j g_{j,t}(\cdot\,;\hat{\theta}_{j,t})\right\}_{t=T_i}^{T_i+P-1}$, by setting $\boldsymbol{w}$ to be the plug-in estimator of $\boldsymbol{w}^*_{\hat{\lambda}}$:

$$\hat{\boldsymbol{w}}_{i,\hat{\lambda}} = \arg\min_{\boldsymbol{w}\in\mathbb{W}} \boldsymbol{w}^\top \hat{\boldsymbol{\Omega}}_{i,\hat{\lambda}} \boldsymbol{w},$$

where $\hat{\boldsymbol{\Omega}}_{i,\hat{\lambda}} = \kappa_{\hat{\lambda}}\hat{\boldsymbol{A}}_i + \hat{\boldsymbol{B}}_i$, $\hat{\boldsymbol{A}}_i$ is a $J \times J$ matrix with its (j,k) element satisfying

$$
\begin{aligned}
(T_i - 1)\hat{\boldsymbol{A}}_{i,(j,k)} = \sum_{s=1}^{T_i-1} & \frac{g_{j,s}(y_{s+1};\hat{\theta}_{j,T_i})g_{k,s}(y_{s+1};\hat{\theta}_{k,T_i})}{[g_s(y_{s+1};\hat{\theta}_{T_i})]^2} \\
& \cdot \left[\nabla_{\theta_j} \log\left\{g_{j,s}(y_{s+1};\hat{\theta}_{j,T_i})\right\}\right]^\top \hat{\boldsymbol{M}}_{i,jj}^{-1}\hat{\boldsymbol{W}}_{i,jk}\hat{\boldsymbol{M}}_{i,kk}^{-1}\nabla_{\theta_k}\log\left\{g_{k,s}(y_{s+1};\hat{\theta}_{k,T_i})\right\},
\end{aligned}
$$

$$
\hat{\boldsymbol{B}}_i = \hat{\boldsymbol{\vartheta}}_{T_i}^\top \left[\frac{1}{T_i - 1}\sum_{s=1}^{T_i-1} \nabla_\theta \log\left\{g_s(y_{s+1};\hat{\theta}_{T_i})\right\}\left[\nabla_\theta \log\left\{g_s(y_{s+1};\hat{\theta}_{T_i})\right\}\right]^\top\right]\hat{\boldsymbol{\vartheta}}_{T_i},
$$

$$
\begin{aligned}
& (T_i - 1)\hat{\boldsymbol{M}}_{i,jj} \\
= & \sum_{s=1}^{T_i-1}\left[\frac{g_{j,s}(y_{s+1};\hat{\theta}_{j,T_i})}{g_s(y_{s+1};\hat{\theta}_{T_i})}\nabla_{\theta_j}\log\left\{g_{j,s}(y_{s+1};\hat{\theta}_{j,T_i})\right\}\left[\nabla_{\theta_j}\log\left\{g_{j,s}(y_{s+1};\hat{\theta}_{j,T_i})\right\}\right]^\top\right]
\end{aligned}
$$

and

$$
(T_i - 1)\hat{\boldsymbol{W}}_{i,jk} = \sum_{s=1}^{T_i-1}\left\{\nabla_{\theta_j}\log\left\{g_{j,s}(y_{s+1};\hat{\theta}_{j,T_i})\right\}\left(\nabla_{\theta_k}\log\left\{g_{k,s}(y_{s+1};\hat{\theta}_{k,T_i}))\right\}\right)^\top\right\},
$$

for $j,k = 1,2,\ldots,J$.

3.5 Comparison with Existing Methods

In the mean-forecast context, Liao et al. (2019) and Liao and Tsay (2020) also proposed MA methods for VAR forecast combination and established the asymptotic optimality of their methods using the fixed-parameter asymptotics under the assumption that the true model is VAR(∞). Our method is different from theirs in terms of the asymptotic analysis and optimality. Following Chen and Liu (2023) and others, our method is built on the local-asymptotics setting. Lohmeyer et al. (2019) also proposed an MA method using the local asymptotics in the context of VAR models. However, their MA method is designed for estimating a focused parameter rather than for forecast combination. Our method is also different from these methods in terms of model flexibility. Specifically, it includes the MA method of Chen and Liu (2023) as a special case where $n = 1$ and the candidate models are linear predictive regressions, and is applicable to a generalized context where the mean forecasts could be generated from univariate or multivariate conditional mean models or from linear or nonlinear conditional mean models. This flexibility is essential because it allows for a wider scope of empirical applications.

In the volatility-forecast context, the related frequentist MA methods are relatively rare. Qiu et al. (2019) considered an MA method for forecasting volatility of financial returns.

Their method is based on the LS estimation of heterogeneous autoregression models of realized volatility, and the asymptotic optimality is defined by minimizing the MSFE. Liu et al. (2020) proposed an MA method for forecasting volatility based on GARCH-type models, and defined the asymptotic optimality of their method by minimizing the KL divergence. In both studies, the MA methods are univariate and based on the fixed-parameter asymptotics. Our method is different from theirs in terms of the asymptotic analysis, optimality and dimension flexibility. In particular, our method is applicable to not only univariate volatility models but also multivariate volatility models. Considering multivariate volatility models is essential for forecasting volatility *and* co-volatility that are of considerable interest in financial applications such as the time-varying β-risk in asset pricing and the time-varying optimal hedge ratio in risk management.[11]

In the probabilistic-forecast context, Zhang et al. (2016) and Ando and Li (2017) considered MA methods for generalized linear models estimated by the ML method, and established the asymptotic optimality of their methods by minimizing the KL divergence based on the fixed-parameter asymptotics; see also Zhang and Liu (2023) for a related MA method. However, these methods commonly assume that the dependent variables and covariates are random samples, and hence are not designed for the probabilistic forecast combination in time series. In comparison, our method is applicable to probabilistic forecasts in the time-series context. In empirical applications, our method complements the existing methods of forecasting recession, such as those mentioned in the introduction, by combining the probabilistic forecasts generated from different binary-response models in an asymptotically optimal way when $n = 2$. It is also applicable to forecasting the recession/expansion status of multiple economies, or the bull/bear status of multiple markets, at the same time by suitably defining the one-hot encoded forecasting target in the case where $n > 2$.

In the density-forecast context, there is a growing interest in establishing MA methods for forecast combination. Hall and Mitchell (2007) established their MA method by maximizing the average log score of the combined density forecast, and Geweke and Amisano (2011) further explored theoretical properties of this method. Jore et al. (2010) and Pauwels and Vasnev (2016) proposed related MA methods based on the exponential weighting scheme and the smoothed log predictive scores, respectively, while Conflitti et al. (2015) and Diebold et al. (2023) proposed a simple iterative algorithm and a regularization method to compute the weights. Kapetanios et al. (2015) allowed the combination weights to depend on the forecasting target, and Pauwels et al. (2023) proposed an approach to compute the optimal

[11] In these applications, the multivariate GARCH has become the workhorse model of the conditional covariance matrix. For the time-varying β-risk, see Bollerslev et al. (1988) and Ng (1991), among others; for the time-varying optimal hedge ratio, see Baillie and Myers (1991), Lien and Tse (2002), Park and Jei (2010), and the references therein.

weights subject to additional higher moments restrictions. Kapetanios et al. (2015) and Pauwels et al. (2023) further investigated theoretical properties of their method based on the fixed-parameter asymptotics. Our method is different from theirs in terms of asymptotic methods and theoretical designs. In particular, unlike these methods, we formulate the density forecast combination as a limiting case of the implied probabilistic forecast combination as n goes to infinity. This highlights the connection between the two seemingly different forecasting problems that are both essential in empirical applications.

In addition, most of the aforementioned MA methods are designed for combining single forecasts in the case where $P = 1$. Following Chen and Liu (2023), our methods are designed for combining forecast sequences with a general P. The MA methods proposed in Sections 3.1-3.4 are built on the recursive scheme. Since our approach is also applicable to the fixed scheme (or the rolling scheme) that corresponds to the case where $P = 1$, we may also apply our MA methods to combining single forecasts by setting $\boldsymbol{w}$ to be the 'fixed' estimator:

$$\hat{\boldsymbol{w}}_{i,0} = \arg\min_{\boldsymbol{w}\in\mathbb{W}} \boldsymbol{w}^\top \hat{\boldsymbol{\Omega}}_{i,0} \boldsymbol{w},$$

where $\hat{\boldsymbol{\Omega}}_{i,0} = \hat{\boldsymbol{A}}_i + \hat{\boldsymbol{B}}_i$. As mentioned by Chen and Liu (2023), considering the combination of forecast sequences is essential not only for generalization but also for reflecting the fact that econometric models may be routinely applied to generating economic forecasts in a long period rather than at one single time. Moreover, it has important implications on the asymptotic risk of our MA method that are not shared by focusing on the combination of single forecasts. In particular, it is important to distinguish between the recursive scheme and the fixed scheme in combining forecast sequences because the former is theoretically better than the latter by allowing (the estimation-uncertainty part of) the asymptotic risk to decrease with the asymptotic P/T ratio $\bar{\lambda}$, as implied by Theorem 1.

More importantly, unlike all the aforementioned existing methods, our MA methods are generated from the Bregman MA approach in a unified way. We utilize the examples in this section to illustrate that our approach is useful for generating new MA methods in different forecasting contexts.

4 Simulation

In this section, we conduct two simulation experiments to compare our approach with alternative MA methods in their finite-sample forecasting performance. We consider a problem of combining VAR forecasts in Section 4.1 and a problem of combining density forecasts in Section 4.2.

4.1 VAR Forecast Combination

In this experiment, we extend the simulation of Chen and Liu (2023) from a univariate context of mean forecast combinations to a multivariate context of VAR forecast combinations. We consider J candidate models: VAR(j), for $j = 1, \ldots, J$, and a set of MA methods for VAR forecast combination that generate the following $\boldsymbol{w}$'s at $t = T_i$:

- Plug-In: $\hat{\boldsymbol{w}}_{i,\hat{\lambda}}$;

- Fixed: $\hat{\boldsymbol{w}}_{i,0}$;

- LsoMA: $\hat{\boldsymbol{w}}_{i,LsoMA} \coloneqq \arg\min_{\boldsymbol{w}\in\mathbb{W}} CV_i(\boldsymbol{w})$, where
$$CV_i(\boldsymbol{w}) \coloneqq \operatorname{tr}(\tilde{\epsilon}_i(\boldsymbol{w})\hat{\boldsymbol{\Sigma}}_i^{-1}\tilde{\epsilon}_i(\boldsymbol{w})^\top),$$
$\tilde{\epsilon}_i(\boldsymbol{w}) \coloneqq \sum_{j=1}^J w_j\tilde{\epsilon}_i(j)$, $\hat{\boldsymbol{\Sigma}}_i \coloneqq (T_i - J - nJ - 1)^{-1}\hat{\epsilon}_i(J)^\top\hat{\epsilon}_i(J)$, $\tilde{\epsilon}_i(j)$ is a $(T_i - J)\times n$ matrix of the 'leave-subject-out residuals' of VAR(j), and $\hat{\epsilon}_i(J)$ is a $(T_i - J)\times n$ matrix of the LS residuals of VAR(J); see Liao et al. (2019, pp. 37-38);

- MMMA: $\hat{\boldsymbol{w}}_{i,MMMA} \coloneqq \arg\min_{\boldsymbol{w}\in\mathbb{W}} C_i(\boldsymbol{w})$, where
$$C_i(\boldsymbol{w}) \coloneqq \operatorname{tr}(\hat{\epsilon}_i(\boldsymbol{w})\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\epsilon}_i(\boldsymbol{w})^\top) + 2n(n\boldsymbol{J} + \boldsymbol{1})^\top\boldsymbol{w},$$
$\boldsymbol{J} \coloneqq (1, \ldots, J)^\top$, $\boldsymbol{1} \coloneqq (1, \ldots, 1)^\top$, and $\hat{\epsilon}_i(\boldsymbol{w}) = \sum_{j=1}^J w_j\hat{\epsilon}_i(j)$ is a $(T_i - J)\times n$ matrix of the weighted LS residuals; see Liao and Tsay (2020, pp. 1107-1108);

- SBIC: $\hat{\boldsymbol{w}}_{i,SBIC} \coloneqq (\hat{w}_{i1,SBIC}, \ldots, \hat{w}_{iJ,SBIC})^\top$, where
$$\hat{w}_{ij,SBIC} \coloneqq \frac{\exp(-\frac{1}{2}BIC_{ij})}{\sum_{l=1}^J \exp(-\frac{1}{2}BIC_{il})},$$
$BIC_{ij} \coloneqq \log\left\{\det\left(\hat{\boldsymbol{\Sigma}}_i(j)\right)\right\} + \log\{T_i\}(jn^2 + n)/T_i$ and $\hat{\boldsymbol{\Sigma}}_i(j) = (T_i - J)^{-1}\hat{\epsilon}_i(j)^\top\hat{\epsilon}_i(j)$ is the residual covariance matrix of VAR(j); see Liao et al. (2019, p. 42);

- SHQC: $\hat{\boldsymbol{w}}_{i,SHQC} \coloneqq (\hat{w}_{i1,SHQC}, \ldots, \hat{w}_{iJ,SHQC})^\top$, where
$$\hat{w}_{ij,SHQC} \coloneqq \frac{\exp(-\frac{1}{2}HQC_{ij})}{\sum_{l=1}^J \exp(-\frac{1}{2}HQC_{il})},$$
and $HQC_{ij} \coloneqq \log\left\{\det\left(\hat{\boldsymbol{\Sigma}}_i(j)\right)\right\} + 2(\log\{\log(T_i)\})(jn^2 + n)/T_i$;

- Equal: $\hat{\boldsymbol{w}}_{i,Equal} \coloneqq (J^{-1}, \ldots, J^{-1})^\top$.

Among these methods, 'Plug-In' and 'Fixed' are new MA methods generated by our approach for combining VAR forecasts, as defined in Sections 3.1 and 3.5. 'LsoMA' is the MA estimator of Liao et al. (2019) based on a leave-subject-out cross-validation criterion, and 'MMMA' is the MA estimator of Liao and Tsay (2020) based on a multivariate Mallows criterion. 'SBIC' and 'SHQC' are, respectively, the smoothed Bayesian information criterion (BIC) and the smoothed Hannan–Quinn criterion (HQC). We consider 'SBIC' because it is a popular approximation to the Bayesian MA based on the diffuse priors, and consider 'SHQC' because it is also commonly used for the choice of VAR. 'Equal' is the equal-weight averaging, which is widely used for forecast combination. Note that 'SBIC,' 'SHQC' and 'Equal' are rule-based MA methods, and 'Plug-In,' 'Fixed,' 'LsoMA' and 'MMMA' are optimized MA methods that are defined in different ways for VAR forecasts.

We set $n = 2$ and let $\hat{\theta}_{j,t}$ be the recursive LS estimator of VAR(j) for $j = 1, 2, 3$ and $J = 3$. Following the simulation of Lohmeyer et al. (2019), we consider the DGP:

$$f_{t+1} = \theta_1^* \begin{pmatrix} 0.5 & 0 \\ 0.5 & 0.5 \end{pmatrix} f_t + \theta_2^* \begin{pmatrix} 1 & 0 \\ 0.5 & 1 \end{pmatrix} f_{t-1} + \theta_3^* \begin{pmatrix} 1 & 0 \\ 0.5 & 1 \end{pmatrix} f_{t-2} + \epsilon_{t+1}^*, \tag{4}$$

where $\theta^* = (\theta_1^*, \theta_2^*, \theta_3^*)^\top$ is a local-parameter vector with $\theta_\ell^* = cT^{-1/2}$ for $\ell = 1, 2, 3$ and for some constant $c > 0$, and $\{\epsilon_t^*\}$ is a sequence of independent and normally distributed random variables with mean zero and variance-covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.17 \\ 0.17 & 0.33 \end{pmatrix}.$$

Under this DGP, the unrestricted model VAR(3) is correctly specified whereas the submodels (i.e., VAR(1) and VAR(2)) are locally misspecified. The specification biases of the submodels and their differences increase with c. We set $c = 0.5$, 1, 1.5, or 2.

To measure the forecasting performance of $\{\hat{f}_{t+1|t}(\boldsymbol{w})\}_{t=T_i}^{T_i+P_i-1}$ generated by an MA method, we define the empirical MSFE (EMSFE):

$$EMSFE_i(\boldsymbol{w}) \coloneqq \frac{1}{P_i} \sum_{t=T_i}^{T_i+P_i-1} \left(f_{t+1} - \hat{f}_{t+1|t}(\boldsymbol{w})\right)^\top V^{-1} \left(f_{t+1} - \hat{f}_{t+1|t}(\boldsymbol{w})\right) \tag{5}$$

where $P_i \coloneqq [\lambda T_i]$ is the nearest integer of λT_i for $i = 1, \ldots, R$, and V is an $n \times n$ weighting matrix. Note that $EMSFE_i(\boldsymbol{w})$ is unbiased for the averaged Bregman risk of $\hat{f}_{t+1|t}(\boldsymbol{w})$ in the mean-forecast context when $\boldsymbol{w}$ is fixed. We computed (5) by setting $\boldsymbol{w} = \hat{\boldsymbol{w}}_{i,\hat{\lambda}}$, $\hat{\boldsymbol{w}}_{i,0}$, $\hat{\boldsymbol{w}}_{i,LsoMA}$, $\hat{\boldsymbol{w}}_{i,MMMA}$, $\hat{\boldsymbol{w}}_{i,SBIC}$, $\hat{\boldsymbol{w}}_{i,SHQC}$ and $\hat{\boldsymbol{w}}_{i,Equal}$ for 'Plug-In,' 'Fixed,' 'LsoMA,' 'MMMA,' 'SBIC,' 'SHQC' and 'Equal,' respectively, and define the simulated average of (5):

$$\widehat{EMSFE}_i \coloneqq \frac{1}{B} \sum_{b=1}^{B} EMSFE_{i,b}(\boldsymbol{w}_b), \tag{6}$$

where $\boldsymbol{w}_b$ and $EMSFE_{i,b}(\boldsymbol{w}_b)$ are, respectively, the counterparts of $\boldsymbol{w}$ and $EMSFE_i(\boldsymbol{w})$ based on the bth replication of the simulation for $b = 1, 2, \ldots, B$.

In this simulation, we set $T = 50$ or 100, $\lambda = T_i^{-1}$ ($P_i = 1$) for combining single forecasts and $\lambda = 1$, 2 or 4 for combining forecast sequences, the maximum R/T ratio 0.8, and the number of simulation replications $B = 1000$. Given the VAR-DGP in (4) and the setting of $V = I_n$, we show the performance measure $\{\widehat{EMSFE}_i\}_{i=1}^R$ in Figure 1 for $T = 50$ and in Figure 2 for $T = 100$. In each cell of these figures, the horizonal axis is the R/T ratio, and the performance measures of different MA methods are presented in different patterns and colors. The colored figures are available in the electronic version of this paper. The main simulation findings are summarized as follows.

First, the sequence $\{\widehat{EMSFE}_i\}_{i=1}^R$ is quite volatile when $\lambda = 1/T_i$, but becomes smoother when $\lambda = 1$, 2 or 4. This reflects the fact that $EMSFE_i(\boldsymbol{w})$ reduces to a squared forecast error when $\lambda = 1/T_i$ (that is, when $P_i = 1$), but is a sample average of the squared forecast errors when $\lambda = 1$, 2 or 4; see also Chen and Liu (2023) for this evidence in the univariate case. Thus, we focus on the case where $\lambda \geq 1$ in the following discussions.

Second, the rule-based methods have quite similar forecasting performance. Despite their predictive superiority over 'Plug-In' when $c = 0.5$, 'Plug-In' catches up these methods in this case provided that the R/T ratio is sufficiently large. By contrast, 'Plug-In' generally outperforms the rule-based methods when $c = 1$, 1.5 or 2. Moreover, the relative advantage of 'Plug-In' over these methods obviously increases with c and tends to increase with λ or the R/T ratio. This shows that although 'Plug-In' may be outperformed by the rule-based methods in finite samples when the candidate models are relatively close to each other (as in the case $c = 0.5$), it outperforms the latter when the candidate models are distinguishable (as in the case where $c \geq 1$). This result holds for both $T = 50$ and $T = 100$.

Third, although 'Plug-In' is outperformed by the other optimized methods being considered when $c = 0.5$, their difference diminishes in this case when the R/T ratio increases. In comparison, 'Plug-In' generally outperforms these methods when $c = 1$, 1.5 or 2, and the difference tends to increase with c or λ. In particular, the relative advantage of 'Plug-In' over 'Fixed' increases with λ. This finding is consistent with the theoretical results that the difference between 'Plug-In' and 'Fixed' increases with λ and that the asymptotic optimality holds for 'Plug-In,' rather than 'Fixed,' in this simulation experiment. This result also holds for both $T = 50$ and $T = 100$.

In this experiment, we also considered alternative simulation settings, including replacing the setting of V by $V = \hat{\boldsymbol{\Sigma}}_i$, replacing the performance measure by the following statistic:

$$\widehat{AMSFE}_i := \frac{1}{B}\sum_{b=1}^{B} \widetilde{AMSFE}_{i,b}(\boldsymbol{w}_b), \tag{7}$$

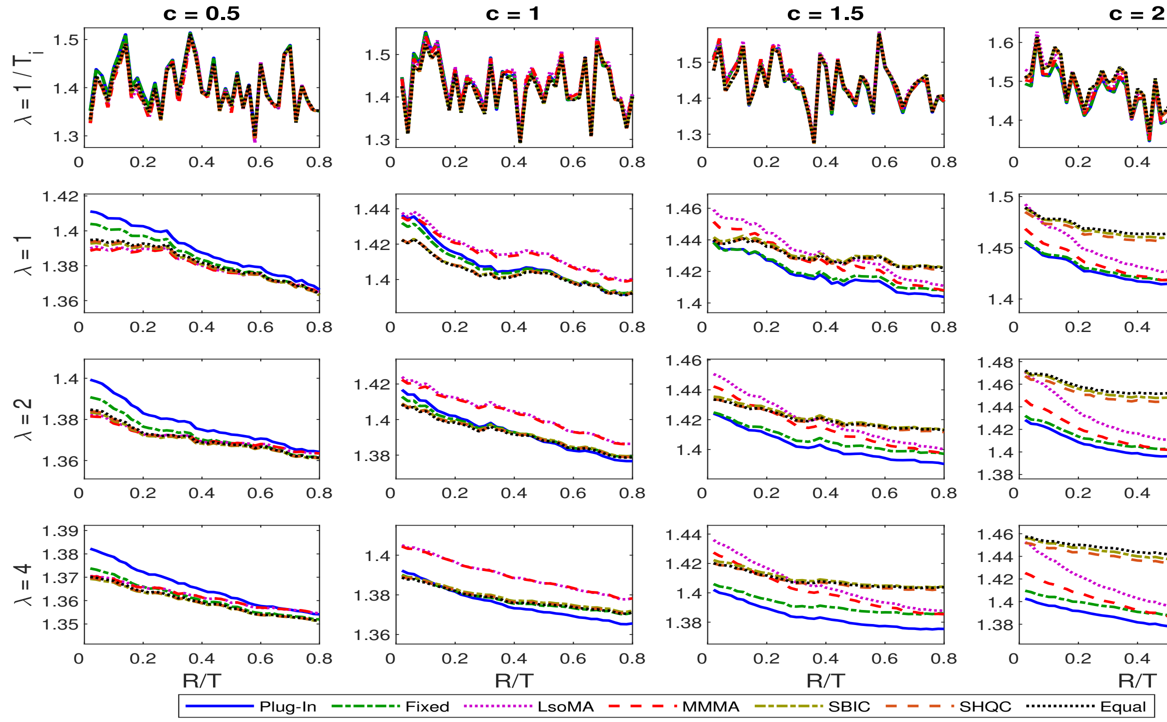

Figure 1: $\{\widehat{EMSFE_i}\}_{i=1}^R$ for $T = 50$ and $V = I_n$ in the setting of VAR(3).

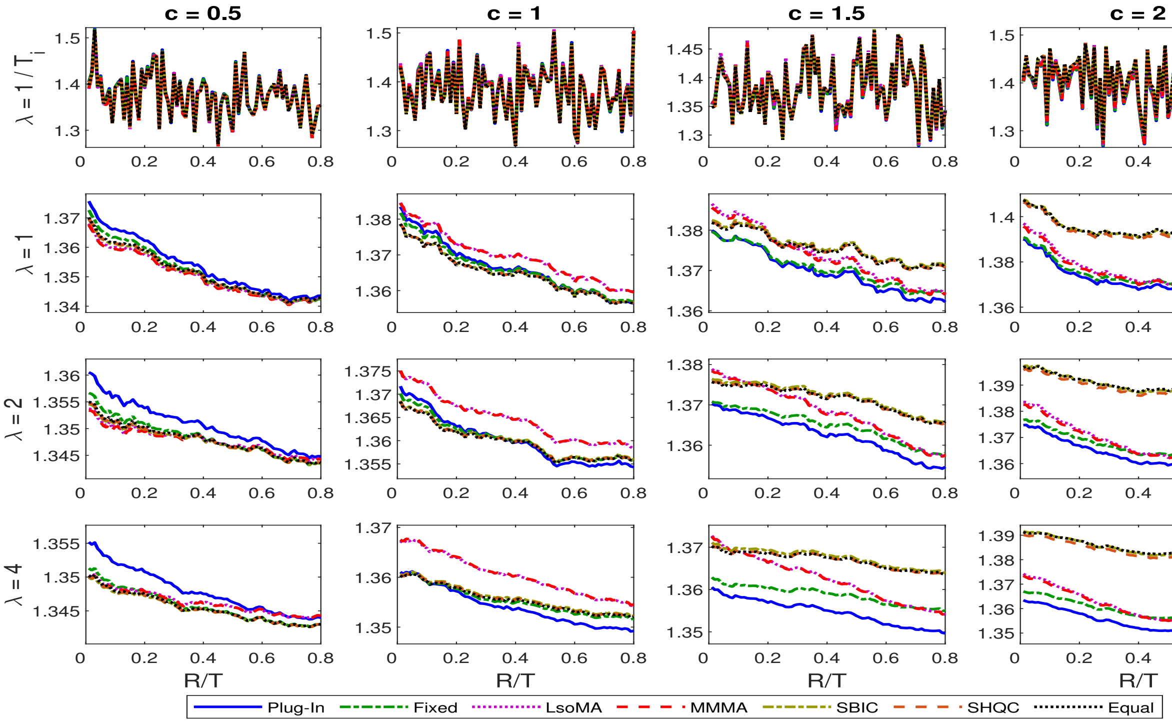

Figure 2: $\{\widehat{EMSFE_i}\}_{i=1}^R$ for $T = 100$ and $V = I_n$ in the setting of VAR(3).

where $\widetilde{AMSFE}_{i,b}(\boldsymbol{w}_b)$ is the counterpart of the statistic:

$$\widetilde{AMSFE}_i(\boldsymbol{w}) := \frac{T_i}{P_i} \sum_{t=T_i}^{T_i+P_i-1} \left(\hat{f}_{t+1|t}(\boldsymbol{w}) - f^*_{t+1|t}\right)^\top V^{-1} \left(\hat{f}_{t+1|t}(\boldsymbol{w}) - f^*_{t+1|t}\right) \tag{8}$$

based on the bth simulation replication, as well as replacing the DGP by a VAR moving average (VARMA) process used in the simulation of Liao and Tsay (2020). Note that $\widetilde{AMSFE}_i(\boldsymbol{w})$ is asymptotically unbiased for the normalized asymptotic MSFE (AMSFE), which is defined by the $\mathcal{R}(\boldsymbol{w})$ in the mean-forecast context, when $\boldsymbol{w}$ is fixed. In addition, all candidate models, including VAR(J), are misspecified under the VARMA process. For the sake of brevity, we present the counterparts of Figures 1 and 2 under these alternative settings in the online appendix. The additional results show that the relative forecasting performance of these methods is generally robust to the replacement of V and the performance measure. For the DGP in which all candidate models are misspecified, we observe that 'Plug-In' is outperformed by other methods when $T = 50$ and the R/T ratio is small, but 'Plug-In' catches up with these methods when the R/T ratio increases and generally outperforms these methods when $T = 100$ and c is large.

4.2 Density Forecast Combination

In this experiment, we consider a set of MA methods for density forecast combination that generate the following $\boldsymbol{w}$'s at $t = T_i$:

- Plug-In: $\hat{\boldsymbol{w}}_{i,\hat{\lambda}}$;

- Fixed: $\hat{\boldsymbol{w}}_{i,0}$;

- HM2007: $\hat{\boldsymbol{w}}_{i,HM} := \arg\min_{\boldsymbol{w}\in\mathbb{W}} KL_i(\boldsymbol{w})$, where

  $$KL_i(\boldsymbol{w}) := \frac{1}{T_i}\sum_{t=1}^{T_i} \log\left\{\sum_{j=1}^{J} w_j g_{j,t}(y_t;\hat{\theta}_{j,T_i})\right\}$$

  for $j = 1, \ldots, J$;

- PV2016: $\hat{\boldsymbol{w}}_{i,PV} := (\hat{w}_{i1,PV}, \ldots, \hat{w}_{iJ,PV})^\top$, where

  $$\hat{w}_{ij,PV} := \frac{1/|S_{ij}|}{\sum_{l=1}^{J} 1/|S_{il}|}$$

  and $S_{ij} := T_i^{-1}\sum_{t=1}^{T_i} \log\left\{g_{j,t}(y_t;\hat{\theta}_{j,T_i})\right\}$ for $j = 1, \ldots, J$;

- SBIC: $\hat{\boldsymbol{w}}_{i,SBIC} \coloneqq (\hat{w}_{i1,SBIC}, \ldots, \hat{w}_{iJ,SBIC})^\top$, where

  $$\hat{w}_{ij,SBIC} \coloneqq \frac{\exp(-\frac{1}{2}BIC_{ij})}{\sum_{l=1}^{J} \exp(-\frac{1}{2}BIC_{il})}$$

  $BIC_{ij} \coloneqq 2\mathcal{L}(\hat{\theta}_{j,T_i}) + k_j \log\{T_i\}$, $\mathcal{L}(\hat{\theta}_{j,T_i})$ is the negative log-likelihood function of the jth model, and k_j is the number of parameters of the jth model, for $j = 1, \ldots, J$;

- Equal: $\hat{\boldsymbol{w}}_{i,Equal} = (J^{-1}, \ldots, J^{-1})^\top$.

Among these methods, 'Plug-In' and 'Fixed' are new MA methods generated by our approach for combining density forecasts, as defined in Sections 3.4 and 3.5. 'HM2007' is the density-forecast combination method proposed by Hall and Mitchell (2007), which is based on the average log score of the combined density forecast; see also Geweke and Amisano (2011) for theoretical and numerical properties of this method. We solve $\hat{\boldsymbol{w}}_{i,HM}$ using the numerical optimization algorithm '`fmincon`' of MATLAB. 'PV2016' is the density-forecast combination method proposed by Pauwels and Vasnev (2016) based on a smoothed log predictive scores of the candidate models. 'SBIC' is the smoothed BIC defined by replacing the VAR models with the candidate models for density forecast, and 'Equal' is the equal-weight averaging as defined previously. Note that 'PV2016,' 'SBIC' and 'Equal' are rule-based MA methods, and 'Plug-In,' 'Fixed' and 'HM2007' are different optimized MA methods for density forecast.

We consider the univariate case where $n = 1$, and let $\hat{\theta}_{j,t}$ be the recursive QML estimator based on the 'normal model' (for $j = 1$), 'the t model' (for $j = 2$) or 'the skewed-t model' (for $j = 3$) with $J = 3$. These models share the same GARCH(1,1) specification: for all t's,

$$\begin{aligned} y_t &= \epsilon_t \sigma_t, \\ \sigma_t^2 &= \alpha + \beta \sigma_{t-1}^2 + \gamma y_{t-1}^2, \end{aligned}$$

where $\{\epsilon_t\}$ is an IID sequence of standardized errors, but have different distributional specifications of ϵ_t. The distribution of ϵ_t is specified as the standard normal for the normal model, the standardized t distribution with the degrees of freedom $\eta > 2$ for the t model, or the skewed-t distribution of Hansen (1994) with the tail parameter $\eta > 2$ and the asymmetry parameter $-1 < \xi < 1$ for the skewed-t model. The skewed-t model includes the t model as a special case where $\xi = 0$, and the t model includes the normal model as a limiting case where $\eta \to \infty$. We set the DGP to be the skewed-t model with the GARCH parameters $(\alpha, \beta, \gamma) = (0.05, 0.9, 0.05)$ and one of the two distributional parameters

- Case 1: $\eta = \eta_0 + c_1 T^{-1/2}$ and $\xi = 0$,

- Case 2: $\eta = \eta_0$ and $\xi = c_2 T^{-1/2}$, where $\eta_0 = 30$, $c_1 = -45$, -95, -145, or -195, and $c_2 = -1$, -3, -5, or -7. Thus, the skewed-t model is correctly specified in both cases, the t model is correctly specified in Case 1 but locally misspecified in Case 2, and the normal model is misspecified in both cases. In Case 1, the skewed-t distribution is symmetric and obviously heavy-tailed when $T = 50$ and $c_1 = -195$, but is close to the standard normal in other settings. In Case 2, the skewed-t distribution is left-skewed, but gets closer to the standard normal when $|c_1|$ decreases; see Figure C.1 of the online appendix for these distributions.

To measure the forecasting performance of $\left\{\sum_{j=1}^{J} w_j g_{j,t}(\cdot\,;\hat{\theta}_{j,t})\right\}_{t=T_i}^{T_i+P_i-1}$, we denote the empirical log predictive score function:

$$Score_i(\boldsymbol{w}) := \frac{1}{P_i}\sum_{t=T_i}^{T_i+P_i-1} \log\left\{\sum_{j=1}^{J} w_j g_{j,t}(\cdot\,;\hat{\theta}_{j,t})\right\}, \tag{9}$$

for $i = 1, \ldots, R$. Note that $Score_i(\boldsymbol{w})$ is unbiased for the model component of the KL divergence when $\boldsymbol{w}$ is fixed. We compute (9) by setting $\boldsymbol{w} = \hat{\boldsymbol{w}}_{i,\hat{\lambda}}$, $\hat{\boldsymbol{w}}_{i,0}$, $\hat{\boldsymbol{w}}_{i,HM}$, $\hat{\boldsymbol{w}}_{i,PV}$, $\hat{\boldsymbol{w}}_{i,SBIC}$ and $\hat{\boldsymbol{w}}_{i,Equal}$ for 'Plug-In,' 'Fixed,' 'HM2007,' 'PV2016,' 'SBIC' and 'Equal,' respectively, and define the simulated average of (9):

$$\widehat{Score}_i := \frac{1}{B}\sum_{b=1}^{B} Score_{i,b}(\boldsymbol{w}_b), \tag{10}$$

where $\boldsymbol{w}_b$ and $Score_{i,b}(\boldsymbol{w}_b)$ are, respectively, the counterparts of $\boldsymbol{w}$ and $Score_i(\boldsymbol{w})$ based on the bth replication of the simulation for $b = 1, 2, \ldots, B$. Since the empirical log predictive scores are generally negative, a larger $\widehat{Score}_i$ means a better density forecast performance.

Corresponding to the previous experiment, we conduct a small-T setting: $T = 50$ or 100. Since each replication of this experiment involves the numerical optimization for computing the QML estimators at each $t \in [T_i, T_i + P_i - 1]$ for $i = 1, 2, \ldots, R$, the computational cost is rather high if we set T to be large. This small-T setting is thus useful for reducing the computational cost of this experiment. In addition, we set $\lambda = T_i^{-1}$ ($P_i = 1$) for combining single forecasts and $\lambda = 1$, 2 or 4 for combining forecast sequences, the maximum R/T ratio 0.8, and the number of simulation replications $B = 1000$ in this experiment. We show the performance measure $\{\widehat{Score}_i\}_{i=1}^{R}$ in Figure 3 (Figure 4) for Case 1 and $T = 50$ ($T = 100$) and Figure 5 (Figure 6) for Case 2 and $T = 50$ ($T = 100$). The main simulation findings are summarized as follows.

First, among the rule-based MA methods, 'PV2016' and 'Equal' are essentially indistinguishable in terms of their forecasting performance. These two methods perform quite well in Case 1 with $c_1 = -45$, -95 or -145 and Case 2 with $c_2 = -1$. In these cases, the candidate models are quite close to each other. However, they are obviously outperformed by 'SBIC'

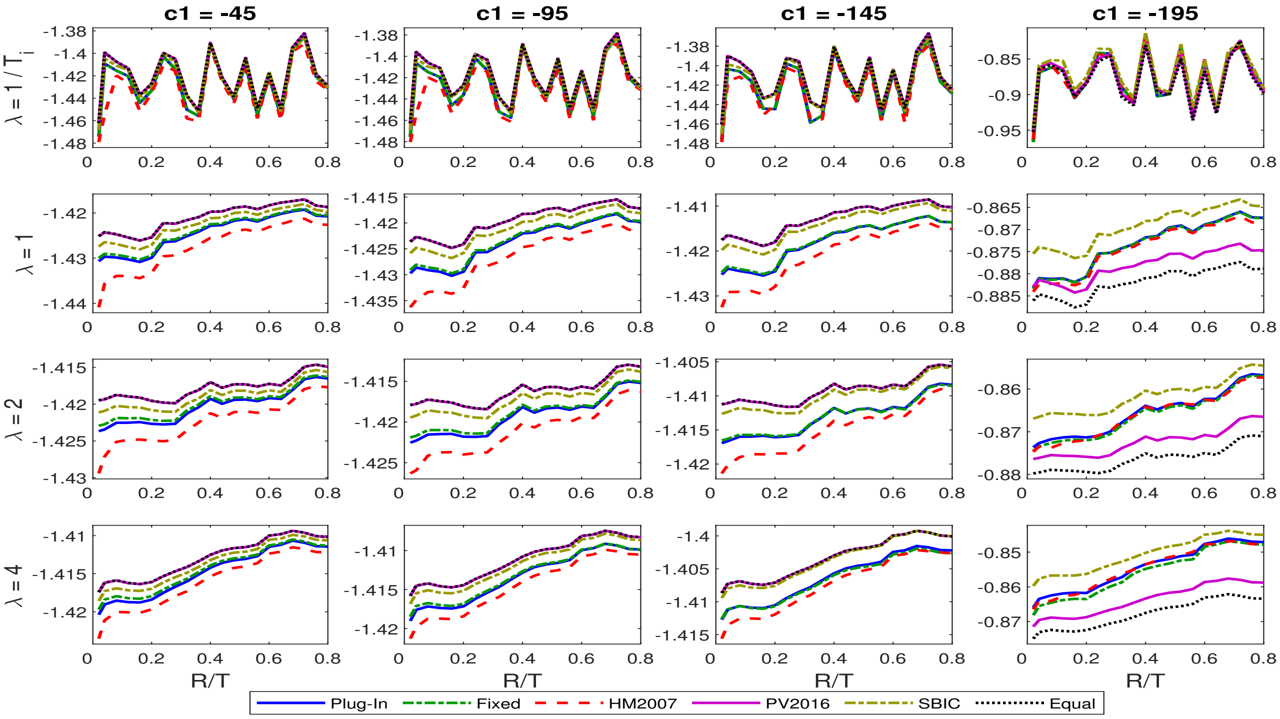

Figure 3: $\{\widehat{Score_i}\}_{i=1}^{R}$ for $T = 50$ in Case 1.

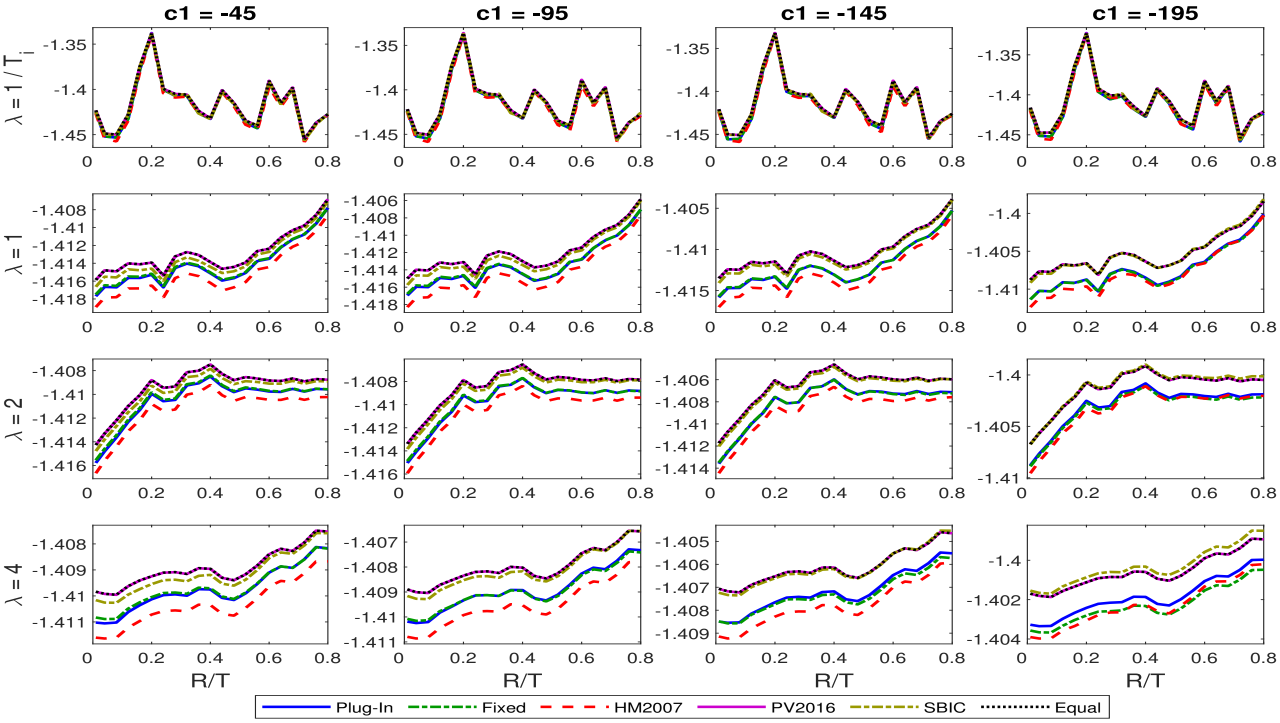

Figure 4: $\{\widehat{Score_i}\}_{i=1}^{R}$ for $T = 100$ in Case 1.

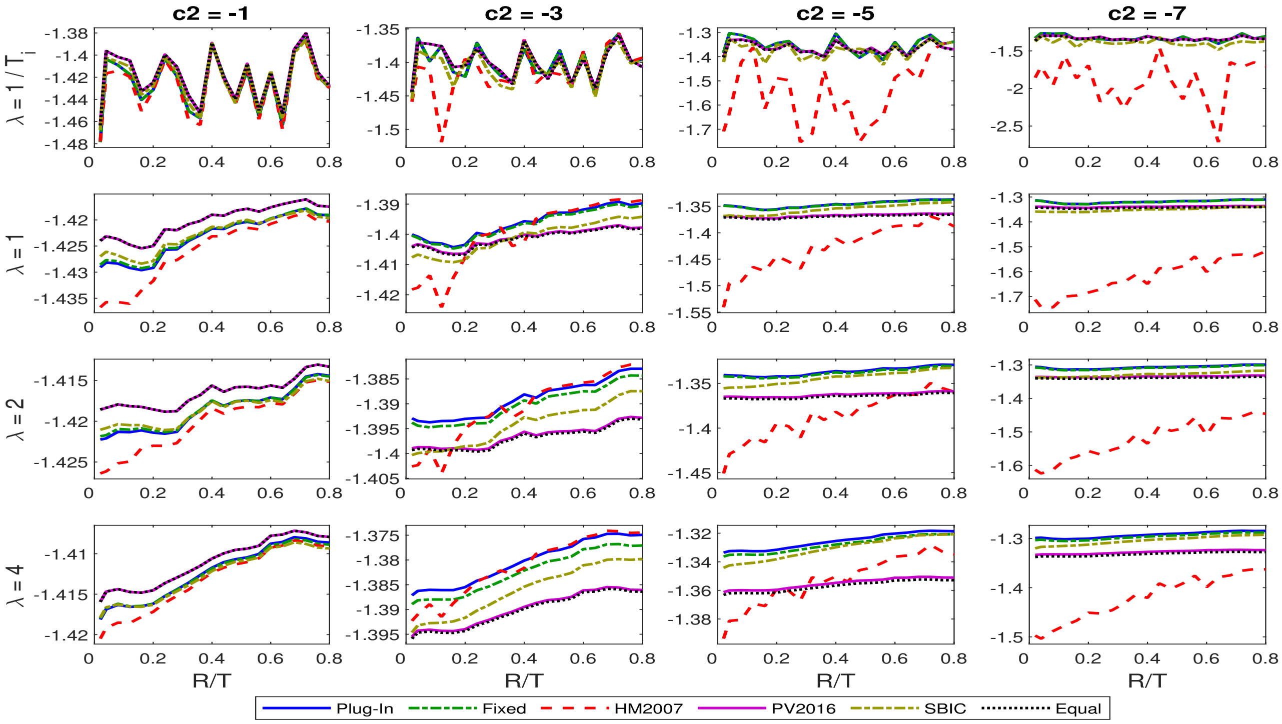

Figure 5: $\{\widehat{Score_i}\}_{i=1}^{R}$ for $T = 50$ in Case 2.

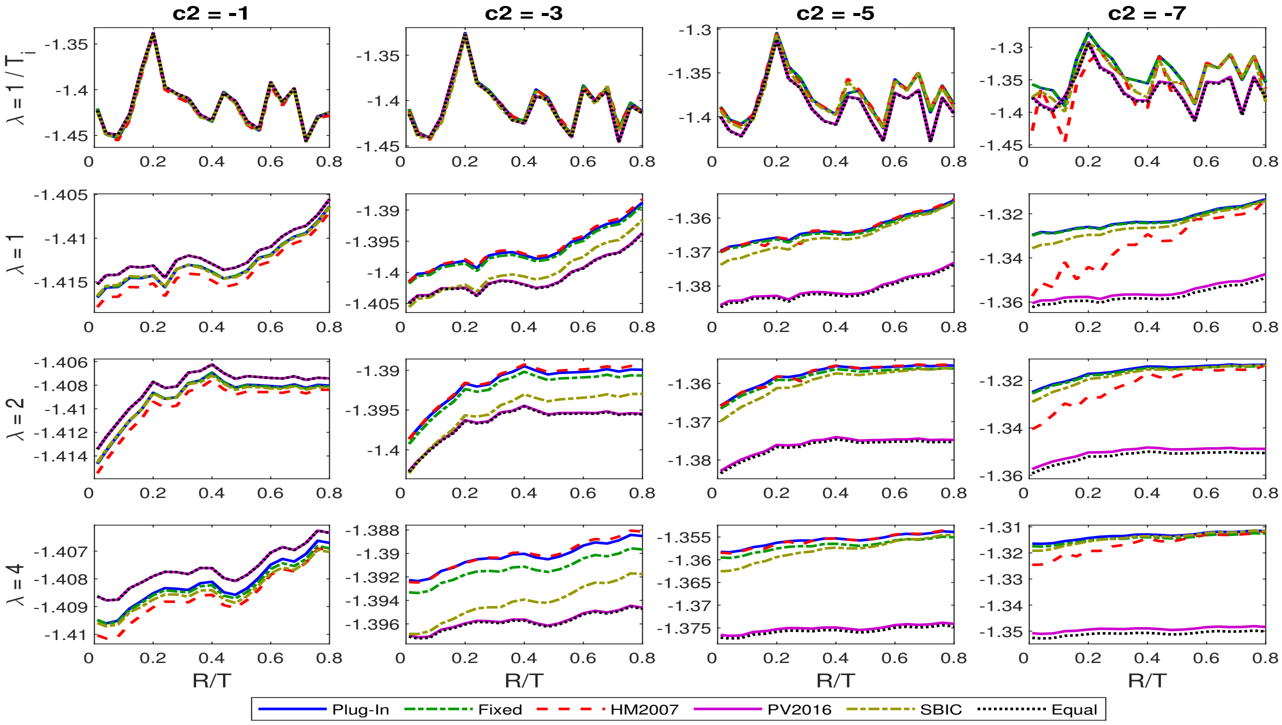

Figure 6: $\{\widehat{Score_i}\}_{i=1}^{R}$ for $T = 100$ in Case 2.

and other MA methods in Case 1 with $c_1 = -195$ (Case 2 with $c_2 = -3$, -5 and -7) where the normal model (the t model) is obviously misspecified. Thus, similar to what we have observed in the previous experiment, 'Equal' performs well when it is difficult to distinguish between the candidate models; however, this 'forecast combination puzzle' disappears when the differences between the candidate models become relevant.

Second, among the optimized MA methods, 'Plug-In' and 'Fixed' have quite similar forecasting performance in most settings. However, 'Plug-In' tends to outperform 'Fixed' in Case 2 with $c_2 = -3$ especially when λ increases to 2 or 4. In addition, 'Plug-In' substantially outperforms 'HM2007' in Case 1 when $c_1 = -45$, -95 or -145 for both T's and in Case 2 when $T = 50$ for all c_2's as well as when $T = 100$ and $c_2 = -1$ or -7. Nonetheless, 'HM2007' tends to catch up 'Plug-In' in these cases when the R/T ratio increases. Moreover, they have similar forecasting performances in Case 1 when $c_1 = -195$ and in Case 2 when $T = 100$ and $c_2 = -3$ or -5. This result suggests that 'Plug-In' tends to outperform 'HM2007' in the case where the estimation uncertainty is more relevant and that the specification bias is less relevant, while they tend to have similar performance when the specification bias become more relevant. This finding is consistent with the fact that, compared to 'HM2007,' 'Plug-In' accounts for not only the specification bias of candidate models but also the uncertainty of parameter uncertainty in combining density forecasts.

Third, 'Plug-In' is generally outperformed by the rule-based MA methods in Case 1 and Case 2 with $c_2 = -1$, but generally outperforms the latter in Case 2 with the other c_2's. In particular, 'Plug-In' substantially outperforms 'PV2016' and 'Equal' in Case 2 when $c_2 \geq 1$ and obviously outperforms 'SBIC' in this case when $c_2 = -3$ or -5. Thus, similar to what we have observed in the previous experiment, 'Plug-In' compares favorably to the rule-based methods when the differences between the candidate models are relevant.

In this experiment, we also considered replacing the parameter setting of DGP with

- Case 3: $\eta = \eta_0 + c_1 T^{-1/2}$ and $\xi = c_2 T^{-1/2}$,

and replacing the setting of $J = 3$ with the setting of $J = 2$ that does not include the normal model as a candidate model. We present the counterparts of Figures 3-6 generated by these alternative simulation settings in the online appendix. The additional results show that the finite-sample forecasting performance of these methods in Case 3 is similar to those in Case 2. This is because the skewed-t distribution in Case 3 is quite similar to that in Case 2, as shown in Figure C.1 of the online appendix. We also observe that the relative forecasting performance of the MA methods is robust to this change of J.

Generally speaking, this simulation shows that our method has proper finite-sample performance in comparison with related existing methods, and the simulation findings are consistent with the theoretical results.

5 Empirical Illustration

In this section, we further compare 'Plug-In' with alternative MA methods in their forecasting performance using real data. Corresponding to the simulation, we consider an empirical example of combining VAR forecasts in Section 5.1 and an empirical example of combining density forecasts in Section 5.2.

5.1 VAR Forecast Combination

In this empirical example, motivated by Stock and Watson (2001), we consider a problem of VAR forecast for macroeconomic variables. We set $n = 3$ and $f_t = (\pi_t, u_t, R_t)'$, where π_t, u_t and R_t are, respectively, the inflation rate, the unemployment rate and the federal funds rate of the U.S., rescaled by multiplying 100, in the tth month of the full sampling period from July 1954 to December 2019. This sampling period includes 786 months.[12] We also consider the case where $n = 2$ and $f_t = (\pi_t, u_t)'$. Following the simulation, we set the candidate models to be the VAR(j)'s that include intercepts for $j = 1, 2, 3$ and $J = 3$, and implement the VAR forecasts using the recursive LS method. We also set $T = 100$, 150 or 200, $P_i = [\lambda T_i]$ with $\lambda = 1/T_i$, 0.5, 1, 1.5 or 2, $i = 1, \ldots, R$, and $R = 1$, 100, 150 or 200. The largest R and λ vary with T because of the limitation of the full-sample length. In addition, we set $\boldsymbol{w} = \hat{\boldsymbol{w}}_{i,0}$, $\hat{\boldsymbol{w}}_{i,LsoMA}$, $\hat{\boldsymbol{w}}_{i,MMMA}$, $\hat{\boldsymbol{w}}_{i,SBIC}$, $\hat{\boldsymbol{w}}_{i,SHQC}$ or $\hat{\boldsymbol{w}}_{i,Equal}$ at $t = T_i$, and evaluate the forecasting performance of an MA method using the empirical counterpart of (6):

$$\widehat{EMSFE}_i^* := \frac{1}{B_i} \sum_{b=1}^{B_i} EMSFE_{i,b}^*(\boldsymbol{w}_b), \tag{11}$$

where $\boldsymbol{w}_b$ is the counterpart of $\boldsymbol{w}$ at $t = T_i + b - 1$,

$$EMSFE_{i,b}^*(\boldsymbol{w}_b) := \frac{1}{P_i} \sum_{t=T_i}^{T_i+P_i-1} \left(f_{t+b} - \hat{f}_{t+b|t+b-1}(\boldsymbol{w}_b) \right)^\top V^{-1} \left(f_{t+b} - \hat{f}_{t+b|t+b-1}(\boldsymbol{w}_b) \right),$$

$B_i := \bar{T} - (T_i + P_i) + 1$, and $\bar{T}$ denotes the full-sample size. After computing the time series of π_t, we have $\bar{T} = 774$. Correspondingly, we compare 'Plug-In' with an alternative MA method in terms of their forecasting performance using the relative EMSFE (RMSFE):

$$RMSFE := \frac{\widehat{EMSFE}_R^* \text{ of 'Plug-In'}}{\widehat{EMSFE}_R^*}. \tag{12}$$

[12] The inflation rate is defined as $\pi_t = 100 \times ((P_t - P_{t-12})/P_{t-12})$, where P_t is the monthly consumer price index. The time series data used in this empirical example is obtained from the Federal Reserve Economic Data (FRED). We download the price index from `https://fred.stlouisfed.org/series/CPIAUCSL`, the unemployment rate from `https://fred.stlouisfed.org/series/UNRATE` and the federal funds rate from `https://fred.stlouisfed.org/series/FEDFUNDS`. See Figure C.2 of the online appendix for the time series of these three variables.

Given $V = I_n$, we report the RMSFEs in Table 1. The main empirical findings are summarized as follows.

First, the rule-based MA methods: 'SBIC,' 'SHQC' and 'Equal' have quite similar forecasting performance in comparison with 'Plug-In.' In the case where $n = 3$, these methods generally outperform 'Plug-In' with the RMSFEs greater than one when $R = 1$ for all (T, λ)'s. However, this evidence becomes less relevant when R increases to 100, and is reversed when $R = 200$ for $T = 100$ ($R = 150$ for $T = 150$ or 200). In the case where $n = 2$, the rule-based methods are all dominated by 'Plug-In' with the RMSFEs smaller than one for all (T, R, λ)'s. These results show that although we are unable to preclude the forecast combination puzzle in this empirical example, 'Plug-In' compares favorably to the rule-based methods when R is sufficiently large or when $n = 2$. Similarly, although 'Plug-In' tends to be slightly outperformed by 'Fixed' when $n = 3$, it generally outperforms 'Fixed' when $n = 2$. (Recall that 'Plug-In' reduces to 'Fixed' when $\lambda = 1/T_i$.) These results commonly suggest that the dimension n plays an important role in determining the relative forecasting performance of 'Plug-In' in finite samples. The importance of n might be attributed to the fact that the estimation quality of VAR is dependent on the dimension n for a fixed T. In particular, the number of unknown parameters of VAR(J) substantially increases from 14 when $n = 2$ to 30 when $n = 3$ in this empirical example.

Second, the relative forecasting performance of 'Plug-In' in comparison with 'LsoMA' or 'MMMA' is mixed when $n = 2$. Given $n = 3$, 'Plug-In' outperforms 'LsoMA' except for the case where $T = 150$, $R = 1$ and $\lambda = 0.5$, and outperforms 'MMMA' except for certain settings especially when $R = 1$. It is difficult to interpret the differences among the empirical performance of these methods in a complete way because these methods are established in different theoretical contexts; moreover, the sample size is limited, and the DGP is unknown. Nonetheless, 'Plug-In' is comparable to these two methods in terms of forecasting performance in this empirical example.

We also considered replacing the setting of $V = I_n$ with $V = \hat{\mathbf{\Sigma}}_i$, and present the counterparts of Table 1 generated by this replacement in the online appendix. We observe that the relative forecasting performance of 'Plug-In' over other methods slightly improves when $V = I_n$ is replaced by $V = \hat{\mathbf{\Sigma}}_i$.

5.2 Density Forecast Combination

In this empirical example, following Bao et al. (2007) and Opschoor et al. (2017), among many others, we consider a problem of density forecast for the daily returns of the S&P 500 index. The full sampling period is from 2nd January 1998 to 31th December 2019, and includes the number of observations 5534.[13] Corresponding to the simulation, we set the candidate models to be the t model and the skewed-t model in the case where $J = 2$ as well as the normal model in the case where $J = 3$, and implement the density forecasts using the recursive QML method. However, following the related literature, we consider a large-T setting: $T = 1250$, 1500 or 1750 in this empirical example. In addition, we set $P_i = [\lambda T_i]$ with $\lambda = 1/T_i$, 0.5, 1, 1.5 or 2, $i = 1, \ldots, R$, and $R = 1$, 250, 500 or 750. The largest R and λ also vary with T because the full-sample size is fixed. We set $\boldsymbol{w} = \hat{\boldsymbol{w}}_{i,0}$, $\hat{\boldsymbol{w}}_{i,HM}$, $\hat{\boldsymbol{w}}_{i,PV}$, $\hat{\boldsymbol{w}}_{i,SBIC}$, or $\hat{\boldsymbol{w}}_{i,Equal}$ at $t = T_i$, and evaluate the forecasting performance of an MA method using the empirical counterpart of (10):

$$\widehat{Score}_i^* := \frac{1}{B_i} \sum_{b=1}^{B_i} Score_{i,b}^*(\boldsymbol{w}_b), \tag{13}$$

where $\boldsymbol{w}_b$ is the counterpart of $\boldsymbol{w}$ at $t = T_i + b - 1$,

$$Score_{i,b}^*(\boldsymbol{w}_b) := \frac{1}{P_i} \sum_{t=T_i}^{T_i+P_i-1} \log \left\{ \sum_{j=1}^{J} w_{b,j} g_{j,t}(\cdot\,; \hat{\theta}_{j,t+b-1}) \right\},$$

$w_{b,j}$ is the jth element of $\boldsymbol{w}_b$, $B_i := \bar{T} - (T_i + P_i) + 1$, and $\bar{T}$ denotes the full-sample size. In this example, $\bar{T} = 5533$. We compare 'Plug-In' with an alternative MA method in terms of their forecasting performance using the relative score (rScore):

$$rScore := \frac{\widehat{Score}_R^* \text{ of 'Plug-In'}}{\widehat{Score}_R^*}. \tag{14}$$

Since the log score is generally negative, 'Plug-In' outperforms (is outperformed by) the alternative method if the rScore is less (greater) than one. We report the rScores in Table 2. The main findings are summarized as follows.

First, 'Plug-In' outperforms the rule-based MA methods: 'PV2016,' 'SBIC' and 'Equal' for all (T, R, λ)'s and for both $J = 2$ and $J = 3$. In addition, 'Plug-In' tends to outperform 'Fixed' provided that λ is sufficiently large. This finding also holds for all (T, R)'s and for both $J = 2$ and $J = 3$.

Second, 'Plug-In' is outperformed by 'HM2007' for all the settings of (T, R, λ) in the case where $J = 2$. However, the relative forecasting performance of 'Plug-In' improves when T increases in the case where $J = 3$. In particular, 'Plug-In' tends to outperform 'HM2007' when R and λ are sufficiently large in the case where $J = 3$. Although the relative forecasting

[13] Let P_t be the close price of the S&P500 index of the tth date in the sampling period. We define the daily return of the tth date as 100 × $(\log\{P_t\} - \log\{P_{t-1}\})$, and download the stock index data from the website of Wall Street Journal: https://www.wsj.com/market-data/quotes/index/SPX/historical-prices. See Figure C.3 of the online appendix for this return sequence.

performance of these two methods in this empirical large-T setting is generally different from its counterpart in the simulated small-T setting, the simulation also shows that these two methods could have similar forecasting performance in certain cases when T is larger.

6 Conclusion

In the recent literature, researchers have proposed different frequentist MA methods for forecast combination. Most of these methods are established in the conventional context where the forecasting target is a univariate random variable, the candidate models are linear predictive regressions, the parameters are estimated by the LS method, the loss function of forecast error is quadratic, and the asymptotic risk is built on the MSFE. Because the existing frequentist MA methods are context-specific, the asymptotic optimality of these methods may not hold in other forecasting contexts. However, from the empirical viewpoint, it is a general rule rather than an exception that economic forecasts are likely to be beyond the conventional context. Different types of economic forecast may involve different forecasting targets, models, estimation methods *or* loss functions of forecast error. It is undoubtedly essential to have a unified MA approach that is flexibly applicable to establishing asymptotically optimal forecast combinations in different forecasting contexts. We contribute to the literature by proposing such a unified approach. Our approach is established in a generalized context where the forecasting target, the candidate models and the estimation method are not pre-specified in theoretical analysis, and the asymptotic risk is built on the BD, which serves as a generalized loss function of forecast errors. We establish the asymptotic optimality of our approach under the local-asymptotics setting, and apply this approach to generate new MA methods for mean forecast combination, volatility forecast combination, probabilistic forecast combination and density forecast combination. Monte Carlo simulations and empirical applications show that the proposed MA methods perform reasonably well in comparison with related existing methods.

Table 1: Forecasting performance of ‘Plug-In’ relative to alternative MA methods for the VAR forecast: $V = I_n$.

			$n = 2$						$n = 3$					
T	R	λ	Fixed	LsoMA	MMMA	SBIC	SHQC	Equal	Fixed	LsoMA	MMMA	SBIC	SHQC	Equal
100	1	$1/T_i$	1.000	0.988	0.995	0.992	0.993	0.992	1.000	0.962	1.008	1.018	1.019	1.017
100	1	0.5	1.000	0.957	0.990	0.994	0.995	0.995	1.002	0.993	1.008	1.017	1.018	1.016
100	1	1.0	0.997	0.938	0.983	0.992	0.993	0.993	1.003	0.987	1.006	1.011	1.013	1.011
100	1	1.5	0.994	0.923	0.977	0.987	0.988	0.989	1.002	0.981	1.003	1.008	1.010	1.010
100	1	2.0	0.993	0.923	0.977	0.982	0.983	0.983	1.001	0.965	0.998	1.007	1.009	1.009
100	100	$1/T_i$	1.000	0.978	1.001	0.991	0.992	0.991	1.000	0.986	0.998	1.000	1.001	0.999
100	100	0.5	0.996	0.958	0.999	0.989	0.990	0.990	1.000	0.992	0.998	0.999	1.001	0.999
100	100	1.0	0.995	0.975	0.998	0.984	0.985	0.984	1.000	0.988	0.998	0.999	1.000	0.998
100	100	1.5	0.996	0.989	0.997	0.989	0.990	0.989	1.001	0.987	0.998	1.002	1.003	1.002
100	100	2.0	0.994	0.990	0.999	0.993	0.994	0.993	1.001	0.989	0.998	1.003	1.004	1.003
100	150	$1/T_i$	1.000	0.992	1.003	0.991	0.991	0.990	1.000	0.989	0.994	1.000	1.000	0.998
100	150	0.5	0.997	0.979	1.002	0.987	0.987	0.986	0.999	0.992	0.998	0.995	0.997	0.994
100	150	1.0	0.997	0.996	0.999	0.991	0.991	0.990	1.001	0.991	0.994	0.999	1.000	0.998
100	150	1.5	0.995	1.005	1.001	0.994	0.994	0.994	1.001	0.992	0.995	0.999	1.000	0.999
100	150	2.0	0.993	1.002	1.002	0.994	0.995	0.994	0.998	0.992	0.998	0.998	0.999	0.997
100	200	$1/T_i$	1.000	0.994	1.005	0.990	0.990	0.989	1.000	0.995	1.002	0.989	0.990	0.986
100	200	0.5	0.998	0.998	1.001	0.987	0.988	0.986	0.999	0.989	0.993	0.995	0.995	0.993
100	200	1.0	0.997	1.003	1.001	0.993	0.994	0.993	1.001	0.991	0.989	0.995	0.996	0.994
100	200	1.5	0.995	1.004	1.003	0.999	0.999	0.998	0.997	0.984	0.991	0.989	0.990	0.988
150	1	$1/T_i$	1.000	0.976	0.994	0.989	0.990	0.989	1.000	0.985	1.002	1.009	1.010	1.008
150	1	0.5	0.998	0.931	0.987	0.989	0.990	0.990	1.001	1.001	1.001	1.006	1.007	1.006
150	1	1.0	0.996	0.924	0.985	0.982	0.983	0.983	1.000	0.993	0.997	1.004	1.005	1.004
150	1	1.5	0.996	0.950	0.988	0.980	0.981	0.981	1.001	0.988	0.997	1.003	1.005	1.004
150	1	2.0	0.997	0.972	0.989	0.984	0.985	0.984	1.002	0.988	0.999	1.007	1.009	1.008
150	100	$1/T_i$	1.000	0.988	0.999	0.987	0.987	0.986	1.000	0.989	0.995	1.000	1.001	0.998
150	100	0.5	0.998	0.975	0.998	0.982	0.983	0.982	1.000	0.992	0.998	0.996	0.997	0.994
150	100	1.0	0.998	0.994	0.997	0.989	0.989	0.988	1.002	0.993	0.996	1.001	1.002	1.000
150	100	1.5	0.997	1.003	0.999	0.992	0.992	0.991	1.002	0.994	0.997	1.001	1.002	1.001
150	100	2.0	0.996	0.999	1.000	0.991	0.992	0.992	1.000	0.993	0.999	0.998	0.999	0.998
150	150	$1/T_i$	1.000	0.990	1.001	0.985	0.986	0.985	1.000	0.991	0.998	0.984	0.986	0.982
150	150	0.5	0.999	0.996	0.999	0.984	0.985	0.983	1.000	0.990	0.994	0.995	0.996	0.993
150	150	1.0	0.999	1.002	0.999	0.992	0.992	0.991	1.002	0.994	0.991	0.997	0.998	0.996
150	150	1.5	0.997	1.002	1.001	0.996	0.997	0.996	1.000	0.985	0.992	0.989	0.990	0.988
200	1	$1/T_i$	1.000	0.972	0.996	0.985	0.986	0.985	1.000	0.988	0.999	1.002	1.003	1.001
200	1	0.5	0.998	0.951	0.992	0.981	0.982	0.982	1.001	0.994	0.999	1.001	1.002	1.000
200	1	1.0	0.998	0.971	0.994	0.980	0.981	0.979	1.001	0.990	1.000	1.001	1.002	1.001
200	1	1.5	0.998	0.987	0.995	0.987	0.987	0.987	1.003	0.992	1.002	1.006	1.007	1.006
200	1	2.0	0.998	0.987	0.996	0.990	0.991	0.991	1.003	0.993	1.002	1.006	1.008	1.006
200	100	$1/T_i$	1.000	0.988	0.999	0.984	0.984	0.983	1.000	0.989	0.996	0.983	0.985	0.981
200	100	0.5	0.999	0.995	0.998	0.983	0.984	0.982	1.001	0.991	0.995	0.996	0.997	0.994
200	100	1.0	0.999	1.002	0.999	0.991	0.992	0.991	1.003	0.996	0.993	0.999	1.000	0.998
200	100	1.5	0.998	1.001	1.000	0.995	0.996	0.995	1.001	0.985	0.992	0.990	0.991	0.989
200	150	$1/T_i$	1.000	0.991	0.999	0.987	0.987	0.985	1.000	0.982	0.992	0.998	0.999	0.995
200	150	0.5	0.999	1.002	0.999	0.986	0.986	0.985	1.001	0.995	0.992	0.996	0.996	0.994
200	150	1.0	0.998	1.001	0.999	0.990	0.991	0.990	1.002	0.992	0.991	0.996	0.996	0.994

Note: The entries in the columns are the RMSFE’s. In particular, $\boldsymbol{w} = \hat{\boldsymbol{w}}_{i,0}$ for the column ‘Fixed’, $\boldsymbol{w} = \hat{\boldsymbol{w}}_{i,LsoMA}$ for the column ‘LsoMA’, $\boldsymbol{w} = \hat{\boldsymbol{w}}_{i,MMMA}$ for the column ‘MMMA’, $\boldsymbol{w} = \hat{\boldsymbol{w}}_{i,SBIC}$ for the column ‘SBIC’, $\boldsymbol{w} = \hat{\boldsymbol{w}}_{i,SHQC}$ for the column ‘SHQC’, and $\boldsymbol{w} = \hat{\boldsymbol{w}}_{i,Equal}$ for the column ‘Equal’.

Table 2: Forecasting performance of 'Plug-In' relative to alternative MA methods for the density forecast.

			$J=2$					$J=3$				
T	R	λ	Fixed	HM2007	PV2016	SBIC	Equal	Fixed	HM2007	PV2016	SBIC	Equal
1250	1	$1/T_i$	1.0000	1.0003	0.9992	0.9995	0.9992	1.0000	1.0003	0.9968	0.9997	0.9967
1250	1	0.5	0.9999	1.0004	0.9994	0.9993	0.9994	0.9999	1.0004	0.9967	0.9996	0.9967
1250	1	1.0	0.9999	1.0006	0.9996	0.9993	0.9996	0.9998	1.0005	0.9973	0.9995	0.9973
1250	1	1.5	0.9998	1.0007	0.9997	0.9991	0.9997	0.9997	1.0006	0.9974	0.9993	0.9974
1250	1	2.0	0.9997	1.0009	0.9997	0.9988	0.9997	0.9995	1.0006	0.9973	0.9991	0.9973
1250	250	$1/T_i$	1.0000	1.0003	0.9992	0.9995	0.9992	1.0000	1.0003	0.9966	0.9997	0.9966
1250	250	0.5	0.9999	1.0004	0.9995	0.9994	0.9995	0.9999	1.0002	0.9965	0.9996	0.9965
1250	250	1.0	0.9999	1.0005	0.9995	0.9991	0.9995	0.9998	1.0003	0.9969	0.9993	0.9969
1250	250	1.5	0.9998	1.0007	0.9994	0.9988	0.9994	0.9997	1.0002	0.9968	0.9990	0.9968
1250	250	2.0	0.9996	1.0009	0.9996	0.9980	0.9996	0.9993	1.0002	0.9970	0.9984	0.9969
1250	500	$1/T_i$	1.0000	1.0003	0.9991	0.9996	0.9991	1.0000	1.0002	0.9961	0.9998	0.9960
1250	500	0.5	0.9999	1.0003	0.9994	0.9993	0.9994	0.9999	1.0001	0.9965	0.9995	0.9965
1250	500	1.0	0.9999	1.0004	0.9993	0.9990	0.9993	0.9998	1.0000	0.9966	0.9992	0.9966
1250	500	1.5	0.9998	1.0005	0.9993	0.9985	0.9993	0.9996	0.9997	0.9964	0.9987	0.9964
1250	500	2.0	0.9995	1.0010	0.9996	0.9974	0.9996	0.9992	0.9998	0.9964	0.9980	0.9964
1250	750	$1/T_i$	1.0000	1.0003	0.9991	0.9997	0.9991	1.0000	1.0001	0.9959	0.9999	0.9959
1250	750	0.5	1.0000	1.0003	0.9994	0.9994	0.9994	0.9999	1.0000	0.9964	0.9995	0.9964
1250	750	1.0	0.9999	1.0004	0.9992	0.9990	0.9992	0.9998	0.9998	0.9962	0.9992	0.9962
1250	750	1.5	0.9997	1.0007	0.9993	0.9977	0.9993	0.9995	0.9995	0.9959	0.9981	0.9959
1500	1	$1/T_i$	1.0000	1.0003	0.9992	0.9994	0.9992	1.0000	1.0002	0.9965	0.9996	0.9965
1500	1	0.5	1.0000	1.0004	0.9994	0.9993	0.9994	0.9999	1.0002	0.9965	0.9995	0.9964
1500	1	1.0	0.9999	1.0005	0.9994	0.9991	0.9994	0.9998	1.0002	0.9968	0.9992	0.9968
1500	1	1.5	0.9998	1.0006	0.9994	0.9987	0.9994	0.9997	1.0001	0.9967	0.9989	0.9967
1500	1	2.0	0.9996	1.0008	0.9995	0.9979	0.9995	0.9994	1.0001	0.9968	0.9983	0.9968
1500	250	$1/T_i$	1.0000	1.0003	0.9990	0.9996	0.9990	1.0000	1.0001	0.9960	0.9997	0.9960
1500	250	0.5	1.0000	1.0003	0.9994	0.9993	0.9994	0.9999	1.0000	0.9965	0.9994	0.9964
1500	250	1.0	0.9999	1.0004	0.9993	0.9990	0.9993	0.9998	0.9999	0.9965	0.9991	0.9965
1500	250	1.5	0.9998	1.0005	0.9992	0.9984	0.9992	0.9997	0.9996	0.9963	0.9986	0.9963
1500	250	2.0	0.9995	1.0009	0.9995	0.9973	0.9995	0.9992	0.9996	0.9962	0.9978	0.9962
1500	500	$1/T_i$	1.0000	1.0003	0.9991	0.9997	0.9991	1.0000	1.0001	0.9959	0.9998	0.9958
1500	500	0.5	1.0000	1.0003	0.9994	0.9993	0.9994	0.9999	1.0000	0.9963	0.9994	0.9963
1500	500	1.0	0.9999	1.0003	0.9991	0.9990	0.9991	0.9998	0.9998	0.9961	0.9991	0.9961
1500	500	1.5	0.9997	1.0006	0.9992	0.9976	0.9992	0.9996	0.9994	0.9957	0.9980	0.9957
1500	750	$1/T_i$	1.0000	1.0003	0.9990	0.9997	0.9990	1.0000	1.0001	0.9959	0.9998	0.9959
1500	750	0.5	1.0000	1.0002	0.9992	0.9995	0.9992	0.9999	0.9998	0.9960	0.9996	0.9960
1500	750	1.0	0.9999	1.0002	0.9990	0.9992	0.9990	0.9999	0.9994	0.9956	0.9993	0.9955
1750	1	$1/T_i$	1.0000	1.0002	0.9990	0.9995	0.9990	1.0000	1.0001	0.9960	0.9996	0.9959
1750	1	0.5	1.0000	1.0003	0.9994	0.9993	0.9994	0.9999	1.0000	0.9964	0.9994	0.9964
1750	1	1.0	0.9999	1.0003	0.9992	0.9989	0.9992	0.9999	0.9999	0.9965	0.9991	0.9964
1750	1	1.5	0.9998	1.0004	0.9991	0.9983	0.9991	0.9997	0.9995	0.9962	0.9985	0.9962
1750	1	2.0	0.9995	1.0008	0.9994	0.9972	0.9994	0.9993	0.9994	0.9961	0.9976	0.9961
1750	250	$1/T_i$	1.0000	1.0002	0.9990	0.9996	0.9990	1.0000	1.0000	0.9958	0.9997	0.9958
1750	250	0.5	1.0000	1.0002	0.9993	0.9993	0.9993	0.9999	0.9999	0.9963	0.9994	0.9963
1750	250	1.0	0.9999	1.0003	0.9991	0.9990	0.9991	0.9998	0.9997	0.9961	0.9991	0.9960
1750	250	1.5	0.9997	1.0005	0.9991	0.9976	0.9991	0.9996	0.9993	0.9956	0.9978	0.9956
1750	500	$1/T_i$	1.0000	1.0002	0.9990	0.9997	0.9990	1.0000	1.0000	0.9959	0.9997	0.9958
1750	500	0.5	1.0000	1.0001	0.9992	0.9995	0.9992	1.0000	0.9998	0.9960	0.9995	0.9960
1750	500	1.0	0.9999	1.0002	0.9989	0.9992	0.9989	0.9999	0.9994	0.9955	0.9993	0.9955

Note: The entries in the columns are the rScore's. In particular, $\boldsymbol{w} = \hat{\boldsymbol{w}}_{i,0}$ for the column 'Fixed', $\boldsymbol{w} = \hat{\boldsymbol{w}}_{i,HM}$ for the column 'HM2007', $\boldsymbol{w} = \hat{\boldsymbol{w}}_{i,PV}$ for the column 'PV2016', $\boldsymbol{w} = \hat{\boldsymbol{w}}_{i,SBIC}$ for the column 'SBIC', and $\boldsymbol{w} = \hat{\boldsymbol{w}}_{i,Equal}$ for the column 'Equal'.

References

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003): "Modeling and Forecasting Realized Volatility," *Econometrica*, 71, 579–625.

Ando, T. and K. Li (2017): "A Weight-Relaxed Model Averaging Approach for High-Dimensional Generalized Linear Models," *Annals of Statistics*, 45, 2654–2679.

Ang, A. and M. Piazzesi (2003): "A No-Arbitrage Vector Autoregression of Term Structure Dynamics with Macroeconomic and Latent Variables," *Journal of Monetary Economics*, 50, 745–787.

Baillie, R. T. and R. J. Myers (1991): "Bivariate GARCH Estimation of the Optimal Commodity Futures Hedge," *Journal of Applied Econometrics*, 6, 109–124.

Banerjee, A., I. S. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha (2007): "A Generalized Maximum Entropy Approach to Bregman Co-clustering and Matrix Approximation," *Journal of Machine Learning Research*, 8, 1919–1986.

Banerjee, A., X. Guo, and H. Wang (2005a): "On the Optimality of Conditional Expectation as a Bregman Predictor," *IEEE Transactions on Information Theory*, 51, 2664–2669.

Banerjee, A., S. Merugu, I. S. Dhillon, and J. Ghosh (2005b): "Clustering with Bregman Divergences," *Journal of Machine Learning Research*, 6, 1705–1749.

Bao, Y., T.-H. Lee, and B. Saltoğlu (2007): "Comparing Density Forecast Models," *Journal of Forecasting*, 26, 203–225.

Barendse, S. and A. J. Patton (2022): "Comparing Predictive Accuracy in the Presence of a Loss Function Shape Parameter," *Journal of Business & Economic Statistics*, 40, 1057–1069.

Bates, J. M. and C. W. J. Granger (1969): "The Combination of Forecasts," *Journal of the Operational Research Society*, 20, 451–468.

Bauer, M. D., G. D. Rudebusch, and J. C. Wu (2012): "Correcting Estimation Bias in Dynamic Term Structure Models," *Journal of Business & Economic Statistics*, 30, 454–467.

Beck, A. (2017): *First-Order Methods in Optimization*, Society for Industrial and Applied Mathematics.

Berkowitz, J. (2001): "Testing Density Forecasts, with Applications to Risk Management," *Journal of Business & Economic Statistics*, 19, 465–474.

Bollerslev, T., R. F. Engle, and J. M. Wooldridge (1988): "A Capital Asset Pricing Model with Time-Varying Covariances," *Journal of Political Economy*, 96, 116–131.

Boyd, S. and L. Vandenberghe (2004): *Convex Optimization*, Cambridge University Press.

Bregman, L. (1967): "The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming," *USSR Computational Mathematics and Mathematical Physics*, 7, 200–217.

Chen, Y.-T. and C.-A. Liu (2023): "Model Averaging for Asymptotically Optimal Combined Forecasts," *Journal of Econometrics*, 235, 592–607.

Cheng, X. and B. E. Hansen (2015): "Forecasting with Factor-Augmented Regression: A Frequentist Model Averaging Approach," *Journal of Econometrics*, 186, 280–293, High Dimensional Problems in Econometrics.

Claeskens, G. and N. L. Hjort (2008): *Model Selection and Model Averaging*, Cambridge University Press.

Clemen, R. T. (1989): "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting*, 5, 559–583.

Collins, M., S. Dasgupta, and R. E. Schapire (2001): "A Generalization of Principal Components Analysis to the Exponential Family," in *Advances in Neural Information Processing Systems*, ed. by T. Dietterich, S. Becker, and Z. Ghahramani, MIT Press, vol. 14.

Conflitti, C., C. De Mol, and D. Giannone (2015): "Optimal Combination of Survey Forecasts," *International Journal of Forecasting*, 31, 1096–1103.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998): "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, 39, 863–883.

Diebold, F. X. and J. A. Lopez (1996): "Forecast Evaluation and Combination," in *Statistical Methods in Finance*, Elsevier, vol. 14 of *Handbook of Statistics*, chap. 8, 241–268.

Diebold, F. X., M. Shin, and B. Zhang (2023): "On the Aggregation of Probability Assessments: Regularized Mixtures of Predictive Densities for Eurozone Inflation and Real Interest Rates," *Journal of Econometrics*, 237, 105321.

Elliott, G., A. Gargano, and A. Timmermann (2013): "Complete Subset Regressions," *Journal of Econometrics*, 177, 357–373.

Elliott, G. and A. Timmermann (2008): "Economic Forecasting," *Journal of Economic Literature*, 46, 3–56.

——— (2016a): *Economic Forecasting*, Princeton University Press.

——— (2016b): "Forecasting in Economics and Finance," *Annual Review of Economics*, 8, 81–110.

Engle, R. (2002): "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models," *Journal of Business & Economic Statistics*, 20, 339–350.

Engle, R. F. and K. F. Kroner (1995): "Multivariate Simultaneous Generalized ARCH," *Econometric Theory*, 11, 122–150.

Erdogan, O., P. Bennett, and C. Ozyildirim (2015): "Recession Prediction Using Yield Curve and Stock Market Liquidity Deviation Measures," *Review of Finance*, 19, 407–422.

Estrella, A. and F. S. Mishkin (1998): "Predicting U.S. Recessions: Financial Variables as Leading Indicators," *Review of Economics and Statistics*, 80, 45–61.

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

Gneiting, T. (2011): "Making and Evaluating Point Forecasts," *Journal of the American Statistical Association*, 106, 746–762.

Gneiting, T. and A. E. Raftery (2007): "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378.

Granger, C. W. J. (1989): "Combining Forecasts—Twenty Years Later," *Journal of Forecasting*, 8, 167–173.

Granger, C. W. J. and R. Ramanathan (1984): "Improved Methods of Combining Forecasts," *Journal of Forecasting*, 3, 197–204.

Hall, S. G. and J. Mitchell (2007): "Combining Density Forecasts," *International Journal of Forecasting*, 23, 1–13.

Hansen, B. (2022): *Econometrics*, Princeton University Press.

Hansen, B. E. (1994): "Autoregressive Conditional Density Estimation," *International Economic Review*, 35, 705–730.

——— (2008): "Least-Squares Forecast Averaging," *Journal of Econometrics*, 146, 342–350.

——— (2010): *Multi-Step Forecast Model Selection*, University of Wisconsin-Madison, Working Paper.

——— (2014): "Model Averaging, Asymptotic Risk, and Regressor Groups," *Quantitative Economics*, 5, 495–530.

Hansen, P. R. and A. Lunde (2005): "A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?" *Journal of Applied Econometrics*, 20, 873–889.

Harville, D. A. (2008): *Matrix Algebra from a Statistician's Perspective*, Springer.

Hirano, K. and J. H. Wright (2017): "Forecasting with Model Uncertainty: Representations and Risk Reduction," *Econometrica*, 85, 617–643.

Hjort, N. L. and G. Claeskens (2003): "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899.

Jore, A. S., J. Mitchell, and S. P. Vahey (2010): "Combining Forecast Densities from VARs with Uncertain Instabilities," *Journal of Applied Econometrics*, 25, 621–634.

Kapetanios, G., J. Mitchell, S. Price, and N. Fawcett (2015): "Generalised Density Forecast Combinations," *Journal of Econometrics*, 188, 150–165.

Kauppi, H. and P. Saikkonen (2008): "Predicting U.S. Recessions with Dynamic Binary Response Models," *Review of Economics and Statistics*, 90, 777–791.

Laurent, S., J. V. Rombouts, and F. Violante (2013): "On Loss Functions and Ranking Forecasting Performances of Multivariate Volatility Models," *Journal of Econometrics*, 173, 1–10.

Leamer, E. E. (1978): *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Wiley.

Lee, T.-H. (2007): *Loss Functions in Time Series Forecasting*, University of California, Riverside, Working Paper.

Liao, J., X. Zong, X. Zhang, and G. Zou (2019): "Model Averaging Based on Leave-Subject-Out Cross-Validation for Vector Autoregressions," *Journal of Econometrics*, 209, 35–60.

Liao, J.-C. and W.-J. Tsay (2020): "Optimal Multistep VAR Forecast Averaging," *Econometric Theory*, 36, 1099–1126.

Lien, D. and Y. K. Tse (2002): "Some Recent Developments in Futures Hedging," *Journal of Economic Surveys*, 16, 357–396.

Liu, C.-A. (2015): "Distribution Theory of the Least Squares Averaging Estimator," *Journal of Econometrics*, 186, 142–159.

Liu, C.-A. and B.-S. Kuo (2016): "Model Averaging in Predictive Regressions," *Econometrics Journal*, 19, 203–231.

Liu, Q., Q. Yao, and G. Zhao (2020): "Model Averaging Estimation for Conditional Volatility Models with an Application to Stock Market Volatility Forecast," *Journal of Forecasting*, 39, 841–863.

Lohmeyer, J., F. Palm, H. Reuvers, and J.-P. Urbain (2019): "Focused Information Criterion for Locally Misspecified Vector Autoregressive Models," *Econometric Reviews*, 38, 763–792.

Negri, I. and Y. Nishiyama (2017): "Moment Convergence of Z-Estimators," *Statistical Inference for Stochastic Processes*, 20, 387–397.

Ng, L. (1991): "Tests of the CAPM with Time-Varying Covariances: A Multivariate GARCH Approach," *Journal of Finance*, 46, 1507–1521.

Nishiyama, Y. (2010): "Moment Convergence of M-Estimators," *Statistica Neerlandica*, 64, 505–507.

Opschoor, A., D. van Dijk, and M. van der Wel (2017): "Combining Density Forecasts Using Focused Scoring Rules," *Journal of Applied Econometrics*, 32, 1298–1313.

Park, S. Y. and S. Y. Jei (2010): "Estimation and Hedging Effectiveness of Time-Varying Hedge Ratio: Flexible Bivariate GARCH Approaches," *Journal of Futures Markets*, 30, 71–99.

Patton, A. J. (2011): "Volatility Forecast Comparison Using Imperfect Volatility Proxies," *Journal of Econometrics*, 160, 246–256.

——— (2020): "Comparing Possibly Misspecified Forecasts," *Journal of Business & Economic Statistics*, 38, 796–809.

Patton, A. J. and K. Sheppard (2009): "Evaluating Volatility and Correlation Forecasts," in *Handbook of Financial Time Series*, ed. by T. Mikosch, J.-P. Kreiß, R. A. Davis, and T. G. Andersen, Springer, 801–838.

Patton, A. J. and A. Timmermann (2007a): "Properties of Optimal Forecasts under Asymmetric Loss and Nonlinearity," *Journal of Econometrics*, 140, 884–918.

——— (2007b): "Testing Forecast Optimality under Unknown Loss," *Journal of the American Statistical Association*, 102, 1172–1184.

Pauwels, L. L., P. Radchenko, and A. L. Vasnev (2023): *High Moment Constraints for Predictive Density Combination*, Working Paper.

Pauwels, L. L. and A. L. Vasnev (2016): "A Note on the Estimation of Optimal Weights for Density Forecast Combinations," *International Journal of Forecasting*, 32, 391–397.

Petropoulos, F., D. Apiletti, V. Assimakopoulos, M. Z. Babai, D. K. Barrow, S. Ben Taieb, C. Bergmeir, R. J. Bessa, J. Bijak, J. E. Boylan, J. Browell, C. Carnevale, J. L. Castle, P. Cirillo, M. P. Clements, C. Cordeiro, F. L. Cyrino Oliveira, S. De Baets, A. Dokumentov, and ... Ziel, F. (2022): "Forecasting: Theory and Practice," *International Journal of Forecasting*, 38, 705–871.

Qiu, Y., X. Zhang, T. Xie, and S. Zhao (2019): "Versatile HAR Model for Realized Volatility: A Least Square Model Averaging Perspective," *Journal of Management Science and Engineering*, 4, 55–73.

Stock, J. H. and M. W. Watson (2001): "Vector Autoregressions," *Journal of Economic Perspectives*, 15, 101–115.

Timmermann, A. (2006): "Forecast Combinations," Elsevier, vol. 1 of *Handbook of Economic Forecasting*, chap. 4, 135–196.

Tong, H. (1990): *Non-Linear Time Series: A Dynamical System Approach*, Oxford University Press.

Tse, Y. K. and A. K. C. Tsui (2002): "A Multivariate Generalized Autoregressive Conditional Heteroscedasticity Model with Time-Varying Correlations," *Journal of Business & Economic Statistics*, 20, 351–362.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge: Cambridge University Press.

Wan, A. T., X. Zhang, and S. Wang (2014): "Frequentist Model Averaging for Multinomial and Ordered Logit Models," *International Journal of Forecasting*, 30, 118–128.

Wang, X., R. J. Hyndman, F. Li, and Y. Kang (2023): "Forecast Combinations: An over 50-Year Review," *International Journal of Forecasting*, 39, 1518–1547.

Wang, Y., F. Ma, Y. Wei, and C. Wu (2016): "Forecasting Realized Volatility in a Changing World: A Dynamic Model Averaging Approach," *Journal of Banking & Finance*, 64, 136–149.

Zhang, C., Y. Jiang, and Y. Chai (2010): "Penalized Bregman Divergence for Large-Dimensional Regression and Classification," *Biometrika*, 97, 551–566.

Zhang, C., Y. Jiang, and Z. Shang (2009): "New Aspects of Bregman Divergence in Regression and Classification with Parametric and Nonparametric Estimation," *Canadian Journal of Statistics*, 37, 119–139.

Zhang, T. (2003): "Sequential Greedy Approximation for Certain Convex Optimization Problems," *IEEE Transactions on Information Theory*, 49, 682–691.

Zhang, X. and C.-A. Liu (2023): "Model Averaging Prediction by K-Fold Cross-Validation," *Journal of Econometrics*, 235, 280–301.

Zhang, X., D. Yu, G. Zou, and H. Liang (2016): "Optimal Model Averaging Estimation for Generalized Linear Models and Generalized Linear Mixed-Effects Models," *Journal of the American Statistical Association*, 111, 1775–1790.