Predictions with dynamic Bayesian predictive synthesis are exact minimax

Kōsaku Takanashi* & Kenichiro McAlinn†

Abstract

We analyze the combination of multiple predictive distributions for time series data when all forecasts are misspecified. We show that a specific dynamic form of Bayesian predictive synthesis—a general and coherent Bayesian framework for ensemble methods—produces exact minimax predictive densities with regard to Kullback-Leibler loss, providing theoretical support for finite sample predictive performance over existing ensemble methods. A simulation study that highlights this theoretical result is presented, showing that dynamic Bayesian predictive synthesis is superior to other ensemble methods using multiple metrics.

Keywords: Ensemble methods; Exact minimax; Time series; Bayesian analysis; $M$-open.
1 Introduction

Prediction and decision making is a central task for statistical methods in many fields, including economics, climatology, and epidemiology, where predictions of future events directly impact policy decision making. In much, if not all, of these applications, assuming that there is a “true” model, or forecast, is not realistic. Assuming such a model exists, or utilizing methodology that, implicitly or explicitly, assumes it for its theoretical justification, can be detrimental in actual application and decision making. As the saying goes, “all models are wrong,” but, if all models are wrong, yet the validity of a method hinges on one model being “right,” then the decision maker will always be misguided. This has spurred recent research into developing methodologies that take into account the fact that all models are misspecified; a setting referred to as $M$-open (as opposed to $M$-closed, when the true model is nested in the set of models).

One way to mitigate misspecification and produce useful predictions in an $M$-open setting is to ensemble multiple sources of information (e.g. predictive distributions). Depending on the field and context, this process is called forecast combination, model averaging, or ensemble learning, though we will use ensemble methods as an umbrella term in this paper. While the idea to ensemble has been around for at least a century (Galton 1907), the seminal paper by Bates and Granger (1969) gave the first theoretical justification for it. Recently, ensemble methods have seen a surge in interest due, partly, to an increase in usage and availability of more complex models, and computational tools that allow for fast, simultaneous calculation of a large class of models. In statistics, particularly Bayesian, Bayesian model averaging (BMA: Raftery et al. 1997; Hoeting et al. 1999) has been a staple; having been integrated into many toolboxes and methodologies. Although BMA is ideal in an $M$-closed setting (Madigan and Raftery 1994), it is known to converge to the “wrong” model in $M$-open settings. Recent developments, such as Bayesian stacking (Yao et al. 2018), aim to address this issue. In machine learning, ensemble learning, including bagging (Breiman 1996), boosting (Schapire 2003), and stacking (Džeroski and Ženko 2004), have been used extensively, although the ensemble method itself is rather basic (often just equal weight averaging). In econometrics, the increase in usage of models and forecasters that produce full density forecasts has stimulated developments into formal forecasting models that utilize this information to improve policy decision making. A number of ideas for density combination strategies have emerged in response (e.g. Hall and Mitchell 2007; Geweke and Amisano 2011; Billio et al. 2012; Aastveit et al. 2014; Kapetanios et al. 2015; Pettenuzzo and Ravazzolo 2016; Negro et al. 2016; Aastveit et al. 2018a,b; Diebold and Shin 2019). It has been convincingly demonstrated in many applications, across multiple domains, that ensemble methods result in improved forecast performance and decision making.

Responding to a need for a general, coherent Bayesian framework, McAlinn and West (2019) extended the work on expert opinion analysis (Genest and Schervish 1985; West and Crosse 1992; West 1992) to develop Bayesian predictive synthesis (BPS). While it can be shown that most, if not
all, ensemble methods are a special case of BPS, McAlinn and West (2019) developed a specific form—a dynamic latent (agent) factor model—to ensemble multiple predictive densities for predict-
ing time series data. Multiple applications (Bianchi and McAlinn 2018; McAlinn et al. 2020; Capek et al. 2020; McAlinn 2021) have shown that BPS outperforms other benchmarks and state-of-the-
art ensemble methods in terms of predictive performances, though the “why” has been unknown.

In this paper, we develop a novel theoretical strategy based on stochastic processes. Utilizing this strategy, we show that dynamic BPS in McAlinn and West (2019) produces exact minimax predictive distributions with regard to Kullback-Leibler loss, providing theoretical support for its finite sample predictive performance. As far as we are aware, this paper is the first to define the Kullback-Leibler risk for non-stationary time series in an $\mathcal{M}$-open setting, and to provide finite sample theoretical justification for an ensemble method.

2 Preliminaries

2.1 Bayesian predictive synthesis

A decision maker, $D$, is interested in forecasting a quantity $y \in \mathbb{R}$, and solicits $J$ agents, where agents encompass models, forecasters, institutions, etc. Agents are denoted as $A_j, j \in J$. Each agent, $A_j$, produces a predictive distribution, $h_j(\hat{y}_j)$, which comprises the set, $\mathcal{H} = \{h_1(\cdot), \ldots, h_J(\cdot)\}$. While BPS theory does not restrict the agent predictions to be predictions of the quantity of interest (e.g. using agent predictions of U.S. inflation to predict Japanese inflation), we consider the special case where all agents predict the same quantity of interest, without loss of generality (in McAlinn and West 2019, the agent predictions are denoted as $h_j(x_j)$, to express this generality).

In the BPS framework, the set, $\mathcal{H}$, is synthesized via Bayesian updating with the posterior of the form,

$$p(y|\mathcal{H}) = \int_{\hat{y}} \alpha(y|\hat{y}) \prod_{j=1,J} h_j(\hat{y}_j) d\hat{y},$$

(1)

where $\hat{y} = \hat{y}_{1:J} = (\hat{y}_1, \ldots, \hat{y}_J)^T$ is a $J$-dimensional latent vector and $\alpha(y|\hat{y})$ is a conditional probability density function for $y$ given $\hat{y}$, called the synthesis function. Eq. (1) is only a coherent Bayesian posterior if it satisfies the consistency condition (see, Genest and Schervish 1985; West and Crosse 1992; West 1992; McAlinn and West 2019). Critical to the BPS framework, the theory does not imply a specific form of synthesis function, $\alpha(y|\hat{y})$. McAlinn and West (2019) show that other ensemble methods, such as equal weight averaging, Bayesian model averaging (Hoeting et al. 1999), and optimal linear pools (Geweke and Amisano 2011), are special cases of BPS; given a specific synthesis function and prior specifications.

Specifically, the consistency condition states that given $D$’s prior, $p(y)$, and her prior expectation of what the agents will produce before observing the forecasts, $m(\hat{y})$, her priors have to be consistent: $p(y) = \int_{\hat{y}} \alpha(y|\hat{y}) m(\hat{y}) d\hat{y}$, for eq. (1) to be a coherent Bayesian posterior.
To improve prediction and information flow for decision making with time series data, McAlinn and West (2019) developed a dynamic specification of BPS. In the dynamic setting, $D$ is sequentially predicting a time series $y_t, t = 1, 2, \ldots$, and receives, for all periods, forecast densities from $A_j, j \in J$. At each time $t$, $D$ aims to forecast $y_{t+1}$ and receives the set $H_{t+1} = \{h_{1,t+1}(\hat{y}_{1,t+1}), \ldots, h_{J,t+1}(\hat{y}_{J,t+1})\}$, a collection of agent predictive densities produced at $t$, from the agents. The full information set used by $D$ is thus $\{y_{1:t}, H_{1:t+1}\}$. As time passes, $D$ learns about characteristics of the agents (bias, dependencies, etc.). Thus, the Bayesian model will involve parameters that define the BPS framework and for which $D$ updates information over time. From eq. (1), $D$ has a time $t$ distribution for $y_{t+1}$ of the form

$$p(y_{t+1}|\Phi_{t+1}, y_{1:t}, H_{1:t+1}) = \int \alpha_t(y_{t+1}|\Phi_{t+1}, \Phi_{t+1}) \prod_{j=1:J} h_{j,t+1}(\hat{y}_{j,t+1}) d\hat{y}_{j,t+1},$$

(2)

where now $\Phi_{t+1}$ represents time-varying parameters defining the synthesis probability density function for which $D$ has current beliefs represented in terms of a current (time $t$) posterior $p(\Phi_{t+1}|y_{1:t}, H_{1:t})$. Again, the time-varying synthesis function, $\alpha_t(y_{t+1}|\hat{y}_{t+1}, \Phi_{t+1})$, need not be a specific form. McAlinn and West (2019) specifies a dynamic synthesis function,

$$\alpha_t(y_{t+1}|\hat{y}_{t+1}, \Phi_{t+1}) = N(y_{t+1}|\theta_{0,t+1} + F_{t+1}^T \theta_{t+1}, \nu_{t+1})$$

(3)

with $F_{t+1} = (\hat{y}_{1,t+1}, \ldots, \hat{y}_{J,t+1})^T$ and $\theta_{t+1} = (\theta_{1,t+1}, \ldots, \theta_{J,t+1})^T$,

the latter being the $J$–vector of time-varying calibration coefficients, $\theta_{0,t+1}$ being the time-varying calibration intercept, and the conditional variance $\nu_{t+1}$ defining residual variation beyond that explained by the regression on latent agent factors. The functional model parameters are now $\Phi_{t+1} = (\theta_{0,t+1}, \theta_{t+1}, \nu_{t+1})$ at each time $t$. This BPS specification defines the standard dynamic linear model (DLM: West and Harrison 1997; Prado and West 2010):

$$y_{t+1} = \theta_{0,t+1} + F_{t+1}^T \theta_{t+1} + \nu_{t+1}, \quad \nu_{t+1} \sim N(0, \nu_{t+1})$$

(4a)

$$\theta_{t+1}' = \theta_{t}' + \omega_{t+1}, \quad \omega_{t+1} \sim N(0, \nu_{t+1} W_{t+1})$$

(4b)

where $\theta_{t+1}' = [\theta_{0,t+1}, \theta_{t+1}]$ evolves in time according to a linear/normal random walk with innovations variance matrix $\nu_{t+1} W_{t+1}$ at time $t + 1$, and $\nu_{t+1}$ is the residual variance in predicting $y_{t+1}$ based on past information and the set of agent forecast distributions. The residuals $\nu_{t+1}$ and evolution innovations $\omega_{s+1}$ are independent over time and mutually independent for all $t, s$. Eqns. (4) define a dynamic latent (agent) factor model, where the $\hat{y}_{t+1}$ vectors in each $F_{t+1}$ are latent variables.
Specifically, the predictive distribution is produced as follows. The likelihood function is

\[
\hat{p}_t (y_{t+1} | \theta^*_t + 1, \hat{y}_{t+1}) = \frac{1}{\sqrt{2\pi v_t^2}} \exp \left( -\frac{(y_{t+1} - (\theta^*_t + 1, \hat{y}_{t+1}))^2}{2v_t^2} \right),
\]

where \( \hat{y}_{t+1} = [1, \hat{y}_{t+1}] \). The predictive distribution of BPS is

\[
p(y_{t+1} | \Phi_{t+1}, y_{1:t}, H_{1:t+1}) = \int \int \hat{p}_t (y_{t+1} | \theta^*_t + 1, \hat{y}_{t+1}) \pi (\theta^*_t + 1 | F_t) d\theta^*_t + 1 h_{t+1} (\hat{y}_{t+1}) d\hat{y}_{t+1},
\]

where \( F_t \) is the smallest \( \sigma \)-algebra that makes all stochastic variables at \( t \) measurable. Here,

\[
\pi (\theta^*_t + 1 | F_t) = \int \pi (\theta^*_t + 1 | \theta_t^*) \pi (\theta_t^* | F_t) d\theta_t^*,
\]

where \( \pi (\theta_t^* | F_t) \) is the posterior distribution of \( \theta_t^* \) at time \( t \), after observing \( y_t \) and \( \hat{y}_{t+1} \). At \( t = 0 \), we denote the (initial) prior distribution of \( \theta^*_0 \) as \( \rho (\theta^*_0) \). Under this (initial) prior, the predictive distribution is denoted as \( \hat{p}_t^\rho (y_{t+1} | F_t) \), and the posterior distribution as \( \pi^\rho (\theta_t^* | F_t) \), both at time \( t \).

### 2.2 Semi-Martingale processes

As a preliminary to our theoretical analysis, we assume that the predictive process, \( y_t \), and the agent (predictive) process, \( \hat{y}_{j,t}, j \in J \), are both square integral processes and filtration \( F_t \)-adapted. Then, both \( y_t, \hat{y}_{j,t}, j \in J \) are uniquely decomposed, from the Doob decomposition, as the following:

\[
X_t = X_0 + M_t + V_t, \quad M_0 = V_0 = 0
\]

where \( X_0 \) is a \( F_0 \)-measurable stochastic variable, \( M \) is a square integrable martingale, and \( V \) is square integrable and predictable, thus \( F_{t-1} \)-adapted. Further, the optimal approximate process of \( y_{t+1} \), utilizing the agent process, \( \hat{y}_{j,t+1}, j \in J \), can be represented as (Föllmer-Schweizer (FS) decomposition; [Schweizer 1995]),

\[
y_{t+1} = a^* + \sum_{s=1}^{t} \langle \theta^*_s + 1, \Delta \hat{y}_s \rangle + z_{t+1},
\]

where \( (a^*, \theta^*_s + 1) \) are the true parameters of the FS decomposition (due to \( \theta^*_s + 1 \) being determined at \( t \)) and \( \Delta \hat{y}_s = \hat{y}_{s+1} - \hat{y}_s \). When \( \{a\} \) is constant, but \( \{\theta_s\}_{1 \leq s \leq t+1} \) vary, then \( z_{t+1} \) is a martingale process, which is orthogonal to the martingale term of \( \hat{y}_t \).

**Remark 2.1.** A martingale process, \( z_t \), is orthogonal to \( M_t \) (the martingale component of \( \hat{y}_t \)) when \( M_t z_t \) is a martingale. Here, \( \Delta M_t \Delta z_t, \Delta M_t z_t, \) and \( M_t \Delta z_t \) all have expectation zero.
Remark 2.2. When \( y_t, \hat{y}_t, \) and \( z_t \) are all square-integrable martingales, \( \langle \eta, M_t \rangle \) are martingales with regard to \( \forall \eta \in \mathbb{R}^J \), and assume,

\[
\mathbb{E} \left[ \langle \eta, \Delta M_t \rangle^2 \bigg| \mathcal{F}_t \right] = \langle \eta, \Sigma (t) \eta \rangle,
\]

where \( \Sigma (t) \) is a \( \mathcal{F}_t \)-measurable \( \mathbb{R}^{J \times J} \) matrix value function. Then,

\[
\mathbb{E} \left[ z_{t+1} \big| \hat{y}_{t+1}, \mathcal{F}_t \right] = 0.
\]

Therefore, we have

\[
\mathbb{E} \left[ y_{t+1} \big| \hat{y}_{t+1}, \mathcal{F}_t \right] = a^* + \sum_{s=1}^{t} \left( \theta^*_s, \Delta \hat{y}_s \right).
\]

where eq. (7) is the optimal approximate projection to the joint plane of agent processes, \( \hat{y}_{j,t+1}, \) \( j \in J \), and the extra term \( a^* \).

3 Exact predictive minimaxity of dynamic Bayesian predictive synthesis

3.1 Predictive minimaxity under Kullback-Leibler risk

Predictive minimaxity, as examined in this paper, is defined as the Kullback-Leibler (KL) risk of the target: the transition probability of the “best” possible synthesis, given the available information, under different synthesis functions. While this means that we are considering the KL risk on the functional space that defines the synthesis function, it is sufficient to consider the synthesis function as defined as having a linear combination of coefficients/weights, \( \theta_{t+1} \), and an extra term, \( a \), from Itô’s lemma and FS decomposition. This is done because the goal is to determine the predictive ability of BPS, not the geometric form of the synthesis function. In other words, with regard to its predictive ability, the synthesis function space that defines the KL risk is solely determined by the \( J + 1 \) parameters, \( (a, \theta_{t+1}) \), since we are only concerned with the one-step ahead forecasts, and \( (a, \theta_{t+1}) \) need not be seen as a function of \( t \). This allows us to define the finite sample minimax KL risk for predictive synthesis in a succinct manner.

Given eq. (7), the best possible transition probability of \( y_{t+1} \) can be written as

\[
p_t \left( y_{t+1} \big| \mathcal{F}_t, a^*, \theta^*_{t+1} \right) = \int_{\mathbb{R}^J} p_t \left( y_{t+1} \big| \mathcal{F}_t, a^*, \theta^*_{t+1}, \hat{y}_{t+1} - \hat{y}_t \right) h_{t+1} (\hat{y}_{t+1}) \, d\hat{y}_{t+1}.
\]

The goal is to show that the Bayesian predictive distribution of BPS, denoted as \( \hat{p}_t \left( y_{t+1} \big| \mathcal{F}_t \right) \), is minimax in terms of the KL risk with regard to the transition probability, \( p_t \left( y_{t+1} \big| \mathcal{F}_t, a^*, \theta^*_{t+1} \right) \).
First, the KL divergence is defined as
\[
\text{KL} \left( p_t \left( y_{t+1} \mid \mathcal{F}_t, a^*_t, \theta^*_t \right) \mid \hat{p}_t \left( y_{t+1} \mid \mathcal{F}_t \right) \right) = \int_{\mathbb{R}} p_t \left( y_{t+1} \mid \mathcal{F}_t, a^*, \theta^*_t \right) \log \frac{p_t \left( y_{t+1} \mid \mathcal{F}_t, a^*, \theta^*_t \right)}{\hat{p}_t \left( y_{t+1} \mid \mathcal{F}_t \right)} dy_{t+1}.
\]

Then, the KL risk is
\[
R_{\text{KL}} \left( (a^*, \theta^*_t), \hat{p}_t \right) = \int_{Y} \text{KL} \left( p_t \left( y_{t+1} \mid \mathcal{F}_t, a^*, \theta^*_t \right) \mid \hat{p}_t \left( y_{t+1} \mid \mathcal{F}_t \right) \right) d\mu_Y \left( y_t, \cdots, y_0 \right).
\]

Here, the probability measure, \( \mu_Y \left( y_t, \cdots, y_0 \right) \), is a cylindrical measure of the process, \( \{y_t\} \). Thus, finding the predictive distribution, \( \hat{p}_t \), that achieves
\[
R_{\text{minimax}} = \min_{\hat{p}_t} \max_{a^*, \theta^*_{t+1}} R_{\text{KL}} \left( (a^*, \theta^*_{t+1}), \hat{p}_t \right),
\]
at time \( t \), is minimax in terms of KL risk.

Utilizing the FS decomposition allows us to define the minimax risk of the averaged estimator. Since the interest is on the comparison of synthesized forecasts, it is appropriate and preferable to consider the minimax risk with regard to the best synthesized estimator, rather than assuming some parametric model for the data generating process and considering the minimax risk with regard to those parameters.

### 3.2 Main theorem

We consider some assumptions on \( y_t \), in preparation for minimax analysis. For the FS decomposition of \( y_t \) (eq. 7), we assume the martingale component of \( \hat{y}_t \) and the increment of its orthogonal martingale process, \( z_t \), are square-integrable for all \( t \). The probability distribution function of \( z_t \) is denoted as \( p_{t,z} (\cdot) \), where \( p_{t,z} (\cdot) \) can depend on \( t \), given the subscript \( t \). Note that \( p_{t,z} (\cdot) \) can be considered as a probability density conditional on \( \Delta \hat{y}_t \), from [Williams (1991)], Section 9.5, Section 14.14.

**Theorem 3.1.** Suppose the following conditions hold;

(i) The predictive process, \( y_t \), and agent processes, \( \hat{y}_{j,t}, j \in J \), are square integrable processes, and are filtration \( \mathcal{F}_t \)-adapted;

(ii) For the martingale component \( M_t \) of \( \hat{y}_t \), \( \langle \eta, M_t \rangle \) are martingales with regard to \( \forall \eta \in \mathbb{R}^J \), and assume there exists a \( \mathcal{F}_t \)-measurable \( \mathbb{R}^{J \times J} \) matrix value function, \( \Sigma (t) \), and

\[
\mathbb{E} \left[ \left( \langle \eta, \Delta M_t \rangle \right)^2 \mid \mathcal{F}_t \right] = \langle \eta, \Sigma (t) \eta \rangle.
\]

Then, the BPS predictive distribution, \( \hat{p}_t^m (y_{t+1} \mid \mathcal{F}_t) \), with a Lebesgue measure prior, \( m \), for the initial
prior \((t = 0)\), is exact minimax with regard to the Kullback-Leibler risk (eq. 8):

\[
R_{KL} \left( (a, \theta_{t+1}), \hat{p}^m_t \right) = \min_{\hat{p}_t} \max_{a, \theta_{t+1}} R_{KL} \left( (a, \theta_{t+1}), \hat{p}_t \right).
\]

The proof of this theorem is done by proving the following two conditions, according to the equalizer rule (Berger 1985, Chapter 5.3.2.), following Corollary 1. of George et al. (2006):

1. The KL risk of \(\hat{p}^m_t (y_{t+1} | F_t)\) is constant for all sets of parameters, \((a^*, \theta^*_{t+1})\):

\[
R_{KL} \left( (a^*, \theta^*_{t+1}), \hat{p}^m_t \right) = c, \quad \forall \ (a^*, \theta^*_{t+1}).
\]

2. \(\hat{p}^m_t (y_{t+1} | F_t)\) is extended Bayes. In other words, the Bayes risk is approximated by the prior sequence, \(\{\rho_k\}\), where

\[
\lim_{k \to \infty} B (\rho_k, \hat{p}^m_k) = B (\rho_k, \hat{p}^m_t).
\]

The Bayes risk is defined as

\[
B (\rho_k, \hat{p}^m_k) = \int R_{KL} \left( (a^*, \theta^*_{t+1}), \hat{p}^m_t \right) \rho_k (a^*, \theta^*_{t+1}),
\]

which is written with regard to the posterior parameters, \((a^*, \theta^*_{t+1})\), and not the prior on \(\theta'_0\). However, this Bayes risk is effectively equivalent to the prior specification of \(\theta'_0\), due to recursive updating. More specifically, the posterior distribution, \(\pi \left( \theta'_t | F_t \right)\), is given by the generalized Bayes formula for filtering (see Appendix C for notation),

\[
\pi \left( \theta'_t | F_t, \{x_s\}_{s=1}^{t+1} \right) = \frac{\int_{(\mathbb{R}^j)^{t-1}} \beta_t (\theta', y) d\mu^0_\theta (\theta_0, \cdots, \theta_{t-1})}{\int_{(\mathbb{R}^j)^{t}} \beta_t (\theta', y) d\mu^0_\theta (\theta_0, \cdots, \theta_t)},
\]

though because we assume a random walk (eq. 9b) for the process, \(\{\theta'_t\}\), the cylindrical measure, \(\mu^0_\theta (\theta_0, \cdots, \theta'_t)\), is determined by the prior distribution of \(\theta'_0\): \(\rho (\theta'_0)\).

### 3.3 Risk constant

First, we show that the predictive distribution of BPS is risk constant. While the parameters of interest are \((a^*, \theta^*_{t+1})\), because the synthesis function for BPS in McAlinn and West (2019) is a dynamic linear model (DLM), we first transform the observation equation to a discrete DLM form. Consider the random walk DLM,

\[
\begin{align*}
y_{t+1} &= \theta_{0,t+1} + \langle \theta_{t+1}, \tilde{y}_{t+1} \rangle + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim N (0, v_{t+1}), \\
\theta'_{t+1} &= \theta'_t + \omega_{t+1}, \quad \omega_{t+1} \sim N (0, v_{t+1} W_{t+1}),
\end{align*}
\]
where \( \theta'_t \) is the vector \([\theta_{0,t}, \theta_t]\), with intercept, \(\theta_{0,t}\), \(\hat{y}'_t\) is \([1, \hat{y}_t]\), and \(\omega_{t+1}\) is a \(J + 1\)-dimensional vector of normal random numbers. We will now correspond the observational equation (eq. 9a) to its FS decomposition (eq. 7). First, noting that,

\[
\langle \theta_{t+1}, \hat{y}_{t+1} \rangle - \langle \theta_t, \hat{y}_t \rangle = \langle \Delta \theta_t, \hat{y}_t \rangle + \langle \theta_{t+1}, \Delta \hat{y}_t \rangle
\]

we have

\[
y_{t+1} = \theta_{0,t+1} + \langle \theta_t, \hat{y}_t \rangle + \langle \Delta \theta_t, \hat{y}_t \rangle + \langle \theta_{t+1}, \Delta \hat{y}_t \rangle + \varepsilon_{t+1},
\]

by setting

\[
\theta_{0,t+1} = \mathbb{E} [\theta_{0,t} \mid \mathcal{F}_t] - \langle \Delta \theta_t, \hat{y}_t \rangle,
\]

\[
\mathbb{E} [\theta_{0,t} \mid \mathcal{F}_t] = \theta_{0,0} + y_t - \langle \theta_t, \hat{y}_t \rangle,
\]

where \(\theta_{0,0}\) is an initial value of \(\{\theta_{0,t}\}\). From this, we have

\[
\mathbb{E} [\theta_{0,t+1} \mid \mathcal{F}_t] = \theta_{0,0} + y_t - \langle \theta_t, \hat{y}_t \rangle,
\]

where \(\theta_{0,t}\) follows a random walk. Therefore, eq. (9a) is transformed to

\[
y_{t+1} = y_t + \theta_{0,0} + \langle \theta_{t+1}, \Delta \hat{y}_t \rangle + \varepsilon_{t+1},
\]  

(10)

where the parameters considered in eq. (7), \((a^*, \theta_{t+1}^*)\), correspond to \((\theta'_t)\) in eq. (10) through

\[
a \leftrightarrow \theta_{0,0} + y_t - \sum_{s=1}^{t-1} \langle \theta_{s+1}^*, \Delta \hat{y}_s \rangle - z_t,
\]

\[
\theta_{t+1}^* \leftrightarrow \theta_{t+1}.
\]

Further, we assume the error, \(\varepsilon_{t+1}\), in eq. (10), follows a normal distribution. In general, the distribution of \(\varepsilon_{t+1}\) and \(z_{t+1}\), obtained from the FS decomposition of the data generating process, need not be the same. Since \(z_{t+1}\) is a process that includes omitted variables, it is unrealistic to exactly identify it.

Noting that \(y_{t+1} = y_t + \Delta y_t\), the probability density function of the increment of \(y_t\), can be written using the probability density function of its martingale process, \(z_{t+1}\), as

\[
p_t (y_{t+1} \mid \mathcal{F}_t, a^*, \theta_{t+1}^*, \Delta \hat{y}_t) = p_{t,z} (y_{t+1} - (Y_t + \langle \theta_{t+1}^*, \Delta \hat{y}_t \rangle)).
\]

Here, \(Y_t = \mathbb{E} [y_{t+1} \mid \mathcal{F}_t]\) is \(\mathcal{F}_t\)-measurable. Then, the transition probability can be obtained by
integrating out the agents’ predictive processes, $\hat{y}_{t+1}$:

$$
pt \left(y_{t+1} \mid F_t, a^*, \theta^*_{t+1} \right) = \int_{\mathbb{R}^d} pt \left(y_{t+1} - \left(Y_t + (\theta^*_{t+1}, \Delta \hat{y}_t) \right) \right) \sum_j h_{j,t+1} (\hat{y}_{j,t+1}) d\hat{y}_{j,t+1}. 
$$

(11)

If we shift transform this transition probability with $y_{t+1} \rightarrow y_{t+1} - \left( Y_t + \langle \theta^*_{t+1}, \Delta \hat{y}_t \rangle \right)$, we have,

$$
pt \left(y_{t+1} - \left(Y_t + \langle \theta^*_{t+1}, \Delta \hat{y}_t \rangle \right) \mid F_t, a^*, \theta^*_{t+1} \right) = pt \left(y_{t+1} \mid F_t, 0, 0 \right).
$$

If we shift transform as $y_{t+1} \rightarrow \tilde{y}_{t+1} = y_{t+1} - \left( Y_t + \langle \theta^*_{t+1}, \Delta \hat{y}_t \rangle \right)$, we have

$$
\tilde{y}_{t+1} = y_t - Y_t + \langle \theta_{t+1} - \theta^*, \Delta \hat{y}_t \rangle + \varepsilon_{t+1}, \varepsilon_{t+1} \sim N(0, V_t).
$$

Therefore, if we set $\tilde{\theta}_t = \theta_t - \theta^*$, then

- $\tilde{\theta}_{0,t+1} \mid F_t = \theta_{0,0} + y_t - \left( \tilde{\theta}_t, \hat{y}_t \right) - \left( \Delta \theta_t, \hat{y}_t \right) - Y_t$,
- $\tilde{\theta}_t \mid F_t = \theta_{0,0} + y_t - \left( \tilde{\theta}_t, \hat{y}_t \right) - Y_t$.

From this, we have

$$
\tilde{y}_{t+1} = \tilde{\theta}_{0,t+1} + \left( \tilde{\theta}_{t+1}, \hat{y}_{t+1} \right) + \varepsilon_{t+1}, \varepsilon_{t+1} \sim N(0, V_t).
$$

The predictive distribution is given as

$$
\hat{p}_t \left( \tilde{y}_{t+1} \mid F_t \right) = \int \int \hat{p}_t \left( \tilde{y}_{t+1} \mid \tilde{\theta}^t_{t+1}, \hat{y}^t_{t+1} \right) \pi \left( \tilde{\theta}^t_{t+1} \mid F_t \right) d\tilde{\theta}^t_{t+1} h_{t+1} \left( \tilde{y}_{t+1} \right) d\hat{y}_{t+1}.
$$

Here, the prior distribution of $\theta^t$, at $t = 0$, is a Lebesgue measure, $m(\cdot)$, and the state space evolves following a random walk (eq. 9b). Then, the following lemma holds.

**Lemma 3.1.** Consider a state space, $\theta^t$, that is updated through a random walk (eq. 9b) evolution, with the prior distribution for $\theta^t$ at $t = 0$ being a Lebesgue measure, $m(\cdot)$. The state space, $\theta^t$, is equivalent in law with regard to the shift transform, $\theta^t \rightarrow \theta^t - \theta^* = \tilde{\theta}^t$:

$$
\theta_{t+1} = \theta_t + \omega_t \implies \tilde{\theta}_{t+1} = \tilde{\theta}_t + \omega_t,
$$

$$
\theta_{0,t+1} = \theta_{0,t} + \omega_t^0 \implies \tilde{\theta}_{0,t+1} = \tilde{\theta}_{0,t} + \omega_t^0.
$$

Therefore, the KL risk is constant for any set of true parameters, $(a^*, \theta^*_{t+1})$:

$$
R_{KL} ((a^*, \theta^*_{t+1}) \mid \tilde{p}_t^m) = R_{KL} ((0, 0) \mid \tilde{p}_t^m) = c.
$$

9
The KL risk is constant between the transition probability of the data generating process, \( p_t (y_{t+1} \mid F_t, a^*, \theta^*_{t+1}) \), and the predictive distribution, \( q^m (y_{t+1} \mid F_t) \).

### 3.4 Extended Bayes

We now consider an extended Bayes strategy given an initial prior, \( \rho_k (\theta'_0) = N (0, \sigma^2_k I) \), and show that the Bayes risk limit is

\[
B (\rho_k, \hat{\rho}_k) \rightarrow B (\rho_k, \hat{\rho}_k^m), \quad (k \rightarrow \infty),
\]

when \( \sigma^2_k \rightarrow \infty \) as \( k \rightarrow \infty \).

The difference in KL risk of the above is

\[
R_{KL} \left( (a^*, \theta^*_{t+1}) \; \mid \; \hat{\rho}_k^m \right) - R_{KL} \left( (a^*, \theta^*_{t+1}) \; \mid \; \hat{\rho}_k^{m*} \right) = \int_Y \log \frac{\hat{\rho}_k^m (y_{t+1} \mid F_t)}{\hat{\rho}_k^{m*} (y_{t+1} \mid F_t)} d\mu_Y (y_0, \ldots, y_t).
\]

The difference between the predictive distributions, \( \hat{\rho}_k^m (y_{t+1} \mid F_t) \) and \( \hat{\rho}_k^{m*} (y_{t+1} \mid F_t) \), is the posterior parameter, \( \pi (\theta'_t \mid F_t) \). The two conditional probabilities, \( \pi^m (\theta'_t \mid F_t) \) and \( \pi^{m*} (\theta'_t \mid F_t) \), can be obtained via the generalized Bayes formula for filtering \cite{Liptser/Shiryaev2001}. Therefore, the posterior distribution is,

\[
\pi^\rho (\theta'_t \mid F_{t_k}) = \frac{ \int_{\mathbb{R}^n} \beta_n (\theta', y) d\mu_\theta (\theta'_0, \ldots, \theta'_{t_k-1}) }{ \int_{\mathbb{R}^n} \beta_n (\theta', y) d\mu_\theta (\theta'_0, \ldots, \theta'_{t_k}) } d\mu_Y (y_1, \ldots, y_t).
\]

Here, \( \beta_n (\theta', y) \) is defined as

\[
\beta_n (\theta', y) = \exp \left\{ \sum_{k=1}^n \langle \theta'_{t_k}, \hat{y}'_{t_k} \rangle y_{t_k} - \frac{1}{2} \sum_{k=1}^n \langle \theta'_{t_k}, \hat{y}'_{t_k} \rangle^2 \right\}
= \prod_{k=1}^n \exp \left\{ \langle \theta'_{t_k}, \hat{y}'_{t_k} \rangle y_{t_k} - \frac{1}{2} \langle \theta'_{t_k}, \hat{y}'_{t_k} \rangle^2 \right\},
\]

where each \( \mu_\theta, \mu_Y \) is a cylindrical measure of processes, \( \{ \theta'_t \} \) and \( \{ y_t \} \), at each \( t = t_k \), and the superscript, \( \rho \), in \( \mu_\theta^\rho \) specifies the distribution of \( \theta'_0 \) as \( \rho (\cdot) \).

To evaluate the difference in KL risk, we define the predictive distribution, \( \hat{\rho}_k^\rho (y_{t+1} \mid F_t, \{ \hat{y}_s \}_{s=1}^{t+1}) \),
for each process, \( \{ \hat{y}_t \} \), the path-wise predictive distribution– as
\[
\hat{p}_t^\rho \left( y_{t+1} \mid F_t, \{ \hat{y}_s \}_{s=1}^{t+1} \right) = \int_{\Omega_{t+1}} \int_{\Omega_{t+1}} \hat{p}_t \left( y_{t+1} \mid \theta_{t+1}', \theta_1', \{ \hat{y}_s \}_{s=1}^{t+1} \right) \pi (\theta_{t+1}' | \theta_1') \pi^\rho (\theta_{t+1} | F_t, \{ \hat{y}_t \}) \, d\theta_1' d\theta_{t+1}'.
\]

Here, \( \pi^\rho (\theta_{t+1} | F_t, \{ \hat{y}_t \}) \) is
\[
\pi^\rho (\theta_{t+1} | F_t, \{ \hat{y}_t \}) = \int_{(\mathbb{R}^n)^{a+1}} \beta_n (\theta', y) \, d\mu_\rho (\theta_0', \ldots, \theta_{t-1}')
\]
which is the path-wise posterior distribution for each process, \( \{ \hat{y}_t \} \), i.e., for each \( t \) with fixed \( \hat{y}_t \). Then, \( \hat{p}_t^\rho \left( y_{t+1} \mid F_t, \{ \hat{y}_s \}_{s=1}^{t+1} \right) \) can be expressed as the following.

**Proposition 3.1.** Consider the initial prior, \( \rho (\theta_0) = \mathcal{N} (0, W_0) \). The path-wise predictive distribution (for each process, \( \{ \hat{y}_t \} \)), \( \hat{p}_t^\rho \left( y_{t+1} \mid F_t, \{ \hat{y}_s \}_{s=1}^{t+1} \right) \), can be expressed as,
\[
\hat{p}_t^\rho \left( y_{t+1} \mid F_t, \{ \hat{y}_s \}_{s=1}^{t+1} \right) = \hat{p}_t^m \left( y_{t+1} \mid F_t, \{ \hat{y}_s \}_{s=1}^{t+1} \right) \times \exp \left\{ - (\hat{\vartheta}_1)' (\hat{\Sigma}_0)^{-1} (\hat{\vartheta}_1) \right\} \sqrt{\left( 2\pi | \hat{\Sigma}_0 | \right)^d \left( 2\pi | \hat{\Omega}_1 | \right)^d},
\]
where \( \hat{p}_t^m \left( y_{t+1} \mid F_t, \{ \hat{y}_s \}_{s=1}^{t+1} \right) \) is the predictive distribution when the initial prior is a Lebesgue measure, and
\[
\hat{\Sigma}_0 = (\hat{\Omega}_1^{-1} + W_0^{-1})^{-1}, \quad \hat{\Omega}_0 = (\hat{\Omega}_1^{-1} + W_0^{-1})^{-1}, \quad \hat{\vartheta}_1 = \left( \hat{y}'_1 V_1^{-1} \hat{y}_1' + \hat{\Omega}_2^{-1} \right)^{-1} \left( \hat{y}'_1 V_1^{-1} \hat{y}_1' \hat{\vartheta}_1 + \hat{\Omega}_2^{-1} \hat{\vartheta}_2 \right),
\]
\[
\hat{\Omega}_0 = \hat{\Omega}_1 + W_0, \quad \hat{\vartheta}_1 = \left( \hat{y}'_1 V_1^{-1} \hat{y}_1' + \hat{\Omega}_2^{-1} \right)^{-1} \left( \hat{y}'_1 V_1^{-1} \hat{y}_1' \hat{\vartheta}_1 + \hat{\Omega}_2^{-1} \hat{\vartheta}_2 \right).
\]

See Appendix B for the details of the iterations.

**Proof.** See Appendix C. \( \square \)

From this, we immediately have the following corollary and proposition.

**Corollary 3.1.** Consider the initial prior, \( \rho_k (\theta_0) = \mathcal{N} (0, \sigma_k^2 I) \), where the variance parameter, \( \sigma_k \), is \( \sigma_k \to \infty \) as \( k \to \infty \). For each process \( \{ \hat{y}_t \} \), the predictive distribution, \( \hat{p}_t^{\rho_k} \left( y_{t+1} \mid F_t, \{ \hat{y}_s \}_{s=1}^{t+1} \right) \), pointwise converges to the predictive distribution, \( \hat{p}_t^m \left( y_{t+1} \mid F_t, \{ \hat{y}_s \}_{s=1}^{t+1} \right) \):
\[
\hat{p}_t^{\rho_k} \left( y_{t+1} \mid F_t, \{ \hat{y}_s \}_{s=1}^{t+1} \right) \to \hat{p}_t^m \left( y_{t+1} \mid F_t, \{ \hat{y}_s \}_{s=1}^{t+1} \right) \quad (k \to \infty).
\]
Proposition 3.2. Assume, for all processes, \( \{ \hat{y}_t \} \), some functional, \( \varphi (\{ \hat{y}_t \}) \), exists and satisfies,
\[
\max \left( \hat{p}_t^{\rho_k} (y_{t+1} \mid \mathcal{F}_t, \{ \hat{y}_s \}_{s=1}^{t+1}), \hat{p}_t^m (y_{t+1} \mid \mathcal{F}_t, \{ \hat{y}_s \}_{s=1}^{t+1}) \right) < \varphi (\{ \hat{y}_t \}) < \infty.
\]
For the initial prior, \( \rho_k (\theta'_0) = N (0, \sigma_k^2 I) \), with variance parameter, \( \sigma_k \), where \( \sigma_k \to \infty \) as \( k \to \infty \), the following holds:
\[
\int_{\hat{Y}} \hat{p}_t^{\rho_k} (y_{t+1} \mid \mathcal{F}_t, \{ \hat{y}_s \}_{s=1}^{t+1}) \, d\mu_{\hat{Y}} \to \int_{\hat{Y}} \hat{p}_t^m (y_{t+1} \mid \mathcal{F}_t, \{ \hat{y}_s \}_{s=1}^{t+1}) \, d\mu_{\hat{Y}}, \quad (k \to \infty).
\]
Proof. From the Lebesgue dominant convergence theorem,
\[
\lim_{k \to \infty} \int_{\hat{Y}} \hat{p}_t^{\rho_k} (y_{t+1} \mid \mathcal{F}_t, \{ \hat{y}_s \}_{s=1}^{t+1}) \, d\mu_{\hat{Y}} \to \int_{\hat{Y}} \hat{p}_t^m (y_{t+1} \mid \mathcal{F}_t, \{ \hat{y}_s \}_{s=1}^{t+1}) \, d\mu_{\hat{Y}}, \quad (k \to \infty),
\]
and Corollary 3.1, the proposition holds. \( \square \)

From this, we see that the difference in Bayes risk is
\[
B (\rho_k, \hat{p}_t^{\rho_k}) \to B (\rho_k, \hat{p}_t^m), \quad (k \to \infty),
\]
thus \( \hat{p}_t^m (y_{t+1} \mid \mathcal{F}_t) \) is extended Bayes.

3.5 Discussion

Since the predictive distribution of BPS, \( \hat{p}_t^m \), is risk constant (section 3.3) and extended Bayes (section 3.4), from Liang and Barron (2004) Theorem 1., it is exact minimax with regard to the KL risk. This result can be interpreted as follows: under the most unfavorable prior distribution, the Bayes decision function is a minimax estimator. On the other hand, if the state stochastic process, \( \{ \theta'_t \} \), is stationary, even if the initial prior distribution is a Lebesgue measure, the updated posterior distribution converges in law to a proper probability measure, due to the individual ergodic theorem. Under this setting, the predictive distribution cannot be minimax, since it is not shift-invariant. Therefore, for minimaxity, the state stochastic process, \( \{ \theta'_t \} \), must not be stationary (e.g. a random walk).

Finally, the linear combination of forecasts (e.g. equal weight averaging, Bayesian model averaging, etc.) cannot produce predictive distributions that are exact minimax. This is because 
\[
\max_a \text{KL} ((a, \theta_{t+1}), \hat{p}_t) \]
can be infinite due to the drift term in the FS decomposition, \( a \), being under-parameterized for linear combination methods.
4 Simulation study

To exemplify the theoretical results, we will use a simple, yet pertinent, simulation study to compare predictive performances between dynamic BPS, equal weight averaging, BMA, and Mallows $C_p$ averaging (Hansen 2007).

4.1 Simulation set up

We construct a simulation study that captures the characteristics encountered in real empirical applications; namely dependence amongst agents and misspecification. First, the data generating process for the target, $y_t$, is generated as follows:

$$y_t = a_t + \sum_{i=1}^{4} \theta_{ti} x_{ti} + \exp(g_t/2) \nu, \quad \nu \sim N(0, 1),$$

$$g_t = g_{t-1} + \eta, \quad \eta \sim N(0, \sigma^2),$$

$$a_t = a_{t-1} + \omega_a, \quad \omega_a \sim N(0, \sigma^2),$$

$$\theta_t = \theta_{t-1} + \omega_{\theta}, \quad \omega_{\theta} \sim N(0, \sigma^2),$$

where the time varying parameters follow a random walk and the observation error has stochastic volatility. We initialize $\theta_0 = 1$ and $a_0 = 0$ and discard the first 50 samples in order to allow random starting points, and initialize $\exp(g_0/2) = 0.1$. The covariates $x_{1:4}$ are generated as i.i.d. samples from $N(0, \sigma^2)$. For the simulation, $\sigma = 0.01$.

At all points, only $\{y, x_{1:3}\}$ are observable by the agents and $x_4$ is omitted, making all models misspecified. We construct two agents, $\{A_1, A_2\}$, with each only observing either $\{x_1, x_3\}$ or $\{x_2, x_3\}$, thus allowing for dependencies. Both agents forecast $y_t$ using a standard conjugate random walk DLM (Section 4, West and Harrison 1997). Prior specifications for the DLM state vector and discount volatility model in each of the two agent models is based on $\theta_0|v_0 \sim N(0, (v_0/s_0)I)$ and $1/v_0 \sim G(n_0/2, n_0s_0/2)$ with $n_0 = 2, s_0 = 0.01$. Discount factors are set to $(\beta, \delta) = (0.99, 0.95)$.

The BPS synthesis function follows eq. (4), with priors $a|v_0 \sim N(0, v_0/s_0)$, $\theta_0|v_0 \sim N(m_0, (v_0/s_0)^2)$ with $m_0 = (0, 1'/J)'$ and $1/v_0 \sim G(n_0/2, n_0s_0/2)$ with $n_0 = 10, s_0 = 0.002$. The discount factors are set to $(\beta, \delta) = (0.95, 0.99)$. Priors and discount factors are identical to McAlinn and West (2019).

For the study, we generate 350 samples from the data generating process. The first 50 are used to sequentially estimate the agent models, with the latter 25 of the 50 are used to simultaneously calibrate the BPS synthesis function. Forecasts are done sequentially over the remaining 300 samples, where agent models and BPS are recalibrated for each $t$ after observing new data and forecasts. We compare the mean squared forecast error (MSFE) and the log predictive density ratio.
Table 1: Predictive evaluation using mean squared forecast error (MSFE) and log predictive density ratio (LPDR) for equal weight averaging (EW), Bayesian model averaging (BMA), Mallows averaging ($C_p$), and Bayesian predictive synthesis (BPS), averaged over the 100 simulations. MSFE is evaluated at $t = \{100, 200, 300\}$. Predictive comparisons (%) were done by calculating $\text{MSFE}_{BPS}/\text{MSFE}_*$, where * denotes the method compared against.

|        | EW   | BMA  | $C_p$ | BPS  |
|--------|------|------|-------|------|
| t = 100 | MSFE$_{1:t}$ | 0.00296 | 0.00294 | 0.00298 | 0.00131 |
|        | (%)  | 44.13 | 44.53 | 43.94 | – |
|        | LPDR$_{1:t}$ | -16.92 | -16.21 | -16.96 | – |
| t = 200 | MSFE$_{1:t}$ | 0.00333 | 0.00331 | 0.00335 | 0.00139 |
|        | (%)  | 41.70 | 41.91 | 41.47 | – |
|        | LPDR$_{1:t}$ | -39.11 | -37.66 | -39.17 | – |
| t = 300 | MSFE$_{1:t}$ | 0.00365 | 0.00364 | 0.00368 | 0.00139 |
|        | (%)  | 38.18 | 38.26 | 37.89 | – |
|        | LPDR$_{1:t}$ | -69.53 | -67.30 | -69.32 | – |

(LPDR), evaluated at $t = \{100, 200, 300\}$. The log predictive density ratios (LPDR) for each $t$ is

$$\text{LPDR}_{1:t} = \sum_{i=1:t} \log \left\{ \frac{p_*(y_{t+1}|y_{1:t})}{p_{BPS}(y_{t+1}|y_{1:t})} \right\}$$

where $p_*(y_{t+1}|y_{1:t})$ is the predictive density of the model being compared with. Comparing both the MSFE and LPDR provides a more holistic assessment of the predictive performance. We repeat the simulation 100 times and report the average.

4.2 Simulation results

Comparing MSFE and LPDR (Table 1), we find significant improvements in favor of BPS, with BPS cutting down the MSFE by at least half compared to the competing strategies. Additionally, the improvements, which is calculated as $\text{MSFE}_{BPS}/\text{MSFE}_*$, where * denotes the method compared against, increase with $t$. The results are consistent with the LPDR, also improving as $t$ increases.

Looking at the other ensemble strategies, we find all to be roughly the same, though BMA edges out the rest in terms of MSFE, and Mallows $C_p$ edges out the rest in terms of LPDR. The order is consistent throughout $t$. This result is somewhat surprising, as Mallows $C_p$, at least theoretically, should give the optimal weights in terms of MSFE. However, this optimality is assumed under i.i.d. conditions, which does not hold for our analysis (and many applications).

The simulation results confirm and reinforce the results from the theoretical analysis in Section 3, dynamic BPS improves forecasts over linear combination. This can be seen holistically in the improved predictive performance in the simulation study. Empirical studies outside of this paper further reinforce this conclusion, with comparisons mirroring the simulation results.
References

Aastveit, K. A., Gerdrup, K. R., Jore, A. S., Thorsrud, L. A., 2014. Nowcasting GDP in real time: A density combination approach. Journal of Business & Economic Statistics 32, 48–68.

Aastveit, K. A., Mitchell, J., Ravazzolo, F., van Dijk, H. K., 2018a. The evolution of forecast density combinations in economics. Tech. rep., Tinbergen Institute Discussion Paper.

Aastveit, K. A., Ravazzolo, F., Van Dijk, H. K., 2018b. Combined density nowcasting in an uncertain economic environment. Journal of Business & Economic Statistics 36 (1), 131–145.

Bates, J. M., Granger, C. W. J., 1969. The combination of forecasts. Operational Research Quarterly 20, 451–468.

Berger, J. O., 1985. Statistical decision theory and Bayesian analysis. Springer Science & Business Media.

Bianchi, D., McAlinn, K., 2018. Divide and conquer: Financial ratios and industry returns predictability. Available at SSRN 3136368.

Billio, M., Casarin, R., Ravazzolo, F., van Dijk, H. K., 2012. Combination schemes for turning point predictions. Quarterly Review of Finance and Economics 52, 402–412.

Breiman, L., 1996. Bagging predictors. Machine learning 24 (2), 123–140.

Capek, J., Crespo Cuaresma, J., Hauzenberger, N., Reichel, V., 2020. Macroeconomic forecasting in the euro area using predictive combinations of dsge models.

Diebold, F. X., Shin, M., 2019. Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. International Journal of Forecasting 35 (4), 1679–1691.

Džeroski, S., Ženko, B., 2004. Is combining classifiers with stacking better than selecting the best one? Machine learning 54 (3), 255–273.

Galton, F., 1907. Vox populi.

Genest, C., Schervish, M. J., 1985. Modelling expert judgements for Bayesian updating. Annals of Statistics 13, 1198–1212.

George, E. I., Liang, F., Xu, X., 2006. Improved minimax predictive densities under kullback–leibler loss. The Annals of Statistics 34 (1), 78–91.

Geweke, J. F., Amisano, G. G., 2011. Optimal prediction pools. Journal of Econometrics 164, 130–141.
Hall, S. G., Mitchell, J., 2007. Combining density forecasts. International Journal of Forecasting 23, 1–13.

Hansen, B. E., 2007. Least squares model averaging. Econometrica 75 (4), 1175–1189.

Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T., 1999. Bayesian model averaging: a tutorial. Statistical science, 382–401.

Kapetanios, G., Mitchell, J., Price, S., Fawcett, N., 2015. Generalised density forecast combinations. Journal of Econometrics 188, 150–165.

Liang, F., Barron, A., 2004. Exact minimax strategies for predictive density estimation, data compression, and model selection. IEEE Transactions on Information Theory 50 (11), 2708–2726.

Liptser, R. S., Shiryaev, A. N., 2001. Statistics of random processes: I. General theory. Vol. 1. Springer Science & Business Media.

Madigan, D., Raftery, A. E., 1994. Model selection and accounting for model uncertainty in graphical models using occam's window. Journal of the American Statistical Association 89 (428), 1535–1546.

McAlinn, K., 2021. Mixed frequency bayesian predictive synthesis for economic nowcasting. Journal of the Royal Statistical Society. Series C (Applied Statistics) forthcoming.

McAlinn, K., Aastveit, K. A., Nakajima, J., West, M., 2020. Multivariate bayesian predictive synthesis in macroeconomic forecasting. Journal of the American Statistical Association 115 (531), 1092–1110.

McAlinn, K., West, M., 2019. Dynamic bayesian predictive synthesis in time series forecasting. Journal of Econometrics 210 (1), 155–169.

Negro, M. D., Hasegawa, R. B., Schorfheide, F., 2016. Dynamic prediction pools: an investigation of financial frictions and forecasting performance. Journal of Econometrics 192(2), 391–405.

Pettenuzzo, D., Ravazzolo, F., 2016. Optimal portfolio choice under decision-based model combinations. Journal of Applied Econometrics 31 (7), 1312–1332.

Prado, R., West, M., 2010. Time Series: Modelling, Computation & Inference. Chapman & Hall/CRC Press.

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

Schapire, R. E., 2003. The boosting approach to machine learning: An overview. In: Nonlinear estimation and classification. Springer, pp. 149–171.
Schweizer, M., 1995. On the minimal martingale measure and the möllmer-schweizer decomposition. Stochastic analysis and applications 13 (5), 573–599.

West, M., 1992. Modelling agent forecast distributions. Journal of the Royal Statistical Society (Series B: Methodological) 54, 553–567.

West, M., Crosse, J., 1992. Modelling of probabilistic agent opinion. Journal of the Royal Statistical Society (Series B: Methodological) 54, 285–299.

West, M., Harrison, P. J., 1997. Bayesian Forecasting & Dynamic Models, 2nd Edition. Springer Verlag.

Williams, D., 1991. Probability with martingales. Cambridge university press.

Yao, Y., Vehtari, A., Simpson, D., Gelman, A., et al., 2018. Using stacking to average bayesian predictive distributions (with discussion). Bayesian Analysis 13 (3), 917–1003.
Supplementary material for
Predictions with dynamic Bayesian predictive synthesis are exact minimax

A Proof of Lemma 3.1

The posterior distribution, $\pi(\theta_t' | F_t)$, for $\theta_t'$ is updated as

$$\pi(\theta_t' | F_t) = \frac{q_t(y_t | \theta_t', \hat{y}_t') \pi(\theta_t' | F_{t-1})}{\int q_t(y_t | \theta_t', \hat{y}_t') \pi(\theta_t' | F_{t-1}) \, d\theta_t'},$$

via the Bayes theorem, and the predictive distribution for $\theta_t'$ (the prior distribution, $\pi(\theta_t' | F_{t-1})$, at $t$) is given as

$$\pi(\theta_t' | F_{t-1}) = \int \pi(\theta_t' | \theta_{t-1}') \pi(\theta_{t-1}' | F_{t-1}) \, d\theta_{t-1}'.$$

Since we assume that $\theta_t'$ follows a random walk process (eq. 9b), the transition probability is $\pi(\theta_t' | \theta_{t-1}') = N(\theta_{t-1}', W_t)$ and $\theta_t'$ is invariant with regard to the shift transformation, $\theta_t' \rightarrow \tilde{\theta}_t'$:

$$\pi(\theta_t' | \theta_{t-1}') = \pi(\tilde{\theta}_t' | \theta_{t-1}').$$

Further, since we assume that the prior distribution of $\theta_t'$, at $t = 0$, is a Lebesgue measure, $m(\cdot)$, from the shift invariance of Lebesgue measures,

$$m(\theta_{0,0}) = m(\tilde{\theta}_{0,0}), \quad m(\theta_0) = m(\tilde{\theta}_0),$$

and iterating through the Bayes theorem, we have

$$\pi(\theta_t' | F_t) = \pi(\tilde{\theta}_t' | F_t).$$

Therefore,

$$q_t^m(y_{t+1} | y_t) = q_t^m(y_{t+1} | \hat{y}_t),$$

and thus the KL risk is constant for all $(a^*, \theta_t')$:

$$R_{KL}((a^*, \theta_t'), q_t^m) = R_{KL}((0, 0), q_t^m) = c.$$
B Details for Proposition 3.1

For the equation in Proposition 3.1, we have,

\[
\tilde{\Sigma}_1 = \left( \Sigma_1^{-1} + W_1^{-1} \right)^{-1}, \quad \Sigma_1 = \left( \hat{y}_1' V_1^{-1} \hat{y}_1' + \tilde{\Omega}_2^{-1} \right)^{-1}, \\
\tilde{\Omega}_1 = \Sigma_1 + W_1, \quad \tilde{\Omega}_1 = \left( \hat{y}_1' V_1^{-1} \hat{y}_1' \right)^{-1} + \tilde{\Omega}_2, \\
\hat{\Sigma}_1 = \left( \hat{\Sigma}_1^{-1} + W_1^{-1} \right)^{-1}, \quad \hat{\Sigma}_1 = \left( \hat{y}_1' V_1^{-1} \hat{y}_1' + \tilde{\Omega}_2^{-1} \right)^{-1}, \\
\hat{\Omega}_1 = \hat{\Sigma}_1 + W_1, \quad \hat{\Omega}_1 = \left( \hat{y}_1' V_1^{-1} \hat{y}_1' \right)^{-1} + \tilde{\Omega}_2.
\]

The iteration of each matrix for \( k = 1, \ldots, n \) is

\[
\tilde{\Sigma}_k = \left( \Sigma_k^{-1} + W_k^{-1} \right)^{-1}, \quad \Sigma_k = \left( \hat{y}_k' V_k^{-1} \hat{y}_k' + \tilde{\Omega}_{k+1}^{-1} \right)^{-1}, \\
\tilde{\Omega}_k = \Sigma_k + W_k, \quad \tilde{\Omega}_k = \left( \hat{y}_k' V_k^{-1} \hat{y}_k' \right)^{-1} + \tilde{\Omega}_{k+1}, \\
\hat{\Sigma}_n+1 = \left( \hat{y}_{n+1}' V_{n+1}^{-1} \hat{y}_{n+1}' + W_{n+1}^{-1} \right)^{-1}, \\
\hat{\Omega}_n+1 = \left( \hat{y}_{n+1}' V_{n+1}^{-1} \hat{y}_{n+1}' \right)^{-1} + W_{n+1},
\]

and for \( k = 1, \ldots, n - 1 \) is

\[
\tilde{\Sigma}_k = \left( \Sigma_k^{-1} + W_k^{-1} \right)^{-1}, \quad \hat{\Sigma}_k = \left( \hat{y}_k' V_k^{-1} \hat{y}_k' + \tilde{\Omega}_{k+1}^{-1} \right)^{-1}, \\
\tilde{\Omega}_k = \Sigma_k + W_k, \quad \hat{\Omega}_k = \left( \hat{y}_k' V_k^{-1} \hat{y}_k' \right)^{-1} + \tilde{\Omega}_{k+1}, \\
\Sigma_n = \left( \hat{y}_n' V_n^{-1} \hat{y}_n' + W_n^{-1} \right)^{-1}, \\
\tilde{\Omega}_n = \left( \hat{y}_n' V_n^{-1} \hat{y}_n' \right)^{-1} + W_n.
\]

C Proof of Proposition 3.1

For the random walk DLM,

\[
y_t = \hat{y}_t' \theta_t' + v_t, v_t \sim N(0, V_t), \\
\theta_t' = \theta_{t-1}' + w_t, w_t \sim N(0, W_t),
\]

19
the conditional density function for the observation and state equation is given as

\[
y_t | \theta'_t \sim N \left( \hat{y}_t^\top \theta'_t, V_t \right) = \frac{1}{\sqrt{2\pi |V_t|}} \exp \left\{ -\left( y_t - \hat{y}_t^\top \theta'_t \right)^\top V_t^{-1} \left( y_t - \hat{y}_t^\top \theta'_t \right) \right\},
\]

\[
\theta'_t | \theta'_{t-1} \sim N \left( \theta'_{t-1}, W_t \right) = \frac{1}{\sqrt{2\pi |W_t|}} \exp \left\{ -\left( \theta'_t - \theta'_{t-1} \right)^\top W_t^{-1} \left( \theta'_t - \theta'_{t-1} \right) \right\}.
\]

If we set \( \hat{\theta}_n = \left( \hat{y}_t^\top \hat{y}_t \right)^{-1} \hat{y}_t y_n \) the conditional density function of the observation equation can be transformed to

\[
\frac{1}{\sqrt{2\pi |V_t|}} \exp \left\{ -\left( y_t - \hat{y}_t^\top \theta'_t \right)^\top V_t^{-1} \left( y_t - \hat{y}_t^\top \theta'_t \right) \right\}
\]

\[
= \frac{1}{\sqrt{2\pi |W_t|}} \exp \left\{ -\left( y_t - \hat{y}_t^\top \hat{\theta}_t \right)^\top W_t^{-1} \left( y_t - \hat{y}_t^\top \hat{\theta}_t \right) \right\}
\]

\[
\times \exp \left\{ -\left( \hat{\theta}_t - \theta'_t \right)^\top \hat{y}_t V_t^{-1} \hat{y}_t^\top \left( \hat{\theta}_t - \theta'_t \right) \right\}.
\]

The predictive distribution, \( q'_t \left( y_{t+1} | \mathcal{F}_t, \{ \hat{y}_s \}_{s=1}^{t+1} \right) \), is given as,

\[
q'_t \left( y_{t+1} | \mathcal{F}_t, \{ \hat{y}_s \}_{s=1}^{t+1} \right) = \frac{\int_{(\mathbb{R}^j)^{t+1}} \beta_t+1 \left( \theta', y \right) d\mu_0^{(t)} \left( \theta'_0, \cdots, \theta'_{t+1} \right)}{\int_{(\mathbb{R}^j)^{t}} \beta_t \left( \theta', y \right) d\mu_0^{(t)} \left( \theta'_0, \cdots, \theta'_t \right)}.
\] (12)

Here,

\[
\beta_t \left( \theta', y \right) = \exp \left\{ \sum_{k=1}^{t} \langle \theta'_k, \hat{y}'_k \rangle y_k - \frac{1}{2} \sum_{k=1}^{n} (\langle \theta'_k, \hat{y}'_k \rangle)^2 \right\}
\]

\[
= \prod_{k=1}^{n} \exp \left\{ \langle \theta'_k, \hat{y}'_k \rangle y_k - \frac{1}{2} (\theta'_k, \hat{y}'_k)^2 \right\}.
\]

We first rewrite the generalized Bayes formula to the sequential formula. Given,

\[
g_t \left( y_t | \theta'_t \right) = N \left( \hat{y}_t^\top \theta'_t, V_t \right),
\]

\[
P_t \left( \theta'_t | \theta'_{t-1} \right) = N \left( \theta'_{t-1}, W_t \right),
\]
the numerator of the left hand side of eq. (12) can be written as

$$
\int \left[ \cdots \int [g_{t+1} (y_{t+1} | \theta'_{t+1}) \, P_{t+1} (\theta'_{t+1} | \theta'_t) \, d\theta'_{t+1}] 
\times g_t (y_t | \theta'_t) \, P_t (\theta'_t | \theta'_{t-1}) \, d\theta'_t] 
\times g_{t-1} (y_{t-1} | \theta'_{t-1}) \, P_{t+1} (\theta'_{t-1} | \theta'_{t-2}) \, d\theta'_{t-1}] 
\vdots 
\times g_1 (y_1 | \theta'_1) \, P_t (\theta'_1 | \theta'_0) \, d\theta'_1 
\times \rho (\theta'_0) \, d\theta'_0 \right].
$$

From the law of iteration of conditional expectations, we integrate from \( t + 1 \) backwards to sequentially remove the conditionals. Note, however, that we do not integrate over quantities that are not necessary for the proof.

First, for \( \int g_{t+1} (y_{t+1} | \theta'_{t+1}) \, P_{t+1} (\theta'_{t+1} | \theta'_t) \, d\theta'_{t+1} \), we have,

$$
g_{t+1} (y_{t+1} | \theta'_{t+1}) \, P_{t+1} (\theta'_{t+1} | \theta'_t) 
= \frac{1}{\sqrt{2 \pi | W_{t+1} |}} \exp \left\{ - \left( y_{t+1} - \hat{y}_{t+1}^T \hat{\theta}_{t+1} \right)^T V_{t+1}^{-1} \left( y_{t+1} - \hat{y}_{t+1}^T \hat{\theta}_{t+1} \right) \right\} 
\times \exp \left\{ - \left( \theta'_{t+1} - \theta'_t \right)^T \hat{y}_{t+1} V_{t+1}^{-1} \hat{y}_{t+1}^T \left( \hat{\theta}_{t+1} - \theta'_t \right) \right\} 
\times \frac{1}{\sqrt{2 \pi | W_{t+1} |}} \exp \left\{ - \left( \theta'_{t+1} - \theta'_t \right)^T W_{t+1}^{-1} \left( \theta'_{t+1} - \theta'_t \right) \right\}.
$$

Completing the square inside the exponent, we have

$$
\left( \theta_{t+1} - \theta'_{t+1} \right)^T \hat{y}_{t+1} V_{t+1}^{-1} \hat{y}_{t+1}^T \left( \hat{\theta}_{t+1} - \theta'_t \right) + \left( \theta'_{t+1} - \theta'_t \right)^T W_{t+1}^{-1} \left( \theta'_{t+1} - \theta'_t \right) 
= \left( \theta'_{t+1} - \theta_{t+1} \right)^T \left( \Sigma_{t+1} \right)^{-1} \left( \theta'_{t+1} - \theta_{t+1} \right) + \left( \theta_{t+1} - \theta'_t \right)^T \left( \hat{\Sigma}_{t+1} \right)^{-1} \left( \theta_{t+1} - \theta'_t \right),
$$

where,

$$
\hat{\theta}_{t+1} = \left( \hat{y}_{t+1} V_{t+1}^{-1} \hat{y}_{t+1}^T \right)^{-1} \left( \hat{y}_{t+1} V_{t+1}^{-1} \hat{y}_{t+1}^T \hat{\theta}_{t+1} + W_{t+1}^{-1} \theta'_t \right), \\
\hat{\Sigma}_{t+1} = \left( \hat{y}_{t+1} V_{t+1}^{-1} \hat{y}_{t+1}^T \right)^{-1}, \\
\hat{\Omega}_{t+1} = \left( \hat{y}_{t+1} V_{t+1}^{-1} \hat{y}_{t+1}^T \right)^{-1} + W_{t+1}.
$$

Since \( \exp \left\{ - \left( \theta_{t+1} - \theta'_t \right)^T \left( \hat{\Omega}_{t+1} \right)^{-1} \left( \theta_{t+1} - \theta'_t \right) \right\} \) does not affect the integration of \( d\theta'_{t+1} \), we move it one step to the past.

21
Next, we integrate the portion concerning $d\theta'_t$:

$$
\int \exp \left\{ - \left( \theta_{t+1} - \theta'_t \right)^\top \left( \tilde\Omega_{t+1} \right)^{-1} \left( \hat\theta_{t+1} - \theta'_t \right) \right\} g_t(y_t | \theta'_t) \ P_t(\theta'_t | \theta'_{t-1})
\ d\theta'_t.
$$

The equation inside the exponent is

$$
\left( \hat\theta_{t+1} - \theta'_t \right)^\top \left( \tilde\Omega_{t+1} \right)^{-1} \left( \hat\theta_{t+1} - \theta'_t \right) + \left( y_t - \hat y_t^\top \hat\theta_t \right)^\top V_t^{-1} \left( y_t - \hat y_t^\top \hat\theta_t \right)
\ + \left( \theta_t - \theta'_t \right)^\top \hat y_t V_t^{-1} \hat y_t^\top \left( \theta_t - \theta'_t \right)
\ + \left( \theta'_t - \theta'_{t-1} \right)^\top W_t^{-1} \left( \theta'_t - \theta'_{t-1} \right).
$$

Then, we have,

$$
\left( \theta_{t+1} - \theta'_t \right)^\top \left( \tilde\Omega_{t+1} \right)^{-1} \left( \hat\theta_{t+1} - \theta'_t \right) + \left( \theta_t - \theta'_t \right)^\top \hat y_t V_t^{-1} \hat y_t^\top \left( \theta_t - \theta'_t \right)
\ = \left( \theta'_t - \tilde\theta_t \right)^\top \left( \tilde\Sigma_t \right)^{-1} \left( \theta'_t - \tilde\theta_t \right) + \left( \theta_{t+1} - \tilde\theta_t \right)^\top \left( \tilde\Omega_t \right)^{-1} \left( \hat\theta_{t+1} - \tilde\theta_t \right),
$$

where,

$$
\tilde\theta_t = \left( \hat y_t V_t^{-1} \hat y_t^\top + \hat\Omega_{t+1}^{-1} \right)^{-1} \left( \hat y_t V_t^{-1} \hat y_t^\top \hat\theta_t + \hat\Omega_{t+1}^{-1} \hat\theta_{t+1} \right),
$$

$$
\tilde\Sigma_t = \left( \hat y_t V_t^{-1} \hat y_t^\top + \hat\Omega_{t+1}^{-1} \right)^{-1},
$$

$$
\tilde\Omega_t = \left( \hat y_t V_t^{-1} \hat y_t^\top \right)^{-1} + \hat\Omega_{t+1}^{-1}.
$$

Next, we have,

$$
\left( \theta'_t - \tilde\theta_t \right)^\top \left( \tilde\Sigma_t \right)^{-1} \left( \theta'_t - \tilde\theta_t \right) + \left( \theta'_t - \theta'_{t-1} \right)^\top W_t^{-1} \left( \theta'_t - \theta'_{t-1} \right)
\ = \left( \theta'_t - \hat\theta_t \right)^\top \left( \hat\Sigma_t \right)^{-1} \left( \theta'_t - \hat\theta_t \right) + \left( \theta_t - \theta'_{t-1} \right)^\top \left( \hat\Omega_t \right)^{-1} \left( \hat\theta_t - \theta'_{t-1} \right),
$$

where,

$$
\hat\theta_t = \left( \hat\Sigma_t^{-1} + W_t^{-1} \right)^{-1} \left( \hat\Sigma_t^{-1} \hat\theta_t + W_t^{-1} \theta'_{t-1} \right),
$$

$$
\hat\Sigma_t = \left( \hat\Sigma_t^{-1} + W_t^{-1} \right)^{-1},
$$

$$
\hat\Omega_t = \hat\Sigma_t + W_t.
$$

Since $\exp \left\{ - \left( \tilde\theta_t - \theta'_{t-1} \right)^\top \left( \tilde\Omega_t \right)^{-1} \left( \tilde\theta_t - \theta'_{t-1} \right) \right\}$ does not affect the integration of $d\theta'_t$, we move it one step to the past.
Likewise, we sequentially integrate backwards. Then, for the integration of \(d\theta_t\), we have,

\[
\int \exp \left\{ - (\bar{\theta}_1 - \theta_0')^\top \left( \hat{\Omega}_1 \right)^{-1} (\bar{\theta}_1 - \theta_0') \right\} \rho(\theta_0') \, d\theta_0'.
\]

If we take a Lebesgue measure for \(\rho\), this integrated value is \(\sqrt{(2\pi|\hat{\Omega}_1|)^J}\). If \(\rho = N(0, W_0)\), then inside the exponent is

\[
(\theta_1 - \theta_0')^\top \left( \hat{\Omega}_1 \right)^{-1} (\theta_1 - \theta_0') + (\theta_0' - \bar{\theta}_0) + (\bar{\theta}_1)^\top \left( \hat{\Omega}_1 \right)^{-1} (\bar{\theta}_1),
\]

where,

\begin{align*}
\bar{\theta}_0 &= \left( \Omega_1^{-1} + W_0^{-1} \right)^{-1} \hat{\Omega}_1^{-1} \bar{\theta}_1, \\
\hat{\Sigma}_0 &= \left( \Omega_1^{-1} + W_0^{-1} \right)^{-1}, \\
\hat{\Omega}_0 &= \hat{\Omega}_1 + W_0.
\end{align*}

Therefore, the integrated value is

\[
\exp \left\{ - (\bar{\theta}_1)^\top \left( \hat{\Omega}_0 \right)^{-1} (\bar{\theta}_1) \right\} \sqrt{\left(2\pi \left| \hat{\Sigma}_0 \right| \right)^J}.
\]

From this, the numerator of eq. (12) can be expressed as

\[
\int_{(\mathbb{R}^J)^{t+1}} \beta_{t+1}(\theta', y) \, d\mu_{t}^\rho(\theta'_0, \cdots, \theta'_{t+1})^{(10)}
\]

\[
= \int_{(\mathbb{R}^J)^{t+1}} \beta_{t+1}(\theta', y) \, d\mu_{t}^\rho(\theta'_0, \cdots, \theta'_{t+1}) \times \frac{\exp \left\{ - (\bar{\theta}_1)^\top \left( \hat{\Omega}_0 \right)^{-1} (\bar{\theta}_1) \right\} \sqrt{\left(2\pi \left| \hat{\Sigma}_0 \right| \right)^J}}{\sqrt{\left(2\pi \left| \hat{\Omega}_1 \right| \right)^J}}. \tag{13}
\]

We now integrate the denominator of eq. (12). This integration is not as simple as offsetting the subscript from the results of the numerator, as the multiple integration of the denominator concerns one less variable than the integration of the numerator.

For the integration with regard to \(d\theta'_t\),

\[
\int g_t(y_t | \theta'_t) \, P_t(\theta'_t | \theta'_{t-1}) \, d\theta'_t,
\]

23
the equation inside the exponent is
\[
\begin{align*}
&\left( y_t - \hat{y}_t^T \hat{\theta}_t \right)^T V_t^{-1} \left( y_t - \hat{y}_t^T \hat{\theta}_t \right) \\
&+ \left( \hat{\theta}_t - \theta'_t \right)^T \hat{y}_t^T V_t^{-1} \hat{y}_t^T \left( \hat{\theta}_t - \theta'_t \right) + \left( \theta'_t - \theta'_{t-1} \right)^T W_t^{-1} \left( \theta'_t - \theta'_{t-1} \right) \\
&= (\theta'_t - \hat{\theta}_t)^T (\hat{\Sigma}_t)^{-1} (\theta'_t - \hat{\theta}_t) + \left( \hat{\theta}_t - \theta'_{t-1} \right)^T (\hat{\Omega}_t)^{-1} \left( \hat{\theta}_t - \theta'_{t-1} \right),
\end{align*}
\]
where,
\[
\begin{align*}
\hat{\theta}_t &= \left( \hat{y}_t^T V_t^{-1} \hat{y}_t^T + W_t^{-1} \right)^{-1} \left( \hat{y}_t^T V_t^{-1} \hat{y}_t^T \hat{\theta}_t + W_t^{-1} \theta'_{t-1} \right), \\
\hat{\Sigma}_t &= \left( \hat{y}_t^T V_t^{-1} \hat{y}_t^T + W_t^{-1} \right)^{-1}, \\
\hat{\Omega}_t &= \left( \hat{y}_t^T V_t^{-1} \hat{y}_t^T \right)^{-1} + W_t.
\end{align*}
\]

We move \( \exp \left\{ - \left( \hat{\theta}_t - \theta'_{t-1} \right)^T (\hat{\Omega}_t)^{-1} \left( \hat{\theta}_t - \theta'_{t-1} \right) \right\} \) back one step to the past.

Next, for the integration concerning \( d\theta'_{t-1} \),
\[
\int \exp \left\{ - \left( \hat{\theta}_t - \theta'_{t-1} \right)^T (\hat{\Omega}_t)^{-1} \left( \hat{\theta}_t - \theta'_{t-1} \right) \right\} \, \gamma_t(y_t | \theta'_{t-1}) \, P_{t-1}(\theta'_{t-1} | \theta'_{t-2}) \, d\theta'_{t-1},
\]
the equation inside the exponent is
\[
\begin{align*}
&\left( \hat{\theta}_t - \theta'_{t-1} \right)^T (\hat{\Omega}_t)^{-1} \left( \hat{\theta}_t - \theta'_{t-1} \right) + \left( \gamma_t(y_t | \theta'_{t-1}) \right)^T V_{t-1}^{-1} \left( \gamma_t(y_t | \theta'_{t-1}) \right) \\
&+ \left( \hat{\theta}_{t-1} - \theta'_{t-1} \right)^T \hat{y}_{t-1}^T V_{t-1}^{-1} \hat{y}_{t-1}^T \left( \hat{\theta}_{t-1} - \theta'_{t-1} \right) + \left( \theta'_{t-1} - \theta'_{t-2} \right)^T W_{t-1}^{-1} \left( \theta'_{t-1} - \theta'_{t-2} \right).
\end{align*}
\]
Then, we have,
\[
\begin{align*}
&\left( \hat{\theta}_t - \theta'_{t-1} \right)^T (\hat{\Omega}_t)^{-1} \left( \hat{\theta}_t - \theta'_{t-1} \right) + \left( \hat{\theta}_{t-1} - \theta'_{t-1} \right)^T \hat{y}_{t-1}^T V_{t-1}^{-1} \hat{y}_{t-1}^T \left( \hat{\theta}_{t-1} - \theta'_{t-1} \right) \\
&= (\theta'_{t-1} - \hat{\theta}_{t-1})^T (\hat{\Sigma}_{t-1})^{-1} (\theta'_{t-1} - \hat{\theta}_{t-1}) + \left( \hat{\theta}_{t-1} - \hat{\theta}_t \right)^T (\hat{\Omega}_t)^{-1} \left( \hat{\theta}_{t-1} - \hat{\theta}_t \right),
\end{align*}
\]
where,
\[
\begin{align*}
\hat{\theta}_{t-1} &= \left( \hat{y}_{t-1}^T V_{t-1}^{-1} \hat{y}_{t-1}^T + \hat{\Omega}_{t-1}^{-1} \right)^{-1} \left( \hat{y}_{t-1}^T V_{t-1}^{-1} \hat{y}_{t-1}^T \hat{\theta}_{t-1} + \hat{\Omega}_{t-1}^{-1} \hat{\theta}_t \right), \\
\hat{\Sigma}_{t-1} &= \left( \hat{y}_{t-1}^T V_{t-1}^{-1} \hat{y}_{t-1}^T + \hat{\Omega}_{t-1}^{-1} \right)^{-1}, \\
\hat{\Omega}_{t-1} &= \left( \hat{y}_{t-1}^T V_{t-1}^{-1} \hat{y}_{t-1}^T \right)^{-1} + \hat{\Omega}_t.
\end{align*}
\]
Next, we have,
\[
(\theta_{t-1}' - \hat{\theta}_{t-1})^\top (\hat{\Sigma}_{t-1})^{-1} (\theta_{t-1}' - \hat{\theta}_{t-1}) + (\theta_{t-1}' - \theta_{t-2}')^\top W_{t-1}^{-1} (\theta_{t-1}' - \theta_{t-2}')
\]
\[
= (\theta_{t-1}' - \hat{\theta}_{t-1})^\top (\hat{\Sigma}_{t-1})^{-1} (\theta_{t-1}' - \hat{\theta}_{t-1}) + (\hat{\theta}_{t-1} - \theta_{t-2}')^\top \hat{\Omega}_{t-1}^{-1} (\hat{\theta}_{t-1} - \theta_{t-2}'),
\]
where,
\[
\hat{\theta}_{t-1} = (\hat{\Sigma}_{t-1}^{-1} + W_{t-1}^{-1})^{-1} (\hat{\Sigma}_{t-1}^{-1} \hat{\theta}_{t-1} + W_{t-1}^{-1} \theta_{t-2}'),
\]
\[
\hat{\Sigma}_{t-1} = (\hat{\Sigma}_{t-1}^{-1} + W_{t-1}^{-1})^{-1},
\]
\[
\hat{\Omega}_{t-1} = \hat{\Sigma}_{t-1} + W_{t-1}.
\]

We move \(\exp \left\{- (\hat{\theta}_{t-1} - \theta_{t-2}')^\top \hat{\Omega}_{t-1}^{-1} (\hat{\theta}_{t-1} - \theta_{t-2}')\right\}\) back one step to the past.

Likewise, we sequentially integrate backwards. Then, for the integration of \(d\theta_0',\) we have,
\[
\int \exp \left\{- (\hat{\theta}_{1} - \theta_0')^\top (\hat{\Omega}_1)^{-1} (\hat{\theta}_{1} - \theta_0')\right\} \rho (\theta_0') d\theta_0'.
\]

If we take a Lebesgue measure for \(\rho,\) the integrated value is \(\sqrt{(2\pi |\hat{\Sigma}_1|)^J}.\) If \(\rho = N(0, W_0),\) then the inside of the exponent is
\[
(\hat{\theta}_{1} - \theta_0')^\top (\hat{\Sigma}_0)^{-1} (\hat{\theta}_{1} - \theta_0') + (\theta_0')^\top W_0^{-1} (\theta_0')
\]
\[
= (\theta_0' - \hat{\theta}_0)^\top (\Sigma_0)^{-1} (\theta_0' - \hat{\theta}_0) + (\hat{\theta}_1)^\top (\hat{\Omega}_0)^{-1} (\hat{\theta}_1),
\]
where,
\[
\hat{\theta}_0 = (\hat{\Omega}_1^{-1} + W_0^{-1})^{-1} \hat{\Sigma}_1^{-1} \hat{\theta}_1,
\]
\[
\Sigma_0 = (\Sigma_0^{-1} + W_0^{-1})^{-1},
\]
\[
\Omega_0 = \hat{\Omega}_1 + W_0.
\]

Therefore, the integrated value is
\[
\exp \left\{- (\hat{\theta}_1)^\top (\hat{\Sigma}_0)^{-1} (\hat{\theta}_1)\right\} \sqrt{(2\pi |\Sigma_0|)^J}.
\]
Therefore, the denominator of eq. (12) can be expressed as

\[ \int_{(\mathbb{R}^j)^t} \beta_t (\theta', y) \, d\mu_{\theta} (\theta'_0, \cdots, \theta'_t) \]

\[ = \int_{(\mathbb{R}^j)^t} \beta_t (\theta', y) \, d\mu_{\theta} (\theta'_0, \cdots, \theta'_{t+1}) \times \frac{\exp \left\{ -\left( \dot{\theta}_1 \right)^T (\hat{\Omega}_0)^{-1} \left( \dot{\theta}_1 \right) \right\} \sqrt{(2\pi |\Sigma_0|)^J}}{\sqrt{(2\pi |\hat{\Omega}_1|)^J}}. \]  

(14)

Finally, from eq. (13) and eq. (14), the corollary holds. Q.E.D.