ISSN 1440-771X

Department of Econometrics and Business Statistics

http://business.monash.edu/econometrics-and-business-statistics/research/publications

Loss-Based Variational Bayes Prediction

David T. Frazier, Ruben Loaiza-Maya, Gael M. Martin
and Bonsoo Koo

May 2021

Working Paper 08/21

Loss-Based Variational Bayes Prediction*

David T. Frazier†, Ruben Loaiza-Maya, Gael M. Martin and Bonsoo Koo

Department of Econometrics and Business Statistics, Monash University and Australian Centre of Excellence in Mathematics and Statistics

Abstract

We propose a new method for Bayesian prediction that caters for models with a large number of parameters and is robust to model misspecification. Given a class of high-dimensional (but parametric) predictive models, this new approach constructs a posterior predictive using a variational approximation to a loss-based, or Gibbs, posterior that is directly focused on predictive accuracy. The theoretical behavior of the new prediction approach is analyzed and a form of optimality demonstrated. Applications to both simulated and empirical data using high-dimensional Bayesian neural network and autoregressive mixture models demonstrate that the approach provides more accurate results than various alternatives, including misspecified likelihood-based predictions.

Keywords: Loss-based Bayesian forecasting; variational inference; Gibbs posteriors; proper scoring rules; Bayesian neural networks; M4 forecasting competition

MSC2010 Subject Classification: 62F15, 60G25, 62M20

JEL Classifications: C11, C53, C58.

*We would like to thank various participants at the 14th International Conference in Computational and Financial Econometrics, December 2020, and the 'ABC in Svalbard' Workshop, April 2021, for helpful comments on earlier drafts of the paper. This research has been supported by Australian Research Council (ARC) Discovery Grants DP170100729 and DP200101414. Frazier was also supported by ARC Early Career Researcher Award DE200101070.

†Corresponding author: david.frazier@monash.edu.

1 Introduction

The conventional paradigm for Bayesian prediction is underpinned by the assumption that the true data generating process is either equivalent to the predictive model adopted, or spanned by a finite set of models over which we average. Of late however, recognition of the unrealistic nature of such an assumption, allied with an increased interest in driving prediction by problem-specific measures of accuracy, or loss, have led to alternative approaches. Whilst antecedents of these new principles are found in the 'probably approximately correct' (PAC)-Bayes approach to prediction in the machine learning literature (see Guedj, 2019, for a review), it is in the statistics and econometrics literature that this 'loss-based prediction' has come to more formal maturity, including in terms of its theoretical validation. This includes Bayesian work on weighted combinations of predictions, such as, for example, Billio *et al.* (2013), Casarin *et al.* (2015), Pettenuzzo and Ravazzolo (2016), Bassetti *et al.* (2018), Batrk *et al.* (2019), McAlinn and West (2019) and McAlinn *et al.* (2020), where weights are updated via various predictive criteria, and the true model is not assumed to be one of the constituent models - i.e. an $\mathcal{M}$-open state of the world (Bernardo and Smith, 1994) is implicitly adopted. It also includes a recent contribution by Loaiza-Maya *et al.* (2020a), in which both single models and predictive mixtures are used to generate accurate Bayesian predictions in the presence of model misspecification; with both theoretical and numerical results highlighting the ability of the approach to out-perform conventional likelihood-based prediction.

The current paper contributes to this evolving literature by providing a new method for producing accurate loss-based predictions in high-dimensional problems. We begin by defining a class of flexible predictive models, conditional on a set of unknown parameters, that are a plausible mechanism for generating probabilistic predictions. A prior distribution is placed over the parameters of this predictive class, and the prior then updated to a posterior via a criterion function that captures a user-specified measure of predictive accuracy. That is, the conventional, and potentially misspecified likelihood-based update is eschewed, in favour of a function that is tailored to the predictive problem at hand; the ultimate goal being to produce accurate predictions according to the measure that matters, without requiring knowledge of the true data generating mechanism.

In the spirit of the various generalized Bayesian *inferential* methods, in which likelihood functions are also replaced by alternative updating mechanisms (*inter alia*, Zhang, 2006a, Zhang, 2006b, Jiang and Tanner, 2008, Bissiri *et al.*, 2016, Giummolè *et al.*, 2017, Knoblauch *et al.*, 2019, Miller and Dunson, 2019, Syring and Martin, 2019, and Pacchiardi and Dutta, 2021), we adopt a coherent update based on the exponential of a scaled sample loss. As in Loaiza-Maya *et al.* (2020a) the loss is, in turn, defined by a proper scoring rule (Gneiting *et al.*, 2007; Gneiting and Raftery, 2007) that rewards a given form of predictive accuracy; for example, accurate prediction of extreme values. Given the high-dimensional nature of the resultant posterior, numerical treatment via 'exact' Markov Chain Monte Carlo (MCMC) is computationally challenging; hence we adopt an 'approximate' approach using variational principles. Since the posterior that results from an exponentiated loss was first denoted as a 'Gibbs posterior' by

Zhang (2006a), we refer to the variational approximation of this posterior as the *Gibbs variational posterior*, and the (marginal) predictive distribution that results from this posterior, via the standard Bayesian calculus, as the *Gibbs variational predictive* (hereafter, GVP). With a slight abuse of terminology, and when it is clear from the context, we also use the abbreviation GVP to reference the method of Gibbs variational prediction, or loss-based variational prediction *per se.*

Two key questions are addressed in the paper: *i)* Does the Gibbs variational posterior asymptotically concentrate onto the point in the parameter space that defines the model with the highest predictive accuracy, as measured by the chosen scoring rule (i.e. the 'optimal' predictive)? *ii)* Does the use of this posterior approximation 'matter' in terms of predictive accuracy? That is, what is lost, if anything, by representing parameter uncertainty via the Gibbs variational posterior rather than by the exact Gibbs posterior?

While several authors, such as Alquier *et al.* (2016), Alquier and Ridgway (2020), and Yang *et al.* (2020), have answered *i)* in the affirmative for certain settings, existing results generally hinge on the satisfaction of a sub-Gaussian concentration inequality for the resulting risk function used to produce the Gibbs posterior, such as a Hoeffding or Bernstein-type inequality. Given that our goal is prediction in possibly non-Markovian, non-stationary, and/or complex models, and with a loss function based specifically on a general scoring rule, it is not clear that such results remain applicable. Therefore, we deviate from the earlier referenced approaches and, instead, affirm *i)* using arguments based on quadratic expansions of the loss function, and appropriate control of remainder terms. The resulting approach does not deliver so-called 'oracle' inequalities, as in the PAC-Bayes literature on Gibbs posteriors; however, it allows us to validate this methodology for a wide variety of models used to produce predictions in commonly encountered time series settings, including models with non-Markovian features.

In terms of question *ii)* above, we do not know of any existing results that rigorously compare the theoretical behavior of the GVP and the potentially infeasible exact Gibbs predictive (built from the potentially inaccessible exact Gibbs posterior). Having demonstrated concentration of the Gibbs variational posterior onto the parameter (vector) that defines the 'optimal predictive model', where optimality is with respect to a given user-specified scoring rule, we theoretically demonstrate that the GVP delivers predictions that are just as accurate as those obtained from the Gibbs predictive when measured according to the score used to construct the Gibbs posterior. We do this by proving that the GVP 'merges' (in the sense of Blackwell and Dubins, 1962) with the optimal predictive, to which the exact Gibbs predictive also merges under regularity.

In addition, in an artificially simple example in which MCMC sampling of the Gibbs posterior is feasible, we illustrate that approximation of the posterior leads to negligible differences between the out-of-sample results yielded by the GVP and those produced by the predictive based on MCMC sampling from the exact Gibbs posterior. We establish this result under both correct specification, in which the true data generating process matches the adopted predictive model, and under misspecification of the predictive model. The correspondence between the 'approximate' and 'exact' predictions in the correct

specification case mimics that documented in Frazier *et al.* (2019), in which posterior approximations are produced by approximate Bayesian computation (ABC) and for the log score update (only). In the misspecified case, 'strictly coherent' predictions are produced (Martin *et al.*, 2020), whereby a given GVP, constructed via the use of a particular scoring rule, is shown to perform best out-of-sample according to that same score when compared with a GVP constructed via some alternative scoring rule; with the numerical values of the average out-of-sample scores closely matching those produced by the exact Gibbs predictive. That is, building a Gibbs posterior via a given scoring rule yields superior predictive accuracy in that rule *despite any inaccuracy* in the measurement of posterior uncertainty that is induced by the variational approximation.

We then undertake more extensive Monte Carlo experiments to highlight the power of the approach in genuinely high-dimensional problems, with predictives based on: an autoregressive mixture model with 20 mixture components and a neural network model used for illustration. An empirical analysis, in which GVP is used to produce accurate prediction intervals for the 4227 daily time series used in the M4 forecasting competition, illustrates the applicability of the method to reasonably large, and realistic data sets.

The remainder of the paper is structured as follows. In Section 2 the loss-based paradigm for Bayesian prediction is defined, with its links to related segments of the literature (as flagged briefly above) detailed. In Section 3 we detail how to construct the GVP and the stochastic gradient ascent (SGA) method that we use in its implementation. Further computational details are included in an appendix to the paper. An illustration of the ability of the variational method to yield essentially equivalent predictive accuracy to that produced via MCMC sampling is given, using a low-dimensional example. We then proceed with the numerical illustration of the method in high-dimensional settings - using both artificially simulated and empirical data - in Sections 4 and 5 respectively. With the numerical illustrations all highlighting that the method 'works' and 'works well', we then provide its theoretical verification in Section 6. We conclude in Section 7 with discussion of the implications of our findings, including for future research directions. The computational details associated with all variational approximations, including all prior specifications, and the proofs of all theoretical results, are included in appendices to the paper.

2 Setup and Loss-Based Bayesian Prediction

2.1 Preliminaries and notation

Consider a stochastic process $\{Y_t : \Omega \to \mathcal{Y}, t \in \mathbb{N}\}$ defined on the complete probability space $(\Omega, \mathcal{F}, P_0)$. Let $\mathcal{F}_t := \sigma(Y_1, \ldots, Y_t)$ denote the natural sigma-field, and let P_0 denote the infinite-dimensional distribution of the sequence $Y_1, Y_2, \ldots$. Let $y_{1:n} = (y_1, \ldots, y_n)'$ denote a vector of realizations from the stochastic process.

Our goal is to use a particular collection of statistical models, adapted to $\mathcal{F}_n$, that describe the behav-

ior of the observed data, to construct accurate predictions for the random variable Y_{n+1}. For every $n \geq 1$, the parameters of the model are denoted by θ_n, the parameter space by $\Theta_n \subseteq \mathbb{R}^{d_n}$, where the dimension d_n could grow as $n \to \infty$ and $\Theta_1 \subseteq \Theta_2 \subseteq ... \subseteq \Theta_n$. For the notational simplicity, we drop the dependence of θ_n and Θ_n on n in what follows. Let $\mathcal{P}^{(n)}$ be a generic class of one-step-ahead predictive models for Y_{n+1}, conditioned on the information $\mathcal{F}_n$ available at time n, such that $\mathcal{P}^{(n)} := \{P_\theta^{(n)}, \theta \in \Theta\}$.[1] When $P_\theta^{(n)}(\cdot)$ admits a density with respect to the Lebesgue measure, we denote it by $p_\theta^{(n)}(\cdot) \equiv p_\theta(\cdot|\mathcal{F}_n)$. The parameter θ thus indexes values in the predictive class, with θ taking values in the complete probability space $(\Theta, \mathcal{T}, \Pi)$, and where Π measures our beliefs - either prior or posterior - about the unknown parameter θ, and when they exist we denote the respective densities by $\pi(\theta)$ and $\pi(\theta|y_{1:n})$.

Denoting the likelihood function by $p_\theta(y_{1:n})$, the conventional approach to Bayesian prediction updates prior beliefs about θ via Bayes rule, to form the Bayesian posterior density,

$$\pi(\theta|y_{1:n}) = \frac{p_\theta\left(y_{1:n}\right)\pi(\theta)}{\int_\Theta p_\theta\left(y_{1:n}\right)\pi(\theta)d\theta}, \tag{1}$$

in which we follow convention and abuse notation by writing this quantity as a density even though, strictly speaking, the density may not exist. The one-step-ahead predictive distribution is then constructed as

$$P_\Pi^{(n)} := \int_\Theta P_\theta^{(n)}\pi(\theta|y_{1:n})d\theta. \tag{2}$$

However, as has recently been argued by some authors (e.g. Lacoste-Julien et al., 2011; Loaiza-Maya *et al.*, 2020a), when the class of predictive models indexed by Θ does not contain the *true* predictive distribution there is no sense in which this conventional approach remains the 'gold standard'. Instead, in this realistic scenario of model misspecification the predictive paradigm needs to change; with the log-score loss that underpins (2) replaced by the *particular* predictive loss that matters for the problem at hand. That is, our prior beliefs about θ and, hence, about the elements $P_\theta^{(n)}$ in $\mathcal{P}^{(n)}$, need to be updated via a criterion function defined by a user-specified measure of predictive loss; hence the nomenclature: 'loss-based Bayesian prediction'.

2.2 Bayesian updating based on scoring rules

Loaiza-Maya *et al.* (2020a) propose a method for producing Bayesian predictions using loss functions that specifically capture the accuracy of density forecasts. For $\mathcal{P}^{(n)}$ a convex class of predictive distributions on $(\Omega, \mathcal{F})$, density prediction accuracy can be measured using the scoring rule $s : \mathcal{P}^{(n)} \times \mathcal{Y} \mapsto \mathbb{R}$, where the expected scoring rule under the true distribution P_0 is defined as

$$\mathbb{S}(\cdot, P_0) := \int_{y\in\Omega} s(\cdot, y)\mathrm{d}P_0(y). \tag{3}$$

[1]The treatment of scalar Y_t and one-step-ahead prediction is for the purpose of illustration only, and all the methodology that follows can easily be extended to multivariate Y_t and multi-step-ahead prediction in the usual manner.

Since $\mathbb{S}(\cdot, P_0)$ is unattainable in practice, a sample estimate based on $y_{1:n}$ is used to define a sample criterion: for a given $\theta \in \Theta$, define sample average score as

$$S_n(\theta) := \sum_{t=0}^{n-1} s(P_\theta^{(t)}, y_{t+1}). \tag{4}$$

Adopting the *generalized* updating rule proposed by Bissiri *et al.* (2016) (see also Giummolè *et al.*, 2017, Holmes and Walker, 2017, Guedj, 2019, Lyddon *et al.*, 2019, and Syring and Martin, 2019), Loaiza-Maya *et al.* (2020a) distinguish elements in Θ using

$$\pi_w(\theta|y_{1:n}) = \frac{\exp\left[wS_n\left(\theta\right)\right]\pi(\theta)}{\int_\Theta \exp\left[wS_n\left(\theta\right)\right]\pi(\theta)\mathrm{d}\theta}, \tag{5}$$

where the scale factor w is obtained in a preliminary step using measures of predictive accuracy, and the same comment regarding abuse of density notation made above, applies here. This posterior can be referred to a *Gibbs* posterior, in the spirit of Zhang (2006a,b) and Jiang and Tanner (2008). It explicitly places weight on elements of Θ that lead to predictive models, $P_\theta^{(n)}$, with higher predictive accuracy in the scoring rule $s(\cdot,\cdot)$. As such, the one-step-ahead predictive,

$$P_{\Pi_w}^{(n)} := \int_\Theta P_\theta^{(n)} \pi_w(\theta|y_{1:n})\mathrm{d}\theta, \tag{6}$$

will often outperform, in the chosen rule $s(\cdot,\cdot)$, the likelihood (or log-score)-based predictive $P_\Pi^{(n)}$ in cases where the model is misspecified. Given its explicit dependence on a Gibbs posterior we refer to the predictive in (6) as the (exact) Gibbs predictive.

3 Gibbs Variational Prediction

3.1 Overview

While the Gibbs posterior measure $\Pi_w(\cdot|y_{1:n})$ is our *ideal* posterior, it becomes difficult to sample when the dimension of θ is large, which occurs in situations with a large number of predictors, or in flexible models. Therefore, the exact Gibbs predictive itself is not readily available in such cases. However, this does not invalidate the tenants on which the predictive $P_{\Pi_w}^{(n)}$ is constructed. Viewed in this way, we see that the problem of predictive inference via $P_{\Pi_w}^{(n)}$ could be solved if one were able to construct an accurate enough approximation to $\Pi_w(\cdot|y_{1:n})$. Herein, we construct such an approximation using variational Bayes (VB).

VB approximates $\Pi_w(\cdot|y_{1:n})$ by attempting to find the closest member in a class $\mathcal{Q}$ of probability measures, and requires a choice of divergence measures for measuring the discrepancy between the

elements of $\mathcal{Q}$ and $\Pi_w(\cdot|y_{1:n})$. The most common class of divergences for VB is the class of Renyi divergences.

Definition 3.1. Let $\gamma > 0$, with $\gamma \neq 1$, and let $P, Q \in \mathcal{P}$ for $\mathcal{P}$ a convex class of probability measures. The γ-Renyi divergence between Q and P is defined as

$$\mathrm{D}_\gamma\left(Q\|P\right) = \begin{cases} \frac{1}{\gamma-1}\log\int\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)^{\gamma-1}\mathrm{d}Q, & \text{if } Q \ll P \\ +\infty, & \text{otherwise.} \end{cases}$$

The Kullback-Leibler divergence, defined as $\mathrm{KL}(Q||\Pi) := \lim_{\gamma\to 1} D_\gamma(Q||\Pi)$, is the most commonly used divergence in the literature.

Given a particular member of the Renyi divergence, we can then define a variational predictive approximation of $P^{(n)}_{\Pi_w}$, say $P^{(n)}_Q$, as follows.

Definition 3.2. For $\mathcal{Q}$ a variational family of distributions, the variational posterior $\widehat{Q}$ satisfies

$$\inf_{Q\in\mathcal{Q}} \mathrm{D}_\gamma\left[Q||\Pi_w(\cdot|y_{1:n})\right] + o_p(1) \geq \mathrm{D}_\gamma\left[\widehat{Q}\|\Pi_w(\cdot|y_{1:n})\right],$$

and the loss-based variational predictive is defined as

$$P^{(n)}_Q := \int_\Theta P^{(n)}_\theta \mathrm{d}\widehat{Q}(\theta).$$

The variational predictive $P^{(n)}_Q$ circumvents the need to construct $P^{(n)}_{\Pi_w}$ via sampling from the posterior $\Pi_w(\cdot|y_{1:n})$. In essence, we replace the sampling problem with an optimization problem for which reliable methods exist even if Θ is high-dimensional, and which in turn yields an approximation to $\Pi_w(\cdot|y_{1:n})$. Consequently, even though, as discussed earlier, it is not always feasible to sample $\Pi_w(\cdot|y_{1:n})$ in situations where Θ is high-dimensional, and/or in cases where the loss is not conjugate with the prior, optimizing a divergence between $Q \in \mathcal{Q}$ and $\Pi_w(\cdot|y_{1:n})$ remains feasible using variational approaches.

Considering the KL divergence between Q and $\Pi_w(\cdot|y_{1:n})$, in this case we obtain the following variational objective function

$$\mathrm{KL}\left(Q||\Pi_w\right) = \int \log\left[\mathrm{d}Q/\mathrm{d}\Pi_w(\cdot|y_{1:n})\right]\mathrm{d}Q$$

which, if Π_w and Q both admit densities, and with the latter density denoted by q, yields

$$\mathrm{KL}\left(Q||\Pi_w\right) = \int \log[q(\theta)]q(\theta)d\theta - \int \log\left\{\exp\left[wS_n(\theta)\right]\pi(\theta)\right\}q(\theta)\mathrm{d}\theta + \log\int\exp\left[wS_n(\theta)\right]\pi(\theta)\mathrm{d}\theta. \tag{7}$$

Defining, in turn, the so-called evidence lower bound (ELBO) as

$$\text{ELBO}[Q||\Pi_w] := \mathbb{E}_q[\log\{\exp[wS_n(\theta)]\,\pi(\theta)\}] - \mathbb{E}_q[\log\{q(\theta)\}], \tag{8}$$

and recognizing that the (inaccessible) final term in (7) is constant with respect to (w.r.t.) $q(\theta)$, minimizing $\text{KL}(Q||\Pi_w)$ w.r.t $q(\theta)$ is seen to be equivalent to maximizing $\text{ELBO}[Q||\Pi_w]$ w.r.t. $q(\theta)$; and this is the way in which the variational problem is solved in practice.

This variational approach to prediction is related to the 'generalized variational inference' approach of Knoblauch *et al.* (2019), but with two main differences. Firstly, our approach is focused on prediction, not on parameter inference *per se*. Our only goal is to produce density forecasts that are as accurate as possible for the given scoring rule $s(\cdot, y)$, with the quantification of parameter uncertainty only a step taken towards that goal. Secondly, our approach follows the idea of Bissiri *et al.* (2016) and Loaiza-Maya *et al.* (2020a) and targets the predictions built from the Gibbs posterior in (5), rather than the exponentiated form of some general loss as in Knoblauch *et al.* (2019). This latter point is certainly critical if inferential accuracy were still deemed to be important, since without the tempering constant w that defines the posterior in (5), the exponentiated loss function can be very flat and the posterior $\Pi_w(\cdot|y_{1:n})$ uninformative about θ.[2] It is our experience however (Loaiza-Maya *et al.*, 2020a), that in settings where predictive accuracy is the only goal, the choice of w actually has little impact on the generation of accurate predictions via $P^{(n)}_{\Pi_w}$, as long as the value of w is not too small and the sample size is reasonably large. Prelimary experimentation in the current setting, in which the variational predictive $P^{(n)}_Q$ is the target, suggests that this finding remains relevant, and as a consequence we have adopted a default value of $w = 1$ in all numerical work. Further research is needed to deduce the precise impact of w on $P^{(n)}_Q$, in particular for smaller sample sizes, and we leave this important topic for future work.

3.2 A toy example: Predicting a financial return

3.2.1 Simulation design

Before proceeding to a more comprehensive demonstration of GVP in both artificial and empirical settings (Sections 4 and 5), plus the proof of its theoretical properties (Section 6), we demonstrate the ability of the method to produce accurate predictions in a simple toy example for a financial return, Y_t, in which the exact Gibbs predictive is also accessible.[3] In brief, we show that driving the Bayesian update by a particular scoring rule does reap benefits - in terms of accuracy out-of-sample in that given rule - relative to prediction based on a misspecified likelihood function; i.e. we demonstrate that GVP yields strictly

[2] See Bissiri *et al.* (2016), Giummolè *et al.* (2017), Holmes and Walker (2017), Lyddon *et al.* (2019), Syring and Martin (2019) and Pacchiardi and Dutta (2021) for various approaches to the setting of w in inferential settings.

[3] We reiterate here, and without further repetition, that all details of the prior specifications and the numerical steps required to produce the variational approximations, for this and the following numerical examples, are provided in Appendices A to D.

coherent predictions in the misspecified setting, as hoped. Crucially, this conclusion holds *despite* the approximation of the posterior via VB. That is, the GVP yields almost equal (average) out-of-sample predictions to those produced by the exact Gibbs posterior accessed via MCMC.

The predictive class, $P^{(t)}$, is defined by a generalized autoregressive conditional heteroscedastic GARCH(1,1) model with Gaussian errors, $Y_t = \theta_1 + \sigma_t \varepsilon_t$, $\varepsilon_t \overset{i.i.d.}{\sim} N(0,1)$, $\sigma_t^2 = \theta_2 + \theta_3 (Y_{t-1} - \theta_1)^2 + \theta_4 \sigma_{t-1}^2$, with $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)'$. We adopt three alternative specifications for the true data generating process (DGP) : 1) a model that matches the Gaussian GARCH(1,1) predictive class; 2) a stochastic volatility model with leverage:

$$
\begin{aligned}
Y_t &= \exp\left(\frac{h_t}{2}\right)\varepsilon_t \\
h_t &= -2 + 0.7\left(h_{t-1} - (-2)\right) + \eta_t \\
(\varepsilon_t, \eta_t)' &\overset{i.i.d.}{\sim} N\left(0, \begin{bmatrix} 1 & -0.35 \\ -0.35 & 0.25 \end{bmatrix}\right);
\end{aligned}
$$

and 3) a stochastic volatility model with a smooth transition in the volatility autoregression:

$$
\begin{aligned}
Y_t &= \exp\left(\frac{h_t}{2}\right)\varepsilon_t \\
h_t &= 0.9 g\left(h_{t-1}\right) h_{t-1} + \eta_t \\
\eta_t &\overset{i.i.d.}{\sim} N(0, 0.25),
\end{aligned}
$$

where $g(x) = \left(1 + \exp\left(-2x\right)\right)^{-1}$. DGP 1) defines a *correct specification* setting, whilst DGPs 2) and 3) characterize different forms of *misspecification.*

Denoting the predictive density function associated with the Gaussian GARCH (1,1) model, evaluated at the observed y_{t+1}, as $p_\theta(y_{t+1}|\mathcal{F}_t)$, we implement GVP using four alternative forms of (postively-oriented) scoring rules:

$$s^{\text{LS}}\left(P_\theta^{(t)}, y_{t+1}\right) = \log p_\theta(y_{t+1}|\mathcal{F}_t), \tag{9}$$

$$s^{\text{CRPS}}\left(P_\theta^{(t)}, y_{t+1}\right) = -\int_{-\infty}^{\infty}\left[P_\theta^{(t)} - I(y \geq y_{t+1})\right]^2 dy, \tag{10}$$

$$s^{\text{CLS-A}}\left(P_\theta^{(t)}, y_{t+1}\right) = \log p_\theta(y_{t+1}|\mathcal{F}_t) I\left(y_{t+1} \in A\right) + \left[\ln \int_{A^c} p_\theta(y|\mathcal{F}_t) dy\right] I\left(y_{t+1} \in A^c\right), \tag{11}$$

$$s^{\text{MSIS}}\left(P_\theta^{(t)}, y_{t+1}\right) = (u_{t+1} - l_{t+1}) + \frac{2}{\alpha}\left(l_{t+1} - y_{t+1}\right) I\left(y_{t+1} < l_{t+1}\right) + \frac{2}{\alpha}\left(y_{t+1} - u_{t+1}\right) I\left(y_{t+1} > u_{t+1}\right), \tag{12}$$

where l_{t+1} and u_{t+1} denote the $100(\frac{\alpha}{2})\%$ and $100(1-\frac{\alpha}{2})\%$ predictive quantiles. The log-score (LS) in (9) is a 'local' scoring rule, attaining a high value if the observed value, y_{t+1}, is in the high density region of

$p_\theta(y_{t+1}|F_t)$. The continuously ranked probability score (CRPS) in (10) (Gneiting and Raftery, 2007) is, in contrast, sensitive to distance, and rewards the assignment of high predictive mass near to the realized y_{t+1}, rather than just at that value. The score in (11) is the censored likelihood score (CLS) of Diks *et al.* (2011), which rewards predictive accuracy over any pre-specified region of interest A (A^c denoting the complement). We use the score for A defining the lower and upper tails of the predictive distribution, as determined in turn by the 10%, 20%, 80% and 90% percentiles of the empirical distribution of Y_t, labelling these cases hereafter as CLS10, CLS20, CLS80 and CLS90. A high value of this score in any particular instance thus reflects a predictive distribution that accurately predicts extreme values of the financial return. The last score considered is the mean scaled interval score (MSIS) in (12), which is designed to measure the accuracy of the $100(1-\alpha)\%$ predictive interval where $\alpha = 0.05$. This score rewards narrow intervals with accurate coverage. All components of (11) have closed-form solutions for the (conditionally) Gaussian predictive model, as does the integral in (10) and the bounds in (12).

In total then, seven distinct scoring rules are used to define the sample criterion function in (4). In what follows we reference these seven criterion functions, and the associated Gibbs posteriors using the notation $S_n^j(\theta) = \sum_{t=0}^{n-1} s^j\left(P_\theta^{(t)}, y_{t+1}\right)$, for $j = \{$LS, CRPS, CLS10, CLS20, CLS80, CLS90, MSIS$\}$.

3.2.2 Estimation of the Gibbs predictives

Given the low dimension of the predictive model it is straightforward to use an MCMC scheme to sample from the exact Gibbs posterior

$$\pi_w^j(\theta|y_{1:n}) \propto \exp\{wS_n^j(\theta)\}\pi(\theta),$$

where $\pi_w^j(\theta|y_{1:n})$ is the exact Gibbs posterior density in (5) computed under scoring rule j. As noted earlier, in this and all following numerical work we set $w = 1$. For each of the j posteriors, we initialize the chains using a burn-in period of 20000 periods, and retain the next $M = 20000$ draws $\theta^{(m,j)} \sim \pi_w^j(\theta|y_{1:n})$, $m = 1, \ldots, M$. The posterior draws are then used to estimate the exact Gibbs predictive in (6) via

$$\hat{P}_{\Pi_w}^{(n,j)} = \frac{1}{M}\sum_{m=1}^{M} P_{\theta^{(m,j)}}^{(n)}.$$

To perform GVP, we first need to produce the variational approximation of $\pi_w^j(\theta|y_{1:n})$. This is achieved in several steps. First, the parameters of the GARCH(1,1) model are transformed to the real line. With some abuse of notation, we re-define here the GARCH(1,1) parameters introduced in the previous section with the superscript r to signal 'raw'. The parameter vector θ is then re-defined as $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)' = \left(\theta_1^r, \log(\theta_2^r), \Phi_1^{-1}(\theta_3^r), \Phi_1^{-1}(\theta_4^r)\right)'$, where Φ_1 denotes the (univariate) normal cumulative distribution function (cdf). The next step involves approximating $\pi_w^j(\theta|y_{1:n})$ for the re-defined θ. We adopt the mean-field variational family (see for example Blei *et al.*, 2017), with a product-form Gaussian density $q_\lambda(\theta) = \prod_{i=1}^4 \phi_1(\theta_i; \mu_i, d_i^2)$, where $\lambda = (\mu', d')'$ is the vector of variational parameters,

comprised of mean and variance vectors $\mu = (\mu_1, \mu_2, \mu_3, \mu_4)'$ and $d = (d_1, \ldots, d_4)'$ respectively, and ϕ_1 denotes the (univariate) normal probability density function (pdf). Denote as Q_λ and Π_w^j the distribution functions associated with $q_\lambda(\theta)$ and $\pi_w^j(\theta|y_{1:n})$, respectively. The approximation is then calibrated by solving the minimization problem

$$\tilde{\lambda} = \underset{\lambda \in \Lambda}{\arg\min}\ \text{KL}\left[Q_\lambda || \Pi_w^j\right], \tag{13}$$

which, as highlighted earlier, is equivalent to maximizing the ELBO. Now defining

$$\mathcal{L}(\lambda) := \text{ELBO}\left[Q_\lambda || \Pi_w^j\right],$$

and remembering the use of the notation $S_n^j(\theta)$ to denote (4) for scoring rule j, (8) (for case j) becomes

$$\mathcal{L}(\lambda) = \mathbb{E}_{q_\lambda}\left[w S_n^j(\theta) + \log \pi(\theta) - \log q_\lambda(\theta)\right]. \tag{14}$$

Optimization is performed via stochastic gradient ascent methods (SGA). SGA maximizes $L(\lambda)$ by first initializing the variational parameters at some vector of values $\lambda^{(0)}$, and then sequentially iterating over

$$\lambda^{i+1} = \lambda^i + \Delta\lambda^{i+1}\left[\widehat{\nabla_\lambda \mathcal{L}(\lambda^{(i)})}, \rho\right].$$

The step size $\Delta\lambda^{i+1}$ is a function of an unbiased estimate of the ELBO gradient, $\widehat{\nabla_\lambda \mathcal{L}(\lambda^{(i)})}$, and the set of tuning parameters that control the learning rates in the optimization problem, denoted by ρ. Throughout this paper we employ the ADADELTA method of Zieler (2012) to update $\Delta\lambda^{i+1}$.

Key to achieving *fast* maximization of $\mathcal{L}(\lambda)$ via SGA is the use of a variance-reduction technique in producing the unbiased gradient estimate $\widehat{\nabla_\lambda \mathcal{L}(\lambda^{(i)})}$. Here, we follow Kingma and Welling (2014) and Rezende et al. (2014), and make use of the 'reparameterization trick'. In this approach, a draw θ from q_λ is written as a direct function of λ, and a set of random variables ε that are invariant with respect to λ. For the mean-field approximation used for this example we can write $\theta(\lambda, \varepsilon) = \mu + d \circ \varepsilon$, with $\varepsilon \sim N(0_4, I_4)$. This reparametrization allows us to re-write $\mathcal{L}(\lambda)$ as

$$\mathcal{L}(\lambda) = \mathbb{E}_\varepsilon\left[w S_n^j(\theta(\lambda, \varepsilon)) + \log \pi(\theta(\lambda, \varepsilon)) - \log q_\lambda(\theta(\lambda, \varepsilon))\right] \tag{15}$$

and its gradient as

$$\nabla_\lambda \mathcal{L}(\lambda) = \mathbb{E}_\varepsilon\left(\frac{\partial \theta(\lambda, \varepsilon)}{\partial \lambda}'\left[w \nabla_\theta S_n^j(\theta(\lambda, \varepsilon)) + \nabla_\theta \log \pi(\theta(\lambda, \varepsilon)) - \nabla_\theta \log q_\lambda(\theta(\lambda, \varepsilon))\right]\right). \tag{16}$$

A low-variance unbiased estimate of $\nabla_\lambda \mathcal{L}(\lambda)$ can be constructed by numerically computing the expectation in (16) using the (available) closed-form expressions for the derivatives $\frac{\partial \theta(\lambda, \varepsilon)}{\partial \lambda}$, $\nabla_\theta \log q_\lambda(\theta(\lambda, \varepsilon))$,

$\nabla_\theta \log \pi(\theta(\lambda, \varepsilon))$, and $\nabla_\theta S_n^j(\theta(\lambda, \varepsilon))$ for any j. We follow Kingma and Welling (2019) and use a single draw of ε for the construction of an estimate of (16). These expressions are provided in Appendix B. Once calibration of $\tilde{\lambda}$ is completed, the GVP is estimated as

$$\hat{P}_Q^{(n,j)} = \frac{1}{M} \sum_{m=1}^{M} P_{\theta^{(m,j)}}^{(n)} \tag{17}$$

with $\theta^{(m,j)} \overset{i.i.d.}{\sim} q_{\tilde{\lambda}}(\theta)$. To calibrate $\tilde{\lambda}$ 10000 VB iterations are used, and to estimate the variational predictive we set $M = 1000$.

To produce the numerical results we conduct the following steps:

1. Generate a times series of length $T = 6000$ from the true DGP;
2. Set $n = 1000$;
3. For $j \in \{$LS, CRPS, CLS10, CLS20, CLS80, CLS90, MSIS$\}$ construct the predictive densities $\hat{P}_{\Pi_w}^{(n,j)}$ and $\hat{P}_Q^{(n,j)}$ as outlined above;
4. For $(i, j) \in \{$LS, CRPS, CLS10, CLS20, CLS80, CLS90, MSIS$\}$ compute the measures of out-of-sample predictive accuracy $S_{\Pi_w}^{i,j,n} = s^i\left(\hat{P}_{\Pi_w}^{(n,j)}, y_{n+1}\right)$ and $S_Q^{i,j,n} = s^i\left(\hat{P}_Q^{(n,j)}, y_{n+1}\right)$;
5. If $n < T - 1$ set $n = n + 1$ and go to Step 3. If $n = T - 1$ go to Step 6.
6. Compute the average out-of-sample scores $S_{\Pi_w}^{i,j} = \frac{1}{5000}\sum_{n=1000}^{5999} S_{\Pi_w}^{i,j,n}$ and $S_Q^{i,j} = \frac{1}{5000}\sum_{n=1000}^{5999} S_Q^{i,j,n}$.

The results are tabulated and discussed in the the following section.

3.2.3 Results

The results of the simulation exercise are recorded in Table 1. Panel A records the results for the case of correct specification (Scenario 1)), with the average out-of-sample scores associated with the exact Gibbs predictive (estimated via MCMC) appearing in the left-hand-side panel (A.1) and the average scores for GVP appearing in the right-hand-side panel (A.2). Panel B records the corresponding results for Scenarios 2) and 3), associated with the two different forms of misspecification. All values on the diagonal of each sub-panel correspond to the case where $i = j$; i.e. the score used to compute the out-of-sample average is the same as that used to define the Gibbs posterior. In the misspecified case, numerical validation of the asymptotic result that the GVP concentrates onto the optimal predictive (to be proven in Section 6) occurs if the largest values (bolded) in a column appear on the diagonal. That is, if using a particular score to drive the update produces the best out-of-sample performance according to that same measure ('strict coherence' in the language of Martin *et al.*, 2020). In the correctly specified case, in which all proper scoring rules will, for a large enough sample, pick up the one true model (Gneiting and

Raftery, 2007), we would expect all values in given column to be very similar to one another, differences reflecting sampling variation only. Finally, validation of the theoretical property of merging between the exact and variational Gibbs predictive (also to be proven in Section 6) occurs if the corresponding results in all left- and right-hand-side panels are equivalent (again, up to sampling error).

As is clear, there is almost uniform numerical validation of all theoretical results. Under misspecification Scenario 2), (Panel B) the GVP results are strictly coherent (i.e. all bold values lie on the diagonal), with the exact Gibbs predictive results equivalent to the corresponding GVP results to three or four decimal places. The same broad findings obtain under misspecification scenario 3), apart from the fact that the MSIS updates are second best (to the log-score; and then only just) in terms of the out-of-sample MSIS measure. In Panel A on the other hand, we see the expected (virtual) equivalence of all results in a given column, reflecting the fact that all Gibbs predictives (however estimated) are concentrating onto the true predictive model and, hence, have identical out-of-sample performance. Of course, for a large but still finite number of observations, we would expect the log-score to perform best, due to the efficiency of the implicit maximum likelihood estimator underpinning the results and, to all intents and purposes this is exactly what we observe in Panels A.1 and A.2.

In summary, GVP performs as anticipated, and reaps distinct benefits in terms of predictive accuracy. Any inaccuracy in the measurement of parameter uncertainty also has negligible impact on the finite sample performance of GVP relative to an exact comparator. In Section 4 we extend the investigation into design settings that mimic the high-dimensional problems to which we would apply the variational approach in practice, followed by an empirical application - again using a high-dimensional predictive model - in Section 5. The theoretical exposition then follows in Section 6.

4 Simulation Study

4.1 Simulation Design

In this section we demonstrate the application of GVP in two realistic simulated examples. In both cases the assumed predictive model is high-dimensional and the exact Gibbs posterior, even if accessible in principle via MCMC, is challenging from a computational point of view. The mean-field class is adopted in both cases to produce variational approximations to the Gibbs posterior. The simulation design for each example (including the choice of $w = 1$) mimics that for the toy example, apart from the obvious changes made to the true DGP and the assumed predictive model, some changes in the size of the estimation and evaluation periods, plus the absence of comparative exact results. For reasons of computational burden we remove the CRPS update from consideration in the first example.

Table 1: Predictive accuracy of GVP using a Gaussian GARCH(1,1) predictive model for a financial return. Panel A corresponds to the correctly specified case, and Panels B and C to the two different misspecified settings as described in the text. The rows in each panel refer to the update method (U.method) used. The columns refer to the out-of-sample measure used to compute the average scores. The figures in bold are the largest average scores according to a given out-of-sample measure.

Panel A. True DGP: GARCH(1,1)

	A.1: Exact Gibbs predictive Average out-of-sample score							**A.2: GVP Average out-of-sample score**						
U.method	LS	CLS10	CLS20	CLS80	CLS90	CRPS	MSIS	LS	CLS10	CLS20	CLS80	CLS90	CRPS	MSIS
LS	**-0.0433**	**-0.2214**	**-0.3087**	**-0.3029**	-0.2150	**-0.1438**	**-1.1853**	**-0.0434**	**-0.2216**	**-0.3088**	**-0.3028**	**-0.2150**	**-0.1438**	**-1.1857**
CLS10	-0.0485	-0.2222	-0.3094	-0.3069	-0.2190	-0.1441	-1.2009	-0.0496	-0.2227	-0.3100	-0.3075	-0.2199	-0.1441	-1.2053
CLS20	-0.0482	-0.2221	-0.3094	-0.3066	-0.2185	-0.1440	-1.2014	-0.0493	-0.2225	-0.3097	-0.3072	-0.2193	-0.1440	-1.2052
CLS80	-0.0558	-0.2305	-0.3198	-0.3032	-0.2150	-0.1446	-1.2087	-0.0567	-0.2311	-0.3204	-0.3033	-0.2152	-0.1446	-1.2107
CLS90	-0.0495	-0.2258	-0.3143	**-0.3029**	-0.2150	-0.1441	-1.1973	-0.0502	-0.2265	-0.3149	-0.3031	-0.2152	-0.1441	-1.1993
CRPS	-0.0462	-0.2239	-0.3116	-0.3027	**-0.2147**	**-0.1438**	-1.1874	-0.0474	-0.2239	-0.3112	-0.3041	-0.2162	-0.1439	-1.1918
MSIS	-0.0438	-0.2216	-0.3089	-0.3031	-0.2153	**-0.1438**	-1.1877	-0.0440	-0.2218	-0.3091	-0.3031	-0.2153	**-0.1438**	-1.1882

Panel B. True DGP: Stochastic volatility with leverage

	B.1: Exact Gibbs predictive Average out-of-sample score							**B.2: GVP Average out-of-sample score**						
U.method	LS	CLS10	CLS20	CLS80	CLS90	CRPS	MSIS	LS	CLS10	CLS20	CLS80	CLS90	CRPS	MSIS
LS	**-0.5636**	-0.3753	-0.5452	-0.3536	-0.2512	-0.2313	-2.3468	**-0.5633**	-0.3752	-0.545	-0.3535	-0.2511	-0.2313	-2.3467
CLS10	-1.0193	**-0.3336**	-0.5021	-0.8379	-0.7339	-0.3679	-3.447	-1.0156	**-0.3336**	-0.502	-0.834	-0.7302	-0.3659	-3.4207
CLS20	-0.806	-0.3354	**-0.4968**	-0.6291	-0.5267	-0.2863	-2.9923	-0.8055	-0.3355	**-0.4969**	-0.6282	-0.5259	-0.2861	-2.9853
CLS80	-0.9203	-0.7372	-0.9311	**-0.329**	**-0.2292**	-0.2402	-3.3135	-0.9357	-0.7514	-0.9463	**-0.329**	-0.2292	-0.2402	-3.3248
CLS90	-0.9575	-0.7615	-0.9649	-0.3294	**-0.2292**	-0.2425	-3.4213	-0.9959	-0.7969	-1.0033	-0.3293	**-0.2291**	-0.2426	-3.4476
CRPS	-0.5692	-0.4029	-0.5671	-0.3431	-0.2419	**-0.23**	-2.4312	-0.5649	-0.3985	-0.5626	-0.3432	-0.2419	**-0.2301**	-2.4338
MSIS	-0.6552	-0.3986	-0.6111	-0.3713	-0.248	-0.2604	**-2.203**	-0.655	-0.3985	-0.6109	-0.3712	-0.2479	-0.2603	**-2.2033**

Panel C. True DGP: Stochastic volatility with smooth transition

	C.1: Exact Gibbs predictive Average out-of-sample score							**C.2: GVP Average out-of-sample score**						
U.method	LS	CLS10	CLS20	CLS80	CLS90	CRPS	MSIS	LS	CLS10	CLS20	CLS80	CLS90	CRPS	MSIS
LS	**-1.6858**	-0.4239	-0.672	-0.6751	-0.4369	-0.7196	**-7.0726**	**-1.686**	-0.424	-0.6721	-0.6752	-0.4371	-0.7196	**-7.0742**
CLS10	-1.8173	**-0.4172**	-0.6707	-0.8016	-0.5468	-0.821	-7.8217	-1.8145	**-0.4172**	-0.6706	-0.7987	-0.5437	-0.8183	-7.7717
CLS20	-1.7173	-0.419	**-0.6674**	-0.7067	-0.4616	-0.7425	-7.2516	-1.7171	-0.4192	**-0.6676**	-0.7063	-0.4611	-0.742	-7.2481
CLS80	-1.733	-0.465	-0.7199	**-0.6681**	-0.4304	-0.7511	-7.3087	-1.7339	-0.4654	-0.7204	**-0.6684**	-0.4307	-0.7513	-7.3159
CLS90	-1.8777	-0.5927	-0.8579	-0.6731	**-0.4288**	-0.8707	-8.1299	-1.8819	-0.5952	-0.8608	-0.674	**-0.4295**	-0.8734	-8.1211
CRPS	-1.6938	-0.4283	-0.6762	-0.6812	-0.4435	**-0.7184**	-7.2059	-1.6938	-0.4286	-0.6765	-0.6809	-0.4432	**-0.7185**	-7.2121
MSIS	-1.6879	-0.4244	-0.6726	-0.6754	-0.4374	-0.7208	-7.0915	-1.6881	-0.4245	-0.6728	-0.6754	-0.4374	-0.7208	-7.093

4.2 Example 1: Autoregressive mixture predictive model

In this example we adopt a true DGP in which Y_t evolves according to the logistic smooth transition autoregressive (LSTAR) process proposed in Teräsvirta (1994):

$$Y_t = \rho_1 Y_{t-1} + \rho_2 \left\{ \frac{1}{1 + \exp\left[-\gamma \left(Y_{t-1} - c\right)\right]} \right\} y_{t-1} + \sigma_\varepsilon \varepsilon_t, \tag{18}$$

where $\varepsilon_t \overset{i.i.d.}{\sim} t_\nu$, and t_ν denotes the standardised Student-t distribution with ν degrees of freedom. This model has the ability to produce data that exhibits a range of complex features. For example, it not only allows for skewness in the marginal density of Y_t, but can also produce temporal dependence structures that are asymmetric. We thus use this model as an illustration of a complex DGP whose characteristics are hard to replicate with simple parsimoneous models. Hence the need to adopt a highly parameterized predictive model; plus the need to acknowledge that even that representation will be misspecified.

The assumed predictive model is based on the flexible Bayesian non-parametric structure proposed in Antoniano-Villalobos and Walker (2016). The predictive distribution for Y_t, conditional on the observed y_{t-1}, is constructed from a mixture of $K = 20$ Gaussian autoregressive (AR) models of order one as follows:

$$P_\theta^{(t-1)} = \sum_{k=1}^{K} \tau_{k,t} \Phi_1 \left[Y_t - \mu; \beta_{k,0} + \beta_{k,1}(y_{t-1} - \mu), \sigma_k^2 \right], \tag{19}$$

with time-varying mixture weights

$$\tau_{k,t} = \frac{\tau_k \phi_1 \left(y_{t-1} - \mu; \mu_k, s_k^2\right)}{\sum_{j=1}^{K} \tau_j \phi_1 \left(y_{t-1} - \mu; \mu_j, s_j^2\right)},$$

where $\mu_k = \frac{\beta_{k,0}}{1-\beta_{k,1}}$ and $s_k^2 = \frac{\sigma_k^2}{1-\beta_{k,1}^2}$. Denoting $\tau = (\tau_1, \ldots, \tau_K)'$, $\beta_0 = (\beta_{1,0}, \ldots, \beta_{K,0})'$, $\beta_1 = (\beta_{1,1}, \ldots, \beta_{K,1})'$ and $\sigma = (\sigma_1, \ldots, \sigma_K)'$, then the full vector of unknown parameters is $\theta = \big(\mu, \tau', \beta_0', \beta_1', \sigma'\big)'$, which comprises $1 + (4 \times 20) = 81$ elements. GVP is a natural and convenient alternative to exact Gibbs prediction in this case.

4.3 Example 2: Bayesian neural network predictive model

In the second example we consider a true DGP in which the dependent variable Y_t has a complex non-linear relationship with a set of covariates. Specifically, the time series process $\{Y_t\}_{t=1}^T$, is determined by a three-dimensional stochastic process $\{X_t\}_{t=1}^T$, with $X_t = (X_{1,t}, X_{2,t}, X_{3,t})'$. The first two covariates are jointly distributed as $(X_{1,t}, X_{2,t})' \overset{i.i.d.}{\sim} N(0, \Sigma)$. The third covariate $X_{3,t}$, independent of the former two, is distributed according to an AR(4) process so that $X_{3,t} = \sum_{i=1}^{4} \alpha_i X_{3,t-i} + \sigma \varepsilon_t$, with $\varepsilon_t \overset{i.i.d.}{\sim} N(0, 1)$. The variable Y_t is then given by $Y_t = X_t' \beta_t$, where $\beta_t = (\beta_{1,t}, \beta_{2,t}, \beta_{3,t})'$ is a three dimensional vector of time-varying coefficients, with $\beta_{i,t} = b_i + a_i F(X_{3,t})$, and F denotes the marginal distribution function

of $X_{3,t}$ induced by the AR(4) model.

This particular choice of DGP has two advantages. First, given the complex dependence structure of the DGP (i.e. non-linear cross-sectional dependence as well as temporal dependence), it would be difficult, once again, to find a simple parsimoneous predictive model that would adequately capture this structure; hence motivating the need to adopt a high-dimensional model and to resort to GVP. Second, because it includes several covariates, we can assess how GVP performs for varying informations sets. For example, we can establish if the overall performance of GVP improves with an expansion of the conditioning information set, as we would anticipate; and if the inclusion of a more complete conditioning set affects the occurrence of strict coherence, or not.

A flexible model that has the potential to at least partially capture several features of the true DGP is a Bayesian feed-forward neural network that takes Y_t as the dependent variable and some of its lags, along with observed values of X_t, $x_t = (x_{1,t}, x_{2,t}, x_{3,t})'$, as the vector of independent inputs, z_t (see for instance Hernández-Lobato and Adams, 2015). The predictive distribution is defined as $P_\theta^{(t-1)} = \Phi_1\left(Y_t; g\left(z_t;\omega\right), \sigma_y^2\right)$. The mean function $g\left(z_t;\omega\right)$ denotes a feed-forward neural network with $q = 2$ layers, $r = 3$ nodes in each layer, a p-dimensional input vector z_t, and a d-dimensional parameter vector ω with $d = r^2(q-1) + r(p+q+1) + 1$. It allows for a flexible non-linear relationship between Y_t and the vector of observed covariates z_t. Defining $c = \log(\sigma_y)$, the full parameter vector of this predictive class, $\theta = (\omega', c)'$, is of dimension $d + 1$.

4.4 Simulation results

4.4.1 Example 1

A time series of length $T = 2500$ for Y_t is generated from the LSTAR model in (18), with the true parameters set as: $\rho_1 = 0$, $\rho_2 = 0.9$, $\gamma = 5$, $c = 0$, $\sigma_\varepsilon = 1$ and $\nu = 3$. With exception of the CRPS, which cannot be evaluated in closed-form for the mixture predictive class and is not included in the exercise as a consequence, the same scores in the toy example are considered. With reference to the simulation steps given earlier, the initial estimation window is 500 observations; hence the predictive accuracy results are based on an evaluation sample of size 2000.

The results in Table 2 are clear-cut. With one exception (that of CLS10), the best out-of-sample performance, according to a given measure, is produced by the version of GVP based on that same scoring rule. That is, the GVP results are almost uniformly strictly coherent: a matching of the update rule with the out-of-sample measure produces the best predictive accuracy in that measure, almost always.

4.4.2 Example 2

In this case, we generate a time series of length $T = 4000$ from the model discussed in Section 4.3, with specifications: $\Sigma_{11} = 1$, $\Sigma_{22} = 1.25$, $\Sigma_{12} = \Sigma_{21} = 0.5$, $\sigma^2 = 0.2$, $\alpha_1 = 0.5$, $\alpha_2 = 0.2$, $\alpha_3 = 0.15$, $\alpha_4 = 0.1$, $a_1 = 1.3$, $b_1 = 0$, $a_2 = -2.6$, $b_2 = 1.3$, $a_3 = -1.5$ and $b_3 = 1.5$. These settings generate

Table 2: Predictive accuracy of GVP using a autoregressive mixture model for data generated from an LSTAR model. The rows in each panel refer to the update method (U.method) used. The columns refer to the out-of-sample measure used to compute the average scores. The figures in bold are the largest average scores according to a given out-of-sample measure.

	Average out-of-sample score					
U.method	LS	CLS10	CLS20	CLS80	CLS90	MSIS
LS	**-1.253**	-0.345	-0.497	-0.452	-0.263	-5.589
CLS10	-1.445	-0.346	-0.512	-0.555	-0.321	-7.279
CLS20	-1.445	**-0.344**	**-0.496**	-0.589	-0.349	-6.674
CLS80	-1.333	-0.414	-0.571	**-0.450**	-0.260	-5.831
CLS90	-1.330	-0.407	-0.564	-0.451	**-0.259**	-5.730
MSIS	-1.410	-0.401	-0.558	-0.474	-0.282	**-5.550**

data with a negatively skewed empirical distribution, a non-linear relationship between the observations on Y_t and $X_{3,t}$, and autoregressive behavior in Y_t.

To assess if the performance of GVP is affected by varying information sets, we consider four alternative specifications for the input vector z_t in the assumed predictive model. These four specifications (labelled as Model 1, Model 2, Model 3 and Model 4, respectively) are: $z_t = y_{t-1}$, $z_t = (y_{t-1}, x_{1,t})'$, $z_t = (y_{t-1}, x_{2,t})'$ and $z_t = (y_{t-1}, x_{1,t}, x_{2,t})'$. The dimension of the parameter vector for each of these model specifications is $d + 1 = 23$, $d + 1 = 26$, $d + 1 = 26$ and $d + 1 = 29$, respectively. Given that the assumed predictive class is Gaussian, all scoring rules used in the seven updates, and for all four model specifications, can be evaluated in closed form. Referencing the simulation steps given earlier, the initial estimation window is 2000 observations; hence the predictive accuracy results are based on an evaluation sample of size 2000 as in the previous example.

With reference to the results in Table 3 we can make two observations. First, as anticipated, an increase in the information set produces higher average scores out-of-sample, for all updates; with the corresponding values increasing steadily and uniformly as one moves from the results in Panel A (based on $z_t = y_{t-1}$) to those in Panel D (based on the largest information, $z_t = (y_{t-1}, x_{1,t}, x_{2,t})'$) set. Secondly however, despite the improved performance of all versions of GVP as the information set is increased to better match that used in the true DGP, strict coherence still prevails. That is, 'focusing' on the measure that matters in the construction of the GVP update, still reaps benefits, despite the reduction in misspecification.

Table 3: Predictive accuracy of GVP using a Bayesian neural network for data generated from the dynamic non-linear regression model discussed in Section 4.3. Panels A to D document results for the four different versions of z_t, in which varying numbers of input variables are included in the predictive model. The rows in each panel refer to the update method (U.method) used. The columns refer to the out-of-sample measure used to compute the average scores. The figures in bold are the largest average scores according to a given out-of-sample measure.

Panel A: Model 1 ($z_t = y_{t-1}$)

	Average out-of-sample score						
U.method	LS	CLS10	CLS20	CLS80	CLS90	CRPS	MSIS
LS	-1.607	-0.489	-0.761	-0.520	-0.310	**-0.659**	-6.579
CLS10	-1.914	**-0.460**	-0.738	-0.852	-0.617	-0.914	-8.174
CLS20	-1.766	**-0.460**	**-0.732**	-0.704	-0.478	-0.772	-7.448
CLS80	-1.637	-0.514	-0.795	**-0.514**	**-0.303**	-0.678	-6.822
CLS90	-1.686	-0.534	-0.831	-0.521	-0.307	-0.718	-6.912
CRPS	-1.620	-0.513	-0.783	**-0.514**	-0.304	-0.660	-6.752
MSIS	**-1.601**	-0.476	-0.753	-0.516	-0.307	-0.660	**-6.338**

Panel B: Model 2 ($z_t = (y_{t-1}, x_{1,t})'$)

	Average out-of-sample score						
U.method	LS	CLS10	CLS20	CLS80	CLS90	CRPS	MSIS
LS	**-1.451**	-0.494	-0.733	-0.424	-0.254	**-0.550**	-6.059
CLS10	-1.858	**-0.435**	-0.686	-0.849	-0.626	-0.866	-7.707
CLS20	-1.683	-0.438	**-0.679**	-0.686	-0.476	-0.708	-6.963
CLS80	-1.551	-0.595	-0.850	**-0.403**	**-0.241**	-0.587	-6.878
CLS90	-1.539	-0.540	-0.811	-0.406	**-0.241**	-0.605	-6.369
CRPS	-1.510	-0.585	-0.822	-0.411	-0.244	-0.553	-6.649
MSIS	-1.452	-0.473	-0.728	-0.414	-0.248	0.562	**-5.671**

Panel C: Model 3 ($z_t = (y_{t-1}, x_{2,t})'$)

	Average out-of-sample score						
U.method	LS	CLS10	CLS20	CLS80	CLS90	CRPS	MSIS
LS	-1.536	-0.428	-0.688	-0.521	-0.314	-0.617	**-5.975**
CLS10	-1.775	**-0.396**	-0.658	-0.793	-0.555	-0.805	-6.704
CLS20	-1.646	-0.397	**-0.654**	-0.664	-0.439	-0.694	-6.277
CLS80	-1.698	-0.566	-0.858	-0.520	**-0.301**	-0.714	-7.282
CLS90	-1.744	-0.581	-0.889	-0.530	-0.304	-0.758	-7.361
CRPS	**-1.535**	-0.438	-0.695	**-0.517**	-0.310	**-0.615**	-6.044
MSIS	-1.549	-0.438	-0.7030	**-0.517**	-0.310	-0.624	-6.054

Panel D: Model 4 ($z_t = (y_{t-1}, x_{1,t}, x_{2,t})'$)

	Average out-of-sample score						
U.method	LS	CLS10	CLS20	CLS80	CLS90	CRPS	MSIS
LS	**-1.390**	-0.483	-0.716	-0.389	-0.229	**-0.511**	-5.890
CLS10	-1.758	**-0.394**	-0.650	-0.784	-0.548	-0.790	-6.659
CLS20	-1.656	**-0.394**	**-0.621**	-0.732	-0.515	-0.712	-6.236
CLS80	-2.127	-1.049	-1.438	**-0.356**	-0.196	-0.736	-9.798
CLS90	-2.157	-0.989	-1.401	-0.379	**-0.192**	-0.782	-9.545
CRPS	-1.519	-0.626	-0.872	-0.380	-0.218	-0.537	-6.922
MSIS	-1.673	-0.592	-0.934	-0.381	-0.203	-0.710	**-5.493**

5 Empirical Application: M4 competition

The Makridakis 4 (M4) forecasting competition was a forecasting event first organised by the University of Nicosia and the New York University Tandon School of Engineering in 2018. The competition sought submissions of point and interval predictions at different time horizons, for a total of 100,000 times series of varying frequencies. The winner of the competition in a particular category (i.e. point prediction or interval prediction) was the submission that achieved the best average out-of-sample predictive accuracy according to the measure of accuracy that defined that category, over all horizons and all series.[4]

We gauge the success of our method of *distributional* prediction in terms of the measure used to rank the interval forecasts in the competition, namely the MSIS. We focus on one-step-ahead prediction of each of the 4227 time series of daily frequency. Each of these series is denoted by $\{Y_{i,t}\}$ where $i = 1, \ldots, 4227$ and $t = 1, \ldots, n_i$, and the task is to construct a predictive interval for Y_{i,n_i+1} based on GVP. We adopt the mixture distribution in (19) as the predictive model, being suitable as it is to capture the stylized features of high frequency data. However, the model is only appropriate for stationary time series and most of the daily time series exhibit non-stationary patterns. To account for this, we model the differenced series $\left\{Z_{i,t} = \Delta^{d_i} Y_{i,t}\right\}$, where d_i indicates the differencing order. The integer d_i is selected

[4]Details of all aspects of the competition can be found via the following link:
https://www.m4.unic.ac.cy/wp-content/uploads/2018/03/M4-Competitors-Guide.pdf

by sequentially applying the KPSS test (Kwiatkowski *et al.*, 1992) to $\Delta^j y_{i,t}$ for $j = 0, \ldots, d_i$, with d_i being the first difference at which the null hypothesis of no unit root is not rejected.[5] To construct the predictive distribution of Z_{i,n_i+1} we first obtain $M = 5000$ draws $\{\theta_i^{(m)}\}_{m=1}^M$ from the Gibbs variational posterior $q_{\tilde{\lambda}}(\theta_i)$ that is based on the MSIS.[6] Draws $\{z_{i,n_i+1}^{(m)}\}_{m=1}^M$ are then obtained from the corresponding predictive distributions $\{P_{\theta_i^{(m)}}^{(n_i+1)}\}_{m=1}^M$. The draws of Z_{i,n_i+1} are then transformed into draws of Y_{i,n_i+1} and a predictive distribution for Y_{i,n_i+1} produced using kernel density estimation; assessment is performed in terms of the accuracy of the prediction interval for Y_{i,n_i+1}.

Table 4 documents the predictive performance of our approach relative to the competing methods (see Makridakis *et al.*, 2020 for details on these methods). The first column corresponds to the average MSIS. In terms of this measure the winner is the method proposed by Smyl (2020), while our method is ranked 11 out of 12. The ranking is six out of 12 in terms of the median MSIS, recorded in column two. However, it is worth remembering that our method aims to achieve high individual, and not aggregate (or average) predictive interval accuracy across series. A more appropriate way of judging the effectiveness of our approach is thus to count the number of series for which each method performs best. This number is reported in column three of the table. In terms of this measure, with a total number of 858 series, our approach is only outperformed by the method proposed in Doornik *et al.* (2020). In contrast, although Smyl is the best method in terms of average MSIS, it is only best at predicting 130 of the series in the dataset. It is also important to mention that most of the competing approaches have more complex assumed predictive classes than our mixture model. Even the simpler predictive classes like the ARIMA and ETS models have model selection steps that allow them to capture richer Markovian processes than the mixture model which, by construction, is based on an autoregressive process of order one.

In summary, *despite* a single specification being defined for the predictive class for all 4227 daily series, the process of driving the Bayesian update via the MSIS rule has still yielded the *best* predictive results in a very high number of cases, with GVP beaten in this sense by only one other competitor. Whilst the predictive model clearly matters, designing a bespoke updating mechanism to produce the predictive distribution is shown to still reap substantial benefits.

6 Theoretical Properties

As the above examples have demonstrated, GVP - by driving the updating mechanism by the measure of predictive accuracy that matters - produces accurate predictions in that measure, without requiring correct model specification. Moreover, it has been shown in one example that the accuracy is equivalent to that achieved by the exact Gibbs predictive; i.e., the approximation of the Gibbs posterior did not entail any loss in predictive accuracy. Critically, GVP is computationally feasible in predictive problems where generating reliable samples from the Gibbs posterior is infeasible, or too computationally burdensome.

[5] To do this we employ the autoarima function in the forecast R package.

[6] Once again, a value of $w = 1$ is used in defining the Gibbs posterior and, hence, in producing the posterior approximation.

Table 4: One-step-ahead predictive performance for the 4,227 daily series from the M4 competition. The columns 'Mean MSIS' and Median MSIS' respectively record the mean and median values of average out-of-sample MSMS, across all 4,227 series, for each competiting method/team. The column 'Series' reports the number of series for which the method/team produces the largest MSIS value. Some of the competing methods are individually described in Doornik *et al.* (2020), Fiorucci and Louzada (2020), Smyl (2020), Petropoulos and Svetunkov (2020), and Montero-Manso *et al.* (2020). For the remaining methods see Makridakis *et al.* (2020). Our GVP method based on the mixture model is referred to as 'Mixture'.

	Out-of-sample results		
	Mean MSIS	Median MSIS	Series
Method			
Doornik *et al.*	-9.36	-4.58	2060
Mixture	-16.66	-5.85	858
Trotta	-11.41	-7.19	376
ARIMA	-10.07	-5.67	278
Fiorucci and Louzada	-9.20	-5.73	163
Smyl	-8.73	-6.65	139
Roubinchtein	-13.08	-7.35	97
Ibrahim	-38.81	-6.33	88
ETS	-9.13	-5.71	67
Petropoulos and Svetunkov	-9.77	-5.68	59
Montero-Manso, *et al.*	-10.24	-6.90	23
Segura-Heras, *et al.*	-15.29	-8.48	19

In general, for any given finite sample, the discrepancy between the exact Gibbs predictive and the GVP will be problem dependent. However, if the forecasting problems under analysis are regular enough, we can demonstrate that the two approaches will indeed produce predictions that are equivalent in large samples.

We maintain the following notation throughout the remainder of the paper. We let $d : \Theta \times \Theta \to \mathbb{R}_{x \geq 0}$ denote a generic divergence measure that satisfies $d(\theta_1, \theta_2) \geq 0$ and $d(\theta_1, \theta_2) = 0$ implies $\theta_1 = \theta_2$. When quantities are Euclidean valued, we take $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ to be the Euclidean norm and inner product for vectors. We let C denote a generic constant independent of n that can vary from line to line. For sequences x_n, y_n and some C we write $x_n \lesssim y_n$ when $x_n \leq C y_n$. If $x_n \lesssim y_n$ and $y_n \lesssim x_n$ we write $x_n \asymp y_n$. For some positive sequence δ_n, the notation $o_p(\delta_n)$ and $O_p(\delta_n)$ have their usual connotation. Unless otherwise stated, all limits are taken as $n \to \infty$, so that when no confusion will result we use $\lim_n$ to denote $\lim_{n \to \infty}$. The random variable Y_n takes values in $\mathcal{Y}$ for all $n \geq 1$, and for a given n, the vector of observations $y_{1:n}$ are generated from the true probability measure P_0, not necessarily in $\mathcal{P}^{(n)}$, and whose dependence on n we supress to avoid confusion with the predictive model $P_\theta^{(n)}$.

6.1 Preliminary Result: Posterior Concentration

Before presenting our main result on the behavior of the GVP, we first state and prove a result regarding the posterior concentration of the variational approximation to the Gibbs posterior. Following Gneiting *et al.* (2007), define the 'optimal' point estimator within the class $\mathcal{P}^{(n)}$, and based on the scoring rule $s(\cdot, y)$ as

$$\hat{\theta}_n := \arg\max_{\theta \in \Theta} S_n(\theta), \text{ where } S_n(\theta) := \sum_{t=0}^{n-1} s(P_\theta^{(t)}, y_{t+1}). \tag{20}$$

In addition, define the limit counterpart to the optimal point estimator as

$$\theta_\star := \arg\max_{\theta \in \Theta} \mathcal{S}(\theta), \text{ where } \mathcal{S}(\theta) = \lim_{n \to \infty} \mathbb{E}\left[S_n(\theta)/n\right]. \tag{21}$$

Define $P_\star^{(n)} := P(\cdot | \mathcal{F}_n, \theta_\star)$ as the 'optimal' predictive. Our ultimate goal is to demonstrate that the predictions based on the GVP, $P_Q^{(n)}$, are equivalent to those made using $P_\star^{(n)}$. To demonstrate such a result, we must first establish convergence results for the exact Gibbs posterior and the variational approximation to the Gibbs posterior. We do so under the following high-level assumptions.

For some positive sequence $\epsilon_n \to 0$, define $\Theta(\epsilon_n) := \{\theta \in \Theta : d(\theta, \theta_\star) \leq \epsilon_n\}$ as an n-dependent neighbourhood of $\theta_\star$ and let $\Theta^c(\epsilon_n)$ denote its complement.

Assumption 6.1. *(i) There exists a non-random function $\mathcal{S}(\theta)$ such that $\sup_{\theta \in \Theta} \|S_n(\theta)/n - \mathcal{S}(\theta)\| \to 0$ in probability. (ii) For any $\epsilon \geq \epsilon_n$, there exists some $C > 0$ such that $\sup_{\Theta^c(\epsilon)} \{\mathcal{S}(\theta) - \mathcal{S}(\theta_\star)\} \leq -C\epsilon^2$.*

Remark 6.1. Assumption 6.1 places regularity conditions on the sample and limit counterpart of the expected scoring rules, and is used to deduce the asymptotic behavior of the Gibbs posterior. The first

part of the assumption is a standard uniform law of large numbers, and the second part amounts to an identification condition for $\theta_\star$.

In what follows, we investigate the rate of convergence of the exact Gibbs posterior. We note here that the non-likelihood nature of $S_n(\theta)$, and the temporal dependence of the observations, ensures that existing results on the concentration of Gibbs posteriors, and/or their variational approximation, of which we are aware are not applicable in our setting; e.g., the approaches of Zhang and Gao (2020), Alquier and Ridgway (2020) and Yang *et al.* (2020) are not directly applicable with data that is temporally dependent, and potentially non-Markovian.

Instead, we use smoothness of $S_n(\theta)$ (and the underlying model $P_\theta^{(n)}$) to deduce a posterior concentration result.

Assumption 6.2. *There exists a sequence of $d \times d$-dimensional matrices Σ_n, and a $d \times 1$-random vector Δ_n such:*

(i) $S_n(\theta) - S_n(\theta_\star) = \langle \sqrt{n}\,(\theta - \theta_\star)\,, \Delta_n/\sqrt{n} \rangle - \frac{n}{2}\|\Sigma_n^{1/2}\sqrt{n}\,(\theta - \theta_\star)\,\|^2 + R_n(\theta)$.

(ii) For any $\epsilon > \epsilon_n$, any $M > 0$, with $M/\sqrt{n} \to 0$, $\limsup P_0\left[\sup_{d(\theta,\theta_\star) \leq M/\sqrt{n}} |R_n(\theta)| > \epsilon\right] = 0$.

(iii) $\left\|\Sigma_n^{-1/2}\Delta_n\right\| = O_p(1)$.

We require the following a tail control condition on the prior.

Assumption 6.3. *For any $\epsilon > \epsilon_n$,* $\log\left\{\Pi\left[\Theta(\epsilon)\right]\right\} \gtrsim -n\epsilon^2$.

Remark 6.2. The above assumptions ultimately preclude cases where Θ can be partitioned into global parameters that drive the model, which are fixed for all n, and local parameters that represent time-dependent latent variables in the model and grow in a one-for-one manner with the sample size. As such, these assumptions are not designed to be applicable to variational inference in classes such as hidden Markov models.

The following is an intermediate result that gives a posterior concentration rate for the Gibbs posterior.

Lemma 6.1. *Under Assumptions 6.1-6.3, for $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$, and any $M_n \to \infty$, for $w_n \in (\underline{w}, \overline{w})$ with probability one, where $0 < \underline{w} < \overline{w} < \infty$,*

$$\Pi_w\left(d(\theta, \theta_\star) > M_n\epsilon_n | y_{1:n}\right) \lesssim \exp\left(-CM_n^2\epsilon_n^2 n\right),$$

with probability converging to one.

To transfer the posterior concentration of the Gibbs posterior to its variational approximation, we must restrict the class $\mathcal{Q}$ that is used to produce the variational approximation. Following Zhang and Gao (2020), we define κ_n to be the approximation error rate for the variational family:

$$\kappa_n^2 = \frac{1}{n} P_0 \inf_{Q \in \mathcal{Q}} \mathrm{D}_\gamma \left\{ Q \| \Pi_w(\cdot | y_{1:n}) \right\},$$

and where $D_\gamma(\cdot\|\cdot)$ is the Renyi divergence defined in 3.1.

Theorem 6.1. *Under the Assumptions of Lemma 6.1, for any sequence $M_n \to \infty$,*

$$\widehat{Q} \left\{ d(\theta, \theta_\star) \geq M_n n \left(\epsilon_n^2 + \kappa_n^2 \right) | y_{1:n} \right\} \to 0$$

in probability.

Remark 6.3. Theorem 6.1 gives a similar result to Theorem 2.1 in Zhang and Gao (2020) but for the Gibbs variational posterior and demonstrates that the latter concentrates onto $\theta_\star$, with the rate of concentration given by the slower of ϵ_n and κ_n. We note that, in cases where the dimension of Θ grows with the sample size, the rate ϵ_n is ultimately affected by the rate of growth of Θ and care must be taken to account for this dependence.

Remark 6.4. Verification of Theorem 6.1 in any practical example requires choosing the variational family $\mathcal{Q}$. In the case of observation-driven time series models, where the predictive model $P_\theta^{(n)}$ can be constructed analytically, and no explicit treatment of 'local' parameters is needed, the variational family can be taken to be a sufficiently rich parametric class. For example, consider the case where the dimension of Θ is fixed and compact, and $P_\theta^{(n)}$ satisfies the Assumptions of Lemma 6.1. Then, we can consider as our variational family the mean-field class:

$$\mathcal{Q}_{MF} := \left\{ Q : q(\theta) = \prod_{i=1}^{d_\theta} q_i(\theta_i) \right\},$$

or the Gaussian family

$$\mathcal{Q}_G := \{ \lambda = (\mu', \mathrm{vech}(\Sigma)')' : q(\theta) \propto \mathcal{N}(\theta; \mu, \Sigma), \},$$

where $\mu \in \mathcal{R}^d$, and Σ is a $d \times d$ positive-definite matrix. In either case, the approximation results of Zhang and Gao (2020) or Yang *et al.* (2020) for these variational families can be used to obtain the rate κ_n. In particular, when d_θ is fixed these results demonstrate that $\kappa_n \lesssim \epsilon_n$, and the Gibbs variational posterior converges at the same rate as the Gibbs posterior.[7]

[7]The rate of concentration obtained by applying Lemma 6.1 in such an example will be $\epsilon_n = \log(n)/\sqrt{n}$, which is slightly slower than the optimal parametric rate $\epsilon_n = 1/\sqrt{n}$. The parametric rate can be achieved by sufficiently modifying the prior

As seen from Theorem 6.1, the variational posterior places increasing mass on the value $\theta_\star$ that maximizes the limit of the expected scoring rule. Theorem 6.1 does not rely on the existence of exponential testing sequences as in much of the Bayesian nonparametrics literature (see, e.g., the treatment in Ghosal and Van der Vaart, 2017). The need to consider an alternative approach is due to the loss-based, i.e., non-likelihood-based, nature of the Gibbs posterior. Instead, the arguments used to demonstrate posterior concentration rely on controlling the behavior of a suitable quadratic approximation to the criterion function in a neighborhood of $\theta_\star$.

6.2 Key Asymptotic Result for Gibbs Variational Prediction

The result of Theorem 6.1 allows us to demonstrate that the predictions made using the GVP, $P_Q^{(n)}$, are just as accurate as those obtained by an agent that uses the 'optimal predictive' $P_\star^{(n)}$. The following is the main result of this paper, and relates the accuracy of the variational predictive $P_Q^{(n)}$, the optimal predictive $P_\star^{(n)}$, and the Gibbs predictive $P_{\Pi_w}^{(n)}$. Let $d_{\text{TV}}\{P, Q\}$ denote the total variation distance between the probability measures P and Q.

Theorem 6.2. *Under the Assumptions in Theorem 6.1,* $d_{TV}\{P_Q^{(n)}, P_\star^{(n)}\} \to 0$ *and* $d_{TV}\{P_Q^{(n)}, P_{\Pi_w}^{(n)}\} \to 0$ *in probability.*

Theorem 6.2 states that the variational predictive and the optimal frequentest predictive $P_\star^{(n)}$ agree in the rule $s(\cdot, \cdot)$ in large samples. This type of result is colloquially known as a 'merging' result (Blackwell and Dubins, 1962) for the corresponding predictives. Heuristically, the result suggests that the predictions obtained by the GVP and those obtained by a frequentest making predictions according to an optimal score estimator converge as the sample size diverges.

Furthermore, the theorem also demonstrates that the difference between predictions made using the GVP and the exact Gibbs predictive also merge. This latter result means that, if we take the exact Gibbs predictive as our gold standard for Bayesian prediction, the GVP yields predictions that are just as accurate as the exact Gibbs predictive (in large samples).

7 Discussion

We have developed a new approach for conducting loss-based Bayesian prediction in high-dimensional models. Based on a variational approximation to the Gibbs posterior defined, in turn, by the predictive loss that is germane to the problem at hand, the method is shown to produce predictions that minimize that loss out-of-sample. 'Loss' is characterized in this paper by positive-oriented proper scoring rules designed to reward the accuracy of predictive probability density functions for a continuous random variable. Hence, loss minimization translates to maximization of an expected score. However, in principle,

condition in Assumption 6.3 to cater for the parametric nature of the model. However, given that this is not germane to the main point off the paper, we do not discuss such situations here.

any loss function in which predictive accuracy plays a role could be used to define the Gibbs posterior. Critically, we have proven theoretically, and illustrated numerically, that for a large enough sample there is no loss incurred in predictive accuracy as a result of approximating the Gibbs posterior.

In comparison with the standard approach based on a likelihood-based Bayesian posterior, our Gibbs variational predictive approach is ultimately aimed at generating accurate predictions in the realistic empirical setting where the predictive model and, hence, the likelihood function is misspecified. Gibbs variational prediction enables the investigator to break free from the shackles of likelihood-based prediction, and to drive predictive outcomes according to the form of predictive accuracy that matters for the problem at hand; and all with theoretical validity guaranteed.

We have focused in the paper on particular examples where the model used to construct the Gibbs variational predictive is observation-driven. Extensions to parameter-driven models (i.e., hidden Markov models, or state space models) require different approaches with regard to both the implementation of variational Bayes, and in establishing the asymptotic properties of the resulting posteriors and predictives. This is currently the subject of on-going work by the authors, and we reference Tran *et al.* (2017), Quiroz *et al.* (2018), Koop and Korobilis (2018), Chan and Yu (2020), Gunawan *et al.* (2020) and Loaiza-Maya *et al.* (2020b) for some treatments of this problem that exist in the literature.

This paper develops the theory of Gibbs variational prediction for classes of models that are general and flexible enough to accommodate a wide range of data generating processes, and which, under the chosen loss function, are smooth enough to permit a quadratic expansion. This latter condition restricts the classes of models, and loss functions, under which our results are applicable. For example, the discrete class of models studied in Douc *et al.* (2013) may not be smooth enough to deduce the validity of such an expansion. See Cooper *et al.* (2021) for an approach to validating (likelihood-based) variational prediction in a more general class of observation-driven models, that differs from the approach adopted herein.

References

Alquier, P. and Ridgway, J. (2020). Concentration of tempered posteriors and of their variational approximations. *Annals of Statistics*, 48(3):1475–1497. 3, 22

Alquier, P., Ridgway, J., and Chopin, N. (2016). On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414. 3

Antoniano-Villalobos, I. and Walker, S. G. (2016). A nonparametric model for stationary time series. *Journal of Time Series Analysis*, 37(1):126–142. 15

Bassetti, F., Casarin, R., and Ravazzolo, F. (2018). Bayesian nonparametric calibration and combination of predictive distributions. *Journal of the American Statistical Association*, 113(522):675–685. 2

Batrk, N., Borowska, A., Grassi, S., Hoogerheide, L., and van Dijk, H. (2019). Forecast density combinations of dynamic models and data driven portfolio strategies. *Journal of Econometrics*, 210(1):170–186. 2

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

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130. 2, 6, 8

Blackwell, D. and Dubins, L. (1962). Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, 33(3):882–886. 3, 24

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877. 10

Casarin, R., Leisen, F., Molina, G., and ter Horst, E. (2015). A Bayesian beta Markov random field calibration of the term structure of implied risk neutral densities. *Bayesian Analysis*, 10(4):791–819. 2

Chan, J. C. and Yu, X. (2020). Fast and accurate variational inference for large Bayesian VARs with stochastic volatility. *CAMA Working Paper*. 25

Cooper, A. D. R., Frazier, D. T., Koo, B., and Martin, G. M. (2021). Variational forecasts for observation-driven models. *Working Paper*. 25

Diks, C., Panchenko, V., and Van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2):215–230. 10

Doornik, J. A., Castle, J. L., and Hendry, D. F. (2020). Card forecasts for M4. *International Journal of Forecasting*, 36(1):129–134. 19, 20

Douc, R., Doukhan, P., and Moulines, E. (2013). Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stochastic Processes and their Applications*, 123(7):2620–2647. 25

Fiorucci, J. A. and Louzada, F. (2020). Groec: Combination method via generalized rolling origin evaluation. *International Journal of Forecasting*, 36(1):105–109. 20

Frazier, D. T., Maneesoonthorn, W., Martin, G. M., and McCabe, B. P. (2019). Approximate Bayesian forecasting. *International Journal of Forecasting*, 35(2):521–539. 4

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge University Press. 24

Giummolè, F., Mameli, V., Ruli, E., and Ventura, L. (2017). Objective Bayesian inference with proper scoring rules. *TEST*, pages 1–28. 2, 6, 8

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268. 2, 21

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378. 2, 10, 12

Guedj, B. (2019). A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*. 6

Gunawan, D., Kohn, R., and Nott, D. (2020). Variational approximation of factor stochastic volatility models. *arXiv preprint arXiv:2010.06738*. 25

Hernández-Lobato, J. M. and Adams, R. (2015). Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869. PMLR. 16

Holmes, C. and Walker, S. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503. 6, 8

Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data-mining. *Annals of Statistics*, 36(5):2207–2231. 2, 6

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*. 12

Knoblauch, J., Jewson, J., and Damoulas, T. (2019). Generalized variational inference: Three arguments for deriving new posteriors. *arXiv preprint arXiv:1904.02063*. 2, 8

Koop, G. and Korobilis, D. (2018). Variational Bayes inference in high-dimensional time-varying parameter models. 25

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1):159–178. 19

Loaiza-Maya, R., Martin, G. M., and Frazier, D. T. (2020a). Focused Bayesian prediction. *Forthcoming. Journal of Applied Econometrics*. 2, 5, 6, 8

Loaiza-Maya, R., Smith, M. S., Nott, D. J., and Danaher, P. J. (2020b). Fast and accurate variational inference for models with many latent variables. *arXiv preprint arXiv:2005.07430*. 25

Lyddon, S., Holmes, C., and Walker, S. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478. 6, 8

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74. 19, 20

Martin, G. M., Loaiza-Maya, R., Worapree Maneesoonthorn, D. T. F., and Hassan, A. R. (2020). Optimal probabilistic forecasts: When do they work? 4, 12

McAlinn, K., Aastveit, K. A., Nakajima, J., and West, M. (2020). Multivariate Bayesian predictive synthesis in macroeconomic forecasting. *Journal of the American Statistical Association*, 115(531):1092–1110. 2

McAlinn, K. and West, M. (2019). Dynamic Bayesian predictive synthesis in time series forecasting. *Journal of Econometrics*, 210(1):155–169. 2

Miller, J. W. and Dunson, D. B. (2019). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125. 2

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., and Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, 36(1):86–92. 20

Pacchiardi, L. and Dutta, R. (2021). Generalized Bayesian likelihood-free inference using scoring rules estimators. *arXiv preprint arXiv:2104.03889*. 2, 8

Petropoulos, F. and Svetunkov, I. (2020). A simple combination of univariate models. *International journal of forecasting*, 36(1):110–115. 20

Pettenuzzo, D. and Ravazzolo, F. (2016). Optimal portfolio choice under decision-based model combinations. *Journal of Applied Econometrics*, 31(7):1312–1332. 2

Quiroz, M., Nott, D. J., and Kohn, R. (2018). Gaussian variational approximation for high-dimensional state space models. *arXiv preprint arXiv:1801.07873*. 25

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714. 41, 42

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 36(1):75–85. 19, 20

Syring, N. and Martin, R. (2019). Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486. 2, 6, 8

Syring, N. and Martin, R. (2020). Gibbs posterior concentration rates under sub-exponential type losses. *arXiv preprint arXiv:2012.04505*. 41

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the american Statistical association*, 89(425):208–218. 15

Tran, M.-N., Nott, D. J., and Kohn, R. (2017). Variational Bayes with intractable likelihood. *Journal of Computational and Graphical Statistics*, 26(4):873–882. 25

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. 41

Yang, Y., Pati, D., Bhattacharya, A., et al. (2020). alpha-variational inference with statistical guarantees. *Annals of Statistics*, 48(2):886–905. 3, 22, 23

Zhang, F. and Gao, C. (2020). Convergence rates of variational posterior distributions. *Annals of Statistics*, 48(4):2180–2207. 22, 23, 42, 43

Zhang, T. (2006a). From eps-entropy to kl entropy: analysis of minimum information complexity density estimation. *Annals of Statistics*, 34:21802210. 2, 3, 6

Zhang, T. (2006b). Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):13071321. 2, 6

A Details for the Implementation of all Variational Approximations

A.1 General notational matters

Throughout this appendix we will employ the following notation. All gradients are column vectors and their notation starts with the symbol ∇. For two generic matrices $A_{d_1 \times d_2}$ and $B_{d_3 \times d_4}$ we have that

$$\frac{\partial A}{\partial B} = \frac{\partial \text{vec}(A)}{\partial \text{vec}(B)},$$

where vec is the vectorization operation and $\frac{\partial A}{\partial B}$ is a matrix of dimension $(d_1 d_2) \times (d_3 d_4)$. Throughout this appendix scalars are treated as matrices of dimension 1×1.

The gradient of the log density of the approximation is computed as

$$\nabla_\theta \log q_\lambda(\theta) = -\left(D^2\right)^{-1}(\theta - \mu).$$

To compute $\frac{\partial \theta}{\partial \lambda}$ note that

$$\frac{\partial \theta}{\partial \lambda} = \left[\frac{\partial \theta}{\partial \mu}, \frac{\partial \theta}{\partial d}\right],$$

where

$$\frac{\partial \theta}{\partial \mu} = I_r, \qquad \frac{\partial \theta}{\partial d} = \text{diag}(e).$$

B The GARCH predictive class (Section 3.2)

The predictive class, $\mathcal{P}^{(t)}$, is defined by a generalized autoregressive conditional heteroscedastic GARCH(1,1) model with Gaussian errors, $Y_t = \theta_1^r + \sigma_t \varepsilon_t$, $\varepsilon_t \overset{i.i.d.}{\sim} N(0,1)$, $\sigma_t^2 = \theta_2^r + \theta_3^r (Y_{t-1} - \theta_1^r)^2 + \theta_4^r \sigma_{t-1}^2$, with $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)' = \left(\theta_1^r, \log(\theta_2^r), \Phi_1^{-1}(\theta_3^r), \Phi_1^{-1}(\theta_4^r)\right)'$. Note that throughout this section θ_1 and θ_1^r can be used interchangeably.

B.1 Priors

We employ the following priors for each of the parameters of the model:

$$p(\theta_1) \propto 1, \qquad p(\theta_2^r) \propto \frac{1}{\theta_2^r} I(\theta_2^r > 0), \qquad \theta_3^r \sim U(0,1), \quad \text{and} \quad \theta_4^r \sim U(0,1).$$

For the implementation of variational inference, all the parameters are transformed into the real line as follows:

$$\text{(i) } \theta_2^r \text{ is transformed to } \theta_2 = \log(\theta_2^r); \qquad \text{(ii) } \theta_3^r \text{ is transformed to } \theta_3 = \Phi_1^{-1}(\theta_3^r);$$
$$\text{(iii) } \theta_4^r \text{ is transformed to } \theta_4 = \Phi_1^{-1}(\theta_4^r).$$

After applying these transformations, we have that the prior densities are:

$$\text{(i) } p(\theta_1) \propto 1; \qquad \text{(ii) } p(\theta_2) \propto 1; \qquad \text{(iii) } p(\theta_3) = \phi_1(\theta_3); \qquad \text{(iv) } p(\theta_4) = \phi_1(\theta_4).$$

The gradient of the logarithm of the prior is $\nabla_\theta \log p(\theta) = (0, 0, -\theta_3, -\theta_4)'$.

B.2 Derivation of $\nabla_\theta S_n(\theta)$ for all scoring rules

We can show that $\nabla_\theta S_n(\theta) = \sum_{t=1}^n \nabla_\theta s\left(P_\theta^{(t-1)}, y_t\right)$. Thus, we must find an expression for $\nabla_\theta s\left(P_\theta^{(t-1)}, y_t\right)$ for each of the scores. The gradients from all the scores can be written as a function of the recursive derivatives:

$$\text{(i) } \frac{\partial \sigma_t^2}{\partial \theta_1} = -2\theta_3^r(y_{t-1} - \theta_1) + \theta_4^r \frac{\partial \sigma_{t-1}^2}{\partial \theta_1}; \qquad \text{(ii) } \frac{\partial \sigma_t^2}{\partial \theta_2^r} = 1 + \theta_4^r \frac{\partial \sigma_{t-1}^2}{\partial \theta_2^r};$$
$$\text{(iii) } \frac{\partial \sigma_t^2}{\partial \theta_3^r} = (y_{t-1} - \theta_1)^2 + \theta_4^r \frac{\partial \sigma_{t-1}^2}{\partial \theta_3^r}; \qquad \text{(iv) } \frac{\partial \sigma_t^2}{\partial \theta_4^r} = \theta_4^r \frac{\partial \sigma_{t-1}^2}{\partial \theta_4^r} + \sigma_{t-1}^2.$$

with $\frac{\partial \sigma_0^2}{\partial \theta_1} = 0$, $\frac{\partial \sigma_0^2}{\partial \theta_2^r} = 0$, $\frac{\partial \sigma_0^2}{\partial \theta_3^r} = 0$ and $\frac{\partial \sigma_0^2}{\partial \theta_4^r} = 0$.

B.2.1 Logarithmic score (LS)

The gradient for the LS can be written as

$$\nabla_\theta s^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right) = \left(\nabla_{\theta_1} s^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right)', \nabla_{\theta_2} s^{\text{LS}}\left(P_\theta^{t-1}, y_t\right), \nabla_{\theta_3} s^{\text{LS}}\left(P_\theta^{t-1}, y_t\right), \nabla_{\theta_4} s^{\text{LS}}\left(P_\theta^{t-1}, y_t\right)\right)',$$

with

$$\nabla_{\theta_1} S^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right) = -\frac{1}{2\sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta_1} + \frac{1}{2}\frac{(y_t - \theta_1)^2}{\sigma_t^4}\frac{\partial \sigma_t^2}{\partial \theta_1} + \frac{(y_t - \theta_1)}{\sigma_t^2}, \quad \nabla_{\theta_2} S^{\text{LS}}\left(P_\theta^{t-1}, y_t\right) = \theta_2^r\left[-\frac{1}{2\sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta_2^r} + \frac{1}{2}\frac{(y_t - \theta_1)^2}{\sigma_t^4}\frac{\partial \sigma_t^2}{\partial \theta_2^r}\right],$$
$$\nabla_{\theta_3} S^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right) = \phi_1(\theta_3)\left[-\frac{1}{2\sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta_3^r} + \frac{1}{2}\frac{(y_t - \theta_1)^2}{\sigma_t^4}\frac{\partial \sigma_t^2}{\partial \theta_3^r}\right] \quad \text{and} \quad \nabla_{\theta_4} S^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right) = \phi_1(\theta_4)\left[-\frac{1}{2\sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta_4^r} + \frac{1}{2}\frac{(y_t - \theta_1)^2}{\sigma_t^4}\frac{\partial \sigma_t^2}{\partial \theta_4^r}\right]$$

B.2.2 Continuously ranked probability score (CRPS)

For the Gaussian GARCH(1,1) predictive class the CRPS can be expressed as

$$s^{\text{CRPS}}\left(P_\theta^{(t-1)}, y_t\right) = -\sigma_t B_t,$$

with $B_t = z_t\left(2\Phi_1(z_t) - 1\right) + 2\phi_1(z_t) - \frac{1}{\sqrt{\pi}}$ and $z_t = \frac{y_t - \theta_1}{\sigma_t}$. The gradient of the CRPS can be written as

$$\nabla_\theta s^{\text{CRPS}}\left(P_\theta^{(t-1)}, y_t\right) = \left(\nabla_{\theta_1} s^{\text{CRPS}}\left(P_\theta^{t-1}, y_t\right), \nabla_{\theta_2} s^{\text{CRPS}}\left(P_\theta^{t-1}, y_t\right), \nabla_{\theta_3} s^{\text{CRPS}}\left(P_\theta^{t-1}, y_t\right), \nabla_{\theta_4} s^{\text{CRPS}}\left(P_\theta^{t-1}, y_t\right)\right)'.$$

The elements of this gradient are given by

$$\begin{aligned}
\nabla_{\theta_1} s^{\text{CRPS}}\left(P_\theta^{(t-1)}, y_t\right) &= -\sigma_t \frac{\partial B_t}{\partial z_t}\left[\frac{\partial z_t}{\partial \sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta_1} - \frac{1}{\sigma_t}\right] - \frac{B_t}{2\sigma_t}\frac{\partial \sigma_t^2}{\partial \theta_1} \\
\nabla_{\theta_2} s^{\text{CRPS}}\left(P_\theta^{(t-1)}, y_t\right) &= -\sigma_t \frac{\partial B_t}{\partial z_t}\frac{\partial z_t}{\partial \sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta_2^r} - \frac{B_t}{2\sigma_t}\frac{\partial \sigma_t^2}{\theta_2^r}\theta_2^r \\
\nabla_{\theta_3} s^{\text{CRPS}}\left(P_\theta^{(t-1)}, y_t\right) &= -\sigma_t \frac{\partial B_t}{\partial z_t}\frac{\partial z_t}{\partial \sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta_3^r} - \frac{B_t}{2\sigma_t}\frac{\partial \sigma_t^2}{\theta_3^r}\phi_1(\theta_3) \\
\nabla_{\theta_4} s^{\text{CRPS}}\left(P_\theta^{(t-1)}, y_t\right) &= -\sigma_t \frac{\partial B_t}{\partial z_t}\frac{\partial z_t}{\partial \sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta_4^r} - \frac{B_t}{2\sigma_t}\frac{\partial \sigma_t^2}{\theta_4^r}\phi_1(\theta_4)
\end{aligned}$$

with $\frac{\partial B_t}{\partial z_t} = 2\Phi_1(z_t) - 1$ and $\frac{\partial z_t}{\partial \sigma_t^2} = -\frac{z_t}{2\sigma_t^2}$.

B.2.3 Censored logarithmic score (CLS)

For some threshold value y_q, denote the upper tail support of predictive distribution as $A = \{y_t : y_t > y_q\}$. The gradient for the upper tail CLS can be written as

$$\begin{aligned}
&\nabla_\theta s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right) \\
= \;& \left(\nabla_{\theta_1} s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right), \nabla_{\theta_2} s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right), \nabla_{\theta_3} s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right), \nabla_{\theta_4} s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right)\right)',
\end{aligned}$$

with

$$\begin{aligned}
\nabla_{\theta_1} s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right) &= \nabla_{\theta_1} s^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right) I(y_t \in A) + \frac{\phi_1\left(\frac{y_q - \theta_1}{\sigma_t}\right)}{P_\theta^{(t-1)}(y_q)}\left(-\frac{y_q - \theta_1}{2\sigma_t^3}\frac{\partial \sigma_t^2}{\partial \theta_1} - \frac{1}{\sigma_t}\right) I(y_t \in A^c) \\
\nabla_{\theta_2} s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right) &= \nabla_{\theta_2} s^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right) I(y_t \in A) + \frac{\phi_1\left(\frac{y_q - \theta_1}{\sigma_t}\right)}{P_\theta^{(t-1)}(y_q)}\left(-\frac{y_q - \theta_1}{2\sigma_t^3}\frac{\partial \sigma_t^2}{\partial \theta_2^r}\right)\theta_2^r I(y_t \in A^c) \\
\nabla_{\theta_3} s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right) &= \nabla_{\theta_3} s^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right) I(y_t \in A) + \frac{\phi_1\left(\frac{y_q - \theta_1}{\sigma_t}\right)}{P_\theta^{(t-1)}(y_q)}\left(-\frac{y_q - \theta_1}{2\sigma_t^3}\frac{\partial \sigma_t^2}{\partial \theta_3^r}\right)\phi_1(\theta_3) I(y_t \in A^c) \\
\nabla_{\theta_4} s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right) &= \nabla_{\theta_4} s^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right) I(y_t \in A) + \frac{\phi_1\left(\frac{y_q - \theta_1}{\sigma_t}\right)}{P_\theta^{(t-1)}(y_q)}\left(-\frac{y_q - \theta_1}{2\sigma_t^3}\frac{\partial \sigma_t^2}{\partial \theta_4^r}\right)\phi_1(\theta_4) I(y_t \in A^c)
\end{aligned}$$

To obtain the gradient expressions for the lower tail CLS simply replace the term $\frac{\phi_1(\frac{y_q-\theta_1}{\sigma_t})}{P_\theta^{(t-1)}(y_q)}$ by the term $-\frac{\phi_1(\frac{y_q-\theta_1}{\sigma_t})}{1-P_\theta^{(t-1)}(y_q)}$, and redefine the set A.

B.2.4 Mean scaled interval score (MSIS)

The gradient of the MSIS can be written as

$$\nabla_\theta s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right) = \left(\nabla_{\theta_1} s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right), \nabla_{\theta_2} s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right), \nabla_{\theta_3} s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right), \nabla_{\theta_4} s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right)\right).$$

For $\theta_i \in \{\theta_1, \theta_2, \theta_3, \theta_4\}$, the elements of the gradient can be written as

$$\nabla_{\theta_i} s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right) = -\left[1 - \frac{2}{1-q} I(y_t > u_t)\right]\frac{\partial u_t}{\partial \theta_i} - \left[\frac{2}{1-q} I(y_t < l_t) - 1\right]\frac{\partial l_t}{\partial \theta_i},$$

where the derivative $\frac{\partial u_t}{\partial \theta_i}$ can be evaluated as $\frac{\partial u_t}{\partial \theta_i} = -\frac{\partial u_t}{\partial \alpha_{u,t}}\frac{\partial \alpha_{u,t}}{\partial \theta_i}$ with $\alpha_{u,t} = P_\theta^{(t-1)}(u_t)$. The first term can be computed as $\frac{\partial u_t}{\partial \alpha_{u,t}} = \frac{1}{p(u_t|y_{1:t-1},\theta)}$. The second term is parameter specific, and can be computed as

(i) $\frac{\partial \alpha_{u,t}}{\partial \theta_1} = \phi_1\left(\frac{u_t-\theta_1}{\sigma_t}\right)\left[\left(-\frac{u_t-\theta_1}{2\sigma_t^3}\right)\frac{\partial \sigma_t^2}{\partial \theta_1} - \frac{1}{\sigma_t}\right]$; (ii) $\frac{\partial \alpha_{u,t}}{\partial \theta_2} = \phi_1\left(\frac{u_t-\theta_1}{\sigma_t}\right)\left(-\frac{u_t-\theta_1}{2\sigma_t^3}\right)\frac{\partial \sigma_t^2}{\partial \theta_2^r}\theta_2^r$;

(iii) $\frac{\partial \alpha_{u,t}}{\partial \theta_3} = \phi_1\left(\frac{u_t-\theta_1}{\sigma_t}\right)\left(-\frac{u_t-\theta_1}{2\sigma_t^3}\right)\frac{\partial \sigma_t^2}{\partial \theta_3^r}\phi_1(\theta_3)$; (iv) $\frac{\partial \alpha_{u,t}}{\partial \theta_4} = \phi_1\left(\frac{u_t-\theta_1}{\sigma_t}\right)\left(-\frac{u_t-\theta_1}{2\sigma_t^3}\right)\frac{\partial \sigma_t^2}{\partial \theta_4^r}\phi_1(\theta_4)$.

The derivative $\frac{\partial l_t}{\partial \theta_i}$ is evaluated in the same fashion as described for $\frac{\partial u_t}{\partial \theta_i}$.

C Autoregressive mixture predictive class (Section 4.2)

C.1 Priors

We employ the following priors for each of the parameters of the model:

$$\beta_{k,0} \sim N(0, 10000^2), \quad \beta_{k,1} \sim U(-1,1), \quad v_k \sim \text{Beta}(1,2), \quad \left(\sigma_k^2\right)^{-1} \sim \mathcal{G}(1,1), \text{ and } \quad \mu \sim N(0, 100^2),$$

where the parameter v_k is given by the stick breaking decomposition of the mixture weights, which sets $\tau_k = v_k \prod_{j=1}^{k-1}(1-v_j)$, for $k = 1, \ldots, K$ and where $v_K = 1$. For the implementation of variational inference, all the parameters are transformed into the real line as follows:

(i) $\beta_{k,1}$ is transformed to $\eta_k = \Phi_1^{-1}\left(\frac{1}{2}(\beta_{k,1}+1)\right)$; (ii) v_k is transformed to $\psi_k = \Phi_1^{-1}(v_k)$;

(iii) σ_k is transformed to $\kappa_k = \log \sigma_k$.

After applying these transformations, we have that the prior densities are:

$$\text{(i) } p(\beta_{k,0}) = \phi_1\left(\beta_{k,1}; 0, 10000^2\right); \qquad \text{(ii) } p(\eta_k) = \phi_1\left(\eta_k; 0, 1\right);$$
$$\text{(iii) } p(\psi_k) \propto \left[1 - \Phi_1\left(\psi_k\right)\right]\phi_1\left(\psi_k\right); \qquad \text{(iv) } p(\kappa_k) \propto 2\exp\left(-2\kappa_k\right)\exp\left[-\exp\left(-2\kappa_k\right)\right];$$
$$\text{(v) } p(\mu) = \phi_1\left(\mu; 0, 100^2\right).$$

Computation of $\nabla_\theta \log p\left(\theta\right)$

Denoting as $\beta_0 = (\beta_{1,0}, \ldots, \beta_{K,0})'$, $\eta = (\eta_1, \ldots, \eta_K)'$, $\psi = (\psi_1, \ldots, \psi_K)'$ and $\kappa = (\kappa_1, \ldots, \kappa_K)'$, the gradient of the prior density is

$$\nabla_\theta \log p\left(\theta\right) = \left[\frac{\partial \log p\left(\theta\right)}{\partial \beta_0}, \frac{\partial \log p\left(\theta\right)}{\partial \eta}, \frac{\partial \log p\left(\theta\right)}{\partial \psi}, \frac{\partial \log p\left(\theta\right)}{\partial \kappa}, \frac{\partial \log p\left(\theta\right)}{\mu}\right]',$$

with

$$\text{(i) } \frac{\partial \log p\left(\theta\right)}{\partial \beta_0} = -(10000)^{-2}\beta_0'; \qquad \text{(ii) } \frac{\partial \log p\left(\theta\right)}{\partial \eta} = -\eta';$$
$$\text{(iii) } \frac{\partial \log p\left(\theta\right)}{\partial \psi} = -(1 - v')^{-1}\frac{\partial v}{\partial \psi} - \psi'; \qquad \text{(iv) } \frac{\partial \log p\left(\theta\right)}{\partial \kappa} = 2\tilde{\kappa}' - 2;$$
$$\text{(v) } \frac{\partial \log p\left(\theta\right)}{\partial \mu} = -(100)^{-2}\mu.$$

and where $(1 - v')^{-1} = ([1 - v_1]^{-1}, \ldots, [1 - v_{K-1}]^{-1})$ and $\tilde{\kappa} = (\exp[-2\kappa_1], \ldots, \exp[-2\kappa_K])'$.

C.2 Derivation of $\nabla_\theta S_n\left(\theta\right)$ for all scoring rules

The focused Bayesian update uses the term $S_n\left(\theta\right) = \sum_{t=1}^n s\left(P_\theta^{(t-1)}, y_t\right)$, thus variational inference requires evaluation of

$$\nabla_\theta S_n\left(\theta\right) = \sum_{t=1}^n \nabla_\theta s\left(P_\theta^{(t-1)}, y_t\right).$$

For the mixture example we consider three alternative choices for $s\left(P_\theta^{(t-1)}, y_t\right)$, namely, the MSIS, the CLS and LS. Here, we derive an expression for $\nabla_\theta s\left(P_\theta^{(t-1)}, y_t\right)$ for each of these scores. As will be shown later, the gradient $\nabla_\theta S\left(P_\theta^{(t-1)}, y_t\right)$ for all the scores can be expressed solely in terms of $\nabla_\theta p\left(y_t|\mathcal{F}_{t-1}, \theta\right)$ and $\nabla_\theta P_\theta^{(t-1)}$, thus we focus on the derivation of these two expressions. Below, we denote $\epsilon_t = y_t - \mu$ and use the scalar r to denote the number of elements in θ. For ease of notation we rewrite

$$p\left(y_t|\mathcal{F}_{t-1}, \theta\right) = \frac{c_{2,t}}{c_{1,t}} \quad \text{and} \quad P_\theta^{(t-1)} = \frac{c_{3,t}}{c_{1,t}},$$

where $c_{1,t,k} = \frac{\tau_k}{s_k}\phi_1\left(\frac{\epsilon_{t-1}-\mu_k}{s_k}\right)$, $c_{2,t,k} = \frac{1}{\sigma_k}\phi_1\left(\frac{\epsilon_t-\beta_{k,0}-\beta_{k,1}\epsilon_{t-1}}{\sigma_k}\right)$, $c_{3,t,k} = \Phi_1\left(\frac{\epsilon_t-\beta_{k,0}-\beta_{k,1}\epsilon_{t-1}}{\sigma_k}\right)$, $c_{1,t} = \sum_{k=1}^K c_{1,t,k}$, $c_{2,t} = \sum_{k=1}^K c_{1,t,k}c_{2,t,k}$ and $c_{3,t} = \sum_{k=1}^K c_{1,t,k}c_{3,t,k}$. The elements of the gradients $\nabla_\theta p\left(y_t|\mathcal{F}_{t-1},\theta\right) = \left[\nabla_{\theta_1} p\left(y_t|\mathcal{F}_{t-1},\theta\right),\ldots,\nabla_{\theta_r} p\left(y_t|\mathcal{F}_{t-1},\theta\right)\right]'$ and $\nabla_\theta P_\theta^{(t-1)} = \left[\nabla_{\theta_1} P_\theta^{(t-1)},\ldots,\nabla_{\theta_r} P_\theta^{(t-1)}\right]'$, can then be computed as

$$\nabla_{\theta_i} p\left(y_t|\mathcal{F}_{t-1},\theta\right) = \sum_{k=1}^K \left[\tau_{k,t}\frac{\partial c_{2,t,k}}{\partial\theta_i} + \frac{c_{2,t,k} - p\left(y_t|\mathcal{F}_{t-1},\theta\right)}{c_{1,t}}\frac{\partial c_{t,k,1}}{\partial\theta_i}\right] \tag{22}$$

and

$$\nabla_{\theta_i} P_\theta^{(t-1)} = \sum_{k=1}^K \left[\tau_{k,t}\frac{\partial c_{3,t,k}}{\partial\theta_i} + \frac{c_{3,t,k} - P_\theta^{(t-1)}}{c_{1,t}}\frac{\partial c_{t,k,1}}{\partial\theta_i}\right]. \tag{23}$$

Table 5 provides the expressions $\frac{\partial c_{1,t,k}}{\partial\theta_i}$, $\frac{\partial c_{2,t,k}}{\partial\theta_i}$ and $\frac{\partial c_{3,t,k}}{\partial\theta_i}$ for $\theta_i \in \{\beta_0,\eta,\kappa,\mu\}$. For $\theta_i \in \psi$, the gradients can be evaluated as

$$\nabla_\psi p\left(y_t|\mathcal{F}_{t-1},\theta\right) = \left[\nabla_\tau p\left(y_t|\mathcal{F}_{t-1},\theta\right)'\frac{\partial\tau}{\partial v}\frac{\partial v}{\partial\psi}\right]',$$

and

$$\nabla_\psi P_\theta^{(t-1)} = \left[\nabla_\tau P_\theta^{(t-1)\prime}\frac{\partial\tau}{\partial v}\frac{\partial v}{\partial\psi}\right]',$$

where $\frac{\partial v}{\partial\psi}$ is a diagonal matrix with entries $\frac{\partial v_k}{\partial\psi_k} = \phi_1\left(\psi_k\right)$, and $\frac{\partial\tau}{\partial v}$ is a lower triangular matrix with diagonal elements $\frac{\partial\tau_k}{\partial v_k} = \prod_{j=1}^{k-1}(1-v_j)$ and off-diagonal elements $\frac{\partial\tau_k}{\partial v_s} = -\frac{\tau_k}{(1-v_s)}$, for $s<k$. The expressions needed to evaluate $\nabla_\tau p\left(y_t|\mathcal{F}_{t-1},\theta\right) = \left[\nabla_{\tau_1} p\left(y_t|\mathcal{F}_{t-1},\theta\right),\ldots,\nabla_{\tau_K} p\left(y_t|\mathcal{F}_{t-1},\theta\right)\right]'$ and $\nabla_\tau P_\theta^{(t-1)} = \left[\nabla_{\tau_1} P_\theta^{(t-1)},\ldots,\nabla_{\tau_K} P_\theta^{(t-1)}\right]'$ are also provided in Table 5.

With expressions (22) and (23), we can now derive expression for $\nabla_\theta s\left(P_\theta^{(t-1)},y_t\right)$ for the three scores considered.

C.2.1 Logarithmic score (LS)

Denote the gradient of the LS in (9) as

$$\nabla_\theta s^{\text{LS}}\left(P_\theta^{(t-1)},y_t\right) = \left(\nabla_{\theta_1} s^{\text{LS}}\left(P_\theta^{(t-1)},y_t\right),\ldots,\nabla_{\theta_r} s^{\text{LS}}\left(P_\theta^{(t-1)},y_t\right)\right)'.$$

The element $\nabla_{\theta_i} s^{\text{LS}}\left(P_\theta^{(t-1)},y_t\right)$ of this gradient can be evaluated as

$$\nabla_{\theta_i} s^{\text{LS}}\left(P_\theta^{(t-1)},y_t\right) = \frac{1}{p\left(y_t|\mathcal{F}_{t-1},\theta\right)}\nabla_{\theta_i} p\left(y_t|\mathcal{F}_{t-1},\theta\right)$$

where the derivative $\nabla_{\theta_i} p\left(y_t|\mathcal{F}_{t-1},\theta\right)$ can be computed using (22).

C.2.2 Censored logarithmic score (CLS)

For some threshold value y_q, denote the upper tail support of predictive distribution as $A = \{y_t : y_t > y_q\}$. The gradient for the upper tail CLS in (11) can be written as

$$\nabla_\theta s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right) = \left(\nabla_{\theta_1} s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right), \ldots, \nabla_{\theta_r} s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right)\right)'.$$

We can compute the element $\nabla_{\theta_i} s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right)$ of the upper CLS as:

$$\nabla_{\theta_i} s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right) = \nabla_{\theta_i} s^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right) I(y_t \in A) + \frac{1}{P_\theta^{(t-1)}(y_q)} \nabla_{\theta_i} P_\theta^{(t-1)}(y_q) I(y_t \in A^c).$$

The derivatives $\nabla_{\theta_i} P_\theta^{(t-1)}(y_q)$ can be computed using (23). For the lower tail CLS, where $A = \{y_t : y_t < y_q\}$ we have that

$$\nabla_{\theta_i} s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right) = \nabla_{\theta_i} s^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right) I(y_t \in A) - \frac{1}{1 - P_\theta^{(t-1)}(y_q)} \nabla_{\theta_i} P_\theta^{(t-1)}(y_q) I(y_t \in A^c).$$

C.2.3 Mean scaled interval score (MSIS)

The gradient of the MSIS in (12) can be written as

$$\nabla_\theta s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right) = \left(\nabla_{\theta_1} s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right), \ldots, \nabla_{\theta_r} s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right)\right)',$$

with elements $\nabla_{\theta_i} s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right)$ defined as:

$$\nabla_{\theta_i} s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right) = -\left[1 - \frac{2}{1-q} I(y_t > u_t)\right] \frac{\partial u_t}{\partial \theta_i} - \left[\frac{2}{1-q} I(y_t < l_t) - 1\right] \frac{\partial l_t}{\partial \theta_i}$$

As such, computation of $\nabla_{\theta_i} s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right)$ entails evaluation of $\frac{\partial u_t}{\partial \theta_i}$ and $\frac{\partial l_t}{\partial \theta_i}$. The derivative $\frac{\partial u_t}{\partial \theta_i}$ can be evaluated by first noting that $\alpha_{u,t} = P_\theta^{(t-1)}(u_t)$. Then, using the triple product rule we know that

$$\frac{\partial u_t}{\partial \theta_i} = -\frac{\partial u_t}{\partial \alpha_{u,t}} \frac{\partial \alpha_{u,t}}{\partial \theta_i},$$

where the first term can be computed as $\frac{\partial u_t}{\partial \alpha_{u,t}} = \frac{1}{p(u_t | \mathcal{F}_{t-1}, \theta)}$. The second term $\frac{\partial \alpha_{u,t}}{\partial \theta_i} = \nabla_{\theta_i} P_\theta^{(t-1)}(u_t)$ is evaluated using (23). The derivative $\frac{\partial l_t}{\partial \theta_i}$ is evaluated in the same fashion as described for $\frac{\partial u_t}{\partial \theta_i}$.

Table 5: Derivatives required to implement reparameterization trick for the mixture model. Other required expressions also include $\frac{\partial \mu_k}{\partial \beta_{1,k}} = \frac{\beta_{0,k}}{(1-\beta_{1,k})^2}$, $\frac{\partial s_k}{\partial \beta_{1,k}} = \frac{\beta_{1,k}\sigma_k}{(1-\beta_{1,k}^2)^{3/2}}$, $\frac{\partial s_k}{\partial \sigma_k} = \frac{1}{(1-\beta_{1,k}^2)^{1/2}}$, $\frac{\partial \sigma_k}{\partial \kappa_k} = \exp(\kappa_k)$, $\frac{\partial \beta_{1,k}}{\partial \eta_k} = 2\phi_1(\eta_k)$. The cross-component derivatives $\frac{\partial c_{i,t,k}}{\partial \tau_j}$, $\frac{\partial c_{i,t,k}}{\partial \beta_{0,j}}$, $\frac{\partial c_{i,t,k}}{\partial \eta_j}$, $\frac{\partial c_{i,t,k}}{\partial \kappa_j}$ are all zero for $j \neq k$. Finally, in this table we denote $\phi_1'(x) = \frac{\partial \phi_1(x)}{\partial x}$.

Expressions for $\frac{\partial}{\partial \theta_i} c_{1,t,k}$

$$\frac{\partial c_{1,t,k}}{\partial \tau_k} = \frac{1}{s_k}\phi_1\left(\frac{\epsilon_{t-1}-\mu_k}{s_k}\right)$$

$$\frac{\partial c_{1,t,k}}{\partial \beta_{0,k}} = -\frac{\tau_k}{s_k}\phi_1'\left(\frac{\epsilon_{t-1}-\mu_k}{s_k}\right)\frac{1}{(1-\beta_{k,1})s_k}$$

$$\frac{\partial c_{1,t,k}}{\partial \eta_k} = \left[-\frac{\tau_k}{s_k}\phi_1'\left(\frac{\epsilon_{t-1}-\mu_k}{s_k}\right)\left(\frac{1}{s_k}\frac{\partial \mu_k}{\partial \beta_{1,k}} + \frac{1}{s_k^2}(\epsilon_{t-1}-\mu_k)\frac{\partial s_k}{\partial \beta_{1,k}}\right) - \phi_1\left(\frac{\epsilon_{t-1}-\mu_k}{s_k}\right)\frac{\tau_k}{s_k^2}\frac{\partial s_k}{\partial \beta_{1,k}}\right]\frac{\partial \beta_{1,k}}{\partial \eta_k}$$

$$\frac{\partial c_{1,t,k}}{\partial \kappa_k} = \left[\frac{\tau_k}{s_k}\phi_1'\left(\frac{\epsilon_{t-1}-\mu_k}{s_k}\right)\left(\frac{1}{s_k}\frac{\partial \mu_k}{\partial \sigma_k} + \frac{1}{s_k^2}(\epsilon_{t-1}-\mu_k)\frac{\partial s_k}{\partial \sigma_k}\right) - \phi_1\left(\frac{\epsilon_{t-1}-\mu_k}{s_k}\right)\frac{\tau_k}{s_k^2}\frac{\partial s_k}{\partial \sigma_k}\right]\frac{\partial \sigma_k}{\partial \kappa_k}$$

$$\frac{\partial c_{1,t,k}}{\partial \mu} = -\frac{\tau_k}{s_k}\phi_1'\left(\frac{\epsilon_{t-1}-\mu_k}{s_k}\right)\frac{1}{s_k}$$

Expressions for $\frac{\partial}{\partial \theta_i} c_{2,t,k}$

$$\frac{\partial c_{2,t,k}}{\partial \tau_k} = 0$$

$$\frac{\partial c_{2,t,k}}{\partial \beta_{0,k}} = -\frac{1}{\sigma_k^2}\phi_1'\left(\frac{\epsilon_t-\beta_{0,k}-\beta_{1,k}\epsilon_{t-1}}{\sigma_k}\right)$$

$$\frac{\partial c_{2,t,k}}{\partial \eta_k} = -\frac{\epsilon_{t-1}}{\sigma_k^2}\phi_1'\left(\frac{\epsilon_t-\beta_{0,k}-\beta_{1,k}\epsilon_{t-1}}{\sigma_k}\right)\frac{\partial \beta_{1,k}}{\partial \eta_k}$$

$$\frac{\partial c_{2,t,k}}{\partial \kappa_k} = -\frac{\epsilon_t-\beta_{0,k}-\beta_{1,k}\epsilon_{t-1}}{\sigma_k^3}\phi_1'\left(\frac{\epsilon_t-\beta_{0,k}-\beta_{1,k}\epsilon_{t-1}}{\sigma_k}\right)\frac{\partial \sigma_k}{\partial \kappa_k} - \frac{1}{\sigma_k^2}\phi_1\left(\frac{\epsilon_t-\beta_{0,k}-\beta_{1,k}\epsilon_{t-1}}{\sigma_k}\right)\frac{\partial \sigma_k}{\partial \kappa_k}$$

$$\frac{\partial c_{2,t,k}}{\partial \mu} = \frac{-1+\beta_{1,k}}{\sigma_k^2}\phi_1'\left(\frac{\epsilon_t-\beta_{0,k}-\beta_{1,k}\epsilon_{t-1}}{\sigma_k}\right)$$

Expressions for $\frac{\partial}{\partial \theta_i} c_{3,t,k}$

$$\frac{\partial c_{3,t,k}}{\partial \tau_k} = 0$$

$$\frac{\partial c_{3,t,k}}{\partial \beta_{0,k}} = -\frac{1}{\sigma_k}\phi_1\left(\frac{\epsilon_t-\beta_{0,k}-\beta_{1,k}\epsilon_{t-1}}{\sigma_k}\right)$$

$$\frac{\partial c_{3,t,k}}{\partial \eta_k} = -\frac{\epsilon_{t-1}}{\sigma_k}\phi_1\left(\frac{\epsilon_t-\beta_{0,k}-\beta_{1,k}\epsilon_{t-1}}{\sigma_k}\right)\frac{\partial \beta_{1,k}}{\partial \eta_k}$$

$$\frac{\partial c_{3,t,k}}{\partial \kappa_k} = -\frac{\epsilon_t-\beta_{0,k}-\beta_{1,k}\epsilon_{t-1}}{\sigma_k^2}\phi_1\left(\frac{\epsilon_t-\beta_{0,k}-\beta_{1,k}\epsilon_{t-1}}{\sigma_k}\right)\frac{\partial \sigma_k}{\partial \kappa_k}$$

$$\frac{\partial c_{3,t,k}}{\partial \mu} = \frac{-1+\beta_{1,k}}{\sigma_k}\phi_1\left(\frac{\epsilon_t-\beta_{0,k}-\beta_{1,k}\epsilon_{t-1}}{\sigma_k}\right)$$

D Bayesian neural network predictive class (Section 4.3)

D.1 Priors

The priors of the model parameters are set as $\omega_k \sim N(0, \text{Inf})$ and $p(\sigma_y^2) \propto \frac{1}{\sigma_y^2}$. The standard deviation parameter σ_y is transformed to the real line as $c = \log(\sigma_y)$, so that the parameter vector is $\theta = (\omega', c)'$. Then, the prior density can be written as $p(\theta) = p(c)\prod_{k=1}^{d} p(\omega_k)$, where $p(\omega_k) \propto 1$ and $p(c) \propto 1$. From this prior density we have that $\nabla_\theta \log p(\theta) = 0$.

D.2 Derivation of $\nabla_\theta S_n(\theta)$ for all scoring rules

As in the previous appendix we can show that $\nabla_\theta S_n(\theta) = \sum_{t=1}^{n} \nabla_\theta s\left(P_\theta^{(t-1)}, y_t\right)$. Thus, we must find an expression for $\nabla_\theta S\left(P_\theta^{(t-1)}, y_t\right)$ for each of the scores.

D.2.1 Logarithmic score (LS)

The gradient for the LS can be written as

$$\nabla_\theta s^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right) = \left(\nabla_\omega^{\text{LS}} s\left(P_\theta^{(t-1)}, y_t\right)', \nabla_c s^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right)\right)',$$

with

$$\nabla_\omega s^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right) = \frac{1}{\sigma_y^2}\left[y_t - g(z_t;\omega)\right]\nabla_\omega g(z_t;\omega) \text{ and } \nabla_c s^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right) = -1 + \frac{1}{\sigma_y^2}\left[y_t - g(z_t;\omega)\right]^2.$$

The term $\nabla_\omega g(z_t;\omega)$ can be evaluated analytically through back propagation.

D.2.2 Censored logarithmic score (CLS)

For the upper tail CLS the gradient can be written as

$$\nabla_\theta s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right) = \nabla_\theta s^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right) I(y_t \in A) + \frac{1}{P_\theta^{(t-1)}(y_q)} \nabla_\theta P_\theta^{(t-1)}(y_q) I(y_t \in A^c),$$

where $\nabla_\theta P_\theta^{(t-1)}(y_q) = \left(\nabla_\omega P_\theta^{(t-1)}(y_q)', \nabla_c P_\theta^{(t-1)}(y_q)\right)'$, $\nabla_\omega P_\theta^{(t-1)}(y_q) = -\phi_1\left(y_q; g(z_t;\omega), \sigma_y^2\right)\nabla_\omega g(z_t;\omega)$ and $\nabla_c P_\theta^{(t-1)}(y_q) = -\phi_1\left(y_q; g(z_t;\omega), \sigma_y^2\right)(y_q - g(z_t;\omega))$. The gradient for the lower tail CLS is

$$\nabla_\theta s^{\text{CLS-A}}\left(P_\theta^{(t-1)}, y_t\right) = \nabla_\theta s^{\text{LS}}\left(P_\theta^{(t-1)}, y_t\right) I(y_t \in A) - \frac{1}{1 - P_\theta^{(t-1)}(y_q)} \nabla_\theta P_\theta^{(t-1)}(y_q) I(y_t \in A^c).$$

D.2.3 Mean scaled interval score (MSIS)

The gradient of the MSIS can be written as

$$\nabla_\theta s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right) = \left(\nabla_\omega s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right), \nabla_c s_{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right)\right)',$$

with elements $\nabla_\omega s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right)$ and $\nabla_c s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right)$ defined as:

$$\nabla_\omega s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right) = -\left[1 - \frac{2}{1-q} I(y_t > u_t)\right] \frac{\partial u_t}{\partial \omega}' - \left[\frac{2}{1-q} I(y_t < l_t) - 1\right] \frac{\partial l_t}{\partial \omega}'$$
$$\nabla_c s^{\text{MSIS}}\left(P_\theta^{(t-1)}, y_t\right) = -\left[1 - \frac{2}{1-q} I(y_t > u_t)\right] \frac{\partial u_t}{\partial c} - \left[\frac{2}{1-q} I(y_t < l_t) - 1\right] \frac{\partial l_t}{\partial c}$$

The derivative $\frac{\partial u_t}{\partial \omega}$ can be evaluated by first noting that $\alpha_{u,t} = P_\theta^{(t-1)}(u_t)$. Then, using the triple product rule we know that

$$\frac{\partial u_t}{\partial \omega} = -\frac{\partial u_t}{\partial \alpha_{u,t}} \frac{\partial \alpha_{u,t}}{\partial \omega},$$

where the first term can be computed as $\frac{\partial u_t}{\partial \alpha_{u,t}} = \frac{1}{p(u_t|\mathcal{F}_{t-1},\theta)}$. The second term $\frac{\partial \alpha_{u,t}}{\partial \omega} = \nabla_\omega P_\theta^{(t-1)}(u_t)$ is evaluated as in the subsection above. The derivative $\frac{\partial l_t}{\partial \omega}$ is evaluated in the same fashion as described for $\frac{\partial u_t}{\partial \omega}$. The corresponding derivatives for parameter c can also be computed using similar steps.

E Proofs of Main Theoretical Results

The assumed predictive model is denoted as $P_\theta^{(n)}$. While the true model probability measure may depend on n, we suppress this dependence on n and simply refer to the true probability measure as P_0. If a result holds with P_0-probability converging to one, we abbreviate this as wpc1.

Proof of Lemma 6.1. The posterior density of θ given $y_{1:n}$ is $\pi_w(\theta|y_{1:n}) \propto \mathrm{d}\Pi(\theta)\exp\left[w_n\{S_n(\theta)\}\right]$,where we recall that, by hypothesis, $w_n \in (\underline{w}, \overline{w}) \subset (0,\infty)$ with probability one. Define the quasi-likelihood ratio

$$Z_n(\theta) := \exp\left[w_n\left\{S_n(\theta) - S_n(\theta_\star) - \frac{1}{2}\Delta_n'\Sigma_n^{-1}\Delta_n\right\}\right],$$

where, by Assumption 6.2 (iii), $\Sigma_n^{-1/2}\Delta_n = O_p(1)$. For $M > 0$, the posterior probability over $A_n := \{\theta : d(\theta,\theta_\star) > M\epsilon_n\}$ is

$$\Pi_w(A_n|y_{1:n}) = \frac{\int_{A_n}\mathrm{d}\Pi(\theta)\,Z_n(\theta)}{\int_\Theta \mathrm{d}\Pi(\theta)\,Z_n(\theta)} = \frac{N_n(A_n)}{D_n}.$$

From Lemma E.1, and for G_n as in that result, for a sequence $t_n \to 0$ with $t_n \le \epsilon_n^2$,

$$D_n \ge \frac{1}{2}\Pi(G_n)e^{-2nt_nw_n} \gtrsim e^{-C_1nt_n}e^{-2nt_nw_n} \gtrsim e^{-(C_1+2\overline{w})nt_n}$$

wpc1. Use the above equation to bound the posterior as

$$\Pi_w(A_n|y_{1:n}) \lesssim N_n(A_n)e^{(C_1+2\overline{w})nt_n}$$

(wpc1). To bound $N_n(A_n)$ from above, use Assumption 6.1 to deduce that, for any positive ϵ_n,

$$\begin{aligned}\sup_{d(\theta,\theta_\star)\ge\epsilon_n} n^{-1}\left\{S_n(\theta) - S_n(\theta_\star)\right\} \le& 2\sup_{\theta\in\Theta}\{S_n(\theta)/n - \mathcal{S}(\theta)\} + \sup_{d(\theta,\theta_\star)\ge\epsilon}\{\mathcal{S}(\theta) - \mathcal{S}(\theta_\star)\}\\ \le& o_p(1) - C_2\epsilon_n^2. \end{aligned} \tag{24}$$

For some $M > 0$, we decompose $N_n(A_n)$:

$$N_n(A_n) = N_n(A_n \cap \Theta_n(M)) + N_n(A_n \cap \Theta_n^c(M)) = N_n^{(1)} + N_n^{(2)},$$

where $\Theta_n(M)$ is a compact set of θ by construction and $\Theta_n^c(M)$ is its complement.

Then, $N_n^{(1)} \le e^{-C_2M\overline{w}n\epsilon_n^2}$ from (24) due to the compactness of $\Theta_n(M)$. For $N_n^{(2)}$, for a sequence $t_n \to 0$ with $t_n \le \epsilon_n^2$,

$$P_0(N_n^{(2)} > e^{-M\overline{w}nt_n/2})$$

$$\leq e^{M\overline{w}nt_n/2}\int_{\mathcal{Y}}\int_{A_n\cap\Theta_n^c(M)}\exp\left\{-\frac{w_n n}{2}(\theta-T_n)'[\Sigma_n/n](\theta-T_n)-R_n(\theta)\right\}\mathrm{d}\Pi(\theta)\mathrm{d}P_0(y_{1:n})$$
$$\leq e^{M\overline{\omega}nt_n/2}\Pi\left\{A_n\cap\Theta_n^c(M)\right\}\leq e^{-MC_2\overline{w}nt_n}$$

The first inequality comes from Markov inequality and the definition of $Z_n(\theta)$ and the second and the third inequalities are due to Assumptions 6.2 and 6.3 respectively.

Combining all the above, we can deduce that (wpc1)

$$\Pi_w(A_n|y_{1:n})\lesssim N_n(A_n)e^{(C_1+2\overline{w})nt_n}\lesssim e^{-\{MC_2\overline{w}-C_1-2\overline{w}\}n\epsilon_n^2},$$

since $t_n\leq\epsilon_n^2$. The right-hand-side vanishes for M large enough, and the result follows.

□

Remark E.1. As seen from the proof of Lemma 6.1, the stated result requires controlling the behavior of

$$\sup_{d(\theta,\theta_\star)\geq\epsilon}n^{-1}\left\{S_n(\theta)-S_n(\theta_\star)\right\}$$

for any $\epsilon>0$. When the parameter space is compact, this control can be guaranteed by controlling the bracketing entropy (see, e.g., van der Vaart and Wellner, 1996). If the parameter space is not compact, this control can be achieved by considering more stringent conditions on the prior than Assumption 6.3. In particular, results where the parameter space is not compact can be obtained by modifying Theorem 4 in Shen and Wasserman (2001) for our setting.

The following lemma bounds the denominator of the Gibbs posterior following a similar approach to Lemma 1 in Shen and Wasserman (2001) (see, also, Syring and Martin, 2020).

Lemma E.1. *Define* $T_n:=\theta_\star+\Sigma_n^{-1}\Delta_n$, *and* $G_n:=\{\theta\in\Theta:\frac{1}{2}(\theta-T_n)'\left[\Sigma_n/n\right](\theta-T_n)\leq t_n\}$. *For* $t_n\to 0$, *and* $t_n\leq\epsilon_n^2$. *Under the Assumptions of Lemma 6.1, wpc1*

$$D_n=\int_\Theta\mathrm{d}\Pi(\theta)Z_n(\theta)>\frac{1}{2}\Pi(G_n)e^{-2\overline{w}nt_n}.$$

Proof. Over G_n, use Assumption 6.2(i) to rewrite $\log\{Z_n(\theta)\}$ as

$$\begin{aligned}\log\{Z_n(\theta)\}&=\log\left[\exp\left\{w_n\left[S_n(\theta)-S_n(\theta_\star)-\frac{1}{2}\Delta_n'\Sigma_n^{-1}\Delta_n\right]\right\}\right]\\&=-\frac{w_n n}{2}(\theta-T_n)'\left[\Sigma_n/n\right](\theta-T_n)-R_n(\theta).\end{aligned}\tag{25}$$

Define the following sets: $\mathcal{C}_n:=\{(\theta,y_{1:n}):|R_n(\theta)|\geq nt_n\}$, $\mathcal{C}_n(\theta):=\{y_{1:n}:(\theta,y_{1:n})\in\mathcal{C}_n\}$, and $\mathcal{C}_n(y_{1:n}):=\{\theta:(\theta,y_{1:n})\in\mathcal{C}_n\}$. On the set $G_n\cap\mathcal{C}_n^c(y_{1:n})$, $Z_n(\theta)$ is bounded in probability, and we can

bound D_n as follows:

$$\begin{aligned} D_n &\geq \int_{G_n \cap \mathcal{C}_n^c(y_{1:n})} \exp\left\{-\frac{w_n n}{2}(\theta - T_n)'[\Sigma_n/n](\theta - T_n) - R_n(\theta)\right\} \mathrm{d}\Pi(\theta) \\ &\geq \exp\left(-2\overline{\omega} n t_n\right) \Pi\left\{G_n \cap \mathcal{C}_n^c(y_{1:n})\right\} \\ &\geq \left[\Pi(G_n) - \Pi\left\{G_n \cap \mathcal{C}_n(y_{1:n})\right\}\right] \exp\left(-2\overline{\omega} n t_n\right). \end{aligned} \tag{26}$$

From the bound in (26), the remainder follows almost identically to Lemma 1 in Shen and Wasserman (2001). In particular,

$$\begin{aligned} P_0\left\{\Pi\left[G_n \cap \mathcal{C}_n(y_{1:n})\right]\right\} &= \int_{\mathcal{Y}} \int_{\Theta} \mathbb{1}\left[G_n \cap \mathcal{C}_n(y_{1:n})\right] \mathrm{d}\Pi(\theta) \mathrm{d}P_0(y_{1:n}) \\ &= \int_{\mathcal{Y}} \int_{\Theta} \mathbb{1}\left[G_n\right] \mathbb{1}\left[\mathcal{C}_n(y_{1:n})\right] \mathrm{d}\Pi(\theta) \mathrm{d}P_0(y_{1:n}) \\ &\leq \frac{1}{n t_n} \Pi(G_n), \end{aligned}$$

where the last line follows from Markov's inequality and the definition of $\mathcal{C}_n(y_{1:n})$. Lastly, consider the probability of the set

$$\begin{aligned} P_0\left\{D_n \leq \frac{1}{2}\Pi(G_n) e^{-2\overline{w} n t_n}\right\} &\leq P_0\left(e^{-2\overline{w} n t_n}\left[\Pi(G_n) - \Pi\left\{G_n \cap \mathcal{C}_n(y_{1:n})\right\}\right] \leq \frac{1}{2}\Pi(G_n) e^{-2\overline{w} n t_n}\right) \\ &= P_0\left[\Pi\{G_n \cap \mathcal{C}_n(y_{1:n})\} \geq \frac{1}{2}\Pi(G_n)\right] \\ &\leq 2 P_0\left\{G_n \cap \mathcal{C}_n(y_{1:n})\right\} / \Pi(G_n) \\ &\leq \frac{2}{n t_n}. \end{aligned}$$

Hence, $P_0\left\{D_n \geq \frac{1}{2}\Pi(G_n) e^{-2\overline{w} n t_n}\right\} \geq 1 - 2(n t_n)^{-1}$ and for $n t_n \to \infty$, we have that

$$D_n \geq \frac{1}{2}\Pi(G_n) e^{-2\overline{w} n t_n} \tag{27}$$

except on sets of P_0-probability converging to zero. □

Proof of Theorem 6.1. The result begins with a similar approach to Corollary 2.1 of Zhang and Gao (2020), but requires particular deviations given the non-likelihood-based version of our problem.

First, we note that Lemma B.2 in the supplement to Zhang and Gao (2020) can be directly applied in this setting: for any $a > 0$ and $n \geq 1$, given observations $y_{1:n}$,

$$a\widehat{Q}\left\{d(\theta, \theta_\star)\right\} \leq D\left[\widehat{Q} \| \Pi_w\left(\cdot | y_{1:n}\right)\right] + \log \Pi_w\left(\exp\left\{a d(\theta, \theta_\star)\right\} | y_{1:n}\right).$$

Similarly, for any $Q \in \mathcal{Q}$, $\mathrm{D}\left[Q \| \Pi_w(\cdot|y_{1:n})\right] + o_p(1) \geq \mathrm{D}\left[\widehat{Q} \| \Pi_w(\cdot|y_{1:n})\right]$ by construction, so that

$$a\widehat{Q}\left\{d(\theta,\theta_\star)\right\} \leq \inf_{Q\in\mathcal{Q}} D\left[Q\|\Pi_w\left(\cdot|y_{1:n}\right)\right] + \log \Pi_w\left(\exp\left\{ad(\theta,\theta_\star)\right\}|y_{1:n}\right). \tag{28}$$

Taking expectations on both sides of (28) , and re-arranging terms, yields

$$\begin{aligned} P_0\widehat{Q}\left\{d(\theta,\theta_\star)\right\} &\leq \frac{1}{a} P_0\left\{\inf_{Q\in\mathcal{Q}} D\left[Q\|\Pi_w\left(\cdot|y_{1:n}\right)\right] + \log \Pi_w\left(\exp\left\{ad(\theta,\theta_\star)\right\}|y_{1:n}\right)\right\} \\ &\leq \frac{1}{a} n\kappa_n^2 + \frac{1}{a}\log\left[P_0\Pi_w\left(\exp\left\{ad(\theta,\theta_\star)\right\}|y_{1:n}\right)\right], \end{aligned} \tag{29}$$

where the second inequality follows from the definition of κ_n^2, and Jensen's inequality.

The second term in equation (29) is bounded by applying Lemma 6.1. In particular, for all $\alpha \geq \alpha_0 > 0$ and any $0 < a \leq \frac{1}{2}c_1$, $c_1 > 0$, by Lemma 6.1 (wpc1),

$$P_0\Pi_w\left(\exp\left\{ad(\theta,\theta_\star)\right\} > \alpha_0|y_{1:n}\right) \lesssim \exp\left(-ac_1\alpha\right).$$

Then, appealing to Lemma B.4 in the supplement to Zhang and Gao (2020) we obtain

$$P_0\Pi_w\left(\exp\left\{ad(\theta,\theta_\star)\right\}|y_{1:n}\right) \lesssim \exp\left(ac_1\alpha_0\right),$$

for all $a \leq \min\{c_1, 1\}$. Taking $a = \min\{c_1, 1\}$ and $\alpha_0 = n\epsilon_n^2$, and applying the above in equation (29), yields, for some $M > 0$,

$$P_0\widehat{Q}\left\{d(\theta,\theta_\star)\right\} \leq \frac{n\kappa_n^2 + \log\left(c_1 + e^{ac_1 n\epsilon_n^2}\right)}{a} \lesssim n\kappa_n^2 + n\epsilon_n^2 + o(1) \lesssim n\left(\kappa_n^2 + \epsilon_n^2\right) + o(1).$$

The stated result then follows from Markov's inequality,

$$P_0\widehat{Q}\left(d(\theta,\theta_\star) > M_n n\left(\epsilon_n^2 + \kappa_n^2\right)\right) \leq \frac{P_0\widehat{Q}\left\{d(\theta,\theta_\star)\right\}}{M_n n\left(\epsilon_n^2 + \kappa_n^2\right)} \lesssim M_n^{-1} \to 0.$$

□

Proof of Theorem 6.2. For probability measures P and Q, let $d_H(P,Q)$ denote the Hellinger distance between P and Q. Fix $\epsilon > 0$, and define the set $A_n(\epsilon) := \{P_\theta^{(n)} \in \mathcal{P},\ \theta \in \Theta : d_H^2(P_\theta^{(n)}, P_\star^{(n)}) \geq \epsilon\}$. By convexity of $d_H^2(\cdot,\cdot)$ in the first argument and Jensen's inequality,

$$\begin{aligned} d_H^2(P_Q^{(n)}, P_\star^{(n)}) &\leq \int_\Theta d_H^2(P_\theta^{(n)}, P_\star^{(n)})\mathrm{d}\widehat{Q}(\theta) \\ &= \int_{A_n^c(\epsilon^2)} d_H^2(P_\theta^{(n)}, P_\star^{(n)})\mathrm{d}\widehat{Q}(\theta) + \int_{A_n(\epsilon^2)} d_H^2(P_\theta^{(n)}, P_\star^{(n)})\mathrm{d}\widehat{Q}(\theta) \end{aligned}$$

$$\leq \epsilon^2 + \sqrt{2}\widehat{Q}\{A_n(\epsilon^2)\}.$$

By Theorem 6.1, for any $\epsilon > 0$, $\widehat{Q}\{A_n(\epsilon)\} = o_p(1)$, so that $d_H^2(P_Q^{(n)}, P_\star^{(n)}) \leq \epsilon^2$ (wpc1). Using the above and the definition $d_H(P_Q^{(n)}, P_\star^{(n)}) = \sqrt{d_H^2(P_Q^{(n)}, P_\star^{(n)})}$, we can conclude that, with probability converging to one,

$$d_H(P_Q^{(n)}, P_\star^{(n)}) \leq \epsilon.$$

Since the above holds for arbitrary $\epsilon > 0$, it also holds for $\epsilon \to 0$, which yields the first stated result.

Now, a similar argument to the above yields

$$d_H^2(P_{\Pi_w}^{(n)}, P_\star^{(n)}) \leq \int_\Theta d_H^2(P_\theta^{(n)}, P_\star^{(n)}) \mathrm{d}\Pi_w(\theta|\mathbf{y}) \leq \epsilon^2 + \sqrt{2}\Pi_w\{A_n(\epsilon^2)|\mathbf{y}\} \leq \epsilon^2,$$

where the last line follows from the convergence in Lemma 6.1 and holds wpc1. To obtain the second result, apply the triangle inequality, and the relationship between the Hellinger and it's square, to see that (wpc1)

$$\begin{aligned} d_H(P_Q^{(n)}, P_{\Pi_w}^{(n)}) &\leq d_H(P_Q^{(n)}, P_\star^{(n)}) + d_H(P_\theta^{(n)}, P_\star^{(n)}) \\ &= \sqrt{d_H^2(P_\theta^{(n)}, P_\star^{(n)})} + \sqrt{d_H^2(P_\theta^{(n)}, P_\star^{(n)})} \\ &\leq 2\epsilon. \end{aligned}$$

Since ϵ is arbitrary, the above holds for $\epsilon \to 0$, and we can conclude that $d_H(P_Q^{(n)}, P_{\Pi_w}^{(n)}) = o_p(1)$. The stated result then follows by applying the relationship between total variation and Hellinger distance: $d_{TV}(P_Q^{(n)}, P_{\Pi_w}^{(n)}) \leq \sqrt{2} d_H(P_Q^{(n)}, P_{\Pi_w}^{(n)})$.

□