The predictive view of Bayesian inference

Chung Hang Edwin Fong

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2021

Acknowledgements

I would like to start by thanking my supervisor, Chris Holmes, whose guidance supported me tremendously through my DPhil. His breath of knowledge and brilliant creativity continue to impress me, and I have greatly enjoyed my DPhil because of his supervision.

I would like to thank Stephen Walker, without whom Chapter 2 of this thesis would not have been possible. His technical expertise and clarity of vision have both guided and motivated me. I extend my thanks to my other collaborators, Brieuc Lehmann and Simon Lyddon, for the many interesting conversations. I would also like to thank my funding body, the Alan Turing Institute, for making all of this possible.

I have had the privilege of meeting many wonderful colleagues at the Department of Statistics at Oxford and the Alan Turing Institute. I am grateful for their friendship and solidarity throughout the DPhil experience.

Thank you to my friends in Hong Kong and London for their perspective during difficult times. A special thank you to my parents and brother for their unconditional support. Finally, I thank my partner, Chae Yeon, for always being there for me.

Abstract

This thesis considers the direct connection between the prediction of future observations and Bayesian inference. Using prediction as a guide, we generalize the Bayesian framework and introduce new methodologies for parameter inference and model selection which improve scalability and performance under model misspecification.

We begin with an introduction in Chapter 1 on methodologies for generalizing Bayesian inference, including the Bayesian bootstrap and general Bayesian updating. A summary of the current role of prediction in Bayesian inference is then provided.

In Chapter 2, we present a novel perspective on Bayesian inference which points to missing observations as the source of statistical uncertainty. We formally connect the Bayesian posterior on the parameter with the joint predictive distribution on the unobserved remainder of the population. Using this connection, we introduce the martingale posterior, which generalizes the Bayesian posterior. The martingale posterior only requires the elicitation of a predictive model, thus removing the need for the likelihood and prior. We discuss notions of predictive coherence and further introduce new nonparametric predictive models based on a bivariate copula update.

In Chapter 3, we investigate the computational benefits of Bayesian nonparametric learning using the Dirichlet process. This is closely related to the Bayesian bootstrap and is a special case of the martingale posterior. We find that the method is robust to model misspecification and is highly scalable due to its parallelizable nature. We further demonstrate that the posterior bootstrap, which approximately samples from the nonparametric posterior, is particularly proficient at targeting multimodal posteriors.

In Chapter 4, we introduce a notion of coherent model scoring under general Bayesian updating. We then explore the formal connection between the Bayesian marginal likelihood and cross-validation, and introduce a cumulative cross-validation score which alleviates some of the deficiencies of the marginal likelihood.

We provide a summary and discussion of future work in Chapter 5.

Contents

Chapter 1

Introduction

1.1 Motivation

With the advent of machine learning, the role of *prediction* in data analysis has grown tremendously. Powerful nonlinear models such as those based on random forests (Breiman, 2001a) or deep neural networks (LeCun et al., 2015) continue to dominate in tasks involving large, high-dimensional and complex datasets. These methods often contain the same familiar traits: make few assumptions about the data generating distribution, digest a large volume of data, and use the predictive accuracy of the model to evaluate and tune the algorithm. Optimization of parameters is often utilized, and the forecasting of new observations are of primary concern, leaving inference and uncertainty quantification in the backseat.

On the other side of the spectrum, we have the careful elicitation of the Bayesian method. One begins with the assumption of a statistical model for the data generating distribution, often parametric, followed by the elicitation of the prior distribution on the parameters indexing the statistical model. Quantifying uncertainty on the parameters of interest is available here, but the luxury of this inference comes at a cost. The traditional Bayesian must make the strong assumption of the model being well-specified, and the subjective prior distribution can be difficult to elicit in the absence of prior

information. Finally, the computational costs of inference can be demanding when Monte Carlo methods (Robert and Casella, 2013) must be invoked.

The above discussion is the distinction of the 'two cultures' of Breiman (2001b), namely the 'algorithmic modelling' and 'data modelling' cultures respectively. Prediction is an intuitive task as observations are grounded in reality, and it is easy to verify models through methods like cross-validation (Stone, 1974; Geisser, 1974). However, the inference and uncertainty quantification of unobserved parameters are often of paramount importance in many statistical problems where scientific understanding or interpretation is desired.

In this thesis, we will explore ways to bring the latter 'data modelling' school, specifically the Bayesian, closer to the first. Our hope is that the modern Bayesian can get the best of both worlds, resulting in uncertainty quantification on parameters of interest with weaker assumptions, scalable computation and strong predictive performance. At the core of our contribution is the fundamental and direct connection between Bayesian prediction and inference, which offers insight into existing methodologies and facilitates the development of novel ones. One common theme throughout this thesis is the acknowledgement of model misspecification, which guides us towards model-free or nonparametric methods that often have frequentist undertones. Computational scalability and practicality of the methodology will also be a focus throughout.

In the remainder of this chapter, we will briefly introduce the foundations of Bayesian inference and discuss the difficulties faced by Bayesians under modern settings. We consider the current generalizations of the Bayesian methods as a solution to the aforementioned problems, with a focus on the Bayesian bootstrap of Rubin (1981) and the general Bayesian updating framework of Bissiri et al. (2016). We then provide a review on the current research landscape of the role of prediction in Bayesian inference and model selection, followed by a summary of the contributions of the thesis.

1.2 Generalizations of Bayesian inference

1.2.1 Preliminaries

We begin by introducing the setting and notation. We assume that we have observed n observations, $y_{1:n} = (y_1, \ldots, y_n)$, where $y_i \in \mathcal{Y}$. Typical examples of the observation space are $\mathcal{Y} = \mathbb{R}^+$ when y_i is a survival time or $\mathcal{Y} = \{0, 1\}$ when y_i represents treatment success or failure. For the majority of this thesis, we will be considering the observations as realizations of random variables which are conditionally independent and identically distributed (i.i.d.), that is $Y_{1:n} \overset{\text{iid}}{\sim} F_0$ and we have observed $Y_{1:n} = y_{1:n}$. Here, F_0 is the unknown true sampling distribution that we condition on, and usually we assume F_0 has a density f_0. The parametric model being considered is then a family of probability densities which we write as $\mathcal{F}_\Omega = \{f_\theta(y); \theta \in \Omega \subseteq \mathbb{R}^p\}$, where Ω is the parameter space. Under the $\mathcal{M}$-closed world assumption (Bernardo and Smith, 2009), the traditional Bayesian believes their parametric model, $\mathcal{F}_\Omega$, is well-specified, in that there exists some true but unknown parameter $\theta_0 \in \Omega$ such that $f_{\theta_0} = f_0$. The Bayesian would then assign a prior distribution Π with density $\pi(\theta)$ on the unknown θ_0, and the posterior density is computed through

$$\pi(\theta \mid y_{1:n}) = \frac{\prod_{i=1}^n f_\theta(y_i)\, \pi(\theta)}{p(y_{1:n})}$$

where $p(y_{1:n}) = \int \prod_{i=1}^n f_\theta(y_i)\, \pi(\theta)\, d\theta$ is the marginal likelihood of the observations.

1.2.2 Challenges of Bayesian inference

As datasets and models grow in size and complexity, a fundamental issue which is exacerbated in Bayesian inference is model misspecification, where the true sampling distribution deviates significantly from the parametric model. A model is *misspecified* if there does not exist a $\theta_0 \in \Omega$ such that $F_{\theta_0} = F_0$, which is known as the $\mathcal{M}$-open setting (Bernardo and Smith, 2009). As "all models are wrong" (Box, 1976),

this leads us to question the validity of the inferences returned from traditional Bayesian methodology, which is formally conditioned on the model being well-specified. Furthermore, traditional Bayesian inference may perform suboptimally or poorly under misspecification; see for example Müller (2013) and Grünwald and van Ommen (2017).

Another complication that arises in Bayesian inference with growing datasets and more complex models is computational scalability. With intractable posteriors, the Bayesian may resort to Markov chain Monte Carlo (MCMC) methods for inference, but these methods are serial, computationally expensive, and often afflicted with convergence issues; see Green et al. (2015); Dunson and Johndrow (2020) for a review. A popular scalable alternative is posterior approximation with variational Bayes (VB) (Blei et al., 2017), but it is difficult to assess the level of approximation and VB can often provide unreliable uncertainty quantification (Giordano et al., 2015). For Bayesian model selection, the marginal likelihood (Jeffreys, 1961) is also notoriously difficult to estimate; see Robert and Wraith (2009) for a survey on Monte Carlo methods.

A possible solution to the above difficulties is to consider generalizations of Bayesian inference where we deviate from the usual prior-likelihood update. There have been a wide variety of suggested methodology in the literature, and Bochkina (2021) provides a thorough summary with a focus on asymptotic properties. For the purposes of this thesis, we will focus on two particular lines of work. The first direction we discuss stems from the *Bayesian bootstrap* of Rubin (1981), which has strong connections with the rich field of Bayesian nonparametrics (Ghosal, 2010). The Bayesian bootstrap will be a main focus in Chapter 3. The second direction we consider is the *general Bayesian updating* framework of Bissiri et al. (2016), which is based on arguments of *coherence*. We will then consider connections of this framework to model selection in Chapter 4.

1.2.3 Parameter of interest

Before discussing the various frameworks for generalizing Bayes, it is convenient to introduce the framework for statistical inference in the $\mathcal{M}$-open case. The following

discussion is likely to be quite familiar to a nonparametric frequentist due to the connections with M-estimators (van der Vaart, 1998, Chapter 5), though it may be less familiar for the traditional Bayesian. As we now assume that the observations arise from an unknown F_0 which no longer lies in a parametric model, we can define our parameter of interest as a functional $\theta(\cdot)$ of F_0, that is $\theta_0 = \theta(F_0)$. The form of this functional can be quite general, but in most cases in this thesis we will consider the form

$$\theta(F) = \arg\min_{\theta} \int \ell(\theta, y)\, dF(y) \tag{1.1}$$

where $\ell : \Omega \times \mathcal{Y} \to \mathbb{R}$ is a loss function which connects the observation to our parameter of interest. For example, the choice of $\ell(\theta, y) = (\theta - y)^2$ would return the mean of F as $\theta(F)$. One particular loss of interest is the negative log-likelihood of a parametric model, that is $\ell(\theta, y) = -\log f_\theta(y)$, which results in $\theta(F_0)$ minimizing the Kullback-Leibler (KL) divergence between F_0 and F_θ. We will see how one can use parametric models in the $\mathcal{M}$-open setting shortly.

1.2.4 Bayesian bootstrap

The Bayesian bootstrap (BB) was introduced by Rubin (1981), and was motivated as the "Bayesian analogue of the bootstrap". Assume we are interested in some functional of the unknown sampling distribution, $\theta_0 = \theta(F_0)$, as previously discussed. Given observations $y_{1:n}$, a sample $\theta^{(j)}$ from the BB posterior can be drawn as follows:

$$\begin{aligned} w_{1:n}^{(j)} &\sim \text{Dirichlet}(1, \ldots, 1) \\ F^{(j)} &= \sum_{i=1}^{n} w_i^{(j)}\, \delta_{y_i} \\ \theta^{(j)} &= \theta(F^{(j)}) \end{aligned}$$

where δ_{y_i} is the Dirac measure centred at y_i. Here, $F^{(j)}$ can be viewed as a random sample from the BB posterior, which induces a distribution on $\theta(F^{(j)})$. This is indeed

similar to Efron's bootstrap (Efron, 1979), where the weights $w_{1:n}$ would instead be drawn from a multinomial instead of the flat Dirichlet. The original motivation given by Rubin (1981) was to elicit a categorical likelihood with support on $y_{1:n}$ with respective probabilities $p_{1:n}$, and elicit the prior distribution $\pi(p_{1:n}) \propto \prod_{i=1}^{n} p_i^{-1}$. Although this choice of prior distribution is somewhat questionable, we will see now that the BB corresponds to the limiting case of a well-defined nonparametric prior.

1.2.4.1 Connections to Bayesian nonparametrics

As discussed earlier, the parametric Bayesian model family $\mathcal{F}_\Omega$ is unlikely to contain the true F_0, which leads to the ramifications of model misspecification. The solution offered by Bayesian nonparametrics is to expand the parametric model to one that is *infinite-dimensional* so that the assumption of F_0 being contained in the support of the Bayesian prior is much more reasonable. In this setting, it is more natural to consider the prior Π as being a distribution over probability measures on the sample space $\mathcal{Y}$.

One of the most well-known and utilized Bayesian nonparametric priors is the *Dirichlet process* of Ferguson (1973). The Dirichlet process (DP) distribution can be written as $\mathrm{DP}(\alpha, F_\pi)$, where α is the scalar concentration term and F_π is the base measure. These can be intuitively thought of as the inverse variance (or data pseudo-count) and the mean of the DP respectively. Draws of F from the DP are almost surely discrete, although the weak support of the DP is usually very large. We refer the reader to Ghosal (2010, Chapter 4) for more technical details. A key reason for the popularity of the DP is its conjugacy if the observations are i.i.d. conditional on F. Formally, if we have

$$[Y_{1:n} \mid F] \overset{\text{iid}}{\sim} F, \quad F \sim \mathrm{DP}(\alpha, F_\pi),$$

then the posterior distribution of F, given $Y_{1:n} = y_{1:n}$, is also a DP of the form

$$\mathrm{DP}\left(\alpha + n,\, \frac{\alpha}{\alpha + n} F_\pi + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{y_i}\right).$$

One interesting property of the DP is the induced predictive distribution of $Y_{n+1} \mid Y_{1:n} = y_{1:n}$ when F is marginalized out, which for $n > 1$ takes the form

$$[Y_{n+1} \mid Y_{1:n} = y_{1:n}] \sim \frac{\alpha}{\alpha + n} F_\pi + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{y_i} \tag{1.2}$$

and $Y_1 \sim F_\pi$. This is known as the *Pólya urn scheme* (Blackwell and MacQueen, 1973), where the predictive distribution is a weighted sum of F_π and the empirical distribution. We will investigate the implications of the Pólya urn scheme for predictive inference in detail in Chapter 2. Looking at the form of the posterior DP, we can consider taking $\alpha \to 0$ which corresponds to reducing the prior pseudo-count to 0. One can show that the posterior DP indeed converges weakly to the BB posterior distribution on F. This shows us as that the BB posterior distribution arises as the 'noninformative limit' of the posterior DP with the desirable property of being computationally simple.

When the BB is viewed as returning a posterior distribution on F which then induces a posterior on $\theta(F)$ defined through (1.1), Lyddon et al. (2019) terms this as the 'loss-likelihood bootstrap'. More generally, we term the process of returning a nonparametric posterior on F, and thus on $\theta(F)$, as *Bayesian nonparametric learning*. Bayesian nonparametric learning (NPL) was first coined in Lyddon et al. (2018), where the prior on F is the mixture of DPs (Antoniak, 1974) which centers the nonparametric model on a parametric Bayesian model. The mixture of DPs (MDP) is defined as

$$[F \mid \theta] \sim \mathrm{DP}(\alpha, F_\theta), \quad \theta \sim \pi(\theta)$$

where F_θ is a parametric centering measure with density f_θ. Here, f_θ and $\pi(\theta)$ are the parametric sampling density and prior density respectively of the traditional Bayesian model. The posterior on F, given i.i.d. observations $Y_{1:n} = y_{1:n}$ as before is also a

mixture of DPs:

$$[F \mid \theta, y_{1:n}] \sim \text{DP}\left(\alpha + n,\ \frac{\alpha}{\alpha+n} F_\theta + \frac{1}{\alpha+n}\sum_{i=1}^{n} \delta_{y_i}\right), \quad \theta \sim \pi(\theta \mid y_{1:n}),$$

where $\pi(\theta \mid y_{1:n})$ is the traditional Bayesian posterior density. The scalar α thus controls the strength of the centering on the traditional Bayesian model compared to the nonparametric model, with $\alpha \to \infty$ returning the parametric Bayesian posterior on $\theta(F)$, and $\alpha \to 0$ returning the BB. In Chapter 3, we will forgo the parametric model entirely and utilize the DP directly for NPL, which has significant computational advantages. In both cases, one can view NPL as a generalization of the BB through the incorporation of prior information on F, where taking $\alpha \to 0$ returns the BB in the noninformative limit.

1.2.4.2 Approximating the Bayesian posterior

The use of the BB and its extensions has also been explored in the parametric context in order to approximate the traditional parametric Bayesian posterior. The *weighted likelihood bootstrap* of Newton and Raftery (1994) was the first to consider utilizing the BB to approximate the parametric Bayesian posterior, which corresponds to setting $\ell(\theta, y) = -\log f_\theta(y)$. Here, the Dirichlet weights are viewed as perturbations of each datum when computing the maximum likelihood estimate. Extending this to incorporate the prior has been investigated by Newton et al. (2020) and Ng and Newton (2020) through the random weighting of 'prior' penalties of the log-likelihood such as the Lasso penalty (Tibshirani, 1996). Nie and Ročková (2020) instead suggests a random location shift of the spike-and-slab Lasso prior (Ročková and George, 2018). Pompe (2021) investigates the inclusion of prior information through a scaled prior penalty on the log-likelihood determined by Edgeworth expansions.

1.2.4.3 Properties of the Bayesian bootstrap

One of the biggest strengths of the BB is the ease of computation. Drawing Dirichlet weights is expedient, so the computational burden lies with the optimization of (1.1) which has seen many recent advances. The BB and related methods are parallelizable and scalable, and even deal well with multimodality as we will see in Chapter 3. In the context of NPL, drawing exact samples from the nonparametric posterior on F may be difficult but approximations exist which are operationally very similar to the BB. Muliere and Secchi (1996) introduced the proper Bayesian bootstrap approximation for the DP, which is based on resampling the base measure of the posterior DP, and Hjort (1991) discusses a similar method. The *posterior bootstrap* is introduced in Lyddon et al. (2018) and Chapter 3 for the MDP and DP respectively, which utilizes reweighted prior pseudo-samples in addition to the observations when optimizing. Other approximations for the DP can be found in Ghosal (2010, Chapter 4).

The BB also has many interesting theoretical properties which follow from the connections to the DP and Efron's bootstrap. Lo (1987) formally identified the first-order asymptotic equivalence between the BB and the frequentist bootstrap, as well as the connection to noninformative limit of the DP. Others that have investigated the asymptotic properties of the BB include Praestgaard and Wellner (1993); Gasparini (1995) and Choudhuri (1998). Similar to the DP, draws of F from the BB are always discrete and supported at $y_{1:n}$, and we consider settings where this is problematic in Chapter 2. Despite this discreteness, the BB tends to perform well under model misspecification. Specifically, Lyddon et al. (2019) show that the BB posterior of $\theta(F)$ as defined in (1.1), under regularity conditions on ℓ, is asymptotically normal with the sandwich covariance matrix. This differs from the asymptotic variance of the traditional Bayesian posterior, and this discrepancy is highlighted by Müller (2013) who shows that an artificial posterior with the sandwich covariance matrix gives a lower asymptotic frequentist risk than the traditional Bayesian posterior under model misspecification. This asymptotic robustness has been utilized in the work of Fushiki et al. (2005) and

Lyddon et al. (2019) to show that the BB is superior to traditional Bayes in predictive risk when the loss function is the log-likelihood and the model is misspecified, which we will also discuss briefly in Chapter 3. A comparison of predictive risk of the BB to tempered Bayesian posteriors is also given in Pompe (2021). In summary, the BB posterior is robust to model misspecification due to its nonparametric nature.

1.2.4.4 Applications and related work

Recently, the BB and NPL has seen some usage in the applied domain due to issues of scalability and model misspecification in traditional Bayesian inference. Hashimoto and Sugasawa (2020) use the BB within Gibbs sampling when utilizing robust alternatives to the likelihood, and Filipovic et al. (2021); Goncharov et al. (2021) utilize the posterior bootstrap in a medical imaging context. Knoblauch and Vomfell (2020) explores NPL with the total variation distance as the objective function, and Galvani et al. (2021) applies the proper Bayesian bootstrap to decision trees to generate an ensemble. Beyond the simple i.i.d. setting, Pompe (2021) considers hierarchical models, and Pompe and Jacob (2021) applies the posterior bootstrap to Bayesian cut models. In the survival setting, Hjort (1991) and Lo (1993) consider the BB under right-censoring, and Arfè and Muliere (2020) extend this to include prior information in the BB through the beta-Stacy process of Walker and Muliere (1997). Beyond the BB, there are a few related nonparametric generalizations of Bayesian inference. One interesting direction is the connection of the BB to the Bayesian empirical likelihood method of Lazar (2003), which is exploited by Kim and Lee (2003) to extend the BB to the Cox model in survival analysis; see Lazar (2021) for a review of Bayesian empirical likelihood methods. Another direction is the BayesBag framework of Bühlmann (2014), which averages multiple traditional posterior distributions fit to bootstrapped copies of the observations. Huggins and Miller (2019, 2020) show that BayesBag is more robust than traditional Bayesian inference for inference and model selection.

1.2.5 General Bayesian updating

The general Bayesian update of Bissiri et al. (2016) is a model-free framework for posterior inference under $\mathcal{M}$-open. Instead of expanding the model space like in the nonparametric approach, we do away with the notion of a statistical model entirely. The starting point is again the parameter of interest $\theta_0 = \theta(F_0)$ in (1.1) which minimizes the expected loss function $\ell(\theta, y)$, where the loss function now takes the place of the statistical model. The decision maker has a prior distribution Π_{G} with density $\pi_{\mathrm{G}}(\theta)$ on the value on the parameter, and given an observation $Y = y$, they would like to compute the update

$$\pi_{\mathrm{G}}(\theta \mid y) = \psi\{\ell(\theta, y), \pi_{\mathrm{G}}(\theta)\},$$

where $\pi_{\mathrm{G}}(\theta \mid y)$ is the density of the general Bayesian posterior distribution $\Pi_{\mathrm{G}}(\cdot \mid y)$. In the absence of the likelihood, the key to determining the form of ψ is a *coherence* condition.

Coherence of subjective beliefs in the traditional Bayesian method is well-understood, and is often a driving motivation behind the Bayesian philosophy. Intuitively, coherence is the notion that subjective beliefs are internally consistent, that is no paradoxes arise. There are multiple versions of coherence, some based on decision-making axioms (Bernardo and Smith, 2009) and others based on betting and Dutch books (e.g. Heath and Sudderth (1978)). A summary of the various notions of coherence can be found in Robins and Wasserman (2000) and Robert (2007).

We now discuss the coherence condition considered in Bissiri et al. (2016). For any two data points (y_1, y_2), the update function ψ is coherent if it satisfies

$$\psi[\ell(\theta, y_2), \psi\{\ell(\theta, y_1), \pi_{\mathrm{G}}(\theta)\}] = \psi\{\ell(\theta, y_1) + \ell(\theta, y_2), \pi_{\mathrm{G}}(\theta)\}. \tag{1.3}$$

In summary, the above condition ensures the general posterior $\pi_{\mathrm{G}}(\theta \mid y_1, y_2)$ should not change depending on if we update with y_1 followed by y_2, or if we update simultaneously with both (y_1, y_2). The additivity of the loss also ensures that we are invariant to the

ordering of the data points, which has connections to exchangeability.

1.2.5.1 The general Bayesian posterior

To derive the form of the general posterior, we can utilize decision theory by treating $\pi_{\mathrm{G}}(\theta \mid y)$ as an optimal 'action'. For ease of exposition, we assume throughout that the appropriate densities (with respect to the Lebesgue measure) exist. Specifically, the general posterior is the distribution Q which minimizes the loss function below,

$$L_{\mathrm{G}}(q; \pi_{\mathrm{G}}, y) = L_{\mathrm{data}}(q, y) + L_{\mathrm{prior}}(q, \pi_{\mathrm{G}}), \tag{1.4}$$

where q is the density of Q, and Q lies in the space of probability distributions over θ that are absolutely continuous with respect to the prior Π_{G}. Here, L_{G} consists of a term L_{data} which ensures the general posterior fits the data well, and a prior regularization term L_{prior} which ensures the general posterior and prior are not too far apart. The solution $\hat{q} = \arg\min_q L_{\mathrm{G}}(q; \pi_{\mathrm{G}}, y)$ is then the general posterior density. A natural choice for the first term is the expected loss,

$$L_{\mathrm{data}}(q, y) = w \int \ell(\theta, y)\, q(\theta)\, d\theta,$$

where w is a positive scalar. We include the term w as there are no constraints on the scaling of ℓ, so w controls the relative weighting of the data and prior. Bissiri et al. (2016) then shows that for the coherency condition (1.3) to hold, and under some other natural assumptions, the prior regularization term must take the form of the KL divergence,

$$L_{\mathrm{prior}}(q, \pi_{\mathrm{G}}) = d_{\mathrm{KL}}(q, \pi_{\mathrm{G}}) := \int \log \frac{q(\theta)}{\pi_{\mathrm{G}}(\theta)}\, q(\theta)\, d\theta.$$

Putting this together, it is relatively straightforward to verify that under regularity

conditions, minimizing $L_{\mathrm{G}}(q; \pi_{\mathrm{G}}, y)$ returns us the general posterior density of the form

$$\pi_{\mathrm{G}}(\theta \mid y) \propto \exp\{-w\ell(\theta, y)\}\, \pi_{\mathrm{G}}(\theta). \tag{1.5}$$

It is also straightforward to show that for multiple data points $y_{1:n}$, the loss is additive so we can substitute $\ell(\theta, y)$ with $\sum_{i=1}^{n} \ell(\theta, y_i)$ in the above. In summary, the general Bayesian posterior returns us uncertainty without the need for a statistical model, whilst maintaining coherence. The form (1.5) has arisen under different names before in the literature, such as the quasi-posterior (Chernozhukov and Hong, 2003) or the Gibbs posterior (Jiang and Tanner, 2008); see Bissiri et al. (2016) and Lyddon (2018) for more details. Note that $w = 1$ and $\ell(\theta, y) = -\log f_\theta(y)$ corresponds to the model f_θ being well-specified, as we recover the traditional Bayesian posterior.

1.2.5.2 Calibrating the loss function

The scaling term w in (1.5) determines the weighting of the loss to the prior, which plays an important role as it controls the variance of the general Bayesian posterior. In the case where $\ell(\theta, y) = -\log f_\theta(y)$, the general Bayesian posterior is equivalent to the *tempering* of the likelihood, that is raising it to a fractional power when $0 < w < 1$. This can be viewed as robustifying Bayesian inference against model misspecification, as we are downweighting the contribution of the data. Some examples of tempering include the SafeBayes framework (Grünwald and van Ommen, 2017), fractional posteriors (Bhattacharya et al., 2019) and coarsening (Miller and Dunson, 2019).

In the general Bayesian context, there have been multiple directions to select the loss scale w, such as through expected information gain (Holmes and Walker, 2017) or the calibration of posterior credible intervals (Syring and Martin, 2019). More examples can be found in Syring (2017) and Lyddon (2018). One particular scheme of interest for this thesis is that of Lyddon et al. (2019), which connects the general Bayesian framework to the Bayesian bootstrap (BB) of Section 1.2.4. Lyddon et al.

(2019) suggests setting w by matching the information in the asymptotic posterior for the BB and general Bayes, in particular using the Fisher information number which depends on the trace of the asymptotic variance in this case. In this way, we combine the robustness of the BB with the incorporation of prior information of the general Bayesian posterior. If θ is 1-dimensional, this scheme simply corresponds to matching the asymptotic variances. However, in the multivariate case, the scalar w is somewhat restrictive as we cannot adjust the covariance structure of the general Bayesian posterior to match the robust sandwich covariance matrix.

1.2.5.3 Extensions and related work

There have been some interesting extensions of the general Bayesian updating framework of Bissiri et al. (2016). For robust $\mathcal{M}$-open inference with a parametric model, Jewson et al. (2018) considers using the general Bayesian framework with different choices of the loss function ℓ that correspond to alternative divergences between f_0 and f_θ, instead of just the KL divergence with $\ell(\theta, y) = -\log f_\theta(y)$. Knoblauch et al. (2019) treat (1.4) as the starting point of inference, and consider more general choices of L_{prior} beyond the KL divergence. They introduce the generalized variational inference framework, which constrains the optimization of $L_{\text{G}}(q; \pi_{\text{G}}, y)$ such that q lies in some variational family. Jewson and Rossell (2021) and Matsubara et al. (2021) utilize the general Bayesian framework to tackle intractable likelihoods using the Hyvärinen score (Hyvärinen and Dayan, 2005) and the Stein discrepancy (Gorham and Mackey, 2015) respectively to specify the loss function. Beyond the general Bayesian framework, other methods consider the replacement of the likelihood with more tractable alternatives. Examples include the pseudo/composite likelihood of Besag (1974) and Dryden et al. (2002) which replace an intractable likelihood with an approximation; see Park and Haran (2018) for a review. Another example is the coreset of Huggins et al. (2016) and Campbell and Broderick (2019), which replaces the log-likelihood sum of the full dataset with a weighted log-likelihood sum of a smaller subset to ease computation.

1.3 Prediction in Bayesian inference

In this previous section, we have focused mainly on obtaining (and generalizing) the Bayesian posterior distribution on parameters for inference. However, there are many settings, such as those encountered in econometrics or biostatistics, where the forecasting of future observables may be of direct interest. A useful object for this task is the 1-step ahead *posterior predictive* density, defined as

$$p(y_{n+1} \mid y_{1:n}) = \int f_\theta(y_{n+1})\, \pi(\theta \mid y_{1:n})\, d\theta.$$

The above allows us to assign the subjective probability of a new observation Y_{n+1}, incorporating the statistical model, prior knowledge, and observations $Y_{1:n} = y_{1:n}$ in a natural way. A similar object is the marginal likelihood, defined as

$$p(y_{1:n}) = \int \prod_{i=1}^{n} f_\theta(y_i)\, \pi(\theta)\, d\theta.$$

This is sometimes denoted as the *prior predictive* density, as it is the subjective probability of the observations under the prior. In this section, we exhibit the foundational and practical role of the prior and posterior predictive distributions within the Bayesian framework.

1.3.1 Predictive inference

Although the starting point of Bayesian inference is often the prior distribution on the unknown parameter θ, the role of observables in Bayesian inference has a rich history. The most well-known proponent of the importance of observables is de Finetti, who pioneered, among many other things, the concept of *exchangeability* in his seminal publication of de Finetti (1937). To begin, we define an infinite sequence of binary random variables, $(Y_1, Y_2, \ldots)$, where $Y_i \in \{0, 1\}$. This infinite sequence is termed *exchangeable* if the joint probability mass function p of the sequence $Y_{1:N}$ is invariant

to permutations of $Y_{1:N}$ for every N, that is the following holds,

$$p(y_1, \ldots, y_N) = p(y_{\pi(1)}, \ldots, y_{\pi(N)})$$

for all permutations π of the indices $\{1, \ldots, N\}$, for every N. De Finetti's representation theorem ensures us that if the infinite sequence is exchangeable, there exists a cumulative distribution function Π such that the joint probability mass function p of any finite sequence $Y_{1:N}$ takes the form

$$p(y_{1:N}) = \int \prod_{i=1}^{N} \theta^{y_i} (1-\theta)^{1-y_i} \, d\Pi(\theta).$$

Furthermore, there exists some random variable Θ taking values in Ω which satisfies

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} Y_i = \Theta$$

almost surely, where Θ is distributed according to Π, and we have

$$\Pi(\theta) = \lim_{N\to\infty} P\left(\frac{1}{N} \sum_{i=1}^{N} Y_i \leq \theta \right).$$

The observations $Y_{1:N}$ can thus be thought of as being i.i.d. from the Bernoulli distribution conditional on some random Θ. As a result, the representation theorem provides justification for the traditional likelihood-prior specification of the Bayesian model when one subjectively views the observables as exchangeable. Roberts (1965) provides an exposition of Bayesian inference and prediction, and summarizes de Finetti's viewpoint succinctly: the parameter Θ can be viewed as an intermediate tool for obtaining the predictive distribution, which is the primary object of interest that is cumbersome to specify directly. On a more fundamental level, we can interpret the prior distribution Π as "beliefs about the limiting relative frequency of 1's" as stated by Bernardo and Smith (2009, Chapter 4.2). This connects the uncertainty in the "lurking"

parameter Θ, which de Finetti criticized as lacking objective meaning, to the prior predictive uncertainty in the observables, which are grounded in reality. In Chapter 2, we will investigate this connection in significant detail, and will instead utilize the work of Doob (1949) to generalize beyond the exchangeable Bayesian model. There have been many extensions to de Finetti's representation theorem, such as that of Hewitt and Savage (1955). We refer the reader to Bernardo and Smith (2009, Chapter 4) and references therein for more details.

Many others have adopted this predictive view, including the seminal work of Dawid (1984). Dawid argues that the main goal of statistical inference is to make forecasts on future observables, and derives a framework adhering to this principle called the *prequential* framework, which is a portmanteau of 'probability/predictive' and 'sequential'. Under the prequential framework, the statistical model is replaced with the prequential forecasting system (PFS), which is a sequence of 1-step ahead predictive distributions for Y_{n+1} given $Y_{1:n}$, for each n. Properties of the PFS are entirely with respect to the forecasts that are made, including predictive notions of consistency and efficiency. Connections of the prequential framework to model selection will be discussed in Section 1.3.2. Dawid (1985) explores the role of extrinsic and intrinsic parameters, where extrinsic parameters have a physical meaning beyond indexing the parametric model and intrinsic does not. Dawid also considers parameters as limiting functions of the observables utilizing the notions of exchangeability and repetitive structures, which has connection to the work of Lauritzen (1988).

Another proponent of the emphasis on prediction is the work of Geisser; a thorough summary of Bayesian predictive inference is summarized in Geisser (1993). Geisser (1982, 1983) emphasizes the role of observables in statistical inference, in particular the unseen, finite collection $Y_{n+1:N}$. For an action a, Geisser introduces the loss function $\ell(a, Y_{n+1:N})$, which takes as input $Y_{n+1:N}$ instead of the usual θ in traditional Bayesian

decision theory. He further argues that the joint posterior predictive density,

$$p(y_{n+1:N} \mid y_{1:n}) = \int \prod_{i=n+1}^{N} f_\theta(y_i)\, \pi(\theta \mid y_{1:n})\, d\theta,$$

should be the basis of all inferences on $Y_{n+1:N}$ and any functions thereof. Taking a function of $Y_{n+1:N}$ has a similar flavour to the 'limiting relative frequency' of de Finetti in a finite population context, which we formalize in Chapter 2. A similar connection of Bayesian prediction to finite populations was also explored in Roberts (1965) and Ericson (1969). An interesting connection to the Bayesian bootstrap (BB) of the previous section is the finite BB of Ghosh and Meeden (1997) and Lo (1988), which extends the BB to finite populations by defining the joint predictive $p(y_{n+1:N} \mid y_{1:n})$ via the Pólya urn scheme of (1.2) with $\alpha = 0$.

The predictive distribution also plays an important role in the Bayesian nonparametrics literature. A recent review of de Finetti's view on prediction and its connections to Bayesian nonparametrics is given in Fortini and Petrone (2014). Fortini et al. (2000) and Fortini and Petrone (2012) consider the construction of exchangeable Bayesian parametric and nonparametric models respectively through the specification of 1-step ahead predictive distributions, which have connections to de Finetti's representation theorem. We will see a similar approach in Chapter 2, but we extend Bayesian inference by utilizing the *conditionally identically distributed* (c.i.d.) setting of Berti et al. (2004), which is a weakening of exchangeability. More recently, the works of Hahn (2015); Hahn et al. (2018) and Berti et al. (2020, 2021) also consider the construction of nonparametric c.i.d. Bayesian models - we leave the details of c.i.d. sequences and related works for Chapter 2.

1.3.2 Predictive model evaluation and selection

1.3.2.1 Bayes factors

The prior and posterior predictive also play a crucial role in Bayesian model evaluation and selection. For two competing Bayesian models, $\mathcal{M}_1$ and $\mathcal{M}_2$, we write

$$p_{\mathcal{M}_j}(y_{1:n}) = \int \prod_{i=1}^{n} f_{\theta_j}(y_i)\, \pi_j(\theta_j)\, d\theta_j$$

where $\theta_j \in \Omega_j$ for $j \in \{1, 2\}$. In the $\mathcal{M}$-closed framework, we write the hypotheses as

$$\mathcal{M}_j : \quad f_0 \text{ lies in } \{f_{\theta_j} : \theta_j \in \Omega_j\},$$

that is either $\mathcal{M}_1$ or $\mathcal{M}_2$ contains the true sampling density f_0 from which $Y_{1:n}$ arose. The Bayesian would then elicit a prior on the model space, that is $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$ such that $p(\mathcal{M}_1) + p(\mathcal{M}_2) = 1$. The Bayes factor (Jeffreys, 1961; Kass and Raftery, 1995) of $\mathcal{M}_1$ to $\mathcal{M}_2$ computes the ratio of posterior to prior odds, returning

$$B_{12} = \frac{p_{\mathcal{M}_1}(y_{1:n})}{p_{\mathcal{M}_2}(y_{1:n})}.$$

Intuitively, B_{12} measures the evidence for $\mathcal{M}_1$ over $\mathcal{M}_2$, where the marginal likelihood $p_{\mathcal{M}_i}$ is the probability of observing the data under the respective Bayesian model. However, computing the marginal likelihood is a challenging Monte Carlo problem (Robert and Wraith, 2009), and its value is sensitive to noninformative priors; see Lindley's paradox (Lindley, 1957). More details on Bayes factors and model selection can be found in Chapters 5 and 7 of Robert (2007) as well as Chapter 4 of this thesis.

1.3.2.2 Cross-validation

In the $\mathcal{M}$-open setting, the evaluation of the predictive performance of Bayesian models on future observations is particularly intuitive, in particular using the posterior

predictive density $p(y \mid y_{1:n})$. This is discussed in general for example in Geisser (1993, Chapter 4), Gelman et al. (2013, Chapter 7) and Vehtari and Ojanen (2012), which we summarize now. In an ideal world, we would have access to the sampling distribution f_0 of a future observation Y_{n+1}, allowing us to evaluate the predictive performance of the Bayesian model. In particular, the use of the log posterior predictive density as the score would yield the quantity which Gelman et al. (2013) terms the expected log predictive density for a new data point (elpd):

$$S_0(y_{1:n}) = \int \log p(y \mid y_{1:n})\, f_0(y)\, dy.$$

This quantity is particularly interesting as the elpd is equal to $-d_{\mathrm{KL}}\{f_0, p(\cdot \mid y_{1:n})\}$ up to a constant, where we have treated the posterior predictive density as an approximation of the unknown f_0 (Geisser, 1971). More general scoring functions beyond the log score for probabilistic forecasts is considered in Gneiting and Raftery (2007).

In practice, we do not have access to the f_0 or future observations, and so we may be tempted to approximate the elpd with

$$S_{\mathrm{train}}(y_{1:n}) = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i \mid y_{1:n}).$$

Unfortunately, S_{train} is a biased estimate of the elpd, as the test datum y_i is not independent from the dataset $y_{1:n}$. One solution to decrease this bias is to partition the observations into a training and held-out test set. In particular, the approach of *cross-validation* has seen immense popularity, which was introduced generally in Stone (1974); Geisser (1974) and specifically for Bayesian models in Geisser and Eddy (1979). Once again using the log posterior predictive density as the score, the leave-p-out cross-validation score is defined as

$$S_{\mathrm{CV}}(y_{1:n}; p) = \frac{1}{\binom{n}{p}} \sum_{t=1}^{\binom{n}{p}} \frac{1}{p} \sum_{j=1}^{p} \log p\left(\tilde{y}_j^{(t)} \mid y_{1:n-p}^{(t)}\right)$$

where $\tilde{y}^{(t)}_{1:p}$ is the held-out test set of size p and $y^{(t)}_{1:n-p}$ is the corresponding training set so that $y_{1:n} = \left\{\tilde{y}^{(t)}, y^{(t)}\right\}$. The superscript t indicates each of the n-choose-p train-test splits, but this can be a hefty number so in practice the above is approximated with Monte Carlo. A popular alternative to the above is *k-fold cross-validation*, where the dataset is divided into k separate subsets with each subset acting as the held-out test set with the remaining $k-1$ subsets as training data, so we only carry out k model fits. With cross-validation however, the training sets are of smaller size than the full dataset which leads to another source of bias; a remedy is discussed in Burman (1989). Furthermore, the need to fit the model k times can be quite expensive.

When $p = 1$, we recover the leave-one-out (LOO) cross-validation score,

$$S_{\mathrm{LOO}}(y_{1:n}) := S_{\mathrm{CV}}(y_{1:n}; 1) = \frac{1}{n}\sum_{i=1}^{n} \log p\left(y_i \mid y_{-i}\right),$$

where $y_{-i} = y_{1:n} \setminus y_i$. This is a particularly interesting case as the bias induced by utilizing y_{-i} instead of $y_{1:n}$ is relatively small if n is large. Furthermore, the computation of the LOO score for Bayesian inference is particularly expedient, as importance sampling can be used to refit the Bayesian model to y_{-i} given samples from the full posterior distribution (Gelfand, 1996), so only a single MCMC run is required. Stabilizing the computation of S_{LOO} for Bayesian models using Pareto-smoothing is investigated in Vehtari et al. (2015, 2017), and scaling to large datasets using approximations is discussed in Magnusson et al. (2019). An alternative to cross-validation is to subtract a bias term from S_{train}, which results in the commonly used information criteria; Vehtari et al. (2017) provides a comparison of LOO to the widely applicable information criterion (WAIC). A thorough review of information criteria and the role of prediction in model selection in general can be found in Vehtari and Lampinen (2002).

There are some interesting connections between the use of the posterior predictive density and the marginal likelihood for model selection. For example, the prequential

score of Dawid (1984) with the log posterior predictive density takes on the form

$$S_{\text{preq}} = \sum_{i=1}^{n} \log p(y_i \mid y_{1:i-1})$$

where we write $p(y_1 \mid y_{1:0}) = p(y_1)$. Intuitively, the prequential score judges a predictive model based on how well it has sequentially predicted the sequence of observations. From the chain rule, we clearly have that $S_{\text{preq}} = \log p(y_{1:n})$, which corresponds to the Bayesian marginal likelihood of the observations. When there is no natural ordering of the data, Gneiting and Raftery (2007) consider the connection of the marginal likelihood and 'random-fold' cross-validation; we extend and investigate this result for exchangeable Bayesian models in more detail in Chapter 4.

1.3.2.3 Predictive model checks

The posterior predictive is also useful for model checking, which ensures that the inferences produced by the Bayesian model are reasonable. Beyond model evaluation based on cross-validation, one can simulate from the posterior predictive density, $Y^* \sim p(y \mid y_{1:n})$, and compare the induced distribution of a scalar test statistic $T(Y^*)$ to the observed statistic $T(y_{1:n})$ (Rubin, 1984). In particular, Gelman et al. (1996) terms the quantity

$$p_B = P(T(Y^*) \geq T(y_{1:n}) \mid y_{1:n})$$

as the posterior predictive p-value, where an extreme value close to 0 or 1 would arise suspicion in the model's adequacy. Gelman et al. (1996) also generalizes the above test statistic to be both a function of Y^* and θ. More details on the selection of T and interpretation of p_B can be found in Gelman et al. (2013, Chapter 6). The above model checks also have close connections to the notion of *calibration*, which considers the frequentist coverage properties of Bayesian probability statements. Calibration is considered in works such as Dawid (1982); Rubin (1984) for Bayesian models and more generally in Gneiting et al. (2007).

1.4 Thesis outline

This thesis investigates novel generalizations of the traditional Bayesian framework under model misspecification which rely on the model-agnostic nature of *prediction* as the foundation. Themes discussed earlier such as nonparametric methods and coherence play a crucial role in the development of the methodology.

Chapter 2 In this chapter, we introduce a novel perspective on Bayesian inference based on the idea of predictive imputation, which does not require the notion of model well-specification. We argue that the Bayesian fundamentally elicits a joint predictive distribution on the missing data given what has been observed. In the conditionally i.i.d. case, we thus specify the predictive distribution on the remainder of the population $Y_{n+1:\infty}$ conditioned on the observed $Y_{1:n}$, which in turn induces a distribution on any statistic of interest. Doob's theorem ensures that this distribution on the parameter is indeed the usual posterior when using the Bayesian posterior predictive. Given this connection, we are now free to relax the construction of the predictive distribution from the traditional likelihood and prior, instead relying on martingales to enforce a notion of predictive coherence. This gives us the *martingale posterior distribution*, which we can sample through an imputation scheme we call *predictive resampling*. In particular, we utilize a bivariate copula update, and extend recent copula-based methods to multivariate density estimation, regression and classification. This framework also has strong connections to the Bayesian bootstrap, which ties together with Chapter 3.

Chapter 3 *Bayesian nonparametric learning* (NPL) (Lyddon et al., 2018) is a method for posterior inference on model parameters for conditionally i.i.d. data without assuming the model is well-specified, and has desirable practical and theoretical properties. In this chapter, we investigate NPL using the Dirichlet process (DP) prior, which is a special case of the martingale posterior of Chapter 2. By utilizing a DP prior on the sampling distribution and a parameter defined by minimizing an expected loss,

we can sample from the NPL posterior through the *posterior bootstrap*. This involves optimizations of randomized objective functions with prior pseudo-data in a similar fashion to the Bayesian bootstrap (Rubin, 1981). As we have decoupled the NPL posterior from the Bayesian posterior, this sampling scheme offers massive scalability with parallel computing and allows for accurate sampling from high-dimensional and multimodal posteriors.

Chapter 4 The marginal likelihood is a tool in Bayesian model selection which measures the joint probability of the data conditioned on model correctness. On the other hand, cross-validation can be applied to evaluate the predictive ability of a Bayesian model (Vehtari and Lampinen, 2002), which makes no assumption about the model being well-specified. In this chapter, we show the equivalence of the marginal likelihood to a *cumulative cross-validation* (CCV) score with the log posterior predictive as the scoring rule, leading to a model misspecified interpretation of the marginal likelihood. This highlights the sensitivity of the marginal likelihood to the prior, and allows us to introduce a score that lies in between the marginal likelihood and leave-one-out cross-validation. We further show that the log posterior predictive is uniquely coherent under data exchangeability in a general Bayesian framework (Bissiri et al., 2016), where coherence implies invariance to data permutation and is required for the equivalence result.

This integrated thesis consists of 3 papers, followed by a discussion in Chapter 5. Each chapter contains the article in full, beginning with an additional preamble and ending with a statement of the author contributions. Technical background details are provided in each respective chapter. Chapter 2 is currently under review as a journal submission (Fong et al., 2021), whereas Chapters 3 and 4 have been published in the *36th International Conference on Machine Learning* (Fong et al., 2019) and *Biometrika* (Fong and Holmes, 2020) respectively.

Chapter 2

Martingale posterior distributions

2.1 Preamble

This chapter contains the most recent work of this thesis, and has the widest scope of the 3 main chapters. Although the methodology in Chapter 3 came first chronologically, it is a special case of the martingale posterior framework that will be introduced in this chapter, so we exhibit this more general framework first. The martingale posterior arose from an attempt to extend the Bayesian bootstrap (Rubin, 1981) to more general settings. Extending the BB under the 'randomized objective function' view (Lyddon et al., 2018) was met with resistance, but the connection between Bayesian prediction and inference allowed us to interpret the Bayesian bootstrap in a different light. This insight developed into the novel framework for predictive Bayesian inference which is discussed in this chapter.

The content of this chapter is a self-contained manuscript with its supplementary material. Details of the preprint, which is currently under review, are given below:

> Fong, E., Holmes, C., and Walker, S. G. (2021). Martingale posterior distributions. *arXiv preprint arXiv:2103.15671.*

A statement of authorship is provided at the end of this chapter.

2.2 Abstract

The prior distribution is the usual starting point for Bayesian uncertainty. In this paper, we present a different perspective which focuses on missing observations as the source of statistical uncertainty, with the parameter of interest being known precisely given the entire population. We argue that the foundation of Bayesian inference is to assign a distribution on missing observations conditional on what has been observed. In the i.i.d. setting with an observed sample of size n, the Bayesian would thus assign a predictive distribution on the missing $Y_{n+1:\infty}$ conditional on $Y_{1:n}$, which then induces a distribution on the parameter. We utilize Doob's theorem, which relies on martingales, to show that choosing the Bayesian predictive distribution returns the conventional posterior as the distribution of the parameter. Taking this as our cue, we relax the predictive machine, avoiding the need for the predictive to be derived solely from the usual prior to posterior to predictive density formula. We introduce the *martingale posterior distribution*, which returns Bayesian uncertainty on any statistic via the direct specification of the joint predictive. To that end, we introduce new predictive methodologies for multivariate density estimation, regression and classification that build upon recent work on bivariate copulas.

2.3 Introduction

Statistical uncertainty in a parameter of interest arises due to missing observations. If a complete population is observed, then the parameter of interest can be assumed to be known precisely. In this paper, we argue that the Bayesian accounts for this uncertainty by constructing a distribution on the missing observations conditional on what has been observed. This in turn induces a distribution on the parameter given the observed data, which we will see is the posterior distribution. In this work, we will describe and generalize this framework in detail for the case where the observations

are independent and identically distributed (i.i.d.), and we will also briefly consider other data structures.

In the i.i.d. case, given $Y_{1:n} \overset{\text{iid}}{\sim} F_0$ where F_0 is the unknown true sampling distribution, the missing observations are the remaining $Y_{n+1:\infty}$, and as such we focus our modelling efforts directly on the predictive density

$$p(y_{n+1:\infty} \mid y_{1:n}). \tag{2.1}$$

Here, the construction of the predictive density is for parameter inference, and not for forecasting future observations as is more usual. For inference, we assume that the object of interest is fully defined once all the observations have been viewed, which we write as $\theta_\infty = \theta(Y_{1:\infty})$. It is clear then that (2.1) induces a distribution on θ_∞, and we call this scheme of imputing $Y_{n+1:\infty}$ and computing θ_∞ as *predictive resampling*. A key observation is that $Y_{1:\infty}$ will always contain the observed $Y_{1:n} = y_{1:n}$ as the predictive Bayesian considers the observed sample to be fixed, in contrast to the frequentist consideration of other possible values of $Y_{1:n}$.

For i.i.d. observations, the traditional Bayesian approach is to elicit a prior density $\pi(\theta)$ and sampling density $f_\theta(y)$, derive the posterior $\pi(\theta \mid y_{1:n})$, then compute the predictive density through

$$p(y \mid y_{1:n}) = \int f_\theta(y)\, \pi(\theta \mid y_{1:n})\, d\theta. \tag{2.2}$$

A concise summary of our approach is the following: while de Finetti (1937) provided a representation of Bayesian inference which relies on exchangeability and the prior distribution, we will introduce a framework based on the results of Doob (1949) which relies solely, in the i.i.d. case, on the predictive distribution. We will see that this framework based on Doob's results is more flexible and the mathematical requirement amounts to the construction of a martingale - it is this flexibility which we exploit in this paper. In fact, through Doob's theorem, we will see that predictive resampling as

described above is identical to posterior sampling when using (2.2) as the predictive and θ indexes the sampling density, in which case $\theta_\infty \sim \pi(\theta \mid y_{1:n})$. Denoting by $p(y)$ the prior predictive, this connection is illustrated below for the traditional Bayesian case:

$$f_\theta(y), \pi(\theta) \xrightarrow[\text{Bayes' rule}]{} \pi(\theta \mid y_{1:n}) \xrightarrow[\int f_\theta(y)\,\pi(\theta|y_{1:n})\,d\theta]{\text{posterior predictive}} p(y \mid y_{1:n})$$

$$\pi(\theta \mid y_{1:n}) \xleftarrow[Y_{n+1:\infty} \sim p(\cdot|y_{1:n})]{\text{Doob's theorem}} p(y \mid y_{1:n}) \xleftarrow[\text{predictive update}]{} p(y)$$

However, the traditional Bayesian focus on the prior on θ makes no appeal to the underlying cause of the uncertainty, that is the unobserved part of the study population $Y_{n+1:\infty}$. Furthermore, the traditional prior to posterior computation is becoming increasingly strained as model complexity and data sizes grow. In our work, we advocate the predictive resampling strategy - given $y_{1:n}$, our starting point is directly the predictive model (2.1) and the target statistic of interest θ_∞, noting now that θ_∞ is no longer restricted to indexing the sampling density. We relax de Finetti's assumption of exchangeability, but we must now take care to construct (2.1) so that θ_N is indeed convergent to some θ_∞, where $\theta_N = \theta(Y_{1:N})$ can be viewed as an estimator. We highlight here that we use n and N for the size of the observed dataset and the imputed population respectively. In the spirit of Doob, we rely heavily on martingales, which also aid in ensuring that expectations of limits coincide with fixed quantities seen at the sample of size n. This can be regarded as a predictive coherency condition, and we designate the distribution of θ_∞ as the *martingale posterior*. Our choice of (2.1) will be density estimators based on recent ideas in the literature, specifically the *conditionally identically distributed* (c.i.d.) sequence of Berti et al. (2004) and bivariate copula update of Hahn et al. (2018).

We now discuss why one would want to go through the route of obtaining the martingale posterior via the induced distribution of θ_∞ from (2.1) rather than the

traditional likelihood–prior construction. Firstly, predictive models are probabilistic statements on observables, which removes the need to elicit subjective probability distributions on parameters which may have no real-world interpretations and only index the sampling density. Secondly, the martingale posterior establishes a direct connection between prediction and statistical inference, opening up the possibility of using modern probabilistic predictive methods for inference (Breiman, 2001b), and transparently acknowledges the source of statistical uncertainty as the missing $Y_{n+1:\infty}$. Thirdly, working directly with predictive distributions is highly practical. For an elicited 1-step ahead predictive, we can predictively resample by carrying out the recursive update

$$\{p(y \mid y_{1:N-1}), y_N\} \mapsto p(y \mid y_{1:N})$$

to sample $Y_{n+1:N}$ for a large enough N such that θ_N has effectively converged to a sample from the martingale posterior, or N matches a known finite study population size. In complex scenarios such as multivariate density estimation and regression, we introduce new copula-based methodologies where our computations remain exact, GPU-friendly and parallelizable, returning us Bayesian uncertainty without any reliance on Markov chain Monte Carlo (MCMC). Finally, a predictive approach more clearly delineates the core similarities and differences between Bayesian and frequentist uncertainty.

We will focus on the i.i.d. data setting in this work, which corresponds to exchangeable traditional Bayesian models. In this setting, the martingale posterior can indeed be regarded as a generalization of the traditional Bayesian model, as the class of c.i.d. models is more general and contains the class of exchangeable models which we will see in Section 2.5.2. In more complex data structures beyond i.i.d. data, such as those encountered in hierarchical modelling or time series, our framework would still apply. In this case, the missing observations we require may no longer be $Y_{n+1:\infty}$, and model elicitation would no longer only involve a sequence of predictive distributions. For example, a simple hierarchical setting is the observation process $Y_i \sim p(y_i|\theta_i)$, where

θ_i is itself drawn from an unknown G_0 and we may be interested in some functional $\gamma(G_0)$. Here, we only observe $Y_{1:n} = y_{1:n}$, so the missing observations of interest are the unobserved random effects $\theta_{1:\infty}$. We can thus seek to impute $\theta_{1:n} \sim p(\theta_{1:n} \mid y_{1:n})$ from the data, followed by the missing remainder $\theta_{n+1:\infty} \sim p(\theta_{n+1:\infty} \mid \theta_{1:n})$. Computing $\gamma(\theta_{1:\infty})$ would then return us a posterior sample. For the remainder of the paper, we will focus on the i.i.d. case and leave the details of non-i.i.d. settings for future work.

In Section 2.4, we formally investigate the connection between predictive and posterior inference, and introduce a predictive framework for inference and the resulting martingale posterior. We then utilize the bootstrap as a canonical example to distinctly compare Bayesian and frequentist uncertainty. We postpone discussion of related work until Section 2.4.5 in order to provide context beforehand. In Section 2.5, we discuss predictive coherence conditions for martingale posteriors, utilizing c.i.d. sequences. In Section 2.6, we revisit the bivariate copula methodology of Hahn et al. (2018) for univariate density estimation, and extend it to obtain the martingale posterior. We then generalize this copula-based method to multivariate density estimation, regression and classification. Section 2.7 then provides a thorough demonstration of the above methods through examples. In Section 2.8, we discuss some theoretical properties of the martingale posterior with the copula-based methodology. Finally, we discuss our results in Section 2.9.

2.4 A predictive framework for inference

2.4.1 Doob's theorem and Bayesian uncertainty

Uncertainty quantification lies at the core of statistical inference, and Bayesian inference is one framework for handling uncertainty in a formal manner. The Bayesian begins with the random variables $(\Theta, Y_1, Y_2, \ldots)$, where $(Y_1, Y_2, \ldots)$ are the observables of interest, and Θ is the parameter which indexes the sampling density $f_\theta(y)$. We assume throughout that the appropriate densities exist. For i.i.d. data, the Bayesian elicits a

joint probability model for the observables and parameter with joint density

$$p(\theta, y_{1:N}) = \pi(\theta) \prod_{i=1}^{N} f_\theta(y_i) \tag{2.3}$$

for each N. Here, the density $\pi(\theta)$ represents prior knowledge about the parameter which generates the observations, and under a Subjectivist point of view, $\Pi(A) = \int_A \pi(\theta)\, d\theta$ represents the subjective probability that the generating parameter value Θ lies in the set A. Marginalizing out Θ gives the joint density of the observables,

$$p(y_{1:N}) = \int \prod_{i=1}^{N} f_\theta(y_i)\, d\Pi(\theta). \tag{2.4}$$

De Finetti however argued that the direct likelihood–prior interpretation of the Bayesian model was insufficient, as Θ is of a "metaphysical" nature and probability statements should only be on observables (Bernardo and Smith, 2009). This then motivated the notion of exchangeability of the infinite sequence $(Y_1, Y_2, \ldots)$, where the joint probability P of the finite sequence of observables $Y_{1:N} = (Y_1, \ldots, Y_N)$ is invariant to the ordering of Y_i for all N. Through de Finetti's representation theorem (de Finetti, 1937) and extensions thereof (e.g. Hewitt and Savage (1955)), the assumption of exchangeability induces the likelihood–prior form of the joint density in (2.4) (where Π may not have a density), which motivates such a specification of the Bayesian model. The representation theorem however is only part of the story. As alluded to in the Section 2.3, the source of statistical uncertainty is the lack of the infinite dataset $Y_{n+1:\infty}$ with which we could pin down any quantity of interest precisely. Bayesian uncertainty through the lens of the prior is still opaque in this regard, even with the aforementioned representation theorem.

The key to understanding the source of uncertainty lies in the predictive imputation of observables, for which we require a result from Doob. Doob (1949) established consistency of the Bayesian method when the observations are distributed according to (2.4). For this result, we require that the model is identifiable, that is $F_\theta \neq F_{\theta'}$

whenever $\theta \neq \theta'$, where F_θ is the cumulative distribution function of f_θ. Let us assume that data has yet to be observed, so the missing observations are $Y_{1:\infty}$. Following the discussion in Section 2.3, one can regard (2.4) as the joint predictive density on the missing population, and can estimate the parameter indexing the sampling density as a function of the imputed $Y_{1:N}$. An appropriate and intuitive point estimate for the Bayesian is the posterior mean, which we write as

$$\bar{\theta}_N = E\left[\Theta \mid Y_{1:N}\right].$$

We now use a secondary result of Doob (1949) to confirm that the prior uncertainty in Θ arises from the predictive uncertainty in $Y_{1:\infty}$.

Theorem 2.1 (Doob (1949))**.** *Assume Θ is in a linear space with $E\left[|\Theta|\right] < \infty$, and $(\Theta, Y_1, Y_2, \ldots)$ is distributed according to* (2.3)*, so $\Theta \sim \Pi$. Under identifiability and measurability conditions on F_θ, we have*

$$\bar{\theta}_N \to \Theta \quad a.s.$$

For the above result, the key is to rely on $\bar{\theta}_N$ being a martingale, that is

$$E\left[\bar{\theta}_N \mid Y_{1:N-1}\right] = \bar{\theta}_{N-1}$$

almost surely. Doob's martingale convergence theorem then ensures that $\bar{\theta}_N$ converges to a limit almost surely. The identifiability condition ensures that the parameter is recoverable from the infinite sample so that the limit of $\bar{\theta}_N$ is indeed Θ. For Θ in more general metric spaces, consistency results with general notions of posterior expectations are provided in Ghosal and van der Vaart (2017, Theorem 6.8). As an aside, we highlight that Doob (1949) provided a more general result: the Bayesian posterior distribution converges weakly to the Dirac measure δ_Θ almost surely for Π-almost every Θ as $N \to \infty$. The technical details of a more general version of this

result can be found in Ghosal and van der Vaart (2017, Theorem 6.9). In the Bayesian nonparametric case where Θ is a probability density function, we have a nonparametric extension of the above results (Lijoi et al., 2004).

Returning to the task at hand, we can summarize the above by considering two distinct methods of sampling Θ from the prior Π before seeing any data. The first is to draw $\Theta \sim \Pi$ directly, which is the opaque view of the inherently random parameter that we are trying to shed light on. The second, which inspires the remainder of our paper, begins with sequentially imputing the unseen observables $Y_1, Y_2, Y_3 \ldots$ from the sequence of predictive densities

$$Y_1 \sim p(\cdot), \quad Y_2 \sim p(\cdot \mid y_1), \quad Y_3 \sim p(\cdot \mid y_2, y_1), \quad \ldots$$

until we have the complete information $Y_{1:\infty}$ in the limit. Given this random infinite dataset, the limiting point estimate $\bar{\theta}_\infty = \lim_{N\to\infty} \bar{\theta}_N$, that is the posterior mean computed on the entire dataset, is in fact distributed according to Π. This equivalence highlights the fact that *a priori* uncertainty in Θ is a consequence of the uncertainty in $Y_{1:\infty}$, and the function $\bar{\theta}$ provides a means to precisely recover our quantity of interest when all information is made available to us.

Of course, such an interpretation is equally valid *a posteriori*, that is after we have observed $Y_{1:n} = y_{1:n}$. Here, sampling $\Theta \sim \Pi(\cdot \mid y_{1:n})$ is equivalent to sampling $Y_{n+1:\infty}$ conditional on $y_{1:n}$ and computing $\bar{\theta}_\infty$ as if we have observed the infinite dataset, noting that $Y_{1:n} = y_{1:n}$ is now fixed. This can be seen by simply substituting the prior π in (2.3), (2.4) and Theorem 2.1 with the posterior $\pi(\cdot \mid y_{1:n})$. In conclusion, Doob's result highlights that the Bayesian seeks to simulate what is needed to pin down the parameter but is missing from reality, that is $Y_{n+1:\infty}$ in the i.i.d. case, and we find this to be a compelling justification for the Bayesian approach.

We now conclude this section with a concrete demonstration of the equivalence between posterior sampling and the forward sampling of $Y_{n+1:\infty}$ through a simple

normal model with unknown mean based on an example from Hahn (2015).

Example 2.1

Let $f_\theta(y) = \mathcal{N}(y \mid \theta, 1)$, with $\pi(\theta) = \mathcal{N}(\theta \mid 0, 1)$. Given an observed dataset $y_{1:n}$, the tractable posterior density takes on the form $\pi(\theta \mid y_{1:n}) = \mathcal{N}(\theta \mid \bar{\theta}_n, \bar{\sigma}_n^2)$ where

$$\bar{\theta}_n = \frac{\sum_{i=1}^n y_i}{n+1}, \quad \bar{\sigma}_n^2 = \frac{1}{n+1}.$$

The posterior predictive density then takes on the form $p(y \mid y_{1:n}) = \mathcal{N}(y \mid \bar{\theta}_n, 1 + \bar{\sigma}_n^2)$. For observed data, we generated $y_{1:n} \overset{\text{iid}}{\sim} f_\theta(y)$ for $n = 10$ with $\theta = 2$, giving $\bar{\theta}_n = 1.84$.

We can plot the independent sample paths for the posterior mean, $\bar{\theta}_{n+1:N}$, as we recursively forward sample $Y_{n+1:N}$, where $N = n + 1000$ in this example. In Figure 2.1, we see that the sample paths of $\bar{\theta}_{n+i}$ each converge to a random Θ as i increases, with the density of $\bar{\theta}_N$ very close to the analytic posterior. From Doob's consistency theorem, we know this is exact for $N \to \infty$.

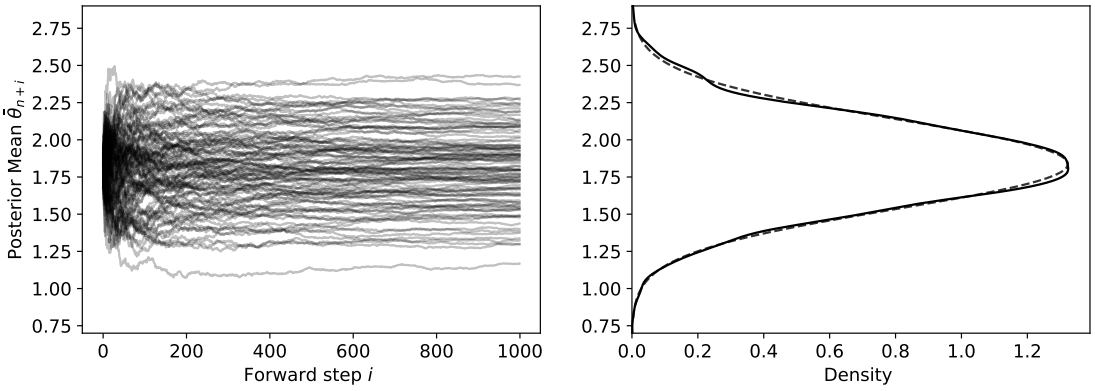


Figure 2.1: (a) Sample paths of $\bar{\theta}_{n+i}$ through forward sampling; (b) Kernel density estimate of $\bar{\theta}_N$ samples (—) and analytical posterior density $\pi(\theta \mid y_{1:n})$ (- - -)

2.4.2 The methodological approach

Through Doob's result in Theorem 2.1, we have demonstrated the predictive view of Bayesian inference as a means to understand how the posterior uncertainty in Θ arises from the missing information $Y_{n+1:\infty}$. The predictive view of Bayesian inference

partitions posterior sampling into two distinct tasks. The first is the simulation of $Y_{n+1:\infty}$ through the sequence of 1-step ahead predictive distributions to assess the uncertainty that arises from the missing observables. The second is the recovery of the parameter of interest Θ from the simulated complete information, which is facilitated by the limiting posterior mean point estimate $\bar{\theta}_\infty$. The uncertainty in Θ then flows from the uncertainty in $Y_{n+1:\infty}$. Inspired by this, we will now demonstrate the practical importance of this interpretation by introducing a predictive framework for inference built exactly on these two tasks. This framework eliminates the need for the usual likelihood–prior construction of the Bayesian model, and as such generalizes the traditional Bayesian posterior to the martingale posterior.

2.4.2.1 Sampling the missing data

For the predictive Bayesian, the role of the posterior $\pi(\theta \mid y_{1:n})$ is to aid in the updating of the predictive density, $p(\cdot \mid y_{1:N-1}) \mapsto p(\cdot \mid y_{1:N})$ after observing Y_N, and the likelihood and prior can be viewed as merely intermediate tools to construct the sequence of predictives (Roberts, 1965). To obviate the need of a likelihood–prior specification, our proposal is to specify the sequence of 1-step ahead predictive densities $\{p(\cdot \mid y_{1:N})\}_{N \geq n}$ directly, which implies a joint density through the factorization

$$p(y_{n+1:N} \mid y_{1:n}) = \prod_{i=n+1}^{N} p(y_i \mid y_{1:i-1}). \tag{2.5}$$

However, we must take care in our elicitation of $\{p(\cdot \mid y_{1:N})\}_{N \geq n}$ to ensure the existence of the limit θ_∞. As this is technical, we defer a formal discussion of this choice and the conditions required to Section 2.5. For now, we point out that a sufficient condition is for the 1-step ahead predictive densities to satisfy a martingale condition similar to that of Doob, with details given in Section 2.5.2. It may seem that constructing this sequence will incur too much complexity, but we will show this is in fact feasible and desirable. One key idea is to utilize a general sequential updating procedure

whereby given an observed $Y_N = y_N$, we have a direct and tractable iterative update $\{p(\cdot \mid y_{1:N-1}), y_N\} \mapsto p(\cdot \mid y_{1:N})$.

2.4.2.2 Recovering the quantity of interest

We now discuss the second task: given a sample $Y_{n+1:\infty}$, we require a procedure to recover the quantity of interest. In a traditional parametric Bayesian model, the quantity of interest is usually the unknown parameter θ that indexes the sampling density, and as shown by Doob, the limiting posterior mean $\bar{\theta}_\infty$ serves this purpose. A more general framework is the decision task discussed in Bissiri et al. (2016), where the aim is to minimize a functional of an unknown distribution function F_0 from which samples $Y_{1:n}$ are i.i.d.. For some loss function $\ell(\theta, y)$, the quantity of interest θ is now defined as

$$\theta_0 = \arg\min_\theta \int \ell(\theta, y)\, dF_0(y). \tag{2.6}$$

More details can be found for example in Huber (2004) and Bissiri et al. (2016). Typical examples are $\ell(\theta, y) = |\theta - y|$ for the median, $\ell(\theta, y) = (\theta - y)^2$ for the mean, and $\ell(\theta, y) = -\log f_\theta(y)$ for the Kullback-Leibler minimizer between some parametric density f_θ and the sampling density f_0. The choice of the negative log–likelihood is also interesting as it allows us to target the parameters of a parametric model without the assumption that the model is well–specified (Walker, 2013; Bissiri et al., 2016). While misspecification under our framework is still an open question, the Bayesian bootstrap has particularly desirable theoretical and practical properties under misspecification (Lyddon et al., 2018, 2019; Fong et al., 2019). We will also consider more general forms of θ_0, e.g. the density of F_0.

Working now in the space of probability distributions, the traditional Bayesian approach would be to elicit a prior on F, perhaps nonparametric, and derive the posterior $\Pi(dF \mid y_{1:n})$. Here, F represents the Bayesian's subjective belief on the unknown true F_0. A posterior sample of θ is then obtained as follows: draw $F \sim \Pi(dF \mid y_{1:n})$ and compute the θ minimizing $\int \ell(\theta, y)\, dF(y)$. For our generalization beyond the

likelihood–prior construction, we do not have a posterior mean nor a posterior F, and thus require an alternative to recover the quantity of interest given a sample of $Y_{n+1:\infty}$ conditioned on $y_{1:n}$. Our proposal is to construct the random limiting empirical distribution function

$$F_\infty(y) = \lim_{N\to\infty} \frac{1}{N} \left\{ \sum_{i=1}^{n} \mathbb{1}(y_i \leq y) + \sum_{i=n+1}^{N} \mathbb{1}(Y_i \leq y) \right\}$$

and take θ to minimize $\int \ell(\theta, y)\, dF_\infty(y)$. Here, our F_∞ takes the place of the posterior draw of F, and its existence will rely on the martingale condition as mentioned above. We can write θ_∞, $\theta(F_\infty)$ or $\theta(Y_{1:\infty})$ interchangeably for the parameter of interest computed from the completed information. If we specify $p(\cdot \mid y_{1:n})$ through the usual likelihood–prior construction, then sampling F from the posterior in fact yields the same random distribution function as F_∞ almost surely; this theoretical justification for the limiting empirical distribution function F_∞ is in Appendix 2.10.3.2.

2.4.3 The martingale posterior

Our framework for predictive inference is summarized as follows. Suppose we observe $Y_{1:n}$ i.i.d. from some unknown F_0 and are interested in the θ_0 defined by (2.6). We specify a sequence of predictive densities $\{p(\cdot \mid y_{1:n})\}_{n\geq 0}$ which satisfies the martingale condition to be discussed in Section 2.5.2 and implies a joint distribution through (2.5). We then impute an infinite future dataset through

$$Y_{n+1} \sim p(\cdot \mid y_{1:n}), \quad Y_{n+2} \sim p(\cdot \mid y_{1:n+1}), \quad \ldots, \quad Y_N \sim p(\cdot \mid y_{1:N-1})$$

for $N \to \infty$. Given the infinite random dataset $Y_{n+1:\infty}$ and the corresponding empirical distribution function F_∞, we compute $\theta_\infty = \theta(F_\infty)$. We designate the distribution of θ_∞ as the martingale posterior, where we use the notation Π_∞ for comparability to traditional Bayes.

Definition 2.1 (Martingale posterior)**.** *The martingale posterior distribution is defined as*

$$\Pi_\infty(\theta_\infty \in A \mid y_{1:n}) = \int \mathbb{1}\{\theta(F_\infty) \in A\}\, d\Pi(F_\infty \mid y_{1:n})\,, \tag{2.7}$$

for measurable set A, which is a subset of the parameter space.

Drawing samples of θ_∞ from the martingale posterior involves repeating the above simulation procedure given above. We refer to this Monte Carlo scheme as predictive resampling, which has strong connections with the Bayesian bootstrap of Rubin (1981), as we will see in Section 2.4.4. In practice however, we may be unable to simulate $N \to \infty$, or the study population may be of finite size N. In this case, we can instead impute $Y_{n+1:N}$ for finite N, giving us the analogous empirical distribution function F_N and parameter $\theta_N = \theta(F_N)$ or $\theta(Y_{1:N})$.

Definition 2.2 (Finite martingale posterior)**.** *The finite martingale posterior is similarly defined as*

$$\Pi_N(\theta_N \in A \mid y_{1:n}) = \int \mathbb{1}\{\theta(y_{1:N}) \in A\}\, p(y_{n+1:N} \mid y_{1:n})\, dy_{n+1:N}\,.$$

In the finite form, the role of the two constituent elements, $p(y_{n+1:N} \mid y_{1:n})$ and $\theta(y_{1:N})$, is even clearer. For infinite populations, we also highlight that the value of θ_N varies around θ_∞, but this may be negligible for sufficiently large N. If the population is actually finite and of size N, then θ_N would be the actual target and thus not an approximation. Finally, we reiterate that the martingale posterior (2.7) is equivalent to the traditional Bayesian posterior when using (2.2) as the predictive. A summary of the notation and an illustration of the imputation scheme is provided respectively in Appendices 2.10.1, 2.10.2.

2.4.4 The Bayesian bootstrap

The resemblance of the martingale posterior to a bootstrap estimator should not have gone unnoticed, as both involve repeated sampling of observables followed by computing estimates from the sampled dataset. The Bayesian bootstrap of Rubin (1981) is often described as the Bayesian version of the frequentist bootstrap. After observing $y_{1:n}$, one draws a random distribution function from the posterior through

$$w_{1:n} \sim \text{Dirichlet}(1, \ldots, 1), \quad F(y) = \sum_{i=1}^{n} w_i \, \mathbb{1}(y_i \leq y).$$

A posterior sample of the statistic of interest can then be computed as $\theta(F)$. One interpretation of the Dirichlet weights is to generate uncertainty through the randomization of the objective function (Newton and Raftery, 1994; Jin et al., 2001; Newton et al., 2020; Ng and Newton, 2020). Closer to our perspective are the connections to Bayesian nonparametric inference, which have been explored in much detail within the literature as it is the non-informative limit of a posterior Dirichlet process (Lo, 1987; Muliere and Secchi, 1996; Ghosal and van der Vaart, 2017). Recent work has exploited the computational advantages of the Bayesian bootstrap for scalable nonparametric inference; see Saarela et al. (2015); Lyddon et al. (2018); Fong et al. (2019); Newton et al. (2020); Knoblauch and Vomfell (2020); Nie and Ročková (2020).

2.4.4.1 The empirical predictive

Within the framework of martingale posteriors, the Bayesian bootstrap has a particularly elegant interpretation that follows from the equivalence to the Pólya urn scheme (Blackwell and MacQueen, 1973; Lo, 1988). The Bayesian bootstrap is equivalent to the martingale posterior if we define our sequence of predictive probability distribution functions to be the sequence of empirical distribution functions, that is

$$P(Y_{n+1} \leq y \mid y_{1:n}) = F_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(y_i \leq y). \tag{2.8}$$

This is easy to see as sampling $Y_{n+1} \sim F_n(y)$ amounts to drawing with replacement 1 of n colours with probability $1/n$ from the urn, and updating to $F_{n+1}(y)$ is equivalent to reinforcing the urn, that is

$$F_{n+1}(y) = \frac{n}{n+1} F_n(y) + \frac{1}{n+1} \mathbb{1}(y_{n+1} \leq y).$$

Continuing on to ∞, the proportions of colours converge in distribution to the Dirichlet distribution. Interestingly, this choice of predictive implies an exchangeable future sequence from the connection to the Dirichlet process. The atomic support of the predictive is however slightly problematic if F_0 is continuous, as any new observations from F_0 will be assigned a predictive probability of zero; we will introduce methodology that remedies this in Section 2.6. Generalizations to other atomic predictives can for example be found in Zabell et al. (1982); Muliere et al. (2000).

One can consider the empirical distribution function as the simplest nonparametric predictive for i.i.d. data, and can thus regard the Bayesian bootstrap as the simplest Bayesian nonparametric model. The uncertainty from the Bayesian bootstrap arises not from the random weights, but from the sequence of empirical predictive distributions. We resample with replacement, treating each resampled point as a new observed datum; this fundamental observation is our motivation for the term predictive resampling.

2.4.4.2 Comparison to the frequentist bootstrap

Throughout this section, we have assumed the existence of an underlying F_0 from which $Y_{1:n}$ are i.i.d., which in turn implies the existence of an unknown true θ_0 much like the frequentist. This has some connections to frequentist consistency under our framework, which we discuss in Section 2.8.3. The posterior random variable θ_∞ then represents our subjective uncertainty in θ_0 after observing $Y_{1:n} = y_{1:n}$. The Bayesian bootstrap and Efron's bootstrap (Efron, 1979) are then ideal vessels for the contrasting of Bayesian and frequentist uncertainty. Both methods are nonparametric and begin by

Algorithm 1: Bayesian bootstrap

Set F_n from the observed data $y_{1:n}$
for $j \leftarrow 1$ **to** B **do**
  for $i \leftarrow n+1$ **to** ∞ **do**
    Sample $Y_i \sim F_{i-1}$
    Update $F_i \leftarrow\!\shortmid \{F_{i-1}, Y_i\}$
  end
  Compute F_∞ from $\{y_{1:n}, Y_{n+1:\infty}\}$
  Evaluate $\theta_\infty^{(j)} = \theta(F_\infty)$
end
Return $\{\theta_\infty^{(1)}, \ldots, \theta_\infty^{(B)}\}$

Algorithm 2: Efron's bootstrap

Set F_n from the observed data $y_{1:n}$
for $j \leftarrow 1$ **to** B **do**
  for $i \leftarrow 1$ **to** n **do**
    Sample $Y_i^* \sim F_n$
    No update to F_n
  end
  Compute F_n^* from $\{Y_{1:n}^*\}$
  Evaluate $\theta_n^{(j)} = \theta(F_n^*)$
end
Return $\{\theta_n^{(1)}, \ldots, \theta_n^{(B)}\}$

constructing the empirical predictive F_n as in (2.8) from the atoms of $y_{1:n}$ as an estimate of F_0, and both involve resampling. The key difference lies in how the resampling is carried out.

The frequentist draws a dataset of size n i.i.d. from F_n, which we write as $Y_{1:n}^*$ with corresponding empirical distribution function F_n^*, and computes $\theta(F_n^*)$ as a random sample of the estimator. The Bayesian on the other hand draws an infinite future dataset $Y_{n+1:\infty}$ through predictive resampling, and computes $\theta(F_\infty)$ as a random sample of the estimand, where F_∞ is the limiting empirical distribution function of $\{y_{1:n}, Y_{n+1:\infty}\}$, noting again that the Bayesian holds $y_{1:n}$ fixed. This is summarized in Algorithms 1 and 2. Notably, the specification in both bootstraps are equivalent: it is merely the elicitation of $F_n(y)$, which entirely characterizes both types of uncertainty.

2.4.5 Related work

There have been many others that shared de Finetti's view on the emphasis on observables for inference. The work of Dawid (1984, 1992a,b) on prequential statistics, a portmanteau of probability/predictive and sequential, is one such example. In his work, Dawid focuses on the importance of forecasting, and introduces statistical methodology that assign predictive probabilities and assesses these methods on their agreement with the observed data. In particular, Dawid (1984) recommends eliciting a sequence of 1-step ahead predictive distributions as we do, but motivates this by

arguing that forecasting is the main statistical task. As pointed out in Section 2.3, this is in contrast to our case where parameter inference is the main task of interest and the sequence of predictives is mainly a convenient tool to construct the joint predictive on future observations. We will see in Section 2.5.2 that stricter conditions are required on this sequence of predictives for inference. Another strong proponent of the predictive approach is the work of Geisser: he believed that the prediction of observables was of much greater importance than the estimation of parameters, which he described as "artificial constructs" (Geisser, 1975). His emphasis on the predictive motivated cross-validation (Stone, 1974; Geisser, 1974), which is now popular for Bayesian model evaluation (Vehtari and Lampinen, 2002; Gelman et al., 2014). Works such as Dawid (1985); Lauritzen (1988) also consider parameters as functions of the infinite sequence of observations using the notion of repetitive structures. Finally, the work of Rubin on both the potential outcomes model (Rubin, 1974) and multiple imputation (Rubin, 2004) highlights the idea of inference via imputation.

An early application of what is essentially finite predictive resampling and martingale posteriors is Bayesian inference for finite populations, first discussed in Roberts (1965); Ericson (1969) and later by Geisser (1982, 1983). A finite population Bayesian bootstrap is described in Lo (1988), in which a finite Pólya urn is used to simulate from the posterior. The 'Pólya posterior' of Ghosh and Meeden (1997) uses the same approach following an admissibility argument. These methods have applications in survey sampling or the interim monitoring of clinical trials (Saville et al., 2014).

There have been recent exciting directions of work that investigate the predictive view of Bayesian nonparametrics (BNP). Fortini et al. (2000) investigate under what conditions parametric models arise from the sequence of predictives using the concept of predictive sufficiency, and derive conditions such that the joint distribution is exchangeable. Fortini and Petrone (2012, 2014) discuss the construction of a range of popular exchangeable BNP priors through a sequence of predictive distributions, motivated through a predictive de Finetti's representation theorem (Fortini and Petrone,

2012, Theorem 2). Berti et al. (2020, 2021) then generalize the nonparametric approach to c.i.d. sequences; we will later see that c.i.d. sequences, as introduced in Berti et al. (2004), play a crucial role in our work. However, the previously described methods are mostly constrained to the discrete case. Hahn (2015) and Hahn et al. (2018) construct c.i.d. models through a predictive sequence for univariate density estimation, respectively utilizing the kernel density estimator and the bivariate copula. Hahn (2015) also discusses the connection of Bayesian uncertainty and prediction with a weaker argument, and gives a similar example to our Example 2.1. Predictive resampling is then used to sample nonparametric densities from a finite martingale posterior; however Hahn (2015) instead specifies the predictive distribution P_N for large N and works backwards to find the sequence of predictives. Fortini and Petrone (2020) analyze the predictive recursion algorithm of Newton et al. (1998) and the implied underlying quasi-Bayesian model. In their work, they carry out predictive resampling to simulate from the prior law of the mixing distribution in an example, and obtain its asymptotic distribution under the c.i.d. model, that is an asymptotic approximation to the martingale posterior. An interesting aside is the recent work of Waudby-Smith and Ramdas (2020) which utilizes adaptive betting with martingale conditions for the purpose of constructing frequentist confidence intervals. We aim to unify these related strands of research under a single framework.

2.5 Predictive resampling for martingale posteriors

For the martingale posterior, we now embark on the task of eliciting the general 1-step ahead predictive distributions, with the traditional Bayesian posterior predictive as a special case. For notational convenience, we write the sequence of predictive probability distribution functions estimated after observing $Y_{1:i} = y_{1:i}$ as

$$P_i(y) := P(Y_{i+1} \leq y \mid y_{1:i}), \quad i \in \{1, 2, \ldots\} \tag{2.9}$$

which may have corresponding density functions $p_i(y)$. The subscript indicates the length of the conditioning sequence, and there may be a $P_0(y)$ as some initial choice. For a general sequence of predictives, where exchangeability no longer necessarily holds, we instead define our joint distribution on $y_{1:N}$ through this sequence of 1-step ahead predictives and the chain rule as in (2.5). The Ionescu-Tulcea theorem (Kallenberg, 1997, Theorem 5.17) guarantees the existence of such a joint distribution as we take $N \to \infty$, which has been pointed out by works such as Dawid (1984); Fortini and Petrone (2012); Berti et al. (2020).

Beyond the traditional Bayesian posterior predictive, there is good justification for specifying the model with 1-step ahead predictives, instead of say m-step ahead. It is simple to interpret and estimate a 1-step ahead predictive as the decision maker's best estimate of the unknown sampling distribution function F_0, and methods such as maximum likelihood estimation already do this. Finally, we will see that a 1-step update of the predictive allows for the enforcing of the c.i.d. condition for predictive coherence.

While the prescription of (2.9) remains a subjective task, we find it to be no more subjective than the selection of a sampling density. There is no longer a need to elicit subjective distributions on parameters which merely index the sampling distribution with no physical meaning, which has been described as 'intrinsic' (Dawid, 1985). In nonparametric inference, we also do not need to elicit priors directly on the space of probability distributions, which can be cumbersome. The uncertainty arises simply from the elicitation of (2.9). It is clear that we can still use external information and subjective judgement not provided by the data $y_{1:n}$ in this construction.

2.5.1 A practical algorithm for uncertainty

Given the model specification (2.9), suppose we wish to undertake inference on a statistic of interest $\theta(F_0)$, defined through a loss function $\ell(\theta, y)$ as in (2.6). We can obtain finite martingale posterior samples through predictive resampling given in

Algorithm 3, noting the similarity to the Bayesian bootstrap algorithm.

Algorithm 3: Predictive resampling

```
Compute P_n from the observed data y_{1:n}
N > n is a large integer
for j ← 1 to B do
    for i ← n + 1 to N do
        Sample Y_i ~ P_{i−1}
        Update P_i ↤ {P_{i−1}, Y_i}
    end
    Compute F_N from {y_{1:n}, Y_{n+1:N}}
    Evaluate θ_N^(j) = θ(F_N) or θ_N^(j) = θ(P_N)
end
Return {θ_N^(1), ..., θ_N^(B)} ~iid Π_N(· | y_{1:n})
```

In summary, we run a forward simulation starting at $P_n(y)$ by consecutively sampling from the 1-step ahead predictives and updating as we go. For large N, we now have a random dataset $\{y_{1:n}, Y_{n+1:N}\}$ from which we can compute the empirical distribution function $F_N(y)$ and statistic of interest $\theta(F_N)$. In particular, when the sequence of predictives takes on the form (2.2), combined with the self-information loss, $-\log f_\theta(y)$, is this procedure equivalent to traditional Bayesian inference.

The empirical distribution is atomic, which may be problematic if the object of interest θ_0 requires the limiting F_∞ to be continuous, for example if θ_0 is the probability density of F_0 or a tail probability. In this case, we can instead compute $\theta(P_N)$, where P_N is the random predictive distribution function conditioned on $\{y_{1:n}, Y_{n+1:N}\}$, which would typically be continuous. We can regard P_N as the finite approximation to the limiting predictive distribution function $P_\infty := \lim_{N\to\infty} P_N$, which serves the same purpose as the limiting empirical F_∞ in Section 2.4.2.2. In fact, P_∞ and F_∞ coincide for traditional Bayesian models, and even for the more general c.i.d. sequence of predictives that we will consider shortly. We discuss this in Appendix 2.10.3, borrowing results from Doob (1949), Berti et al. (2004) and Lijoi et al. (2004).

Some experimental and theoretical guidance for selecting a sufficiently large N to estimate P_∞ is given in Sections 2.7 and 2.8. However, it is also interesting to

consider a finite population, where the F_0 of interest is indeed the empirical distribution function of a population of size N, as discussed in Sections 2.4.3 and 2.4.5. In this case, truncating predictive resampling at N indeed returns the correct uncertainty in any parameter of interest $\theta(Y_{1:N})$ of the finite population.

2.5.2 Predictive coherence and conditionally identically distributed sequences

The notion of coherence on one's belief on the parameter θ is key to the subjective Bayesian, where coherence may be defined in a decision-theoretic sense (Bernardo and Smith, 2009, Chapter 2.3) or through Dutch book arguments (e.g. Heath and Sudderth (1978)). Extensions of coherence to forecasting are given in Lane and Sudderth (1984); Berti et al. (1998), and more examples of coherence in general can be found in Robins and Wasserman (2000); Eaton and Freedman (2004). More recently, the notion of coherence of belief updating was introduced in Bissiri et al. (2016), where a belief update on a statistic of interest θ is coherent if the update is equivalent whether computed sequentially with y_1 followed by y_2 or with $\{y_1, y_2\}$ in tandem through an additive loss condition. In bypassing the traditional likelihood–prior construction, we must forsake the usual coherence of belief updating and exchangeability. Instead, we specify conditions for a valid martingale posterior entirely in terms of the predictive distribution function, which we term *predictive coherence*.

Suppose we observe $Y_{1:n}$ i.i.d. from some F_0 and construct $P_n(y)$ as in (2.9). We can then view the predictive machine $P_n(y)$ as the best estimate of the unknown distribution function F_0 from which the data arose, incorporating all observed data and any possible subjective knowledge. The first minimal condition is that the sequence of predictive distribution functions $P_{n+1}(y), P_{n+2}(y) \ldots$ converges to a random distribution function. Secondly, we would ensure that predictive resampling does not introduce any new information or bias, as P_n is already our best summary of the observed $y_{1:n}$, and

the procedure should merely return uncertainty. Formally, we write these conditions respectively as follows:

Condition 2.1 (Existence)**.** *The sequence* $P_{n+1}(y), P_{n+2}(y), \ldots$ *converges to a random* $P_\infty(y)$ *almost surely for each* $y \in \mathbb{R}$*, where* P_∞ *is a random probability distribution function.*

Condition 2.2 (Unbiasedness)**.** *The posterior expectation of the random distribution function satisfies*

$$E\left[P_\infty(y) \mid y_{1:n}\right] = P_n(y)$$

almost surely for each $y \in \mathbb{R}$*.*

Under Condition 2.1, P_∞ is defined through the sequence of predictives, and we can thus treat P_∞ directly as the random distribution function without the need for an underlying Bayes' rule representation. This in turn gives us the posterior uncertainty in any statistic $\theta(P_\infty)$. Condition 2.2 is stricter, and implies that P_n is our best estimate of F_0 and is equal to the posterior mean.

Fortunately, Conditions 2.1 and 2.2 are satisfied if the sequence $Y_{n+1}, Y_{n+2}, \ldots$ is *conditionally identically distributed* (c.i.d.), as introduced and studied in Berti et al. (2004). Many useful properties of c.i.d. sequences have been shown in their work, which we now summarize. The sequence $Y_{n+1}, Y_{n+2}, \ldots$ is c.i.d if we have

$$P\left(Y_{i+k} \leq y \mid y_{1:i}\right) = P_i(y), \quad \forall k > 0$$

almost surely for each $y \in \mathbb{R}$. This states that conditional on $y_{1:i}$, any future data points will be identically distributed according to the predictive P_i. This predictive invariance is particularly natural as a minimal predictive coherence condition, and serves as an analogue to de Finetti's exchangeability assumption in the predictive framework. In fact, as shown in Kallenberg (1988), the c.i.d. condition is a weakening of exchangeability, and Berti et al. (2004) also show that c.i.d. sequences are asymptotically exchangeable,

which we state formally in Theorem 2.3 in Section 2.8.1.

An equivalent formulation of c.i.d. sequences which connects closely to the predictive coherency conditions is that $P_i(y)$ is a martingale for $i \in \{n+1, n+2, \ldots\}$, that is

$$E\left[P_i(y) \mid y_{1:i-1}\right] \equiv \int P_i(y)\, dP_{i-1}(y_i) = P_{i-1}(y) \tag{2.10}$$

almost surely for each $y \in \mathbb{R}$, noting that P_i depends on y_i as in (2.9). Relying again on Doob's martingale convergence theorem (Doob, 1953), the sequence $P_n(y), P_{n+1}(y), \ldots$ converges to $P_\infty(y)$ almost surely for each $y \in \mathbb{R}$, and P_∞ can be shown to be a random probability distribution function (Berti et al., 2004); we state this formally in Theorem 2.4 in Section 2.8.1. In this case, we also designate the distribution of P_∞ as the martingale posterior when we do not specify θ_∞. Condition 2.2 is then satisfied as the sequence $P_{n+1}(y), P_{n+2}(y), \ldots$ is uniformly integrable. Furthermore, we are guaranteed the existence of the limiting empirical distribution function F_∞ as required in Section 2.4.2.2, and in fact $F_\infty(y) = P_\infty(y)$ almost surely so the interchangeability of $\theta(F_\infty)$ and $\theta(P_\infty)$ is justified. This equivalence, as well as the convergence of $\theta(Y_{1:N})$ with N for a certain class of parameters, is discussed in Appendix 2.10.3.1. Although not explored here, connections of the c.i.d. property to other notions of coherence, such as those given at the start of this subsection, would be interesting to investigate especially given the absence of the prior distribution.

Although the above predictive coherence conditions are for a valid martingale posterior, we still need to specify a sequence of predictive distributions. Clearly the traditional Bayesian posterior predictive satisfies the above conditions, but in the interest of computational expediency or the desire to bypass the likelihood–prior construction, we may wish to consider more general predictive machines. The remainder of this paper will consider recursive predictive densities using bivariate copulas.

2.6 Recursive predictives with bivariate copulas

In this section, we focus primarily on the elicitation of the sequence of predictives (2.9) in the continuous case, where $p_i(y)$ is the density of $P_i(y)$ in (2.9). Analogous predictives are derivable for the discrete case, and these are obtained in Berti et al. (2020). In particular, we investigate the prescription of this sequence of predictives through a recursive manner, that is for $i \in \{0, 1, \ldots\}$

$$p_{i+1}(y) = \psi_{i+1}^{\rho} \{p_i(y), y_{i+1}\}$$

where ψ_i^{ρ} is a sequence of update functions, possibly parameterized by a hyperparameter ρ. In this case, we require an initial guess $p_0(y)$ for our recursion, which plays the role of a prior guess on f_0. A recursive update of this form is not necessary for a martingale posterior, but it allows for simple satisfaction of conditions for predictive coherence as discussed in Section 2.5.2, and computations for predictive resampling will also be significantly easier. Furthermore, when one is only interested in estimating $p_n(y)$, recursive updates may have computational advantages as one does not need to explicitly estimate the posterior.

Recursive updates have previously been motivated as a fast alternative to MCMC in Dirichlet process mixture models (DPMM). The predictive recursion algorithm was first introduced by Newton et al. (1998), which estimates the mixing distribution through a recursive update, and its properties have been studied in detail in the literature; see Martin (2018) for a thorough review. One interesting property shown in Fortini and Petrone (2020) is that the sequence of observables in Newton's algorithm is c.i.d.; however, the computation of the predictive densities is intractable and requires numerical integration, so we will not discuss this method further here. Direct recursive updates for the predictive density were then introduced in Hahn (2015); Hahn et al. (2018); Berti et al. (2020), all of which satisfy the c.i.d. condition. The bivariate copula method of Hahn et al. (2018) is particularly tractable and well motivated, and we will

now build on this method in this section.

2.6.1 Bivariate copula update

To satisfy the c.i.d. condition required for predictive coherence, we can extend the martingale condition to hold for the sequence of densities $p_n, p_{n+1}, \ldots$ such that for $i \in \{n+1, n+2, \ldots\}$

$$E\left[p_i(y) \mid y_{1:i-1}\right] \equiv \int p_i(y)\, p_{i-1}(y_i)\, dy_i = p_{i-1}(y) \tag{2.11}$$

almost surely for each $y \in \mathbb{R}$, assuming the expectations exist. We highlight again that p_i depends on y_i as it is the density of (2.9). The above is a sufficient condition for (2.10) to hold, so our sequence is c.i.d. and the existence and unbiasedness conditions are satisfied giving us a valid martingale posterior. In fact, the martingale convergence theorem shows that $p_i(y) \to p_\infty(y)$ almost surely for each $y \in \mathbb{R}$, but more assumptions are needed to show that p_∞ is the density of $P_\infty(y)$; we explore this in Theorem 2.5 in Section 2.8.1.

One particular tractable form of update rule ψ_i^ρ that satisfies (2.10) is the bivariate copula (Nelsen, 2007) update interpretation of Bayesian inference first introduced in Hahn et al. (2018) for univariate data. A bivariate copula is a bivariate cumulative distribution function $C : [0,1]^2 \to [0,1]$ with uniform marginal distributions, and in the cases we consider it will have a probability density function $c : [0,1]^2 \to \mathbb{R}$. The bivariate copula can be regarded as characterizing the dependence between two random variables independent of their marginals, which can be seen through Sklar's theorem in the bivariate case.

Theorem 2.2 (Sklar (1959))**.** *For a bivariate cumulative distribution function $F(y_1, y_2)$ with continuous marginals $F_1(y_1), F_2(y_2)$, there exists a unique bivariate copula C such that*

$$F(y_1, y_2) = C\{F_1(y_1), F_2(y_2)\}.$$

Furthermore, if F has a density f with marginal densities f_1, f_2, we can write

$$f(y_1, y_2) = c\{F_1(y_1), F_2(y_2)\}\, f_1(y_1)\, f_2(y_2)$$

where c is the density of C.

This holds for higher dimensions, but we state it for $d = 2$ as this is what we will be working with. From this, we can see that the bivariate copula can model the dependence structure between consecutive predictive densities, and thus we have the following corollary, with the proof given in Appendix 2.10.4.1.

Corollary 2.1. *The sequence of conditional densities $p_0, p_1, \ldots$ satisfies the martingale condition* (2.11) *if and only if there exists a unique sequence of bivariate copula densities $c_1, c_2, \ldots$ such that*

$$p_{i+1}(y) = c_{i+1}\{P_i(y), P_i(y_{i+1})\}\, p_i(y) \tag{2.12}$$

for $i \in \{0, 1, \ldots\}$ and P_i is the distribution function of p_i.

In the univariate case, we can thus elicit a c.i.d. model through a sequence of copulas, that is we have (2.12) as our update function ψ^{ρ}_{i+1}. We highlight that c_{i+1} is the bivariate copula density that models the dependence between $\{Y_{i+1}, Y_{i+2}\}$ conditioned on $Y_{1:i}$. Although the sequence c_{i+1} can technically depend arbitrarily on $y_{1:i}$ (and the sample size $i+1$) without violating the martingale condition, we will later constrain this dependence. As all exchangeable Bayesian models are c.i.d., there exists a unique sequence of copulas which may or may not be tractable that characterize the model (Hahn et al., 2018). This sequence takes on exactly the form

$$p_{i+1}(y) = \underbrace{\frac{\int f_\theta(y)\, f_\theta(y_{i+1})\, \pi(\theta \mid y_{1:i})\, d\theta}{p_i(y)\, p_i(y_{i+1})}}_{c_{i+1}\{P_i(y), P_i(y_{i+1})\}}\, p_i(y). \tag{2.13}$$

The copula density arises following Theorem 2.2 as the numerator in (2.13) is the joint density $p_i(y, y_{i+1})$ with marginal densities $p_i(y)$ and $p_i(y_{i+1})$. Instead of specifying the sampling distribution and prior, we will now consider the specification of the sequence of copulas c_i directly. The form for c_i inspired by the DPMM is particularly attractive, and serves well as the canonical extension of the Bayesian bootstrap predictive to continuous random variables. In the remainder of this section, we will first review the method of Hahn et al. (2018) for univariate density estimation, and extend the methodology to include predictive resampling and hyperparameter selection. We then introduce analogous copula updates for more advanced data settings, including multivariate density estimation, regression and classification.

2.6.2 Univariate case

Tractable forms of this sequence of copulas in Bayesian models are investigated in Hahn et al. (2018), which correspond to conjugate priors. The update of particular interest is that of the DPMM (Escobar and West, 1995) of the particular form

$$f_G(y) = \int \mathcal{N}(y \mid \theta, 1)\, dG(\theta), \quad G \sim \mathrm{DP}\left(a, G_0\right), \quad G_0 = \mathcal{N}(\theta \mid 0, \tau^{-1}),$$

where $a > 0$ is the scalar precision parameter that we set to $a = 1$. The model is nonparametric, making it a strong candidate for a predictive update, but only the copula update for $i = 0$ is tractable. Inspired by this first update step, Hahn et al. (2018) suggest that the general update to compute the density $p_i(y)$ after observing $y_{1:i}$ for $i \in \{0, \ldots, n-1\}$ takes on the form

$$\begin{aligned} p_{i+1}(y) &= (1 - \alpha_{i+1})\, p_i(y) + \alpha_{i+1} c_\rho \left\{P_i(y), P_i(y_{i+1})\right\} p_i(y) \\ P_{i+1}(y) &= (1 - \alpha_{i+1}) P_i(y) + \alpha_{i+1} H_\rho \left\{P_i(y), P_i(y_{i+1})\right\} \end{aligned} \tag{2.14}$$

where $P_i(y)$ is the distribution function of $p_i(y)$. Here $c_\rho(u, v)$ is the bivariate Gaussian copula density and $H_\rho(u, v)$ is the conditional Gaussian copula of the forms:

$$\begin{aligned} c_\rho(u,v) &= \frac{\mathcal{N}_2\left\{\Phi^{-1}(u), \Phi^{-1}(v) \mid 0, 1, \rho\right\}}{\mathcal{N}\{\Phi^{-1}(u) \mid 0, 1\}\,\mathcal{N}\{\Phi^{-1}(v) \mid 0, 1\}} \\ H_\rho(u,v) &= \Phi\left\{\frac{\Phi^{-1}(u) - \rho\Phi^{-1}(v)}{\sqrt{1-\rho^2}}\right\} \end{aligned} \tag{2.15}$$

where Φ^{-1} is the standard inverse normal distribution function and $\mathcal{N}_2$ is the standard bivariate density with correlation $\rho \in (0, 1)$. The role of ρ as a bandwidth will be explored shortly. The update (2.14) is then a mixture of the independent copula density and the Gaussian copula density, and the sequence $\alpha_i = \mathcal{O}\left(i^{-1}\right)$ ensures the update approaches the independent copula as $i \to \infty$. Although asymptotic independence is not necessary for the martingale condition, this property holds for Bayesian sequences of copulas (Hahn et al., 2018), and is indeed important for frequentist consistency when estimating p_n as we will see in Section 2.8.3. We will see the specific suggested form of α_i at the end of this subsection.

Note the similarity of the update in (2.14) to the generalized Pólya urn for the Dirichlet process, which for $c = 1$ has the update $P_{i+1}(y) = (1-\alpha_{i+1})P_i(y) + \alpha_{i+1}\,\mathbb{1}(y_{i+1} \leq y)$. We can thus interpret (2.14) as a smooth generalization of the Bayesian bootstrap update for continuous distributions. One can also interpret (2.14) as a Bayesian kernel density estimate (KDE) that satisfies the c.i.d. condition, as the regular KDE cannot satisfy this condition (West, 1991). The update can be visualized in Figure 2.2, where for convenience we write $u_i = P_i(y)$, $v_i = P_i(y_{i+1})$. The Gaussian copula kernel $c_\rho\left(u_i, v_i\right) p_i(y)$ is a data dependent kernel roughly centered at y_{i+1}, as shown in the left. The kernel becomes sharper as ρ increases, and we recover the Bayesian bootstrap in the limit of $\rho \to 1$ (with $\alpha_i = 1/i$). The update is then a mixture of $p_i(y)$ and the copula kernel, which gives us $p_{i+1}(y)$ in the right panel.

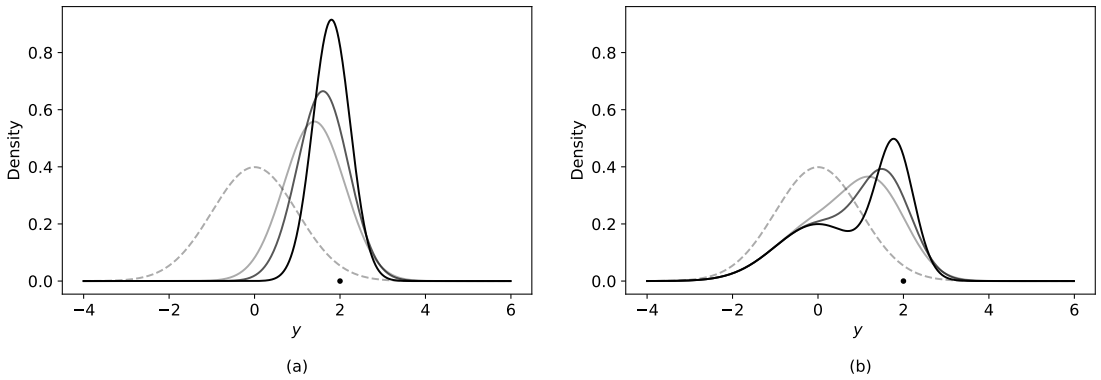


Figure 2.2: Current predictive density $p_i(y)$ (- - -) and new datum y_{i+1} (•); (a) Copula kernel $c_\rho(u_i, v_i)\, p_i(y)$ for correlation $\rho = 0.7, 0.8, 0.9$ (—,—,—); (b) Corresponding updated predictive density $p_{i+1}(y)$ (—,—,—) for $\alpha_{i+1} = 0.5$; note that we write $u_i = P_i(y), v_i = P_i(y_{i+1})$

The recursive update was first introduced to compute $p_n(y)$, but properties of the update make it a highly suitable candidate for predictive resampling. Firstly, by Corollary 2.1, this update is guaranteed to provide a c.i.d. sequence and hence satisfy the existence and unbiasedness conditions. Secondly, the update of the predictive distribution is online, and does not require an expensive recomputation of the predictive distribution at each step. Finally, the predictive resampling update is particularly computationally elegant as $y_{i+1} \sim P_i(y)$ implies that $P_i(y_{i+1}) \sim \mathcal{U}[0,1]$, so all that is required is the simulation of uniform random variables. The forward sampling step then involves simulating $V_i \sim \mathcal{U}[0,1]$ and computing

$$p_{i+1}(y) = [1 - \alpha_{i+1} + \alpha_{i+1} c_\rho \{P_i(y), V_i\}]\, p_i(y)$$
$$P_{i+1}(y) = (1 - \alpha_{i+1}) P_i(y) + \alpha_{i+1} H_\rho \{P_i(y), V_i\}$$

iterated over $i \in \{n, \ldots, N\}$, which gives us a random $p_N(y)$ at the end. There is no need to actually sample $Y_{i+1} \sim P_i(y)$, which is possible but is more computationally expensive. In Section 2.8, we will see that this update form allows easy analysis of the theoretical properties of predictive resampling.

The bandwidth ρ controls the smoothness of the density estimate, which we can set in a data-dependent manner as we show in Section 2.6.5.2. On the other hand,

the sequence α_i is responsible for the uncertainty as we will see in Section 2.8, so extra care must be taken when eliciting this. Hahn et al. (2018) suggest the form $\alpha_i = (i+1)^{-1}$ inspired from the stick-breaking process of the posterior DP as in the Bayesian bootstrap, which works well for estimating $p_n(y)$ but we find this performs poorly when predictive resampling, giving too little uncertainty. This was also observed in Fortini and Petrone (2020) in the case of Newton's recursive method. However, it should be observed that the posterior over the mixing distribution G is actually a mixture of DPs, that is

$$[G \mid \theta_{1:n}, y_{1:n}] \sim \mathrm{DP}\left(a+n, \frac{aG_0 + \sum_{i=1}^n \delta_{\theta_i}}{a+n}\right), \quad [\theta_{1:n} \mid y_{1:n}] \sim \pi(\theta_{1:n} \mid y_{1:n})$$

where $\pi(\theta_{1:n} \mid y_{1:n})$ is intractable. As shown in Appendix 2.10.5.1, we only require the simplifying assumption of $\pi(\theta_{1:n} \mid y_{1:n}) = \prod_{i=1}^n G_0(\theta_i)$, which corresponds to each datum belonging to its own cluster in a similar spirit to the KDE. This then returns us the same copula update as (2.14) with

$$\alpha_i = \left(2 - \frac{1}{i}\right)\frac{1}{i+1}. \tag{2.16}$$

Intuitively, the additional mixing over $\theta_{1:n}$ results in the inflated value compared to $\alpha_i = (i+1)^{-1}$. Note this is still $\mathcal{O}(i^{-1})$, matches with initial update step for $i = 1$, and works much better in practice as it approaches 0 more slowly. We use this sequence for the remainder of the copula methods.

2.6.3 Multivariate case

In this section, we extend the univariate method to multivariate data $\mathbf{y} \in \mathbb{R}^d$, allowing us to both learn $p_n(\mathbf{y})$ recursively, and retain the c.i.d. sequence so we can predictively resample to obtain uncertainty. Even without predictive resampling, a general multivariate density estimator $p_n(\mathbf{y})$ is of interest, as the KDE is known to perform

poorly in high dimensions; see Wang and Scott (2019) for a review. Computation for the multivariate DPMM (MacEachern, 1994; Escobar and West, 1995; Neal, 2000) may scale poorly as the number of dimensions grows. Variational inference (VI) is a quicker approximation as demonstrated in Blei and Jordan (2006), but there is strong dependence on the optimization procedure, which may impair performance in high dimensions. A copula method for bivariate data is suggested in the appendix of Hahn et al. (2018), but it does not scale well with dimensionality and is not c.i.d.. A recursive method for multivariate density estimation is introduced in Cappello and Walker (2018), but numerical integration on a grid is still required, which scales exponentially with d, or a Monte Carlo scheme is required. Fortini and Petrone (2020) propose a multivariate extension of Newton's recursive method, but it also requires an approximate Monte Carlo scheme to evaluate the predictive density.

Extending the above argument in Corollary 2.1 to multivariate data is not as straightforward, as we would like to factorize the joint density into $p_i(\mathbf{y}, \mathbf{y}_{i+1}) = k(\mathbf{y}, \mathbf{y}_{i+1})p_i(\mathbf{y})p_i(\mathbf{y}_{i+1})$, which does not have the copula interpretation like in the 2-dimensional case. Furthermore, building high-dimensional copulas is a difficult task, and bivariate copulas are good building blocks for higher dimensional dependency (Joe and Xu, 1996; Bedford and Cooke, 2001; Aas et al., 2009).

2.6.3.1 Factorized kernel

With the above in mind, we now consider the first step update of a multivariate DPMM below

$$f_G(\mathbf{y}) = \int \prod_{j=1}^{d} \mathcal{N}(y^j \mid \theta^j, 1)\, dG(\boldsymbol{\theta}), \quad G \sim \mathrm{DP}\left(a, G_0\right), \quad G_0(\boldsymbol{\theta}) = \prod_{j=1}^{d} \mathcal{N}(\theta^j \mid 0, \tau^{-1})$$

where y^j is the j-th dimension of $\mathbf{y}$, and likewise for θ^j. Note the factorized normal kernel and independent priors for each θ^j. From this, we see that we can factorize $p_0(\mathbf{y}) = \prod_{j=1}^{d} p_0(y^j)$. It is shown in Appendix 2.10.5.2 that the first update step takes

on the form

$$p_1(\mathbf{y}) = \left[1 - \alpha_1 + \alpha_1 \prod_{j=1}^{d} c_\rho \left\{P_0(y^j), P_0(y_1^j)\right\}\right] p_0(\mathbf{y})$$

where y_i^j is the j-th dimension of the i-th data point. However, naively using this update for $i > 1$ will result in the sequence $p_i(\mathbf{y})$ no longer satisfying the martingale condition in (2.11), and we also find that it performs poorly empirically. A simple but key extension allows us to retain the c.i.d. sequence:

$$p_{i+1}(\mathbf{y}) = \left\{1 - \alpha_{i+1} + \alpha_{i+1} \prod_{j=1}^{d} c_\rho \left(u_i^j, v_i^j\right)\right\} p_i(\mathbf{y}) \tag{2.17}$$

where

$$u_i^j = P_i\left(y^j \mid y^{1:j-1}\right), \quad v_i^j \;\; = P_i\left(y_{i+1}^j \mid y_{i+1}^{1:j-1}\right).$$

The input to the bivariate normal copula is now the *conditional* cumulative distribution function at $\mathbf{y}$ and $\mathbf{y}_{i+1}$ for a particular dimension ordering, and this change ensures many desirable properties. First, we can verify that the martingale condition (2.11) now holds through a multivariate change of variables from $\mathbf{y}_{i+1}$ to $v_i^{1:d}$, so the c.i.d. condition is satisfied. By marginalizing $y^d, y^{d-1}, \ldots, y^{k+1}$ in descending order, we also have that the marginals for a single ordering of dimensions has the same update

$$p_{i+1}\left(y^{1:k}\right) = \left\{1 - \alpha_{i+1} + \alpha_{i+1} \prod_{j=1}^{k} c_\rho \left(u_i^j, v_i^j\right)\right\} p_i\left(y^{1:k}\right). \tag{2.18}$$

From this, we can update the conditional distribution functions via

$$u_{i+1}^k = \left\{(1 - \alpha_{i+1})u_i^k + \alpha_{i+1} H_\rho\left(u_i^k, v_i^k\right) \prod_{j=1}^{k-1} c_\rho \left(u_i^j, v_i^j\right)\right\} \frac{p_i\left(y^{1:k-1}\right)}{p_{i+1}\left(y^{1:k-1}\right)} \tag{2.19}$$

and likewise for v_{i+1}^k. As a result, all terms in the update (2.17) can be computed tractably, with no need for numerical integration or approximations, allowing us to extend this method to any number of dimensions as computation complexity is linear

in d. Notably, we must specify an ordering of the dimensions of $\mathbf{y}$, which at first may seem undesirable. However, it is not an assumption on dependence, and the only implication is that the subset of ordered marginal distributions continue to satisfy (2.18), that is a sort of marginal coherence. Interestingly, the form of (2.18) suggests that $p_i\left(y^{1:k}\right)$ depends only on the first k dimensions of $\mathbf{y}_{1:i}$. Practically, we find the dimension ordering makes little difference, and we recommend selecting the ordering such that any conditional or marginal distributions of interest remain tractable. In Appendix 2.10.5.3 we provide an extension to the above for mixed-type data.

Predictive resampling again takes on a simple form due to the nature of the update (2.17). We can imagine drawing each dimension of $\mathbf{Y} \sim P_i(\cdot)$ in a sequential nature, that is

$$[Y^1] \sim P_i\left(y^1\right), \quad [Y^2 \mid y^1] \sim P_i\left(y^2 \mid y^1\right), \quad \ldots, \quad [Y^d \mid y^{1:d-1}] \sim P_i\left(y^d \mid y^{1:d-1}\right). \tag{2.20}$$

Letting V_i^j denote $P_i\left(Y^j \mid Y^{1:j-1}\right)$, we then have that $V_i^j \overset{\text{iid}}{\sim} \mathcal{U}[0,1]$ for $j = \{1, \ldots, d\}$, which we can substitute into (2.17) and (2.19), similar to the univariate case. Predictive resampling again only requires sampling d independent uniform random variables for each forward step and computing the update.

2.6.4 Regression

We now consider extending the copula method and predictive resampling to the regression setting, where we have univariate $y_i \in \mathbb{R}$ (which can be easily extended to multivariate) with corresponding covariates $\mathbf{x}_i \in \mathcal{X}$, where for example $\mathcal{X} = \mathbb{R}^d$. We will later also consider binary regression, where $y_i \in \{0, 1\}$. One assumption is that the covariates are random, where we write $\{y_i, \mathbf{x}_i\} \overset{\text{iid}}{\sim} f_0(y, \mathbf{x})$, and we are interested in $f_0(y_i \mid \mathbf{x}_i)$. We term this the 'joint method', as we infer the full joint $f_0(y_i, \mathbf{x}_i)$ from which the conditional then follows. Examples of this are Müller et al. (1996); Shahbaba and Neal (2009); Hannah et al. (2011), where the prior on $f_0(y_i, \mathbf{x}_i)$ is a DPMM. The

second type of assumption, which we call the 'conditional method', is the more common framework. Here we assume that $\mathbf{x}_{1:n}$ are fixed design points and the randomness arises from the response $y_{1:n}$, so we infer a family of conditional densities $\{f_{\mathbf{x}}(y) : \mathbf{x} \in \mathcal{X}\}$. The most common framework is the additional assumption of $y_i = g(\mathbf{x_i}) + \epsilon_i$, where ϵ_i are independent zero-mean noise, and a prior on the mean function g is assumed, e.g. a Gaussian process (Rasmussen, 2003). Alternatively, one can elicit a prior on $\{f_{\mathbf{x}}(y) : \mathbf{x} \in \mathcal{X}\}$ directly, for example with mixture models based on the dependent Dirichlet process (MacEachern, 1999). We recommend Wade (2013); Wade et al. (2014); Quintana et al. (2020) for thorough reviews.

2.6.4.1 Joint method

The joint method follows easily from the multivariate: we first estimate the joint predictive density $p_{i+1}(y, \mathbf{x})$, then compute the conditional $p_{i+1}(y \mid \mathbf{x}) = p_{i+1}(y, \mathbf{x})/p_{i+1}(\mathbf{x})$. Utilizing (2.18), we have the tractable update for the conditional density

$$p_{i+1}(y \mid \mathbf{x}) = p_i(y \mid \mathbf{x}) \frac{\left\{1 - \alpha_{i+1} + \alpha_{i+1} c_{\rho_y}(q_i, r_i) \prod_{j=1}^{d} c_{\rho_x}\left(u_i^j, v_i^j\right)\right\}}{\left\{1 - \alpha_{i+1} + \alpha_{i+1} \prod_{j=1}^{d} c_{\rho}\left(u_i^j, v_i^j\right)\right\}} \tag{2.21}$$

where

$$\begin{aligned} q_i &= P_i(y \mid \mathbf{x}), & r_i &= P_i(y_{i+1} \mid \mathbf{x}_{i+1}) \\ u_i^j &= P_i\left(x^j \mid x^{1:j-1}\right), & v_i^j &= P_i\left(x_{i+1}^j \mid x_{i+1}^{1:j-1}\right). \end{aligned} \tag{2.22}$$

Here, we can have separate bandwidths for y and $\mathbf{x}$, and even one for each dimension of $\mathbf{x}$. The updates for $q_{i+1}, r_{i+1}, u_{i+1}^j, v_{i+1}^j$ are the same as in (2.19), and again all terms are tractable. Predictive resampling in this case requires simulating both $\{Y, \mathbf{X}\} \sim P_i(y, \mathbf{x})$ just like in (2.20).

2.6.4.2 Conditional method

When $\mathbf{x}$ is high-dimensional, it may be cumbersome to model $p_n(\mathbf{x})$ when we are only interested in the conditional density. The conditional method models $p(y \mid \mathbf{x})$ directly,

and we turn to the dependent Dirichlet process (DDP) and its extensions for inspiration. In particular, consider the general covariate-dependent stick-breaking mixture model

$$f_{G_{\mathbf{x}}}(\mathbf{y}) = \int \mathcal{N}(y \mid \theta, 1)\, dG_{\mathbf{x}}(\theta), \quad G_{\mathbf{x}} = \sum_{k=1}^{\infty} w_k(\mathbf{x})\, \delta_{\theta_k^*}$$

where $w_k(\mathbf{x})$ follows an $\mathbf{x}$-dependent stick-breaking process, and $\theta_k^* \overset{\text{iid}}{\sim} \mathcal{N}(\theta \mid 0, \tau^{-1})$. A full derivation is provided in Appendix 2.10.5.5. We can show that the update step of the predictive takes the form

$$p_{i+1}(y \mid \mathbf{x}) = \left\{1 - \alpha_{i+1}(\mathbf{x}, \mathbf{x}_{i+1}) + \alpha_{i+1}(\mathbf{x}, \mathbf{x}_{i+1})\, c_{\rho_y}\left(q_i, r_i\right)\right\} p_i(y \mid \mathbf{x}) \tag{2.23}$$

where $\alpha_1(\mathbf{x}, \mathbf{x}') = \sum_{k=1}^{\infty} E\left[w_k(\mathbf{x}) w_k(\mathbf{x}')\right]$, $\rho_y = 1/(1+\tau)$ and q_i, r_i are as in (2.22). The term $\alpha_1(\mathbf{x}, \mathbf{x}')$ is tractable for some choices of the construction of $w_k(\mathbf{x})$, e.g. the kernel stick-breaking process (Dunson and Park, 2008). Unfortunately this does not provide guidance on how to generalize to $\alpha_i(\mathbf{x}, \mathbf{x}')$. Instead, we turn to the joint copula method in the previous section for inspiration, which can be written as (2.23) with

$$\alpha_i(\mathbf{x}, \mathbf{x}') = \frac{\alpha_i \prod_{j=1}^{d} c_{\rho_x}\left(u_{i-1}^j, v_{i-1}^j\right)}{1 - \alpha_i + \alpha_i \prod_{j=1}^{d} c_{\rho_x}\left(u_{i-1}^j, v_{i-1}^j\right)}.$$

This form of $\alpha_i(\mathbf{x}, \mathbf{x}')$ can be viewed as a distance measure between $\mathbf{x}$ and $\mathbf{x}'$ that is dependent on $P_n(\mathbf{x})$ which is updated in parallel. To avoid modelling $P_n(\mathbf{x})$, we can simplify the above and consider the following as a distance function directly:

$$\alpha_i(\mathbf{x}, \mathbf{x}') = \frac{\alpha_i \prod_{j=1}^{d} c_{\rho_{x^j}}\left\{\Phi\left(x^j\right), \Phi\left(x'^j\right)\right\}}{1 - \alpha_i + \alpha_i \prod_{j=1}^{d} c_{\rho_{x^j}}\left\{\Phi\left(x^j\right), \Phi\left(x'^j\right)\right\}} \tag{2.24}$$

which is equivalent to the joint method but leaving $P_n(\mathbf{x}) = P_0(\mathbf{x})$ without updating, providing us an increase in computational speed. This form requires $\mathbf{x}_{1:n}$ to be standardized for good performance, and we find that specifying independent bandwidths for each dimension in $\mathbf{x}$ works well. This method is similar to the normalized covariate-

dependent weights of Antoniano-Villalobos et al. (2014).

If $\mathbf{x}_{1:n}$ is indeed a subsequence of a deterministic sequence of design points $\mathbf{x}_1, \mathbf{x}_2, \ldots$, then predictive resampling simply involves selecting $\mathbf{x}_i$ for $i > n$ from this sequence, and drawing $[Y_{i+1} \mid \mathbf{x}_{i+1}] \sim P_i(y \mid \mathbf{x}_{i+1})$. If $\mathbf{X}_{1:n}$ is actually random and we have chosen the conditional approach simply for convenience, then we can draw the future $\mathbf{X}_{n+1:N}$ from the sequence of empirical predictives as in the Bayesian bootstrap. We have however noticed some numerical sensitivity to this choice of $P_n(\mathbf{x})$ in the uncertainty in $p_n(y \mid \mathbf{x})$ for $\mathbf{x}$ far from the observed dataset; this is illustrated in Appendices 2.10.7.5 and 2.10.7.6. Once again, conditional on $\mathbf{X}_{i+1} = \mathbf{x}_{i+1}$, we have that $P_i(Y_{i+1} \mid \mathbf{x}_{i+1}) \sim \mathcal{U}[0,1]$, so predictive resampling only consists of simulating independent uniform random variables and updating. An example of using the Bayesian bootstrap for the covariates is provided in Appendix 2.10.7.6.

2.6.4.3 Classification

For classification, both the joint and conditional approach generalize easily to when $y_i \in \{0,1\}$. To this end, we can derive the copula update for a beta-Bernoulli mixture. As shown in Appendix 2.10.5.6, this gives

$$d_{\rho_y}\{q_i, r_i\} = \begin{cases} 1 - \rho_y + \rho_y \dfrac{q_i \wedge r_i}{q_i\, r_i} & \text{if } y = y_{i+1} \\ 1 - \rho_y + \rho_y \dfrac{q_i - \{q_i \wedge (1 - r_i)\}}{q_i\, r_i} & \text{if } y \neq y_{i+1} \end{cases}$$

where $q_i = p_i(y \mid \mathbf{x}), r_i = p_i(y_{i+1} \mid \mathbf{x}_{i+1})$ and $\rho_y \in (0,1)$. We can simply replace the bivariate Gaussian copula density $c_{\rho_y}(q_i, r_i)$ in (2.21) and (2.23) with $d_{\rho_y}(u_i, v_i)$. One can check that q_i is indeed a martingale when predictive resampling, and forward sampling can be done directly as drawing binary Y_{n+1} from the Bernoulli predictive is straightforward. Unfortunately, we do not have the useful property of $P_i(y_{i+1}) \sim \mathcal{U}[0,1]$ in the discrete case, so predictive resampling beyond the Bayesian bootstrap for $\mathbf{X}_{n+1:N}$ is computationally expensive at $\mathcal{O}(N^2)$, or approximation via a grid is required. The

Bayesian bootstrap for $\mathbf{X}_{n+1:N}$ is still feasible as we only need to compute $p_N(y \mid \mathbf{x})$ at the observed $\mathbf{x}_{1:n}$. An example of this method is provided in Appendix 2.10.7.5.

2.6.5 Practical considerations

In this subsection, we discuss some practical considerations. Further details, such as those regarding sampling and optimization, are given in Appendix 2.10.6.

2.6.5.1 Initial density

For the copula methods, we require an initial guess $p_0(\mathbf{y})$ to begin our recursive updates, which can contain prior information. As it is a statement on observables, it is easier to elicit than a traditional Bayesian prior. In practice, we recommend standardizing each variable in the data $y^j_{1:n}$ to have mean 0 and variance 1 and using the default initialization $\mathcal{N}(y^j \mid 0, 1)$ for each dimension in an empirical Bayes fashion. For discrete variables, a suitable default choice is the uniform distribution over the classes. Finally, in the regression case, we can include prior information on the regression function, e.g. $p_0(y \mid \mathbf{x}) = \mathcal{N}(y \mid \beta^{\mathrm{T}}\mathbf{x}, 1)$. However, $p_0(y \mid \mathbf{x}) = \mathcal{N}(y \mid 0, 1)$ tends to work well as a default choice.

2.6.5.2 Hyperparameters

As we recommend the fixed form of α_i in (2.16), the only hyperparameter in the copula update is the constant ρ which parameterizes the bivariate normal copula in (2.15). While Hahn et al. (2018) suggest a default choice for ρ, we prefer a data-driven approach. Fortunately, there is an obvious method to select ρ using the prequential log score of Dawid (1984), that is to maximize $\sum_{i=1}^{n} \log p_{i-1}(\mathbf{y}_i)$ for density estimation or $\sum_{i=1}^{n} \log p_{i-1}(y_i \mid \mathbf{x}_i)$ for regression, which is related to a cross-validation metric (Gneiting and Raftery, 2007; Fong and Holmes, 2020). This fits nicely into our simulative framework, as ρ is selected on how well the sequence of predictives forecasts consecutive data points, which then informs us on the future predictives for predictive resampling.

We can also specify a separate ρ_j for each dimension, which corresponds to differing length scales for the update from each conditional distribution. For optimization, gradients with respect to ρ can be computed quickly using automatic differentiation.

2.6.5.3 Permutations

Due to our relaxation of exchangeability in Section 2.5.2, one downside to the copula update and c.i.d. sequences in general is the dependence of p_n on the permutation of $y_{1:n}$ when there is no natural ordering of the data. For permutation invariance, we can average p_n and the corresponding prequential log-likelihood over M random permutations of $y_{1:n}$. We find in practice that $M = 10$ is sufficient, which is computationally feasible for moderate n due to the speed of the copula update, and the method is also parallelizable over permutations. For predictive resampling, we then begin with the permutation averaged p_n and forward sample with the copula update. From asymptotic exchangeability in Theorem 2.3 in Section 2.8.1, averaging over permutations is not required for forward sampling provided N is chosen to be sufficiently large. Theoretical properties of permutation averaging are explored in Tokdar et al. (2009); Dixit and Martin (2019), which we do not consider here.

2.6.5.4 Computational complexity

For computing $p_n(\mathbf{y})$ in the multivariate copula method, there is an overhead of first computing v_i^j for $j \in \{1, \ldots, d\}$, $i \in \{0, \ldots, n-1\}$ using (2.19), which requires $\mathcal{O}(n^2 d)$ operations, followed by $\mathcal{O}(nd)$ operations to compute $p_n(\mathbf{y})$ at a single $\mathbf{y}$ (which is then parallelizable). After computing $p_n(\mathbf{y})$, predictive resampling N future observables requires $\mathcal{O}(Nd)$ for each sample of $p_N(\mathbf{y})$; this is fully parallelizable across test points and posterior samples. Interestingly, we first compute $p_n(\mathbf{y})$ and only predictively resample after if uncertainty is desired, allowing for large computational savings if we are only interested in prediction. The regression methods have a similar computational cost.

2.7 Illustrations

In this section, we demonstrate the martingale posteriors induced by the copula methods of the previous section. Code for all experiments is available online at `https://github.com/edfong/MP`. We will demonstrate the copula method on examples where θ_0 is the density itself or the loss function induces a simple parameter, e.g. quantiles. However, any θ_0 of interest (as in Section 2.4.2.2) can technically be computed directly from the density or from $y_{1:n}$ and samples of $Y_{n+1:\infty}$, although this may require a high-dimensional grid or relatively expensive sampling. As a result, for cases with complex loss functions that do not rely on the smoothness of F_∞ (e.g. a parametric log-likelihood), we recommend the Bayesian bootstrap instead as a computationally efficient predictive resampling approach. For examples regarding the Bayesian bootstrap, we refer the reader to the references in Section 2.4.4, and we qualitatively compare the Bayesian bootstrap and the copula methods in Section 2.9.

For all examples, we follow the recommendations of Section 2.6.5 for P_0 and averaging over permutations. We will demonstrate the monitoring of convergence to P_∞, but we set $N = n + 5000$ as a standard default for the number of forward samples, where n is the size of the dataset. All copula examples are implemented in JAX (Bradbury et al., 2018), which is a Python package popular in the machine learning community. JAX is ideal for our copula updates: its just-in-time compilation facilitates a dramatic speed-up for our iterative updates especially on a GPU, and its efficient automatic differentiation allows for quick hyperparameter selection. Note that the first execution of code induces an overhead compilation time of between 10-20 seconds for all examples. We carry out all copula experiments on an Azure NC6 Virtual Machine, which has a one-half Tesla K80 GPU card. The copula methods consist of many parallel simple computations on a matrix of density values, which is very suitable for a GPU, unlike traditional MCMC. The DPMM with MCMC examples are implemented in the `dirichletprocess` package (Ross and Markwick, 2018), which utilizes Gibbs

sampling. Other benchmarks are implemented in `sklearn` (Pedregosa et al., 2011). Unless otherwise stated, default hyperparameter values are set for baselines. As the baseline packages are designed for CPU usage, we run them on a 2.6 GHz 6-Core Intel Core i7-8850H CPU. Further details can be found in Appendix 2.10.7.2.

2.7.1 Density estimation

2.7.1.1 Univariate Gaussian mixture model

We begin by demonstrating the validity of the martingale posterior uncertainty returned from predictive resampling by comparing to a traditional DPMM in a simulated example, where the true density is known. We also discuss the monitoring of convergence of predictive resampling. For the data, we simulate $n = 50$ and $n = 200$ samples from a Gaussian mixture model with density $f_0(y) = 0.8\,\mathcal{N}(y \mid -2, 1) + 0.2\,\mathcal{N}(y \mid 2, 1)$. For all plots, we compute the copula predictive $p_n(y)$ on an even grid of size 160. Figures 2.3 and 2.4 show the martingale posterior density using the copula method for $n = 50$ and $n = 200$ respectively, compared to the traditional DPMM of Escobar and West (1995) with MCMC. We draw $B = 1000$ samples for both methods. We see that the resulting uncertainty and posterior means are comparable between the copula and DPMM, and the uncertainty decreases as n increases. The true density is largely contained within the 95% credible intervals.

For predictive resampling with the copula method, we judge convergence by considering the L_1 distance between the forward sampled p_N and initial p_n. This is demonstrated in Figure 2.5 for a single forward sample for $n = 50$. On the left, we have a numerical estimate of $\|p_N - p_n\|_1$ which converges to a constant, and likewise for $\|P_N - P_n\|_1$ on the right, where $\|\cdot\|_1$ is the L_1 norm and is computed on the grid. We see in this example that $N = n + 5000$ is sufficiently large for p_N to approximate p_∞. When we are not plotting on a grid and instead predicting over some test set, we

may instead monitor

$$\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |p_N(y_i) - p_n(y_i)|.$$

Optimization of the prequential log-likelihood gives us the optimal hyperparameter $\rho = 0.77$ and 0.78 for $n = 50$ and 200 respectively. The prequential log-likelihood is returned easily from the copula method, allowing for easy hyperparameter selection. However, computing the marginal likelihood for the DPMM is non-trivial, and thus setting the hyperparameters of the priors in a data-driven way, that is empirical Bayes, remains a difficult task. Here, we select the DPMM hyperparameters to match the smoothness of the posterior mean of the copula method for comparability of the uncertainty.

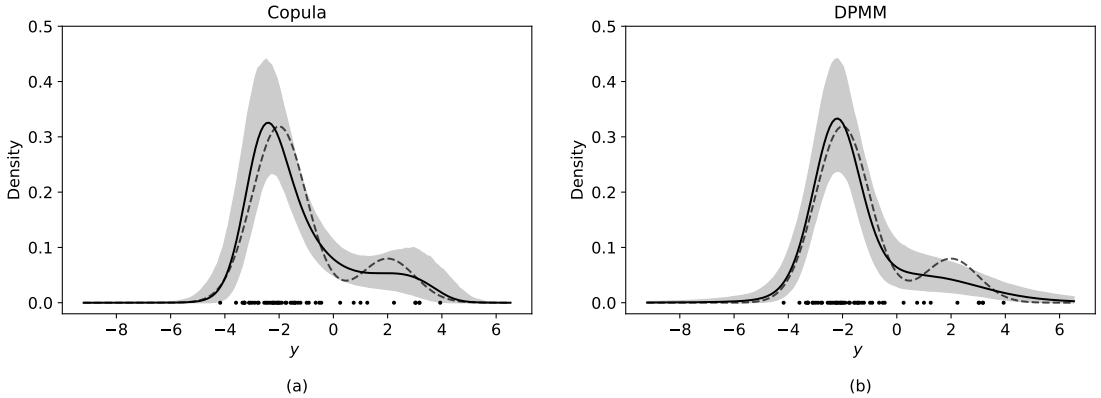


Figure 2.3: Posterior mean (—) and 95% credible interval (■) of (a) $p_N(y)$ for the copula method and (b) $p_\infty(y)$ for the DPMM, for $n = 50$ with true density (- - -) and data (•)

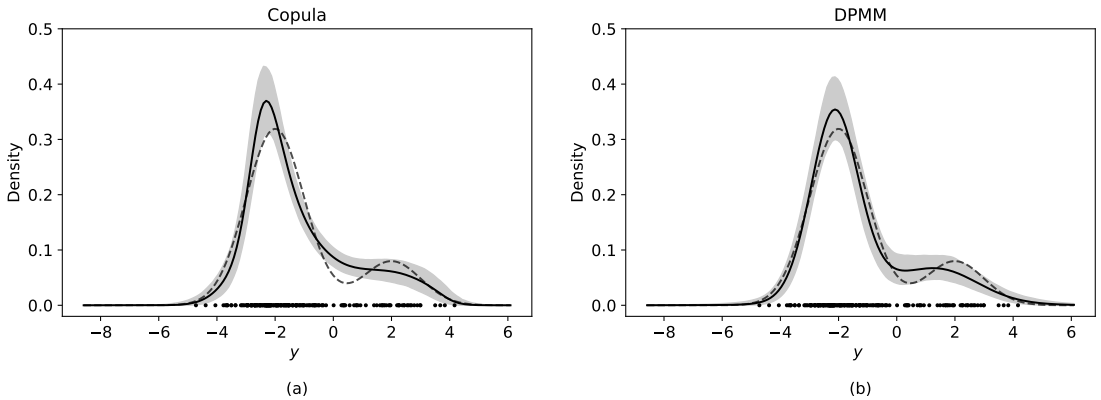


Figure 2.4: Posterior mean (—) and 95% credible interval (■) of (a) $p_N(y)$ for the copula method and (b) $p_\infty(y)$ for the DPMM, for $n = 200$ with true density (- - -) and data (•)

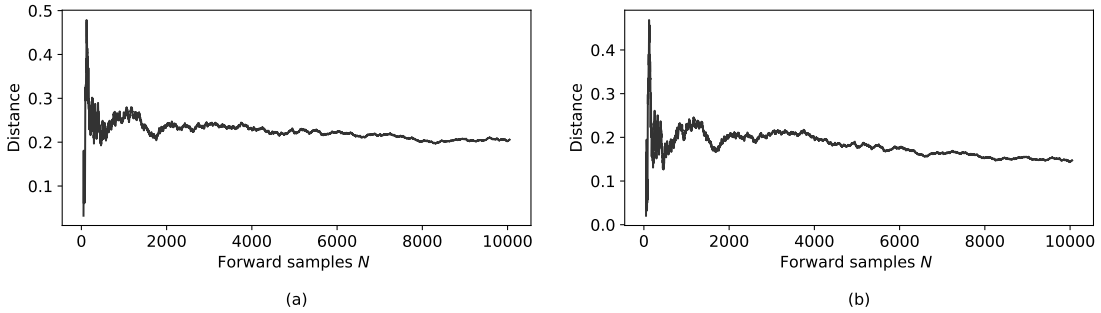


Figure 2.5: Estimated L_1 distance (a) $\|p_N - p_n\|_1$ and (b) $\|P_N - P_n\|_1$ for a single forward sample for $n = 50$

2.7.1.2 Univariate galaxy dataset

We now demonstrate the martingale posterior sampling of a parameter of interest that requires a smooth density, through predictive resampling and the computation of $\theta(P_N)$. We analyze the classic 'galaxy' dataset (Roeder, 1990), thereby extending the example of Hahn et al. (2018) to the predictive resampling framework. The dataset consists of $n = 82$ velocity measurements of galaxies in the Corona Borealis region. For all plots, we compute $p(y)$ on an even grid of size 200, and unnormalize after the copula method so that the scale of y is in km/sec.

Figure 2.6 compares predictive resampling with the copula method for $B = 1000$ posterior samples of p_N, where the selected bandwidth is $\rho = 0.93$. The bandwidth for KDE was computed through 10-fold cross-validation, and DPMM hyperparameters are set to the suggested values in West (1991). The 95% credible intervals and posterior mean of the copula approach are comparable with the DPMM. Excluding compilation times, the optimization for ρ and computation of $p_n(y)$ on the grid of size 200 took 0.5 seconds, and predictive resampling took 2 seconds. In comparison, DPMM with MCMC took 25 seconds for the same number of posterior samples, where the samples are not independent; the plots for MCMC are thus produced with $B = 2000$. Given this random density, we can also compute the statistics of interest θ directly from the grid of density values. Martingale posterior samples of the number of modes and 10% quantiles of the random density are shown in Figure 2.7, with comparison to the

DPMM. Here the copula method tends to prefer 4 modes, while the DPMM prefers 5.

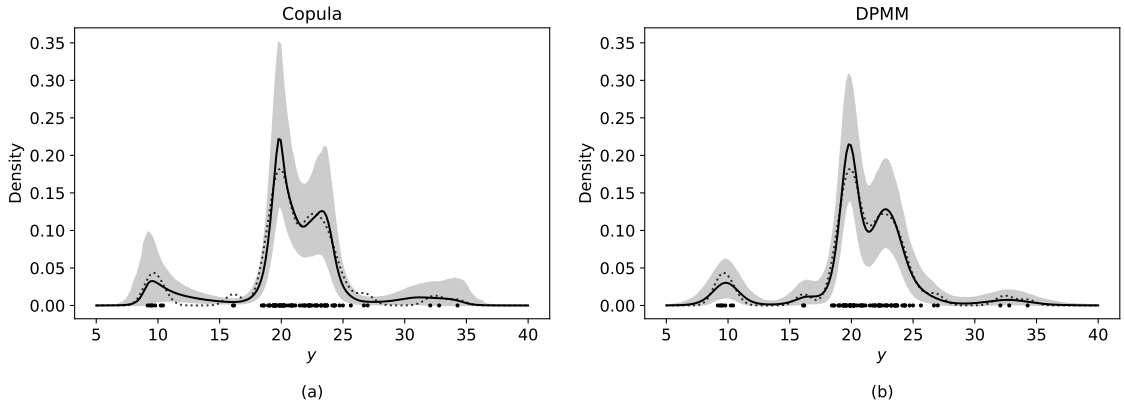


Figure 2.6: Posterior mean (—) and 95% credible interval (■) of (a) $p_N(y)$ for the copula method and (b) $p_\infty(y)$ for the DPMM, with KDE (·····) and data (•)

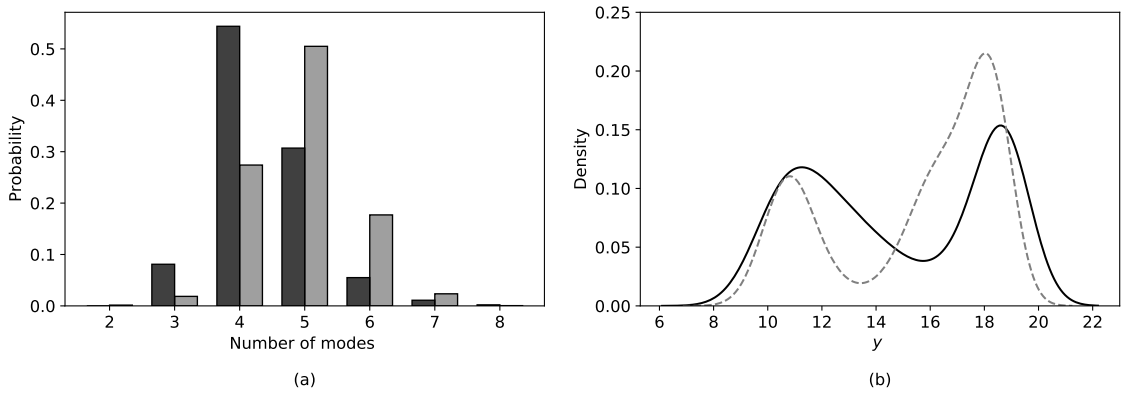


Figure 2.7: (a) Posterior samples of number of modes for the copula method (■) and DPMM (■); (b) Posterior density of 10% quantiles for the copula method (—) and the DPMM (- - -)

2.7.1.3 Bivariate air quality dataset

We demonstrate the martingale posterior for bivariate data using the method of Section 2.6.3.1, which has large computational gains over posterior sampling with DPMM when the density is of interest, where the latter is expensive due to dimensionality. For this, we look at the 'airquality' dataset (Chambers, 2018) from `DPpackage`. The dataset consists of daily ozone and solar radiation measurements in New York, with $n = 111$ completed data points. For all plots, we compute $p_n(\mathbf{y})$ on a grid of size 25×25.

We fit the multivariate copula method of Section 2.6.3.1 with one bandwidth per dimension, and optimizing the prequential log-likelihood returns $\rho = [0.47, 0.82]$.

Predictive resampling $B = 1000$ samples returns us the martingale posterior mean and standard deviation of the bivariate density as shown in Figure 2.8. Again excluding compilation times, the optimization for ρ and computation of $p_n(y)$ on the grid of size 625 took 1 second, and predictive resampling took 10 seconds in total. For comparison, the DPMM with MCMC required 4 minutes for the same number of samples. Further details and comparisons to the DPMM are given in Appendix 2.10.7.4.

Figure 2.9 plots a martingale posterior sample of the density, with the corresponding L_1 distance convergence plot. We see that $N = 5000$ is again sufficient, which suggests a dimension independent convergence rate of $P_N \to P_\infty$. This is justified in the theory in Section 2.8.

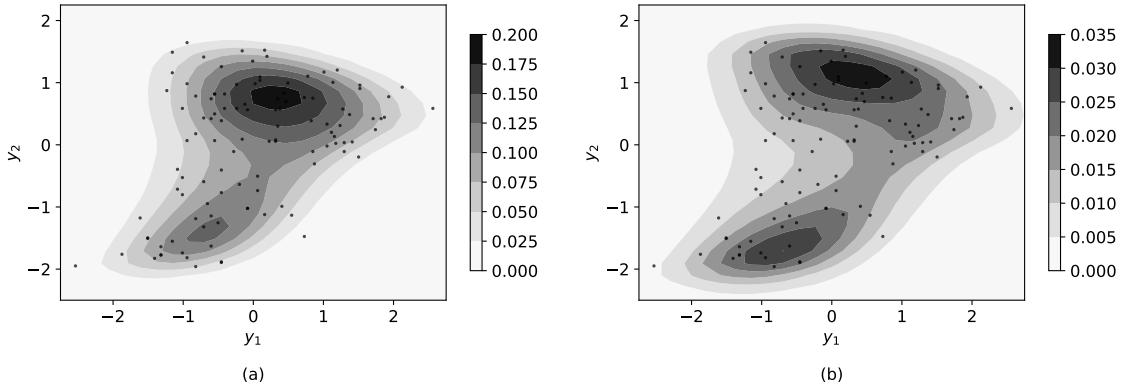


Figure 2.8: Posterior (a) mean and (b) standard deviation of $p_N(\mathbf{y})$ for the copula method with scatter plot of data (•)

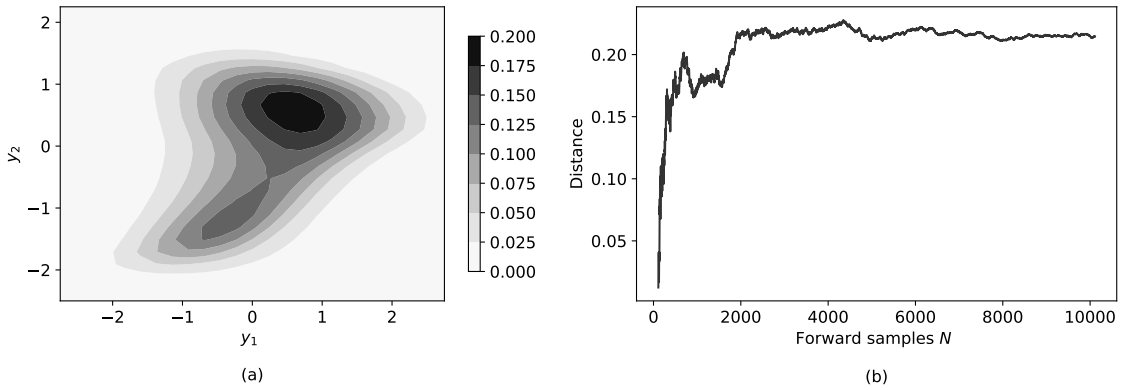


Figure 2.9: (a) Random sample of $p_N(\mathbf{y})$; (b) Corresponding estimated $||p_N - p_n||_1$

2.7.1.4 Multivariate UCI datasets

In this section, we demonstrate the multivariate copula method of Section 2.6.3.1 as a highly effective density estimator compared to the usual DPMM, as we do not need to deal with the posterior sampling or integration over high-dimensional parameters. We demonstrate on multivariate datasets from the UCI Machine Learning Repository (Dua and Graff, 2017). To prevent misleadingly high density values, we remove non-numerical variables and one variable from any pairs with Pearson correlation coefficient greater than 0.98 (e.g. see Tang et al. (2012)). We compare to the KDE, DPMM and multivariate Gaussian, and evaluate the methods with a 50-50 test-train split and average the test log-likelihoods over 10 random splits.

For the copula method, we use a single value of ρ for all dimensions for a fair comparison to the KDE. We find that having distinct $\rho_{1:d}$ slightly improves predictive performance at the cost of higher optimization times. For the KDE, we use a single scalar bandwidth set through 10-fold cross-validation. For the DPMM, we set the Gaussian kernel to have diagonal covariance matrices and use VI (Blei and Jordan, 2006). Using a full covariance matrix kernel is unreliable likely due to local optima for VI, and MCMC is too computationally expensive for large d. For the multivariate Gaussian, we use the empirical mean and covariance.

Dataset	n	d	Gaussian	KDE	DPMM (VI)	Copula
Breast cancer	569	26	−17.8 (0.61)	−25.6 (0.29)	−33.4 (0.80)	**-13.0** (0.26)
Ionosphere	351	32	−49.4 (1.97)	−32.3 (0.79)	−36.5 (0.59)	**-21.5** (1.63)
Parkinsons	195	16	−14.3 (0.54)	−15.6 (0.41)	−25.7 (0.92)	**-9.9** (0.28)
Wine	178	13	−16.1 (0.26)	−15.7 (0.20)	−22.8 (0.61)	**-14.6** (0.17)

Table 2.1: Average test log-likelihood, standard errors (in brackets) and best performance in bold

As shown in Table 2.1, the performance is significantly better on test data for these datasets. The better performance than the KDE is likely due to the regularizing effect of $p_0(\mathbf{y})$, which is important here as n is only of moderate size. The DPMM (VI) likely performs poorly as the diagonal covariance cannot capture dependent structure,

and the number of variational parameters is still high so optimization is difficult. We provide a more detailed analysis of the degradation in performance with dimensionality of the DPMM with VI in Appendix 2.10.7.7, where the copula method remains robust to dimensionality.

Overall, the run-times for the copula method, KDE and DPMM (VI) are similar, all of which are orders of magnitude faster than the DPMM with MCMC. For a single train-test split, the slowest example of the above (Breast cancer) for the copula method required less than 4 seconds in total to optimize ρ, while computing the overhead v_i^j and predicting on the test data required less than 100ms. For the same example, the KDE and DPMM (VI) required around 1.5 and 6 seconds respectively.

2.7.2 Regression and classification

2.7.2.1 Regression in LIDAR dataset

We now demonstrate the joint copula regression method of Section 2.6.4.1 on a non-linear heteroscedastic regression example, where the copula method performs well off-the-shelf. We use the LIDAR dataset from Wasserman (2006), which consists of $n = 221$ observations of the distance travelled by the light and the log ratio of intensity of the measured light from the two lasers; the latter is the dependent variable. For the plots below, we evaluate the conditional density on a y, x grid of 200×40 points.

For the copula method, we optimize the prequential conditional log-likelihood over the $M = 10$ permutations, and get $\rho_y = 0.90, \rho_x = 0.83$. The *predictive* mean and 95% central interval of $p_n(y \mid x)$ are shown in Figure 2.10, compared to the DPMM, and we observe that the copula methods handle the nonlinearity better. The optimization, fitting and prediction on the grid took under 4 seconds for the copula method, compared to 5 minutes for the DPMM with MCMC for the same number of samples.

In Figure 2.11, we see martingale posterior samples of $p_N(y \mid x = 0)$ for the copula method compared to the DPMM. For reference, predictive resampling the $B = 1000$

martingale posterior samples on the y grid for a single x took under 3 seconds. One can see in Figure 2.11 that there is more posterior uncertainty in the density $p_N(y \mid x = 0)$ for the copula methods, as the DPMM has a simpler mean function (weighted sum of linear). Convergence of the conditional density under predictive resampling is now dependent on the value of x. Figure 2.13(b) shows the L_1 distances as before for $x = 0$; however, we find that more forward samples are needed for x far from the data. Figure 2.12 then shows martingale posterior samples of $p_N(y \mid x = -3)$ where x is far from the data, and we see that both the copula and DPMM method have larger uncertainty as expected. However, predictive resampling for the conditional copula method of Section 2.6.4.2 does not always demonstrate this desirable behaviour for outlying x; the joint and conditional methods are compared in Appendix 2.10.7.6 and this undesirable behaviour is also noted in Appendix 2.10.7.5.

One may also be interested in the uncertainty in a point estimate for the function which we write as θ_x, in this case the conditional median. In Figure 2.13(a), we plot the martingale posterior mean and 95% credible interval of the conditional median of $P_N(y \mid x)$, where we see the uncertainty increasing with x . Here we predictively resample on a y, x grid of size 40×40 and compute the median numerically; this took 12 seconds for $B = 1000$ samples.

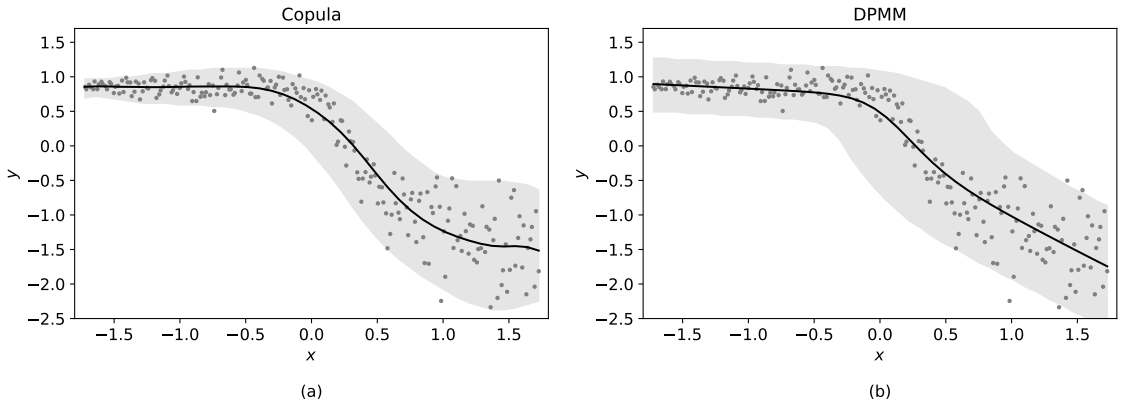


Figure 2.10: $p_n(y \mid x)$ (—) with 95% predictive interval (■) for the (a) joint copula method and (b) joint DPMM, with data (•)

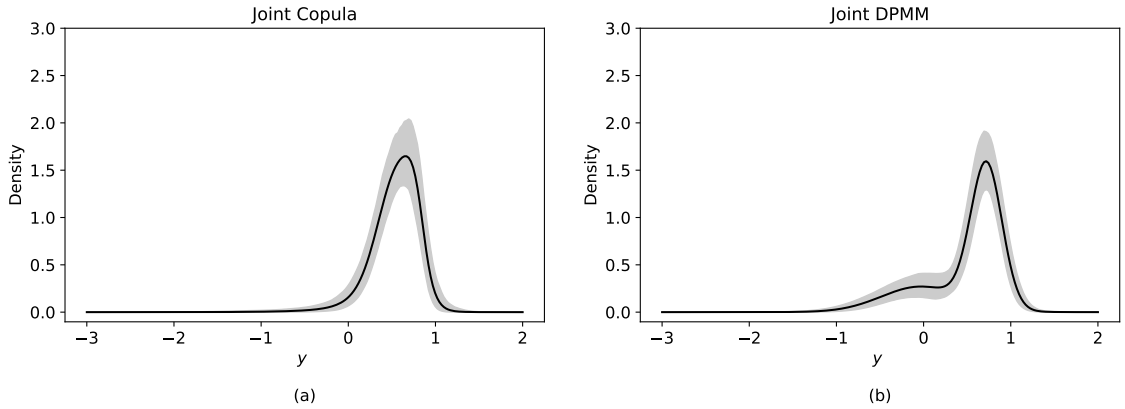


Figure 2.11: Posterior mean (—) and 95% credible interval (■) of (a) $p_N(y \mid x = 0)$ for the joint copula method and (b) $p_\infty(y \mid x = 0)$ for the joint DPMM

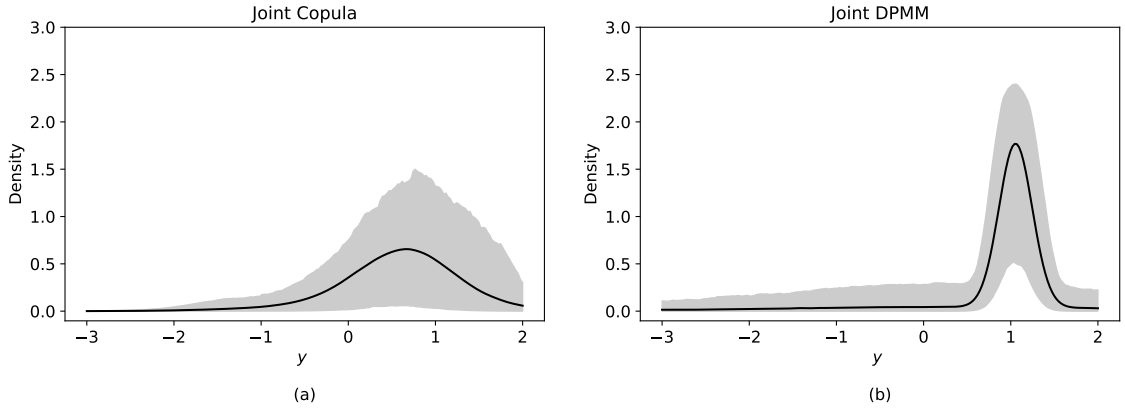


Figure 2.12: Posterior mean (—) and 95% credible interval (■) of (a) $p_N(y \mid x = -3)$ for the joint copula method and (b) $p_\infty(y \mid x = -3)$ for the joint DPMM

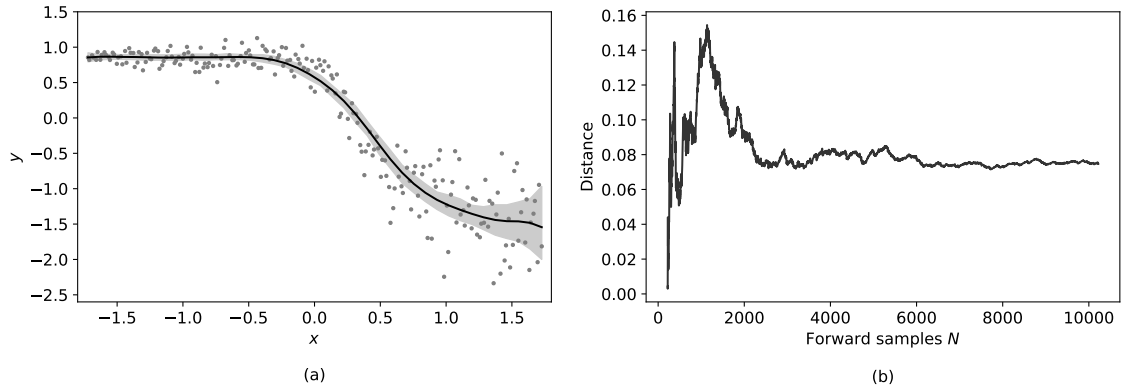


Figure 2.13: (a) Posterior mean (—) and 95% credible interval (■) of the conditional median of $P_N(y \mid x)$, with data (•); (b) Estimated L_1 distance $\|p_N(\cdot \mid x) - p_n(\cdot \mid x)\|_1$ for a single forward sample with $x = 0$

2.7.2.2 Multivariate covariates in UCI datasets

We now demonstrate the conditional copula method for prediction in the regression and classification setting with multivariate covariates, which is of particular interest to the machine learning community. For high-dimensional covariates, the conditional copula method performs better than the joint method, both in terms of computational speed and test log-likelihood. This is likely due to the dominance of estimating $P_n(\mathbf{x})$ in high dimensions, which disrupts the estimate of $P_n(y \mid \mathbf{x})$.

Similar to the multivariate density estimation, we demonstrate the regression and classification conditional copula methods on UCI datasets with scalar y and multivariate $\mathbf{x}$. Again, we evaluate the methods with 10 random 50-50 test-train splits and evaluate the average test conditional log-likelihoods. We convert categorical variables into dummy variables, and report the preprocessed covariate dimensionality in Table 2.2. We compare to Bayesian linear regression and Gaussian processes (GP) with a single length scale RBF kernel as baselines for regression, and similarly to logistic regression and GPs with the logistic link and Laplace approximation for classification. We use the Laplace approximation as it is available off-the-shelf in `sklearn`, and we found that independent kernel length scales (ARD) performed worse due to overfitting given n is moderate. For the conditional copula method, we have distinct bandwidths $\rho_{1:d}$ for each covariate, which we optimize through the prequential log-likelihood over $M = 10$ permutations.

	Dataset	n	d	Linear	GP	Copula
Regression	Boston	506	13	−0.842 (0.043)	−0.404 (0.040)	**−0.351** (0.025)
	Concrete	1030	8	−0.965 (0.008)	**−0.364** (0.014)	−0.445 (0.013)
	Diabetes	442	10	−1.096 (0.017)	−1.089 (0.015)	**−1.003** (0.018)
	Wine Quality	1599	11	−1.196 (0.017)	**−0.497** (0.034)	−1.143 (0.020)
Classification	Breast cancer	569	30	−0.107 (0.005)	−0.105 (0.005)	**−0.096** (0.008)
	Ionosphere	351	33	−0.348 (0.005)	**−0.304** (0.006)	−0.388 (0.016)
	Parkinsons	195	22	−0.352 (0.007)	−0.364 (0.013)	**−0.257** (0.010)
	Statlog	1000	20	**−0.530** (0.009)	−0.542 (0.011)	−0.541 (0.006)

Table 2.2: Average test log-likelihood, standard errors (in brackets) and best performance in bold

In Table 2.2, we see the test log-likelihoods, where the copula method is competitive with the GP, though in general we find that the GP provides a better estimate for the mean function for regression. Again, optimization took the most time due to the d bandwidths, taking on average 30 seconds per fold for the slowest example ('Statlog'). The time for actual fitting and prediction on the test set was under 120ms per fold for all examples. The GP on the slowest examples required around 20 seconds per fold for the marginal likelihood optimizations, but computation time scales as $\mathcal{O}(n^3)$.

2.8 Theory

In this section, we provide a theoretical analysis of the martingale posteriors and predictive resampling using the copula update introduced in Section 2.6. We utilize the theory of c.i.d. sequences from the works of Berti et al. (2004, 2013). We then show frequentist consistency (with little n) under relatively weak conditions for the multivariate copula update by extending the proof of Hahn et al. (2018), and we discuss its implications. All proofs are deferred to Appendix 2.10.4.

2.8.1 Martingale posteriors for copula density estimation

We first analyze the properties under predictive resampling of the multivariate copula recursive update for the martingale posterior. We write $P_i(\mathbf{y})$ as the joint cumulative distribution function of the density $p_i(\mathbf{y})$ with update (2.17), and consider predictive resampling starting at $p_n(\mathbf{y})$ such that $\mathbf{Y}_{i+1} \sim P_i(\mathbf{y})$ for $i = n, n+1 \ldots, N$. As before, n corresponds to the number of observed data points, whereas $N - n$ corresponds to the number of forward samples drawn from predictive resampling. The first two results follow directly from the c.i.d. property of the sequence.

Theorem 2.3. *[Berti et al. (2004, Theorem 2.5)] The sequence* $\mathbf{Y}_{N+1}, \mathbf{Y}_{N+2}, \ldots$ *is asymptotically exchangeable, that is*

$$(\mathbf{Y}_{N+1}, \mathbf{Y}_{N+2}, \ldots) \xrightarrow{d} (\mathbf{Z}_1, \mathbf{Z}_2, \ldots)$$

for $N \to \infty$*, where* $(\mathbf{Z}_1, \mathbf{Z}_2, \ldots)$ *is exchangeable.*

The above justifies that we may not need to average over permutations for sufficiently large N when predictive resampling.

As mentioned in Section 2.5.2, we would like $P_N(\mathbf{y}) \to P_\infty(\mathbf{y})$ at each $\mathbf{y} \in \mathbb{R}^d$, which indeed holds for predictive resampling here from the c.i.d. sequence:

Theorem 2.4. *[Berti et al. (2004, Lemma 2.1, 2.4)] There exists a random probability measure* P_∞ *such that* P_N *converges weakly to* P_∞ *almost surely.*

Specifically for the univariate case of the copula update above, we can strengthen this to convergence in total variation, which also implies that the limiting predictive P_∞ is continuous, following from an interesting result in Berti et al. (2013).

Theorem 2.5. *For* $y \in \mathbb{R}$*, suppose the sequence of probability measures* P_N *has density function* $p_N(y)$ *and cumulative distribution function* $P_N(y)$ *satisfying the updates* (2.14)*. Let us assume that the initial* $P_n(y)$ *is continuous and its density satisfies*

$$\int_K p_n^2(y)\, dy < \infty$$

for all K*, where* K *is a compact subset of* $\mathbb{R}$ *with finite Lebesgue measure. For the sequence*

$$\alpha_i = \left(2 - \frac{1}{i}\right) \frac{1}{i+1},$$

let us assume further that $\rho < 1/\sqrt{3}$*. We then have*

(a) P_∞ *is absolutely continuous with respect to the Lebesgue measure almost surely, with density* p_∞*.*

(b) *P_N converges in total variation to P_∞ almost surely, that is*

$$\lim_{N\to\infty} \int |p_N(y) - p_\infty(y)|\, dy = 0 \quad \text{a.s.}$$

The assumptions hold if $p_n(y)$ is continuous. From this, we are justified in using $p_N(y)$ as an approximate sample of the martingale posterior $p_\infty(y)$. We conjecture that the choice of $\rho < 1/\sqrt{3}$ can be relaxed, and empirically it seems the case. Furthermore, this restriction on ρ is not needed if $\alpha_i = (i+1)^{-1}$. Unfortunately, we have been unable to extend Theorem 2.5 to the multivariate copula update, as the update for $P\left(y^j \mid y^{1:j-1}\right)$ is not as easy to bound. We also conjecture that the L_1 convergence holds true in the multivariate case, and again the empirical results suggest so.

We can also quantify to some degree the convergence rate to P_∞ as we predictively resample. We have the following result from a variant of the Azuma-Hoeffding inequality from McDiarmid (1998).

Proposition 2.1. *For $M > N$ and any $\epsilon \geq 0$, the cumulative distribution function $P_N(\mathbf{y})$ of the density in* (2.17) *satisfies*

$$\sup_{\mathbf{y}} \mathbb{P}\left(|P_M(\mathbf{y}) - P_N(\mathbf{y})| \geq \epsilon\right) \leq 2\exp\left(\frac{-\epsilon^2}{\frac{2\epsilon\alpha_{N+1}}{3} + \frac{1}{2}\sum_{i=N+1}^{M}\alpha_i^2}\right).$$

Taking the limit (superior) as $M \to \infty$ of the above gives insight into the quality of the approximation of P_∞ when we truncate the predictive resampling at P_N. For our choice of α_i from (2.16), we have $\sum_{i=N+1}^{\infty}\alpha_i^2 = \mathcal{O}(N^{-1})$, so the limiting probability of a difference greater than ϵ decreases roughly at rate $\exp(-\epsilon^2 cN)$ for some constant c. Notably, this rate is independent from the dimensionality d, and instead depends only on the sequence α_i. Furthermore, we have some notion of posterior contraction in Proposition 2.1 if we instead consider N as the number of observed data points and M as the number of forward samples.

2.8.2 Martingale posteriors for conditional copula regression

For the regression case where $y \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^d$, we analyze the update given in (2.23) and (2.24). Assuming we have observed $y_{1:n}, \mathbf{x}_{1:n}$, we draw the sequence $\mathbf{X}_{n+1:\infty}$ from the Bayesian bootstrap with $\mathbf{x}_{1:n}$. While this is no longer the traditional c.i.d. setup, we still have that $P_N(y \mid \mathbf{x})$ is a martingale under predictive resampling, so we have that $P_N(y \mid \mathbf{x})$ converges pointwise for each $\mathbf{x}$ almost surely. Fortunately, Berti et al. (2006, Theorem 2.2) assures that the martingale posterior $P_\infty(y \mid \mathbf{x})$ exists.

Theorem 2.6. *For each $\mathbf{x} \in \mathbb{R}^d$, there exists a random probability measure $P_\infty(\cdot \mid \mathbf{x})$ such that $P_N(\cdot \mid \mathbf{x})$ converges weakly to $P_\infty(\cdot \mid \mathbf{x})$ almost surely.*

We also have the appropriate extension to Proposition 2.1 below.

Proposition 2.2. *For $M > N$ and any $\epsilon \geq 0$, the cumulative distribution function $P_N(y \mid \mathbf{x})$ of the density in (2.23) satisfies*

$$\sup_y \mathbb{P}\left(|P_M(y \mid \mathbf{x}) - P_N(y \mid \mathbf{x})| \geq \epsilon\right) \leq 2 \exp\left(\frac{-\epsilon^2}{\frac{4\epsilon C \alpha_{N+1}}{3} + 2C^2 \sum_{i=N+1}^{M} \alpha_i^2}\right)$$

for each $\mathbf{x} \in \mathbb{R}^d$, where C depends only on ρ and $\mathbf{x}$.

It can be shown that C increases as $\mathbf{x}$ moves from the origin. Assuming $x_{1:n}$ is standardized, this implies that the number of forward samples needed for convergence may increase as $\mathbf{x}$ shifts away from the data. The above results can also be easily extended to the classification scenario.

2.8.3 Frequentist consistency of copula density estimation

To simulate from the martingale posterior given $\mathbf{Y}_{1:n}$, we start with the density p_n computed from (2.17), so we would like to verify that it is indeed an appropriate predictive density. In this subsection, we thus concern ourselves with the frequentist notion of consistency, that is we look at the properties of the density estimate p_n

assuming $\mathbf{Y}_{1:n}$ is i.i.d. from some probability distribution with density function f_0 as we take $n \to \infty$. It should be noted that this is distinct from the Doob-type asymptotics of predictive resampling in the previous subsections where we take $N \to \infty$.

The frequentist consistency of the univariate copula method was first discussed in Hahn et al. (2018) based on the 'almost supermartingale' of Robbins and Siegmund (1971). We will now extend the result to the multivariate copula method, of which the univariate method is a special case. The full proof can be found in Appendix 2.10.4.6. Instead of the Kullback-Leibler divergence, we work with the squared Hellinger distance between probability density functions p_1 and p_2 on $\mathbf{y} \in \mathbb{R}^d$, defined as $d_{\mathrm{H}}^2(p_1, p_2) := 1 - \int \sqrt{p_1(\mathbf{y})\, p_2(\mathbf{y})}\, d\mathbf{y}$. We then have the main result.

Theorem 2.7. *For $\mathbf{Y}_{1:n} \overset{\text{iid}}{\sim} f_0$, suppose the sequence of densities $p_n(\mathbf{y})$ satisfies the updates in (2.17). Assume that $\rho \in (0,1)$, $\alpha_i = a(i+1)^{-1}$ where $a < 2/5$, and there exists $B < \infty$ such that $f_0(\mathbf{y})/p_0(\mathbf{y}) \leq B$ for all $\mathbf{y} \in \mathbb{R}^d$. We then have that p_n is Hellinger consistent at f_0, that is*

$$\lim_{n\to\infty} d_{\mathrm{H}}^2(p_n, f_0) = 0 \quad \text{a.s.}$$

Intuitively, the update (2.17) can be regarded as a stochastic gradient descent in the space of probability density functions, where α_{i+1} is the step-size. As is standard in stochastic optimization (Kushner and Yin, 2003), consistency of the copula method relies delicately on the decay of the sequence α_i, which ensures we approach the independent copula at the correct rate. A similar condition is for example discussed in Tokdar et al. (2009) for Newton's algorithm. On the one hand, we require $\sum_{i=1}^{\infty} \alpha_i = \infty$ to ensure that the initialization p_0 is forgotten. On the other hand, we require the sequence α_i to decay sufficiently quickly to 0, that is $\sum_{i=1}^{\infty} \alpha_i^2 < \infty$, for information to accumulate correctly. The requirement on a also ensures the information in later terms decay properly. Notably, the condition on $a < 2/5$ is different to the suggestion for predictive resampling, so a different choice of α_n may be more suitable when consistency

is of primary interest. The second assumption is a regularity condition on the tails of the initial p_0 being heavier than f_0, which motivates a heavy-tailed initial density as also suggested by Hahn et al. (2018). Interestingly, the bounded condition on f_0/p_0 is the only requirement on f_0 for consistency, which follows from the nonparametric update. However, unlike the KDE there are no conditions on the bandwidth ρ, which likely follows from the data-dependence of the copula kernel.

There are a number of unanswered questions when compared to the consistency of traditional Bayes. The first is whether the martingale posterior converges weakly to the Dirac measure at F_0, as we have only shown Hellinger consistency of the posterior mean measure of P_∞. We believe this is likely to be positive, as there is a notion of posterior contraction as in Proposition 2.1. A related inquiry is the rate of convergence of p_n, or the martingale posterior on p_∞, to the true f_0. The second and more ambitious question is whether the above approach provides a general method to prove consistency for other copula models. For the multivariate copula method, we only require the weak tail condition on f_0, but the proof relies heavily on the nonparametric nature of the update. It is still unclear what the conditions would be if the copula sequence corresponded to a parametric Bayesian model, such as the examples given in Hahn et al. (2018). In the absence of the prior under the predictive view, a question of interest is whether an analogue to the Kullback-Leibler property of the traditional Bayesian prior (e.g. Ghosal and van der Vaart (2017, Definition 6.15)) exists, which would highlight a predictive notion of model misspecification.

2.9 Discussion

We see that Bayesian uncertainty at its core is concerned with the missing observations required to know any statistic of interest precisely. In the i.i.d. case, this is $Y_{n+1:\infty}$, and our task is to obtain the joint distribution $p(y_{n+1:\infty} \mid y_{1:n})$, which is simplified through the factorization into a sequence of 1-step ahead predictive densities. One

open question is whether there are more general methods to elicit this joint beyond the likelihood–prior construction and the prequential factorization. For the more general data setting, the Bayesian would be tasked with eliciting $p(y_{\text{mis}} \mid y_{\text{obs}})$, where the missing observations y_{mis} would be specific to the setting and statistic of interest. We highlight that y_{mis} must be sufficiently large to compute the statistic precisely, unlike in multiple imputation (Rubin, 2004) where the imputed data is often finite and for computational convenience. For future work, identifying y_{mis} and extending the methodology in more complex data settings such as time series or hierarchical data is of primary interest.

In terms of practical methodology, it is worth comparing when one would prefer to use the Bayesian bootstrap versus the copula methods. When the data is high-dimensional but a low-dimensional statistic is of interest, the copula methods may not be suitable, as computing the density on a grid or sampling the data directly is required. Fortunately, the Bayesian bootstrap shines in this setting. On the other hand, the discreteness of the Bayesian bootstrap makes it unsuitable for when smoothness is required, for example when the density is directly of interest, or in regression where we rely on smoothness with x. In these settings, the copula methods are highly suitable. Together, the predictive framework allows us to cover a wide variety of settings with practical advantages over the traditional Bayesian approach.

We believe our framework offers interesting insight into the interplay between Bayesian and frequentist approaches. As we have seen through the lens of the Bayesian bootstrap, Bayesians and frequentists are concerned with $Y_{n+1:\infty}$ and $Y_{1:n}$ respectively. Analysis of the frequentist asymptotic properties of martingale posteriors also offers new challenges, as we must work with the predictive distribution directly, and it is unclear if the methods used in our paper generalize to other copula models. For generalizations of our martingale posterior framework, imputing aspects of the population instead of the entire population directly may also help bridge the gap between Bayesian and frequentist methods. In the hierarchical example in Section 2.3, we can in fact treat θ_i

as the mean of population i from which we observe a single sample y_i. We would thus be imputing the means of observation populations (i.e. the random effects) instead of the entire population of observables directly. This interpretation would align well with our philosophy of only imputing what one would need to carry out the statistical task.

Acknowledgements

The authors are grateful for the detailed comments of three referees and the Associate Editor on the previous version of the paper. The authors also thank Sahra Ghalebikesabi, Brieuc Lehmann, Geoff Nicholls, George Nicholson and Judith Rousseau for their helpful comments. Fong is funded by The Alan Turing Institute Doctoral Studentship, under the EPSRC grant EP/N510129/1. Holmes is supported by The Alan Turing Institute, the Health Data Research, U.K., the Li Ka Shing Foundation, the Medical Research Council, and the U.K. Engineering and Physical Sciences Research Council.

2.10 Appendix

2.10.1 Notation

In this section, we summarize the notation used in the main paper. In Table 2.3 below, we provide a summary of the notation introduced in Section 2.3 and provide some concrete examples after.

	Notation	Definition
Data	Y	One data unit as a random variable, e.g. one row in a data table
	y	Observable as a fixed realisation of Y, or input into a PDF/CDF
	n	The size of the data set, or the number of observed data units
	N	The size of the study population (or approximating ∞)
	$Y_{1:N}, Y_{1:\infty}$	The conceptual complete data table for the whole study population
Distributions	F_0	The true, unknown sampling distribution function where $Y_{1:N} \sim F_0$
	F_N, F_∞	The (limiting) empirical distribution function of the imputed population $Y_{1:N}$
	P_N, P_∞	The (limiting) predictive distribution function of the imputed $Y_{1:N}$
Parameters	Θ	Bayesian parameter as a random variable, with distribution $\Pi(\cdot)$
	$\bar{\theta}_N$	Posterior mean of Θ computed from $Y_{1:N}$
	$\Pi(\theta \mid y_{1:n})$	The conventional Bayesian posterior distribution
	θ_0	True parameter or estimand of interest, computed from F_0
	θ_N, θ_∞	The estimate of θ_0, computed from the imputed population $Y_{1:N}, Y_{1:\infty}$
	$\Pi_N(\theta_N \mid y_{1:n})$, $\Pi_\infty(\theta_\infty \mid y_{1:n})$	The (finite) martingale posterior distribution

Table 2.3: Notation for some key values

2.10.1.1 Parameters

Formally, we write $\theta(F)$ as a functional which takes as input a distribution function $F(y)$ and returns a vector in $\mathbb{R}^p$. In some cases, it can be written as

$$\theta(F) = \arg\min_{\theta} \int \ell(\theta, y)\, dF(y)$$

for a loss function of interest, where for example we may have the mean functional as

$$\theta(F) = \int y\, dF(y).$$

The true parameter/statistic of interest is then $\theta_0 = \theta(F_0)$. The input distribution function may also be the atomic empirical distribution F_N, in which case $\theta_N = \theta(F_N)$. We also use the notation $\theta(Y_{1:N})$ interchangeably with $\theta(F_N)$. In the mean example then, we have

$$\theta(Y_{1:N}) = \theta(F_N) = \frac{1}{N}\sum_{i=1}^{N} Y_i.$$

In the limiting case, we write $\theta_\infty = \theta(F_\infty) = \theta(Y_{1:\infty})$, where

$$F_\infty = \lim_{N\to\infty} F_N.$$

In some cases where $\theta(F)$ requires F to be smooth, we may instead pass in a smooth predictive distribution, that is $\theta_N = \theta(P_N)$, and similarly $\theta_\infty = \theta(P_\infty)$.

2.10.2 Bayesian inference as missing data

We illustrate the conceptual idea of treating Bayesian inference as a missing population imputation in the visualization below.

A: Population data table

Unit	A	B	C
1	10	5	8
2	5	21	13
3	13	12	17
4	6	7	10
⋮	⋮	⋮	⋮
N	21	13	11

The study ⟹

B: Observed data

Unit	A	B	C
1	10	5	8
2	?	?	?
3	?	?	?
4	6	7	10
⋮	⋮	⋮	⋮
N	?	?	?

C: Exchangeability

Unit	A	B	C
1	10	5	8
$n = 2$	6	7	10
3	?	?	?
4	?	?	?
⋮	⋮	⋮	⋮
N	?	?	?

Predictively resample ⟹

D: Imputed population 1

Unit	A	B	C
1	10	5	8
2	6	7	10
3	**4**	**20**	**12**
4	**12**	**12**	**18**
⋮	⋮	⋮	⋮
N	**19**	**15**	**12**

Predictively resample ⟹

Imputed population 2

Unit	A	B	C
1	10	5	8
2	6	7	10
3	**6**	**18**	**13**
4	**10**	**9**	**21**
⋮	⋮	⋮	⋮
N	**15**	**12**	**16**

A: The conceptual complete target population data table. If this table was known, then there would be no uncertainty in the statistic of interest, $\theta_0 = \theta(Y_{1:N})$, or in any resulting decision.
B: The experiment or observational study reveals n data units selected at random from the population data table. Uncertainty in $\theta(Y_{1:N})$ arises from the remaining missing data marked (?).
C: Following an assumption of exchangeability, we can relabel the observed units from 1 to n.
D: A predictive model allows us to impute the missing data $Y_{n+1:N} \sim p(\cdot \mid y_{1:n})$ via predictive resampling to create full synthetic data tables. The imputed synthetic data tables gives us corresponding estimates $\{\theta_N^{(1)}, \theta_N^{(2)}, \ldots\}$, which are posterior samples that characterise the Bayesian uncertainty in θ_0 arising from the missing $Y_{n+1:N}$. This notion of Bayesian inference as imputation is connected to the ideas of Rubin (1974, 2008).

2.10.3 Limiting predictive and empirical distribution

In this section, we summarize some asymptotic properties of conditionally identically distributed (c.i.d.) sequences and exchangeable sequences from the literature. In particular, we look at the equivalence of the limiting predictive and empirical distributions which we use in the main paper.

2.10.3.1 Conditionally identically distributed sequences

We begin with the more general class of c.i.d. sequences, of which exchangeable sequences are a subset. We have the following strong law for c.i.d. sequences.

Theorem 2.8. *[Berti et al. (2004, Theorem 2.2)] Suppose the sequence $Y_1, Y_2, \ldots$ is c.i.d. with respective predictive distribution functions $P_0, P_1, \ldots$, where P_N is conditional on $\mathcal{F}_N := \sigma(Y_1, \ldots, Y_N)$. We then have*

$$F_\infty(y) := \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(Y_i \leq y) = P_\infty(y) \quad a.s.$$

for each $y \in \mathbb{R}$, where $P_\infty := \lim_{N\to\infty} P_N$ is the limiting random predictive distribution function from the martingale posterior, obtained through predictive resampling as in Condition 2.1.

In summary, the limiting empirical is equivalent to the limiting predictive distribution for c.i.d. sequences (with exchangeable sequences as a special case), which justifies the interchangeability of F_∞ and P_∞ as discussed in Section 2.5.2.

Convergence of the parameter

Here we consider the convergence of parameters of interest θ_0 which take the form of

$$\theta_0 = \int g(y)\, dF_0(y)$$

for the martingale posterior with c.i.d. sequences. This form of the parameter is a

special case of the more general $\arg\min_\theta \int \ell(\theta, y)\, dF_0(y)$, which is difficult to analyze due to the stronger convergence required of the entire function $\int \ell(\theta, y)\, dF_0(y)$. For a finite predictive sample of size N, we can write the parameter estimate as

$$\theta(F_N) = \int g(y)\, dF_N(y) = \frac{1}{N}\sum_{i=1}^{N} g(y_i).$$

Alternatively, if we work directly with the predictive distribution function, we have

$$\theta(P_N) = \int g(y)\, dP_N(y).$$

The more general strong law for c.i.d. sequences assures us that $\theta(Y_{1:N})$ converges with N in both settings to θ_∞ from the martingale posterior almost surely.

Theorem 2.9. *[Berti et al. (2004, Lemma 2.1, Theorem 2.2)] Suppose the sequence $Y_1, Y_2, \ldots$ is c.i.d. with respective predictive distribution functions $P_0, P_1, \ldots$, where P_N is the predictive distribution function conditioned on $\mathcal{F}_N$. For a measurable function $g : \mathcal{Y} \to \mathbb{R}$ such that $E\left[|g(Y_1)|\right] < \infty$, we have that*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} g(Y_i) = \theta_\infty$$

almost surely and in L_1. Likewise, we have that

$$\lim_{N\to\infty} \int g(y)\, dP_N(y) = \theta_\infty$$

almost surely and in L_1. Furthermore, θ_∞ is integrable and

$$E\left[\theta_\infty \mid \mathcal{F}_N\right] = \theta(P_N).$$

However, we prefer to define the parameter as a function of F_∞ instead of a limiting parameter, that is $\theta_\infty = \theta(F_\infty)$, as F_∞ always exists for c.i.d. sequences.

2.10.3.2 Exchangeable sequences

Given the model specification of the sampling density and prior, we have random variables $(\Theta, Y_1, Y_2, \ldots)$ on some probability space which have the joint density

$$p(\theta, y_{1:N}) = \pi(\theta) \prod_{i=1}^{N} f_\theta(y_i) \tag{2.25}$$

for all N. We write our random sequence of posterior predictive distribution functions as

$$P_N(y) := P(Y_{N+1} \leq y \mid \mathcal{F}_n)$$

for all N. We then have the following equivalence result.

Theorem 2.10. *[Doob (1949); Lijoi et al. (2004)] Suppose that $(\Theta, Y_{1:N})$ are distributed according to P with density (2.25), then the sequence of predictive distribution functions satisfies*

$$P_\infty(y) := \lim_{N\to\infty} P_N(y) = F_\Theta(y) \quad \text{a.s.}$$

for each $y \in \mathbb{R}$. Furthermore, this holds when our parameter space is the family of all densities on $\mathbb{R}$, that is $\Theta = f$ where f is a random density, and we define $F_f(y) = \int_{-\infty}^{y} f(z)\, dz$.

From Theorems 2.8 and 2.10, we then have a strong law result below, which justifies using the limiting empirical distribution function F_∞ as in Section 2.4.2.2.

Theorem 2.11. *[Berti et al. (2004, Theorem 2.2)] Suppose that $(\Theta, Y_{1:N})$ are distributed according to P with density (2.25), then the sequence of empirical distribution functions satisfies*

$$F_\infty(y) := \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(Y_i \leq y) = F_\Theta(y) \quad \text{a.s.}$$

for each $y \in \mathbb{R}$, where $\mathbb{1}(A)$ is the indicator function for the event A. Again, this holds when $\Theta = f$, and $F_f(y) = \int_{-\infty}^{y} f(z)\, dz$.

2.10.4 Proofs

2.10.4.1 Corollary 2.1

From the martingale condition (2.11), we have that the product $p_{i+1}(y)\, p_i(y_{i+1}) = p_i(y, y_{i+1})$ is a bivariate density with marginals $p_i(y)$ and $p_i(y_{i+1})$, so from Theorem 2.2 there exists a bivariate copula density c_{i+1} such that $p_i(y, y_{i+1}) = c_{i+1}\{P_i(y), P_i(y_{i+1})\}\, p_i(y)\, p_i(y_{i+1})$. Dividing both sides by $p_i(y_{i+1})$ gives us the result. The reverse implication requires checking that (2.12) satisfies (2.11), which follows easily through a change of variables with $v = P_i(y_{i+1})$.

2.10.4.2 Theorem 2.5

From Berti et al. (2013, Theorem 4), we require that

$$\sup_N E\left[\int_K p_N^2(y)\, dy\right] < \infty$$

for all compact $K \subset \mathbb{R}$ in order for P_∞ to be absolutely continuous with respect to the Lebesgue measure. If this holds, then from Berti et al. (2013, Theorem 1), we know that $P_N \to P_\infty$ in total variation and $p_N \to p_\infty$ pointwise and in L_1 almost surely, and p_∞ is the density of P_∞ with respect to the Lebesgue measure.

For notational convenience, we assume we are predictive resampling starting at $p_0(y)$. If we look at the second moment conditioned on $y_{1:n}$, we have

$$\begin{aligned} E\left[p_{n+1}(y)^2 \mid y_{1:n}\right] = p_n^2(y)\, \big\{&(1-\alpha_{n+1})^2 + 2\alpha_{n+1}(1-\alpha_{n+1})\, E_v\left[c_\rho\{P_n(y), v\}\right] \\ &+ \alpha_{n+1}^2\, E_v\left[c_\rho^2\{P_n(y), v\}\right]\big\} \end{aligned}$$

where $v \sim \mathcal{U}[0,1]$. We have that

$$\int_0^1 c_\rho(u, v)\, dv = 1$$

and

$$q_\rho(u) := \int_0^1 c_\rho^2(u,v)\,dv = \frac{\exp\left(\frac{\rho^2}{1+\rho^2}\Phi^{-1}(u)^2\right)}{\sqrt{1-\rho^4}}.$$

So we have that

$$E\left[p_{n+1}^2(y) \mid y_{1:n}\right] = p_n^2(y)[1-\alpha_{n+1}^2+\alpha_{n+1}^2\, q_\rho\{P_n(y)\}].$$

From Fubini's theorem, we can write

$$E\left[\int_K p_{n+1}^2(y)\,dy \mid y_{1:n}\right] = \int_K p_n^2(y)\left[1-\alpha_{n+1}^2+\alpha_{n+1}^2 q_\rho\{P_n(y)\}\right]dy.$$

Following Hahn et al. (2018) and (Inglot, 2010, Theorem 2.1), we have

$$\Phi^{-1}(u)^2 \le -2\log(u\wedge\bar{u})$$

where $\bar{u}=1-u$. Using $P_n(y)\ge P_0(y)\prod_{i=1}^n(1-\alpha_i)$ and likewise for $\bar{P}_n$, we can upper bound

$$\begin{aligned} q_n\{P_n(y)\} &\le \frac{1}{\sqrt{1-\rho^4}}\left\{P_n(y)\wedge\bar{P}_n(y)\right\}^{-2\gamma} \\ &\le \frac{1}{\sqrt{1-\rho^4}}\left\{P_0(y)\wedge\bar{P}_0(y)\right\}^{-2\gamma}\prod_{i=1}^n(1-\alpha_i)^{-2\gamma} \end{aligned}$$

where $\gamma=\rho^2/(1+\rho^2)$. As $y\in K$ where K is compact and $P_0(y)$ is continuous, we have that $\left\{P_0(y)\wedge\bar{P}_0(y)\right\}^{-2\gamma}$ is upper bounded by $A<\infty$. Plugging this in, we have

$$E\left[\int_K p_{n+1}^2(y)\,dy \mid y_{1:n}\right] \le \left\{1-\alpha_{n+1}^2+\frac{A}{\sqrt{1-\rho^4}}\,\alpha_{n+1}^2\prod_{i=1}^n(1-\alpha_i)^{-2\gamma}\right\}\int_K p_n^2(y)\,dy.$$

We have that

$$\begin{aligned}\prod_{i=1}^{n}(1-\alpha_i)^{-2\gamma} &= \prod_{i=1}^{n}\left\{1-\left(2-\frac{1}{i}\right)\frac{1}{i+1}\right\}^{-2\gamma} \\ &= \prod_{i=1}^{n}\left(\frac{i+1}{i-1+\frac{1}{i}}\right)^{2\gamma} \\ &\leq 2^{2\gamma}\prod_{i=2}^{n}\left(\frac{i+1}{i-1}\right)^{2\gamma} \\ &= \{n(n+1)\}^{2\gamma} \\ &\leq (n+1)^{4\gamma}\end{aligned}$$

Iterating the expectation, we then have

$$E\left[\int_K p_{n+1}^2(y)\,dy\right] \leq \prod_{i=1}^{n+1}\left(1+\frac{C}{i^\kappa}\right)\int_K p_0^2(y)\,dy$$

where $\kappa = 2-4\gamma$ and $C<\infty$. We have by assumption that $\int_K p_0^2(y)\,dy$ is bounded. Finally, the product term is monotonically increasing and upper bounded by

$$\begin{aligned}\prod_{i=1}^{\infty}\left(1+\frac{C}{i^\kappa}\right) &= \exp\left\{\sum_{i=1}^{\infty}\log\left(1+\frac{C}{i^\kappa}\right)\right\} \\ &\leq \exp\left(C\sum_{i=1}^{\infty}\frac{1}{i^\kappa}\right)\end{aligned}$$

which is bounded if $4\gamma<1$ so $\kappa>1$. This implies that $\rho<1/\sqrt{3}$ is required for boundedness.

2.10.4.3 Proposition 2.1

Theorem 6.1 from Chung and Lu (2006) states that for a martingale X_i relative to $\mathcal{F}_i$, we have

$$\Pr(X_n - E[X_n] \geq \epsilon) \leq \exp\left\{\frac{-\epsilon^2}{2\left(\sum_{i=1}^{n}\sigma_i^2+\frac{M\epsilon}{3}\right)}\right\}$$

where $|X_i - X_{i-1}| \leq M$ and $\sigma_i^2 := E\left[(X_i-X_{i-1})^2 \mid \mathcal{F}_{i-1}\right]$ for $1\leq i\leq n$. The original result is by McDiarmid (1998). Considering the martingale $-X_i$ and applying the

union bound gives the two-sided inequality:

$$\Pr(|X_n - E[X_n]| \geq \epsilon) \leq 2 \exp\left\{ \frac{-\epsilon^2}{2\left(\sum_{i=1}^{n} \sigma_i^2 + \frac{M\epsilon}{3}\right)} \right\}.$$

The cumulative distribution function of the multivariate copula method satisfies

$$P_{i+1}(\mathbf{y}) = P_i(\mathbf{y})\,(1 - \alpha_{i+1}) + \alpha_{i+1} \underbrace{\int_{-\infty}^{\mathbf{y}} \prod_{j=1}^{d} c_\rho\left(u_i^j, v_i^j\right) p_n(\mathbf{y}')\, d\mathbf{y}'}_{Q_{i+1}(\mathbf{y})}$$

so we have

$$|P_{i+1}(\mathbf{y}) - P_i(\mathbf{y})| = \alpha_{i+1}\,|Q_{i+1}(\mathbf{y}) - P_i(\mathbf{y})| \leq \alpha_{i+1}$$

for all $\mathbf{y} \in \mathbb{R}^d$ as $Q_{i+1}(\mathbf{y})$ is also a CDF and lies in the interval $(0, 1)$.

When predictive resampling, we have that $v_i^j \sim \mathcal{U}[0, 1]$ independently across $j \in \{1, \ldots, d\}$, and from the property of copulas we have

$$\int_0^1 c_\rho(u, v_i^j)\, dv_i^j = 1.$$

Defining $\mathcal{F}_i = \sigma(\mathbf{Y}_1, \ldots, \mathbf{Y}_i)$, this implies that

$$E\left[Q_{i+1}(\mathbf{y}) \mid \mathcal{F}_i\right] = P_i(\mathbf{y})$$

almost surely for each $\mathbf{y} \in \mathbb{R}^d$, so $P_i(\mathbf{y})$ is a martingale with respect to $\mathcal{F}_i$.

Lemma 2.1. *For the multivariate copula method, the conditional variance of the martingale satisfies*

$$E\left[\{P_{N+1}(\mathbf{y}) - P_N(\mathbf{y})\}^2 \mid \mathcal{F}_N\right] \leq \frac{\alpha_{N+1}^2}{4} \quad a.s.$$

Proof. We have that

$$\{P_{N+1}(\mathbf{y}) - P_N(\mathbf{y})\}^2 = \alpha_{N+1}^2 \{Q_{N+1}(\mathbf{y}) - P_N(\mathbf{y})\}^2$$

where

$$Q_{N+1}(\mathbf{y}) = \int_{-\infty}^{\mathbf{y}} \prod_{j=1}^{d} c_\rho \left(u_N^j, v_N^j\right) p_N(\mathbf{y}')\, d\mathbf{y}'$$

and $Q_{N+1}(\mathbf{y})$ lies in $[0, 1]$. Putting this together gives us

$$E\left[\{Q_{N+1}(\mathbf{y}) - P_N(\mathbf{y})\}^2 \mid \mathcal{F}_N\right] = \mathrm{Var}\left[Q_{N+1}(\mathbf{y}) \mid \mathcal{F}_N\right] \leq \frac{1}{4}$$

almost surely, which follows from the maximum variance of a random variable on $[0, 1]$. □

The martingale is also bounded in difference, as $|P_{M+1}(\mathbf{y}) - P_M(\mathbf{y})| \leq \alpha_{N+1}$ for all $M \geq N$ as $\alpha_N = \left(2 - \frac{1}{N}\right) \frac{1}{N+1}$ is monotonically decreasing. McDiarmid's theorem then gives us Proposition 2.1. As the bound is independent of $\mathbf{y}$, we can take the supremum of both sides.

Numerically this is tighter than Azuma's inequality due to the extra $1/4$ before σ_i^2. Assuming the sequence $\alpha_i \leq 2(i+1)^{-1}$, for $\epsilon = 0.05$ and $N = 5000$, we have 0.42 for Azuma's inequality and 0.0047 for McDiarmid's inequality. However, decreasing ϵ further makes this bound quite loose.

2.10.4.4 Theorem 2.6

In the regression context, we are interested in the conditional distribution of Y_n given $X_n = x$, so we cannot rely on the c.i.d. result of Berti et al. (2004). Fortunately, we can use Theorem 2.2 of Berti et al. (2006) to show that $P_N(\cdot \mid \mathbf{x})$ from predictive resampling converges weakly to a random probability measure almost surely for each $\mathbf{x} \in \mathbb{R}^d$.

We consider the sequence of random probability measures $\{P_N(\cdot \mid \mathbf{x}), P_{N+1}(\cdot \mid$

$\mathbf{x}), \ldots\}$ on $S = \mathbb{R}$ defined on $(\Omega, \mathcal{A}, P)$, and begin by showing that for any $f \in C_b(S)$, we have that $P_N(f \mid \mathbf{x})$ converges almost surely, where we define

$$P_N(f \mid \mathbf{x}) := \int f(y)\, p_N(y \mid \mathbf{x})\, dy.$$

This is indeed condition (2) of Berti et al. (2006). We further write $Z_i = \{Y_i, \mathbf{X}_i\}$ and $\mathcal{F}_i = \sigma(Z_1, \ldots, Z_i)$. Now taking the conditional expectation, we have from Fubini's theorem

$$\begin{aligned} E\left[P_{N+1}(f \mid \mathbf{x}) \mid \mathcal{F}_N\right] &= \int f(y)\, E\left[p_{N+1}(y \mid \mathbf{x}) \mid \mathcal{F}_N\right] dy \\ &= P_N(f \mid \mathbf{x}) \end{aligned}$$

almost surely for each $\mathbf{x} \in \mathbb{R}^d$, as $p_N(y \mid \mathbf{x})$ is a martingale with respect to $\mathcal{F}_N$ irrespective of how we draw $\mathbf{X}_{n+1:\infty}$. As $|f(y)|$ is bounded by some $B < \infty$, we also have that

$$E\left[|P_N(f \mid \mathbf{x})|\right] \leq B$$

for all N, so $P_N(f \mid \mathbf{x})$ is a martingale with respect to $\mathcal{F}_N$ and converges almost surely. As $\mathbb{R}$ is Radon, Theorem 2.2 of Berti et al. (2006) applies, so there exists a random probability measure P_∞ on S, defined on $(\Omega, \mathcal{A}, P)$ such that $P_N(\cdot \mid \mathbf{x}) \to P_\infty$ weakly almost surely.

2.10.4.5 Proposition 2.2

Following the derivation of Proposition 2.1, we have that

$$|P_{N+1}(y \mid \mathbf{x}) - P_N(y \mid \mathbf{x})| \leq \alpha_{N+1}(\mathbf{x}, \mathbf{x}_{N+1})$$

where

$$\alpha_{N+1}(\mathbf{x}, \mathbf{x}_{N+1}) = \frac{\alpha_{N+1} \prod_{j=1}^{d} c_{\rho_j}\left\{\Phi\left(x^j\right), \Phi\left(x_{N+1}^j\right)\right\}}{1 - \alpha_{N+1} + \alpha_{N+1} \prod_{j=1}^{d} c_{\rho_j}\left\{\Phi\left(x^j\right), \Phi\left(x_{N+1}^j\right)\right\}}.$$

We have the following lemma:

Lemma 2.2. *For the conditional copula method for regression, we have that*

$$\alpha_{N+1}(\mathbf{x}, \mathbf{x}_{N+1}) \leq 2C\alpha_{N+1}$$

where

$$C = \prod_{j=1}^{d} \frac{1}{\sqrt{1-\rho_j^2}} \exp\left\{\frac{x_j^2}{2}\right\}.$$

Proof. As $c_\rho(u, v) \geq 0$, we have

$$\alpha_{N+1}(\mathbf{x}, \mathbf{x}_{N+1}) \leq \frac{\alpha_{N+1}}{1-\alpha_{N+1}} \prod_{j=1}^{d} c_{\rho_j}\left\{\Phi\left(x^j\right), \Phi\left(x_{N+1}^j\right)\right\}.$$

We can then write

$$\begin{aligned} c_\rho\{\Phi(x), \Phi(x')\} &= \frac{1}{\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\rho^2(x^2+x'^2) - 2\rho x x'\right\}\right] \\ &\leq \frac{1}{\sqrt{1-\rho^2}} \exp\left\{\frac{x^2}{2}\right\}. \end{aligned}$$

Finally, noting that $(1-\alpha_{N+1}) \geq 0.5$, we have the result. □

To get the concentration inequality, again writing $Z_i = \{Y_i, \mathbf{X}_i\}$ and $\mathcal{F}_i = \sigma(Z_1, \ldots, Z_i)$. We have that $P_N(y \mid \mathbf{x})$ is a martingale with respect to $\mathcal{F}_N$ so the following lemma and McDiarmid's theorem gives us Proposition 2.2, where the supremum again follows from the bound being independent of y.

Lemma 2.3. *For the conditional regression method, the conditional variance of the martingale satisfies*

$$E\left[\left\{P_{N+1}(y \mid \mathbf{x}) - P_N(y \mid \mathbf{x})\right\}^2 \mid \mathcal{F}_N\right] \leq C^2\alpha_{N+1}^2$$

almost surely for each $\mathbf{x} \in \mathbb{R}^d$

Proof. We have that

$$\begin{aligned}\{P_{N+1}(y \mid \mathbf{x}) - P_N(y \mid \mathbf{x})\}^2 &= \alpha_{N+1}(\mathbf{x}, \mathbf{x}_{N+1})^2 \{Q_{N+1}(y \mid \mathbf{x}) - P_N(y \mid \mathbf{x})\}^2 \\ &\leq 4C^2\alpha_{N+1}^2 \{Q_{N+1}(y \mid \mathbf{x}) - P_N(y \mid \mathbf{x})\}^2\end{aligned}$$

from the above lemma, where

$$Q_{N+1}(y \mid \mathbf{x}) = \int_{-\infty}^{y} c_\rho(q_N, r_N)\, p_N(y' \mid \mathbf{x})\, dy'$$

and $Q_{N+1}(y \mid \mathbf{x})$ lies in $[0,1]$. From predictive resampling, $r_N \sim \mathcal{U}[0,1]$ conditional on $\mathbf{X}_{N+1}$, so from the tower property we have that $E[Q_{N+1}(y \mid \mathbf{x}) \mid \mathcal{F}_N] = P_N(y \mid \mathbf{x})$. Putting this together gives us

$$E\left[\{Q_{N+1}(y \mid \mathbf{x}) - P_N(y \mid \mathbf{x})\}^2 \mid \mathcal{F}_N\right] = \mathrm{Var}\left[Q_{N+1}(y \mid \mathbf{x}) \mid \mathcal{F}_N\right] \leq \frac{1}{4}$$

almost surely for all $\mathbf{x} \in \mathbb{R}^d$, which follows from the maximum variance of a random variable on $[0,1]$. □

2.10.4.6 Theorem 2.7

In this section, we prove frequentist consistency of the multivariate copula update (2.17) in Section 2.6.3.1. For $\mathbf{y} \in \mathbb{R}^d$, $p_n(\mathbf{y})$ is an estimate of the true density $f_0(\mathbf{y})$ from which we observe samples $\mathbf{Y}_{1:n} \overset{\text{iid}}{\sim} f_0(\cdot)$, and we consider $n \to \infty$. For convenience, we restate the following assumptions:

Assumption 2.1. *We have $\rho \in (0,1)$ and $\alpha_i = a(i+1)^{-1}$ where*

$$a < \frac{2}{5}.$$

Assumption 2.2. *There exists $B < \infty$ such that*

$$\frac{f_0(\mathbf{y})}{p_0(\mathbf{y})} \leq B$$

for all $\mathbf{y} \in \mathbb{R}^d$.

The interpretation of the assumptions are discussed in the main paper. We highlight that the assumptions are slightly different to those in Hahn et al. (2018) as the 'almost supermartingale' requires different conditions in the multivariate case. We begin in the same way as Hahn et al. (2018). Define the KL divergence as

$$d_{\mathrm{KL}}(f_0, p_n) = \int \log \frac{f_0(\mathbf{y})}{p_n(\mathbf{y})} \, f_0(\mathbf{y}) \, d\mathbf{y}$$

and define

$$T(p) := \int \int \left\{ \prod_{j=1}^{d} c_\rho \left(u^j, v^j \right) - 1 \right\} f_0(\mathbf{y}) \, f_0(\mathbf{y}') \, d\mathbf{y} \, d\mathbf{y}'$$

where again

$$u^j = P\left(y^j \mid y^{1:j-1} \right), \quad v^j = P\left(y'^j \mid y'^{1:j-1} \right).$$

The inequality $\log(1+x) \geq x - 2x^2, x \approx 0$ can still be used as $\prod_{j=1}^{d} c_\rho(u^j, v^j) \geq 0$ and $\alpha_n \to 0$. We can follow through the same algebra to obtain a multivariate version of equation (17) in Hahn et al. (2018). Writing $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n)$, we have

$$\begin{aligned} &E\left\{ d_{\mathrm{KL}}(f_0, p_n) \mid \mathcal{F}_{n-1} \right\} - d_{\mathrm{KL}}(f_0, p_{n-1}) \\ &\leq -\alpha_n \int \int \left\{ \prod_{j=1}^{d} c_\rho \left(u_{n-1}^j, v_{n-1}^j \right) - 1 \right\} f_0(\mathbf{y}) \, f_0(\mathbf{y}') \, d\mathbf{y} \, d\mathbf{y}' + E(R_n \mid \mathcal{A}_{n-1}) \end{aligned}$$

almost surely, where $u_{n-1}^j = P_{n-1}(y^j \mid y^{1:j-1})$ and $v_{n-1}^j = P_{n-1}(y^{j'} \mid y^{1:j-1'})$, and

$$R_n = 2\alpha_n^2 \int \left\{ \prod_{j=1}^{d} c_\rho \left(u_{n-1}^j, v_{n-1}^j \right) - 1 \right\}^2 f_0(\mathbf{y}) \, d\mathbf{y}.$$

This is an 'almost supermartingale' in the sense of Robbins and Siegmund (1971) if $T(p_n)$ is positive and $\sum_n E[R_n \mid \mathcal{F}_{n-1}] < \infty$ almost surely.

Lemma 2.4. *For a density p with support containing that of $f_0(\mathbf{y})$, we have that $T(p) \geq 0$ with equality if and only if $p = f_0$ Lebesgue-almost everywhere.*

Proof. First we note that the copula density product can be written as

$$\prod_{j=1}^{d} c_\rho(u^j, v^j) = \int \prod_{j=1}^{d} \psi_{\theta^j}(u^j)\, \psi_{\theta^j}(v^j)\, \mathcal{N}(\theta^j \mid 0, \rho)\, d\boldsymbol{\theta}$$

where $\boldsymbol{\theta} = \{\theta^1, \ldots, \theta^d\}$ and

$$\psi_\theta(u) = \frac{\mathcal{N}(\Phi^{-1}(u) \mid \theta, 1-\rho)}{\mathcal{N}(\Phi^{-1}(u) \mid 0, 1)}.$$

This gives us

$$\begin{aligned} T(p) &= \int\int \left\{ \prod_{j=1}^{d} c_\rho\left(u^j, v^j\right) - 1 \right\} f_0(\mathbf{y})\, f_0(\mathbf{y}')\, d\mathbf{y}\, d\mathbf{y}' \\ &= \int \left[\left\{ \int \prod_{j=1}^{d} \psi_{\theta^j}(u^j) f_0(\mathbf{y})\, d\mathbf{y} \right\}^2 - 1 \right] \prod_{j=1}^{d} \mathcal{N}(\theta^j \mid 0, \rho)\, d\boldsymbol{\theta} \\ &= \int \left\{ \int \prod_{j=1}^{d} \psi_{\theta^j}(u^j) f_0(\mathbf{y})\, d\mathbf{y} - 1 \right\}^2 \prod_{j=1}^{d} \mathcal{N}(\theta^j \mid 0, \rho)\, d\boldsymbol{\theta}. \end{aligned}$$

The last line follows from $E(X^2) - E^2(X) = E\{X - E(X)\}^2$ as

$$\prod_{j=1}^{d} \int \psi_{\theta^j}(u^j)\, \mathcal{N}(\theta^j \mid 0, \rho)\, d\theta^j = 1$$

for all $\mathbf{u} = \{u^1, \ldots, u^d\} \in [0,1]^d$. The above shows that $T(p) \geq 0$.

For the second equality result, note that $T(p) = 0$ if and only if $\int \prod_{j=1}^{d} \psi_{\theta^j}(u^j)\, f_0(\mathbf{y})\, d\mathbf{y} = 1$ for Lebesgue-almost all $\boldsymbol{\theta}$. This can be strengthened to

all $\boldsymbol{\theta} \in \mathbb{R}^d$ from the continuity of $\int \prod_{j=1}^d \psi_{\theta^j}(u^j)\, f_0(\mathbf{y})\, d\mathbf{y}$, which follows from the continuity of $\psi_\theta(u)$, the upper bound

$$\psi_\theta(u) \leq \frac{1}{\sqrt{1-\rho}} \exp\left(\frac{\theta^2}{2\rho}\right)$$

and dominated convergence.

To show that $\int \prod_{j=1}^d \psi_{\theta^j}(u^j)\, f_0(\mathbf{y})\, d\mathbf{y} = 1$ for all $\boldsymbol{\theta}$ holds if and only if $p = f_0$, factorize $f_0(\mathbf{y}) = \prod_{j=1}^d {f_0}^j(y^j)$ where ${f_0}^j(y^j) = f_0(y^j \mid y^{1:j-1})$. We carry out a multivariate change of variables from $\mathbf{y}$ to $\mathbf{z} = \{z^1, \ldots, z^j\}$, where $z^j = \Phi^{-1}(u^j)$, and note the Jacobian is triangular. This gives

$$\int \prod_{j=1}^d \frac{{f_0}^j\left\{(P^j)^{-1}(\Phi(z^j))\right\}}{p^j\left\{(P^j)^{-1}(\Phi(z^j))\right\}} \mathcal{N}(z^j \mid \theta^j, 1-\rho)\, d\mathbf{z}$$

where $(P^j)^{-1}$ is the inverse CDF for $P(y^j \mid y^{1:j-1})$, $p^j(y^j) = p(y^j \mid y^{1:j-1})$, and each ratio term with z^j depends on $z^{1:j-1}$. It is clear that $T(f_0) = 0$. We now want to show that $T(p) = 0$ implies the above density ratio is 1 almost everywhere.

To do so, we point out that the multivariate normal location family $\mathcal{N}\{\mathbf{z}; \boldsymbol{\theta}, (1-\rho)I_d\}$ is complete from Lehmann and Romano (2006, Theorem 4.3.1), that is $E[g(\mathbf{z})] = 0$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$ implies $g(\mathbf{z}) = 0$ Lebesgue-almost everywhere. From this, $\int \prod_{j=1}^d \psi_{\theta^j}(u^j)\, f_0(\mathbf{y})\, d\mathbf{y} = 1$ for all $\boldsymbol{\theta}$ implies

$$\prod_{j=1}^d \frac{{f_0}^j\left\{(P^j)^{-1}(\Phi(z^j))\right\}}{p^j\left\{(P^j)^{-1}(\Phi(z^j))\right\}} = 1$$

for Lebesgue-almost all $\mathbf{z}$, so $f_0 = p$ holds Lebesgue-almost everywhere as the product of the conditionals is the joint. □

We highlight this above lemma to show that $T(p)$ has the makings of a probability divergence, which we will use later. We now prove the second requirement for the almost super-martingale.

Lemma 2.5. *Under the assumptions above, we have that* $\sum_n E\left[R_n \mid \mathcal{F}_{n-1}\right] < \infty$ *almost surely.*

Proof. We only need to bound

$$\zeta_n = \int \int \prod_{j=1}^{d} c_\rho \left(u_{n-1}^j, v_{n-1}^j\right)^2 f_0(\mathbf{y})\, f_0(\mathbf{y}')\, d\mathbf{y}\, d\mathbf{y}'.$$

Following the univariate proof of Hahn et al. (2018), we have from the mixture representation of the copula and Cauchy-Schwarz that

$$c_\rho(u,v)^2 \lesssim \exp\left(\lambda z_u^2\right) \exp\left(\lambda z_v^2\right)$$

where $\lesssim$ indicates the inequality up to a constant, $\lambda = \rho/(1+\rho)$ and $z_u = \Phi^{-1}(u)$. This then gives us

$$\zeta_n \lesssim \left[\int \prod_{j=1}^{d} \exp\left\{\lambda (z^j)^2\right\} f_0(\mathbf{y})\, d\mathbf{y}\right]^2.$$

Again applying a change of variables from $\mathbf{y}$ to $\mathbf{z} = \{z^1, \ldots, z^d\}, z^j = \Phi^{-1}(u_{n-1}^j)$ as before, we can write

$$\begin{aligned}
&\int \prod_{j=1}^{d} \exp\left\{\lambda (z^j)^2\right\} f_0(\mathbf{y})\, d\mathbf{y} \\
&\propto \int \exp\left\{\left(\lambda - \frac{1}{2}\right) \sum_{j=1}^{d} (z^j)^2\right\} \frac{f_0(\mathbf{y})}{p_{n-1}(\mathbf{y})}\, d\mathbf{z} \\
&= \int \exp\left\{\left(\lambda - \frac{1}{2}\right) \sum_{j=1}^{d} (z^j)^2\right\} \frac{f_0(\mathbf{y})}{p_0(\mathbf{y}) \prod_{i=1}^{n-1} (1 - \alpha_i + \alpha_i \prod_{j=1}^{d} c_{i,j})}\, d\mathbf{z} \\
&\leq \prod_{i=1}^{n-1} (1-\alpha_i)^{-1} \int \exp\left\{\left(\lambda - \frac{1}{2}\right) \sum_{j=1}^{d} (z^j)^2\right\} \frac{f_0(\mathbf{y})}{p_0(\mathbf{y})}\, d\mathbf{z} \\
&\leq B \prod_{i=1}^{n-1} (1-\alpha_i)^{-1} \int \exp\left\{\left(\lambda - \frac{1}{2}\right) \sum_{j=1}^{d} (z^j)^2\right\} d\mathbf{z}.
\end{aligned}$$

The third line follows from $c_{i,j} := c_\rho\left(u_{i-1}^j, v_{i-1}^j\right) \geq 0$, and the last line from Assumption 2.2. As $\lambda < \frac{1}{2}$, we have that $\int \exp\left\{\left(\lambda - \frac{1}{2}\right)\sum_{j=1}^d (z^j)^2\right\} d\mathbf{z}$ is bounded. Following through, we need to show

$$\sum_n^\infty \alpha_n^2 \left\{\prod_{i=1}^{n-1}(1-\alpha_i)^{-2}\right\} \tag{2.26}$$

converges as in original proof. The product term on the right can be written as

$$\begin{aligned}
\prod_{i=1}^{n-1}(1-\alpha_i)^{-2} &= \exp\left\{-2\sum_{i=1}^{n-1}\log(1-\alpha_i)\right\} \\
&\leq \exp\left\{\frac{2}{1-\alpha_1}\sum_{i=1}^{n-1}\alpha_i\right\} \\
&\leq \exp\left\{\frac{2a}{1-a/2}\sum_{i=1}^{n-1}(i+1)^{-1}\right\} \\
&\leq n^{\frac{4a}{2-a}}
\end{aligned}$$

where the last line follows from $\sum_{i=1}^n i^{-1} \leq \log n + 1$. Finally, for the sum in (2.26) to be finite, we just require

$$\frac{4a}{2-a} \leq 1$$

which is satisfied for $a < 2/5$ as in Assumption 2.1. □

We now have the first main result, which is similar to that in Hahn et al. (2018).

Theorem 2.12. *Let p_n satisfy the update (2.17) for $\mathbf{Y}_{1:n} \overset{\text{iid}}{\sim} f_0(\mathbf{y})$ where f_0 is continuous. Under the assumptions above, we have that the following holds almost surely:*

$$d_{\text{KL}}(f_0, p_n) \to K_\infty \quad \text{and} \quad \sum_{n=1}^\infty \alpha_n T(p_{n-1}) < \infty,$$

where K_∞ is a random variable.

Proof. This follows directly from Robbins and Siegmund (1971, Theorem 1) as we have shown the conditions required in our Lemmas 2.4, 2.5. □

It is not straightforward to show that $K_\infty = 0$ almost surely, even in the univariate case as claimed in Hahn et al. (2018), as their proof by contradiction requires that

$$K_\infty > 0 \implies \liminf_n T(p_{n-1}) > 0$$

which is nontrivial to verify. Nonetheless, we have opted to retain the result that K_∞ exists for completion, but we emphasize that this result is not necessary for the remaining proofs. We now deviate from the original proof and provide the details missing from Hahn et al. (2018).

We will rely on the second result of Theorem 2.12, that is the almost sure boundedness of $\sum_{n=1}^{\infty} \alpha_n T(p_{n-1})$, to prove Hellinger consistency. These are new additional steps that depend on the divergence-like properties of $T(p)$. To proceed, we require two lemmas.

Lemma 2.6. *From Theorem 2.12, we have that*

$$\liminf_n T(p_n) = 0$$

almost surely.

Proof. For contradiction, assume that $\liminf_n T(p_n) = \delta > 0$. Picking $\epsilon < \delta$, there exists N such that for all $n > N$, we have

$$T(p_n) > \delta - \epsilon.$$

However, this implies that

$$\sum_{n=1}^{\infty} \alpha_n T(p_{n-1}) \geq C + (\delta - \epsilon) \sum_{n=N+1}^{\infty} \alpha_n = \infty$$

which is a contradiction as $\sum_{n=1}^{\infty} \alpha_n T(p_{n-1}) < \infty$. □

Let us now consider the density on $\mathbf{u} = \{u^1, \ldots, u^d\} \in [0,1]^d$ defined

$$g_n(\mathbf{u}) = \prod_{j=1}^{d} \frac{f_0^j\{(P_n^j)^{-1}(u^j)\}}{p_n^j\{(P_n^j)^{-1}(u^j)\}} \tag{2.27}$$

where as a reminder

$$\begin{aligned} f_0^j(y^j) &= f_0(y^j \mid y^{1:j-1}) \\ p_n^j(y^j) &= p_n(y^j \mid y^{1:j-1}) \\ P_n^j(y^j) &= P_n(y^j \mid y^{1:j-1}) \\ \left(P_n^j\right)^{-1}(u^j) &= P_n^{-1}(u^j \mid y^{1:j-1}). \end{aligned}$$

We can write $T(p_n)$ as a function of g_n through a change of variables from y^j to $u^j = P_n^j(y^j)$, and similarly for $y^{j'}$ to v^j, for $j = 1, \ldots, d$, that is

$$T(g_n) = \int \int \prod_{j=1}^{d} c_\rho(u^j, v^j)\, g_n(\mathbf{u})\, g_n(\mathbf{v})\, d\mathbf{u}\, d\mathbf{v} - 1.$$

Lemma 2.6 gives us the following.

Lemma 2.7. *The sequence of densities g_n satisfies*

$$g_\infty(\mathbf{u}) := \liminf_n\, g_n(\mathbf{u}) = 1$$

for Lebesgue-almost all $\mathbf{u} \in [0,1]^d$ almost surely.

Proof. Repeated use of Fatou's lemma and the fact that $\liminf_n x_n^2 = (\liminf_n x_n)^2$ for $x_n \geq 0$ gives us

$$
\begin{aligned}
T(g_\infty) &= E_{\boldsymbol{\theta}} \left[\left(\int \prod_{j=1}^{d} \psi_{\theta^j}(u^j)\, g_\infty(\mathbf{u})\, d\mathbf{u} \right)^2 \right] - 1 \\
&\leq E_{\boldsymbol{\theta}} \left[\left(\liminf_n \int \prod_{j=1}^{d} \psi_{\theta^j}(u^j)\, g_n(\mathbf{u})\, d\mathbf{u} \right)^2 \right] - 1 \\
&= E_{\boldsymbol{\theta}} \left[\liminf_n \left(\int \prod_{j=1}^{d} \psi_{\theta^j}(u^j)\, g_n(\mathbf{u})\, d\mathbf{u} \right)^2 \right] - 1 \\
&\leq \liminf_n E_{\boldsymbol{\theta}} \left[\left(\int \prod_{j=1}^{d} \psi_{\theta^j}(u^j)\, g_n(\mathbf{u})\, d\mathbf{u} \right)^2 \right] - 1 \\
&= \liminf_n T(g_n) = 0.
\end{aligned}
$$

As $T(g_\infty)$ is non-negative, it is equal to 0. From the original proof then, $T(g_\infty) = 0$ implies $g_\infty(\mathbf{u}) = 1$ Lebesgue-almost everywhere from Lemma 2.4. □

We now require the squared Hellinger distance between probability density functions g_1 and g_2 on $\mathbf{u} \in [0,1]^d$, which is defined

$$
d^2_{\mathrm{H}}(g_1, g_2) := 1 - \int \sqrt{g_1(\mathbf{u})\, g_2(\mathbf{u})}\, d\mathbf{u}.
$$

A straightforward lemma follows from this.

Lemma 2.8. *The density p_n is Hellinger consistent at f_0 if and only if g_n in* (2.27) *is Hellinger consistent at the uniform density on* $[0,1]^d$.

Proof. Through a change of variables from u_j to $y_j = (P_n^j)^{-1}(u^j)$ for $j = 1, \ldots, d$, it is simple to show that

$$
\begin{aligned}
d^2_{\mathrm{H}}(g_n, 1) &= 1 - \int \sqrt{g_n(\mathbf{u})}\, d\mathbf{u} \\
&= 1 - \int \sqrt{f_0(\mathbf{y})\, p_n(\mathbf{y})}\, d\mathbf{y} \\
&= d^2_{\mathrm{H}}(p_n, f_0).
\end{aligned}
$$

□

We now have the final main result.

Theorem 2.13. *The density g_n converges in Hellinger distance to the uniform, that is*

$$\lim_n d_{\mathrm{H}}^2(g_n, 1) = 0$$

almost surely.

Proof. The limit superior of the Hellinger distance is

$$\limsup_n d_{\mathrm{H}}^2(g_n, 1) = 1 - \liminf_n \int \sqrt{g_n(\mathbf{u})}\, d\mathbf{u}.$$

From Fatou's lemma and the fact that $\liminf_n \sqrt{x_n} = \sqrt{\liminf_n x_n}$ for $x_n \geq 0$, we have

$$\liminf_n \int \sqrt{g_n(\mathbf{u})}\, d\mathbf{u} \geq \int \sqrt{\liminf_n g_n(\mathbf{u})}\, d\mathbf{u} = 1.$$

So we have

$$\limsup_n d_{\mathrm{H}}^2(g_n, 1) = 0$$

and $0 \leq \liminf_n d_{\mathrm{H}}^2(g_n, 1) \leq \limsup_n d_{\mathrm{H}}^2(g_n, 1)$, which gives us the result. □

From Lemma 2.8 and Theorem 2.13, we have Theorem 2.7 in the main paper as desired.

2.10.5 Copula derivations

2.10.5.1 The sequence of weights α_i

We now derive the sequence of weights α_i for the univariate copula update. The actual copula update for $n > 1$ for the posterior DP mixture is

$$p_{n+1}(y) = p_n(y)\, \frac{\int p(y \mid G)\, p(y_{n+1} \mid G)\, d\pi(G \mid y_{1:n})}{p_n(y)\, p_n(y_{n+1})}$$

where $\pi(G \mid y_{1:n})$ is a mixture of Dirichlet processes, that is

$$[G \mid \theta_{1:n}, y_{1:n}] \sim \text{DP}\left(a+n, \frac{aG_0 + \sum_{i=1}^{n} \delta_{\theta_i}}{a+n}\right)$$
$$[\theta_{1:n} \mid y_{1:n}] \sim \pi(\theta_{1:n} \mid y_{1:n}).$$

Usually, samples from the posterior over the means of the cluster assignments $\pi(\theta_{1:n} \mid y_{1:n})$ are obtained through Gibbs sampling. For tractability, we need to modify the term $\pi(\theta_{1:n} \mid y_{1:n})$. Let us instead assume that each cluster mean is drawn independently from the prior, so we can write

$$\pi(\theta_{1:n} \mid y_{1:n}) = \prod_{i=1}^{n} G_0(\theta_i).$$

Now computing the integral term, we have

$$\int p(y \mid G)\, p(y_{n+1} \mid G)\, d\pi(G \mid y_{1:n}) = E\left[\int p(y \mid G)\, p(y_{n+1} \mid G)\, d\pi(G \mid \theta_{1:n})\right] \quad (2.28)$$

where the expectation is over $\theta_{1:n} \sim \prod_{i=1}^{n} G_0(\theta_i)$. We can use the stick-breaking construction of the DP for the term inside the integral to get the familiar form (see Appendix 2.10.5.2). We have the inner term

$$\begin{aligned} &\int p(y \mid G)\, p(y_{n+1} \mid G)\, d\pi(G \mid \theta_{1:n}) \\ &= \left(1 - \frac{1}{a+n+1}\right) \int K(y \mid \theta)\, dG_n(\theta) \int K(y_{n+1} \mid \theta')\, dG_n(\theta') \\ &+ \frac{1}{a+n+1} \int K(y \mid \theta) K(y_{n+1} \mid \theta)\, dG_n(\theta) \end{aligned} \quad (2.29)$$

where we write $K(y \mid \theta) = \mathcal{N}(y \mid \theta, 1)$. Here, G_n is random and defined as

$$G_n = \frac{aG_0 + \sum_{i=1}^{n} \delta_{\theta_i}}{a+n}$$
$$\theta_{1:n} \sim \prod_{i=1}^{n} G_0(\theta_i).$$

Taking expectation of the first term in (2.29), we can write

$$
\begin{aligned}
&E\left[\int K(y \mid \theta)\, dG_n(\theta) \int K(y_{n+1} \mid \theta')\, dG_n(\theta')\right] \\
&= \frac{a^2}{(a+n)^2}\left\{\int K(y \mid \theta)\, dG_0(\theta) \int K(y_{n+1} \mid \theta')\, dG_0(\theta')\right\} \\
&+ \frac{a}{(a+n)^2}\left\{\int K(y \mid \theta)\, dG_0(\theta) \sum_{i=1}^{n} E\left[K(y_{n+1} \mid \theta_i)\right]\right\} \\
&+ \frac{a}{(a+n)^2}\left\{\int K(y_{n+1} \mid \theta)\, dG_0(\theta) \sum_{i=1}^{n} E\left[K(y \mid \theta_i)\right]\right\} \\
&+ \frac{1}{(a+n)^2} \sum_{i=1}^{n}\sum_{j=1}^{n} E\left[K(y \mid \theta_i)\, K(y_{n+1} \mid \theta_j)\right]
\end{aligned}
$$

We use the fact that $\theta_i \sim G_0$ and θ_i, θ_j are independent for $i \neq j$ to simplify the above to:

$$
\begin{aligned}
&\frac{a^2 + 2na + (n^2 - n)}{(a+n)^2}\left\{\int K(y \mid \theta)\, dG_0(\theta) \int K(y_{n+1} \mid \theta')\, dG_0(\theta')\right\} \\
&\qquad + \frac{n}{(a+n)^2} \int K(y \mid \theta) K(y_{n+1} \mid \theta)\, dG_0(\theta).
\end{aligned}
$$

Taking expectation of the second term in (2.29) is much simpler:

$$
E\left[\int K(y \mid \theta)\, K(y_{n+1} \mid \theta)\, dG_n(\theta)\right] = \int K(y \mid \theta)\, K(y_{n+1} \mid \theta)\, dG_0(\theta).
$$

Now plugging this back into (2.28), we have

$$
\begin{aligned}
&E\left[\int p(y \mid G)\, p(y_{n+1} \mid G)\, d\pi(G \mid \theta_{1:n})\right] \\
&= \left\{\frac{a+n}{a+n+1}\right\}\left\{\frac{a^2 + 2na + (n^2 - n)}{(a+n)^2}\right\}\left\{\int K(y \mid \theta)\, dG_0(\theta) \int K(y_{n+1} \mid \theta')\, dG_0(\theta')\right\} \\
&+ \underbrace{\left\{\frac{1}{a+n+1} + \left\{\frac{a+n}{a+n+1}\right\}\frac{n}{(a+n)^2}\right\}}_{\alpha_{n+1}} \int K(y \mid \theta)\, K(y_{n+1} \mid \theta)\, dG_0(\theta).
\end{aligned}
$$

This suggests the update

$$
p_{n+1}(y) = p_n(y)\left[1 - \alpha_{n+1} + \alpha_{n+1} c_\rho\left\{P_n(y), P_n(y_{n+1})\right\}\right]
$$

where for $a = 1$, we have

$$\alpha_n = \frac{1}{n+1} + \frac{n-1}{n(n+1)} = \left(2 - \frac{1}{n}\right) \frac{1}{n+1}.$$

The intuitive reasoning for this discrepancy from the usually suggested $(i+1)^{-1}$ is due to the mixing over the atoms of G_n. Note that for $n = 1$, we still have $\alpha_1 = 0.5$, so the first copula update step still agrees with the DP mixture.

The assumption of $\pi(\theta_{1:n} \mid y_{1:n}) = \prod_{i=1}^{n} G_0(\theta_i)$ is the only simplification required to get the copula update with the above α_i exactly, where $\theta_{1:n}$ are the means of the cluster allocations. For the DPMM, there are usually ties in the posterior samples of $\theta_{1:n}$, so by assuming all $\theta_i \overset{\text{iid}}{\sim} G_0$, we have allocated each y_i to its own cluster. In that sense, the copula update can be viewed as a mixture model where we allocate a new cluster for each data point, similar to the KDE.

2.10.5.2 Multivariate copula method

In this section, we derive the copula update for the multivariate DPMM, focussing on just the first step. One could also follow the argument of Appendix 2.10.5.1 to return the same update with the specific form for α_i. The multivariate DPMM with factorized kernel has the form

$$f_G(\mathbf{y}) = \int \prod_{j=1}^{d} \mathcal{N}(y^j \mid \theta^j, 1)\, dG(\boldsymbol{\theta}), \quad G \sim \text{DP}\left(a, G_0\right), \quad G_0(\boldsymbol{\theta}) = \prod_{j=1}^{d} \mathcal{N}(\theta^j \mid 0, \tau^{-1}).$$

Following the example in Hahn et al. (2018) and (2.13), we want to compute the copula density for the first update step of the DPMM, that is

$$\frac{E\left[f_G(\mathbf{y})\, f_G(\mathbf{y}_1)\right]}{p_0(\mathbf{y})\, p_0(\mathbf{y}_1)}. \tag{2.30}$$

From the stick-breaking representation of the DP, we can write G as

$$G = \sum_{k=1}^{\infty} w_k \, \delta_{\boldsymbol{\theta}_k^*}$$

where $w_k = v_k \prod_{j<k}\{1 - v_j\}$, $v_k \overset{\text{iid}}{\sim} \text{Beta}(1, a)$ and $\boldsymbol{\theta}_k^* \overset{\text{iid}}{\sim} G_0$. We can then write the numerator as

$$\begin{aligned}
&E\left[\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} w_j \, w_k \, K(\mathbf{y} \mid \boldsymbol{\theta}_j^*) \, K(\mathbf{y}_1 \mid \boldsymbol{\theta}_k^*)\right] \\
&= \left(1 - E\left[\sum_{k=1}^{\infty} w_k^2\right]\right) E\left[K(\mathbf{y} \mid \boldsymbol{\theta}^*)\right] E\left[K(\mathbf{y}_1 \mid \boldsymbol{\theta}^*)\right] \\
&+ E\left[\sum_{k=1}^{\infty} w_k^2\right] E\left[K(\mathbf{y} \mid \boldsymbol{\theta}^*) \, K(\mathbf{y}_1 \mid \boldsymbol{\theta}^*)\right]
\end{aligned}$$

where we have used the fact that $\sum_{k=1}^{\infty} w_k = 1$ almost surely. Here, $\theta^* \sim G_0$ and we have written

$$K(\mathbf{y} \mid \boldsymbol{\theta}^*) = \prod_{j=1}^{d} K(y^j \mid \theta^{*j}), \quad K(y^j \mid \theta^{*j}) = \mathcal{N}(y^j \mid \theta^{*j}, 1).$$

It is easy to show that

$$\alpha_1 := E\left[\sum_{k=1}^{\infty} w_k^2\right] = \frac{1}{1+a} \quad \text{a.s.}$$

As $p_0(\mathbf{y}) = E\left[K(\mathbf{y} \mid \boldsymbol{\theta}^*)\right]$, we have that (2.30) can be written as

$$1 - \alpha_1 + \alpha_1 \frac{E\left[K(\mathbf{y} \mid \boldsymbol{\theta}^*) \, K(\mathbf{y}_1 \mid \boldsymbol{\theta}^*)\right]}{p_0(\mathbf{y}) \, p_0(\mathbf{y}_1)}.$$

We note that the kernel K factorizes with independent priors on each dimension, and $p_0(\mathbf{y}) = \prod_{j=1}^d p_0(y^j) = \prod_{j=1}^d \mathcal{N}(y^j \mid 0, 1+\tau^{-1})$, so

$$\frac{E\left[K(\mathbf{y} \mid \boldsymbol{\theta}^*)\, K(\mathbf{y}_1 \mid \boldsymbol{\theta}^*)\right]}{p_0(\mathbf{y})\, p_0(\mathbf{y}_1)} = \prod_{j=1}^d \frac{E\left[K(y^j \mid \theta^{*j})\, K(y_1^j \mid \theta^{*j})\right]}{p_0(y^j)\, p_0(y_1^j)}.$$

Finally, we can compute each univariate term

$$\frac{E\left[K(y \mid \theta^*)\, K(y_1 \mid \theta^*)\right]}{p_0(y)\, p_0(y_1)} = \frac{\mathcal{N}_2(\Phi^{-1}(u), \Phi^{-1}(v) \mid 0, 1, \rho)}{\mathcal{N}(\Phi^{-1}(u) \mid 0, 1)\mathcal{N}(\Phi^{-1}(v) \mid 0, 1)}$$

where $\mathcal{N}_2(\cdot \mid 0, 1, \rho)$ is the bivariate normal density with mean 0, variance 1 and correlation $\rho = 1/(1+\tau)$, and $u = P_0(y), v = P_0(y_1)$. This is of course exactly the Gaussian copula density $c_\rho(u, v)$. Putting the above together gives us the copula update

$$p_1(\mathbf{y}) = \left[1 - \alpha_1 + \alpha_1 \prod_{j=1}^d c_\rho\left\{P_0(y^j), P_0(y_1^j)\right\}\right] p_0(\mathbf{y}).$$

2.10.5.3 Update for categorical data

In this section, we derive a copula type update for categorical data. For y on a countable space, we can derive a copula-type update with the DP prior, that is

$$f_G(y) = g(y), \quad G \sim \mathrm{DP}(a, G_0)$$

where g is the probability mass function of G, and likewise for g_0 and G_0. The predictive probability mass function is

$$p_n(y) = \frac{a g_0(y) + T_y^n}{a + n}.$$

where $T_y^n = \sum_{i=1}^n \mathbb{1}(y_i = y)$. Following a similar calculation of the Dirichlet-categorical model in Hahn et al. (2018), we have that

$$\begin{aligned} d_\rho(y, y_1) = \frac{p_1(y)}{p_0(y)} &= \frac{a}{a+1}\left(1 + \frac{\mathbb{1}(y = y_1)}{a g_0(y)}\right) \\ &= 1 - \rho + \rho \frac{\mathbb{1}(y = y_1)}{g_0(y)} \end{aligned}$$

where we have used $p_0(y) = g_0(y)$ and $\rho = 1/(a+1)$. We can then compute

$$\begin{aligned} D_\rho\{P_0(y), P_0(y_1)\} = P(Y \le y, Y_1 \le y_1) &= \sum_{z \le y, z' \le y_1} d_\rho(z, z')\, p_0(z)\, p_0(z') \\ &= (1-\rho) P_0(y) P_0(y_1) + \rho \left\{P_0(y) \wedge P_0(y_1)\right\}. \end{aligned}$$

which again is the mixture of the independent and Fréchet-Hoeffding copula.

For our updates, we can rewrite $d_\rho(y, y_1)$ as a function of P_0, p_0. Although $p_0(Y = k) = p_0(Y_1 = k)$ in this context, we need to keep the terms $p_0(y), p_0(y_1)$ separate in anticipation of the multivariate case, where this equality may not hold as we will be working with conditionals $p_n(y^j \mid y^{1:j-1})$.

Using the above, we have

$$\begin{aligned} d_\rho(y, y_1) = \{&[D_\rho\{P_0(y), P_0(y_1)\} - D_\rho\{P_0(y), P_0(y_1 - 1)\}] \\ &-[D_\rho\{P_0(y-1), P_0(y_1)\} - D_\rho\{P_0(y-1), P_0(y_1 - 1)\}]\} / \{p_0(y)\, p_0(y_1)\}. \end{aligned} \tag{2.31}$$

Here, $d_\rho(y, y_1)$ is the difference quotient of D_ρ, in a similar way $c_\rho(u, v)$ is the derivative of $C_\rho(u, v)$ for the continuous case. The update for the density and distribution function is then

$$\begin{aligned} p_1(y) &= d_\rho(y, y_1)\, p_0(y) \\ P_1(y) &= [D_\rho\{P_0(y), P_0(y_1)\} - D_\rho\{P_0(y), P_0(y_1 - 1)\}] / p_0(y_1). \end{aligned}$$

For the categorical case where $y \in \{1, \ldots, K\}$, we would have $P_0(y) = 0$ for $y < 1$ and $P_0(y) = 1$ for $y \ge K$.

For mixed data where some dimensions of $\mathbf{y}$ may be discrete, the conditional

factorization allows for an easy extension of the a multivariate mixed copula method. We simply substitute the bivariate Gaussian copula density c_ρ with d_ρ in (2.17) for the respective discrete dimensions, where we may have different bandwidths for the discrete and continuous data. However, in the discrete case, obtaining the martingale posterior may be more computationally difficult as we do not have the property $P_i(y_{i+1}) \overset{\text{iid}}{\sim} \mathcal{U}[0,1]$ as in the continuous case.

2.10.5.4 Martingale for regression

We now show that predictive resampling in the regression context gives us a martingale. For the update

$$p_{i+1}(y \mid \mathbf{x}) = \left\{1 - \alpha_{i+1}(\mathbf{x}, \mathbf{x}_{i+1}) + \alpha_{i+1}(\mathbf{x}, \mathbf{x}_{i+1})\, c_{\rho_y}\left(q_i, r_i\right)\right\} p_i(y \mid \mathbf{x})$$

where

$$q_i = P_i(y \mid \mathbf{x}), \quad r_i = P_i(Y_{i+1} \mid \mathbf{X}_{i+1}),$$

it is straightforward to show the martingale. Conditional on $\mathbf{X}_{i+1} = \mathbf{x}_{i+1}$, we have that $r_i \sim \mathcal{U}[0,1]$, so we can write

$$E\left[c_{\rho_y}\left(q_i, r_i\right) \mid y_{1:i}, \mathbf{x}_{1:i+1}\right] = \int_0^1 c_{\rho_y}\left(q_i, r\right)\, dr = 1.$$

So we have

$$E\left[p_{i+1}(y \mid \mathbf{x}) \mid y_{1:i}, \mathbf{x}_{1:i+1}\right] = p_i(y \mid \mathbf{x})$$

and from the tower rule

$$E\left[p_{i+1}(y \mid \mathbf{x}) \mid y_{1:i}, \mathbf{x}_{1:i}\right] = p_i(y \mid \mathbf{x})$$

almost surely for each $\mathbf{x} \in \mathbb{R}^d$. The martingale holds irrespective of the distribution of $\mathbf{X}_{i+1}$.

2.10.5.5 Conditional regression with dependent stick-breaking

We now derive the regression copula update inspired by the dependent DP. Consider the general covariate-dependent stick-breaking mixture model

$$f_{G_{\mathbf{x}}}(\mathbf{y}) = \int \mathcal{N}(y \mid \theta, 1)\, dG_{\mathbf{x}}(\theta), \quad G_{\mathbf{x}} = \sum_{k=1}^{\infty} w_k(\mathbf{x})\, \delta_{\theta_k^*}.$$

For the weights, we elicit the stick-breaking prior $w_k(\mathbf{x}) = v_k(\mathbf{x}) \prod_{j<k}\{1 - v_j(\mathbf{x})\}$ where $v_k(\mathbf{x})$ is a stochastic process on $\mathcal{X}$ taking values in $[0, 1]$, and is independent across k. For the atoms, we assume they are independently drawn from a normal distribution,

$$\theta_k^* \overset{\text{iid}}{\sim} G_0, \quad G_0 = \mathcal{N}(\theta \mid 0, \tau^{-1}).$$

Once again, we want to compute

$$\frac{E\left[f_{G_{\mathbf{x}}}(y)\, f_{G_{\mathbf{x}_1}}(y_1)\right]}{p_0(y \mid \mathbf{x})\, p_0(y_1 \mid \mathbf{x}_1)}.$$

Following the stick-breaking argument as in Appendix 2.10.5.2, we can write the numerator as

$$\{1 - \alpha_1(\mathbf{x}, \mathbf{x}')\}\, E\left[K(y \mid \theta^*)\right] E\left[K(y_1 \mid \theta^*)\right] + \alpha_1(\mathbf{x}, \mathbf{x}') E\left[K(y \mid \theta^*)\, K(y_1 \mid \theta^*)\right]$$

where we write

$$K(y \mid \theta^*) = \mathcal{N}(y \mid \theta^*, 1), \quad \theta^* \sim G_0,$$

and

$$\alpha_1(\mathbf{x}, \mathbf{x}') = \sum_{k=1}^{\infty} E\left[w_k(\mathbf{x}) w_k(\mathbf{x}')\right].$$

As before, we have

$$\frac{E\left[K(y \mid \theta^*)\, K(y_1 \mid \theta^*)\right]}{p_0(y \mid \mathbf{x})\, p_0(y_1 \mid \mathbf{x}_1)} = c_{\rho_y}\left\{P_0(y \mid \mathbf{x}), P_0(y_1 \mid \mathbf{x}_1)\right\}$$

where $\rho_y = 1/(1+\tau)$. We thus have the copula density as a mixture of the independent and Gaussian copula density. This then implies the first update step of the predictive takes the form

$$p_1(y \mid \mathbf{x}) = [1 - \alpha_1(\mathbf{x}, \mathbf{x}_1) + \alpha_1(\mathbf{x}, \mathbf{x}_1)\, c_\rho \{P_0(y \mid \mathbf{x}), P_0(y_1 \mid \mathbf{x}_1)\}]\; p_0(y \mid \mathbf{x}).$$

2.10.5.6 Beta-Bernoulli copula update for classification

In this section, we derive the copula update for the beta-Bernoulli model. The Bernoulli sampling density with beta prior is a special case of the DP update, where $y \in \{0, 1\}$. We will use the update (2.31) which simplifies drastically for the binary case.

For $y = y_1 = 0$, we have that $d_\rho(y, y_1) = D_\rho\{P_0(y), P_0(y_1)\}$, which directly gives us

$$d_\rho(y, y_1) = 1 - \rho + \rho\, \frac{p_0(y) \wedge p_0(y_1)}{p_0(y)\, p_0(y_1)} \quad \text{if } y = y_1 = 0.$$

For $y = 1, y_1 = 0$ then, we have that any terms in (2.31) with $y_1 - 1$ are 0, giving us

$$\begin{aligned} d_\rho(y, y_1)\, p_0(y)\, p_0(y_1) &= D_\rho\{P_0(y), P_0(y_1)\} - D_\rho\{P_0(y-1), P_0(y_1)\} \\ &= p_0(y_1) - (1-\rho)\, p_0(1-y)\, p_0(y_1) - \rho\{p_0(1-y) \wedge p_0(y_1)\} \\ &= (1-\rho)\, p_0(y)\, p_0(y_1) + \rho\, (p_0(y_1) - [\{1 - p_0(1-y)\} \wedge p_0(y_1)]) \\ &= (1-\rho)\, p_0(y)\, p_0(y_1) + \rho\, (p_0(y) - [p_0(y) \wedge \{1 - p_0(y_1)\}]) \end{aligned}$$

where we have used

$$p_0(y_1) - [\{1 - p_0(y)\} \wedge p_0(y_1)] = p_0(y) - [p_0(y) \wedge \{1 - p_0(y_1)\}].$$

This gives us

$$d_\rho(y, y_1) = 1 - \rho + \rho\, \frac{p_0(y) - [p_0(y) \wedge \{1 - p_0(y_1)\}]}{p_0(y)\, p_0(y_1)} \quad \text{if } y = 1, y_1 = 0.$$

Following the above derivations for the remaining two cases gives us

$$d_\rho(y, y_1) = \begin{cases} 1 - \rho + \rho \dfrac{p_0(y) \wedge p_0(y_1)}{p_0(y)\, p_0(y_1)} & \text{if } y = y_1 \\ 1 - \rho + \rho \dfrac{p_0(y) - [p_0(y) \wedge \{1 - p_0(y_1)\}]}{p_0(y)\, p_0(y_1)} & \text{if } y \neq y_1 \end{cases}$$

returning us the equation in the main paper if we plug in $q_i = p_i(y \mid \mathbf{x})$ and $r_i = p_i(y_{i+1} \mid \mathbf{x}_{i+1})$.

We now provide a quick check that the beta-Bernoulli update for classification indeed satisfies the martingale conditions. Let us write $q_i(1) = p_i(y = 1 \mid \mathbf{x})$, $q_i(0) = 1 - q_i(1)$ and likewise $r_i(1) = p_i(y_{i+1} = 1 \mid \mathbf{x}_{i+1})$, $r_i(0) = 1 - r_i(1)$. We first check that $q_{i+1}(1) + q_{i+1}(0) = 0$ given $q_i(1) + q_i(0) = 0$. For $y_{i+1} = 1$, we just need to check that the following term is equal to 1:

$$\frac{\{q_i(1) \wedge r_i(1)\} + q_i(0) - \{q_i(0) \wedge (1 - r_i(1))\}}{r_i(1)}.$$

If $q_i(0) < r_i(0)$, then $q_i(1) > r_i(1)$ so the numerator is $r_i(1) + q_i(0) - q_i(0) = r_i(1)$. Likewise if $q_i(0) > r_i(0)$, the numerator is $r_i(1) + q_i(0) - q_i(0) = r_i(1)$. The same applies for $y_{i+1} = 0$. Now to check the martingale condition. For $y = 1$, we want the below term to equal 1:

$$\begin{aligned} &\frac{\{q_i(1) \wedge r_i(1)\} + q_i(1) - \{q_i(1) \wedge (1 - r_i(0)\}}{q_i(1)} \\ &= \frac{q_i(1) + \{q_i(1) \wedge r_i(1)\} - \{q_i(1) \wedge r_i(1)\}}{q_i(1)} = 1. \end{aligned}$$

2.10.5.7 Probit copula update for classification

We can follow a similar derivation to the beta-Bernoulli copula update to derive an update for the conjugate probit-normal model. Here, the DPMM kernel is the normal CDF, $P(y = 1 \mid \theta) = \Phi(\theta)$, and the base measure is $G_0 = \mathcal{N}(\theta \mid 0, \tau^{-1})$. The update

term instead takes the form

$$b_{\rho_y}(q_i, r_i) = \begin{cases} \dfrac{C_{\rho_y}(q_i, r_i)}{q_i\, r_i} & \text{if } y = y_{i+1} \\ \dfrac{\{q_i - C_{\rho_y}(q_i, 1 - r_i)\}}{q_i\, r_i} & \text{if } y \neq y_{i+1} \end{cases}$$

where $q_i = p_i(y \mid \mathbf{x}), r_i = p_i(y_{i+1} \mid \mathbf{x}_{i+1})$, and C is the bivariate Gaussian copula distribution function. Again, we have $\rho_y = 1/(1+\tau)$. Note the similarities to the beta-Bernoulli update, where instead of $C_{\rho_y}(u, v)$ we have $D_{\rho_y}(u, v)$. However, this is computationally much more expensive as we need to approximate a bivariate normal integral.

2.10.6 Practical considerations for copula methods

2.10.6.1 Implementation details

For our copula methods, we begin by first computing v_i^j for $i = \{0, \ldots, n-1\}, j = \{1, \ldots, d\}$ where $v_i^j = P_i(y_{i+1}^j \mid y_{i+1}^{1:j-1})$. In practice, this involves iterating the copula methods and computing the conditional CDF values through (2.19) for the observed datapoints $\mathbf{y}_{1:n}$. For the just-in-time compilation, it is generally faster to work with arrays of fixed size, and so it is better to compute $P_{1:n}$ for all $\mathbf{y}_{1:n}$ and extract the needed v_i^j. Given v_i^j, we can compute p_n at any test point $\mathbf{y}$, and predictively resample.

2.10.6.2 Computing quantiles and sampling

For the univariate cases, we may be interested in computing quantiles for $P_n(y)$ or $P_n(y \mid \mathbf{x})$. For example, finding the median $\bar{y}$ such that $P_n(\bar{y} \mid \mathbf{x}) = 0.5$ may be of interest if we are interested in a point estimate of the regression function. Although we cannot compute $\bar{y}$ exactly, we can solve $P_n(y \mid \mathbf{x}) = 0.5$ for y through efficient gradient-based optimization using automatic differentiation. We may also be interested in sampling from $P_n(\mathbf{y})$ or $P_N(\mathbf{y})$, e.g. for computing Monte Carlo estimates of parameters. We can again utilize an optimization-based inverse transform sampling method: we simulate

d independent uniform random variables $U^{1:d}$, and then find the quantiles by solving $u_n^j = U^j$ for $j = \{1, \ldots, d\}$ to obtain a random sample $\mathbf{Y}$. In practice, we can minimize the loss function, $\mathbf{Y} = \arg\min_{\mathbf{Y}} \sum_{j=1}^{d} \{U^j - P(Y^j \mid Y^{1:j-1})\}^2$, which would give 0 loss at the sample.

2.10.6.3 Optimization details

As gradients with respect to ρ are generally difficult to compute, we opt for automatic differentiation, which is available in JAX. For hyperparameter optimization, we use the SLSQP method in `scipy` (Virtanen et al., 2020). However, it would be of interest to investigate the applications of stochastic gradient methods. For computing quantiles/sampling, we implement BFGS with Armijo backtracking line search directly in JAX so that the entire optimization procedure can be JIT-compiled.

2.10.7 Additional experiments

2.10.7.1 Reproducibility of copula experiments

Due to the non-determinism of GPU computations even with random seeds, the reported results for the copula methods in the main paper are timed on the GPU but exact numerical values and plots are computed on the CPU (with the appropriate `jaxlib` version). The results are unsurprisingly very similar on the two platforms, but the CPU allows for exact reproducibility with the provided code at `https://github.com/edfong/MP`. The baselines methods do not have this issue as they are timed and run on the CPU.

2.10.7.2 Baseline details

For the DPMM with MCMC examples, we use the full covariance multivariate normal kernel with conjugate priors from the `dirichletprocess` package (Ross and Markwick, 2018), which implements Gibbs sampling from Neal (2000). The hyperprior on the concentration parameter is $a \sim \text{Gamma}(2, 4)$. For the univariate case, the

conjugate base measure is

$$G_0(\mu, \sigma^2) = \mathcal{N}\left(\mu \mid \mu_0, \frac{\sigma^2}{k_0}\right) \text{Inverse Gamma}(\sigma^2 \mid \alpha, \beta).$$

For the GMM examples, we set $\{\mu_0, k_0, \alpha, \beta\} = \{0, 1, 0.1, 0.1\}$; we select α, β to match the smoothness of the posterior mean with the copula method. For the galaxy example (with unnormalized data), we follow the suggestions of West (1991) and set $\{\mu_0, k_0, \alpha, \beta\} = \{20, 0.07, 2, 1\}$, where again k_0 is set to a point estimate that matches the smoothness to the copula method.

For the multivariate DPMM, the base measure of the DP is the normal-Wishart distribution:

$$G_0(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathcal{N}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, (k_0\boldsymbol{\Lambda})^{-1}) \, \text{Wi}(\boldsymbol{\Lambda} \mid \boldsymbol{\Lambda_0}, \nu)$$

where $\boldsymbol{\Lambda}$ is the precision matrix of the kernel. For both the air quality and LIDAR example, we have $\{\boldsymbol{\mu}_0, k_0, \boldsymbol{\Lambda}_0, \nu\} = \{\mathbf{0}, d, I_d, d\}$ where I_d is the identity matrix and $d = 2$m which is default in the package. For regression, we fit the DPMM to estimate the joint density $p(y, x)$ and compute the implied conditional $p(y \mid x)$. For MCMC, we sample $B = 2000$ posterior samples for the plots with a burn-in chain of length 2000. For the timing, we have the same burn-in chain but only take $B = 1000$ to compare to the copula. Given a posterior sample of the cluster parameters $\theta_{1:n}$, we can draw an approximate posterior sample of $[G_\infty \mid \theta_{1:n}]$, which is conditionally independent from $y_{1:n}$, from the stick-breaking representation (see Key Property 5 of Ross and Markwick (2018)), from which we can compute a random sample of p_∞ by mixing over the kernel with G_∞.

The remainder methods are implemented in `sklearn`. For the DPMM with VI (mean-field approximation), we use the diagonal covariance kernel, with default hyperparameters for the priors. For the variational approximation, we set the upper limit of clusters to $K = 30$, and initialize and optimize 100 times to avoid local minima. For the KDE, the scalar bandwidth is set through 10-fold cross-validation with the

log predictive density over a grid. For Gaussian process regression, we use the RBF kernel with a single length scale, which is set by maximizing the marginal likelihood over 10 repeats. The same is done for Gaussian process classification with the logistic link function, which approximates the posterior with the Laplace approximation. For the linear models, we use Bayesian ridge regression and logistic regression with L_2 regularization with default values.

2.10.7.3 Galaxy

For the galaxy example, convergence to the martingale posterior is assessed in Figure 2.14, where 5000 forward samples is sufficient.

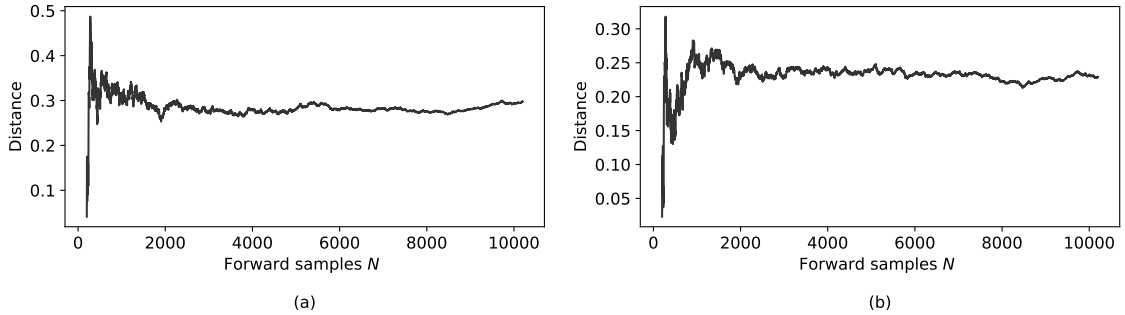


Figure 2.14: Estimated L_1 distance (a) $\|p_N - p_n\|_1$ and (b) $\|P_N - P_n\|_1$ for a single forward sample

2.10.7.4 Bivariate air quality

In Figure 2.15 we show the posterior mean and standard deviation of the density obtained for the DPMM with MCMC, which is comparable to that of the martingale posterior from the copula method.

2.10.7.5 Classification in simulated moon dataset

We demonstrate the conditional method of Section 2.6.4.2 with the beta-Bernoulli copula update for classification on a non-linear classification example, which is particularly interesting as we can predictively resample $Y_{n+1:\infty}$ directly. We analyze the simulated

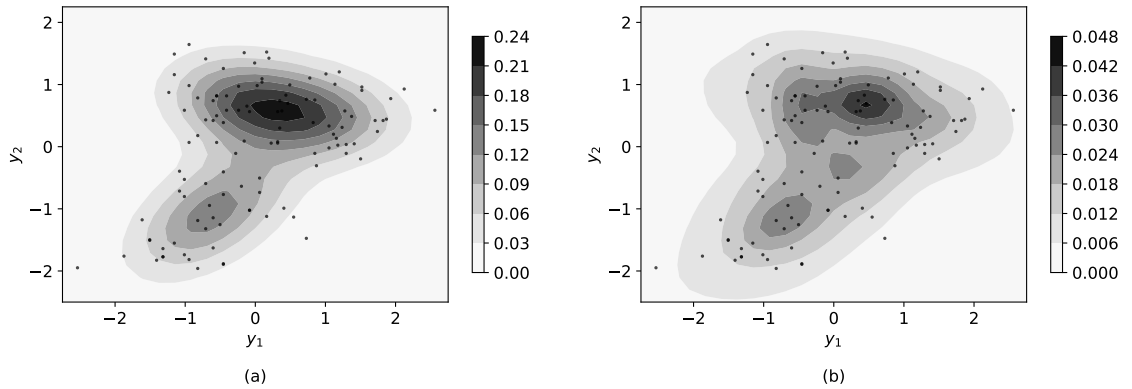


Figure 2.15: Posterior (a) mean and (b) standard deviation of density for DPMM, with data (•)

moon dataset from `sklearn` (Pedregosa et al., 2011) with bivariate covariates. Figure 2.16 shows the decision boundary on the left, and $n = 100$ simulated data points on the right with a fit GP using the logistic link and Laplace approximation. We fit the conditional copula method and get the bandwidths $\rho_y = 0.73, \rho_x = [0.92, 0.74]$; optimizing, fitting and prediction on the $\mathbf{x}$-grid of size 25×25 required 2 seconds, versus around 1 second for the GP.

Predictive resampling is also possible here, where we draw $X_{n+1:N}$ with the Bayesian bootstrap and simulate $Y_{n+1:N}$ directly. Generating $B = 1000$ samples of $Y_{n+1:N}, X_{n+1:N}$ and computing each $p_N(y = 1 \mid \mathbf{x})$ on the 25×25 grid took under 3 seconds. The martingale posterior mean and standard deviations of $p_N(y = 1 \mid \mathbf{x})$ are shown in Figure 2.17. The martingale posterior mean $p_n(y = 1 \mid \mathbf{x})$ is similar to the GP, and we see that the uncertainty is higher in $\mathbf{x}$-regions where both classes are present. However, we must note that the uncertainty decreases to 0 as we move away from the data which is undesirable. This may be due to the Bayesian bootstrap resampling $\mathbf{x}_{n+1:\infty}$ from the observed $\mathbf{x}_{1:n}$, so $\mathbf{x}_N$ in the tails are never sampled. As a result, $\alpha(\mathbf{x}, \mathbf{x}_N)$ remains small for $\mathbf{x}$ in the tails, so there is not much change in $p_N(y \mid \mathbf{x})$. This behaviour is also observed in the conditional copula regression case in Appendix 2.10.7.6. As such, we recommend only looking at inlying $\mathbf{x}$ when using the Bayesian bootstrap for $\mathbf{x}_{n+1:\infty}$ with the conditional copula method. On the left of Figure 2.18, we also plot the martingale

posterior of $p_N(y = 1 \mid \mathbf{x} = [1, -0.8])$, which has mean 0.86 but has relatively large uncertainty. On the right, we plot the absolute difference $|p_N(y = 1 \mid \mathbf{x}) - p_n(y = 1 \mid \mathbf{x})|$ as we carry out one forward sample, which converges relatively quickly with N.

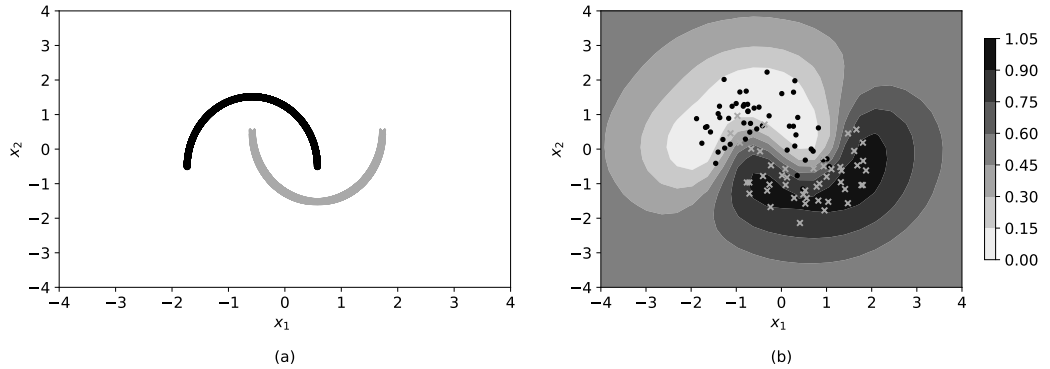


Figure 2.16: (a) Simulated data with no noise with $y = 0$ (•) and $y = 1$ (×); (b) $n = 100$ simulated data points with Gaussian noise (added to covariates) and $p_n(y = 1 \mid \mathbf{x})$ for GP

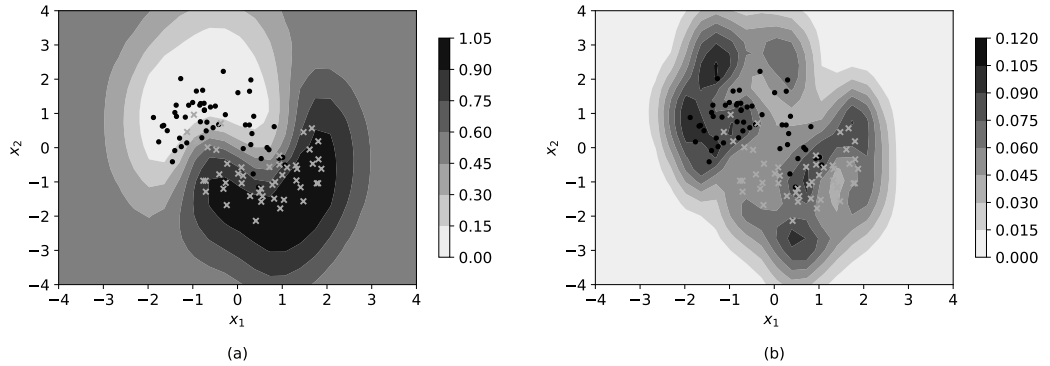


Figure 2.17: Posterior (a) mean and (b) standard deviation of $p_N(y = 1 \mid \mathbf{x})$ for conditional copula method with $y = 0$ (•) and $y = 1$ (×)

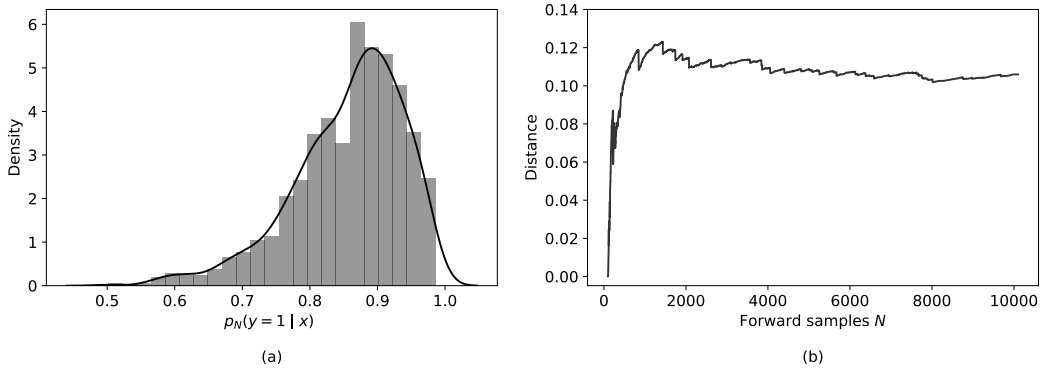


Figure 2.18: (a) Posterior samples of $p_N(y = 1 \mid \mathbf{x})$; (b) Convergence plot $|p_N(y = 1 \mid \mathbf{x}) - p_n(y = 1 \mid \mathbf{x})|$ for $\mathbf{x} = [1, -0.8]$

2.10.7.6 LIDAR

In Figure 2.19, we see the predictive means and 95% central intervals for the conditional copula method on the left and the GP on the right. The conditional copula method still performs well, with a slight bias towards the end of range of x, in a similar way to the Nadaraya-Watson estimator. The GP deals well with the nonlinearity but much more poorly with the heteroscedasticity.

In Figures 2.20, 2.21 we see the difference in the martingale posteriors of $p_N(y \mid x = 0)$ for the joint and conditional copula method. For $x = 0$ which lies within the data, the uncertainty for the joint and conditional methods are similar. However, for $x = -3$ which is far from the data, we see that the joint method has a high amount of uncertainty as expected, but the conditional method does not have any uncertainty. We believe this occurs due to the discrete nature of the Bayesian bootstrap when resampling $x_{n+1:\infty}$ for the conditional method as discussed in Appendix 2.10.7.5. This issue can be mitigated by resampling using a smooth density (e.g. a KDE) for $x_{n+1:\infty}$, but still occurs for suitably distant x. It seems that the choice of resampling distribution of $x_{n+1:\infty}$ affects the uncertainty of outlying x, but we have found much less sensitivity to this choice for inlying x. Interestingly, the joint method does not have this issue irrespective of the distance of x from the dataset, so we recommend the joint method when the x of interest is outlying. We leave a detailed investigation of this for future work.

We can also see the difference in $p_n(y \mid x)$ between the joint and conditional method by comparing Figures 2.19 and 2.10. Although the joint method is not skewed towards the end of the range of x, the prequential log-likelihood is actually higher for the conditional method. The conditional method is only slightly faster in this example, but for higher dimensional covariates, the computational gain of not having to estimate $P_n(\mathbf{x})$ is much higher. For the median plot in Figure 2.13(a) in the main paper, we compute the y, x grid of size 40×40 using the 95% credible bands as the limits for y. The median is then computed numerically from the grid where $P_N(y \mid x) = 0.5$.

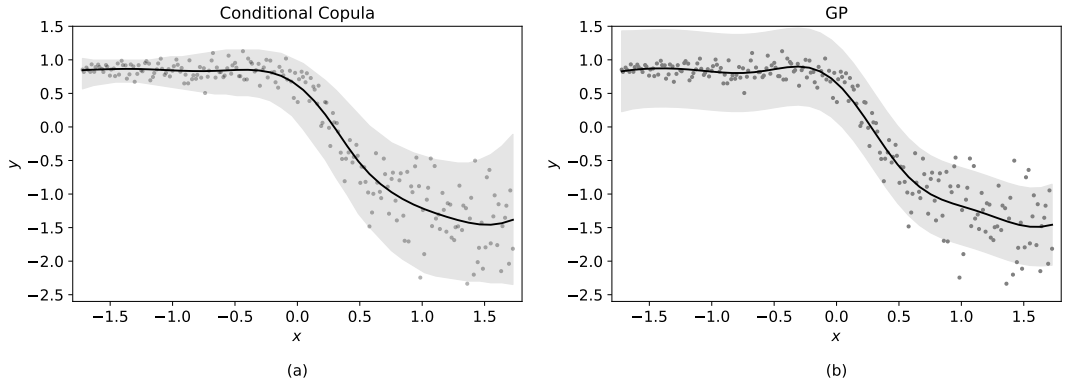


Figure 2.19: $p_n(y \mid x)$ (—) with 95% predictive interval (■) for the (a) conditional copula method and (b) GP, with data (•)

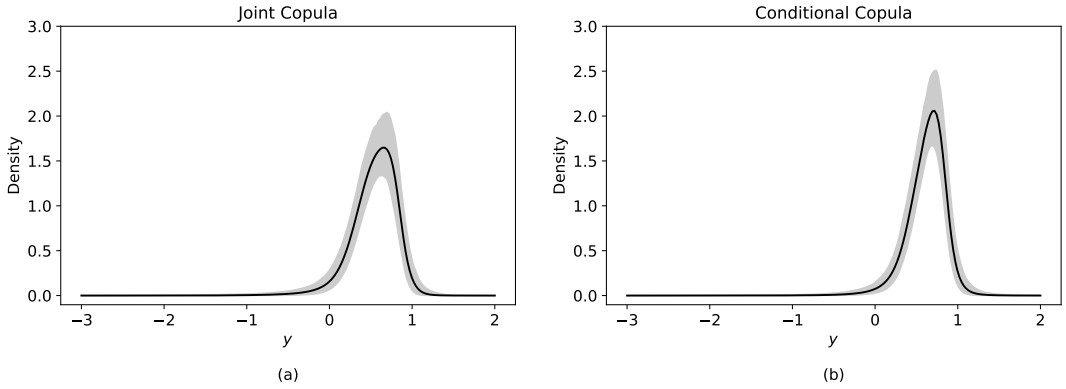


Figure 2.20: Posterior mean (—) and 95% credible interval (■) of $p_N(y \mid x = 0)$ for the (a) joint copula method and (b) conditional copula method

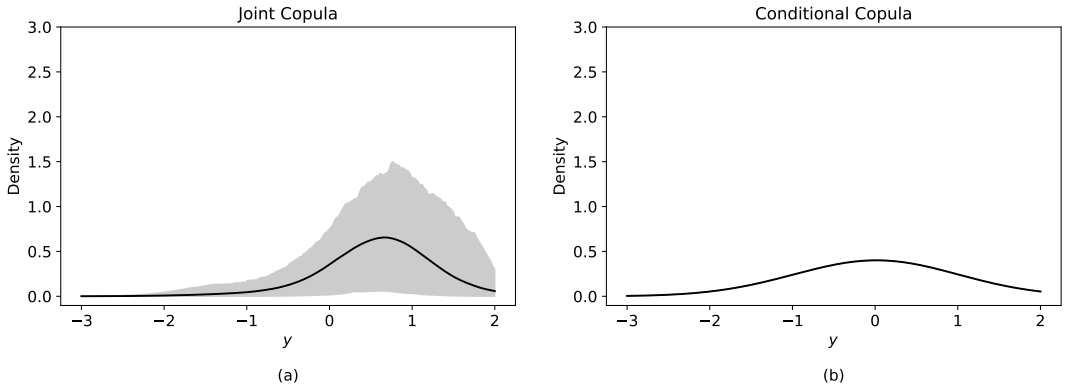


Figure 2.21: Posterior mean (—) and 95% credible interval (■) of $p_N(y \mid x = -3)$ for the (a) joint copula method and (b) conditional copula method

2.10.7.7 High-dimensional GMM

In this section, we consider a simulated example to demonstrate the degradation in performance with dimensionality for the DPMM with VI (with diagonal covariance matrix), which the copula method is robust to. We simulate $n = 100, n_{\text{test}} = 1000$ data points from

$$f_0(\mathbf{y}) = 0.5\mathcal{N}(\mathbf{y} \mid \mu_d^1, I_d) + 0.5\mathcal{N}(\mathbf{y} \mid \mu_d^2, I_d)$$

where $\mu_d^1 = \Big[-1, \ldots, -1\Big]^{\mathrm{T}}$ and $\mu_d^1 = \Big[2, \ldots, 2\Big]^{\mathrm{T}}$ are both d-vectors, and I_d is the $d \times d$ identity matrix. The DPMM is thus well-specified in this example. As before, we normalize the data before fitting the methods.

Below is a plot of the average test log-likelihoods as d increases for each method. The DPMM with VI degrades steeply with dimensionality, likely due to the difficulty of the variational optimization. The KDE performs quite well in this simple example.

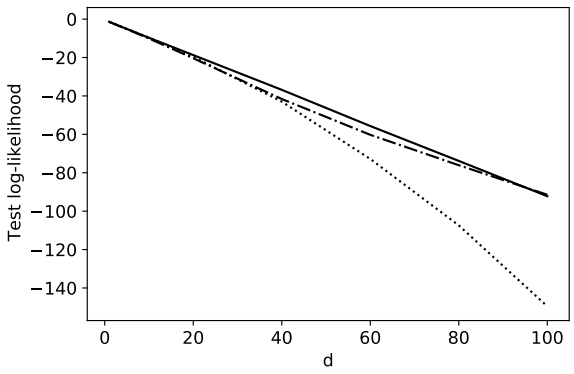


Figure 2.22: Average test log-likelihood with dimensionality for the copula method (—), DPMM with VI (·····) and KDE(-·-)

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

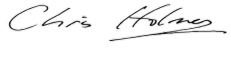
Title of Paper	Martingale posterior distributions
Publication Status	□ Published □ Accepted for Publication **x** Submitted for Publication □ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Fong, E., Holmes, C., and Walker, S. G. (2021). Martingale posterior distributions. *arXiv preprint arXiv:2103.15671.*

Student Confirmation

Student Name:	Chung Hang Edwin Fong	
Contribution to the Paper	○ Lead of project, with co-authors as advisors. ○ Formulated the main idea of the paper. ○ Carried out the literature review. ○ Developed methodology and theoretical results. ○ Implemented the algorithms/experiments in code and interpreted results. ○ Undertook manuscript writing (edited by co-authors).	
Signature 	Date	2nd December, 2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Chris Holmes		
Supervisor comments ○ The manuscript was largely developed by the candidate with guidance from the co-authors. ○ I verify that the above summary of the candidate's contribution is accurate.		
Signature 	Date	3rd December 2021

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 3

Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap

3.1 Preamble

Although the work in this chapter was completed earlier than Chapter 2, it is a special case of the martingale posterior. Specifically, the Dirichlet process prior amounts to eliciting the predictive distribution as the weighted sum of a prior base distribution and the empirical distribution, that is (1.2) in the introduction. We primarily investigate the computational benefits of such a scheme in this chapter.

The content of this chapter is a self-contained manuscript with its supplementary material. Details of the publication are given below:

> Fong, E., Lyddon, S., and Holmes, C. (2019). Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 1952-1962. PMLR.

A statement of authorship is provided at the end of this chapter.

3.2 Abstract

Increasingly complex datasets pose a number of challenges for Bayesian inference. Conventional posterior sampling based on Markov chain Monte Carlo can be too computationally intensive, is serial in nature and mixes poorly between posterior modes. Furthermore, all models are misspecified, which brings into question the validity of the conventional Bayesian update. We present a scalable Bayesian nonparametric learning routine that enables posterior sampling through the optimization of suitably randomized objective functions. A Dirichlet process prior on the unknown data distribution accounts for model misspecification, and admits an embarrassingly parallel posterior bootstrap algorithm that generates independent and exact samples from the nonparametric posterior distribution. Our method is particularly adept at sampling from multimodal posterior distributions via a random restart mechanism, and we demonstrate this on Gaussian mixture model and sparse logistic regression examples.

3.3 Introduction

As datasets grow in complexity and size, Bayesian inference becomes increasingly difficult. The posterior is often intractable, so we resort to simulation methods for inference via Markov chain Monte Carlo (MCMC), which is inherently serial and often too computationally expensive in datasets with a large number of data points (Bardenet et al., 2017). MCMC further struggles with multimodal posteriors which arise in many settings including mixture models (Jasra et al., 2005) or non-convex priors (Seeger et al., 2007), as the MCMC sampler can become trapped in local modes (Rudoy and Wolfe, 2006). Current methods to sample from multimodal posteriors with MCMC include parallel tempering (Neal, 1996) and adaptive MCMC (Pompe et al., 2018), but the associated computational cost is high. Posterior approximation with variational Bayes (VB) (Blei et al., 2017) is a faster alternative, but it is generally difficult to quantify the quality of the approximation, and is thus problematic if accurate uncertainty

quantification is desired (Giordano et al., 2015).

A further methodological issue facing Bayesian inference is the fact that all models are false. The increasing scale of datasets exacerbates the effects of model misspecification (Walker, 2013), as the true sampling distribution is meaningfully different from the parametric family of distributions of the model. There is rarely formal acknowledgement of model misspecification which can lead to inconsistencies (Watson and Holmes, 2016; Grünwald and van Ommen, 2017).

Bayesian nonparametric learning (NPL) introduced by Lyddon et al. (2018) allows for the use of statistical models without assuming the model is true. NPL uses a nonparametric prior centred on a parametric model, and returns a nonparametric posterior over the parameter of interest. The method focuses on accounting for model misspecification and for posterior approximation such as from Variational Bayes (VB) by placing a mixture of Dirichlet processes (Antoniak, 1974) prior on the sampling distribution. In addition to the acknowledgement of model misspecification, the method admits an embarrassingly parallel Monte Carlo sampling scheme consisting of randomized maximizations. However, in most cases this method requires sampling the Bayesian posterior, which is computationally expensive for complex models.

3.3.1 Contribution

In this work, we propose a simplified variant of NPL that utilises a Dirichlet process (DP) prior on F_0 instead of a mixture of Dirichlet processes (MDP) prior. This allows us to perform inference directly and detaches the nonparametric prior from the prior of the model parameter of interest. Instead of centering on a Bayesian posterior, we center the DP on a sampling distribution which encapsulates our prior beliefs. This simpler choice of prior also has desirable theoretical properties and is highly scalable as we no longer need to sample from the Bayesian posterior. Our method can handle a variety of statistical models through the choice of the loss functions, and can be applied to a wide range of machine learning settings as we will demonstrate in Section 3.5.

Our method implies a natural noninformative prior, which may be relevant when the number of data points is substantially larger than the number of parameters.

The posterior bootstrap sampling scheme was introduced by Lyddon et al. (2018) under the NPL framework, and we inherit its computational strengths such as parallelism and exact inference under a Bayesian nonparametric model. Independent samples from the nonparametric posterior are obtained through the optimization of randomized objective functions, and we obtain the weighted likelihood bootstrap (Newton and Raftery, 1994) as a special case. Furthermore, sampling from multimodal posteriors now involves a non-convex optimization at each bootstrap sample that we solve through local search and random restart. We demonstrate that our method recovers posterior multimodality on a Gaussian Mixture Model (GMM) problem. We further show that our method is computationally much faster than conventional Bayesian inference with MCMC, and has superior predictive performance on real sparse classification problems. Finally, we utilize the computational speed of NPL to carry out a Bayesian sparsity-path-analysis for variable selection on a genetic dataset.

3.4 Bayesian nonparametric learning

Assume that we have observed $Y_{1:n} = y_{1:n}$, where $Y_{1:n}$ is a sequence of n i.i.d. observables and F_0 is the unknown sampling distribution from which the observations arose. We may be interested in a parameter $\theta \in \Omega \subseteq \mathbb{R}^p$, which indexes a family of probability densities $\mathcal{F}_\Omega = \{f_\theta(y); \theta \in \Omega\}$. Conventional Bayesian updating of the prior to the posterior via Bayes' theorem formally assumes that F_0 belongs to the model $\mathcal{F}_\Omega$, which is questionable in the presence of complex and large datasets. This assumption is not necessary for NPL. We derive the foundations of NPL by treating parameters as functionals of F_0, with model fitting as a special case.

3.4.1 The parameter of interest

We define our parameter of interest as

$$\theta_0 = \theta(F_0) = \arg\min_{\theta} \int \ell(\theta, y)\, dF_0(y) \tag{3.1}$$

where $\ell(\theta, y)$ is a loss function, and its form can be used to target statistics of interest. For example, setting $\ell(\theta, y) = |y - \theta|$ returns the median and $(y - \theta)^2$ returns the mean.

The loss function of particular interest is $\ell(\theta, y) = -\log f_\theta(y)$, where f_θ is the density of some parametric model. The value of θ_0 minimises the Kullback-Leibler divergence $d_{\mathrm{KL}}(f_0, f_\theta)$, which is the parameter of interest in conventional Bayesian analysis (Walker, 2013; Bissiri et al., 2016). We have not assumed that $\mathcal{F}_\Omega$ contains F_0, and θ_0 in this case does not have any particular generative meaning as it is simply the parameter that satisfies (3.1).

3.4.2 The Dirichlet process prior

As the sampling distribution is unknown, we place a DP prior on F_0

$$[F \mid \alpha, F_\pi] \sim \mathrm{DP}\,(\alpha, F_\pi)$$

where F_π is our prior centering measure, and α is the strength of our belief.

The base measure F_π We encode our prior knowledge about the sampling distribution in the measure F_π. If we believe a particular model f_θ to be accurate, and have prior beliefs about θ encoded in the prior density $\pi(\theta)$, a sensible choice for the density of F_π is $f_\pi(y) = \int f_\theta(y)\, \pi(\theta)\, d\theta$. Alternatively, we could directly specify f_π as a density that accurately represents our beliefs without the burden of defining a joint distribution on (y, θ). In the presence of historical data $\hat{y}_{1:\hat{n}}$, a suitable choice for F_π is the empirical distribution of the historical data, i.e. $F_\pi = \frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \delta_{\hat{y}_i}$ where δ is

the Dirac measure. This is in a similar fashion to power priors (Ibrahim et al., 2000). Further intuition is provided in Appendix 3.7.1.1.

It should be noted that we cannot directly include a prior on the parameter of interest θ_0, only implicitly through (α, F_π). Our prior is selected independently of the model of interest, and this is appropriate under a misspecified model setting since we do not believe there to be a true f_θ. As all parameters of interest are defined as a functional of F_0 as in (3.1), any informative prior on F_0 is thus informative of θ_0.

The concentration α The size of α measures the concentration of the DP about F_π, and a large value corresponds to a smaller variance in a functional of the DP. We see in (3.2) that the DP posterior base measure is a weighted sum of the prior F_π and the empirical distribution $F_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$, with the weights proportional to α and n respectively. We can thus interpret α as the effective sample size from the prior F_π. One method of selecting α is through simulation of the prior distribution of θ via (3.1) and tuning its variance. Alternatively, we can select α through the a priori variance of the mean functional (see Appendix 3.7.1.2). The special case of $\alpha = 0$ corresponds to the Bayesian bootstrap (Rubin, 1981), which in our case corresponds to a natural way to define an noninformative prior about F_0 (see Gelman et al. (2013) for a review on noninformative priors). For $n \gg p$, it may be suitable to set $\alpha = 0$ as the prior should have little influence and the Bayesian bootstrap is more computationally efficient.

3.4.3 The NPL posterior

From the conjugacy of the DP, the posterior of F is

$$\begin{aligned} [F \mid y_{1:n}] &\sim \mathrm{DP}\left(\alpha + n, G_n\right), \\ G_n &= \frac{\alpha}{\alpha + n} F_\pi + \frac{1}{\alpha + n} \sum_{i=1}^n \delta_{y_i}. \end{aligned} \tag{3.2}$$

Our NPL posterior $\widetilde{\Pi}(\cdot \mid y_{1:n})$ is thus

$$\widetilde{\Pi}(\theta \in A \mid y_{1:n}) = \int \Pi(\theta \in A \mid F)\, d\Pi(F \mid y_{1:n})$$

for a measurable set $A \subseteq \Omega$, where $\Pi(\theta \in A \mid F) = \mathbb{1}\{\theta(F) \in A\}$; the indicator arises as θ is a deterministic functional of F as in (3.1). Properties of the NPL posterior follow from properties of the DP, e.g. draws of $F \mid y_{1:n}$ are almost surely discrete, so (3.1) simplifies to

$$\theta(F) = \arg\min_{\theta} \sum_{k=1}^{\infty} w_k\, \ell(\theta, \tilde{y}_k) \tag{3.3}$$

where $w_{1:\infty} \sim \text{GEM}(\alpha + n)$ and $\tilde{y}_{1:\infty} \overset{\text{iid}}{\sim} G_n$ from the stick-breaking construction (Sethuraman, 1994). Formally, the GEM distribution is defined

$$v_k \sim \text{Beta}(1, \alpha + n), \quad w_k = v_k \prod_{j=1}^{k-1}(1 - v_j).$$

We preserve the theoretical advantages from Lyddon et al. (2018) due to the equivalence in the limits of the DP and the MDP for $\alpha \to 0$ and $n \to \infty$, where α also denotes the concentration parameter of the MDP.

Consistency Under mild regularity conditions, the NPL posterior is consistent at θ_0 as defined in (3.1), from the properties of the DP (see van der Vaart (1998); Ghosal (2010); Ghosal and van der Vaart (2017) for details). Interestingly, this is true regardless of the choice of F_π and its support. This is not the case in conventional Bayesian inference through Bayes' rule where the support of the prior must contain θ_0 for posterior consistency. This is particularly reassuring in our misspecified model setting, as inferences about θ_0 are robust to choices of F_π.

Asymptotic dominance The NPL posterior predictive density, $\tilde{p}(y \mid y_{1:n}) = \int f_\theta(y)\, d\widetilde{\Pi}(\theta \mid y_{1:n})$, for $\alpha = 0$ *asymptotically dominates* the conventional Bayesian

posterior predictive density, $p(y \mid y_{1:n}) = \int f_\theta(y)\, \pi(\theta \mid y_{1:n})\, d\theta$, up to $o(n^{-1})$ under mild regularity conditions, i.e.

$$\mathbb{E}_{y_{1:n} \sim q} \left[d_{\mathrm{KL}}\{q(\cdot), p(\cdot \mid y_{1:n})\} - d_{\mathrm{KL}}\{q(\cdot), \tilde{p}(\cdot \mid y_{1:n})\} \right] = K\{q(\cdot)\} + o(n^{-1})$$

for all densities q, where K is a non-negative and possibly positive real-valued functional. This states that compared to the Bayesian posterior predictive, the NPL posterior predictive is closer in expected KL divergence to the true F_0 up to $o(n^{-1})$. The proof for the MDP case is given in Theorem 1 of Lyddon et al. (2018), and the above follows from the equivalence of the MDP and the DP for $\alpha = 0$.

3.4.4 Sampling from the NPL posterior

In almost all cases, $\widetilde{\Pi}(\cdot \mid y_{1:n})$ is not tractable, but lends itself to a parallelizable Monte Carlo sampling scheme. It may be more intuitive to think of sampling F from the posterior DP, then calculating (3.1) to generate the sample from $\widetilde{\Pi}(\cdot \mid y_{1:n})$, as shown in Algorithm 4.

Algorithm 4: NPL posterior sampling

for $i = 1$ ***to*** B **do**
  Draw $F^{(i)} \sim \mathrm{DP}(\alpha + n, G_n)$
  $\theta^{(i)} = \arg\min_\theta \int \ell(\theta, y)\, dF^{(i)}(y)$
end

Here B is the number of posterior bootstrap samples. One advantage of this sampling scheme is that it is embarrassingly parallel as each of the B samples can be drawn independently. We can thus take advantage of increasingly available multi-core computing, unlike in conventional Bayesian inference as MCMC is inherently sequential.

3.4.4.1 The posterior bootstrap

Sampling from the DP exactly requires infinite computation time if F_π is continuous, but approximate samples can be generated by truncation of the sum in (3.3). For example, we could truncate the stick-breaking and set the remaining weights to 0. Alternatively, we could approximate $w_{1:T} \sim \text{Dir}(\alpha/T, \ldots, \alpha/T)$ with the finite Dirichlet distribution for large T. For further details, see Muliere and Secchi (1996); Ishwaran and Zarepour (2002). We opt for the latter suggestion as Dirichlet weights can be generated efficiently, which leads to a simpler variant of the posterior bootstrap algorithm as shown in Algorithm 5.

Algorithm 5: Posterior bootstrap sampling

Define T as truncation limit
Observed samples are $y_{1:n}$
for $i = 1$ ***to*** B **do**
  Draw prior pseudo-samples $\tilde{y}_{1:T}^{(i)} \overset{\text{iid}}{\sim} F_\pi$
  Draw $(w_{1:n}^{(i)}, \tilde{w}_{1:T}^{(i)}) \sim \text{Dir}\,(1, \ldots, 1, \alpha/T, \ldots, \alpha/T)$
  $\theta^{(i)} = \arg\min_\theta \left\{ \sum_{j=1}^{n} w_j^{(i)}\, \ell(\theta, y_j) + \sum_{k=1}^{T} \tilde{w}_k^{(i)}\, \ell(\theta, \tilde{y}_k^{(i)}) \right\}$
end

For $\alpha = 0$, we simply draw $w_{1:n}^{(i)} \sim \text{Dir}\,(1, \ldots, 1)$, which is no longer an approximation and is equivalent to the Bayesian bootstrap. For $\alpha > 0$, the sampling scheme is asymptotically exact for $T \to \infty$, but this is computationally infeasible. We could fix T to a moderate value, or select it adaptively via adaptive NPL, where we use the stick-breaking construction until the remaining probability is less than ϵ.

3.4.5 Tackling multimodal posteriors with initialization

Multimodal posteriors can arise in Bayesian inference if the likelihood function is non-log-concave like in GMMs (Jin et al., 2016; Stephens, 1999), or if the prior is non-log-concave which can arise when selecting sparse priors (Seeger et al., 2007; Park and Casella, 2008; Lee et al., 2010). Unlike the method by Lyddon et al. (2018) with the MDP, our NPL posterior with the DP is now decoupled from the Bayesian posterior.

There is thus no reliance on an accurate representation of the Bayesian posterior with potential multimodality, which MCMC and VB can often struggle to capture. If our loss function in (3.1) is non-convex (e.g. $-\log f_\theta(y)$ of a GMM), our NPL posterior may also be multimodal. This now presents an optimization issue: solving (3.1) requires non-convex optimization. In general, optimizing non-convex objectives is difficult (see Jain and Kar (2017)), but under smoothness assumption of the loss, we can apply convex optimization methods to find local minima.

3.4.5.1 Random restart for multiple modes

Random restart (see G. E. Boender and H. G. Rinnooy Kan (1987)) can be utilized with convex optimization methods to generate a list of potential global minima then selecting the one with the lowest objective. This involves R random initializations of $\theta^{\text{init}} \sim \pi_0$ for each local optimization, and it was shown by Hu et al. (1994) that the uniform measure for π_0 has good properties for convergence. If the number of modes is finite, then the global minimum will be achieved asymptotically in the limit of the $R \to \infty$. The probability of obtaining the correct global minimum for finite R is related to the size of its basin of attraction. Random restart NPL (RR-NPL) is shown in Algorithm 6.

Algorithm 6: RR-NPL posterior sampling

for $i = 1$ ***to*** B **do**
  Draw $F^{(i)} \sim \text{DP}(\alpha + n, G_n)$
  for $r = 1$ ***to*** R **do**
    Draw $\theta_r^{\text{init}} \sim \pi_0$
    $\theta_r^{(i)} = \text{local arg min}_\theta \left(\int \ell(\theta, y)\, dF^{(i)}(y), \theta_r^{\text{init}} \right)$
  end
  $\theta^{(i)} = \arg\min_r \int \ell(\theta_r^{(i)}, y)\, dF^{(i)}(y)$
end

This is particularly suited to NPL with non-convex loss functions for the following reasons. Firstly, random restart can utilize efficient convex optimization techniques such as quasi-Newton methods, and the restarts can be easily implemented in parallel which

is coherent with our parallelizable sampling scheme. Secondly, we can compromise between accuracy and computational cost by selecting R, as computational cost scales linearly with R (though we can parallelize). The repercussions of an insufficiently large R are not severe: our NPL posterior will incorrectly allocate more density to local modes/saddles but all modes will likely still be present for a sufficiently large B. This is demonstrated in Appendix 3.7.6.2. Finally, the uniform initialization can sample from nonidentifiable posteriors with symmetric modes as their basins of attraction are selected with equal probability.

Practically, uniform initialization may not be possible if the support of the parameter is infinite, e.g. the variance σ^2. In this case, we can pick another π_0 (e.g. Gamma for a positive parameter), or sample uniformly from a truncated support. For adaptively setting R, we can utilize stopping rules as discussed in Appendix 3.7.2.

3.4.5.2 Fixed initialization for local modes

We may be interested in targeting local modes of the posterior when we value interpretability of posterior quantities over exact posterior representation. For example in K-component mixture models, there will be $K!$ symmetrical modes (or sets of modes), and label-switching occurs if the sampler travels between these (Jasra et al., 2005) which impedes useful inference in terms of clustering.

We can target one NPL posterior mode through a fixed initialization scheme by taking advantage of the fact that local optimization methods like expectation-maximization (EM) or gradient ascent are hill-climbers. We initialize each maximization step with the same θ^{init}, causing the sampler to stay within the basin of attraction of the local posterior mode with high probability. We can utilize VB's mode-selection to select θ^{init}, assuming the Bayesian and NPL posterior modes are close. Mean-field VB also tends to underestimate posterior variance (Blei et al., 2017), so we are able to obtain accurate local uncertainty quantification of the mode through this scheme. Fixed initialization NPL (FI-NPL) is shown in Algorithm 7.

Algorithm 7: FI-NPL posterior sampling

Select θ^{init} from mode of interest
for $i = 1$ ***to*** B **do**
  Draw $F^{(i)} \sim \text{DP}(\alpha + n, G_n)$
  $\theta^{(i)} = \text{local arg min}_\theta \left(\int \ell(\theta, y)\, dF^{(i)}(y), \theta^{\text{init}} \right)$
end

3.4.6 Loss-NPL

As we cannot define priors on θ_0 directly, we can instead penalize undesirable properties in the loss

$$\ell(\theta, y) = -\log f_\theta(y) + \gamma g(\theta). \tag{3.4}$$

For example, $g(\theta) = |\theta|$ obtains the Bayesian NPL-Lasso, or we can set $g(\theta) = -\log \pi(\theta)$ if we have some prior preference. We recommend $\gamma = \frac{1}{n}$ if we desire roughly the same prior regularization as in Bayesian inference, where n is the size of the training set. The reasoning is outlined in Appendix 3.7.4. We could also tune γ through desired predictive performance or properties of θ. Note that we are no longer encoding prior beliefs, and are instead expressing an alternative parameter of interest that minimizes the expectation of (3.4).

3.4.7 Related work

We build on the work of Lyddon et al. (2018) which specifies an MDP prior on F_0, and recovers conventional Bayesian inference in the limit of $\alpha \to \infty$. Although the foundations of nonparametric learning are unchanged, our NPL posterior is decoupled from the Bayesian model, offering flexibility in prior measure selection, computational scalability and full multimodal exploration.

NPL unsurprisingly overlaps with other nonparametric approaches to inference. We recover the Bayesian bootstrap (Rubin, 1981) if we set $\alpha = 0$, and further setting $\ell(\theta, y) = -\log f_\theta(y)$ gives the weighted likelihood bootstrap (Newton and Raftery, 1994),

as discussed in Lyddon et al. (2019). Setting the loss to (3.4) and $\alpha = 0$ also returns the fixed prior weighted Bayesian bootstrap (Newton et al., 2018). However, these methods were posited as approximations to the true Bayesian posterior, and the Bayesian bootstrap/weighted likelihood bootstrap are unable to incorporate prior information. The NPL posterior on the other hand is exact and distinct to the conventional Bayesian posterior with theoretical advantages, and we are able to incorporate prior information either through F_π or $\ell(\theta, y)$.

Treating parameters as functionals of the sampling distribution is akin to empirical likelihood methods (Owen, 1988), in which parameters are defined through estimating equations of the form $\int m(\theta, y)\, dF_0(y) = 0$. The definition of a parameter of interest through the loss $\ell(\theta, y)$ is also present in general Bayesian updating introduced by Bissiri et al. (2016), where a coherent posterior over a parameter of interest is obtained without the need to specify a joint generative model. Their target parameter is equivalent to (3.1), and their methodology is built on a notion of coherency.

3.5 Examples

We now demonstrate our method on some examples; the code is available online[1]. We compare NPL to conventional Bayesian inference with the No-U-Turn Sampler (NUTS) by Homan and Gelman (2014), and Automatic Differentiation Variational Inference (ADVI) by Kucukelbir et al. (2017) in Stan (Carpenter et al., 2017). We select these as baselines as they are off-the-shelf algorithms that do not require tuning. Similarly, NPL only requires a weighted likelihood optimization procedure. All NPL examples are run on 4 Azure F72s_v2 (72 vCPUs) virtual machines, implemented in Python. The NUTS and ADVI examples cannot be implemented in an embarrassingly parallel manner, so they are run on a single Azure F72s_v2. We avoid running multiple MCMC chains in parallel as the models are multimodal which may impede mixing, and combining

[1] `https://github.com/edfong/npl`

unmixed chains is unprincipled. For tabulated results, each run was repeated 30 times with different seeds, and we report the mean with 1 standard error. We emphasize again that our NPL posterior is distinct to the conventional Bayesian posterior, so we are comparing the two inference schemes and their associated sampling methods. We include additional empirical comparisons to importance sampling and NPL with an MDP prior in Appendices 3.7.6.3, 3.7.6.4.

3.5.1 Gaussian mixture model

We demonstrate the ability of RR-NPL to accurately sample from a multimodal posterior in a K-component, d-dimensional diagonal GMM toy problem, which NUTS and ADVI fail to do. It should be noted that in addition to the $K!$ symmetrical modes present from label-switching, further multimodality is present due to the non-log-concavity of the likelihood. We further show how FI-NPL can be used in a clustering example with real data to provide accurate local uncertainty quantification which ADVI is unable to do. Our conventional Bayesian model for $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, d\}$ and $k \in \{1, \ldots, K\}$ is

$$\begin{aligned} \mathbf{y}_i \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma} &\sim \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{\mu}_k, \operatorname{diag}(\boldsymbol{\sigma}_k^2)\right), \\ \boldsymbol{\pi} \mid a_0 &\sim \operatorname{Dir}(a_0, \ldots, a_0), \\ \mu_{kj} &\sim \mathcal{N}(0, 1), \\ \sigma_{kj} &\sim \operatorname{logNormal}(0, 1). \end{aligned} \tag{3.5}$$

The posterior is multimodal, and we use ADVI and NUTS for inference. For NPL, we are interested in model fitting, so our loss function is simply the negative log-likelihood

$$\ell(\{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}, \mathbf{y}) = -\log \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{y}; \boldsymbol{\mu}_k, \operatorname{diag}(\boldsymbol{\sigma}_k^2)\right).$$

In the case of small n, we may want to include a regularization term in the loss to avoid singularities of the likelihood. We select the DP prior separately for each example.

3.5.1.1 Toy example: implementation and results

We analyze toy data from a GMM with $K = 3$, $d = 1$ and the following parameters:

$$\boldsymbol{\pi}_0 = \{0.1, 0.3, 0.6\}, \;\; \boldsymbol{\mu}_0 = \{0, 2, 4\}, \;\; \boldsymbol{\sigma}_0^2 = \{1, 1, 1\}.$$

We generate $n_{\text{train}} = 1000$ for model fitting and another $n_{\text{test}} = 250$ held-out for model evaluation with different seeds for each of the 30 runs. For the Bayesian model we set $a_0 = 1$, and for NPL we set $\alpha = 0$ as $n \gg p$. We optimize each bootstrap maximization with a weighted EM algorithm (derived in Appendix 3.7.6.1), and implement this in a modified `GaussianMixture` class from `sklearn.mixture` (Pedregosa et al., 2011). For RR-NPL, we initialize $\boldsymbol{\pi} \sim \text{Dir}(1, \ldots, 1)$, $\mu_{kj} \sim \text{unif}(-2, 6)$ and $\sigma_{kj}^2 \sim \text{IG}(1, 1)$ for each restart. For FI-NPL we initialize with one of the posterior modes from RR-NPL. We produce 2000 posterior samples for each method. We evaluate the predictive performance of each method on held-out test data with the mean log pointwise predictive density (LPPD) as suggested by Gelman et al. (2013), which is described in Appendix 3.7.7.1. A larger value is equivalent to a better fit to the test data.

Figure 3.1 shows the posterior kernel density estimates (KDE) of (μ_1, μ_2) for 1 run of each method. RR-NPL clearly recovers the multi-modality of the NPL posterior, including the symmetry about $\mu_1 = \mu_2$ due to the nonidentifiability of the GMM posterior. NUTS and ADVI remain trapped in one local mode of the Bayesian posterior as expected. Even if we carried out random initialization of NUTS/ADVI over multiple runs, each run would only pick out one mode, and there is no general method to combine the posteriors. ADVI also clearly underestimates the marginal posterior uncertainty. FI-NPL remains in a single mode, showing that we can fix label-switching through this

initialization. However, the FI-NPL mode is not identical to a truncated version of the RR-NPL mode, as posterior mass is not reallocated symmetrically from the other modes. We see in Tables 3.1, 3.2 that RR-NPL has similar mean LPPD on toy test data compared to NUTS, and is twice as fast as NUTS.

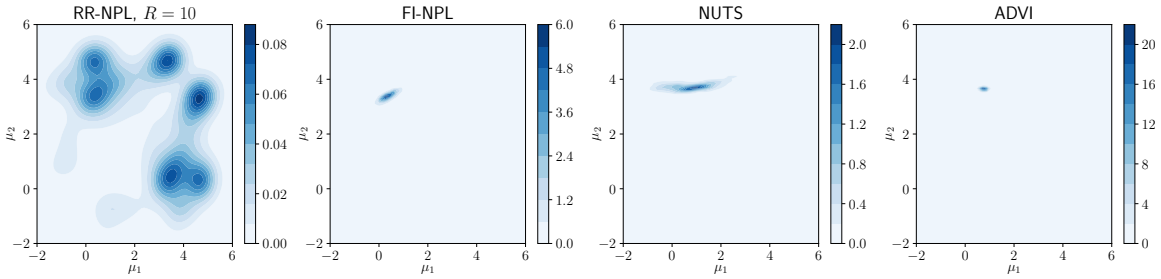


Figure 3.1: Posterior KDE of (μ_1, μ_2) in $K = 3$ toy GMM problem

Table 3.1: Mean LPPD on held-out test data for GMM (with standard error)

Dataset	RR-NPL	FI-NPL	NUTS	ADVI
Toy	**-1.909** (0.007)	-1.911 (0.007)	**-1.908** (0.007)	-1.912 (0.007)
MNIST	/	**2463.4** (4.4)	/	1188.2

Table 3.2: Run-time for 2000 samples for GMM (with standard error)

Dataset	RR-NPL	FI-NPL	NUTS	ADVI
Toy	37.2s (0.8s)	5.5s (0.4s)	1m20s (2.9s)	0.8s (0.1s)
MNIST	/	57.9s (0.2s)	/	5h6m

3.5.1.2 MNIST: implementation and results

We now demonstrate FI-NPL on clustering handwritten digits from MNIST (LeCun and Cortes, 2010), which consists of 28×28 pixel images. In this example $n_{\text{train}} = 10000, n_{\text{test}} = 2500$ and $d = 784$. We normalize all pixel values such that they lie in the interval $[0, 1]$, and set $K = 10$. We believe a priori that many pixels are close to 0, so for ease we elicit a tight normal centering measure for the DP

$$f_\pi(\mathbf{y}) = \prod_{j=1}^{d} \mathcal{N}(y_j; 0, 0.1^2).$$

NUTS is prone to the label-switching problem and is too computationally intensive as ADVI already requires 5 hours, so we only compare FI-NPL to ADVI. We set $a_0 = 1000$ for ADVI, and $\alpha = 1$ for FI-NPL with $T = 500$. We carry out a single run of ADVI to select a local mode, and set θ^{init} of FI-NPL to the ADVI-selected mode. We then carry out 30 repeats of FI-NPL with this initialization, and compare to the original ADVI run. We see in Figure 3.2 that we obtain larger posterior variances in FI-NPL, as ADVI likely underestimates the posterior variances due to the mean-field approximation. Notice the modes are not exactly aligned as the NPL and Bayesian posterior are distinct, and furthermore ADVI is approximate. We conjecture that ADVI does not set components exactly to 0 due to the strong Dirichlet prior. We see in Tables 3.1, 3.2 that FI-NPL is predictively better and runs around 300 times faster than ADVI.

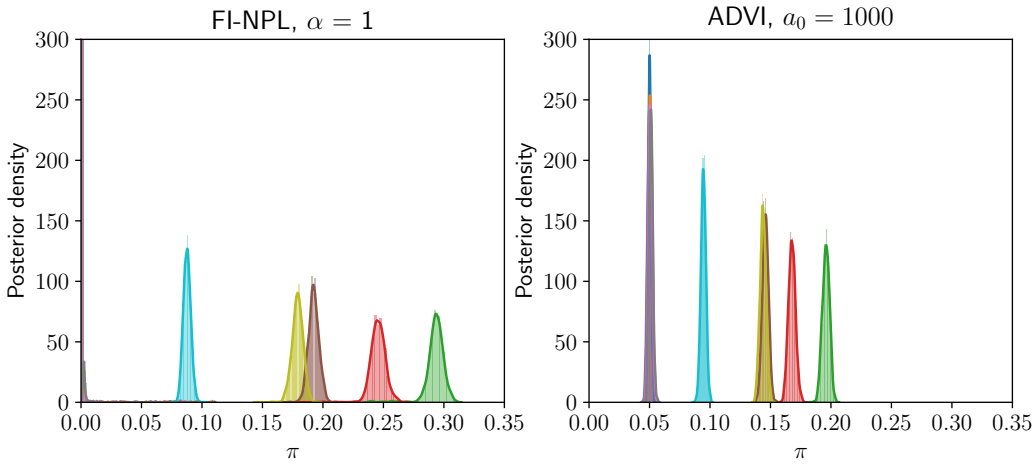


Figure 3.2: Posterior marginal KDEs of $\boldsymbol{\pi}$ for K=10 GMM on MNIST; 5 of the components have been set to 0 for FI-NPL, and likewise to 0.05 for ADVI

3.5.2 Logistic regression with ARD priors

We now demonstrate the predictive performance and computational scalability of loss-NPL in a Bayesian sparse logistic regression example on real datasets. To induce sparsity, we place automatic relevance determination (ARD) priors (MacKay, 1994) on the coefficients with Gamma hyperpriors (Gelman et al., 2008). The conventional

Bayesian model for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, d\}$ is

$$\begin{aligned} y_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \beta_0 &\sim \text{Bernoulli}(\eta_i), \\ \eta_i &= \sigma(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0), \\ \beta_j \mid \lambda_j &\sim \mathcal{N}\left(0, \frac{1}{\lambda_j}\right), \\ \lambda_j \mid a, b &\sim \text{Gamma}\left(a, b\right). \end{aligned}$$

Marginally, the prior is the non-standardized t-distribution with (degrees of freedom, location, squared scale)

$$\beta_j \sim \text{Student-t}\left(2a, 0, \frac{b}{a}\right).$$

This posterior is intractable and potentially multimodal due to the non-log-concavity of the prior, and we carry out conventional Bayesian inference via NUTS and ADVI. When applying loss-NPL to regression, we assume $y, x \overset{\text{iid}}{\sim} F_0$, and place a DP prior on the joint distribution $F_0(y, x)$. We target the parameter which satisfies (3.1) with loss

$$\begin{aligned} \ell(\{\boldsymbol{\beta}, \beta_0\}, \{y, \mathbf{x}\}) = &- \{y \log \eta + (1 - y) \log(1 - \eta)\} \\ &+ \gamma \left(\frac{2a + 1}{2}\right) \sum_{j=1}^{d} \log\left(1 + \frac{\beta_j^2}{2b}\right) \end{aligned}$$

which is the negative sum of the log-likelihood and log-prior, with additional scaling parameter γ. Again our NPL posterior may be multimodal due to the non-convexity of the loss, and so we utilize RR-NPL. It should be noted that our target parameter is now different to conventional Bayesian inference, but our method achieves the common goal of variable selection under a Bayesian framework. For the DP prior, we elicit the centering measure

$$\begin{aligned} f_\pi(y, x) &= f_\pi(y) f_\pi(x), \\ f_\pi(y) &= \text{Bernoulli}(0.5), \\ f_\pi(x) &= \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x). \end{aligned}$$

The prior assumes y, x are independent which is equivalent to assuming $\boldsymbol{\beta} = \mathbf{0}$ a priori. This is appropriate as we believe many components of $\boldsymbol{\beta}$ to be close to 0. The prior on x is its empirical distribution, which is in an empirical Bayes manner where the prior is estimated from the data.

3.5.2.1 Implementation and results

We analyze 3 binary classification datasets from the UCI ML repository (Dheeru and Karra Taniskidou, 2017): 'Adult' (Kohavi, 1996), 'Polish companies bankruptcy 3rd year', (Zikeba et al., 2016), and 'Arcene' (Guyon et al., 2005) with details in Table 3.3. We handle categorical covariates with dummy variables, and normalize all covariates to have mean 0 and standard deviation 1. Missing real values were imputed with the mean, and data with missing categorical values were dropped. We carry out a random stratified train-test split for each of the 30 runs, with 80-20 split for 'Adult', 'Polish' and 50-50 split for 'Arcene' due to the smaller dataset. For both NPL and conventional Bayesian inference, the hyperparameters were set to $a = b = 1$, which was selected by tuning the sparsity of the Bayesian posterior means to a desired value. For NPL, we set $\alpha = 0$ for 'Adult' and 'Polish' as n is sufficiently large, and $\alpha = 1$ for 'Arcene' with $T = 100$ as n is only 100. We set $\gamma = 1/n_{\text{train}}$ for each dataset as explained in Section 3.4.6 for a fair comparison to the conventional Bayesian model. We initialize each optimization with $\beta_j^0 \sim \mathcal{N}(0, 1)$, and select the number of restarts to $R = 1$ for expediency. Optimization was carried out using the L-BFGS-B algorithm (Zhu et al., 1997) implemented in `scipy.optimize` (Jones et al., 01).

We can see in Table 3.4 that loss-NPL is predictively similar or better than NUTS and ADVI, and from Table 3.5 we see that the posterior mean is sparser for loss-NPL. Finally, we see from Table 3.6 that the loss-NPL run-times for 2000 posterior samples are much faster than for NUTS, and comparable to VB. Further measures of predictive performance are provided in Appendix 3.7.7.4.

Table 3.3: UCI datasets descriptions for LogReg

Dataset	Type	d	n_{train}	n_{test}	Positive %
Adult	Cat.	96	36177	9045	24.6
Polish	Real	64	8402	2101	4.8
Arcene	Real	10000	100	100	44.0

Table 3.4: Mean LPPD on held-out test data for LogReg (with standard error)

Dataset	Loss-NPL	NUTS	ADVI
Adult	**-0.326** (0.001)	**-0.326** (0.001)	-0.327 (0.001)
Polish	**-0.229** (0.006)	-3.336 (0.760)	-0.247 (0.009)
Arcene	**-0.449** (0.019)	-0.464 (0.006)	**-0.445** (0.012)

Table 3.5: Percentage of posterior mean $|\beta_j| < \epsilon$ for LogReg (with standard error)

Dataset	ϵ	Loss-NPL	NUTS	ADVI
Adult	0.1	**17.6** (0.5)	16.1 (0.5)	12.1 (0.6)
Polish	0.1	**33.5** (0.9)	15.9 (0.6)	15.8 (0.6)
Arcene	0.01	**87.4** (0.1)	4.7 (0.1)	3.5 (0.1)

Table 3.6: Run-time for 2000 samples for LogReg (with standard error)

Dataset	Loss-NPL	NUTS	ADVI
Adult	2m24s (1.5s)	2h36m (1m)	26.9s (1.3s)
Polish	19.0s (0.7s)	1h20m (4m)	3.3s (0.1s)
Arcene	53.5s (0.2s)	4h31m (10m)	54.2s (0.6s)

3.5.3 Bayesian sparsity-path-analysis

We now utilize loss-NPL to carry out Bayesian sparsity-path-analysis for logistic regression, which allows us to visualize how the responsibility of each covariate changes with the sparsity penalty as discussed by Lee et al. (2012). We use the same ARD prior as Section 3.5.2 with the same initialization scheme, set $\gamma = 1/n$, and elicit a noninformative DP prior with $\alpha = 0$. We found empirically that the results for larger values of R are similar and so the approximation with $R = 1$ is sufficient. We fix a and vary the value of b to favour solutions of different sparsity. This varies the squared scale $c = b/a$ of the Student-t prior with fixed degrees of freedom, where a smaller c corresponds to a heavier sparsity penalty and thus more components are set to 0.

3.5.3.1 Implementation and results

We analyze the genotype/pseudo-phenotype dataset with $n = 500$ as described by Lee et al. (2012), containing patient covariates $\mathbf{x}_i$ which exhibit strong block-like correlations as shown in Figure 3.3. We normalize the covariates to have mean 0 and standard deviation 1. The pseudo-phenotype data is generated by $y_i \sim \text{Bernoulli}(\sigma(\boldsymbol{\beta}^T \mathbf{x}_i))$, where $\boldsymbol{\beta}$ has 5 randomly selected non-zero components out of $d = 50$, with the rest set to 0. Each non-zero component is sampled from $\mathcal{N}(0, 0.2)$, and the exact values of $\boldsymbol{\beta}$ are provided in Appendix 3.7.8.1. We set $a = 1$ and vary $b_t = 0.98^{t-1}$ for $t = \{1, \ldots, 450\}$, and generate 4000 posterior samples for each setting. The posterior medians of the

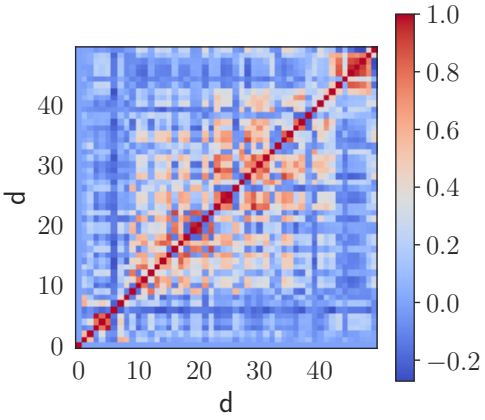


Figure 3.3: Correlations of covariates $\boldsymbol{x}$ from genetic dataset

non-zero components of $\boldsymbol{\beta}$ with 80% central credible interval are shown in Figure 3.4 for a range of $\log c$ values. Both the posterior median and central credible intervals are estimated through the appropriate order statistics of the posterior samples (Gelman et al., 2013). We can see that β_{10}, β_{14} and β_{24} have early predictive power as their credible intervals remain large despite a significant sparsity penalty (small $\log(c)$), whilst the other two coefficients β_{31}, β_{37} are masked. A plot of the absolute medians for all components is included in Appendix 3.7.8.2. For β_{10} and β_{14}, the median is close to 0 but the credible interval is large which is due to the multimodality of the marginal posterior. This multimodality is also responsible for the jitter in the median around $\log(c) = -6.5$ for β_{14} in Figure 3.4 ; the true median likely lies between the two separated modes but the finite posterior sample size causes the sample median

to jump between the two. A posterior marginal KDE plot of β_{14} changing with $\log c$ is shown in Figure 3.5, allowing us to visualize how the importance of the covariate changes with the sparsity penalty. We observe the bimodality in the marginal posterior for $\log(c) < -4$ as expected from the above discussion.

Loss-NPL required 5 minutes 24 seconds to generate all 450×4000 posterior samples. The computational speed of NPL enables fast Bayesian analysis of large datasets with different hyperparameter settings, allowing for Bayesian variable selection analysis.

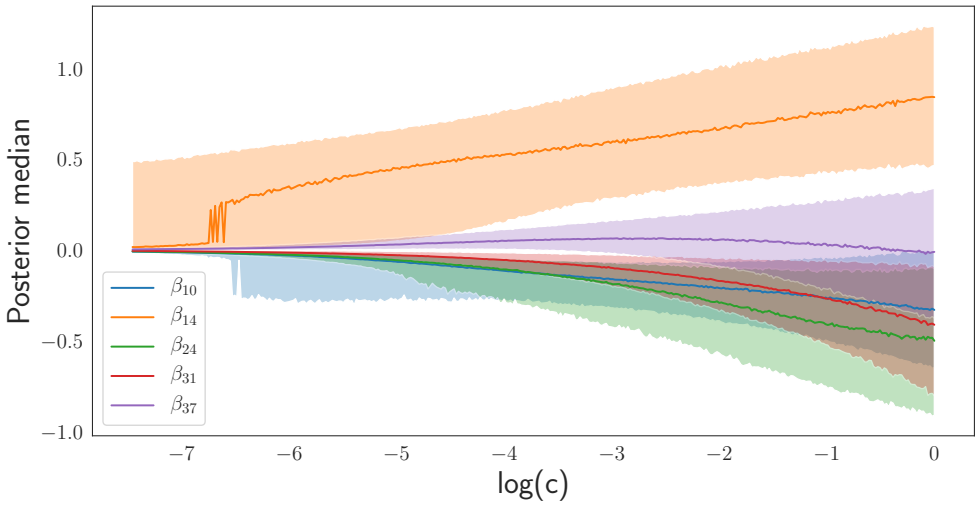


Figure 3.4: Lasso-type plot for posterior medians of non-zero $\boldsymbol{\beta}$ with 80% credible intervals against $\log(c)$ from genetic dataset

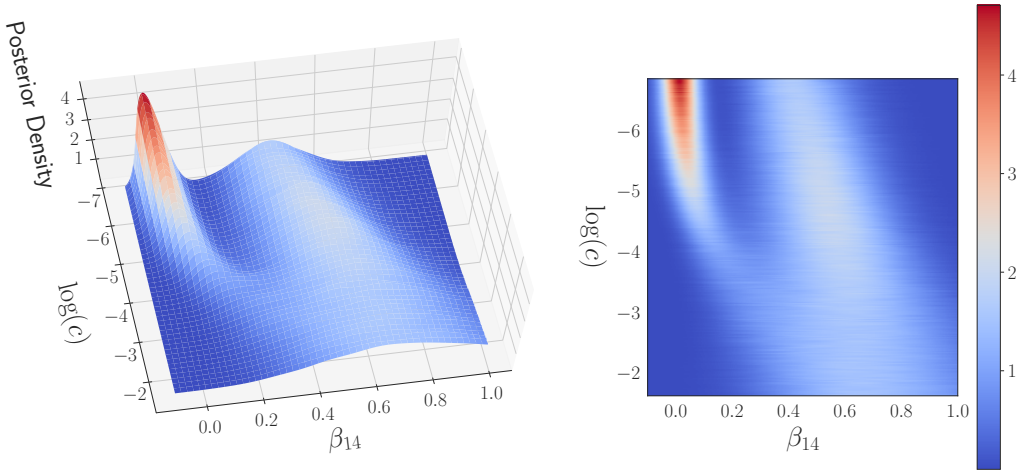


Figure 3.5: Posterior marginal KDE of β_{14} against $\log(c)$ from genetic dataset

3.6 Discussion

We have introduced a variant of Bayesian nonparametric learning (NPL) with a Dirichlet process (DP) prior on the sampling distribution F_0, which leads to highly scalable exact inference under model misspecification, detached from the conventional Bayesian posterior. This method admits a sampling scheme for multimodal posteriors that allows for full mode exploration, which involves a non-convex optimization that we solve through random restart. We demonstrated that NPL can perform predictively better than conventional Bayesian inference, while providing exact uncertainty quantification.

For future work, the small sample performance of NPL could be further explored and compared to conventional Bayesian inference; we currently recommend NPL for moderate to large values of n. The scaling of the number of repeats R with increasing dimension for full mode exploration would also be a future avenue of research.

Acknowledgements

We would like to thank our anonymous reviewers for their careful analyses of our paper and their valuable input. We also thank Luke Kelly for useful discussions, Ho Chung Leon Law for helpful feedback and Anthony Lee for providing the genotype dataset. EF is funded by The Alan Turing Institute Doctoral Studentship, under the EPSRC grant EP/N510129/1. SL is funded by the EPSRC OxWaSP CDT, through EP/L016710/1. CH is supported by The Alan Turing Institute, the HDR-UK, the Li Ka Shing Foundation, and the MRC.

3.7 Appendix

3.7.1 Eliciting the prior Dirichlet process

3.7.1.1 Intuition of the prior F_π

The parameter of interest when model fitting (Walker, 2013) is

$$\begin{aligned}\theta_0 &= \arg\max_\theta \int \log f_\theta(y)\, dF_0(y) \\ &= \arg\min_\theta d_{\mathrm{KL}}(f_0, f_\theta).\end{aligned}$$

The prior on F_0 is

$$[F \mid \alpha, F_\pi] \sim \mathrm{DP}\left(\alpha, F_\pi\right).$$

The effects of the implicit prior on θ_0 due to F_π when model-fitting can be seen in the limit of $\alpha \to \infty$ under regularity conditions:

$$\theta_0 \xrightarrow{p} \arg\min_\theta d_{\mathrm{KL}}(f_\pi, f_\theta). \tag{3.6}$$

In the limit, the prior collapses on one of the points that minimizes the KL divergence between the prior centering density and the model. Intuitively, the prior regularizes θ_0 towards (3.6), and α acts as weighting between F_π and $F_n = \frac{1}{n}\sum_{i=1} \delta_{y_i}$. It is thus a measure of belief that F_π is the true sampling distribution.

3.7.1.2 Selecting α through the mean functional

We can tune α through the a priori variance of the mean functional

$$\begin{aligned}\theta_\mu(F) &= \arg\min_\theta \int (y-\theta)^2\, dF(y) \\ &= \int y\, dF(y).\end{aligned} \tag{3.7}$$

If $F \sim \mathrm{DP}(\alpha, F_\pi)$, then the a prior variance of (3.7) follows from the properties of

the Dirichlet process (DP):

$$\text{Var}\left[\theta_\mu(F)\right] = \frac{\text{Var}_{F_\pi}\left[y\right]}{1+\alpha}$$

and so we can elicit α from a priori knowledge of $\text{Var}\left[\theta_\mu(F)\right]$.

3.7.2 Stopping rules for adaptively selecting R

Although not explored in our paper, we can utilize heuristic stopping rules for adaptively selecting R for full mode exploration when sampling from the NPL posterior. A simple example is to stop the repeats if there have been no improvements in the optimized function value for the last m repeats, where m is the parameter of the stopping rule. More complex methods involve estimating the missing probability mass due to local minima not being observed, and thresholding based on that. See Betrò and Schoen (1987); Dick et al. (2014) for a comparison of some methods. Although there is no clear answer for selecting R, we can also parallelize over restarts to alleviate the computation burden.

3.7.3 Stochastic subsampling

For very large n, we can utilize stochastic gradient methods by subsampling to optimize the weighted loss. The full weighted loss and gradient are defined as

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} w_i\, \ell(\theta, y_i),$$
$$\nabla_\theta L(\theta) = \sum_{i=1}^{n} w_i\, \nabla_\theta \ell(\theta, y_i).$$

If we subsample a mini-batch $\tilde{y}_{1:m} \overset{\text{iid}}{\sim} \sum_{i=1}^{n} w_i\, \delta_{y_i}$, we can then calculate the mini-batch gradient

$$\nabla_\theta L^m(\theta) = \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \ell(\theta, \tilde{y}_i).$$

The mini-batch gradient is unbiased:

$$\mathbb{E}\left[\nabla_\theta L^m(\theta)\right] = \mathbb{E}\left[\nabla_\theta \ell(\theta, \tilde{y})\right] = \sum_{i=1}^{n} w_i \, \nabla_\theta \ell(\theta, y_i).$$

Setting $m = 1$ allows use to use stochastic gradient descent (SGD) and its variants which improves scalability. Furthermore, extensions to SGD such as ADAGRAD (Duchi et al., 2011) and ADAM (Kingma and Ba, 2014) help with escaping saddle points, which can potentially reduce the number of R required for RR-NPL to obtain full mode exploration.

3.7.4 Selecting γ in loss-NPL

For loss-NPL we can set the loss function to

$$\ell(\theta, y) = -\log f_\theta(y) - \gamma \log \pi(\theta).$$

In this case, we recommend the scaling parameter to be $\gamma = 1/n$ if we want roughly the same prior regularization of $\pi(\theta)$ as in traditional Bayesian inference. This can be seen when we look at the expected of $\int \ell(\theta, y)\, dF$ for $\alpha = 0$ (i.e. $F \sim \text{DP}\left(n, \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}\right)$):

$$\mathbb{E}\left[\int \ell(\theta, y)\, dF(y)\right] = -\frac{1}{n} \sum_{i=1}^{n} \log f_\theta(y_i) - \gamma \log \pi(\theta).$$

We obtain the same weighting as in Bayesian inference between the log-likelihood and log-prior for $\gamma = 1/n$.

3.7.5 Toy example: normal location model

We now empirically demonstrate the small sample performance of NPL and the role of the prior concentration α in a toy normal location model problem. Suppose the model of interest is $f_\theta(y) = \mathcal{N}(y; \theta, \sigma^2)$ with known σ^2. Our parameter of interest is defined

as

$$\begin{aligned}\theta_0 &= \arg\max_\theta \int \log f_\theta(y)\, dF_0(y) \\ &= \arg\min_\theta \int (y-\theta)^2\, dF_0(y).\end{aligned}$$

If we set the derivative of the objective to 0, we obtain

$$\theta_0 = \int y\, dF_0(y).$$

If we believe our parametric model to be accurate, we can place a prior $\pi(\theta) = \mathcal{N}(\theta; 0, \tau^2)$ on θ. The centering measure on our DP is thus

$$f_\pi(y) = \int f_\theta(y)\, \pi(\theta)\, d\theta = \mathcal{N}(y; 0, \sigma^2 + \tau^2).$$

When $n = 0$, our NPL prior density $\widetilde{\pi}(\theta)$ is approximately normal (Yamato, 1984) from the properties of the DP:

$$\widetilde{\pi}(\theta) \approx \mathcal{N}\left(\theta; 0, \frac{\sigma^2 + \tau^2}{1+\alpha}\right).$$

3.7.5.1 Implementation and results

We sample the observables $y \sim \mathcal{N}(1, 1^2)$ and set our parametric prior variance to $\tau^2 = 1$. We simulate the NPL posterior in Figure 3.6 for various values of n and $\alpha = 1$, and compare it to the tractable traditional Bayesian posterior with the same model $\{f_\theta, \pi(\theta)\}$. For the NPL posterior bootstrap sampler, we generate $B = 10000$ samples and truncate the DP at $T = 1000$.

We see from Figure 3.6 that the NPL prior is approximately normal ($n = 0$), with same mean and variance due to the choice of α. For large n, the NPL posterior and Bayesian posterior are similar, due to the first order correctness of the weighted likelihood bootstrap (Newton and Raftery, 1994). For smaller values of n, the NPL posterior is non-normal, as our prior is not a conjugate prior on θ. For $n = 1$, the

sample observed is close to 0 so the posterior uncertainty is small despite only observing one sample; this suggests that NPL may be better suited to moderate to large values of n. Figure 3.7 shows the effect on the NPL posterior of increasing prior strength α for $n = 1$, which regularizes the posterior but also causes it to concentrate about 0.

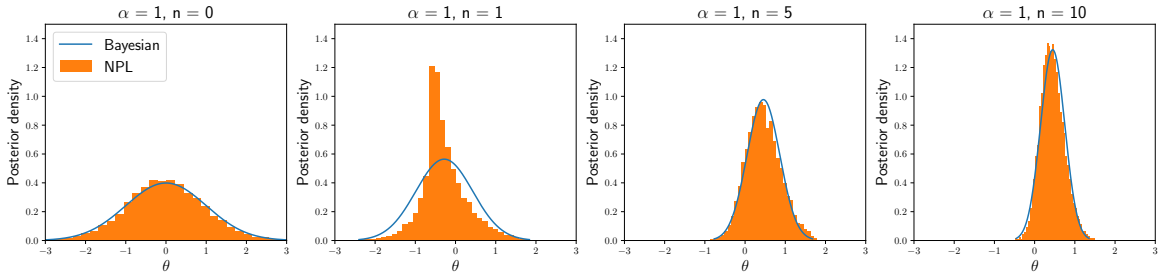


Figure 3.6: NPL posterior and Bayesian posterior for fixed α and increasing n in normal location model

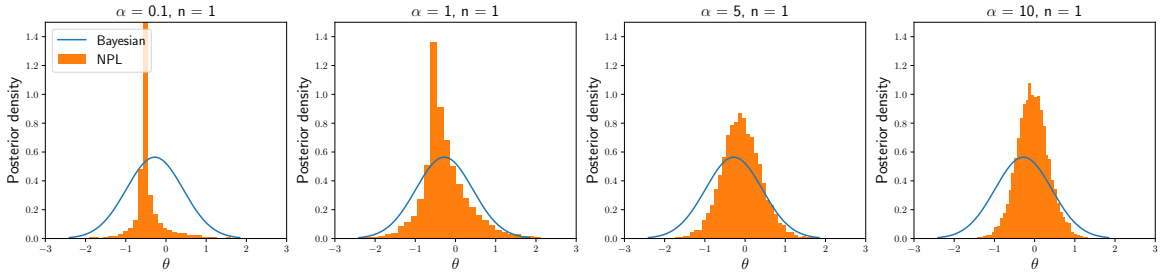


Figure 3.7: NPL posterior and Bayesian posterior for increasing α and fixed n in normal location model

3.7.6 Gaussian mixture model

3.7.6.1 Optimization details

We derive the EM algorithm that maximizes the weighted likelihood of the diagonal-covariance GMM:

$$\begin{aligned}\mathcal{L}^w(\theta) &= \sum_{i=1}^{n} w_i \log f_\theta(\mathbf{y}_i) \\ &= \sum_{i=1}^{n} w_i \left\{\log f_\theta(\mathbf{y}_i, z_i) - \log f_\theta(z_i \mid \mathbf{y}_i)\right\}.\end{aligned}$$

Taking an expectation over the posterior $f_{\theta'}(z_{1:n} \mid \mathbf{y}_{1:n})$, we obtain

$$\begin{aligned}\mathcal{L}^w(\theta) &= \sum_{i=1}^{n} w_i \sum_{z_i} f_{\theta'}(z_i \mid \mathbf{y}_i)\ \log f_\theta(\mathbf{y}_i, z_i) - \sum_{i=1}^{n} w_i \sum_{z_i} f_{\theta'}(z_i \mid \mathbf{y}_i)\ \log f_\theta(z_i \mid \mathbf{y}_i) \\ &= \sum_{i=1}^{n} w_i\, Q^i(\theta \mid \theta') - \sum_{i=1}^{n} w_i\, H^i(\theta \mid \theta').\end{aligned}$$

Taking the difference of the weighted likelihood with θ'

$$\begin{aligned}\mathcal{L}^w(\theta) - \mathcal{L}^w(\theta') &= \sum_{i=1}^{n} w_i \left\{Q^i(\theta \mid \theta') - Q^i(\theta' \mid \theta')\right\} \\ &\quad + \sum_{i=1}^{n} w_i \left\{H^i(\theta' \mid \theta') - H^i(\theta \mid \theta')\right\}.\end{aligned}$$

From Gibbs' inequality,

$$H^i(\theta' \mid \theta') \geq H^i(\theta \mid \theta').$$

As all $w_i \geq 0$,

$$\mathcal{L}^w(\theta) - \mathcal{L}^w(\theta') \geq \sum_{i=1}^{n} w_i \left\{Q^i(\theta \mid \theta') - Q^i(\theta' \mid \theta')\right\}.$$

So by maximizing $\sum_{i=1}^{n} w_i\, Q^i(\theta \mid \theta')$ w.r.t. θ, we cannot decrease the weighted log-

likelihood. As a reminder, the log-likelihood for each datapoint is

$$\log f_\theta(\mathbf{y}_i) = \log \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{y}_i; \boldsymbol{\mu}_k, \operatorname{diag}\left(\boldsymbol{\sigma}_k^2\right)\right).$$

At the expectation step, we calculate

$$f_\theta(z_i = k \mid \mathbf{y}_i) = \frac{\prod_{j=1}^{d} \mathcal{N}(y_{ij}; \mu_{kj}, \sigma_{kj}^2)\, \pi_k}{\sum_{k=1}^{K} \prod_{j=1}^{d} \mathcal{N}(y_{ij}; \mu_{kj}, \sigma_{kj}^2)\, \pi_k}.$$

The maximization step is then:

$$\begin{aligned}
\hat{\pi}_k &= \sum_{i=1}^{n} w_i\, f_{\theta'}(z_i = k \mid \mathbf{y}_i), \\
\hat{\mu}_{kj} &= \frac{\sum_{i=1}^{n} w_i\, f_{\theta'}(z_i = k \mid \mathbf{y}_i)\, y_{ij}}{\hat{\pi}_k}, \\
\hat{\sigma}_{kj}^2 &= \frac{\sum_{i=1}^{n} w_i\, f_{\theta'}(z_i = k \mid \mathbf{y}_i)\, (y_{ij} - \hat{\mu}_{kj})^2}{\hat{\pi}_k}.
\end{aligned}$$

3.7.6.2 Toy example

We see the posterior KDE plots for (π_1, π_2), (μ_1, μ_2) and (σ_1^2, σ_2^2) in Figures 3.8, 3.9, 3.10, and for increasing R in Figures 3.11, 3.12, 3.13. For RR-NPL, we observe multimodality in addition to symmetry about the diagonal due to label-switching. Smaller values of R exhibit an over-representation of local modes/saddles, and the posterior accuracy increases for larger R. We also show the run-times for different R for RR-NPL in Table 3.7, and we see that the run-time increases roughly linearly with R.

Table 3.7: Run-time (seconds) for 2000 posterior samples on Azure for different values of R with RR-NPL (with standard error)

	$R = 1$	$R = 2$	$R = 5$	$R = 10$
Toy Sep	4.9 (0.1)	8.0 (0.2)	19.0 (0.4)	37.2 (0.8)

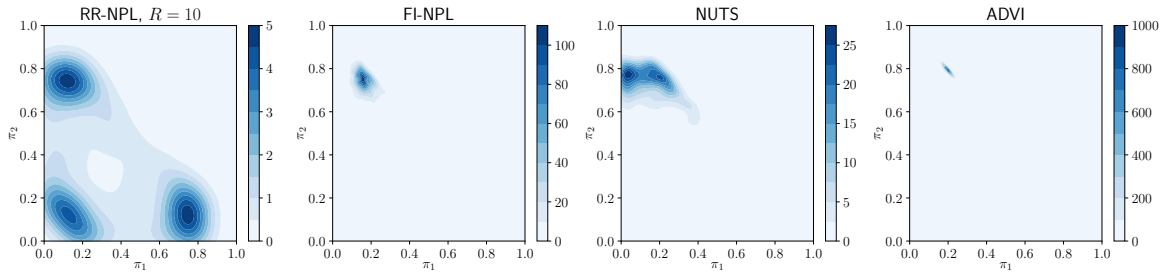


Figure 3.8: Posterior KDE of (π_1, π_2) in $K = 3$ separable toy GMM problem

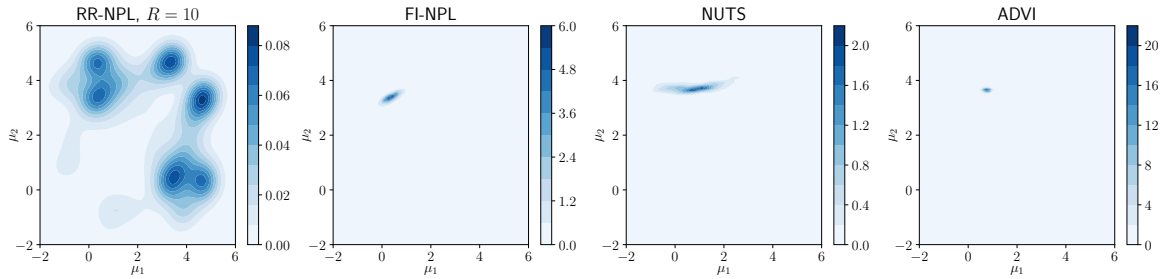


Figure 3.9: Posterior KDE of (μ_1, μ_2) in $K = 3$ separable toy GMM problem

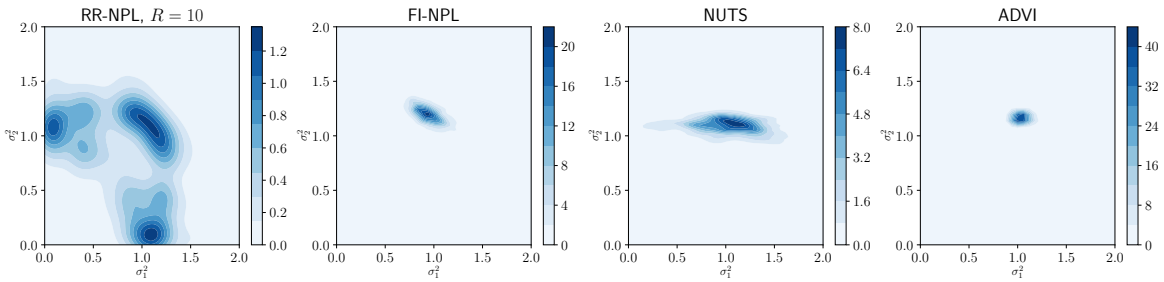


Figure 3.10: Posterior KDE of (σ_1^2, σ_2^2) in $K = 3$ separable toy GMM problem

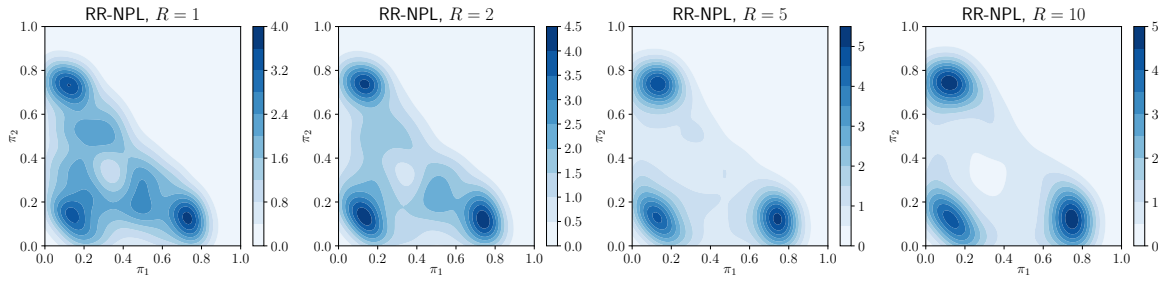


Figure 3.11: Posterior KDE of (π_1, π_2) in $K = 3$ separable toy GMM problem for increasing R

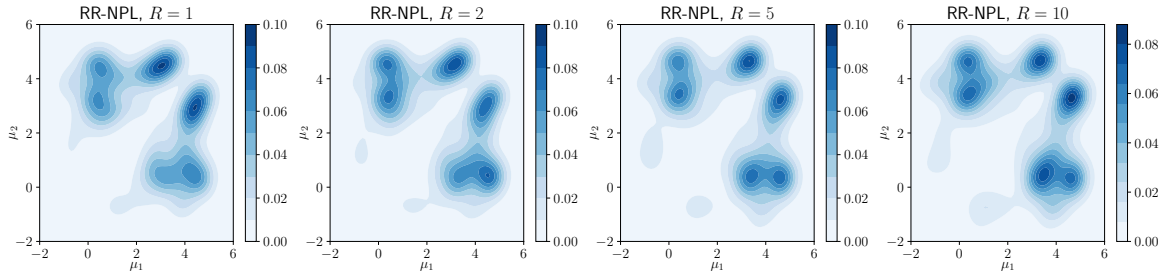


Figure 3.12: Posterior KDE of (μ_1, μ_2) in $K = 3$ separable toy GMM problem for increasing R

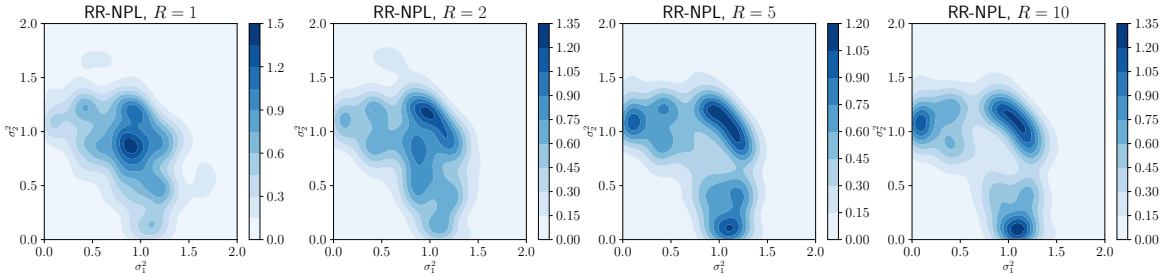


Figure 3.13: Posterior KDE of (σ_1^2, σ_2^2) in $K = 3$ separable toy GMM problem for increasing R

3.7.6.3 Comparison to importance sampling

As suggested by our helpful reviewers and meta-reviewer, we include a discussion and empirical comparison with importance sampling (IS) here. The efficacy of IS hinges on finding a good approximating proposal density, which is in itself a challenging research question. Generic proposals can lead to very large and difficult to detect errors (e.g. see Bishop (2006, pg. 534)). Moreover, the variance of the IS approximation is driven by the variance of the importance weights $p(x)/q(x)$, for $x \sim q(x)$ approximating $p(x)$. The proposal distribution needs to capture all aspects of the target distribution and not just the modes, otherwise the ratio $p(x)/q(x)$ may not be bounded. This makes IS challenging to apply in moderate to high dimensional problems and especially when there is multimodality; we will now demonstrate this in our toy GMM example from Section 3.5.1 of the main paper.

We set the task of estimating the mean log pointwise predictive density (LPPD) (Gelman et al., 2013) of the 250 held-out test data points; the LPPD is defined in Appendix 3.7.7.1. We use self-normalized importance sampling (SNIS) as the posterior is unnormalized, so the estimate of the posterior predictive is defined as

$$p(\tilde{y} \mid y_{1:n}) \approx \frac{\sum_{b=1}^{B} w_b \, f(\tilde{y} \mid \boldsymbol{\theta}_b)}{\sum_{b=1}^{B} w_b} = \sum_{b=1}^{B} \tilde{w}_b \, f(\tilde{y} \mid \boldsymbol{\theta}_b),$$
$$w_b = \frac{f(y_{1:n} \mid \boldsymbol{\theta}_b) \, \pi(\boldsymbol{\theta}_b)}{q(\boldsymbol{\theta}_b)}, \quad \boldsymbol{\theta}_b \sim q(\boldsymbol{\theta})$$

where $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$ is 9-dimensional, and $\{f(y \mid \boldsymbol{\theta}), \pi(\boldsymbol{\theta})\}$ is the GMM likelihood and prior as defined in (3.5) of the main paper. The choice of the proposal is non-trivial, as both the posterior and the function being integrated are multimodal. We use a broad proposal of the same form as the prior as shown below:

$$q(\boldsymbol{\pi}) = \text{Dir}(0.1, \ldots, 0.1), \quad q(\mu_{kj}) = \mathcal{N}(0, 25), \quad q(\sigma_{kj}) = \text{logNormal}(0, 1).$$

The proposal has support in the true value of $\boldsymbol{\theta}$ and thus should have support in regions

of high $f(\tilde{y} \mid \boldsymbol{\theta})$. As we are integrating over n_{test} distinct likelihoods with varying multimodality, a broad proposal is appropriate. We carry out SNIS with $B = 10^7$ implemented on the same virtual machines, where this choice of B is determined by approximately matching the time required to produce 2000 posterior samples with NPL. We repeat 10 runs with the same training and test set but vary the seed for the samplers, and report the mean and standard error (SE) of the mean LPPD. For SNIS, we also report the effective sample size, defined as $\text{ESS} = 1/\sum_{b=1}^{B} \tilde{w}_b^2$. This is shown below in Table 3.8, and we see that the SE for SNIS is an order of magnitude larger than that of RR-NPL. Furthermore, the ESS of SNIS is extremely poor, likely due to the difficulty in selecting a good proposal in this 9-dimensional problem. We notice that most of the weights are very close to 0 except for a few that dominate. These effects will be increasingly amplified in higher dimensions, and is thus why IS fails in problems of even moderate dimensionality.

Table 3.8: Performance on held-out test data for toy GMM (with standard error)

	RR-NPL	SNIS
Mean of LPPD estimate	-1.7984	-1.8070
SE of LPPD estimate	2×10^{-4}	4.4×10^{-3}
ESS	2000	1.75 (0.25)
Run-time	29.3s (0.2s)	31.0s (2.8s)

3.7.6.4 Comparison to NPL with mixture of Dirichlet processes prior

As explained in the main paper, NPL with a mixture of DPs prior (MDP-NPL) as introduced in Lyddon et al. (2018) requires accurate sampling of the Bayesian posterior, which MCMC and VB may not be able to provide. Another difference is the meaning of the parameter α in the two NPL schemes, which is the concentration of the MDP and the DP. In MDP-NPL, the concentration represents the strength of belief that the centering traditional Bayesian model is correctly specified, whereas in NPL with a DP prior (DP-NPL) it is the strength of belief that F_π is the true sampling distribution. We see in Appendix 3.7.1.1 that the limit of $\alpha \to \infty$ gives different results between the two NPL schemes, while $\alpha \to 0$ gives the same limit.

As evident in Figure 3.1 of the main paper, NUTS and ADVI clearly fail at representing the multimodality in our toy GMM problem, and as a result it is not possible to carry out MDP-NPL in that example. We thus compare DP-NPL to MDP-NPL experimentally in an easier toy GMM problem in which NUTS can represent the posterior accurately. We carry out the same experiment with alternative GMM parameters:

$$\boldsymbol{\pi}_0 = \{0.1, 0.3, 0.6\}, \;\; \boldsymbol{\mu}_0 = \{0, 1, 2\}, \;\; \boldsymbol{\sigma}_0^2 = \{1, 1, 1\}.$$

The means are closer together, and so NUTS can mix properly as shown in Figure 3.14. For MDP-NPL, we center the MDP with the Bayesian model given in (3.5) of the main paper, set $\alpha = 1000$ and carry out the posterior bootstrap step using the posterior samples generated by NUTS and maximizing the weighted likelihood. For DP-NPL, we elicit the centering measure $f_\pi(y) = \mathcal{N}(y; 0, 1)$ and set $\alpha = 10$. For both NPL schemes, we carry out 10 random restarts for each posterior sample with the same initialization scheme as in the main paper.

The posterior KDE plots for (π_1, π_2), (μ_1, μ_2) and (σ_1^2, σ_2^2) are shown in Figures 3.14, 3.15, 3.16. The difference between the DP-NPL and MDP-NPL posteriors is small, and both can represent the multimodality well. Predictively, DP-NPL and MDP-NPL perform similarly as shown in Table 3.9, but the run-times are much greater for MDP-NPL as shown in Table 3.10. This is because we first need to generate the Bayesian posterior samples via NUTS before we can proceed to the posterior bootstrap, and so the run-time of NUTS is still the bottleneck.

Table 3.9: Mean LPPD on held-out test data for inseparable GMM (with standard error)

	DP-NPL	MDP-NPL	NUTS	ADVI
Toy	-1.612 (0.007)	-1.610 (0.007)	-1.609 (0.007)	-1.613 (0.007)

Table 3.10: Run-time for 2000 samples for inseparable GMM (with standard error)

	DP-NPL	MDP-NPL	NUTS	ADVI
Toy	52.6s (1.2s)	3m2s (2.4s)	2m12s (2.2s)	0.8s (0.02s)

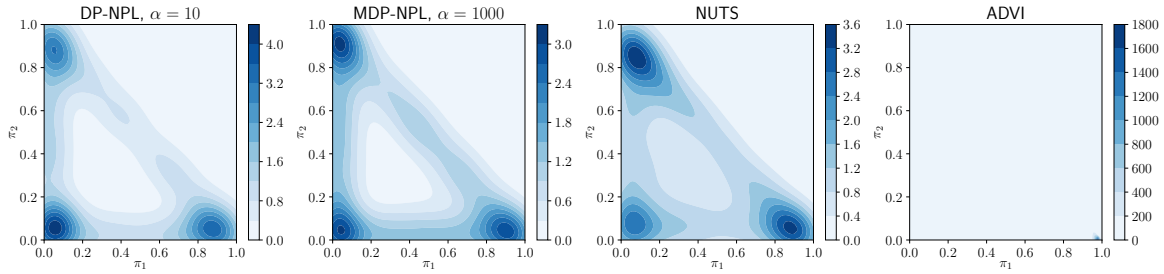


Figure 3.14: Posterior KDE of (π_1, π_2) in $K = 3$ inseparable toy GMM problem

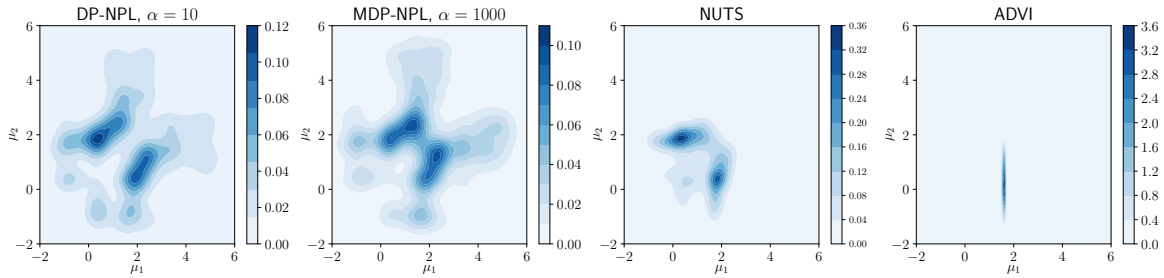


Figure 3.15: Posterior KDE of (μ_1, μ_2) in $K = 3$ inseparable toy GMM problem

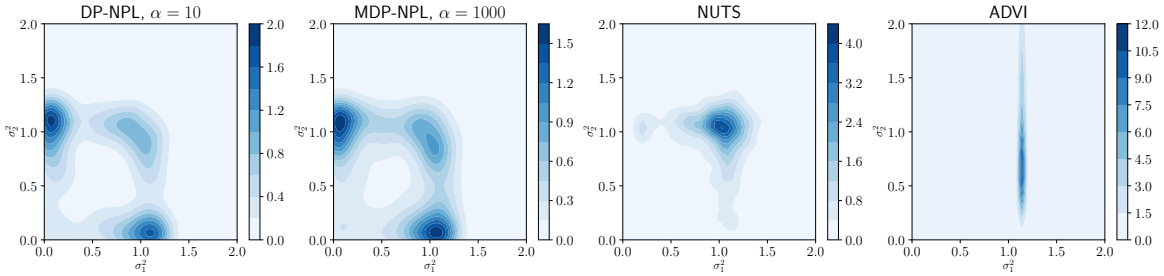


Figure 3.16: Posterior KDE of (σ_1^2, σ_2^2) in $K = 3$ inseparable toy GMM problem

3.7.7 Logistic regression with ARD priors

3.7.7.1 Predictive performance

For all the following measures of predictive performance on held-out test data, we can use a Monte Carlo estimate of the predictive distribution of a test data point $(\tilde{y}, \tilde{x})$:

$$p(\tilde{y} \mid \tilde{x}, y_{1:n}, x_{1:n}) = \int f(\tilde{y} \mid \tilde{x}, \boldsymbol{\beta})\, d\widetilde{\Pi}(\boldsymbol{\beta} \mid y_{1:n}, x_{1:n})$$
$$\approx \frac{1}{B} \sum_{b=1}^{B} f(\tilde{y} \mid \tilde{x}, \boldsymbol{\beta}_b),$$
$$\boldsymbol{\beta}_b \sim \widetilde{\Pi}(\cdot \mid y_{1:n}, x_{1:n}),$$

where $f(y \mid x, \boldsymbol{\beta})$ is the likelihood, $\widetilde{\Pi}(\cdot \mid y_{1:n}, x_{1:n})$ is the NPL or Bayesian posterior, B is the number of posterior samples, and $(y_{1:n}, x_{1:n})$ is the training set. We evaluate the mean LPPD of held-out test data as a measure of predictive performance:

$$\text{Mean LPPD} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \log p(\tilde{y}_i \mid \tilde{x}_i, y_{1:n}, x_{1:n}).$$

Below, we additionally include the mean squared error (MSE) here on held-out test data, defined

$$\text{MSE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (p(\tilde{y}_i \mid \tilde{x}_i, y_{1:n}, x_{1:n}) - \tilde{y}_i)^2.$$

Finally, we also report the percentage accuracy, defined

$$\text{P.a.} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \hat{y}_i^{\tilde{y}_i} (1 - \hat{y}_i)^{(1-\tilde{y}_i)},$$
$$\hat{y}_i = \mathbb{I}(p(\tilde{y}_i \mid \tilde{x}_i, y_{1:n}, x_{1:n}) > 0.5)$$

where $\mathbb{I}$ is the indicator function.

3.7.7.2 Sparsity measure

For the sparsity results, we simply calculate the posterior mean $\hat{\boldsymbol{\beta}} = \frac{1}{B} \sum_{b=1}^{B} \boldsymbol{\beta}_b$, where $\boldsymbol{\beta}_b \sim \widetilde{\Pi}(\cdot \mid y_{1:n}, x_{1:n})$ as above. We then report the percentage of components of $\hat{\boldsymbol{\beta}}$ that have absolute value less than ϵ.

3.7.7.3 Optimization details

L-BFGS-B (Zhu et al., 1997) is a quasi-Newton method which requires the gradient, which for the marginal Student-t distribution is defined for $j \in \{1, \ldots, d\}$ as

$$\frac{\partial \ell(\{\boldsymbol{\beta}, \beta_0\}, \{y, x\})}{\partial \beta_j} = -(y - \eta)\, x_j + \gamma \left(\frac{2a + 1}{2b + \beta_j^2} \right) \beta_j,$$
$$\frac{\partial \ell(\{\boldsymbol{\beta}, \beta_0\}, \{y, x)}{\partial \beta_0} = -(y - \eta)\,.$$

3.7.7.4 Additional results

We can see in Tables 3.11, 3.12 that loss-NPL performs equally or better than NUTS and ADVI predictively in MSE and classification accuracy as well as LPPD. A posterior marginal density plot for β_{13} in the 'Adult' dataset is shown in Figure 3.17 for reference.

Table 3.11: MSE on held-out test data (with standard error)

Dataset	Loss-NPL	NUTS	ADVI
Adult	**0.104** (0.000)	**0.104** (0.000)	0.105 (0.000)
Polish	**0.056** (0.002)	0.524 (0.043)	0.058 (0.002)
Arcene	**0.134** (0.005)	0.152 (0.003)	0.143 (0.004)

Table 3.12: Predictive accuracy % on held-out test data (with standard error)

Dataset	Loss-NPL	NUTS	ADVI
Adult	**84.92** (0.05)	**84.92** (0.05)	84.84 (0.05)
Polish	**93.65** (0.31)	37.27 (4.35)	93.51 (0.28)
Arcene	**81.73** (0.69)	77.80 (0.77)	79.70 (0.62)

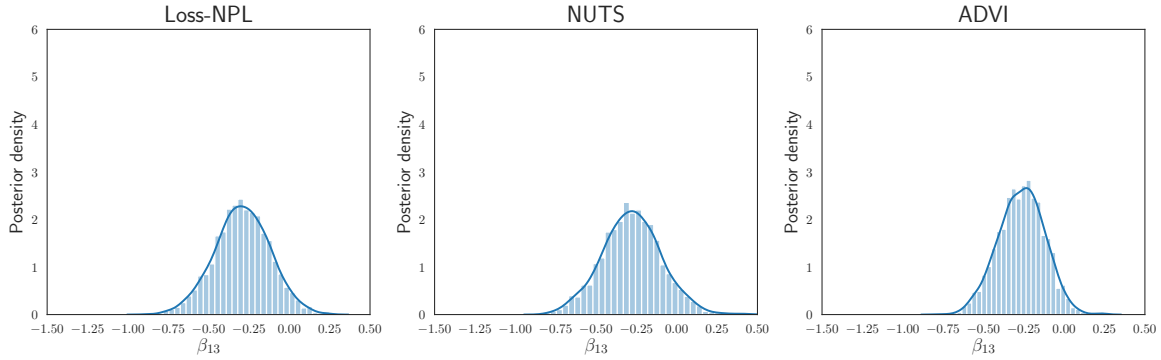


Figure 3.17: Posterior marginal KDE of β_{13} for 'Adult' dataset

3.7.8 Bayesian sparsity-path-analysis

3.7.8.1 Values of β

We follow the setting of $\boldsymbol{\beta}$ in Lee et al. (2012): the 5 non-zero indices and their respective values are

$$\mathcal{I} = \{10, 14, 24, 31, 37\},$$

$$\beta_{\mathcal{I}} = \{-0.2538, 0.4578, -0.1873, -0.1498, 0.0996\}.$$

3.7.8.2 Variable selection

We see more clearly from Figure 3.18 that β_{10}, β_{14} and β_{24} have early predictive power, with one other null-coefficient showing early predictive importance. The other two non-zero coefficients are masked.

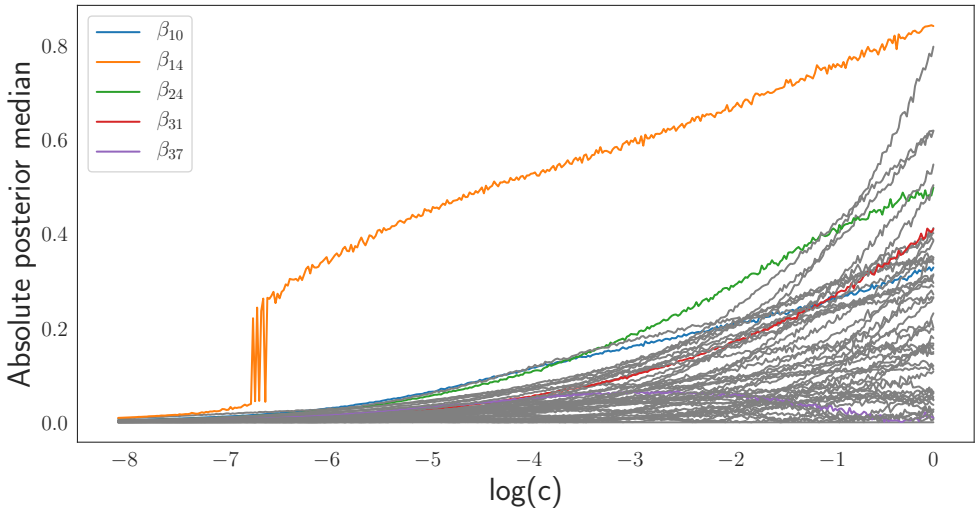


Figure 3.18: Lasso-type plot for absolute posterior medians of $\boldsymbol{\beta}$ against $\log(c)$ from genetic dataset with non-zero components in colour

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap
Publication Status	**x** Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Fong, E., Lyddon, S., and Holmes, C. (2019). Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1952-1962. PMLR.

Student Confirmation

Student Name:	Chung Hang Edwin Fong		
Contribution to the Paper	○ Lead of project, with supervisor as advisor. ○ Formulated the main idea of the paper. ○ Carried out the literature review. ○ Developed methodology and theoretical results. ○ Implemented the algorithms/experiments in code and interpreted results. ○ Undertook manuscript writing (edited by co-authors).		
Signature 		Date	2nd December, 2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Chris Holmes		
Supervisor comments ○ The manuscript was largely developed by the candidate with my guidance. ○ I verify that the above summary of the candidate's contribution is accurate.		
Signature 	Date	3rd December 2021

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 4

On the marginal likelihood and cross-validation

4.1 Preamble

This chapter deviates from the previous 2 chapters and focuses on Bayesian model evaluation within the $\mathcal{M}$-open setting. Once again, prediction and coherence play a crucial role in this work. The first section investigates coherent model evaluation under the general Bayesian framework of Bissiri et al. (2016) and the second section connects the marginal likelihood with cross-validation.

The content of this chapter is a self-contained manuscript with its supplementary material. Details of the publication are given below:

> Fong, E., and Holmes, C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, 107(2), 489-496.

A statement of authorship is provided at the end of this chapter.

4.2 Abstract

In Bayesian statistics, the marginal likelihood, also known as the evidence, is used to evaluate model fit as it quantifies the joint probability of the data under the prior. In contrast, non-Bayesian models are typically compared using cross-validation on held-out data, either through k-fold partitioning or leave-p-out subsampling. We show that the marginal likelihood is formally equivalent to exhaustive leave-p-out cross-validation averaged over all values of p and all held-out test sets when using the log posterior predictive probability as the scoring rule. Moreover, the log posterior predictive is the only coherent scoring rule under data exchangeability. This offers new insight into the marginal likelihood and cross-validation and highlights the potential sensitivity of the marginal likelihood to the choice of the prior. We suggest an alternative approach using cumulative cross-validation following a preparatory training phase. Our work has connections to prequential analysis and intrinsic Bayes factors but is motivated through a different course.

4.3 Introduction

Probabilistic model evaluation and selection is an important task in statistics and machine learning, particularly when multiple models are under initial consideration. In the non-Bayesian literature, models are typically compared using out-of-sample performance criteria such as cross-validation (Geisser and Eddy, 1979; Shao, 1993; Vehtari and Lampinen, 2002), or predictive information (Watanabe, 2010). Computing the leave-p-out cross-validation score requires n-choose-p test set evaluations for n data points, which in most cases is computationally unviable and hence approximations such as k-fold cross-validation are often used instead (Geisser, 1975). A survey is provided by Arlot and Celisse (2010), and a Bayesian perspective on cross-validation by Vehtari and Ojanen (2012); Gelman et al. (2014).

In Bayesian statistics, the marginal likelihood or model evidence is the natural mea-

sure of model fit. For a model $\mathcal{M}$ with sampling density $\{f_\theta(y) : \theta \in \Omega\}$ parameterized by θ, a prior density $\pi(\theta)$, and observations $y_{1:n} \in \mathcal{Y}^n$, the marginal likelihood or the prior predictive density is defined as

$$p_{\mathcal{M}}(y_{1:n}) = \int f_\theta(y_{1:n})\, \pi(\theta)\, d\theta. \tag{4.1}$$

The marginal likelihood can be used to calculate the posterior probability of the model given the data, $p(\mathcal{M} \mid y_{1:n}) \propto p_{\mathcal{M}}(y_{1:n})\, p(\mathcal{M})$, as it is the probability of the data being generated under the prior when the model is correctly specified (Robert, 2007, Chapter 7). The ratio of marginal likelihoods between models is known as the Bayes factor that quantifies the prior to posterior odds on observing the data. The marginal likelihood can be difficult to compute if the likelihood is peaked with respect to the prior, although Monte Carlo solutions exist; see Robert and Wraith (2009) for a survey. Under vague priors, the marginal likelihood may also be highly sensitive to the prior dispersion even if the posterior is not; a well known example is Lindley's paradox (Lindley, 1957; O'Hagan and Forster, 2004; Robert, 2014). As a result, its approximations such as the Bayesian information criterion (Schwarz, 1978) or the deviance information criterion (Spiegelhalter et al., 2002) are widely used, see also Gelman et al. (2014).

For our work, it is useful to note from the property of probability distributions that the log marginal likelihood can be written as the sum of log conditionals,

$$\log p_{\mathcal{M}}(y_{1:n}) = \sum_{i=1}^{n} \log p_{\mathcal{M}}(y_i \mid y_{1:i-1}) \tag{4.2}$$

where $p_{\mathcal{M}}(y_i \mid y_{1:i-1}) = \int f_\theta(y_i)\, \pi(\theta \mid y_{1:i-1})\, d\theta$ is the posterior predictive for $i > 1$, $p_{\mathcal{M}}(y_1 \mid y_{1:0}) = \int f_\theta(y_1)\, \pi(\theta)\, d\theta$, and this representation is true for any permutation of the data indices.

While Bayesian inference formally assumes that the model space captures the truth, in the model misspecified or so called $\mathcal{M}$-open scenario (Bernardo and Smith, 2009, Chapter 6) the log marginal likelihood can be simply interpreted as a predictive

sequential, or prequential (Dawid, 1984), scoring rule of the form $S(y_{1:n}) = \sum_i s(y_i \mid y_{1:i-1})$ with score function $s(y_i \mid y_{1:i-1}) = \log p_{\mathcal{M}}(y_i \mid y_{1:i-1})$. This interpretation of the log marginal likelihood as a predictive score (Kass and Raftery, 1995; Gneiting and Raftery, 2007; Bernardo and Smith, 2009, Chapter 6) has resulted in alternative scoring functions for Bayesian model selection (Dawid and Musio, 2014, 2015; Watson and Holmes, 2016; Shao et al., 2019), and provides insight into the relationship between the marginal likelihood and posterior predictive methods (Vehtari and Ojanen, 2012). Key et al. (1999) considered cross-validation from an $\mathcal{M}$-open perspective and introduced a mixture utility for model selection that trades off fidelity to data with predictive power.

4.4 Uniqueness of the marginal likelihood under coherent scoring

To begin, we prove that under an assumption of data exchangeability, the log posterior predictive is the only prequential scoring rule that guarantees coherent model evaluation. The coherence property under exchangeability, where the indices of the data points carry no information, refers to the principle that identical models on seeing the same data should be scored equally irrespective of data ordering.

In demonstrating the uniqueness of the log posterior predictive, it is useful to introduce the notion of a general Bayesian model (Bissiri et al., 2016), which is a framework for Bayesian updating without the requirement of a true model. Define a parameter of interest by

$$\theta_0 = \arg\min_{\theta} \int \ell(\theta, y)\, dF_0(y) \tag{4.3}$$

where $F_0(y)$ is the unknown true sampling distribution giving rise to the data, and $\ell : \Omega \times \mathcal{Y} \to \mathbb{R}$ is a loss function linking an observation y to the parameter θ. Bissiri et al. (2016) argue that after observing $y_{1:n}$, a coherent update of beliefs about θ_0 from

a prior density $\pi_{\mathrm{G}}(\theta)$ to the posterior density $\pi_{\mathrm{G}}(\theta \mid y_{1:n})$ exists and must take on the form

$$\pi_{\mathrm{G}}(\theta \mid y_{1:n}) \propto \exp\{-w\ell(\theta, y_{1:n})\}\, \pi_{\mathrm{G}}(\theta) \tag{4.4}$$

where $\ell(\theta, y_{1:n}) = \sum_i \ell(\theta, y_i)$ is an additive loss function and $w > 0$ is a loss scale parameter; see Holmes and Walker (2017); Lyddon et al. (2019) on the selection of w. For $w = 1$ and $\ell(\theta, y) = -\log f_\theta(y)$, we obtain traditional Bayesian updating without assuming the model $f_\theta(y)$ is true for some value of θ. From (4.3), $\mathcal{M}$-open Bayesian inference is simply targeting the value of θ that minimizes the Kullback-Leibler divergence between $dF_0(y)$ and $f_\theta(y)$. The form (4.4) is uniquely implied by the assumptions in Theorem 1 of Bissiri et al. (2016), and we now focus on the coherence property of the update rule. An update function $\psi\{\ell(\theta, y), \pi_{\mathrm{G}}(\theta)\} = \pi_{\mathrm{G}}(\theta \mid y)$ is coherent if, for some inputs $y_{1:2}$, it satisfies

$$\psi[\ell(\theta, y_2), \psi\{\ell(\theta, y_1), \pi_{\mathrm{G}}(\theta)\}] = \psi\{\ell(\theta, y_1) + \ell(\theta, y_2), \pi_{\mathrm{G}}(\theta)\}.$$

This coherence condition is natural under an assumption of exchangeability as we expect posterior inferences about θ_0 to be unchanged whether we observe $y_{1:2}$ in any order or all at once, as it is in traditional Bayesian updating.

We now extend this coherence condition to general Bayesian model choice, where the goal is to evaluate the fit of the observed data under the general Bayesian model class $\mathcal{M}_{\mathrm{G}} = \{\ell(\theta, y) : \theta \in \Omega\}$ with a prior density $\pi_{\mathrm{G}}(\theta)$. We treat w as a parameter outside of the model specification, as there are principled methods to select it from the model, prior and data. We define the log posterior predictive score as

$$s_{\mathrm{G}}(\tilde{y} \mid y_{1:n}) = \log \int g\{\ell(\theta, \tilde{y})\}\, \pi_{\mathrm{G}}(\theta \mid y_{1:n})\, d\theta$$

where $g : \mathbb{R} \to (0, \infty)$ is a continuous monotonically decreasing scoring function that transforms $\ell(\theta, y)$ into a predictive score for a test point $\tilde{y}$. We define the cumulative

prequential log score as

$$S_{\mathrm{G}}(y_{1:n}) = \sum_{i=1}^{n} s_{\mathrm{G}}(y_i \mid y_{1:i-1})$$

where $s_{\mathrm{G}}(y_1 \mid y_{1:0}) = \log \int g\{\ell(\theta, y_1)\}\, \pi_{\mathrm{G}}(\theta)\, d\theta$. The cumulative prequential log score sums the log posterior predictive score of each consecutive data point in a prequential manner, where a large score indicates that the model is predicting well. It seems that there are many choices for g, but we will see that all but one violate coherency, as defined below.

Definition 4.1. *The model scoring function $g(l)$ is coherent if it satisfies*

$$\sum_{i=1}^{n} s_{\mathrm{G}}(y_i \mid y_{1:i-1}) = \log \int g\{\ell(\theta, y_{1:n})\}\, \pi_{\mathrm{G}}(\theta)\, d\theta \tag{4.5}$$

for all Ω, $\pi(\theta)$ and $n > 0$, such that $S_{\mathrm{G}}(y_{1:n})$ is invariant to the ordering or partitioning of the observations.

We now present our main result on the uniqueness of the choice of g.

Proposition 4.1. *If the model scoring function $g : \mathbb{R} \to (0, \infty)$ is continuous, monotonically decreasing and coherent, then the unique choice of scoring rule $g(l)$ is*

$$g(l) = \exp(-wl)$$

where w is the loss-scale in the general Bayesian posterior.

Proof. The proof is given in the Appendix 4.8.1. □

This holds irrespective of whether the model is true or not. More importantly for us is the corollary below.

Corollary 4.1. *The marginal likelihood is the unique coherent marginal score for Bayesian inference.*

Proof. Let $w = 1$ and $\ell(\theta, y) = -\log f_\theta(y)$, and hence $g\{\ell(\theta, y)\} = f_\theta(y)$. □

The marginal likelihood arises naturally as the unique prequential scoring rule under coherent belief updating in the Bayesian framework. The coherence of the marginal likelihood implies an invariance to the permutation of the observations $y_{1:n}$ under exchangeability, including independent and identically distributed data, a property that is not shared by other prequential scoring rules, such as Dawid and Musio (2014); Grünwald and van Ommen (2017); Shao et al. (2019).

4.5 The marginal likelihood and cross-validation

4.5.1 Equivalence of the marginal likelihood and cumulative cross-validation

The leave-p-out cross-validation score is defined as

$$S_{\mathrm{CV}}(y_{1:n}; p) = \frac{1}{\binom{n}{p}} \sum_{t=1}^{\binom{n}{p}} \frac{1}{p} \sum_{j=1}^{p} s\left(\tilde{y}_j^{(t)} \mid y_{1:n-p}^{(t)}\right) \tag{4.6}$$

where $\tilde{y}_{1:p}^{(t)}$ denotes the tth of n-choose-p possible held-out test sets, with $y_{1:n-p}^{(t)}$ the corresponding training set, such that $y_{1:n} = \left\{\tilde{y}^{(t)}, y^{(t)}\right\}$, and S_{CV} records the average predictive score per datum. Although leave-one-out cross-validation is a popular choice, it was shown in Shao (1993) that it is asymptotically inconsistent for a linear model selection problem, and requires $(p/n) \to 1$ as $n \to \infty$ for consistency. We will not go into further detail here but instead refer the reader to Arlot and Celisse (2010). Selecting a larger p has the interpretation of penalizing complexity (Vehtari and Ojanen, 2012), as complex models will tend to over-fit to a small training set. However, the number of test set evaluations grows rapidly with p and hence k-fold cross-validation is often adopted for computational convenience.

From a Bayesian perspective it is natural to consider the log posterior predictive as the scoring function, $s(\tilde{y} \mid y) = \log \int f_\theta(\tilde{y})\, \pi(\theta \mid y)\, d\theta$, particularly as we have now

shown that it is the only coherent scoring mechanism, which leads us to the following result.

Proposition 4.2. *The Bayesian marginal likelihood is equivalent to the cumulative leave-p-out cross-validation score using the log posterior predictive as the scoring rule, such that*

$$\log p_{\mathcal{M}}(y_{1:n}) = \sum_{p=1}^{n} S_{\mathrm{CV}}(y_{1:n}; p) \tag{4.7}$$

with $s(\tilde{y}_j \mid y_{1:n-p}) = \log p_{\mathcal{M}}(\tilde{y}_j \mid y_{1:n-p}) = \log \int f_\theta(\tilde{y}_j)\, \pi(\theta \mid y_{1:n-p})\, d\theta$.

Proof. This follows from the invariance of the marginal likelihood under arbitrary permutation of the sequence $y_{1:n}$ in (4.2). We provide a proof and an alternative proof by induction in Appendices 4.8.2, 4.8.3. □

This connection has previously been discussed in Gneiting and Raftery (2007). The Bayesian marginal likelihood is simply n times the average leave-p-out cross-validation score, $n \times (1/n) \sum_{p=1}^{n} S_{\mathrm{CV}}(y_{1:n}; p)$, where the scaling by n is due to (4.6) being a per datum score. Bayesian models are evaluated through out-of-sample predictions on all $(2^n - 1)$ possible held-out test sets whereas cross-validation with fixed p only captures a snapshot of model performance. Evaluating the predictive performance on $(2^n - 1)$ test sets would appear intractable for most applications, but we see through (4.7) and (4.1) that it is computable as a single integral.

4.5.2 Sensitivity to the prior and preparatory training

The representation of the marginal likelihood as a cumulative cross-validation score (4.7) provides insight into the sensitivity to the prior. The last term in the right hand side of (4.7) involves no training data, $S_{\mathrm{CV}}(y_{1:n}; n) = (1/n) \sum_{i=1}^{n} \log \int f_\theta(y_i)\, \pi(\theta)\, d\theta$, which scores the model entirely on how well the analyst is able to specify the prior. In many situations, the analyst may not want this term to contribute to model evaluation. Moreover, there is tension between any desire to specify vague priors to safeguard

their influence and the fact that diffuse priors can lead to an arbitrarily large and negative model score for real valued parameters from (4.7). It may seem inappropriate to penalize a model based on the subjective ability to specify the prior, or to compare models using a score that includes contributions from predictions made using only a handful of training points even with informative priors. For example, we see that 10% of terms contributing to the marginal likelihood come from out-of-sample predictions using, on average, less than 5% of available training data. This is related to the start-up problem in prequential analysis (Dawid, 1992a).

A natural and obvious solution is to begin evaluating the model performance after a preparatory phase, for example using 10% or 50% of the data as preparatory training prior to testing. This leads to a Bayesian cumulative leave-P-out cross-validation score defined as

$$S_{\mathrm{CCV}}(y_{1:n}; P) = \sum_{p=1}^{P} S_{\mathrm{CV}}(y_{1:n}; p) \tag{4.8}$$

with a preparatory cross-validation score $S_{\mathrm{PCV}}(y_{1:n}; P) = \sum_{p=P+1}^{n} S_{\mathrm{CV}}(y_{1:n}; p)$, for $1 \leq P < n$. We suggest setting P to leave out $0.9n$, $0.5n$ or $\max(0.9n, n - 10d)$, where d is the total number of model parameters, as reasonable default choices, but clearly this is situation specific. One may be interested in reporting both S_{CCV} and S_{PCV}, as the latter can be regarded as an evaluation of the prior, but we suggest that only S_{CCV} is used for model evaluation from the arguments above. Although full coherency is now lost, we still have coherency conditioned on a preparatory training set, where permutation of the data within the training and test sets does not affect the score, and so we can write (4.8) as

$$S_{\mathrm{CCV}}(y_{1:n}; P) = \frac{1}{\binom{n}{P}} \sum_{t=1}^{\binom{n}{P}} \log p_{\mathcal{M}} \left(\tilde{y}_{1:P}^{(t)} \mid y_{1:n-P}^{(t)} \right). \tag{4.9}$$

This equivalence is derived in Appendix 4.8.4 in a similar fashion to Proposition 4.2. This has precisely the form of the the log geometric intrinsic Bayes factor of Berger and Pericchi (1996) but motivated by a different route. The intrinsic Bayes factor was developed in an objective Bayesian setting (Berger and Pericchi, 2001), where improper priors cause indeterminacies in the evaluation of the marginal likelihood. The intrinsic Bayes factor remedies this with a partition of the data into $y_{1:l}, y_{l+1:n}$, where $y_{1:l}$ is the minimum training sample used to convert an improper prior $\pi(\theta)$ into a proper prior $\pi(\theta \mid y_{1:l})$. In contrast, we set $n - P$ to provide preparatory training and $\pi(\theta)$ can be subjective. Moreover, in modern applications we often have $d \gg n$ where intrinsic Bayes factors cannot be applied in their original form.

We can approximate (4.9) through Monte Carlo where the training data sets $y^{(t)}_{1:n-P}$ are drawn uniformly at random, and for non-conjugate models the inner term must also be estimated, for example through

$$\hat{S}_{\text{CCV}}(y_{1:n}; P) = \frac{1}{T}\sum_{t=1}^{T} \log\left\{\frac{1}{B}\sum_{b=1}^{B} f_{\theta_b^{(t)}}\left(\tilde{y}^{(t)}_{1:P}\right)\right\} \tag{4.10}$$

where samples $\theta_b^{(t)} \sim \pi\left(\theta \mid y^{(t)}_{1:n-P}\right)$ are obtained via T Markov chain Monte Carlo samplers. If we assume that the number of samples B per chain is sufficiently large, then the variance of the estimate $\hat{S}_{\text{CCV}}$ is approximately of the form τ^2/T. However, fitting T models may be costly, but we can run the chains in parallel. To avoid the need for T Markov chain Monte Carlo chains in (4.10), we can instead take advantage of the fact that the partial posteriors for different training sets will be similar, and utilize importance sampling (Bhattacharya and Haslett, 2007; Vehtari et al., 2017) or sequential Monte Carlo (Bornn et al., 2010) to estimate the posterior predictives for computational savings. We provide further details on efficient computation of (4.10) in Appendix 4.8.5.

4.6 Illustration for the normal linear model

We illustrate the use of Bayesian cumulative cross-validation in a polynomial regression example, where the rth polynomial model is defined as

$$f_\theta(y \mid x, r) = \mathcal{N}\{y; \theta^{\mathrm{T}}\phi_r(x), \sigma^2\}, \quad \phi_r(x) = \begin{bmatrix} 1 & x & \dots & x^{r-1} & x^r \end{bmatrix}^{\mathrm{T}}.$$

We observe the data $\{y_{1:n}, x_{1:n}\}$, and we place a fixed vague prior on the intercept term, $\theta_0 \sim \mathcal{N}(\theta_0; 0, 100^2)$, and $\theta_d \sim \mathcal{N}(\theta_d; 0, s^2)$ for $d \in \{1, \dots, r\}$ on the remaining coefficients. In our example, we have $n = 100$ and the true model is $r = 1$, $\theta = \begin{bmatrix} 1 & 0.5 \end{bmatrix}^{\mathrm{T}}$ with known $\sigma^2 = 1$. For our prior, we vary the value of $s^2 \in \{10^{-1}, 10^0, 10^4\}$ to investigate the impact of the prior tails. For each prior setting, we calculate $\log p_{\mathcal{M}}(y_{1:n})$ and $S_{\mathrm{CCV}}(y_{1:n}; P)$ for models $r \in \{0, 1, 2\}$. In this example, $\log p_{\mathcal{M}}(y_{1:n})$ is tractable, whereas S_{CCV} requires a Monte Carlo average over tractable log posterior predictives. We report the mean over 10 runs of estimating S_{CCV} with $T = 10^6$ random training/test splits. We calculate the Monte Carlo standard error over the 10 runs and report the maximum for each setting of P.

The results are shown in Table 4.1, where $\hat{S}_{\mathrm{CCV}}$ is normalized to be on the same scale as $\log p_r(y_{1:n})$. Under the strong prior $s^2 = 10^{-1}$ and the moderate prior $s^2 = 10^0$, the marginal likelihood correctly identifies the true model, but when we increase s^2 to 10^4 it heavily over-penalizes the more complex models and prefers $r = 0$. In fact, the magnitude of the marginal likelihood and the discrepancy just described can be made arbitrarily large by simply increasing s^2, which should be guarded against when a modeller has weak prior beliefs. This issue is not observed with $\hat{S}_{\mathrm{CCV}}$ for the values of P we consider. The vague prior does not impede the ability of $\hat{S}_{\mathrm{CCV}}$ to correctly identify the true model $r = 1$ and the scores are stable within each column of P.

In Appendix 4.8.6, we present graphical tools for exploring the cumulative cross-validation and the effect of the choice of P on S_{CCV}. We provide an additional example using probit regression on the Pima Indian data set in Appendix 4.8.7.

Table 4.1: Log marginal likelihoods and cumulative cross-validation scores for normal linear model

s^2	Model r	$\log p_r(y_{1:n})$	$\hat{S}_{\mathrm{CCV}}(y_{1:n}; P) \times n/P$ $P = 0.9n$	$P = 0.5n$	$P = 0.1n$
10^{-1}	0	-158.82	-153.80	-153.21	-153.06
	1	-155.57	-150.39	-149.55	-149.27
	2	-156.12	-150.94	-149.81	-149.38
10^{0}	0	-158.82	-153.80	-153.21	-153.06
	1	-156.26	-150.77	-149.66	-149.34
	2	-157.80	-151.90	-150.04	-149.50
10^{4}	0	-158.82	-153.80	-153.21	-153.06
	1	-160.81	-150.91	-149.68	-149.35
	2	-166.93	-152.30	-150.08	-149.53
Maximum standard error			0.001	0.003	0.007

4.7 Discussion

We have shown that for coherence, the unique scoring rule for Bayesian model evaluation in either $\mathcal{M}$-open or $\mathcal{M}$-closed is provided by the log posterior predictive probability, and that the marginal likelihood is equivalent to a cumulative cross-validation score over all training-test data partitions. The coherence flows from the fact that the scoring rule and the Bayesian update both use the same information, namely the likelihood function, which is appropriate as the alternative would be to learn and score under different criteria. If we are interested in an alternative loss function to the log-likelihood, we advocate a general Bayesian update (Bissiri et al., 2016; Lyddon et al., 2019) that targets the parameters minimising the expected loss, with models evaluated using the corresponding coherent cumulative cross-validation score.

Acknowledgement

The authors thank Lucian Chan, George Nicholson, the editor, an associate editor and two referees for their helpful comments. Fong was funded by The Alan Turing Institute. Holmes was supported by The Alan Turing Institute, the Health Data Research, U.K.,

the Li Ka Shing Foundation, the Medical Research Council, and the U.K. Engineering and Physical Sciences Research Council.

4.8 Appendix

4.8.1 Proof of Proposition 4.1

Proof. We look at the case where $\Omega = \{0, 1\}$, so the prior $\pi_{\mathrm{G}}(\theta)$ is parametrized by $p \in [0, 1]$ with $\pi_{\mathrm{G}}(\theta = 0) = p$. We let $n = 2$, denoting the observables as y_1, y_2. We further write $\ell(0, y_1) = l_0$ and $\ell(1, y_1) = l_1$, and likewise $\ell(0, y_2) = h_0$ and $\ell(1, y_2) = h_1$. We write p_1 as the updated $\pi_{\mathrm{G}}(\theta = 0 \mid y_1)$ obtained from the general Bayesian update (4.4). The function $g(l)$ must then satisfy

$$\begin{aligned} \{g(l_0)\, p + g(l_1)\,(1-p)\}\,\{g(h_0)\, p_1 + g(h_1)\,(1-p_1)\} \\ = \{g(l_0 + h_0)\, p + g(l_1 + h_1)\,(1-p)\} \end{aligned} \tag{4.11}$$

for all $0 \leq p \leq 1$ and for all $l_0, l_1, h_0, h_1 \in \mathbb{R}$. If we let $p = 1$, then $p_1 = 1$, so this simplifies to

$$g(l_0)\, g(h_0) = g(l_0 + h_0).$$

As g is continuous and monotonically decreasing, to satisfy (4.5) it must take on the form

$$g(l) = \exp(-\lambda l) \tag{4.12}$$

for $\lambda \geq 0$. We now explicitly write out the form of p_1

$$p_1 = \frac{\exp(-w l_0)\, p}{\exp(-w l_0)\, p + \exp(-w l_1)\,(1-p)} = \frac{\exp(-w l_0)\, p}{Z_1}. \tag{4.13}$$

If we plug (4.12), (4.13) into (4.11), we obtain

$$\begin{aligned}
&\{\exp(-\lambda l_0)\, p + \exp(-\lambda l_1)\,(1-p)\} \times \\
&\{\exp(-\lambda h_0)\, \exp(-w l_0) p + \exp(-\lambda h_1)\, \exp(-w l_1)\,(1-p)\} \\
&= Z_1 \left[\exp\{-\lambda(l_0+h_0)\}\, p + \exp\{-\lambda(l_1+h_1)\}\,(1-p)\right].
\end{aligned}$$

Expanding, cancelling terms, and simplifying we obtain

$$\begin{aligned}
&\exp(-\lambda l_1 - w l_0)\{\exp(-\lambda h_0) - \exp(-\lambda h_1)\} \\
&= \exp(-\lambda l_0 - w l_1)\{\exp(-\lambda h_0) - \exp(-\lambda h_1)\}
\end{aligned}$$

and so we must have $\lambda = 0$ or $\lambda = w$, where only the latter solution is non-trivial. We have thus shown that for $n = 2, |\Omega| = 2$, the unique non-trivial solution to (4.5) is

$$g(l) = \exp(-wl). \tag{4.14}$$

The remainder of the proof involves showing that this choice of g satisfies (4.5) for all $n > 0$ and all Ω and $\pi(\theta)$. Subbing (4.14) into (4.5), we obtain

$$\begin{aligned}
\prod_{i=1}^{n} \exp\left\{s_{\mathrm{G}}(y_i \mid y_{1:i-1})\right\} &= \prod_{i=1}^{n} \int \exp\{-w\ell(\theta, y_i)\}\, \frac{\exp\left\{-w\ell(\theta, y_{1:i-1})\right\}\, \pi_{\mathrm{G}}(\theta)\, d\theta}{\int \exp\left\{-w\ell(\theta', y_{1:i-1})\right\}\, \pi_{\mathrm{G}}(\theta')\, d\theta'} \\
&= \prod_{i=1}^{n} \frac{\int \exp\left\{-w\ell(\theta, y_{1:i})\right\}\, \pi_{\mathrm{G}}(\theta)\, d\theta}{\int \exp\left\{-w\ell(\theta', y_{1:i-1})\right\}\, \pi_{\mathrm{G}}(\theta')\, d\theta'} \\
&= \int \exp\left\{-w\ell(\theta, y_{1:n})\right\}\, \pi_{\mathrm{G}}(\theta)\, d\theta
\end{aligned}$$

where for convenience we write $\ell(\theta, y_{1:0}) = 0$. □

4.8.2 Proof of Proposition 4.2

Proof. Consider the $(n! \times n)$ matrix Z with elements $(Z)_{ti} = \log p_{\mathcal{M}}(y_i^{(t)} \mid y_{1:i-1}^{(t)})$, such that the tth row of Z records the prequential sequence of log posterior predictives under the tth of $n!$ permutations of $y_{1:n}$. By the property of conditional probabilities,

we have that the row sums of Z are equal, $\sum_i (Z)_{ti} = \sum_i (Z)_{t'i}$ for all t, t', and hence

$$\log p_{\mathcal{M}}(y_{1:n}) = \frac{1}{n!} \sum_{t=1}^{n!} \sum_{i=1}^{n} (Z)_{ti} = \sum_{i=1}^{n} \frac{1}{n!} \sum_{t=1}^{n!} (Z)_{ti}.$$

Within each column of Z, the values $(Z)_{ti}$ are invariant to the permutation of $y_{1:i-1}$ in the preceding $i-1$ columns under exchangeability. There are thus n-choose-$(i-1)$ distinct training sets and $n-i+1$ choices for y_i given the training set. For each column $i \in \{1, \ldots, n\}$, we can then write

$$\begin{aligned} \frac{1}{n!} \sum_{t=1}^{n!} (Z)_{ti} &= \frac{1}{\binom{n}{i-1}} \sum_{t=1}^{\binom{n}{i-1}} \frac{1}{n-i+1} \sum_{j=1}^{n-i+1} s\left(\tilde{y}_j^{(t)} \mid y_{1:i-1}^{(t)}\right) \\ &= S_{\mathrm{CV}}(y_{1:n}; n-i+1) \end{aligned}$$

where $s\left(\tilde{y}_j^{(t)} \mid y_{1:i-1}^{(t)}\right) = \log p_{\mathcal{M}}\left(\tilde{y}_j^{(t)} \mid y_{1:i-1}^{(t)}\right)$. We have the result for $p = n-i+1$. □

4.8.3 Alternative proof of Proposition 4.2

To prove Proposition 4.2, we first begin by showing the following proposition.

Proposition 4.3. *For a preparatory cross-validation score, $S_{\mathrm{PCV}}(y_{1:n}; P)$, defined as the sum of cross-validation terms from leave-$(P+1)$-out to leave-n-out,*

$$S_{\mathrm{PCV}}(y_{1:n}; P) = \sum_{p=P+1}^{n} S_{\mathrm{CV}}(y_{1:n}; p),$$

we have the following equivalence relationship

$$S_{\mathrm{PCV}}(y_{1:n}; P) = \frac{1}{\binom{n}{P}} \sum_{t=1}^{\binom{n}{P}} \log p_{\mathcal{M}}\left(y_{1:n-P}^{(t)}\right) \tag{4.15}$$

which states that S_{PCV} is the average log marginal likelihood over all choices of the training set.

Proof. To show this, we use a proof by induction. We see that (4.15) is trivially true for $P = n-1$, as this is simply $S_{\text{CV}}(y_{1:n}; n)$. Assuming (4.15) holds for some $1 \leq P \leq n-1$, we have

$$\begin{aligned} S_{\text{PCV}}(y_{1:n}; P-1) &= S_{\text{PCV}}(y_{1:n}; P) + S_{\text{CV}}(y_{1:n}; P) \\ &= \frac{1}{\binom{n}{P}} \sum_{t=1}^{\binom{n}{P}} \log p_{\mathcal{M}}\left(y_{1:n-P}^{(t)}\right) + \frac{1}{\binom{n}{P}} \sum_{t=1}^{\binom{n}{P}} \frac{1}{P} \sum_{j=1}^{P} \log p_{\mathcal{M}}\left(\tilde{y}_j^{(t)} \mid y_{1:n-P}^{(t)}\right) \\ &= \frac{1}{P\binom{n}{P}} \sum_{t=1}^{\binom{n}{P}} \left\{ P \log p_{\mathcal{M}}\left(y_{1:n-P}^{(t)}\right) + \sum_{j=1}^{P} \log p_{\mathcal{M}}\left(\tilde{y}_j^{(t)} \mid y_{1:n-P}^{(t)}\right) \right\}. \end{aligned}$$

From the properties of conditional probability, we can write

$$S_{\text{PCV}}(y_{1:n}; P-1) = \frac{1}{P\binom{n}{P}} \sum_{t=1}^{\binom{n}{P}} \sum_{j=1}^{P} \log p_{\mathcal{M}}\left(\tilde{y}_j^{(t)}, y_{1:n-P}^{(t)}\right). \tag{4.16}$$

Again, the marginal likelihood is invariant to the permutation of the sequence under data exchangeability, so we have to consider the repetitions in the partitions $\tilde{y}_j^{(t)}, y_{1:n-P}^{(t)}$. For each of the n choose $(n-P+1)$ unordered sequences $y_{1:n-P+1}^{(t')}$, there are $(n-P+1)$ partitions into $\tilde{y}_j^{(t)}, y_{1:n-P}^{(t)}$, so there are $n-P+1$ repetitions of each unordered $y_{1:n-P+1}^{(t')}$ in (4.16). We can thus write

$$\begin{aligned} S_{\text{PCV}}(y_{1:n}; P-1) &= \frac{(n-P+1)}{P\binom{n}{P}} \sum_{t'=1}^{\binom{n}{P-1}} \log p_{\mathcal{M}}\left(y_{1:n-P+1}^{(t')}\right) \\ &= \frac{1}{\binom{n}{P-1}} \sum_{t'=1}^{\binom{n}{P-1}} \log p_{\mathcal{M}}\left(y_{1:n-P+1}^{(t')}\right) \end{aligned}$$

and by induction we have (4.15). □

Proposition 4.2 then follows trivially by setting $P = 0$ in Proposition 4.3.

4.8.4 Derivation of S_{CCV} for Bayesian models

The following corollary follows easily from Propositions 4.2 and 4.3.

Corollary 4.2. *For the cumulative cross-validation score defined as*

$$S_{\mathrm{CCV}}(y_{1:n}; P) = \sum_{p=1}^{P} S_{\mathrm{CV}}(y_{1:n}; p), \tag{4.17}$$

we have the following equivalence relationship

$$S_{\mathrm{CCV}}(y_{1:n}; P) = \frac{1}{\binom{n}{P}} \sum_{t=1}^{\binom{n}{P}} \log p_{\mathcal{M}} \left(\tilde{y}_{1:P}^{(t)} \mid y_{1:n-P}^{(t)} \right). \tag{4.18}$$

Proof. We note that $\log p_{\mathcal{M}}(y_{1:n}) = S_{\mathrm{CCV}}(y_{1:n}; P) + S_{\mathrm{PCV}}(y_{1:n}; P)$ from their definitions and Proposition 4.2. From the permutation invariance of the marginal likelihood, we can write

$$\log p_{\mathcal{M}}(y_{1:n}) = \frac{1}{\binom{n}{P}} \sum_{t=1}^{\binom{n}{P}} \log p_{\mathcal{M}} \left(\tilde{y}_{1:P}^{(t)}, y_{1:n-P}^{(t)} \right). \tag{4.19}$$

By subtracting (4.15) in Proposition 4.3 from (4.19) and regarding each term in the summation, we have

$$\begin{aligned} S_{\mathrm{CCV}}(y_{1:n}; P) &= \frac{1}{\binom{n}{P}} \sum_{t=1}^{\binom{n}{P}} \left\{ \log p_{\mathcal{M}} \left(\tilde{y}_{1:P}^{(t)}, y_{1:n-P}^{(t)} \right) - \log p_{\mathcal{M}} \left(y_{1:n-P}^{(t)} \right) \right\} \\ &= \frac{1}{\binom{n}{P}} \sum_{t=1}^{\binom{n}{P}} \log p_{\mathcal{M}} \left(\tilde{y}_{1:P}^{(t)} \mid y_{1:n-P}^{(t)} \right) \end{aligned}$$

□

4.8.5 Computing S_{CCV}

We note that $\hat{S}_{\mathrm{CCV}}$ in (4.10) is a biased estimate, and Rischard et al. (2018) provides unbiased estimators of $\log p_{\mathcal{M}}(\tilde{y}_{1:P} \mid y_{1:n-P})$ directly through unbiased Markov chain Monte Carlo and path sampling methods.

The arithmetic averaging over training/test splits $\hat{S}_{\text{CCV}}$ may also be inherently unstable, as demonstrated by the following example. Suppose that y is a binary random variable which takes on either 0 or 1 with equal probability, and we are attempting to estimate $S_{\text{CCV}}(y_{1:n}; n/2)$. For large n, it is likely that approximately half of the values in $y_{1:n}$ are equal to 0 and the other half to 1. There will thus exist a permutation of the sequence $y_{1:n}$ such that almost all the first $n/2$ values are equal to 0, with the remaining almost all equal to 1. The model will then be certain that $y = 0$ after observing the training set, and score the remaining $n/2$ points very poorly, giving a large negative log posterior predictive. This suggests that an arithmetic average may be unstable; the median or robust trimmed mean over permutations may be stabler alternatives.

The form in (4.18) relies on the conditional coherency of Bayesian updating and scoring. Without this, S_{CCV} still exists as defined in (4.17), and can be directly estimated for example through

$$\hat{S}_{\text{CCV}}(y_{1:n}; P) = \frac{P}{T} \sum_{t=1}^{T} \frac{1}{p^{(t)}} \sum_{j=1}^{p^{(t)}} s\left(\tilde{y}_j^{(t)} \mid y_{1:n-p^{(t)}}^{(t)}\right)$$

where $p^{(t)} \sim \mathcal{U}\{1, P\}$ and the training set $y_{1:n-p^{(t)}}^{(t)}$ is sampled uniformly at random conditioned on $p^{(t)}$. This facilities alternative choices for the belief updating model and $s(\tilde{y} \mid y)$.

4.8.6 Visualization of cumulative cross-validation

A visualization of the effects of the training/preparatory data size is shown in Figure 4.1 for $s^2 = 1$ in the polynomial regression example. We omit $S_{\text{CV}}(y_{1:n}; n)$ and $S_{\text{CCV}}(y_{1:n}; n)$ for clarity of the plot, as both are significantly more negative than the other values. On the left we see that the individual cross-validation term $S_{\text{CV}}(y_{1:n}; p)$ prefers the simplest $r = 0$ model when the training set is very small as over-fitting is penalized, but as $n - p$ increases, the true $r = 1$ model overtakes it. The $r = 2$ model eventually overtakes the $r = 0$ model too, and we see the discrepancy between $r = 2$ and $r = 1$

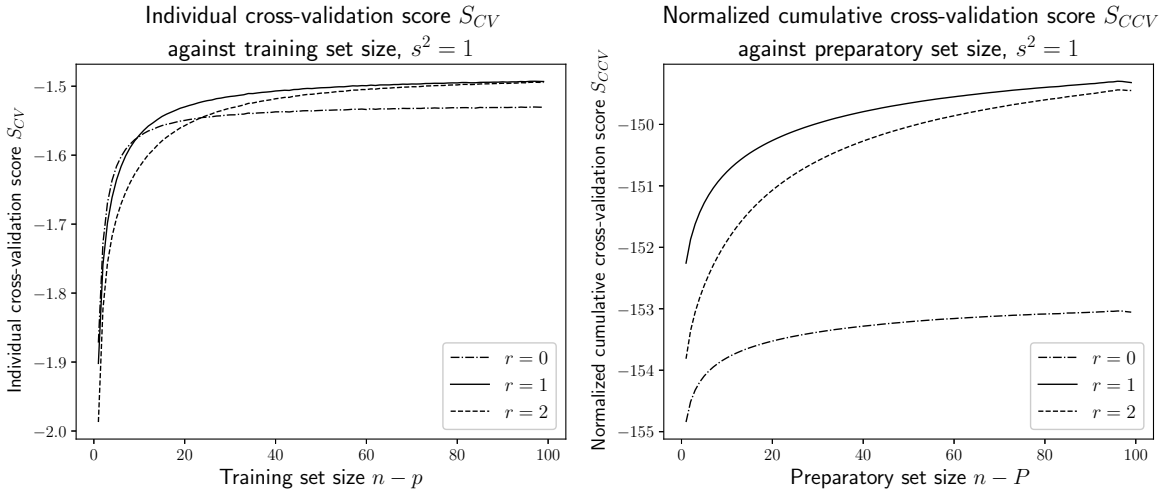


Figure 4.1: Leave-p-out cross-validation score $S_{\text{CV}}(y_{1:n}; p)$ against $n - p$ (left) and normalized cumulative cross-validation score $S_{\text{CCV}}(y_{1:n}; P) \times n/P$ against $n - P$ (right) for $s^2 = 1$ and $p, P \in \{1, \ldots, 99\}$ in the polynomial regression example; the maximum standard error is 0.0003 for S_{CV} and 0.002 for $\hat{S}_{\text{CCV}}$.

decrease as over-fitting is penalized less and less. This latter effect is demonstrative of how leave-one-out cross-validation under-penalizes complex models as argued in Shao (1993), and why a value of $P > 1$ should be preferred. On the right, we observe a similar effect for the cumulative cross-validation score S_{CCV}, but the discrepancy between $r = 2$ and $r = 1$ remains more noticeable for moderate $n - P$ as a cumulative sum of S_{CV} terms is being taken.

4.8.7 Illustration for the probit model

To demonstrate the cumulative cross-validation score in an intractable example, we carry out model selection in the Pima Indian benchmark model with a probit model. We observe binary random variables $y_{1:n}$ with associated r-dimensional covariates $x_{1:n}$, and the probit model is defined as

$$f_\theta(y \mid x) = \{\Phi\left(\theta^{\mathrm{T}} \tilde{x}\right)\}^y \left\{1 - \Phi\left(\theta^{\mathrm{T}} \tilde{x}\right)\right\}^{1-y}$$

where Φ is the standard normal cumulative distribution function and $\tilde{x} = \begin{bmatrix} 1 & x^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$. As suggested in Marin and Robert (2010), we elicit a g-prior $\pi(\theta) = \mathcal{N}\left\{\theta; 0_{r+1}, g(X^{\mathrm{T}}X)^{-1}\right\}$

where 0_{r+1} is a $r+1$ vector of 0s and X is the n by $r+1$ matrix with rows $\tilde{x}_i^{\mathrm{T}}$.

The dataset consists of $n = 332$ data points and we consider $r = 3$ covariates consisting of `glu`, `bp` and `ped`, which correspond to plasma glucose concentration from an oral glucose test, diastolic blood pressure and diabetes pedigree function respectively. We compare the full model $\mathcal{M}_0$: (`glu,bp,ped`) with $\mathcal{M}_1$: (`glu,bp`) through $\log p_{\mathcal{M}}(y_{1:n})$ and $S_{\mathrm{CCV}}(y_{1:n}; P)$ to test for significance of `ped`. We standardize all covariates to have 0 mean and variance 1. We calculate $\log p_{\mathcal{M}}(y_{1:n})$ using importance sampling with a Gaussian proposal with 10^3 samples. The proposal mean is set to the maximum likelihood estimate of θ and proposal covariance to the estimated covariance matrix of the maximum likelihood estimate as suggested in Marin and Robert (2010). For $S_{\mathrm{CCV}}(y_{1:n}; P)$, we estimate each posterior predictive in (4.10) with the same importance sampling scheme where we temper the proposal such that its covariance matrix is divided by $(n-P)/n$. We also use 10^3 proposal samples and average over $T = 10^5$ random train/test splits. We carry out 10 runs of each and report the mean and maximum standard error as before.

We see in Table 4.2 that for $g = n$, the simpler model with `ped` omitted performs worse for both scores, and there is thus strong evidence for `ped`. However, when we set $g = 10n$, we see that comparing models via the marginal likelihood suggests that `ped` is no longer significant, while the cumulative cross-validation score changes little with this increased variance of the prior. As a sanity check, we run a Gibbs sampler targeting $\pi(\theta \mid y_{1:n}, x_{1:n})$ for the two prior settings within the full model $\mathcal{M}_0$, and plot the marginal posterior of θ_{ped} in Figure 4.2. For reference, the posterior means of $\theta_{\mathrm{glu}}, \theta_{\mathrm{bp}}$ are 0.70 and 0.12 respectively. The posteriors of θ_{ped} are indistinguishable for the two prior settings, with a significant mean for θ_{ped}. This agrees well with the cumulative cross-validation score $\hat{S}_{\mathrm{CCV}}$ which is clearly robust to vague priors.

Table 4.2: Log marginal likelihoods and cumulative cross-validation score for probit model

g	Model	$\log p_{\mathcal{M}}(y_{1:n})$	$\hat{S}_{\text{CCV}}(y_{1:n}; P) \times n/P$ $P = 0.9n$
n	(glu,bp,ped)	-168.93	-165.87
	(glu,bp)	-170.00	-167.37
$10n$	(glu,bp,ped)	-173.10	-166.28
	(glu,bp)	-173.05	-167.64
Maximum standard error		0.001	0.006

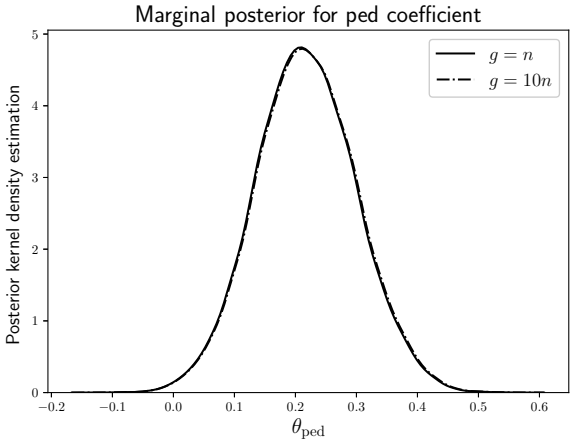


Figure 4.2: Marginal posterior density plots for θ_{ped} for different prior scalings g.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	On the marginal likelihood and cross-validation
Publication Status	**x** Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Fong, E. and Holmes, C. (2020). On the marginal likelihood and cross-validation. Biometrika, 107(2):489-496.

Student Confirmation

Student Name:	Chung Hang Edwin Fong		
Contribution to the Paper	○ Lead of project, with supervisor as advisor. ○ Formulated the main idea of the paper. ○ Carried out the literature review. ○ Developed methodology and theoretical results. ○ Implemented the algorithms/experiments in code and interpreted results. ○ Undertook manuscript writing (edited by supervisor).		
Signature 		Date	2nd December, 2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Chris Holmes		
Supervisor comments ○ The manuscript was largely developed by the candidate with my guidance. ○ I verify that the above summary of the candidate's contribution is accurate.		
Signature 	Date	3rd December 2021

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 5

Discussion

5.1 Summary

In this thesis, we have investigated the role of prediction in interpreting and generalizing Bayesian inference under model misspecification. Through Doob's theorem (Doob, 1949), we have connected the predictive imputation of missing observations directly to the posterior uncertainty in any statistic of interest. With Bayesian inference viewed in this light, the source of uncertainty is clear - in the i.i.d. case, it arises from missing observations $Y_{n+1:\infty}$ on which we assign a subjective predictive distribution. This interpretation also clearly differentiates between the frequentist focus on repeated samples of $Y_{1:n}$ versus the Bayesian focus on $Y_{n+1:\infty}$. Finally, the predictive interpretation of the Bayesian bootstrap (BB) is particularly helpful as we now see that the Dirichlet weights arise from the intuitive specification of the empirical distribution as the 1-step ahead predictive.

Like de Finetti, we argue that the goal of statistical inference is to elicit a joint distribution on all observables, but Doob's result allows us to rely on the prequential factorization instead of the likelihood-prior construction. Practically, we are now free to specify a sequence of 1-step ahead predictive distributions as our 'model' directly, subject to a martingale condition for predictive coherence. To this end, we introduce a

series of novel copula updates, which in a sense extend the BB by replacing the empirical distribution with a smooth nonparametric density. Interestingly, the method allows us to carry out nonparametric Bayesian density estimation and regression, even in multivariate settings, without any reliance on the likelihood/prior nor MCMC. However, under the appropriate settings, the empirical predictive distribution from the Pólya urn scheme of the Dirichlet process (DP) offers even more significant computational advantages. This is because we can compute the limiting empirical distribution F_∞ directly without relying on predictive resampling, which results in the familiar Dirichlet weights. We observe that Bayesian nonparametric learning with the DP results in an inference scheme that is robust to model misspecification, entirely parallelizable, and proficient at tackling multimodal posteriors.

In the last section of the thesis, we explore coherent model scoring under the general Bayesian framework of Bissiri et al. (2016) and show that the marginal likelihood uniquely satisfies a coherence property. We further show equivalence of the marginal likelihood and cross-validation, which allows us to define the cumulative cross-validation score which has connections to the intrinsic Bayes factor of Berger and Pericchi (1996).

To conclude, we find the predictive view of Bayesian inference to be a powerful tool that allows the modern Bayesian to handle model misspecification, scalability and interpretability. This perspective blurs the lines between the two cultures of Breiman (2001b), namely the two approaches usually taken in machine learning and statistics, and also brings the Bayesian closer to the frequentist. We hope that the work in this thesis can lead to fruitful avenues for future research, and we discuss some potential directions in the remainder of this chapter.

5.2 Future work

5.2.1 Beyond i.i.d. data

One major future direction of work is extending the martingale posterior framework to settings beyond i.i.d. data. In more general settings, such as in time series and hierarchical settings, the observed data may not arise from the same population as the missing observations that we require to compute the statistic of interest. Identification of the missing observations in various other settings remains to be investigated, and how one would elicit the predictive distribution $p(y_{\text{mis}} \mid y_{\text{obs}})$ or enforce predictive coherence in other settings is still an open question. We have given a brief answer in Section 2.3 of Chapter 2 for the hierarchical setting, but further investigation, including a practical demonstration, is still required. Extending the martingale posterior to right-censored observations for survival analysis has also been investigated by Fong and Lehmann (2021). In this setting, the missing data includes the remainder of the population, $Y_{n+1:\infty}$, as well as the finite number of observations that have been right-censored. A copula update with a sequential Monte Carlo scheme is then utilized to sample from the martingale posterior.

5.2.2 Properties of the martingale posterior

The copula updates introduced in Chapter 2 based on the work of Hahn et al. (2018) have interesting but non-standard properties, especially compared to traditional Bayesian methods. For example, the predictive density $p(y \mid y_{1:n})$ can be computed exactly and efficiently, but obtaining uncertainty involves the additional simulation of uniform random variables followed by sequential updating up to some truncation value N. There are still a few unanswered theoretical questions on the copula updates for martingale posteriors. Further investigations into the convergence of the truncated predictive P_N to the limiting P_∞, including extending Theorem 2.5 of Chapter 2 to the multivariate or regression setting, would be interesting. Although we have shown

frequentist consistency of the posterior mean density p_n to the true f_0 when $n \to \infty$, we have yet to investigate the convergence rate or a Bernstein-von Mises result for the induced martingale posterior. It would also be interesting to examine whether an analogue to the Kullback-Leibler property of the traditional Bayesian prior (Ghosal and van der Vaart, 2017, Definition 6.15) exists for the martingale posterior in the absence of the prior distribution. On a more practical note, a valuable future direction would be to extend the copula update by incorporating modern machine learning methods. Approximation schemes to reduce the order complexity from $\mathcal{O}(n^2)$ would also be impactful.

5.2.3 Applications of the Bayesian bootstrap

We have seen in Chapter 3 and related works that methods based on the Bayesian bootstrap (BB) have excellent performance in the big data setting and under model misspecification. A key to the BB's computational efficiency is the expediency of drawing the Dirichlet weights, which we know arises from the sequence of empirical distributions in the Pólya urn scheme. However, generalizations of the BB have previously relied on other interpretations. For example, extending the BB to settings with prior information has been investigated by Newton et al. (2020), Nie and Ročková (2020), and Pompe (2021) through the random weighting interpretation. The Bayesian nonparametric view has enabled the application of the BB to survival analysis (Arfè and Muliere, 2020), and the empirical likelihood framework has allowed the extension to the Cox model (Kim and Lee, 2003). A potentially fruitful direction of research is whether the predictive interpretation of the BB, which we find to be intuitive, allows natural extensions into more challenging data settings. Another relatively unexplored direction of work is the application of the BB in model selection and evaluation when using a parametric model to define the loss function, and also for outlier detection in robust statistics.

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, 44(2):182–198.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174.

Antoniano-Villalobos, I., Wade, S., and Walker, S. G. (2014). A Bayesian nonparametric regression model with normalized weights: a study of hippocampal atrophy in alzheimer's disease. *Journal of the American Statistical Association*, 109(506):477–490.

Arfè, A. and Muliere, P. (2020). A general Bayesian bootstrap for censored data based on the beta-stacy process. *arXiv preprint arXiv:2002.04081*.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Bardenet, R., Doucet, A., and Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(1):1515–1557.

Bedford, T. and Cooke, R. (2001). *Mathematical tools for probabilistic risk analysis*. Cambridge University Press.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122.

Berger, J. O. and Pericchi, L. R. (2001). *Objective Bayesian Methods for Model Selection: Introduction and Comparison*, volume 38 of *Lecture Notes–Monograph Series*, pages 135–207. Institute of Mathematical Statistics, Beachwood, OH.

Bernardo, J. and Smith, A. (2009). *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley.

Berti, P., Dreassi, E., Leisen, F., Rigo, P., and Pratelli, L. (2021). Bayesian predictive inference without a prior. *arXiv preprint arXiv:2104.11643*.

Berti, P., Dreassi, E., Pratelli, L., and Rigo, P. (2020). A class of models for Bayesian predictive inference. *Bernoulli*, 27(1):702–726.

Berti, P., Pratelli, L., and Rigo, P. (2004). Limit theorems for a class of identically distributed random variables. *The Annals of Probability*, 32(3):2029–2052.

Berti, P., Pratelli, L., and Rigo, P. (2006). Almost sure weak convergence of random probability measures. *Stochastics and Stochastics Reports*, 78(2):91–97.

Berti, P., Pratelli, L., and Rigo, P. (2013). Exchangeable sequences driven by an absolutely continuous random measure. *The Annals of Probability*, pages 2090–2102.

Berti, P., Regazzini, E., and Rigo, P. (1998). Well calibrated, coherent forecasting systems. *Theory of Probability & Its Applications*, 42(1):82–102.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Betrò, B. and Schoen, F. (1987). Sequential stopping rules for the multistart algorithm in global optimisation. *Mathematical Programming*, 38(3):271–286.

Bhattacharya, A., Pati, D., and Yang, Y. (2019). Bayesian fractional posteriors. *The Annals of Statistics*, 47(1):39–66.

Bhattacharya, S. and Haslett, J. (2007). Importance re-sampling MCMC for cross-validation in inverse problems. *Bayesian Analysis*, 2(2):385–407.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

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.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143.

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.

Bochkina, N. (2021). Bernstein-von mises theorem and misspecified models: a review.

Bornn, L., Doucet, A., and Gottardo, R. (2010). An efficient computational approach for prior sensitivity analysis and cross-validation. *Canadian Journal of Statistics*, 38(1):47–64.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., and Wanderman-Milne, S. (2018). JAX: composable transformations of Python+NumPy programs.

Breiman, L. (2001a). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Bühlmann, P. (2014). Discussion of big Bayes stories and BayesBag. *Statistical science*, 29(1):91–94.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514.

Campbell, T. and Broderick, T. (2019). Automated scalable Bayesian inference via Hilbert coresets. *The Journal of Machine Learning Research*, 20(1):551–588.

Cappello, L. and Walker, S. G. (2018). A Bayesian motivated Laplace inversion for multivariate probability distributions. *Methodology and Computing in Applied Probability*, 20(2):777–797.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32.

Chambers, J. M. (2018). *Graphical methods for data analysis*. CRC Press.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2):293–346.

Choudhuri, N. (1998). Bayesian bootstrap credible sets for multidimensional mean functional. *Ann. Statist.*, 26(6):2104–2127.

Chung, F. and Lu, L. (2006). Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1):79–127.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

Dawid, A. P. (1984). Present position and potential developments: Some personal views; Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, 147(2):278.

Dawid, A. P. (1985). Probability, symmetry and frequency. *The British Journal for the Philosophy of Science*, 36(2):107–128.

Dawid, A. P. (1992a). Prequential analysis, stochastic complexity and Bayesian inference. *Bayesian Statistics*, 4:109–125.

Dawid, A. P. (1992b). Prequential data analysis. *Lecture Notes-Monograph Series*, pages 113–126.

Dawid, A. P. and Musio, M. (2014). Theory and applications of proper scoring rules. *METRON*, 72(2):169–183.

Dawid, A. P. and Musio, M. (2015). Bayesian model selection based on proper scoring rules. *Bayesian Analysis*, 10(2):479–499.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68. [English translation in Studies in Subjective Probability (1980) (H. E. Kyburg and H. E. Smokler, eds.) 53-118. Krieger, Malabar, FL.].

Dheeru, D. and Karra Taniskidou, E. (2017). UCI machine learning repository.

Dick, T., Wong, E., and Dann, C. (2014). How many random restarts are enough? Technical report.

Dixit, V. and Martin, R. (2019). Permutation-based uncertainty quantification about a mixing distribution. *arXiv preprint arXiv:1906.05349*.

Doob, J. L. (1949). Application of the theory of martingales. *Actes du Colloque International Le Calcul des Probabilités et ses applications (Lyon, 28 Juin–3 Juillet 1948), Paris CNRS, 23–27*.

Doob, J. L. (1953). *Stochastic processes*, volume 101. New York Wiley.

Dryden, I., Ippoliti, L., and Romagnoli, L. (2002). Adjusted maximum likelihood and pseudo-likelihood estimation for noisy gaussian markov random fields. *Journal of Computational and Graphical Statistics*, 11(2):370–388.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Dunson, D. B. and Johndrow, J. (2020). The hastings algorithm at fifty. *Biometrika*, 107(1):1–23.

Dunson, D. B. and Park, J.-H. (2008). Kernel stick-breaking processes. *Biometrika*, 95(2):307–323.

Eaton, M. L. and Freedman, D. A. (2004). Dutch book against some objective priors. *Bernoulli*, 10(5):861–872.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, pages 1–26.

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 31(2):195–224.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230.

Filipovic, M., Dautremer, T., Comtat, C., Stute, S., and Barat, E. (2021). Reconstruction, analysis and interpretation of posterior probability distributions of pet images, using the posterior bootstrap. *Physics in Medicine & Biology*.

Fong, E. and Holmes, C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, 107(2):489–496.

Fong, E., Holmes, C., and Walker, S. G. (2021). Martingale posterior distributions. *arXiv preprint arXiv:2103.15671*.

Fong, E. and Lehmann, B. (2021). A predictive approach to Bayesian nonparametric survival analysis. Technical report.

Fong, E., Lyddon, S., and Holmes, C. (2019). Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1952–1962. PMLR.

Fortini, S., Ladelli, L., and Regazzini, E. (2000). Exchangeability, predictive distributions and parametric models. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 86–109.

Fortini, S. and Petrone, S. (2012). Predictive construction of priors in Bayesian nonparametrics. *Brazilian Journal of Probability and Statistics*, 26(4):423–449.

Fortini, S. and Petrone, S. (2014). Predictive distribution (de Finetti's view). *Wiley StatsRef: Statistics Reference Online*, pages 1–9.

Fortini, S. and Petrone, S. (2020). Quasi-Bayes properties of a procedure for sequential learning in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1087–1114.

Fushiki, T. et al. (2005). Bootstrap prediction and Bayesian prediction under misspecified models. *Bernoulli*, 11(4):747–758.

G. E. Boender, C. and H. G. Rinnooy Kan, A. (1987). Bayesian stopping rules for multistart global optimization methods. *Mathematical Programming*, 37:59–80.

Galvani, M., Bardelli, C., Figini, S., and Muliere, P. (2021). A Bayesian nonparametric learning approach to ensemble models using the proper Bayesian bootstrap. *Algorithms*, 14(1):11.

Gasparini, M. (1995). Exact multivariate Bayesian bootstrap distributions of moments. *Ann. Statist.*, 23(3):762–768.

Geisser, S. (1971). The inferential use of predictive distributions. *Foundations of Statistical Inference*, pages 456–469.

Geisser, S. (1974). A predictive approach to the random effect model. *Biometrika*, 61(1):101–107.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328.

Geisser, S. (1982). Aspects of the predictive and estimative approaches in the determination of probabilities. *Biometrics*, pages 75–85.

Geisser, S. (1983). On the prediction of observables: a selective update. Technical report, University of Minnesota.

Geisser, S. (1993). *Predictive inference*, volume 55. CRC press.

Geisser, S. and Eddy, W. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74:153–160.

Gelfand, A. E. (1996). Model determination using sampling-based methods. *Markov chain Monte Carlo in practice*, pages 145–161.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, London, third edition.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24:997–1016.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760.

Ghosal, S. (2010). *The Dirichlet process, related priors and posterior asymptotics*, page 35–79. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Ghosh, M. and Meeden, G. (1997). *Bayesian methods for finite population sampling*, volume 79. CRC Press.

Giordano, R., Broderick, T., and Jordan, M. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1441–1449, Cambridge, MA, USA. MIT Press.

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.

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

Goncharov, F., Barat, É., and Dautremer, T. (2021). Nonparametric posterior learning for emission tomography with multimodal data. *arXiv preprint arXiv:2108.00866*.

Gorham, J. and Mackey, L. (2015). Measuring sample quality with stein's method. *Advances in Neural Information Processing Systems*, 28:226–234.

Green, P. J., Łatuszyński, K., Pereyra, M., and Robert, C. P. (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25(4):835–862.

Grünwald, P. and van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103.

Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G. (2005). Result analysis of the NIPS 2003 feature selection challenge. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 545–552. MIT Press.

Hahn, P. R. (2015). Predictivist Bayes density estimation. Technical report.

Hahn, P. R., Martin, R., and Walker, S. G. (2018). On recursive Bayesian predictive distributions. *Journal of the American Statistical Association*, 113(523):1085–1093.

Hannah, L. A., Blei, D. M., and Powell, W. B. (2011). Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12(Jun):1923–1953.

Hashimoto, S. and Sugasawa, S. (2020). Robust Bayesian regression with synthetic posterior distributions. *Entropy*, 22(6):661.

Heath, D. and Sudderth, W. (1978). On finitely additive priors, coherence, and extended admissibility. *The Annals of Statistics*, pages 333–345.

Hewitt, E. and Savage, L. J. (1955). Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, 80(2):470–501.

Hjort, N. L. (1991). Bayesian and empirical Bayesian bootstrapping. *Preprint series. Statistical Research Report http://urn. nb. no/URN: NBN: no-23420*.

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

Homan, M. D. and Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Hu, X., Shonkwiler, R., and Spruill, M. C. (1994). Random restarts in global optimization. Technical report.

Huber, P. J. (2004). *Robust statistics*, volume 523. John Wiley & Sons.

Huggins, J., Campbell, T., and Broderick, T. (2016). Coresets for scalable Bayesian logistic regression. *Advances in Neural Information Processing Systems*, 29:4080–4088.

Huggins, J. H. and Miller, J. W. (2019). Robust inference and model criticism using bagged posteriors. *arXiv preprint arXiv:1912.07104*.

Huggins, J. H. and Miller, J. W. (2020). Robust and reproducible model selection using bagged posteriors. *arXiv preprint arXiv:2007.14845*.

Hyvärinen, A. and Dayan, P. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4).

Ibrahim, J. G., Chen, M.-H., et al. (2000). Power prior distributions for regression models. *Statistical Science*, 15(1):46–60.

Inglot, T. (2010). Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, 30(2):339–351.

Ishwaran, H. and Zarepour, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30:269 – 283.

Jain, P. and Kar, P. (2017). Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10:142–336.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20(1):50–67.

Jeffreys, H. (1961). *Theory of Probability*. Oxford, Oxford, England, third edition.

Jewson, J. and Rossell, D. (2021). General Bayesian loss function selection and the use of improper models. *arXiv preprint arXiv:2106.01214*.

Jewson, J., Smith, J. Q., and Holmes, C. (2018). Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442.

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

Jin, C., Zhang, Y., Balakrishnan, S., Wainwright, M. J., and Jordan, M. I. (2016). Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4123–4131, USA. Curran Associates Inc.

Jin, Z., Ying, Z., and Wei, L. (2001). A simple resampling method by perturbing the minimand. *Biometrika*, 88(2):381–390.

Joe, H. and Xu, J. J. (1996). The estimation method of inference functions for margins for multivariate models.

Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python.

Kallenberg, O. (1988). Spreading and predictable sampling in exchangeable sequences and processes. *The Annals of Probability*, pages 508–534.

Kallenberg, O. (1997). *Foundations of Modern Probability*, volume 2. Springer.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Key, J., Pericchi, L., and Smith, A. (1999). Bayesian model choice: What and why? (with discussion). *Bayesian Statistics*, 5.

Kim, Y. and Lee, J. (2003). Bayesian bootstrap for proportional hazards models. *Ann. Statist.*, 31(6):1905–1922.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Knoblauch, J., Jewson, J., and Damoulas, T. (2019). Generalized variational inference. *arXiv preprint arXiv:1904.02063*.

Knoblauch, J. and Vomfell, L. (2020). Robust Bayesian inference for discrete outcomes with the total variation distance. *arXiv preprint arXiv:2010.13456*.

Kohavi, R. (1996). Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the second internal conference on knowledge on knowledge discovery and data mining*, pages 202–207. AAAI Press.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18:430–474.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media.

Lane, D. A. and Sudderth, W. D. (1984). Coherent predictive inference. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 166–185.

Lauritzen, S. L. (1988). *Extremal Families and Systems of Sufficient Statistics*, volume 49. Springer Science & Business Media.

Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika*, 90(2):319–326.

Lazar, N. A. (2021). A review of empirical likelihood. *Annual Review of Statistics and its Application*, 8:329–344.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database.

Lee, A., Caron, F., Doucet, A., and Holmes, C. (2010). A hierarchical Bayesian framework for constructing sparsity-inducing priors. *arXiv e-prints*, page arXiv:1009.1914.

Lee, A., Caron, F., Doucet, A., and Holmes, C. (2012). Bayesian sparsity-path-analysis of genetic association signal using generalized t priors. *Statistical applications in genetics and molecular biology*, 11 2.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Lijoi, A., Prünster, I., and Walker, S. G. (2004). Extending Doob's consistency theorem to nonparametric densities. *Bernoulli*, 10(4):651–663.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1-2):187–192.

Lo, A. Y. (1987). A large sample study of the Bayesian bootstrap. *The Annals of Statistics*, 15(1):360–375.

Lo, A. Y. (1988). A Bayesian bootstrap for a finite population. *The Annals of Statistics*, pages 1684–1695.

Lo, A. Y. (1993). A Bayesian bootstrap for censored data. *The Annals of Statistics*, 21(1):100–123.

Lyddon, S. (2018). *Model misspecification and general Bayesian bootstraps*. PhD thesis, University of Oxford.

Lyddon, S., Walker, S., and Holmes, C. C. (2018). Nonparametric learning from Bayesian models with randomized objective functions. In *Advances in Neural Information Processing Systems 31*, pages 2075–2085. Curran Associates, Inc.

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

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3):727–741.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, volume 1, pages 50–55. Alexandria, Virginia. Virginia: American Statistical Association; 1999.

MacKay, D. J. C. (1994). Bayesian nonlinear modeling for the prediction competition. *ASHRAE Trans*, 100.

Magnusson, M., Andersen, M. R., Jonasson, J., and Vehtari, A. (2019). Bayesian leave-one-out cross-validation for large data. *arXiv preprint arXiv:1904.10679*.

Marin, J.-M. and Robert, C. P. (2010). Importance sampling methods for Bayesian discrimination between embedded models. *Frontiers of Statistical Decision Making and Bayesian Analysis*, pages 513–527.

Martin, R. (2018). On nonparametric estimation of a mixing density via the predictive recursion algorithm. *arXiv preprint arXiv:1812.02149*.

Matsubara, T., Knoblauch, J., Briol, F.-X., Oates, C., et al. (2021). Robust generalised Bayesian inference for intractable likelihoods. *arXiv preprint arXiv:2104.07359*.

McDiarmid, C. (1998). Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer.

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

Muliere, P. and Secchi, P. (1996). Bayesian nonparametric predictive inference and bootstrap techniques. *Annals of the Institute of Statistical Mathematics*, 48(4):663–673.

Muliere, P., Walker, S., et al. (2000). Neutral to the right processes from a predictive perspective: a review and new developments. *Metron*, 58:13–30.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83(1):67–79.

Müller, U. K. (2013). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849.

Neal, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 6(4):353–366.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Newton, M., Polson, N. G., and Xu, J. (2018). Weighted Bayesian bootstrap for scalable Bayes. *arXiv e-prints*, page arXiv:1803.04559.

Newton, M. and Raftery, A. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B-Methodological*, 56:3 – 48.

Newton, M. A., Polson, N. G., and Xu, J. (2020). Weighted Bayesian bootstrap for scalable posterior distributions. *Canadian Journal of Statistics*.

Newton, M. A., Quintana, F. A., and Zhang, Y. (1998). Nonparametric Bayes methods using predictive updating. In *Practical nonparametric and semiparametric Bayesian statistics*, pages 45–61. Springer.

Ng, T. L. and Newton, M. A. (2020). Random weighting in Lasso regression. *arXiv preprint arXiv:2002.02629*.

Nie, L. and Ročková, V. (2020). Bayesian Bootstrap spike-and-slab Lasso. *arXiv preprint arXiv:2011.14279*.

O'Hagan, A. and Forster, J. J. (2004). *Kendall's Advanced Theory of Statistics, volume 2B: Bayesian Inference*, volume 2. Arnold.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

Park, J. and Haran, M. (2018). Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, 113(523):1372–1390.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103:681–686.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pompe, E. (2021). Introducing prior information in weighted likelihood bootstrap with applications to model misspecification. *arXiv preprint arXiv:2103.14445*.

Pompe, E., Holmes, C., and Łatuszyński, K. (2018). A framework for adaptive MCMC targeting multimodal distributions. *arXiv e-prints*, page arXiv:1812.02609.

Pompe, E. and Jacob, P. E. (2021). Asymptotics of cut distributions and robust modular inference using posterior bootstrap. *arXiv preprint arXiv:2110.11149*.

Praestgaard, J. and Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, 21(4):2053–2086.

Quintana, F. A., Mueller, P., Jara, A., and MacEachern, S. N. (2020). The dependent Dirichlet process and related models. *arXiv preprint arXiv:2007.06129*.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer.

Rischard, M., Jacob, P. E., and Pillai, N. (2018). Unbiased estimation of log normalizing constants with applications to Bayesian cross-validation. *arXiv preprint arXiv:1810.01382*.

Robbins, H. and Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Elsevier.

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Robert, C. P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, 2nd edition.

Robert, C. P. (2014). On the Jeffreys-Lindley paradox. *Philosophy of Science*, 81(2):216–232.

Robert, C. P. and Wraith, D. (2009). Computational methods for Bayesian model choice. In *AIP conference proceedings*, volume 1193, pages 251–262. AIP.

Roberts, H. V. (1965). Probabilistic prediction. *Journal of the American Statistical Association*, 60(309):50–62.

Robins, J. and Wasserman, L. (2000). Conditioning, likelihood, and coherence: A review of some foundational concepts. *Journal of the American Statistical Association*, 95(452):1340–1346.

Ročková, V. and George, E. I. (2018). The spike-and-slab Lasso. *Journal of the American Statistical Association*, 113(521):431–444.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624.

Ross, G. J. and Markwick, D. (2018). dirichletprocess: An R package for fitting complex Bayesian nonparametric models.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, pages 1151–1172.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, 2(3):808–840.

Rudoy, D. and Wolfe, P. J. (2006). Monte Carlo methods for multi-modal distributions. In *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, pages 2019–2023.

Saarela, O., Stephens, D. A., Moodie, E. E., and Klein, M. B. (2015). On Bayesian estimation of marginal structural models. *Biometrics*, 71(2):279–288.

Saville, B. R., Connor, J. T., Ayers, G. D., and Alvarez, J. (2014). The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. *Clinical Trials*, 11(4):485–493.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Seeger, M., Gerwinn, S., and Bethge, M. (2007). Bayesian inference for sparse generalized linear models. In *Proceedings of the 18th European Conference on Machine Learning*, ECML '07, pages 298–309, Berlin, Heidelberg. Springer-Verlag.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650.

Shahbaba, B. and Neal, R. (2009). Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*, 10(Aug):1829–1850.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494.

Shao, S., Jacob, P. E., Ding, J., and Tarokh, V. (2019). Bayesian model comparison with the Hyvärinen score: Computation and consistency. *Journal of the American Statistical Association*, pages 1–24.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Stephens, M. (1999). Dealing with multimodal posteriors and non-identifiability in mixture models. *Journal of the Royal Statistical Society, Series B*.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 36(2):111–133.

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

Syring, N. A. (2017). *Gibbs Posterior Distributions: New Theory and Applications*. PhD thesis, University of Illinois at Chicago.

Tang, Y., Salakhutdinov, R., and Hinton, G. (2012). Deep mixtures of factor analysers. *arXiv preprint arXiv:1206.4635*.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tokdar, S. T., Martin, R., and Ghosh, J. K. (2009). Consistency of a recursive estimate of mixing distributions. *The Annals of Statistics*, 37(5A):2502–2522.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

Vehtari, A. and Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10):2339–2468.

Vehtari, A. and Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wade, S. (2013). *Bayesian Nonparametric Regression Through Mixture Models*. PhD thesis, Ph. D. thesis, Bocconi University.

Wade, S., Walker, S. G., and Petrone, S. (2014). A predictive study of Dirichlet process mixture models for curve fitting. *Scandinavian Journal of Statistics*, 41(3):580–605.

Walker, S. and Muliere, P. (1997). Beta-stacy processes and a generalization of the Pólya-urn scheme. *The Annals of Statistics*, pages 1762–1780.

Walker, S. G. (2013). Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference*.

Wang, Z. and Scott, D. W. (2019). Nonparametric density estimation for high-dimensional data—algorithms and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(4):e1461.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594.

Watson, J. and Holmes, C. C. (2016). Approximate models and robust decisions. *Statistical Science*, 31(4):465–489.

Waudby-Smith, I. and Ramdas, A. (2020). Estimating means of bounded random variables by betting. *arXiv preprint arXiv:2010.09686*.

West, M. (1991). Kernel density estimation and marginalization consistency. *Biometrika*, 78(2):421–425.

Yamato, H. (1984). Characteristic functions of means of distributions chosen from a Dirichlet process. *The Annals of Probability*, 12(1):262–267.

Zabell, S. L. et al. (1982). WE Johnson's 'Sufficientness' Postulate. *Annals of Statistics*, 10(4):1090–1099.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560.

Zikeba, M., Tomczak, S. K., and Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*.