Scalable Nonparametric Sampling from Multimodal Posteriors with the Posterior Bootstrap

Edwin Fong [1 2] **Simon Lyddon** [1] **Chris Holmes** [1 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.

1. 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 & 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 et al., 2016; Grünwald & 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.

1.1. Our 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

[1]Department of Statistics, University of Oxford, Oxford, United Kingdom [2]The Alan Turing Institute, London, United Kingdom. Correspondence to: Edwin Fong <edwin.fong@stats.ox.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

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. 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 & 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.

2. Bayesian Nonparametric Learning

Assume that we have observed $y_{1:n} \overset{iid}{\sim} F_0$, where $y_{1:n}$ is a sequence of n i.i.d. observables and F_0 is the unknown sampling distribution. We may be interested in a parameter $\theta \in \Theta \subseteq \mathbb{R}^p$, which indexes a family of probability densities $\mathcal{F}_\Theta = \{f_\theta(y); \theta \in \Theta\}$. Conventional Bayesian updating of the prior to the posterior via Bayes' theorem formally assumes that F_0 belongs to the model F_Θ, 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.

2.1. The Parameter of Interest

We define our parameter of interest as

$$\theta_0(F_0) = \arg\min_\theta \int l(y, \theta) dF_0(y) \tag{1}$$

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

The loss function of particular interest is $l(y, \theta) = -\log f_\theta(y)$, where f_θ is the density of some parametric model. The value of θ_0 minimises the Kullback-Leibler divergence $\mathrm{KL}(f_0 || f_\theta)$, which is the parameter of interest in conventional Bayesian analysis (Walker, 2013; Bissiri et al., 2016). We have not assumed that F_Θ contains F_0, and θ_0 in this case does not have any particular generative meaning as it is simply the parameter that satisfies (1).

2.2. The Dirichlet Process Prior

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

$$[F|\alpha, F_\pi] \sim \mathrm{DP}(\alpha, F_\pi) \tag{2}$$

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 $\pi(\theta)$, a sensible choice for the density of F_π is $f_\pi(y) = \int f_\theta(y) d\pi(\theta)$. 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. $F_\pi(y) = \frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \delta_{\hat{y}_i}(y)$ where δ is the Dirac measure. This is in a similar fashion to power priors (Ibrahim et al., 2000). Further intuition is provided in Section A.1 of the Supplementary Material.

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 (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) 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 (1) and tuning its variance. Alternatively, we can select α through the a priori variance of the mean functional (see Section A.2 of the Supplementary Material). 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.

2.3. The NPL Posterior

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

$$[F|y_{1:n}] \sim \text{DP}\left(\alpha + n, G_n\right),$$
$$G_n = \frac{\alpha}{\alpha+n} F_\pi + \frac{1}{\alpha+n}\sum_{i=1}^{n} \delta_{y_i}. \tag{3}$$

Our NPL posterior $\tilde{\pi}(\theta|y_{1:n})$ is thus

$$\tilde{\pi}(\theta|y_{1:n}) = \int \pi(\theta|F) d\pi(F|y_{1:n}) \tag{4}$$

where $\pi(\theta|F) = \delta_{\theta_0(F)}(\theta)$; the delta arises as θ is a deterministic functional of F as in (1). Properties of the NPL posterior follow from properties of the DP, e.g. draws of $F|y_{1:n}$ are almost surely discrete, so (1) simplifies to

$$\theta(F) = \arg\min_\theta \sum_{k=1}^{\infty} w_k l(\tilde{y}_k, \theta) \tag{5}$$

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

$$v_k \sim \text{Beta}(1, \alpha + n), \quad w_k = v_k \prod_{j=1}^{k-1}(1 - v_j). \tag{6}$$

We preserve the theoretical advantages from Lyddon et al. (2018) due to the symmetries 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 regularity conditions, the NPL posterior is consistent at θ_0 as defined in (1), from the properties of the DP (see van der Vaart (1998); Ghosal (2010); Ghosal & 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 $\tilde{\pi}(\cdot|y_{1:n})$ for $\alpha = 0$ *asymptotically dominates* the conventional Bayesian posterior predictive $\pi(\cdot|y_{1:n})$ up to $o(n^{-1})$ under regularity conditions, i.e.

$$\mathbb{E}_{y_{1:n}\sim q}\left[\text{KL}(q(\cdot)||\pi(\cdot|y_{1:n})) - \text{KL}(q(\cdot)||\tilde{\pi}(\cdot|y_{1:n}))\right] = K(q(\cdot)) + o(n^{-1}) \tag{7}$$

for all distributions 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 $\alpha = 0$.

2.4. Sampling from the NPL Posterior

In almost all cases, $\tilde{\pi}(\theta|y_{1:n})$ 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 (1) to generate the sample from $\tilde{\pi}(\theta|y_{1:n})$, as shown in Algorithm 1.

Algorithm 1 NPL Posterior Sampling

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

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.

2.4.1. 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 (5). 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 Ishwaran & 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 2.

Algorithm 2 Posterior Bootstrap Sampling

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

For $\alpha = 0$, we simply draw $w^{(i)}_{1:n} \sim \text{Dir}\left(1, \ldots, 1\right)$, 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.

2.5. 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 & 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 (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 (1) requires non-convex optimization. In general, optimizing non-convex objectives is difficult (see Jain & Kar (2017)), but under smoothness assumption of the loss, we can apply convex optimization methods to find local minima.

2.5.1. Random Restart for Multiple Modes

Random restart (see G. E. Boender & 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 3.

Algorithm 3 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 $\theta_r^{\text{init}} \sim \pi_0$
    $\theta_r^{(i)} = \text{local arg min}_\theta \left(\int l(y, \theta) dF^{(i)}(y), \theta_r^{\text{init}} \right)$
  end for
  $\theta^{(i)} = \arg\min_r \int l(y, \theta_r^{(i)}) dF^{(i)}(y)$
end for

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 Section E.2.2 of the Supplementary Material. 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 Section B of the Supplementary Material.

2.5.2. 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 4.

Algorithm 4 FI-NPL Posterior Sampling

Select θ^{init} from mode of interest
for $i = 1$ **to** B **do**
  Draw $F^{(i)} \sim \text{DP}(\alpha + n, G_n)$
  $\theta^{(i)} = \text{local arg min}_\theta \left(\int l(y, \theta) dF^{(i)}(y), \theta^{\text{init}} \right)$
end for

2.6. Loss-NPL

As we cannot define priors on θ_0 directly, we can instead penalize undesirable properties in the loss

$$l(y, \theta) = -\log f_\theta(y) + \gamma g(\theta). \tag{8}$$

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 Section D of the Supplementary Material. 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 (8).

2.7. 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 $l(y, \theta) = -\log f_\theta(y)$ gives the weighted likelihood bootstrap (Newton & Raftery, 1994), as discussed in Lyddon et al. (2019). Setting the loss to (8) 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 $l(y, \theta)$.

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 $\int m(y, \theta) dF_0(y) = 0$. The definition of a parameter of interest through the loss $l(y, \theta)$ 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 (1), and their methodology is built on a notion of coherency.

3. 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 & 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 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 Sections E.2.3, E.2.4 of the Supplementary Material.

[1] https://github.com/edfong/npl

3.1. 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 $i \in \{1, \dots, n\}$, $j \in \{1, \dots, d\}$ and $k \in \{1, \dots, K\}$ is

$$\begin{aligned} \mathbf{y}_i | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma} &\sim \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{\mu}_k, \text{diag}(\boldsymbol{\sigma}_k^2)\right), \\ \boldsymbol{\pi} | a_0 &\sim \text{Dir}(a_0, \dots, a_0), \\ \mu_{kj} &\sim \mathcal{N}(0, 1), \\ \sigma_{kj} &\sim \text{logNormal}(0, 1). \end{aligned} \tag{9}$$

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

$$l(\mathbf{y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = -\log \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{y}; \boldsymbol{\mu}_k, \text{diag}(\boldsymbol{\sigma}_k^2)\right). \tag{10}$$

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.

3.1.1. Toy Example: Implementation and Results

We analyze toy data from a GMM with $K = 3$, $d = 1$ and the following parameters:

$$\boldsymbol{\pi}_0 = \{0.1, 0.3, 0.6\}, \quad \boldsymbol{\mu}_0 = \{0, 2, 4\}, \quad \boldsymbol{\sigma}_0^2 = \{1, 1, 1\}.$$

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

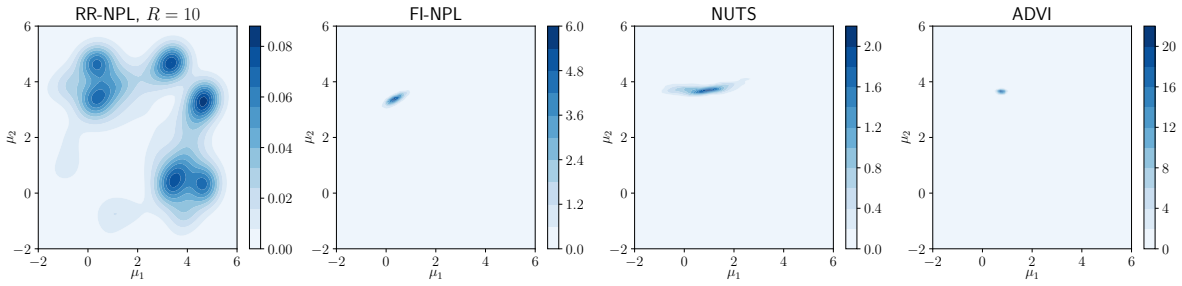

Figure 1. Posterior KDE of (μ_1, μ_2) in $K = 3$ toy GMM problem

each bootstrap maximization with a weighted EM algorithm (derived in Section E.2.1 of the Supplementary Material), 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 $\sigma^2_{kj} \sim \text{IG}(1, 1)$ 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 Section E.3.1 of the Supplementary Material. A larger value is equivalent to a better fit to the test data.

Figure 1 shows the posterior KDEs of (μ_1, μ_2) for 1 run of each method. RR-NPL clearly recovers the multimodality 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 1, 2 that RR-NPL has similar mean LPPD on toy test data compared to NUTS, and is twice as fast as NUTS.

3.1.2. MNIST: Implementation and Results

We now demonstrate FI-NPL on clustering handwritten digits from MNIST (LeCun & 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

$$f_\pi(\mathbf{y}) = \prod_{j=1}^{d} \mathcal{N}(y_j; 0, 0.1^2). \tag{11}$$

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 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 1, 2 that FI-NPL is predictively better and runs around 300 times faster than ADVI.

Table 1. Mean LPPD on held-out test data for GMM

	RR-NPL	FI-NPL	NUTS	ADVI
Toy	**-1.909**$\pm$ 0.040	-1.911 $\pm$ 0.040	**-1.908** $\pm$0.039	-1.912 $\pm$ 0.041
MNIST	/	**2463.4** $\pm$24.1	/	1188.2

Table 2. Run-time for 2000 samples for GMM

	RR-NPL	FI-NPL	NUTS	ADVI
Toy	37.2s $\pm$ 4.5s	5.5$\pm$ 2.2s	1m20s $\pm$ 16s	0.8s $\pm$ 0.1s
MNIST	/	57.9s $\pm$ 1.0s	/	5h6m

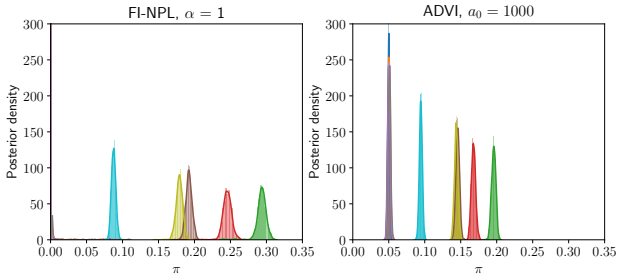

Figure 2. Posterior marginal KDEs of $\boldsymbol{\pi}$ for K=10 GMM on MNIST; 5 of the components have been set to 0 for FI-NPL, and likewise to 0.05 for ADVI

3.2. Logistic Regression with Automatic Relevance Determination 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

$$\begin{aligned} y_i|\mathbf{x}_i, \boldsymbol{\beta}, \beta_0 &\sim \text{Bernoulli}(\eta_i), \\ \eta_i &= \sigma(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0), \\ \beta_j|\lambda_j &\sim \mathcal{N}\left(0, \frac{1}{\lambda_j}\right), \\ \lambda_j|a, b &\sim \text{Gamma}\,(a, b)\,. \end{aligned} \tag{12}$$

Marginally, the prior is the non-standardized t-distribution with (degrees of freedom, location, squared scale)

$$\beta_j \sim \text{Student-t}\left(2a, 0, \frac{b}{a}\right). \tag{13}$$

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 $y, x \overset{iid}{\sim} F_0$, and place a DP prior on the joint distribution $F_0(y, x)$. We target the parameter which satisfies (1) with loss

$$\begin{aligned} l(y, \mathbf{x}, \boldsymbol{\beta}, \beta_0) = &-(y \log \eta + (1-y)\log(1-\eta)) \\ &+\gamma\left(\frac{2a+1}{2}\right)\sum_{j=1}^{d}\log\left(1+\frac{\beta_j^2}{2b}\right) \end{aligned} \tag{14}$$

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

$$\begin{aligned} f_\pi(y, x) &= f_\pi(y) f_\pi(x), \\ f_\pi(y) &= \text{Bernoulli}(0.5), \\ f_\pi(x) &= \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}(x). \end{aligned} \tag{15}$$

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.

3.2.1. Implementation and Results

We analyze 3 binary classification datasets from the UCI ML repository (Dheeru & 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. 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 $\gamma = \frac{1}{n_{\text{train}}}$ for each dataset as explained in Section 2.6 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., 2001–).

We can see in Table 4 that loss-NPL is predictively similar or better than NUTS and ADVI, and from Table 5 we see that the posterior mean is sparser for loss-NPL. Finally, we see from Table 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 Section E.3.4 of the Supplementary Material.

Table 3. UCI datasets descriptions for LogReg

Data Set	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 4. Mean LPPD on held-out test data for LogReg

Data Set	Loss-NPL	NUTS	ADVI
Adult	**-0.326**$\pm$0.004	**-0.326**$\pm$0.004	-0.327 $\pm$ 0.004
Polish	**-0.229**$\pm$ 0.034	-3.336$\pm$ 4.162	-0.247 $\pm$ 0.047
Arcene	**-0.449** $\pm$ 0.104	-0.464 $\pm$ 0.032	**-0.445** $\pm$0.068

Table 5. Percentage of posterior mean $|\beta_j| < \epsilon$ for LogReg

Data Set	ϵ	Loss-NPL	NUTS	ADVI
Adult	0.1	**17.6** $\pm$ 2.8	16.1 $\pm$ 2.7	12.1$\pm$ 3.1
Polish	0.1	**33.5** $\pm$ 4.7	15.9 $\pm$3.3	15.8$\pm$3.5
Arcene	0.01	**87.4** $\pm$ 0.7	4.7 $\pm$ 0.3	3.5 $\pm$ 0.3

Table 6. Run-time for 2000 samples for LogReg

Data Set	Loss-NPL	NUTS	ADVI
Adult	2m24s $\pm$ 8s	2h36m $\pm$ 4m	26.9s $\pm$ 7.3s
Polish	19.0s $\pm$ 4.0s	1h20m $\pm$ 21m	3.3s $\pm$ 0.8s
Arcene	53.5s $\pm$ 1.1s	4h31m $\pm$ 53m	54.2s $\pm$3.3s

3.3. 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 3.2 with the same initialization scheme, set $\gamma = \frac{1}{n}$, 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.

3.3.1. 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. 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 Section E.4.1 of the Supplementary Material. 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.

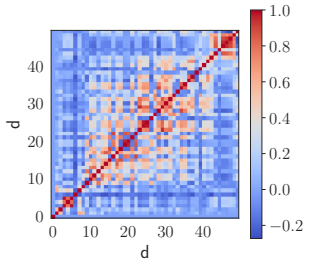

Figure 3. Correlations of covariates $\boldsymbol{x}$ from genetic dataset

The posterior medians of the non-zero components of $\boldsymbol{\beta}$ with 80% central credible interval are shown in Figure 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 Section E.4.2 of the Supplementary Material. 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 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 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.

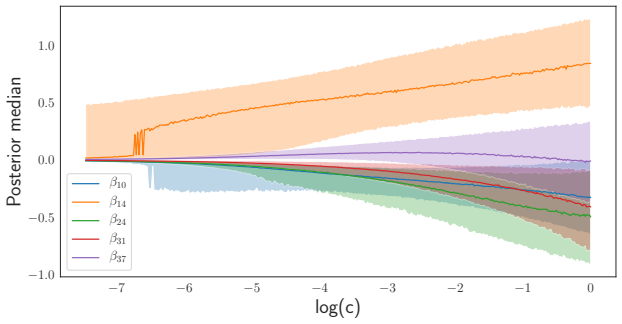

Figure 4. Lasso-type plot for posterior medians of non-zero $\boldsymbol{\beta}$ with 80% credible intervals against $\log(c)$ from genetic dataset

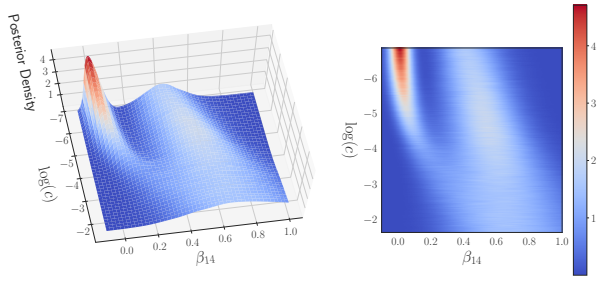

Figure 5. Posterior marginal KDE of β_{14} against $\log(c)$ from genetic dataset

4. 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.

References

Antoniak, C. E. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974. ISSN 0090-5364. doi: 10.1214/aos/1176342871.

Bardenet, R., Doucet, A., and Holmes, C. On Markov chain Monte Carlo methods for tall data. *J. Mach. Learn. Res.*, 18(1):1515–1557, 2017. ISSN 1532-4435.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016. ISSN 13697412. doi: 10.1111/rssb.12158.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32, 2017. ISSN 1548-7660. doi: 10.18637/jss.v076.i01.

Dheeru, D. and Karra Taniskidou, E. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

G. E. Boender, C. and H. G. Rinnooy Kan, A. Bayesian stopping rules for multistart global optimization methods. *Mathematical Programming*, 37:59–80, 1987. doi: 10.1007/BF02591684.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008. doi: 10.1214/08-AOAS191.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.

Ghosal, S. The Dirichlet process, related priors and posterior asymptotics. *Bayesian Nonparametrics*, 28:35, 2010.

Ghosal, S. and van der Vaart, A. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017. doi: 10.1017/9781139029834.

Giordano, R., Broderick, T., and Jordan, M. 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, pp. 1441–1449, Cambridge, MA, USA, 2015. MIT Press.

Grünwald, P. and Van Ommen, T. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017. doi: 10.1214/17-BA1085.

Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G. Result analysis of the NIPS 2003 feature selection challenge. In Saul, L. K., Weiss, Y., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems 17*, pp. 545–552. MIT Press, 2005.

Homan, M. D. and Gelman, A. The no-u-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014. ISSN 1532-4435.

Hu, X., Shonkwiler, R., and Spruill, M. C. Random restarts in global optimization. Technical report, 1994.

Ibrahim, J. G., Chen, M.-H., et al. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000.

Ishwaran, H. and Zarepour, M. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30:269 – 283, 2002. doi: 10.2307/3315951.

Jain, P. and Kar, P. Non-convex optimization for machine learning. *Foundations and Trends in Machine Learning*, 10:142–336, 2017. doi: 10.1561/2200000058.

Jasra, A., Holmes, C. C., and Stephens, D. A. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20(1):50–67, 2005. ISSN 0883-4237. doi: 10.1214/088342305000000016.

Jin, C., Zhang, Y., Balakrishnan, S., Wainwright, M. J., and Jordan, M. I. 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, pp. 4123–4131, USA, 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9.

Jones, E., Oliphant, T., Peterson, P., et al. SciPy: Open source scientific tools for Python, 2001–. URL http://www.scipy.org/.

Kohavi, R. 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*, pp. 202–207. AAAI Press, 1996.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. Automatic differentiation variational inference. *J. Mach. Learn. Res.*, 18(1):430–474, 2017. ISSN 1532-4435.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Lee, A., Caron, F., Doucet, A., and Holmes, C. A hierarchical Bayesian framework for constructing sparsity-inducing priors. *arXiv e-prints*, art. arXiv:1009.1914, 2010.

Lee, A., Caron, F., Doucet, A., and Holmes, C. Bayesian sparsity-path-analysis of genetic association signal using generalized t priors. *Statistical applications in genetics and molecular biology*, 11 2, 2012.

Lyddon, S., Walker, S., and Holmes, C. C. Nonparametric learning from Bayesian models with randomized objective functions. In *Advances in Neural Information Processing Systems 31*, pp. 2075–2085. Curran Associates, Inc., 2018.

Lyddon, S. P., Holmes, C. C., and Walker, S. G. General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 03 2019. to appear, doi: 10.1093/biomet/asz006.

MacKay, D. J. C. Bayesian nonlinear modeling for the prediction competition. *ASHRAE Trans*, 100, 1994.

Neal, R. M. Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 6(4):353–366, 1996. ISSN 0960-3174. doi: 10.1007/BF00143556.

Newton, M. and Raftery, A. Approximate Bayesian inference by the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B-Methodological*, 56:3 – 48, 1994. doi: 10.1111/j.2517-6161.1994.tb01956.x.

Newton, M., Polson, N. G., and Xu, J. Weighted Bayesian bootstrap for scalable Bayes. *arXiv e-prints*, art. arXiv:1803.04559, 2018.

Owen, A. B. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988. doi: 10.1093/biomet/75.2.237.

Park, T. and Casella, G. The Bayesian lasso. *Journal of the American Statistical Association*, 103:681–686, 2008. ISSN 0162-1459. doi: 10.1198/016214508000000337.

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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pompe, E., Holmes, C., and Łatuszyński, K. A framework for adaptive MCMC targeting multimodal distributions. *arXiv e-prints*, art. arXiv:1812.02609, 2018.

Rubin, D. B. The Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981. ISSN 0090-5364. doi: 10.1214/aos/1176345338.

Rudoy, D. and Wolfe, P. J. Monte Carlo methods for multimodal distributions. In *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, pp. 2019–2023, 2006. doi: 10.1109/ACSSC.2006.355120.

Seeger, M., Gerwinn, S., and Bethge, M. Bayesian inference for sparse generalized linear models. In *Proceedings of the 18th European Conference on Machine Learning*, ECML '07, pp. 298–309, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-74957-8. doi: 10.1007/978-3-540-74958-5_29.

Sethuraman, J. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994. ISSN 10170405, 19968507.

Stephens, M. Dealing with multimodal posteriors and non-identifiability in mixture models. *Journal of the Royal Statistical Society, Series B*, 1999.

van der Vaart, A. W. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.

Walker, S. G. Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference*, 2013. ISSN 03783758. doi: 10.1016/j.jspi.2013.05.013.

Watson, J., Holmes, C., et al. Approximate models and robust decisions. *Statistical Science*, 31(4):465–489, 2016.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23 (4):550–560, 1997. ISSN 0098-3500. doi: 10.1145/279232.279236.

Zikeba, M., Tomczak, S. K., and Tomczak, J. M. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, 2016.