Supplementary Material: Scalable Nonparametric Sampling from Multimodal Posteriors with the Posterior Bootstrap

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

A. Eliciting the Prior Dirichlet Process

A.1. Intuition of the Prior F_π

The parameter of interest when model fitting (Walker, 2013) is

$$\begin{aligned}\theta_0(F_0) &= \arg\max_\theta \int \log f_\theta(y) dF_0(y) \\ &= \arg\min_\theta \mathrm{KL}(f_0 || f_\theta).\end{aligned} \tag{A.1}$$

The prior on F_0 is

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

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:

$$\theta_0(F) \xrightarrow{p} \arg\min_\theta \mathrm{KL}(f_\pi || f_\theta). \tag{A.3}$$

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 (A.3), 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.

A.2. Selecting α through the Mean Functional

We can tune α through the a priori variance of the mean functional

$$\begin{aligned}\theta_\mu(F) &= \arg\min_\theta \int (y-\theta)^2 dF(y) \\ &= \int y dF(y).\end{aligned} \tag{A.4}$$

If $F \sim \mathrm{DP}(\alpha, F_\pi)$, then the a prior variance of (A.4) follows from the properties of the Dirichlet process (DP):

$$\mathrm{Var}\left[\theta_\mu(F)\right] = \frac{\mathrm{Var}_{F_\pi}\left[y\right]}{1+\alpha} \tag{A.5}$$

and so we can elicit α from a priori knowledge of $\mathrm{Var}\left[\theta_\mu(F)\right]$.

[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.

B. 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 & 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.

C. 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

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{i=1}^{n} w_i l(y_i, \theta), \\ \nabla_\theta L(\theta) &= \sum_{i=1}^{n} w_i \nabla_\theta l(y_i, \theta). \end{aligned} \tag{C.1}$$

If we subsample a mini-batch $\tilde{y}_{1:m} \overset{iid}{\sim} \sum_{i=1}^{n} w_i \delta_{y_i}$, we can then calculate the mini-batch gradient

$$\nabla_\theta L^m(\theta) = \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta l(\tilde{y}_i, \theta). \tag{C.2}$$

The mini-batch gradient is unbiased:

$$\mathbb{E}\left[\nabla_\theta L^m(\theta)\right] = \mathbb{E}\left[\nabla_\theta l(\tilde{y}, \theta)\right] = \sum_{i=1}^{n} w_i \nabla_\theta l(y_i, \theta). \tag{C.3}$$

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 & Ba, 2014) help with escaping saddle points, which can potentially reduce the number of R required for RR-NPL to obtain full mode exploration.

D. Selecting γ in Loss-NPL

For loss-NPL we can set the loss function to

$$l(y, \theta) = -\log f_\theta(y) - \gamma \log \pi(\theta). \tag{D.1}$$

In this case, we recommend the scaling parameter to be $\gamma = \frac{1}{n}$ 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 $\int l(y, \theta) dF$ for $\alpha = 0$ (i.e. $F \sim \text{DP}\left(n, \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}\right)$):

$$\mathbb{E}\left[\int l(y, \theta) dF\right] = -\frac{1}{n} \sum_{i=1}^{n} \log f_\theta(y_i) - \gamma \log \pi(\theta). \tag{D.2}$$

We obtain the same weighting as in Bayesian inference between the log-likelihood and log-prior for $\gamma = \frac{1}{n}$.

E. Examples

E.1. 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

$$\begin{aligned} \theta_0(F_0) &= \arg\max_\theta \int \log f_\theta(y) dF_0(y) \\ &= \arg\min_\theta \int (y - \theta)^2 dF_0(y). \end{aligned} \tag{E.1}$$

If we set the derivative of the objective to 0, we obtain

$$\theta_0(F_0) = \int y dF_0(y). \tag{E.2}$$

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

$$f_\pi(y) = \int f_\theta(y) d\pi(\theta) = \mathcal{N}(y; 0, \sigma^2 + \tau^2). \tag{E.3}$$

When $n = 0$, our NPL prior $\tilde{\pi}(\theta)$ is approximately normal (Yamato, 1984) from the properties of the DP:

$$\tilde{\pi}(\theta) \approx \mathcal{N}\left(\theta; 0, \frac{\sigma^2 + \tau^2}{1 + \alpha}\right). \tag{E.4}$$

E.1.1. 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 E.1 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 E.1 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 & 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 E.2 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.

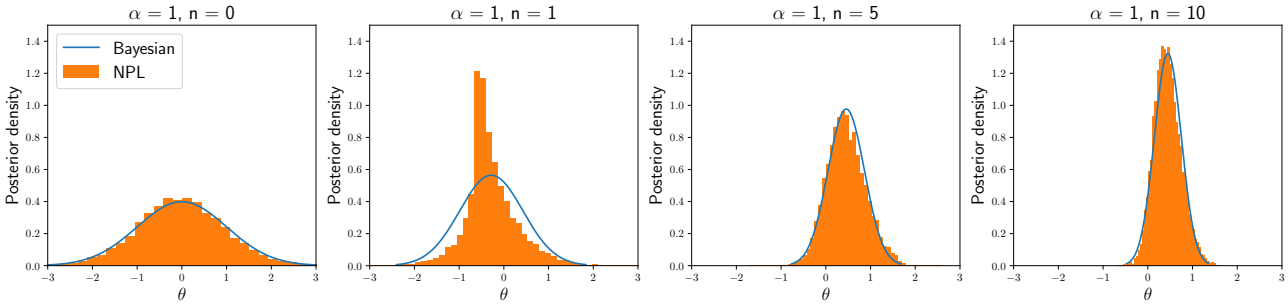

Figure E.1. NPL posterior and Bayesian posterior for fixed α and increasing n in normal location model

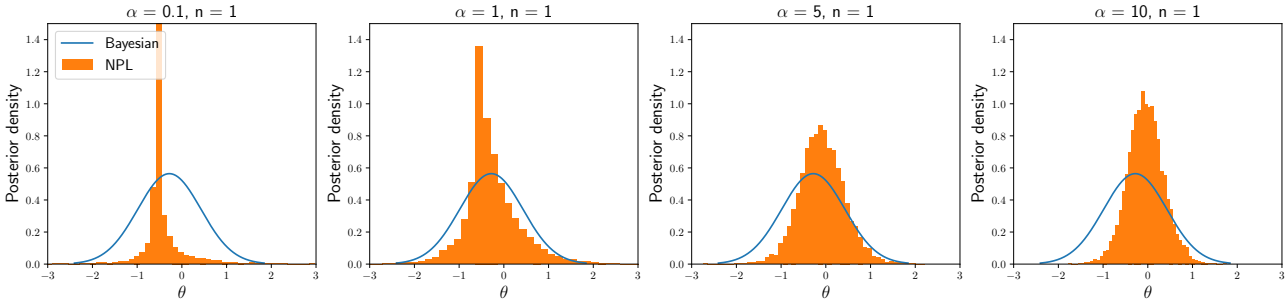

Figure E.2. NPL posterior and Bayesian posterior for increasing α and fixed n in normal location model

E.2. Gaussian Mixture Model

E.2.1. Optimization Details

We derive the EM algorithm that maximizes the weighted likelihood of the diagonal-covariance GMM:

$$\begin{aligned}\mathcal{L}^w(\theta) &= \sum_{i=1}^n w_i \log f_\theta(\mathbf{y}_i) \\ &= \sum_{i=1}^n w_i \left(\log f_\theta(\mathbf{y}_i, z_i) - \log f_\theta(z_i|\mathbf{y}_i)\right).\end{aligned}$$

Taking an expectation over the posterior $f_{\theta'}(z_{1:n}|\mathbf{y}_{1:n})$, we obtain

$$\begin{aligned}\mathcal{L}^w(\theta) &= \sum_{i=1}^n w_i \sum_{z_i} f_{\theta'}(z_i|\mathbf{y}_i)\left(\log f_\theta(\mathbf{y}_i, z_i)\right) - \sum_{i=1}^n w_i \sum_{z_i} f_{\theta'}(z_i|\mathbf{y}_i)\left(\log f_\theta(z_i|\mathbf{y}_i)\right) \\ &= \sum_{i=1}^n w_i Q^i(\theta|\theta') - \sum_{i=1}^n w_i H^i(\theta|\theta').\end{aligned}$$

Taking the difference of the weighted likelihood with θ'

$$\mathcal{L}^w(\theta) - \mathcal{L}^w(\theta') = \sum_{i=1}^n w_i \left(Q^i(\theta|\theta') - Q^i(\theta'|\theta')\right) + \sum_{i=1}^n w_i \left(H^i(\theta'|\theta') - H^i(\theta|\theta')\right).$$

From Gibbs' inequality,

$$H^i(\theta'|\theta') \geq H^i(\theta|\theta').$$

As all $w_i \geq 0$,

$$\mathcal{L}^w(\theta) - \mathcal{L}^w(\theta') \geq \sum_{i=1}^n w_i \left(Q^i(\theta|\theta') - Q^i(\theta'|\theta')\right).$$

So by maximizing $\sum_{i=1}^n w_i Q^i(\theta|\theta')$ w.r.t. θ, we cannot decrease the weighted log-likelihood. As a reminder, the log-likelihood for each datapoint is

$$\log f_\theta(\mathbf{y}_i) = \log\left(\sum_{k=1}^K \pi_k \mathcal{N}\left(\mathbf{y}_i; \boldsymbol{\mu}_k, \operatorname{diag}\left(\boldsymbol{\sigma}_k^2\right)\right)\right). \tag{E.5}$$

At the expectation step, we calculate

$$f_\theta(z_i = k|\mathbf{y}_i) = \frac{\prod_{j=1}^d \mathcal{N}(y_{ij}; \mu_{kj}, \sigma_{kj}^2)\pi_k}{\sum_{k=1}^K \prod_{j=1}^d \mathcal{N}(y_{ij}; \mu_{kj}, \sigma_{kj}^2)\pi_k}. \tag{E.6}$$

The maximization step is then:

$$\begin{aligned}\hat{\pi}_k &= \sum_{i=1}^n w_i f_{\theta'}(z_i = k|\mathbf{y}_i), \\ \hat{\mu}_{kj} &= \frac{\sum_{i=1}^n w_i f_{\theta'}(z_i = k|\mathbf{y}_i) y_{ij}}{\hat{\pi}_k}, \\ \hat{\sigma}_{kj}^2 &= \frac{\sum_{i=1}^n w_i f_{\theta'}(z_i = k|\mathbf{y}_i)(y_{ij} - \hat{\mu}_{kj})^2}{\hat{\pi}_k}.\end{aligned} \tag{E.7}$$

E.2.2. 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 E.3, E.4, E.5, and for increasing R in Figures E.6, E.7, E.8. 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 E.1, and we see that the run-time increases roughly linearly with R.

Table E.1. Run-time (seconds) for 2000 posterior samples on Azure for different values of R with RR-NPL

	$R = 1$	$R = 2$	$R = 5$	$R = 10$
TOY SEP	4.9 ± 0.6	8.0 ± 1.1	19.0 ± 2.1	$37.2\text{s} \pm 4.5\text{s}$

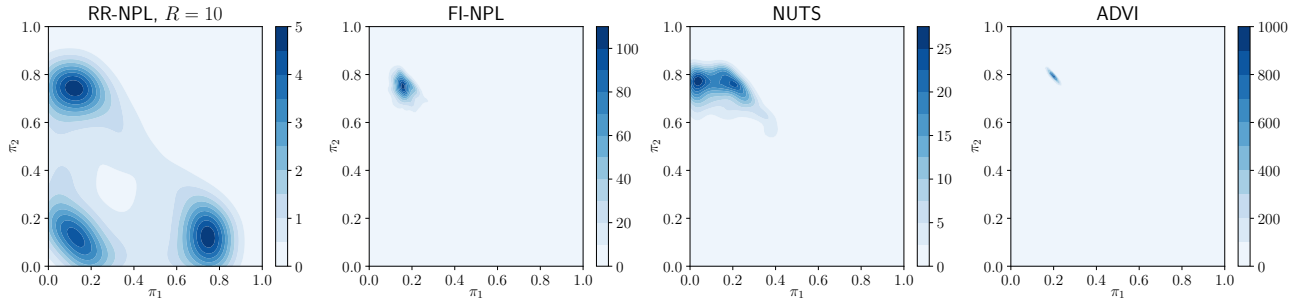

Figure E.3. Posterior KDE of (π_1, π_2) in $K = 3$ toy separable GMM problem

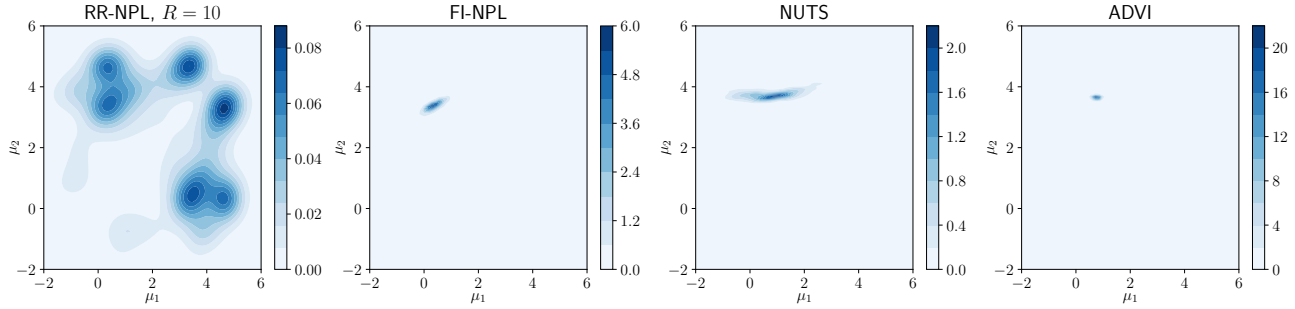

Figure E.4. Posterior KDE of (μ_1, μ_2) in $K = 3$ toy separable GMM problem

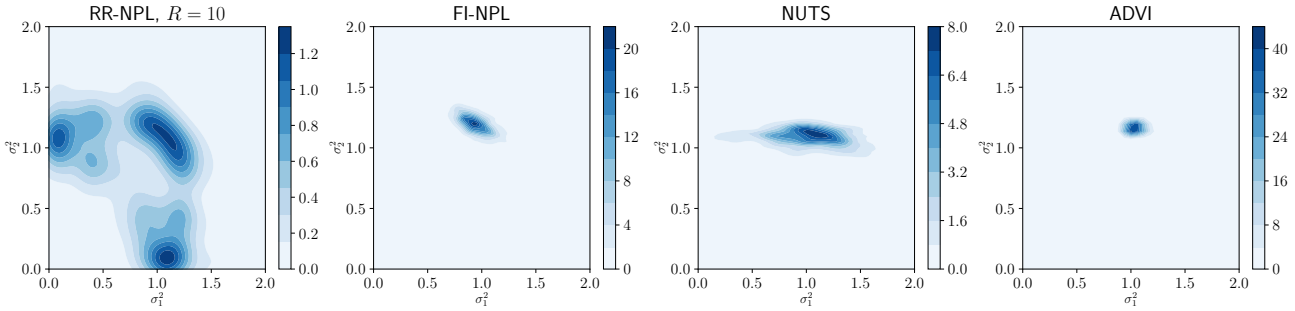

Figure E.5. Posterior KDE of (σ_1^2, σ_2^2) in $K = 3$ separable toy GMM problem

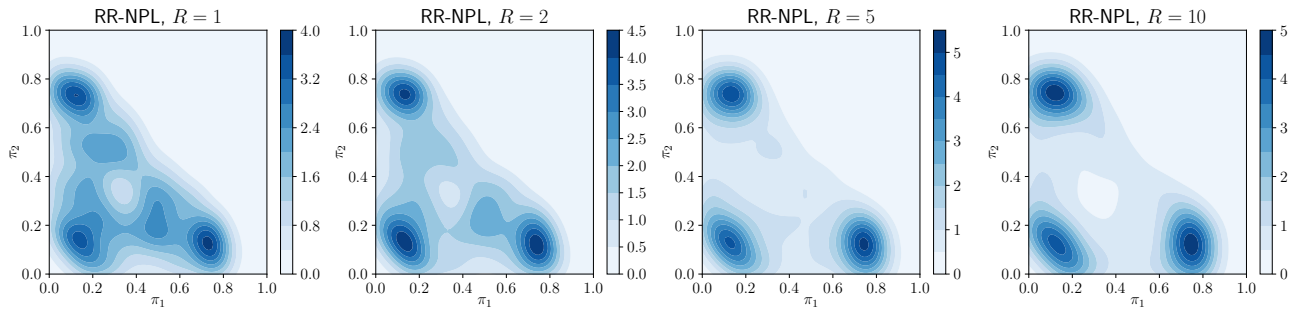

Figure E.6. Posterior KDE of (π_1, π_2) in $K = 3$ separable toy GMM problem for increasing R

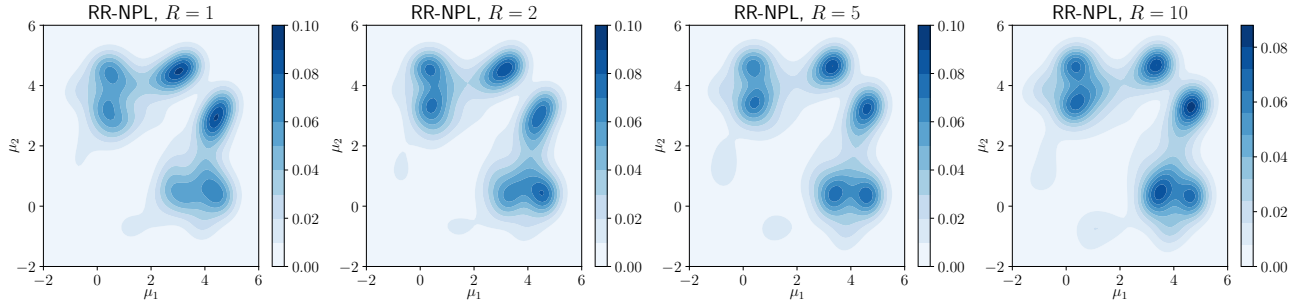

Figure E.7. Posterior KDE of (μ_1, μ_2) in $K = 3$ toy GMM problem for RR-NPL with increasing R

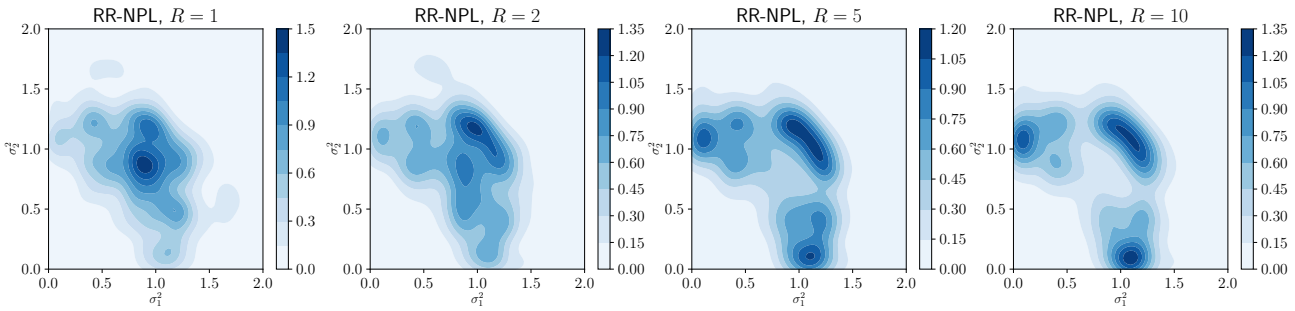

Figure E.8. Posterior KDE of (σ_1^2, σ_2^2) in $K = 3$ separable GMM toy problem for increasing R

E.2.3. 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 3.1 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 Section E.3.1. We use self-normalized importance sampling (SNIS) as the posterior is unnormalized, so the estimate of the posterior predictive is defined as

$$p(\tilde{y}|y_{1:n}) \approx \frac{\sum_{b=1}^{B} w_b f(\tilde{y}|\boldsymbol{\theta}_b)}{\sum_{b=1}^{B} w_b} = \sum_{b=1}^{B} \tilde{w}_b f(\tilde{y}|\boldsymbol{\theta}_b), \quad w_b = \frac{f(y_{1:n}|\boldsymbol{\theta}_b)\pi(\boldsymbol{\theta}_b)}{q(\boldsymbol{\theta}_b)}, \quad \boldsymbol{\theta}_b \sim q(\boldsymbol{\theta}) \tag{E.8}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$ is 9-dimensional, and $\{f(y|\boldsymbol{\theta}), \pi(\boldsymbol{\theta})\}$ is the GMM likelihood and prior as defined in (9) 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:

$$q(\boldsymbol{\pi}) = \text{Dir}(0.1, \ldots, 0.1), \quad q(\mu_{kj}) = \mathcal{N}(0, 25), \quad q(\sigma_{kj}) = \text{logNormal}(0, 1). \tag{E.9}$$

The proposal has support in the true value of $\boldsymbol{\theta}$ and thus should have support in regions of high $f(\tilde{y}|\boldsymbol{\theta})$. 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 E.2, 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 E.2. Performance on held-out test data for toy GMM

	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 ± 0.80
Run-time	29.3s $\pm$ 0.5s	31.0s $\pm$ 8.7s

E.2.4. 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 Section A.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 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:

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

The means are closer together, and so NUTS can mix properly as shown in Figure E.9. For MDP-NPL, we center the MDP with the Bayesian model given in (9) 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 in Figures E.9, E.10, E.11. 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 E.3, but the run-times are much greater for MDP-NPL as shown in Table E.4. 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.

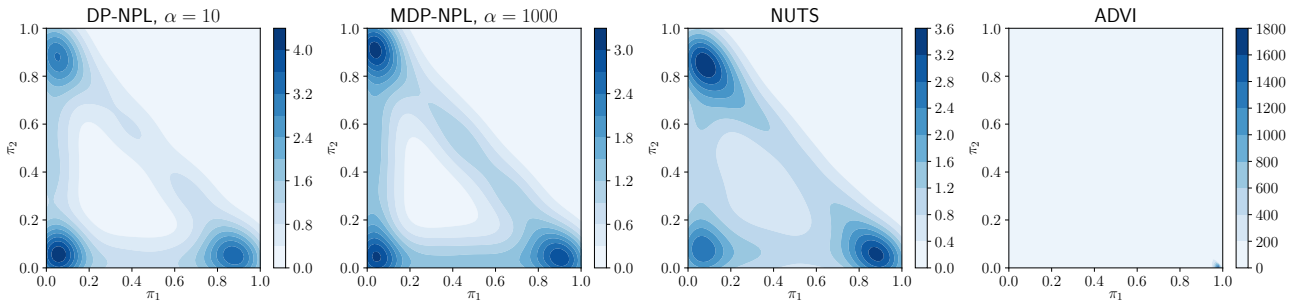

Figure E.9. Posterior KDE of (π_1, π_2) in $K = 3$ inseparable toy GMM problem 2

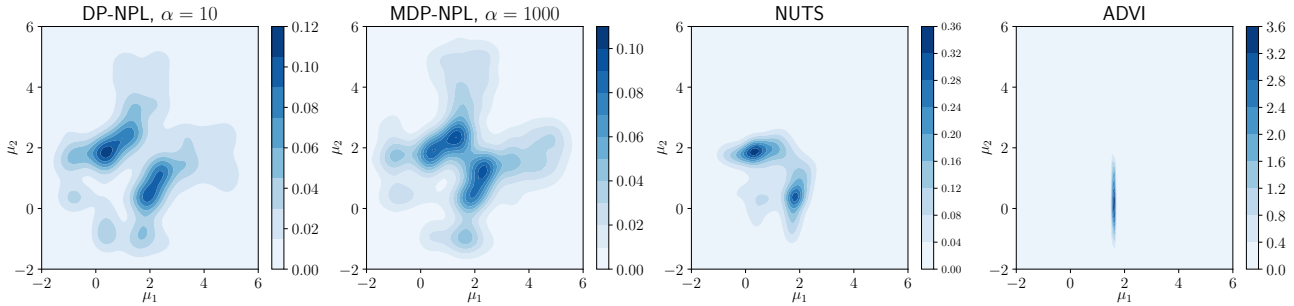

Figure E.10. Posterior KDE of (μ_1, μ_2) in $K = 3$ inseparable toy GMM problem

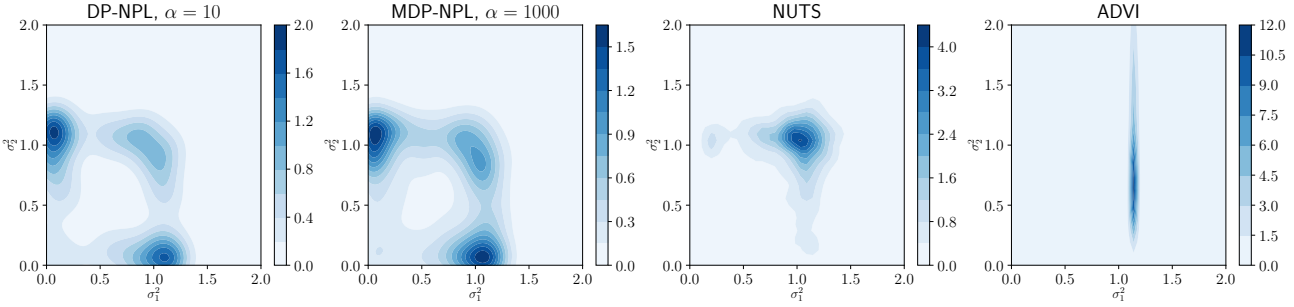

Figure E.11. Posterior KDE of (σ_1^2, σ_2^2) in $K = 3$ inseparable toy GMM problem

Table E.3. Mean LPPD on held-out test data for inseparable GMM

	DP-NPL	MDP-NPL	NUTS	ADVI
TOY	-1.612 ± 0.038	-1.610 ± 0.037	-1.609±0.037	-1.613 ± 0.038

Table E.4. Run-time for 2000 samples for inseparable GMM

	DP-NPL	MDP-NPL	NUTS	ADVI
TOY	52.6S± 6.3S	3M2S ± 13S	2M12S ± 12S	0.8S ± 0.1S

E.3. Logistic Regression with Automatic Relevance Determination Priors

E.3.1. 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 $(\tilde{y}, \tilde{x})$:

$$\begin{aligned} p(\tilde{y}|\tilde{x}, y_{1:n}, x_{1:n}) &= \int f(\tilde{y}|\tilde{x}, \boldsymbol{\beta}) d\tilde{\pi}(\boldsymbol{\beta}|y_{1:n}, x_{1:n}) \\ &\approx \frac{1}{B}\sum_{b=1}^{B} f(\tilde{y}|\tilde{x}, \boldsymbol{\beta}_b), \\ \boldsymbol{\beta}_b &\sim \tilde{\pi}(\boldsymbol{\beta}|y_{1:n}, x_{1:n}), \end{aligned} \tag{E.10}$$

where $f(y|x, \boldsymbol{\beta})$ is the likelihood, $\tilde{\pi}(\boldsymbol{\beta}|y_{1:n}, x_{1:n})$ 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:

$$\text{Mean LPPD} = \frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}} \log p(\tilde{y}_i|\tilde{x}_i, y_{1:n}, x_{1:n}). \tag{E.11}$$

Below, we additionally include the mean squared error (MSE) here on held-out test data, defined

$$\text{MSE} = \frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}} (p(\tilde{y}_i|\tilde{x}_i, y_{1:n}, x_{1:n}) - \tilde{y}_i)^2. \tag{E.12}$$

Finally, we also report the percentage accuracy, defined

$$\begin{aligned} \text{P.a.} &= \frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}} \hat{y}_i^{\tilde{y}_i}(1-\hat{y}_i)^{(1-\tilde{y}_i)}, \\ \hat{y}_i &= \mathbb{I}(p(\tilde{y}_i|\tilde{x}_i, y_{1:n}, x_{1:n}) > 0.5) \end{aligned} \tag{E.13}$$

where $\mathbb{I}$ is the indicator function.

E.3.2. 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 $\boldsymbol{\beta}_b \sim \tilde{\pi}(\boldsymbol{\beta}|y_{1:n}, x_{1:n})$ as above. We then report the percentage of components of $\hat{\boldsymbol{\beta}}$ that have absolute value less than ϵ.

E.3.3. 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

$$\begin{aligned} \frac{\partial l(y, x, \boldsymbol{\beta}, \beta_0)}{\partial \beta_j} &= -(y-\eta)\, x_j + \gamma\left(\frac{2a+1}{2b+\beta_j^2}\right)\beta_j, \\ \frac{\partial l(y, x, \boldsymbol{\beta}, \beta_0)}{\partial \beta_0} &= -(y-\eta)\,. \end{aligned} \tag{E.14}$$

E.3.4. Additional Results

We can see in Tables E.5, E.6 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 E.12 for reference.

Table E.5. MSE on held-out test data

Data Set	Loss-NPL	NUTS	ADVI
Adult	**0.104** $\pm$ 0.001	**0.104** $\pm$ 0.001	0.105$\pm$ 0.002
Polish	**0.056** $\pm$ 0.011	0.524 $\pm$0.236	0.058 $\pm$ 0.011
Arcene	**0.134**$\pm$0.026	0.152$\pm$ 0.014	0.143 $\pm$0.020

Table E.6. Predictive accuracy % on held-out test data

Data Set	Loss-NPL	NUTS	ADVI
Adult	**84.92** $\pm$ 0.29	**84.92**$\pm$ 0.30	84.84 $\pm$ 0.30
Polish	**93.65** $\pm$ 1.68	37.27$\pm$ 23.81	93.51$\pm$ 1.56
Arcene	**81.73** $\pm$3.79	77.80$\pm$4.22	79.70$\pm$ 3.40

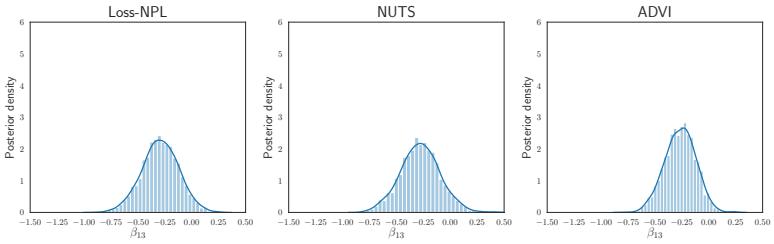

Figure E.12. Posterior marginal KDE of β_{13} for 'Adult' dataset

E.4. Bayesian Sparsity-path-analysis

E.4.1. Values of $\boldsymbol{\beta}$

We follow the setting of $\boldsymbol{\beta}$ in Lee et al. (2012): the 5 non-zero indices and their respective values are

$$\mathcal{I} = \{10, 14, 24, 31, 37\},$$
$$\beta_{\mathcal{I}} = \{-0.2538, 0.4578, -0.1873, -0.1498, 0.0996\}.$$

E.4.2. Variable selection

We see more clearly from Figure E.13 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.

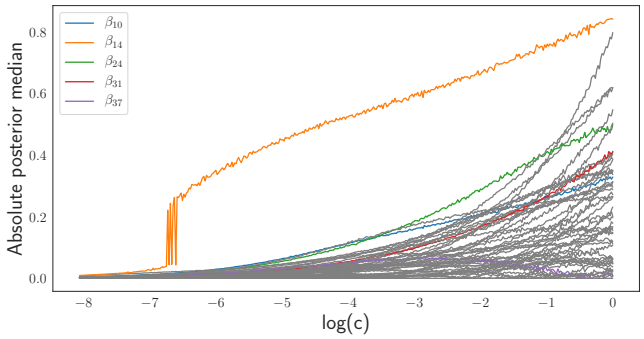

Figure E.13. Lasso-type plot for absolute posterior medians of $\boldsymbol{\beta}$ against $\log(c)$ from genetic dataset with non-zero components in colour

References

Betrò, B. and Schoen, F. Sequential stopping rules for the multistart algorithm in global optimisation. *Mathematical Programming*, 38(3):271–286, 1987. ISSN 1436-4646. doi: 10.1007/BF02592015.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Dick, T., Wong, E., and Dann, C. How many random restarts are enough? Technical report, 2014.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011. ISSN 1532-4435.

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.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

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.

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.

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.

Yamato, H. Characteristic functions of means of distributions chosen from a Dirichlet process. *The Annals of Probability*, 12(1):262–267, 1984. ISSN 0091-1798. doi: 10.1214/aop/1176993389.

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.