WILD POSTERIORS IN THE WILD

Yunyi Shen* **Tamara Broderick***

1 INTRODUCTION

Bayesian posterior approximation is more accessible to practitioners than ever thanks to modern black-box software — such as Stan (Carpenter et al., 2017), Pyro (Bingham et al., 2019), PyMC (Abril-Pla et al., 2023), NIMBLE (de Valpine et al., 2017), and others. While this software offers widely accurate approximation with minimal user effort, it is well known that certain posterior geometries remain challenging for standard approximation schemes. As such, research into alternative approximations continues to thrive. In these papers, it is common for authors to demonstrate that their new approximation works well by testing it on posterior shapes considered to be challenging or "wild." But the shapes are not always directly connected to a practical application where they might arise. In the present note, we provide examples of applications in the wild that give rise to some common benchmark posterior shapes. We hope these connections to applications will be useful for developers of posterior approximations in at least two ways. (1) Understanding the underlying application and model can help a developer understand precisely what the user hopes to get out of the data analysis. For instance, a posterior mean and variance need not always be useful posterior summaries. (2) In cases where a posterior shape can be matched to a modern data analysis, the developer can rest assured that a good approximation will be useful for applied problems. While the present note cannot be exhaustive, we collect further examples at `https://github.com/YunyiShen/weird-posteriors`. And we hope our work inspires developers to track down applications corresponding to their benchmark shapes.

2 WILD POSTERIORS IN THE WILD

Banana. The contours of a posterior distribution can take a "banana" shape when data provides information about the product of two parameters but can only weakly identify the two. As one example, consider an **N-mixture** model (Royle, 2004), used by ecologists counting unmarked animals.

$$\begin{aligned} &\text{observing: } y_{i,n_i} \\ &y_{i,n_i} \overset{ind}{\sim} \text{Binomial}(p, N_i), \quad N_i \overset{iid}{\sim} \text{Poisson}(\lambda) \quad \text{(likelihood)} \\ &p \sim \text{Unif}(0,1), \quad \lambda \sim \text{Gamma}(\alpha, \beta) \quad \text{(prior)} \end{aligned} \tag{1}$$

Here y_{i,n_i} is the count of the animal at location i and "repeat" n_i. The goal is to infer the "abundance" λ and "detection rate" p. If for each location i there are no repeats (i.e., $n_i = 1$), then the data is Poisson distributed with parameter λp, so λ and p are not identified. With some repeats at each location, these two parameters are weakly identified (fig. 1-A). We give the details for all of our simulations in appendix C.

A second example is the **occupancy** model (MacKenzie et al., 2002).

$$\begin{aligned} &\text{observing: } y_{i,n_i} \\ &y_{i,n_i} \overset{ind}{\sim} \text{Bernoulli}(pz_i), \quad z_i \overset{iid}{\sim} \text{Bernoulli}(\psi) \quad \text{(likelihood)} \\ &\psi, p \sim \text{Unif}(0,1) \quad \text{(prior)} \end{aligned} \tag{2}$$

Here y_{i,n_i} equals 1 if one sees an animal at location i in repeat n_i and otherwise $y_{i,n_i} = 0$. The goal is to infer the occupancy rate ψ and the detection rate p. When most of the data observations are 0, there are two competing explanations for the data: either small p or small ψ, so the model is at best weakly identified (fig. 1-B).

Needle. A posterior can be "needle"-shaped when the data provides information about the sum or difference of two parameters but can only weakly identify their values. A familiar example is

*EECS, MIT, {yshen99, tbroderick}@mit.edu

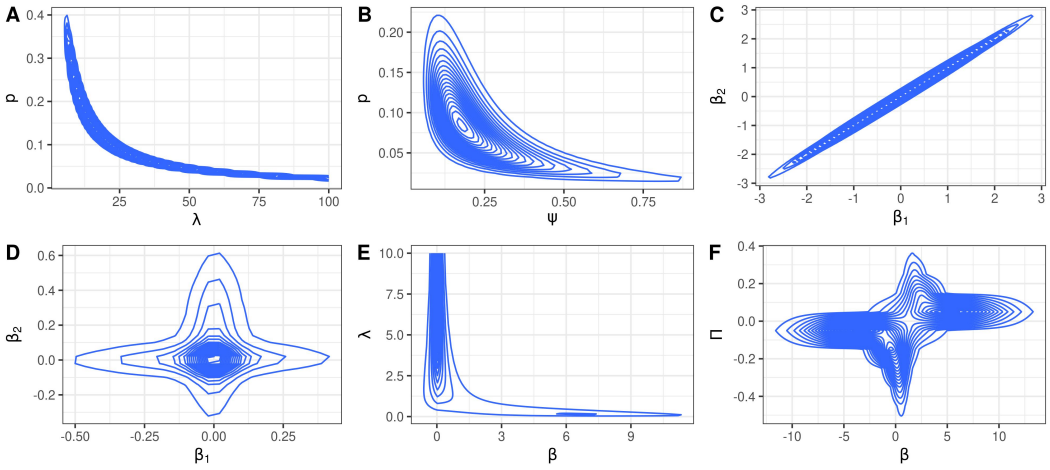

Figure 1: Example unnormalized posterior contour. A: N-mixture model, B: Occupancy model, C: linear regression with collinearity, D: Spike-and-slab prior, E: adaptive LASSO, F: weak instrumental variable.

linear regression with **multicollinearity**. Figure 1-C shows an example with two linear regression coefficients.

Cross. Consider a linear regression where the coefficients have a **spike-and-slab** Gaussian mixture prior (George & McCulloch, 1993). For instance, Kazemi Naeini et al. (2024) used this prior in genome-wide association studies to find variants related to bipolar disorder. When data is limited, the posterior can look like a "cross" (fig. 1-D).

$$\begin{aligned} &\text{observing: } (x_i, y_i) \\ &y_i = x_i^\top \beta + \epsilon_i, \epsilon_i \overset{iid}{\sim} \text{Normal}(0,1) \quad \text{(likelihood)} \\ &\beta_j \overset{ind}{\sim} \text{Normal}(0, 0.1 z_j + 100(1-z_j)), \;\; z_j \overset{iid}{\sim} \text{Bernoulli}(0.1) \quad \text{(prior)} \end{aligned} \tag{3}$$

Multimodal. Multimodality is common in Bayesian variable selection. The Bayesian LASSO (Park & Casella, 2008) has a log-concave posterior and thus is not multimodal. However, the **adaptive LASSO**[1] proposed in Park & Casella (2008) puts a hyperprior on the amount of penalization and can exhibit multimodality even with a single covariate (fig. 1-E).

$$\begin{aligned} &\text{observing: } (x_i, y_i), \sigma \\ &y_i \overset{ind}{\sim} N(x_i\beta, \sigma^2) \quad \text{(likelihood)} \\ &\beta \sim \text{Laplace}(0, \lambda) \quad \lambda \sim \text{Unif}(0.001, 10) \quad \text{(prior)} \end{aligned} \tag{4}$$

Gallo et al. (2022) use this prior (including the uniform hyperprior on λ just above) to analyze how mammals adjust "diel" activity across a gradient of urbanization. Van Erp et al. (2019) reviews the use of this style of prior in psychology, and Banner et al. (2020) critiques it in ecology.

Singularity. Hoogerheide & van Dijk (2008) showed that a simple **instrumental variable** model with a diffuse prior can exhibit challenging posterior behaviors, including point singularities, diverging ridgelines, and multiple modes. Hoogerheide & van Dijk (2008) use this model to analyze the effect of education on income using birth quarter as an instrument (since, in the United States, birth date determines start and duration of schooling).

$$\begin{aligned} &\text{observing: } (x_i, y_i, z_i) \\ &y_i = x_i\beta + \epsilon_i, \;\; x_i = z_i\Pi + v_i, \;\; (\epsilon_i, v_i) \overset{iid}{\sim} \text{Normal}(0, \Sigma) \quad \text{(likelihood)} \\ &p(\beta, \Pi, \Sigma) \propto |\Sigma|^{-3/2} \quad \text{(prior)} \end{aligned} \tag{5}$$

One problem arises when Π is close to 0; then the model is very weakly identified, and a singularity forms around $\Pi = 0$ (fig. 1-F).

[1]This name appears in Wang (2012).

REFERENCES

Oriol Abril-Pla, Virgile Andreani, Colin Carroll, Larry Dong, Christopher J Fonnesbeck, Maxim Kochurov, Ravin Kumar, Junpeng Lao, Christian C Luhmann, Osvaldo A Martin, Michael Osthege, Ricardo Vieira, Thomas Wiecki, and Robert Zinkov. PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science*, 9:e1516, 2023.

Katharine M Banner, Kathryn M Irvine, and Thomas J Rodhouse. The use of Bayesian priors in ecology: The good, the bad and the not great. *Methods in Ecology and Evolution*, 11(8):882–889, 2020.

Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20(28):1–6, 2019.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76:1–32, 2017.

Perry de Valpine, Daniel Turek, Christopher J Paciorek, Clifford Anderson-Bergman, Duncan Temple Lang, and Rastislav Bodik. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26 (2):403–413, 2017.

Travis Gallo, Mason Fidino, Brian Gerber, Adam A Ahlers, Julia L Angstmann, Max Amaya, Amy L Concilio, David Drake, Danielle Gay, Elizabeth W Lehrer, Maureen H Murray, Travis J Ryan, Colleen Cassady St Clair, Carmen M Salsbury, Heather A Sander, Theodore Stankowich, Jaque Williamson, J Amy Belaire, Kelly Simon, and Seth B Magle. Mammals adjust diel activity across gradients of urbanization. *eLife*, 11:e74756, 2022.

Edward I George and Robert E McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Lennart F Hoogerheide and Herman K van Dijk. Possibly ill-behaved posteriors in econometric models. 2008.

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

Maryam Kazemi Naeini, Mahdi Akbarzadeh, Iraj Kazemi, Doug Speed, and Sayed Mohsen Hosseini. Using the Bayesian variational spike and slab model in a genome-wide association study for finding associated loci with bipolar disorder. *Annals of Human Genetics*, 88(3):212–246, 2024.

Sangjoon Kim, Neil Shephard, and Siddhartha Chib. Stochastic volatility: likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, 65(3):361–393, 1998.

Darryl I MacKenzie, James D Nichols, Gideon B Lachman, Sam Droege, J Andrew Royle, and Catherine A Langtimm. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255, 2002.

Radford M Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

Trevor Park and George Casella. The Bayesian LASSO. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

J Andrew Royle. N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115, 2004.

Stan Development Team. Stan reference guide: time-series models: Stochastic volatility models, 2025. URL https://mc-stan.org/docs/stan-users-guide/time-series.html#stochastic-volatility-models. https://mc-stan.org/docs/stan-users-guide/time-series.html#stochastic-volatility-models, Accessed: 2025-02-24.

Sara Van Erp, Daniel L Oberski, and Joris Mulder. Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, 89:31–50, 2019.

Hao Wang. Bayesian graphical LASSO models and efficient posterior computation. *Bayesian Analysis*, 7(4):867 – 886, 2012. doi: 10.1214/12-BA729. URL `https://doi.org/10.1214/12-BA729`.

Benjamin Wee. Comparing MCMC algorithms in stochastic volatility models using simulation based calibration. *arXiv preprint arXiv:2402.12384*, 2024.

A Discussion

Bayesian posterior approximation methods continue to progress. To support the development of new methods, it is useful to have example models and datasets where challenging posteriors arise, complementing standard benchmarking distributions.

However, we note that a complex posterior shape often signals identifiability problems in the model. So in many cases, what may seem like a challenging posterior perhaps need not be. For example, in Gaussian mixture models, multimodality often arises since an ordering must be assigned (in code) to the elements of the fundamentally unordered partition of the data; then we obtain (redundant) modes for each potential labeling of the partition elements. In this case, capturing a single mode among these redundant modes would not only be sufficient, but would in fact be strictly more desirable than capturing the full posterior (cf. the well-known label-switching problem in Markov chain Monte Carlo samplers of mixtures (Jasra et al., 2005)). In some cases, the practitioner might be able to choose an appropriate model that avoids unidentifiability or challenging shapes—such as using a non-centered parameterization for Neal's funnel (Neal, 2003). But in other cases, such as the Gaussian mixture, the best approach might be to address the problem in the approximation software itself, perhaps by focusing on common types of data analyses.

B Additional experiments

Mushroom. The **stochastic volatility** model for stock returns proposed by Kim et al. (1998) can give rise to a mushroom-shaped posterior.

$$
\begin{aligned}
&\text{observing : } y_t \\
&y_t = \epsilon_t e^{\frac{h_t}{2}}, \quad h_{t+1} = \mu + \phi(h_t - \mu) + \delta_t \sigma \\
&h_1 \sim \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{1-\phi^2}}\right), \quad \epsilon_t, \delta_t \overset{iid}{\sim} \text{Normal}(0,1) \quad \text{(likelihood)} \\
&\phi \sim \text{Uniform}(-1,1), \;\; \sigma \sim \text{Cauchy}(0,5), \;\; \mu \sim \text{Cauchy}(0,10) \quad \text{(prior)}
\end{aligned}
\tag{6}
$$

The challenge of this model arises from the posterior behavior when ϕ approaches 1. We next describe why the resulting shape might be seen as mushroom-like, why the mushroom shape arises, and finally why the mushroom shape is challenging.

We first describe what the mushroom shape looks like. When ϕ is very near 1, μ is much heavier tailed than for ϕ substantially smaller than 1. When considering the marginal posterior over ϕ and μ, we can think of the heavy-tailed behavior for ϕ near 1 as corresponding to the hat of the mushroom and the region with ϕ substantially smaller than 1 as corresponding to the stem of the mushroom. We provide a rough illustration in fig. 2. For this model, it is difficult to access the unnormalized posterior marginal density in ϕ and μ due to the need to numerically integrate out all other parameters. Therefore, to create fig. 2, we show samples from Stan (Carpenter et al., 2017). For ϕ near 1, we can see some indication of the widening behavior of μ in the plot.

Next we describe why the mushroom shape arises. First, observe that we can rewrite the formula for h_{t+1} as a function of h_t as follows: $h_{t+1} = (1-\phi)\mu + \phi h_t + \delta_t \sigma$. As $\phi \to 1$, the first term vanishes. So h_{t+1} depends on μ increasingly primarily through h_t (rather than directly). By recursion, we conclude that, as $\phi \to 1$, h_{t+1} comes to depend on μ primarily through h_1. Next, we observe that the dependency of h_1 on μ also becomes weaker as $\phi \to 1$. In particular, as $\phi \to 1$, the

variance of the prior on h_1 diverges. So as $\phi \to 1$, there becomes increasingly little dependence of the data on μ; that is to say, μ is increasingly weakly identified, and the posterior marginal over μ reverts to its (heavy-tailed) prior behavior.

Finally, we discuss why this shape is challenging. The resulting mushroom shape essentially leaves a thin slice (thin across ϕ) is the marginal posterior over ϕ and μ. Thus the stem and hat of the posterior exhibit fundamentally different length scales in terms of the size of the largest sphere that fits into posterior level sets. This challenge is the same one as the one faced by Neal's funnel (Neal, 2003). Reparametrization might help address some of the challenges of sampling in this model (Wee, 2024).

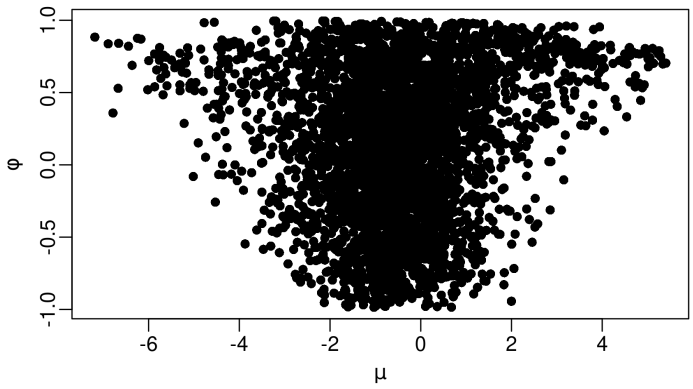

Figure 2: Posterior samples of the stochastic volatility model

C SIMULATION DETAILS

We generate our contour plots in fig. 1 by plotting the unnormalized posteriors in each case.

To generate fig. 1-A, we simulated data from the likelihood in eq. (1) with $\lambda = 30$ and $p = 0.1$. We generated data at 20 locations and 5 repeats at each location i ($\forall i \in \{1, \ldots, 20\}, N_i = 5$).

To generate fig. 1-B, we simulated data from the likelihood in eq. (2) with $\psi = p = 0.1$. We generated data at 100 locations and 8 repeats per location. $n_i = 1, \ldots, 8$ for all i ($\forall i \in \{1, \ldots, 100\}, N_i = 8$).

To generate fig. 1-C, we use the following model where β and x_i are two-dimensional, respectively.

$$\begin{aligned} &\text{observing: } (x_i, y_i) \\ &y_i \overset{ind}{\sim} N(x_i^\top \beta, 1) \quad \text{(likelihood)} \\ &\beta \sim N(0, \sigma^2 I_2) \quad \text{(prior)} \end{aligned} \tag{7}$$

Here, I_2 is the identity matrix of dimension 2. Further, we take the x_i's generated from a normal with high covariance; in particular, we simulate the x_i's from a bivariate normal with variance components on the diagonal equal to 1 and off-diagonal covariance components equal to -0.995. We choose $\beta = (-10, 10)$. And we choose a large σ (100) so that the prior is not very informative. Then the two β's are only weakly identified up to their difference.

To generate fig. 1-D, we simulated 10 data points with $\beta = (0, 0)$ from the likelihood in eq. (3). We drew the x_i values i.i.d. uniformly in $[-2, 2]$.

To generate fig. 1-E, we simulated we simulated 5 data points from the likelihood in eq. (4) with $\beta = 5$ and a known $\sigma = 8$. We let all $x_i = 1$, so the task was Gaussian mean estimation.

To generate fig. 1-F, we simulated 50 data points from the likelihood in eq. (5) with $\beta = \Pi = 0.1$, $\Sigma = I$, and instruments $z_i \overset{iid}{\sim} \text{Bernoulli}(0.75)$.

To generate fig. 2, we simulated from the model in eq. (6) with $\mu = -1.02, \phi = -0.95, \sigma = 0.1$. We generated data for $t = 1, \ldots, 5$. The implementation in Stan was taken from the Stan reference guide (Stan Development Team, 2025).