Bayesian nonparametric sampling and stochastic simulation of probabilistic models [1]

Chris Holmes

Professor of Statistics
University of Oxford

Scientific Director for Health
The Alan Turing Institute

SeRC annual meeting
May 2019

[1] Joint work with Stephen Walker, Simon Lyddon, and Edwin Fong

Overview

- The changing nature of health data science
- The resulting challenges for the information sciences
  - statistics
  - machine learning
  - inductive logic
- Using Approximate Models and Computational decision theory *at scale*
  - formal methods for robust, scalable, decision analysis
- Concluding remarks

AI and Health

- The UK is making significant investment into “AI”, in part following a belief that AI is set to transform medicine
  - By “AI” we take to mean computational statistics and machine learning,
- Alan Turing Institute – the UK’s national institute for data science and AI
  - 13 University Partners
  - 320+ Turing Fellows & Research Fellows
  - 45+ PhD students (plus 20+ on a short-term enrichment placement)
  - 30 Interns (12 week programme)
  - **20+ Research Software Engineers/Data Scientists**

Why the interest in AI?

Changing world

- Data generation and data acquisition is no longer the bottleneck
- Driven by advances in digital measurement technologies
  - Genomes; medical images; electronic health records; wearables; social media
- And resources to capture data in BioBanks and longitudinal cohorts
  - UK Biobank on 500,000 individuals:
    - 100,000 brain images,
    - 100,000 MRI body scans,
    - 100,000 “fitbit” data,
    - all individuals genotyped on 3M marker array, ...
- Coupled to increasing raw computing power (GPUs) that facilitate compute hungry algorithms
- High level (governmental) recognition of data as a resource
- And connectivity across data environments

Impact on Statistics and Information Sciences

- The new era is having a major disruptive effect on Statistics and machine learning
- Driven by the desire to combine information from multiple data-modalities at population scales
- Increasingly fanciful to think that we have anything close to a "true model"
- We need principled approaches to learning from data, that are robust to modelling assumptions
- We need methods that can scale and make use of modern compute environments
- We need to be aware of the consequences of complex studies on reproducibility of research

Reproducible Research

- Reproducible research is fundamental to the scientific method
- The onus should be on me to provide you with the tools to refute my research findings
  - Popper uses falsification as a criterion of demarcation to draw a sharp line between those theories that are scientific and those that are unscientific – Wikipedia
- Yet the increasing complexity of modern (e)science is challenging in this regard
- As a community we need to commit to, and work hard, to ensure our work is reproducible
- This requires a cultural shift from us and planning from day one!
  - There are tools to assist: GitHub; Code capsules; Notebooks

Preamble

- Statistics is the scientific study of uncertainty
  - uncertainty is quantified in units of probability
- Statistics is about being precise about imprecision
  - Bayesian statistics is perhaps more explicit on this matter than other approaches

Foundations of Bayesian inference

- I will present some of our recent work in Bayesian methods that seek to address issues such as robustness to model misspecification and scalability of computational inference (stochastic simulation)
- Bayesian statistics is founded in decision theory and optimal decision making under uncertainty, principally following Savage (1954)
- At the heart of Bayesian inference is the updating rule on parameters of a statistical (probabilistic) model

$$\begin{array}{ccccc} \text{Posterior} & \propto & \text{Prior} & \times & \text{Likelihood} \\ p(\theta \mid x_{1:n}) & \propto & p(\theta) & \times & f_\theta(x_{1:n}) \end{array}$$

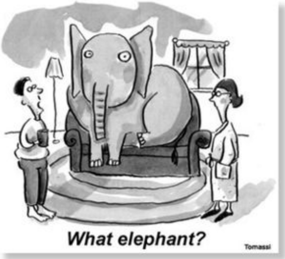
However.....

- Bayesian inference is predicated on the model being true

$$\text{Nature} = f_0(x) = f_\theta(x) \;\; \exists\, \theta \in \Theta$$

  - you have to assume that Nature's true data generating mechanism, $f_0(x)$, is contained under the support of the prior
  - and....

    All of Bayesian statistics is model based



- But increasing $f_0(x)$ is hard to justify or define....how can I define a true generative model over {genomes, medical images, eHRs, . . . }?
- Of course, models are just simply......models.....and it's fanciful to think otherwise, but formally all Bayesian statements of uncertainty are predicated on the model being true

Bayesian Analysis

- Increasingly reliant on approximate methods such as Variational Bayes
- Should we worry about $p(\theta|x)$?

- But if we do just carry on,
  - what does the posterior $p(\theta|x)$ actually represent?
  - should I simply plug $p(\theta|x)$ into decision analysis?

Question: What are we learning about?

- If the model is false then what does the parameter and posterior formally represent?

$$p(\theta|x) \propto f_\theta(x)\, p(\theta)$$

- As more and more data arrives, for most regular {models, priors} the posterior will concentrate around a point, θ_0,

$$p(\theta|x) \xrightarrow[n\to\infty]{} \delta_{\theta_0}$$

  that maximises the expected log-likelihood function (you can think of the negative log-likelihood as a loss function or error function)

$$\theta_0 = \arg\max_\theta \int \log f_\theta(x) dF_0(x)$$

  for data arising from $x \sim F_0(x)$

What are we learning about?

- You can think of this as the optimal value under an infinite sample size

$$\begin{aligned} x_i &\sim F_0(x) \\ \theta_0 &= \arg\max_{\theta} \sum_{i=1}^{\infty} \log f_\theta(x_i) \end{aligned}$$

- θ_0 is the value that minimizes the Kullback-Leibler divergence from the model to Nature's true unknown sampling distribution, $F_0(x)$, irrespective of whether the model is misspecified or not

- θ_0 is the target of inference and the prior $p(\theta)$ should be seen as specifying beliefs in this context
  - so the prior is no longer on the "true value" but rather on the point where the posterior will concentrate as you obtain more data

Updating with incorrect models.....a fairy tale...

Consider the following (imaginary) thought experiment....

- Imagine that you've chosen a parametric (generative) probabilistic model, $f_\theta(x)$, specified a prior $\pi(\theta)$, and obtained a data set $\{x_i\}_{i=1}^n$
- You're just about to update your model
- That is, you are just about to call an algorithm in Stan or WinBUGS (or Variational Bayes) to calculate the posterior

$$p(\theta|\boldsymbol{x}) \propto \prod_i f_\theta(x_i)\, p(\theta)$$

- When someone offers you an exact emulator (computer model) of Nature!
- How would you proceed?

A thought experiment

- With an exact emulator of Nature, $F_0(x)$, you can simply request an infinite sample size, $\tilde{\boldsymbol{x}} = \{\tilde{x}\}_{1:\infty}$ for

  $$\tilde{x}_i \sim F_0(x)$$

  and then update to obtain

  $$p(\theta|\{\tilde{x}\}_{1:\infty}) = \prod_{i=1}^{\infty} f_\theta(\tilde{x}_i)\,\pi(\theta)$$

  and with an infinite sample size, and prior of sufficient support, all uncertainty is removed,

  $$\begin{array}{rcl} p(\theta|\{\tilde{x}\}_{1:\infty}) & \rightarrow & \delta_{\theta_0} \\ \theta_0 & = & \arg\max_\theta \sum_i \log f_\theta(\tilde{x}_i) \\ \tilde{x}_i & \sim & F_0(x) \end{array}$$

- Of course, this assumes that you know F_0!

Bayesian Nonparametric Learning

- In the above story, posterior uncertainty in the optimal value θ_0 can be seen to flow directly from uncertainty in F_0
  - as knowing $F_0(x)$ identifies the target θ_0
  - And θ_0 is the value that minimizes the KL divergence from the model to Nature's $F_0(x)$, **irrespective of whether the model is true**
- F_0 is unknown, but being "Bayesian" we can place a prior directly on it, $p(F)$, for $F \in \mathcal{F}$, that should reflect our honest uncertainty
  - So place a prior directly on the space of distribution functions F rather than θ and learn about θ_0 that way
  - This is the essence of Bayesian Nonparametric Learning – using a Bayesian NP model, $p(F|x)$, to train a parametric model $f_\theta(x)$

Bayesian Nonparametric Learning

- So if we can simulate a nonparametric distribution $F \sim p(F|x)$, we can then use this to train our model

- We will use Bayesian nonparametrics to learn about $p(F|x)$, from which we then learn θ

- Given a sample, $F^{(i)} \sim p(F|\boldsymbol{x})$, then for each $F^{(i)}$ there is no uncertainty in the corresponding optimal parameter values of the model (that minimizes KL to $F^{(i)}$)

$$\theta^{(i)} = \arg\max_{\theta} \int \log f_{\theta}(x) dF^{(i)}(x)$$

- Repeating the operation provides a bag of Monte Carlo samples, $\{\theta^{(1)}, \ldots, \theta^{(T)}\}$, then characterises the marginal posterior distribution $\tilde{p}(\theta|x)$

Computational Algorithm: using nonparametric models to train parametric models

The above leads to the following sampling algorithm for θ:

Assuming $F(x)$ has finite support on the data $\{\tilde{x}\}_j$ on $\mathcal{X}$ then

1. Draw $F \sim p(F|x_{1:n})$
2. Set $\theta(F) = \arg\max_{\theta \in \Theta} \sum_i w_i \log f_\theta(\tilde{x}_i)$

   Repeat

where $w_i = f^{(NP)}(\tilde{x}_i)$, and $\sum_i w_i = 1$

- If the draws of F can be made independently, then samples of θ's can be drawn in parallel using the NP re-weighted objective functions
- If we use a Dirichlet Process DP to model F then the weights w are simply uniform on the simplex
- We replace traditional MCMC with optimization of randomized objective functions

The Bayesian posterior bootstrap

- The case $F^{(i)} \sim DP(F|\boldsymbol{x}, c = 0, G)$ is known as the **Bayesian bootstrap**
- And the fitting of the resulting $\theta^{(i)}$ via

$$\theta^{(i)} = \arg\max_{\theta} \sum_{j} w_j \log f_\theta(x_j)$$

  with $\boldsymbol{w}^{(i)} \sim Uniform(n)$ (Newton & Rafetry, 1984)
- This is simply a randomly re-weighted data maximisation at each step
  - That is, fit the model to a weighted representation of the data
  - Where the weights are stochastic
- This captures the uncertainty in the model fit arising from the finite sample – in a precise manner

Comparison to Efron's Bootstrap

Given dataset $x_{1:n} = (x_1, \dots, x_n)$

Let $\hat{F}_n$ denote the empirical distribution function:

$$\hat{F}_n(\cdot) = \sum_{i=1}^{n} \delta_{x_i}(\cdot)$$

which has atomic support at the data

And utility function $u(\theta, x)$, e.g. $u(\theta, x) = \log f_\theta(x)$

Efron's Bootstrap	Bayesian Bootstrap
For $i = 1, \dots, B$:	For $i = 1, \dots, B$:
	• $F^{(i)} \sim \mathrm{DP}(F; \widehat{F}_n, c = 0)$
• $x^{(i)}_{1:n} \sim \widehat{F}_n$	• $\boldsymbol{w}^{(i)} \sim Dir(1, 1, 1..., 1)$
• $\theta^{(i)}_{\text{Boot}} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log f_\theta(x^{(i)}_i)$	• $\theta^{(i)}_{\text{Bayes}} = \arg\max_{\theta \in \Theta} \sum_{j=1}^{n} w^{(i)}_j \log f_\theta(x^{(i)}_i)$

Posterior bootstrap asymptotics

Theorem (Lyddon, Holmes & Walker (2018))

Let $\tilde{\theta}$ be a NPL sample given a loss function ℓ, such as $\ell = -\log f_{\cdot}(x)$, and n observations $x_{1:n}$. Then under regularity conditions, for any Borel set $A \subset \mathbb{R}^d$, as $n \to \infty$ we have

$$P_{LL}\left\{n^{1/2}\left(\tilde{\theta}_n - \hat{\theta}_n\right) \in A \mid x_{1:n}\right\} \to P(z \in A)$$

a.s. $x_{1:\infty}$, where $z \sim N_d\{0, J^{-1}IJ^{-1}\}$ with

$$V = \int \nabla\ell(\theta, x)\nabla\ell(\theta, x)^T dF_0(x) \quad \textit{and} \quad J = \int \nabla^2\ell(\theta, x) dF_0(x)$$

where ∇ is the gradient operator with respect to θ, and

$$\hat{\theta}_n = \arg\min_{\theta} n^{-1}\sum_{i=1}^{n}\ell(\theta, x_i)$$

.

Asymptotics - interpretation

- Misspecified Bayes posterior has scaled covariance matrix $\Sigma_{\text{Bayes}} = J^{-1}$
- Misspecified MLE has scaled covariance matrix $\Sigma_{\text{MLE}} = J^{-1}VJ^{-1}$
  - same as the NPL posterior
  - when the model is true $V = J$
- $J^{-1}VJ^{-1}$ is referred to as the **sandwich covariance matrix** in the robust statistics literature, for example Royall & Tsou (2003)
- Müller (2013) showed that, under regularity, the sandwich covariance matrix leads to decisions with **lower frequentist risk** than misspecified Bayes

NP-Learning is predictively superior to Bayes

- A natural metric for assessing a posterior distribution is the **predictive risk**, defined as the expected Kullback-Leibler divergence of the posterior predictive to F_0

- We say predictive p_1 **asymptotically dominates** p_2 if for all distributions q there exists a non-negative and possibly positive real-valued functional $K(q)$ such that for $x_{1:n} \sim q$ we have:

$$\mathbb{E}_q d_{\mathsf{KL}}(q(\cdot), p_2(\cdot \mid x_{1:n})) - \mathbb{E}_q d_{\mathsf{KL}}(q(\cdot), p_1(\cdot \mid x_{1:n})) = K(q) + o(n^{-1})$$

Theorem (Lyddon, Walker & Holmes (2018))

The posterior predictive of NP-learning with $c = 0$ asymptotically dominates the standard Bayesian posterior predictive

How to combine with prior information

- So far it's not very Bayesian as there's no prior
- We would like to incorporate prior information into the learning
  - For example, from a mathematical model of the process, or a previous study
- To do so we make use of synthetic data

Priors through synthetic-data

- To do this we rely on the use of synthetic data drawn from a prior sample predictive

$$\begin{array}{ccc} \theta' & \sim & p(\theta) \\ x^*_{1:T} & \sim_{iid} & f_{\theta'}(x) \end{array}$$

  where $p(\theta)$ is prior information (or approximate data source)

- Then combine the synthetic data with the actual data for the update with a draw $F \sim MDP(F|c, x, x^*)$ (Antoniak, 1974) where c is equivalent to an effective sample size in $p(\theta)$, with

$$\tilde{\theta}^{(i)} = \arg\max_{\theta} \left[n \sum_{j=1}^{n} w_j^{(i)} \log f_\theta(x_j) + c \sum_{j=n+1}^{n+T} w_j^{(i)} \log f_\theta(x^*_{j-n}) \right]$$

  with randomized weights $w_{1:n+T}$, where $(\frac{c}{c+n})$ characterises the relative influence of the prior data

- Prior specification through synthetic data is well known in parametric (conjucate) models: Beta-Binomial (Laplace) and Linear regression

E.g: Posterior bootstrap samples for VB inference

- Variational Bayes cover are an important class of approximate models designed for computational tractability and scalable inference

- While prediction maybe good, it is known that inference on parameters is not to be trusted due to (artificial) conditional indepedence structures engineered into the model

  - VB builds an approximation by minimizing KL divergence to an incorrect model. Why not minimize KL to the correct distribution?

- We can use NPL to correct for the known model misspecification

  - Take a fast, approximate, update for $p(\theta|x) \propto f_\theta(x)p(\theta)$, using a Variational Bayes model, $f_{\theta^*}(x)$

  - Use the VB posterior $p(\theta^*|x)$ as a centering model under a nonparametric prior

  - Use a posterior bootstrap to draw samples, $\theta^{(j)}$, that combine information in the data and information in the prior model

Algorithm 1: The Variational Bayes - Posterior Bootstrap

```
Data:  Dataset $x_{1:n} = (x_1, \ldots, x_n)$.
Approximate VB posterior $q(\theta|x_{1:n})$, concentration parameter $c$,
centering model $f_\theta(x)$.
Number of centering model samples $T$.
begin
    for $i = 1, \ldots, B$ do
        Draw VB posterior model parameter $\theta^{(i)*} \sim q(\theta^*|x_{1:n})$;
        Draw posterior synthetic-data $x^{(i)}_{(n+1):(n+T)} \overset{iid}{\sim} f_{\theta^{(i)*}}(x)$;
        Generate weights $(w^{(i)}_1, \ldots, w^{(i)}_n, w^{(i)}_{n+1}, \ldots, w^{(i)}_{n+T}) \sim$
          Dirichlet$(1, \ldots, 1, c/T, \ldots, c/T)$;
        Compute parameter update
              $\tilde{\theta}^{(i)} =$
          $\arg\max_\theta \left\{ \sum_{j=1}^{n} w^{(i)}_j \log f_\theta(x_j) + \sum_{j=1}^{T} w^{(i)}_{n+j} \log f_\theta(x^{(i)}_{n+j}) \right\}$;
    end
    Return NP posterior sample $\{\tilde{\theta}^{(i)}\}_{i=1}^{B}$.
end
```

VB and EP bivariate Gaussian example from Bishop's book

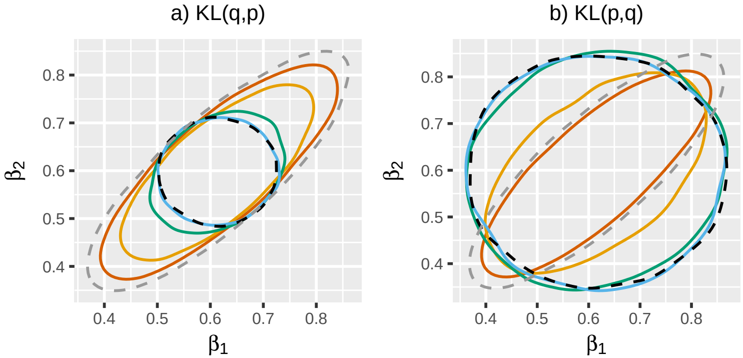


Figure: 95% probability contour for a bivariate Gaussian, comparing **VB-NPL (black dashed)** with NP-Learning for decreasing $c \in \{10^4, 10^3, 10^2, 1\}$

- **Correlation structure** of posterior, lost in mean field approximation, is **recovered** by NP-learning.
- Run-time: 20s for VB-NPL, and 30 mins for MCMC, 1 million samples

Fast, robust, Bayesian logistic regression

- Consider the Bayes logistic regression model

$$\log\left(\frac{p(y=1|x)}{p(y=0|x)}\right) = x\beta$$

- Two challenges for a conventional Bayesian update:
  - It assumes that the model is true – and all interpretation of posterior intervals are predicated on this
  - We have to use (Polya-Gamma) MCMC with a burn-in, thinning, and convergence diagnostics to draw dependent samples approximately $\theta \sim p(\theta|x)$

- Using NP-learning we can draw iid samples in parallel $\tilde{\theta} \sim \tilde{p}(\theta|x)$

Statlog example: german credit data

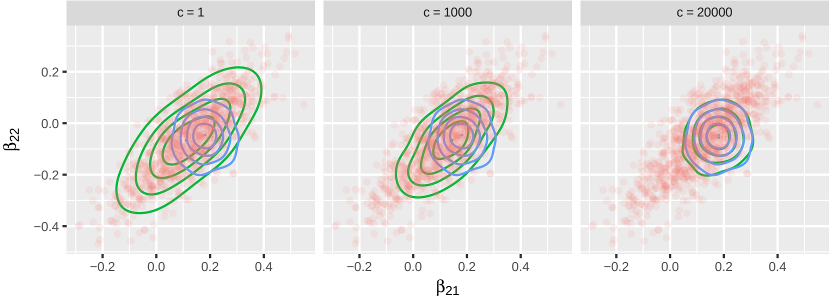


Figure: Posterior contour plot for β_{22} vs β_{21}, for NPL (green) and **VB (blue)**, for three different values of the concentration parameter c. Scatter plot is a sample from a Bayesian logistic posterior (red) via Polya-Gamma scheme.

- The posterior bootstrap corrects the model to exact coverage
- Run-time 1 million samples: **20 seconds for NPL** using AWS, and **30 mins for MCMC**, 95 times speed up
- NPL: no burn-in, no thinning, no need for convergence diagnostics

Gaussian Mixture Models

Consider a Bayesian model for K-component diagonal GMM with non-conjugate prior is:

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

For NPL, we are interested in model fitting, so our loss function is simply the negative log-likelihood:

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

We use an example in 2-d with $K = 3$

Gaussian Mixture Model

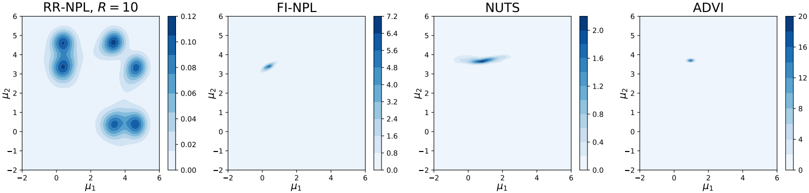


Figure: Posterior KDE of (μ_1, μ_2) in $K=3$ toy GMM problem

Bayes-NPL captures all of the known symmetries in the multi-modal posterior model space at a fraction of the run-time

Optimisation of randomised objective functions is much more efficient than Markov chain Monte Carlo simulation

Conclusions – Bayesian Nonparametric Learning

- Modern applications can be disruptive for traditional statistical methods
- NP-Learning is motivated by large scale applications that do not rely on notions of true models
- It's important to note that **Bayes NP-Learning is not an approximation to the conventional Bayesian posterior**, and

$$\tilde{p}_{NPL}(\theta|\boldsymbol{x}) \neq p(\theta|\boldsymbol{x})$$

- They are targeting the same parameter, $\theta_0 = \arg\min_\theta \mathrm{KL}(F_\theta||F_0)$, but they are conditioning on different states of knowledge
    - in particular conventional Bayes assumes that the model is true – and learns at a rate that is defined by this
- NPL is scalable and trivially parallel on modern compute architectures
    - provides theoretical robustness over conventional Bayes

Thank you!

References

This talk is built on recent work on scalable methods for approximate Bayesian models

- Bissiri, Holmes & Walker, “General Bayesian Updating” (2016) *JRSS-B*
- Lyddon, Holmes & Walker, “Generalized Bayesian Updating and the Loss-Likelihood Bootstrap” (2018) *Biometrika*
- Lyddon, Walker & Holmes, “Nonparametric learning from Bayesian models with randomized objective functions” (2018) *NeurIPS*
- Fong, Lyddon, & Holmes, “Scalable Nonparametric Sampling from Multimodal Posteriors with the Posterior Bootstrap” (2019), to appear, *ICML*