Bayesian learning at scale with approximate models

Chris Holmes,
Professor of Biostatistics,
University of Oxford

CIRM
October 2018

Overview

- Foundations of Bayesian inference
  - with an emphasis on approximate models
- Why Bayesian analysis is challenged by new world data
- Computational decision theory and approximate models
  - formal methods for robust, scalable, decision analysis
- Concluding remarks

Foundations of Bayesian inference

- Bayesian statistics is founded in decision theory and optimal decision making under uncertainty, principally following Savage (1954)[1]
- Savage postulated a set of axioms (Savage 1954), building on the work of Good and others, that motivated the adoption of Bayesian updating as a way to acheive optimal (rationale and coherent) decision making
  - DeGroot (1970); reviewed in Fishburn (1986), Bernardo and Smith (1994)

[1] although note that Cox (R.T.) axioms underpin Bayes Theorem as an extension of logic (Cox, 1946) – predominantly referred to in machine learning (Jaynes, 2003)

A Savage World

- Consider a (terribly dull) World in which everyone is seen to behave rationally
  - never preferring action A to B when it is expected that A leads to a worse outcome
- and coherently
  - in that two individuals with the same starting beliefs, on seeing the same data, arrive at the same conclusion
- Then it is as if they are updating beliefs using probability calculus

- That is, their World can be perfectly modelled using a computer and Bayes theorem
  - In the computer model all uncertainty, on both random components x and unknown constants θ, is specified via a joint probability model

    $$p(x, \theta) = f_\theta(x)p(\theta)$$

    where, in most cases, the choice of "likelihood" (sampling distribution), $f(\cdot)$, is at least as subjective as the prior $p(\theta)$
  - Optimal rationale decisions are taken by maximising expected rewards $R(a, \theta)$ on potential actions a given current state of partial information in $\{x, p(\theta)\}$

    $$\widehat{\text{Action}} = \arg\max_a \int R(a, \theta)p(\theta|x)d\theta$$

- In this way, Bayesian statistics provides a prescriptive, operational, approach to optimal decision making

Key features of Bayesian statistics

Perhaps the most important

All of Bayesian statistics is model based

This has a number of attractions:

- You define a joint probability model to express uncertainty on all relevant unknowns {data, parameters} treated interchangeably
  - using generative model structures
  - often involving hierarchical model building
- Update subjective beliefs *via* the model, often using Bayes theorem
- Bayes separates out the model building (data analysis) from the decision making
- Formally, if the axioms hold, then Bayes uniquely identifies the value of information
  - That is, if the model is "true", then Bayes provides optimal information processing

However.....

- Bayesian inference is predicated on the model being true

$$f_0(x) = f_\theta(x) \;\; \exists\, \theta \in \Theta$$

  - 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....
- Terminology: M-closed refers to when the true sampling distribution is contained within the model space; and M-open when no model is true

Bayesian Analysis

- Clearly a joint model, $f_0(x)$, representing true subjective beliefs is unobtainable particularly for many modern applications
- Should we worry?

- 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? $\widehat{\text{Action}} = \arg\max_a \int R(a,\theta)p(\theta|x)d\theta$?
- It seems clear that sensitivity to modelling assumptions is important and contextual

(Trivial) Example

Suppose that data, $\{x_1, \ldots, x_n\}$, arise from an exponential distribution,

$$x_i \sim \exp(\lambda)$$

Yet the statistician fits a normal model, assuming,

$$x_i \sim N(\mu, \sigma^2)$$

with μ, σ unknown

- Should we (you) be concerned?

Context

- Maybe yes, if the outcome of analysis depends on predicting a future event $Pr(x \in [a, b])$
- Maybe no, if the outcome of analysis depends on inferring $E[X]$ (and n is large)

The point being is that it seems inappropriate to separate out the issue of misspecification (approximation) from the context, use and rationale of the analysis

Models as metaphors

- Of course, models are just simply.....models.....and it's fanciful to think otherwise
- It's unsurprising that others have given formal consideration to this
  - G.E.P. Box – "All models are wrong....."
    - synopsis: update like a Bayesian, critique like a frequentist
  - J. Berger (1980's) – see Berger (1994) for review
    - Sensitivity analysis to prior specification
  - M. Goldstein (2004 – book) – linear Bayes
    - express beliefs on expectation (summary statistic) – a bit ABC like
  - L. P. Hansen and T. J. Sargent (2008) – econometricians and robust control theory
    - sensitivity analysis in the posterior
    - building upon P. Whittle – signal contamination by a malevolent Nature
    - Reviewed, with extensions, in Watson & Holmes (2016) – Statistical Science

General Bayesian Updating

- We have been working on formal methods for Bayesian updating without the assumption that the models are true
  - Bissiri, Holmes & Walker "General Bayesian Updating" (2016) *JRSS-b*
  - Holmes & Walker "Assigning a value to a power likelihood in a general Bayesian model" (2017) *Biometrika*
- And today's talk on
  - Lyddon, Walker & Holmes "Nonparametric learning from Bayesian models with randomized objective functions" (2018) *NIPS*
  - Lyddon, Holmes & Walker "Generalized Bayesian Updating and the Loss-Likelihood Bootstrap" (2018) *archive*

Prediction versus Inference

- If we are purely interested in prediction then issues of model misspecification are easier to deal with, as we can cross calibrate against performance metrics
- But a large part of Statistics is concerned with the isolation and estimation of parameters (that are never observed)
- In this latter task model misspecification is more subtle and more challenging

Q1: What are we learning about?

- To begin, if the model is false then what does the parameter formally represent?
- A famous formula gives

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

- 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

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

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

What are we learning about?

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

$$\begin{aligned} KL(F_\theta || F_0) &= \text{Entropy} + E_{F_0}[-\log f_\theta(x)] \\ \int \log f_\theta(x) dF_0(x) &\approx \sum_{i=1}^{\infty} \log f_\theta(x_i) \\ x_i &\sim F_0(x) \end{aligned}$$

  for Nature's F_0

- θ_0 is the target of inference and the prior $p(\theta)$ should be seen as specifying beliefs in this context

General Bayesian Updating

- In Bissiri et al (2016) we showed that a coherent generalisation of Bayesian updating was available via the posterior

$$\log p^{(GB)}(\theta|x) = C + \lambda \log f_\theta(x) + \log p(\theta)$$

$$p^{(GB)}(\theta|x) \quad \propto \quad \exp[-\lambda R(x,\theta)] \times p(\theta)$$

- This involves replacing the likelihood function with a loss-likelihood $\exp[-\lambda R(x,\theta)]$
- Treating $R(x,\theta) = -\log f_\theta(x)$ as a loss-function (risk, or negative utility) targeting the parameter value of interest, θ_0
  - $-\log f_\theta(x)$ is known as the self-information loss
  - But other loss functions can be used depending on the decision analysis
  - This is not an approximation, or pseudo-Bayes, but a valid subjective representations of beliefs

General Bayesian Updating

- This showed that we can derive a valid Bayesian update in the absence of a "true model", using loss-functions tuned to the decision analysis
  - e.g. this allows for valid Bayesian analysis of log-linear proportional hazard models in survival analysis
- However, once you ackowledge that the true sampling distribution lies outside of the model space you introduce a scale parameter $\lambda \geq 0$ that quantifies the relative information in the data
  - One of the beautiful aspects of conventional (M-closed) Bayesian inference is that this value is precisely specified (Zellner 1988)

Quantifying the value of data

- We will concentrate on the use of self-information loss

$$p^{(GB)}(\theta|x) \quad \propto \quad \exp[-\lambda R(x,\theta)] \times p(\theta)$$

  with $R(x,\theta) = -\log f_\theta(x)$

- The information constant λ determines the extent of the update in how far (in measure space) the posterior will move away from the prior, with $\lambda \to 0$ the posterior doesn't change, and $\lambda \to \infty$ all of the posterior mass accumulates around the MLE
- Learning the learning rate is a non-trivial problem
- Recently we have been exploring Bayesian nonparametric approaches to this

Nonparametric learning for parametric models: a new approach to Bayesian updating

- Imagine that you've chosen a parametric (generative) model, $f_\theta(x)$, that you're about to update using data set $\{x_i\}_{i=1}^n$
- Suppose now that I provide you with an infinite sample $\{\tilde{x}_i\}_{i=1}^\infty$ from Nature's $F_0(x)$

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

- With an infinite sample you would ignore your data and use $f_{\tilde{\theta}}(x)$ where you plug in the value

$$\tilde{\theta} = \arg\max_\theta \sum_i \log f_\theta(\tilde{x}_i)$$

  as you have the perfect update and all uncertainty in θ is removed
- Of course, this assumes that you know F_0

Nonparametric Learning

- Uncertainty in the optimal value θ_0 can be seen to flow directly from uncertainty in F_0
- F_0 is unknown, but being "Bayesian" we can place a prior on it, $p(F)$, for $F \in \mathcal{F}$, that should reflect our honest uncertainty
- Typically the prior should have broad support unless we have special knowledge to hand, which is a problem with a parametric modelling approach that only supports a family of distribution functions indexed by a finite-dimensional parameter
  - parametric Bayes assumes $F_0 \in \mathcal{F}_\theta$ but it learns about

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

- Fortunately there is a whole field of Bayesian nonparametrics for learning F, for learning θ

- Once a prior for F is chosen, the correct way to propagate uncertainty about θ comes naturally from the posterior distribution for the law $\mathcal{L}[\theta(F)|x_{1:n}]$, via $\mathcal{L}[F|x_{1:n}]$, where

$$\theta(F) = \arg\max_{\theta\in\Theta} \int \log f_\theta(x) dF(x)$$

- The posterior for the parameter is then captured in the marginal by treating F as a latent auxilliary probability measure,

$$\tilde{p}(\theta \mid x_{1:n}) = \int_{\mathcal{F}} p(\theta, dF \mid x_{1:n}) = \int_{\mathcal{F}} p(\theta \mid F) p(dF \mid x_{1:n}), \qquad (1)$$

  where $p(\theta|F) = \delta_{\theta(F)}$ assigns probability 1 to $\theta = \theta(F)$
- We use $\tilde{p}$ to denote the NP update to distinguish it from the conventional Bayesian posterior $p(\theta|x_{1:n}) \propto p(\theta) \prod_{i=1}^n f_\theta(x_i)$
- In general the nonparametric posterior $\tilde{p}(\theta \mid x_{1:n})$ will be different to the standard Bayesian update as they are conditioning on different states of prior knowledge. In particular, as stated above, $p(\theta|x_{1:n})$ assumes artificially that $F_0 \in \mathcal{F}_\Theta$.

Computational Algorithm: using nonparametric models to train parametric models

The above leads to the following sampling algorithm for θ

1. Draw $F \sim p(F|x_{1:n})$ using a nonparametric update
2. Set $\theta(F) = \arg\max_{\theta\in\Theta} \int \log f_\theta(x) dF(x)$

   Repeat

Note: if $F(x)$ has finite support $\{\tilde{x}\}_j$ on $\mathcal{X}$ then this becomes

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

Weighted Likelihood Bootstrap

- Newton & Rafetry (1994) introduced the "weighted likelihood bootstrap" that has precisely this form, with randomized weights on the data samples,

$$w \sim Dir(1, 1, ...1)$$

$$\theta^{(i)} = \arg\max \sum_i w_i \log f_\theta(x_i)$$

- which can be considered as drawing F from a degenerate Dirichlet Process (Bayesian Bootstrap, Rubin 1974)

- the WLB was presented as an approximation to a Bayesian model under M-closed
  - the paper was not particularly well received, in part as (a) it coincided with the arrival of MCMC that could provide exact approximation, and (b) it doesn't allow for inclusion of prior information on $p(\theta)$

- However, we would argue that the WLB provides an exact solution to a general Bayesian update in M-open, learning about θ_0 in the absence of prior information and a true model space

Posterior Bootstrap

- We would like to incorporate prior information via a generative model $f_\theta(x)$ into the nonparametric draw $F \sim p(F|x)$
- Note: the simple idea of regularization with the prior doesn't work

$$\theta^{(j)} = \arg\max_\theta \left[\sum_{i=1}^{n} w_i \log f_\theta(x_i) + \log p(\theta) \right]$$

consider $n = 0$

Priors through synthetic-data

- To do this we rely on the use of synthetic samples drawn from

$$\begin{aligned} \theta &\sim p(\theta) \\ x^* &\sim_{iid} f_\theta(x) \end{aligned}$$

  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

$$\tilde{\theta} = \arg\max_\theta \left[n \sum_i w_i \log f_\theta(x_i) + c \sum_j w_j \log f_\theta(x_j^*) \right]$$

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

Data-centric priors via synthetic observations

- To recap, we represent the prior, $p(\theta)$, via (infinite) sample sets of synthetic observations,

$$\begin{aligned} \theta^{(j)} &\sim p(\theta) \\ \widetilde{x}_i^{(j)} &\sim f_{\theta^{(j)}}(x) \quad \text{for } i = 1, ... \end{aligned}$$

  and construct the synthetic data set $\widetilde{X}^{(j)} = \{\widetilde{x}_i^{(j)}\}_{i=1}^{\infty}$

- In practice for each $\theta^{(j)}$ we can sample $\{\widetilde{x}_i\}_{i=1}^{T}$ synthetic observations for T large

- We can then combine the observed samples $\{x_i\}_{i=1}^{n}$ with the pseudo-samples $\{\widetilde{x}_i\}_{i=1}^{T}$

- We can then Bayesian bootstrap the combined data $\{x_1, \ldots, x_n, \widetilde{x}_1^{(j)}, \ldots, \widetilde{x}_T^{(j)}\}$, redrawing synthetic data at each step, and putting weight n on the real obervations and weight c on the synthetic data

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

E.g: Posterior bootstrap samples for VB inference on parameters

- Variational Bayes cover are an important class of approximate models designed for computational tractability and scalable inference
  - as you saw earlier today!
- While prediction maybe good, it is known that inference on parameters is not to be trusted due to (artificial) conditional independence 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?
- Importance sampling correction can show huge variance and can break down especially in high-dimensional models
- We can explore the use of nonparametric correction, via the posterior bootstrap, for the VB approximation

Posterior bootstrap VB correction

- 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
  - Draw synthetic-samples under the VB approximation, $\theta^{(i)*} \sim p(\theta^*|x)$ and $\tilde{\boldsymbol{x}} \sim f_{\theta^{(i)*}}(x)$
  - Use a posterior bootstrap to draw samples, $\tilde{\theta}^{(j)}$, that provide a NP correction, combining information in the data and information in the prior model (through synthetic samples)

Algorithm 1: The Variational Bayes - Posterior Bootstrap

Data: Dataset $x_{1:n} = (x_1, \ldots, x_n)$.
Approximate VB posterior $p(\theta^*|x_{1:n})$, 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 $\theta^{(i)*} \sim p(\theta^*|x_{1:n})$;

    Draw posterior synthetic-sample $x^{(i)}_{(n+1):(n+T)} \overset{iid}{\sim} f_{\theta^{(i)*}}(x)$;

    Generate weights $(w_1^{(i)}, \ldots, w_n^{(i)}, w_{n+1}^{(i)}, \ldots, w_{n+T}^{(i)}) \sim$ Dirichlet$(1, \ldots, 1, c/T, \ldots, c/T)$;

    Compute parameter update

$$\tilde{\theta}^{(i)} = \arg\max_\theta \left\{ \sum_{j=1}^{n} w_j^{(i)} \log f_\theta(x_j) + \sum_{j=1}^{T} w_{n+j}^{(i)} \log f_\theta(x_{n+j}^{(i)}) \right\};$$

  end

  Return NP posterior sample $\{\tilde{\theta}^{(i)}\}_{i=1}^{B}$.

end

VB and EP bivariate Gaussian example from Bishop's book

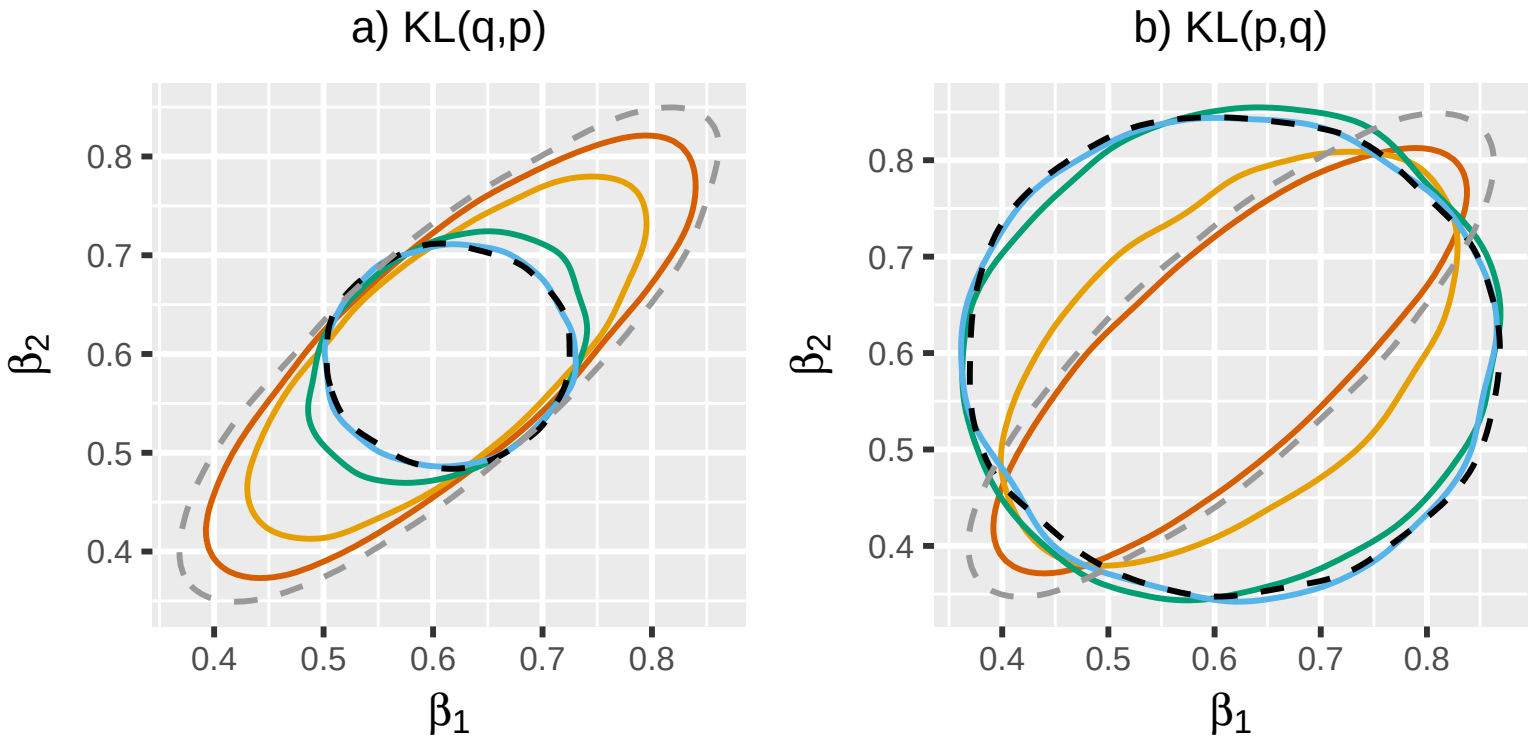

Figure: 95% probability contour for a bivariate Gaussian, comparing VB-NPL with $c \in \{1, 10^2, 10^3, 10^4\}$ (red, orange, green, blue respectively) to Bayes posterior (grey dashed line) and a VB approximation (black dashed line).

- The posterior bootstrap provides a one-shot correction of the VB model to provide exact coverage – it's clear that IS would be awful
- The VB model regularizes and smooths for small data sets

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 and convergence diagnostics to draw dependent samples approximately $\theta \sim p(\theta|x)$
- Using Nonparametric learning we can draw iid samples in parallel $\tilde{\theta} \sim \tilde{p}(\theta|x)$

Statlog example: german credit data

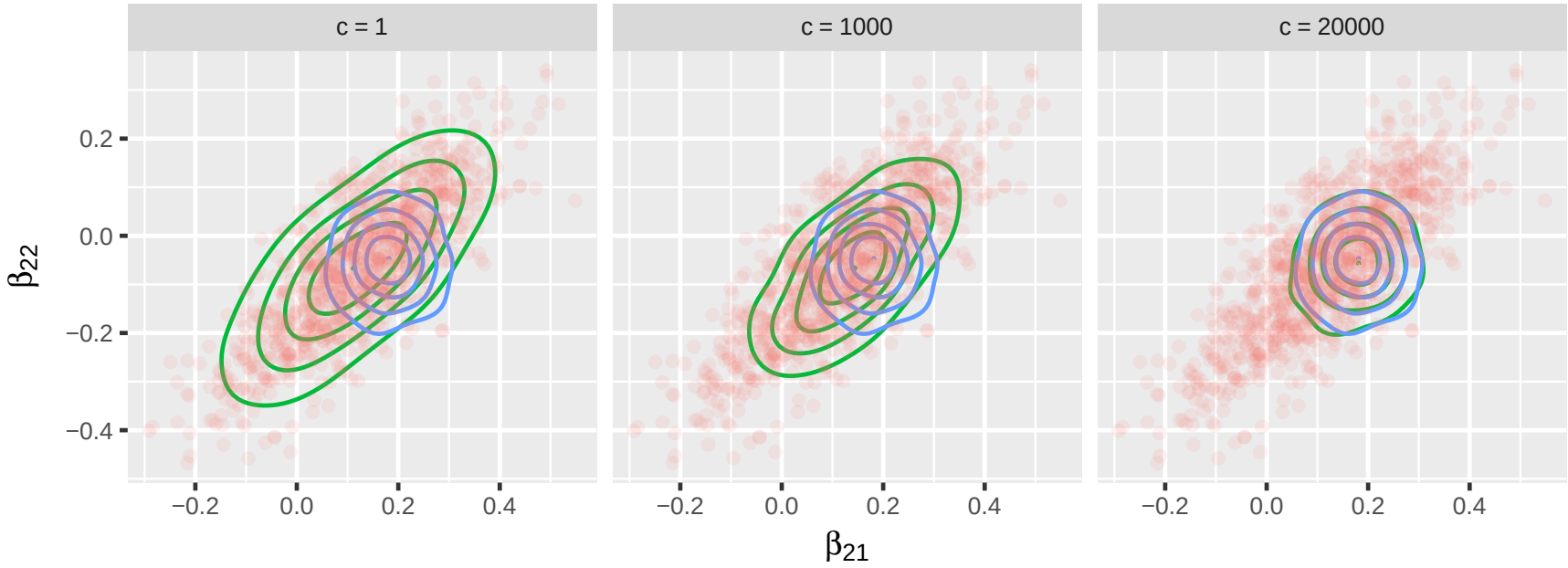

Figure: Posterior contour plot for β_{22} vs β_{21}, for VB-NPL (green) and VB (blue), for three different values of the concentration parameter c. Scatter plot is a sample from a Bayesian logistic posterior (red) via Polya-Gamma scheme.

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

Properties

- Two Theorems show,
  - NPL obtains exact asymptotic (frequentist) interval coverage for $\tilde{p}(\theta|x)$ – conventional Bayes update only does if the model (likelihood) is true
  - NPL obtains exact asymptotic (frequentist) prediction calibration
- The approach is trivially parallizable in correcting for (VB or other) model approximation – but requires an additive log-likelihood function for data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$

$$\log f_\theta(\boldsymbol{x}) = \sum_i \log f_\theta(x_i)$$

i.e., conditional independence of the data given the parameter

Central Idea: randomized objective functions

- More generally the approach extends to Bayesian inference that directly targets functionals of interest

$$\theta_0 = \arg\min \int R(x, \theta) dF_0(x)$$

where we have risk (loss, utility) $R(\cdot, \cdot)$, leading to samples from suitably randomized objective functions

$$\tilde{\theta}^{(j)} = \arg\min \sum_i w_i R(x_i, \theta)$$

for random weights

- Using a mix of real and synthetic data, and where for self-information loss, $R = -\log f_\theta(x)$ we obtain the Bayesian update for parameters indexing a model
- For example, we can learn about the outputs (optimal predictions) from machine learning algorithms $y = g(x)$ with additive objective functions where $g(\cdot)$ is an algorithm that takes x as input and predicts a y

Directly updating from synthetic data

- We considered Random forests (RF) to construct a Bayesian RF (BRF), via NPL with decision trees, under a prior mixing distribution
- This enables the incorporation of prior information, via synthetic data generated from a prior prediction function, in a principled way that is not available to RF
- To demonstrate the ability of BRF to incorporate prior information, we conducted the following experiment
- For 13 binary classification datasets from the UCI ML repository, we constructed a prior; training; and test dataset of equal size
- We measured test dataset predictive accuracy for three methods relative to an RF trained on the training dataset only:
  - BRF $(c = 0)$ (a non-informative BRF with $c = 0$, trained on the training dataset only),
  - BRF $(c > 0)$ (a BRF trained on the training dataset, incorporating prior synthetic-samples from a non-informative BRF trained on the prior dataset
  - RF-all (an RF trained on combined training and prior data).

Results

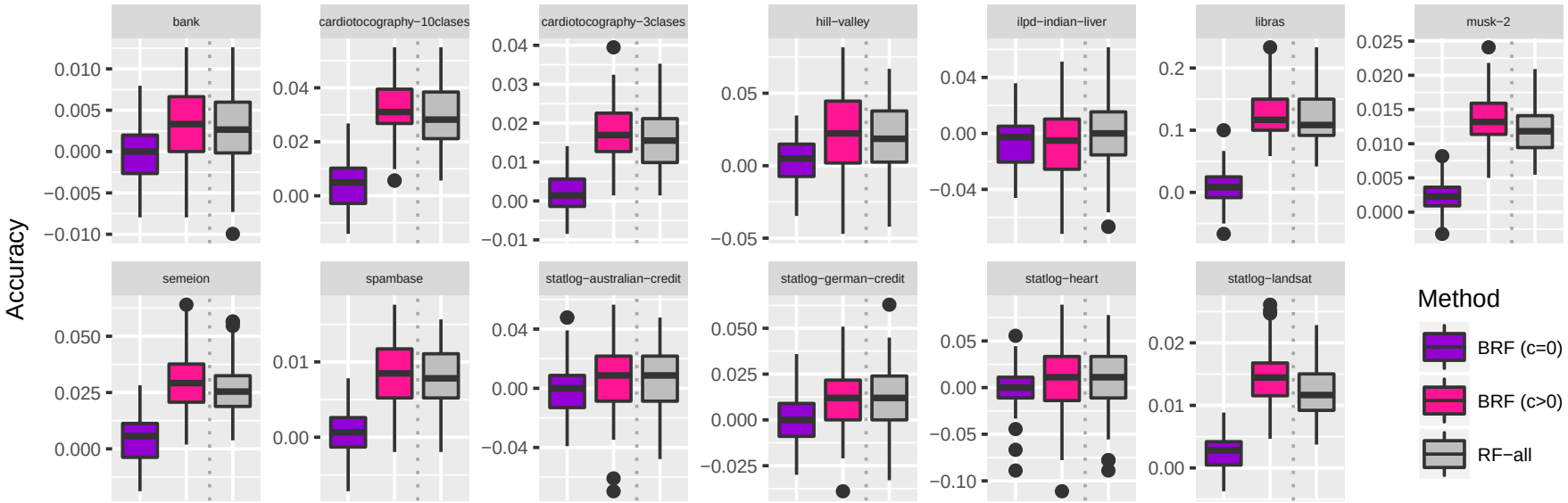

Figure: Boxplot of classification accuracy minus that of RF, for 13 UCI datasets.

- BRF best performance occurs when c is set equal to the number of samples in the prior training dataset, in line with intuition of the role of c as an effective sample size
- BRF$-c$ accuracy is better than that of RF, and close to that of RF-all
- BRF may have privacy benefits over RF-all as it only requires synthetic-samples; the prior data *and model* can be kept private

Conclusions

- We are motivated by large scale applications that do not rely on notions of true models
- We wish to avoid MCMC but make use of approximate models
- We wish for accurate uncertainty quantification on parameters of interest
- Replacing priors with synthetic-samples, and MCMC with randomized objective functions through the MDP (Antoniak, 1974)

$$\tilde{\theta} = \arg\max_{\theta} \left[n \sum_i w_i \log f_\theta(x_i) + c \sum_j w_j \log f_\theta(x_j^*) \right]$$

- On the one hand we are using nonparametrics to correct for parametric approximations, on the other hand we are using parametric models to regularize nonparametric inference

Thank you!

References – of mine

- Watson & Holmes "Approximate Models and Robust Decisions" – with discussion (2016) *Statistical Science*
- Bissiri, Holmes & Walker "General Bayesian Updating" (2016) *JRSS-b*
- Holmes & Walker "Assigning a value to a power likelihood in a general Bayesian model" (2017) *Biometrika*
- Lyddon, Walker & Holmes "Nonparametric learning from Bayesian models with randomized objective functions" (2018) *NIPS*
- Lyddon, Holmes & Walker "Generalized Bayesian Updating and the Loss-Likelihood Bootstrap" (2018) *archive*