From General Bayesian updating to Martingale Posteriors and back again via the Bayesian Bootstrap

Chris Holmes
University of Oxford, and
Ellison Institute of Technology

PostBayes seminar, April 2025

A story of three papers[1]

[Bissiri et al., 2016] – General Bayes

[Lyddon et al., 2018] [Lyddon et al., 2019] – Loss-likelihood Bootstrap

[Fong et al., 2023] – Martingale Posteriors

[1] and friendship, in particular with Stephen Walker; students Pier Giovanni Bissiri, Simon Lyddon, Edwin Fong

Bissiri, Holmes, & Walker, A general framework for updating belief distributions. *JRSS-B*. (2016)

Stephen and I started discussing the idea in 2009 during the second week of the Bayesian Nonparametric workshop (BNP-7) in Turin – it was very exciting

To paraphrase the question Stephen posed:

"Suppose you have beliefs about some parameter, θ, not necessarily indexing a likelihood function, and information (data), y, relevant for learning about the parameter. If you're willing to entertain an **update**, then what are the necessary properties of the update in the general setting outside of a likelihood-prior construction."

$$\{\pi(\theta), y\} \to \pi_u(\theta)$$

Example: estimating the median

Suppose you're interested in inferring the population median of heights of children

The first thing to note is the targeted nature of the question wrt a particular parameter (statistic) of the population

$$\theta_0 = \arg\min_{\theta \in \Theta} \int l(\theta, y)\, dF_0(y)$$

for some population distribution $F_0(y)$ and loss-function $l(\cdot, \cdot)$ targeting the parameter of interest, e.g. absolute loss in this case $l(\theta, y) = |\theta - y|$

This introduces the notion of an estimand, θ_0, the object you are trying to estimate

Inference

You have some data from independent samples of children

Child	1	2	3	4	5
Height (cm)	120	125	130	135	140

and beliefs of plausible values of the median based on knowledge of children, $\pi(\theta)$

We assume that combining the information in $\pi(\theta)$ and the data is possible

"These Bayesians are crazy"

The Bayesian solution solves a more general problem [Vapnik, 1998]

You first specify a sampling distribution (model) for the data

$$f_\theta(y)$$

with some parameters θ, potentially different and of higher dimension to the target estimand, θ_0

You define a prior on the model parameters, $\pi(\theta)$, before collecting data

Update using Bayes rule

$$\pi(\theta \mid y) \propto f_\theta(y)\, \pi(\theta)$$

Recover posterior beliefs on the estimand by assuming that the model is true $F_{\mathcal{M}} = \int_\alpha F_\alpha(\cdot)\pi(\alpha \mid y)\, d\alpha$, where $F_{\mathcal{M}}$ is a proxy for F_0, following which you calculate beliefs on θ_0

Bayes

Bayesian analysis separates out the modelling of the data from the target of inference

- sometimes this is a strength if you wish to ask many questions
- but translating beliefs about the estimand into beliefs about model parameters is non-trivial, e.g. [Bornn et al., 2019]
- solves a more general problem in order to provide direct uncertainty on the value of the estimand

In General Bayesian updating we were seeking an update directly on the target parameter, bypassing the need for a full sampling distribution

Additive information

If prior beliefs on the value of the estimand, $\pi(\theta)$, and data are independent, then the information needs to be additive

This, surprisingly, leads to a unique solution for coherent updating, namely, given three pieces of information $\{\pi(\theta), y_1, y_2\}$
Then, for coherent updating, the sequential update

$$\begin{aligned} \{\pi(\theta), y_1\} &\rightarrow \pi_1(\theta) \\ \{\pi_1(\theta), y_2\} &\rightarrow \pi_{1,2}(\theta) \end{aligned} \tag{1}$$

and the joint update

$$\{\pi(\theta), y_1, y_2\} \rightarrow \pi_{1,2}(\theta) \tag{2}$$

Must be equivalent under all settings

Coherent Updating

It turns out that the unique update with this property [Bissiri et al., 2016] is provided by

$$\pi_u(\theta) \propto \pi(\theta) \exp[-\sum_i l(\theta, y_i)]$$

and given that the loss-function has arbitrary scaling we have

$$\pi_u(\theta) \propto \pi(\theta) \exp[-\alpha \sum_i l(\theta, y_i)]$$

where α is a free parameter related to the information in the data relative to the information in $\pi(\theta)$

Supra-Bayesian pooling

The solution has exactly the same form as a Gibbs Posterior [Zhang, 2006]

The Gibbs Posterior is derived as an upper bound on the expected predictive risk from a randomized algorithm – where θ indexes a prediction model and the user samples $\theta \sim \pi_u(\theta)$ and observes the loss $\sum_i l(\theta, y_i)$

- This isn't Bayesian prediction
- Interesting result, and same solution

If we view $\{\pi(\theta), y\}$ as separate pieces of information that we want to combine, then we can view the General Bayes update as a Supra-Bayesian operation, leading to geometric (logarithmic) pooling of beliefs [Genest et al., 1986, Genst and McConway, 1999]

- We didn't cite McConway in [Bissiri et al., 2016] :(

How to set the learning rate?

In conventional Bayes you define the joint distribution, $p(\theta, y)$, *a priori* and then the learning rate is fixed and information processing is optimal [Zellner, 1988]

- Posterior sufficiency of $\pi(\theta \mid y)$ – you can throw away the data

But you have to define a joint distribution and assume that it's true

- Non-Bayesians have a hard time in viewing θ as a random variable

When you don't have a joint distribution, and you wish to use a General Bayes update, then you need to fix the learning rate, α

- This is especially tricky when the target, θ_0, is multi-dimensional

Calibration of learning

We needed a way to calibrate the information in the data, so as to specify α in the loss-likelihood, $\exp[-\alpha \sum_i l(\theta, y_i)]$

I recall we were looking at methods in "Objective Bayes" but, unsurprisingly, all use self-information loss (negative log-likelihood from a full sampling distribution), and we were trying to get away from likelihoods

Methods in Bayesian nonparametrics drew us to the Weighted-Likelihood Bootstrap [Newton and Raftery, 1994], a computational method that uses the Bayesian bootstrap to approximate a Bayes posterior (without a prior)

- The original paper received some criticism as a method for approximation, as MCMC was just emerging
- We didn't care about that as we weren't looking for an approximation to a Bayes posterior
- Method is purely data driven

Weighted Likelihood Bootstrap

The WLB draws samples, $\theta^{(j)}$ according to

$$\begin{aligned} w^{(j)} &\sim Dir_n(1, \ldots, 1) \\ \theta^{(j)} &= \arg\min_{\theta} -\sum_i w_i \log f_\theta(y_i) \end{aligned}$$

where w is a draw from the Bayesian Bootstrap [Rubin, 1981]

The method is 'prior free', in that you couldn't use a prior with it, which is what we wanted, and had good asymptotic properties

We can replace the self-information loss with a general loss function to create a loss-likelihood bootstrap with

$$\theta^{(j)} = \arg\min_{\theta} -\sum_i w_i l(\theta, y_i)$$

And then calibrate the General Bayes learning rate, α by, say, moment matching $\exp[-\alpha \sum_i l(\theta, y_i)]$ to samples $\{\theta^{(j)}\}_{j=1:J}$ [Lyddon et al., 2019]

Bayesian Bootstrap is amazing!

We started by looking for a way to set $\widehat{\alpha}$, which lead us to the Bayesian Bootstrap, which turned out to be more interesting (in many ways) than the original task

The WLP has lower predictive risk, asymptotically, than the corresponding Bayesian model when in $\mathcal{M}$-open [Lyddon et al., 2018]; using a result from [Fushiki, 2005]

- The BB is also studied outside of a Bayesian context, and without reference to [Rubin, 1981], where it's used to derive confidence intervals for Z-estimators, and shown to be more widely applicable than Efron's Bootstrap [Jin et al., 2001], requiring fewer assumptions

There's also a simple way to incorporate prior beliefs into the WLB using auxiliary, synthetic data, drawn from a prior predictive and combined with the real data, and re-weighted accordingly, [Lyddon et al., 2018]

The Bayesian bootstrap

The BB is quite widely known, but not well understood

- derived as the limiting distribution under a Dirichlet Process $DP(c, G)$, with base measure G and concentration parameter $c \to 0$

The BB and WLB seemed a bit opaque, and connections to Efron's bootstrap are unclear

For us, the key insight into its working came from the Polya urn representation of the DP by [Blackwell and MacQueen, 1973], as sampling with replacement from the empirical distribution $\mathbb{P}_n$

Bayesin bootstrap as a joint predictive

Starting from the empirical distribution function, $\mathbb{P}_n$, as a nonparametric 'prior free' predictive.
Leading to the 1-step predictive:

$$p(y_{N+1} \mid y_{1:N}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{y_i}$$

for $N = n, n+1, \ldots, \infty$, which reinforces an atom at a sampled value

This is a Pólya urn scheme with replacement, leading to a randomized empirical distribution function [Blackwell and MacQueen, 1973]

$$F_\infty := \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \delta_{y_i} = \sum_{i=1}^{n} w_i \delta_{y_i}$$

with $w_{1:n} \sim \text{Dirichlet}(1, \ldots, 1)$.

Picking off the parameter of interest $\theta(F_\infty)$ gives us the Bayesian Bootstrap [Rubin, 1981], where $\theta(\cdot)$ is an estimator

Bootstraps

Algorithm 1: Bayesian bootstrap

Set F_n from observed data $y_{1:n}$
for $j \leftarrow 1$ **to** B **do**
  for $i \leftarrow n+1$ **to** ∞ **do**
    Sample $Y_i \sim F_{i-1}$
    Update $F_i \leftarrow\!\shortmid \{F_{i-1}, Y_i\}$
  end
  Compute F_∞ from $\{y_{1:n}, Y_{n+1:\infty}\}$
  Evaluate $\theta_\infty^{(j)} = \theta(F_\infty)$
end
Return $\{\theta_\infty^{(1)}, \ldots, \theta_\infty^{(B)}\}$

Algorithm 2: Efron's bootstrap

Set F_n from observed data $y_{1:n}$
for $j \leftarrow 1$ **to** B **do**
  for $i \leftarrow 1$ **to** n **do**
    Sample $Y_i^* \sim F_n$
    No update to F_n
  end
  Compute F_n^* from $\{Y_{1:n}^*\}$
  Evaluate $\theta_n^{(j)} = \theta(F_n^*)$
end
Return $\{\theta_n^{(1)}, \ldots, \theta_n^{(B)}\}$

where $\theta(\cdot)$ is an estimator targeting the estimand

Note – in the Bayesian boostrap we don't actually have to simulate $i \leftarrow n+1$ **to** ∞ as we know the limiting distribution is a DP

Bayes analysis as predictive inference

Conventional Bayesian analysis has a similar predictive representation, where posterior uncertainty, $p(\theta \mid y_{1:n})$, arises from missing information through the data you don't have

$$\{y_{1:n}, Y_{n+1:\infty}\}$$

The missing information can be imputed through one-step posterior predictives

$$y_j \sim p(y \mid y_{1:j-1})$$

for $j = n+1 : \infty$, where $p(\cdot \mid y) = \int f_\theta(\cdot)\, \pi(\theta \mid y)\, ds$

The key connection is through [Doob, 1949]

Doob

From the imputed population, we compute the parameter estimate[2]

$$\theta_\infty = \theta(Y_{1:\infty})$$

which has a distribution conditional on $y_{1:n}$, induced by $Y_{n+1:\infty}$.
If we keep going generating ever more samples

Key result: From Doob's consistency theorem, if we use a conventional Bayes posterior predictive then $\theta_\infty \sim \pi(\theta \mid y_{1:n})$.

[2]Specifically, the posterior mean $\bar{\theta}(Y_{1:N}) = \int \theta \, d\pi(\theta \mid Y_{1:N})$

All of inference as missing data

The predictive view makes clear the essential distinction between Bayes and Frequentist inference

- **Frequentist** uncertainty considers variability arising from

$$\theta(Y_{1:n})$$

  with estimator $\theta(\cdot)$ and $Y_i \sim_{iid} F_0(y)$

- **Bayesian** uncertainty considers

$$\theta(y_{1:n}, Y_{n+1:\infty})$$

  for $Y_{n+1:\infty}$ *jointly* from F_0

They both have the same target estimand and same estimator

All of inference as missing data

- **Frequentist** uncertainty considers

$$\theta(Y_{1:n});\ \ Y_i \sim_{iid} F_0(y)$$

- **Bayesian** uncertainty considers

$$\theta(y_{1:n}, Y_{n+1:\infty});\ \ Y_{n+1:\infty} \sim_{joint} F_0$$

Frequentists consider indirect uncertainty in the estimand value through variability in the estimator following replicate draws of size n

Bayesians consider direct uncertainty on the population parameter of interest, targeted by the estimator through the missing $y_{n+1:\infty}$

They answer different questions. Bayes answers a harder question, and a more important question

The prior doesn't enter into the distinction – it's a distraction

Reframing the question

Viewed this way, one way to obtain Bayesian uncertainty on θ is through a joint predictive that conditions on the data you have

$$p(Y_{n+1:N} \mid y_{1:n})$$

Nothing is *a priori*, and allows for

- model checking
- cross-validation

This comes at a price – careful of data dipping, and care in how to construct the joint predictive if not using a likelihood-prior

This was the key motivation for [Fong et al., 2023], in weakening the conditions of exchangeability

The role of prior knowledge

The starting point of the analysis is the conditional predictive $p(y_{mis} \mid y_{obs})$

We don't care about defining a prior on data that has already been observed

- Why model what you have, you have it? Model what you don't have but need to answer the question

Expert judgement goes into defining a predictive model given all available information and existing data

Prior elicitation becomes predictive model evaluation/selection: where objective criteria are well developed and accepted

The search for objective priors becomes a search for objective predictives

Constructing a Predictive

Consider the Bayesian joint predictive under a *prequential* factorization:

$$p(y_{n+1:\infty} \mid y_{1:n}) = \prod_{i=n+1}^{\infty} p(y_i \mid y_{1:i-1}).$$

This factorization is simple law of probability that $p(a, b) = p(a \mid b)\, p(b)$

From which we can impute $Y_{n+1:\infty} \sim p(y_{n+1:\infty} \mid y_{1:n})$ through the recursion

1. Draw $Y_{n+1} \sim p(y \mid y_{1:n})$
2. Draw $Y_{n+2} \sim p(y \mid y_{1:n}, Y_{n+1})$ from the updated predictive, ..., etc

This constructive specification for the joint requires a 1-step predictive $p(y_{n+1} \mid y_{1:n})$ AND the update $p(y_{n+2} \mid y_{1:n}, y_{n+1})$

Imputation through Predictive Resampling

To generate the $Y_{n+1:\infty}$ we use a sequential imputation algorithm for $p_i = p(y_{i+1} \mid y_{1:i})$, that we call *predictive resampling*:

Algorithm 3: Predictive Resampling

Compute predictive p_n from the observed data $y_{1:n}$
$N > n$ is a large integer

for $i \leftarrow n+1$ **to** N **do**
  Sample $Y_i \sim p_{i-1}$
  Update $p_i \leftarrow\!\shortmid \{p_{i-1}, Y_i\}$
end
Evaluate $\theta_N = \theta(Y_{1:N})$

N is set large enough to yield no relevant uncertainty in the parameter of interest, θ, meaning in the context of a particular analysis for all practical purposes the conditional posterior can be replaced with a point estimate $\pi(\theta \mid y_{1:N}) = 1_{\theta(y_{1:N})}$

Comments on predictive resampling

The starting point of $p_n = p(y_{n+1} \mid y_{1:n})$ violates "coherent" belief updating

The update to the predictive p_i at each step necessitates the need for efficient online, continual, learning (model updating)

The algorithm is trivially parallel across samples of θ_N

Can extended to use maximum likelihood predictives at each step [Holmes and Walker, 2023], and also to model uncertainty using a consistent model selection criteria at each step [Shirvaikar et al., 2024]

Joint predictives

We need the joint predictive $p(y_{mis} \mid y_{obs})$ in order to obtain $\pi(\theta \mid y_{obs})$

- **Likelihood-prior**:
  - provides coherent updating and a short-cut to obtain $\pi(\theta \mid y_{1:n})$ directly through Bayes rule
  - need to define the prior-likelihood before you've seen any data
- **Predictive-updating**:
  - more general, allows one to use all of the data to construct the best predictive – nothing *a priori*
  - need to check for a resulting valid joint distribution (convergence and martingale conditions) and compute updates

Not claiming that Martingale posteriors are universally better, rather, they offer a different perspective and can sometimes be useful

Re-targeting

A final note on the story

Exploring the properties of the Bayesian bootstrap moved us away from General Bayesian updating towards predictive inference and Martingale posteriors

However, if interest is on a specific θ_0, such as a median, then having to construct a joint model for $p(Y_{n+1:\infty} \mid y_{1:n})$ may seem wasteful

Moreover, inductive bias in the choice of the predictive model for $p(Y_{n+1:\infty} \mid y_{1:n})$ may affect inference for our target $\pi(\theta \mid y_{1:n})$

[Yiu et al., 2025] considers posterior targeting taking the posterior marginal $\pi(\theta \mid y_{1:n})$ and re-focusing it using semi-parametric inference, using the Bayesian bootstrap!

Conclusions I

General Bayes [Bissiri et al., 2016] relaxed the $\mathcal{M}$-closed model framework to allow for robust inference on low dimensional targets

- Provides direct uncertainty on the estimand $\pi(\theta \mid y_{1:n})$

Setting the learning rate in GB is hard, and led to the development of the loss-likelihood bootstrap [Lyddon et al., 2018, Lyddon et al., 2019], using the Bayesian bootstrap with synthetic data as a 'prior'

The [Blackwell and MacQueen, 1973] representation of the BB provides insight into what the BB is doing, through predictive re-sampling, and [Doob, 1949] shows this is also true of conventional Bayes, allowing for more general constructions [Fong et al., 2023]

Conclusions II

Predictive inference places frequentist uncertainty, $\theta(Y_{1:n})$, and Bayesian uncertainty, $\theta(\{y_{1:n}, Y_{n+1:\infty}\})$, on an equal footing – making clear the essential distinction (which doesn't involve a prior)

Allows for more general constructions starting with the data you have [Fong et al., 2023], but more care is needed than using a likelihood-prior

Posterior corrections can correct for inductive bias in the model predictive that affect a targeted inference for a particular estimand [Yiu et al., 2025]

Thank you!

References I

[Bissiri et al., 2016] Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016).
A general framework for updating belief distributions.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78(5):1103–1130.

[Blackwell and MacQueen, 1973] Blackwell, D. and MacQueen, J. B. (1973).
Ferguson distributions via Pólya urn schemes.
The Annals of Statistics, 1(2):353–355.

[Bornn et al., 2019] Bornn, L., Shephard, N., and Solgi, R. (2019).
Moment conditions and bayesian nonparametrics.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 81(1):5–43.

[Doob, 1949] Doob, J. L. (1949).
Application of the theory of martingales.
Actes du Colloque International Le Calcul des Probabilités et ses applications (Lyon, 28 Juin–3 Juillet 1948), Paris CNRS, 23–27.

[Fong et al., 2023] Fong, E., Holmes, C., and Walker, S. G. (2023).
Martingale posterior distributions.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 85(5):1357–1391.

[Fushiki, 2005] Fushiki, T. (2005).
Bootstrap prediction and bayesian prediction under misspecified models.
Bernoulli, 11(4):747–758.

References II

[Genest et al., 1986] Genest, C., McConway, K. J., and Schervish, M. J. (1986).
Characterization of externally bayesian pooling operators.
The Annals of Statistics, 14(2):487–501.

[Genst and McConway, 1999] Genst, C. and McConway, K. (1999).
Bayesian pooling operators.
Rethinking the Foundations of Statistics, page 314.

[Holmes and Walker, 2023] Holmes, C. C. and Walker, S. G. (2023).
Statistical inference with exchangeability and martingales.
Philosophical Transactions of the Royal Society A, 381(2247):20220143.

[Jin et al., 2001] Jin, Z., Ying, Z., and Wei, L. J. (2001).
A simple resampling method by perturbing the minimand.
Biometrika, 88(2):381–390.

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

[Lyddon et al., 2019] Lyddon, S. P., Holmes, C. C., and Walker, S. G. (2019).
General Bayesian updating and the loss-likelihood bootstrap.
Biometrika, 106(2):465–478.

References III

[Newton and Raftery, 1994] Newton, M. and Raftery, A. (1994).
Approximate Bayesian inference by the weighted likelihood bootstrap.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 56:3 – 48.

[Rubin, 1981] Rubin, D. B. (1981).
The Bayesian bootstrap.
The Annals of Statistics, 9(1):130–134.

[Shirvaikar et al., 2024] Shirvaikar, V., Walker, S. G., and Holmes, C. (2024).
A general framework for probabilistic model uncertainty.
arXiv preprint arXiv:2410.17108.

[Vapnik, 1998] Vapnik, V. N. (1998).
Statistical Learning Theory.
Wiley, New York.

[Yiu et al., 2025] Yiu, A., Fong, E., Holmes, C., and Rousseau, J. (2025).
Semiparametric posterior corrections.
Journal of the Royal Statistical Society Series B: Statistical Methodology, page qkaf005.

[Zellner, 1988] Zellner, A. (1988).
Optimal information processing and bayes's theorem.
The American Statistician, 42(4):278–284.

References IV

[Zhang, 2006] Zhang, T. (2006).
Information-theoretic upper and lower bounds for statistical estimation.
IEEE Transactions on Information Theory, 52(4):1307–1321.