Parametric bootstrap: a way to combine Bayesian and frequentist ideas

Snigdhansu Chatterjee

School of Statistics, University of Minnesota

Joint Work with:
Partha Lahiri, Univ. of Maryland, College Park

JPSM Bayesian Workshop 2008, April 30, 2008

Table of contents

Outline

1. **Prediction in the past millenia**
2. Bayesian prediction
   - A slide (or two) about objective Bayesian statistics
3. Bootstrap prediction
4. And the twain shall (nearly) meet
   - Prediction in multilevel models

Prediction long long time back

(*Credit: Uderzo*)

Compare with:

(*Credit: Rembrandt*)

The basic setup (Bjørnstad, 1990)

- We have an observed sample: $Y_n = (Y_1, \dots, Y_n)$ and an unobserved sample: $Y' = (Y'_1, \dots, Y'_m)$.
- θ is a function of Y', like Y' itself, $\sum Y'_i$ or max Y'_i.
- We want to predict θ, based on Y_n.
- Assume that (Y_n, θ) has a density or mass function $\ell(y_n, \theta; \xi)$ for some parameter ξ.

The basic setup continued

Joint distribution

$(Y_n, \theta) \sim \ell(y_n, \theta; \xi)$.

- Assume $(\theta | Y_n) \sim \pi(\theta | y_n, \xi)$. We want to estimate $\pi(\theta | y_n, \xi_0)$, where ξ_0 is the true value of the parameter.
- Predictors are of the form $L(\theta, y_n)$; obtained from $\ell(\cdot, \xi)$ by eliminating ξ somehow.
- $L(\theta, y_n)$ need not look like $\pi(\theta | y_n, \xi)$.

The strawman predictor

Joint and conditional distribution

$(Y_n, \theta) \sim \ell(y_n, \theta; \xi_0), \quad (Z|Y_n) \sim \pi(\theta|y_n, \xi_0)$.

- **Estimative approach/ naive prediction:**
  Suppose $\hat{\xi}$ is an estimator of ξ based on observed data Y_n.
  Use $\pi(\theta|y_n, \hat{\xi})$ as a predictive density.
- This is misleadingly precise, as variability of $\hat{\xi}$ is not counted.

Example

Y_i's and $\theta(= Y_1')$ are $i.i.d.$ $N_p(\xi, \mathbf{I}_p)$ for some $\xi \in R^p$.
MLE of ξ is $\bar{Y}$.
The naive/estimative predictor is $N_p(\bar{Y}, \mathbf{I}_p)$.

Outline

Bayesian predictive density

Joint and conditional

$(Y_n, \theta) \sim \ell(y_n, \theta; \xi_0), \quad (\theta | Y_n) \sim \pi(\theta | y_n, \xi_0)$.

- Assume $\xi \sim h(\cdot)$, and $(Y_n | \xi) \sim f_n(\cdot | \xi)$.
- The posterior density of $(\xi | Y_n)$ is

$$h_n(\xi | Y_n) \propto f_n(\cdot | \xi) h(\cdot).$$

- The Bayesian predictive density is

$$L_B(\theta | y_n) = \int \pi(\theta | y_n, \xi) h_n(\xi | y_n) d\xi.$$

- Depends on choice of $h(\cdot)$.
- Non-Bayesian approaches use sufficiency or profiling, and have several shortcomings.

Measuring prediction accuracy: KL divergence

Joint and conditional distribution

$(Y_n, \theta) \sim \ell(y_n, \theta; \xi_0), \quad (\theta | Y_n) \sim \pi(\theta | y_n, \xi_0).$

- The Kullback-Leibler discrepancy is

$$KL(\pi, L) = \int \log\left(\frac{\pi(\theta | Y_n, \xi_0)}{L(\theta | Y_n)}\right) \pi(\theta | Y_n, \xi_0) d\theta$$

- This is a function of Y_n.
- We say L_1 is *KL*-better than L_2 if

$$EKL(\pi, L_1) \leq EKL(\pi, L_2).$$

Measuring prediction accuracy: The Renyi-Amari divergences

- The Renyi-Amari divergence criteria are given by

$$A_\beta(\pi, L) = \frac{1}{\beta(1-\beta)} \left[1 - \int \pi^{1-\beta} L^\beta \right].$$

- This is a broad class of divergence functions, one of which is a distance between densities.
- The Kullback Leibler divergence ($\beta \downarrow 0$), Hellinger distance ($\beta = 1/2$), reverse Kullback Leibler divergence ($\beta \to 1$) and χ^2 divergence ($\beta = 2$) are special cases.
- Variously attributed to Renyi (1961), Amari (1981), Cressie-Read (1984), ...

Bayesian versus naive

Example

Y_{ni}'s and θ are *i.i.d.* $N_p(\xi, \mathbf{I}_p)$.
Hence $\pi(\theta | Y_n, \xi) = \phi_p(\theta - \xi)$, the density of $N_p(\xi, \mathbf{I}_p)$.
The naive predictor is the $N_p(\bar{Y}, \mathbf{I}_p)$ density.

- Choose ξ uniformly in R^p, that is $h(\cdot) \equiv 1$. Thus, Bayesian predictor is the integrated likelihood:

$$L_B(\theta | y_n) \propto \int_\xi \phi_p(\theta - \xi) \prod_{i=1}^{n} \phi_p(y_{ni} - \xi) d\xi.$$

This is *KL*-better than the naive predictor. (Aitchison (1975))

Better than integrated likelihood

Integrated likelihood > naive

Data Y_{ni}'s and $\theta = Y'_{11}$ are *i.i.d.* $N(\xi, \mathbf{I}_p)$.
Integrated likelihood $L_B(\theta|y_n) \propto \int_\xi \phi_p(\theta - \xi) \prod_{i=1}^n \phi_p(y_{ni} - \xi) d\xi$ is *KL*-better than the naive predictor $N(\bar{Y}_n, \mathbf{I}_p)$.

- There are other choices of priors (harmonic priors, superharmonic priors) that obtain predictors that are *KL*-better than integrated likelihood. Komaki (1996, ...), George, Liang, Xu (2006).
- Example: Harmonic prior: $h(\xi) \propto ||\xi||^{-(p-2)}$.

A (semi)-final word on the Bayesian story

All this is only for the i.i.d. $N(\xi, \mathbf{I}_p)$ case!

Objective Bayesian probability (O'Hagan and Forster, 2002)

- It is quite possible to follow Bayesian method without holding either a frequentist or subjective view of probability.
- $P(A|H)$ still represents a degree of belief in A based on information H, but it is a unique objective measure of the degree to which A is logically entailed by H.
- This does not require A to be repeatable (not frequentist!), ...
- ... but needs the prior to be the unique distribution implied by prior information.

Near objectivity: an example

Example

Does it matter if instead of $\bar{Y}$, we use as an estimator of mean:

- $\bar{Y} + C/n^{1/2}$?
- Yes. (Asymptotic normality would be affected)
- $\bar{Y} + C/n$?
- Maybe too high a bias, but CLT unaffected.
- $\bar{Y} + C/n^{3/2}$?
- Higher order asymptotics would change, (but hardly matters?).
- ...
- There is a "comfort level" k such that we would not distinguish between $\bar{Y}$ and $\bar{Y} + C/n^{k/2}$.

Moral of the story

We are indifferent between methods whose properties match up to a given level.

Near objectivity: an example

Example

Does it matter if instead of $\bar{Y}$, we use as an estimator of mean:

- $\bar{Y} + C/n^{1/2}$?
- Yes. (Asymptotic normality would be affected)
- $\bar{Y} + C/n$?
- Maybe too high a bias, but CLT unaffected.
- $\bar{Y} + C/n^{3/2}$?
- Higher order asymptotics would change, (but hardly matters?).
- ...
- There is a “comfort level” k such that we would not distinguish between $\bar{Y}$ and $\bar{Y} + C/n^{k/2}$.

Moral of the story

We are indifferent between methods whose properties match up to a given level.

Near objectivity: an example

Example

Does it matter if instead of $\bar{Y}$, we use as an estimator of mean:

- $\bar{Y} + C/n^{1/2}$?
- Yes. (Asymptotic normality would be affected)
- $\bar{Y} + C/n$?
- Maybe too high a bias, but CLT unaffected.
- $\bar{Y} + C/n^{3/2}$?
- Higher order asymptotics would change, (but hardly matters?).
- $\ldots$
- There is a "comfort level" k such that we would not distinguish between $\bar{Y}$ and $\bar{Y} + C/n^{k/2}$.

Moral of the story

We are indifferent between methods whose properties match up to a given level.

Near objectivity: an example

Example

Does it matter if instead of $\bar{Y}$, we use as an estimator of mean:

- $\bar{Y} + C/n^{1/2}$?
- Yes. (Asymptotic normality would be affected)
- $\bar{Y} + C/n$?
- Maybe too high a bias, but CLT unaffected.
- $\bar{Y} + C/n^{3/2}$?
- Higher order asymptotics would change, (but hardly matters?).
- ...
- There is a "comfort level" k such that we would not distinguish between $\bar{Y}$ and $\bar{Y} + C/n^{k/2}$.

Moral of the story

We are indifferent between methods whose properties match up to a given level.

Near objectivity: an example

Example

Does it matter if instead of $\bar{Y}$, we use as an estimator of mean:

- $\bar{Y} + C/n^{1/2}$?
- Yes. (Asymptotic normality would be affected)
- $\bar{Y} + C/n$?
- Maybe too high a bias, but CLT unaffected.
- $\bar{Y} + C/n^{3/2}$?
- Higher order asymptotics would change, (but hardly matters?).
- ...
- There is a "comfort level" k such that we would not distinguish between $\bar{Y}$ and $\bar{Y} + C/n^{k/2}$.

Moral of the story

We are indifferent between methods whose properties match up to a given level.

Near objectivity: an example

Example

Does it matter if instead of $\bar{Y}$, we use as an estimator of mean:

- $\bar{Y} + C/n^{1/2}$?
- Yes. (Asymptotic normality would be affected)
- $\bar{Y} + C/n$?
- Maybe too high a bias, but CLT unaffected.
- $\bar{Y} + C/n^{3/2}$?
- Higher order asymptotics would change, (but hardly matters?).
- ...
- There is a "comfort level" k such that we would not distinguish between $\bar{Y}$ and $\bar{Y} + C/n^{k/2}$.

Moral of the story

We are indifferent between methods whose properties match up to a given level.

Near objectivity: an example

Example

Does it matter if instead of $\bar{Y}$, we use as an estimator of mean:

- $\bar{Y} + C/n^{1/2}$?
- Yes. (Asymptotic normality would be affected)
- $\bar{Y} + C/n$?
- Maybe too high a bias, but CLT unaffected.
- $\bar{Y} + C/n^{3/2}$?
- Higher order asymptotics would change, (but hardly matters?).
- ...
- There is a "comfort level" k such that we would not distinguish between $\bar{Y}$ and $\bar{Y} + C/n^{k/2}$.

Moral of the story

We are indifferent between methods whose properties match up to a given level.

Near objectivity: an example

Example

Does it matter if instead of $\bar{Y}$, we use as an estimator of mean:

- $\bar{Y} + C/n^{1/2}$?
- Yes. (Asymptotic normality would be affected)
- $\bar{Y} + C/n$?
- Maybe too high a bias, but CLT unaffected.
- $\bar{Y} + C/n^{3/2}$?
- Higher order asymptotics would change, (but hardly matters?).
- ...
- There is a "comfort level" k such that we would not distinguish between $\bar{Y}$ and $\bar{Y} + C/n^{k/2}$.

Moral of the story

We are indifferent between methods whose properties match up to a given level.

Outline

Joint and conditional; integration method

$(Y_n, Z) \sim \ell(y_n, \theta; \xi_0)$, $(\theta | Y_n) \sim \pi(\theta | y_n, \xi_0)$.
Bayesian predictor: $L_B(\theta|y_n) = \int \pi(\theta|y_n, \xi) h_n(\xi|y_n) d\xi$.

- Consider an estimator $\hat{\xi}$ of ξ based on the observed data Y_n.
- The estimator $\hat{\xi}$ has a density $I_n(\cdot)$, which may be approximated using bootstrap, say by $I_n^*(\cdot|Y_n)$.
- (Harris (1989)) The bootstrap predictive density is

$$L^*(\theta|y_n) = \int \pi(\theta|y_n, \xi) I_n^*(\xi|y_n) d\xi.$$

- Depends on choice of $\hat{\xi}$.

Recent developments in the bootstrap story

- Fushiki, Komaki, Aihara (2004) showed bootstrap predicton is like Bayesian prediction with the M-prior of Hartigan (1968, . . ., 1990s).
- Whether it is KL-better than the naive predictor depends on several technical conditions.
- Bagging (a recent classification technique) is related to bootstrap prediction (Fushiki, Komaki, Aihara (2005)).
- Efron (in several papers and talks) discussed conection between frequentist (bootstrap) and Bayesian inference.
- Can be extended to complex (non-$i.i.d.$, high dimensional) problems, evaluation by other criteria possible . . ., (C. (2008?)).

Outline

The multilevel model

- $$\begin{aligned} \text{Level I:} \quad & (Y_n|\theta_n, \xi) \sim f_n(\cdot; \theta_n, \xi), \\ \text{Level II:} \quad & (\theta_n|\xi) \sim g_n(\cdot; \xi). \end{aligned}$$
- Some part of ξ denotes regression parameters, other part denotes variance components, we club all these together.
- $$(Y_n|\xi) \sim \int f_n g_n d\theta_n = m_n(\cdot; \xi),$$
  $$(\theta_n|Y_n, \xi) \sim \frac{f_n g_n}{m_n} = \pi_n(\theta_n|Y_n, \xi).$$
- We want to predict $\pi_n(\cdot|Y_n, \xi_0) \equiv \pi_{n0}$.
- Let ξ_n be a good estimator of ξ, based on Y_n, like the MLE, REML.

An example (Laird and Louis, 1987)

Example

$$\begin{aligned} Y_{ni}|\theta_{ni} &\sim N(\theta_{ni}, 1) \\ \theta_{ni} &\sim N(\mu, A) \end{aligned}$$

independently. Here $\xi = (\mu, B)$, where $B = 1/(1 + A)$.

- $$(\theta_n | Y_n, \xi) \sim N_n(B\mu\mathbf{1} + (1 - B)Y_n, (1 - B)\mathbf{I}).$$

  The density of this is $\pi_n(\cdot | Y_n, \xi) \equiv \pi_{n0}$.
- ξ may be estimated by $\hat{\xi} = (\hat{\mu}, \hat{B})$ from the marginal density of Y_n.

Example continued (Laird and Louis, 1987): the strawman

The predictee

$(\theta_n | Y_n, \xi) \sim N_n(B\mu\mathbf{1} + (1 - B)Y_n, (1 - B)\mathbf{I})$.

- **Empirical best/Bayes Predictor** is the estimative (strawman!) predictor $N_n(\hat{B}\hat{\mu}\mathbf{1} + (1 - \hat{B})Y_n, (1 - \hat{B})\mathbf{I})$.
- Corrections from mean squared error angle: Prasad and Rao (1990) and several others, . . ., Lahiri and Rao (2008?), Chatterjee and Lahiri (2008?).
- Correcting from confidence intervals angle using bootstrap: Laird and Louis (1987), Carlin and Gelfand (1990, 1991), few others, . . . Hall and Maiti (2006), Chatterjee, Lahiri, Li (2008?).
- Most of the bootstrap-based confidence interval correction techniques used calibration. Chatterjee, Lahiri, Li uses bootstrap intervals directly, so it's not really a "correction".

The Bayesian predictor

Model and predictee

Level I: $(Y_n|\theta_n, \xi) \sim f_n(\cdot; \theta_n, \xi)$, Level II: $(\theta_n|\xi) \sim g_n(\cdot; \xi)$.
Predictee: $(\theta_n|Y_n, \xi) \sim \pi_n(\theta_n|Y_n, \xi)$.

- Fix the prior

$$\xi \sim h(\cdot) \text{ and compute the posterior } (\xi|Y_n) \sim h_n(\cdot|Y_n)$$

- The Bayesian predictive density is given by

$$g_{nh}(\theta_n; Y_n) = \int \pi_n(\theta_n; Y_n, t) h_n(t; Y_n) dt.$$

The parametric bootstrap predictor

Model and predictee

Level I: $(Y_n|\theta_n,\xi) \sim f_n(\cdot;\theta_n,\xi)$, Level II: $(\theta_n|\xi) \sim g_n(\cdot;\xi)$.
Predictee: $(\theta_n|Y_n,\xi) \sim \pi_n(\theta_n|Y_n,\xi)$.

- Generate data according to the scheme

$$\begin{aligned}(Y_n^*|\theta_n^*,\xi_n) &\sim f_n(\cdot;\theta_n^*,\xi_n),\\ (\theta_n^*|\xi_n) &\sim g_n(\cdot;\xi_n).\end{aligned}$$

- Let ξ_n^* be a bootstrap version of ξ_n, obtained by using the same techniques used to obtain ξ_n, but using Y_n^* instead of Y_n.
- Suppose $(\xi_n^*|Y_n) \sim l_n^*(\cdot|Y_n)$.
- The parametric bootstrap predictive density is given by

$$g_n^*(\theta_n;Y_n) = \int \pi_n(\theta_n;Y_n,t)l_n^*(t|Y_n)dt.$$

The predictee

$(\theta_n | Y_n, \xi) \sim N_n(B\mu\mathbf{1} + (1-B)Y_n, (1-B)\mathbf{I})$ has density $\pi_n(\theta_n; Y_n, \xi)$.
EB: $N_n(\hat{B}\hat{\mu}\mathbf{1} + (1-\hat{B})Y_n, (1-\hat{B})\mathbf{I})$ with density $\pi_n(\theta_n; Y_n, \hat{\xi})$.
Bayesian: $g_{nh}(\theta_n; Y_n) = \int \pi_n(\theta_n; Y_n, t) h_n(t; Y_n) dt$.
Bayesian: $g_n^*(\theta_n; Y_n) = \int \pi_n(\theta_n; Y_n, t) l_n^*(t | Y_n) dt$.

Theorem

(Laird and Louis): Generally, bootstrap and Bayesian predictions do not exactly match.

Can they match closely? Does it matter?

Why should the twain (nearly) meet?

- Coherence and decision theoretic optimality of bootstrap prediction, *at least approximately*.
- Frequentist properties of Bayesian predictors.
- "BUY 1, GET 2" with computations! Use whichever method is easier, get both, *at least approximately*.
- This is a way of obtaining prior for objective Bayesian inference.
- This is a route to extend Bayesian optimality beyond the iid setup, especially to high dimensional cases.

Why should the twain (nearly) meet?

- Coherence and decision theoretic optimality of bootstrap prediction, *at least approximately*.
- Frequentist properties of Bayesian predictors.
- "BUY 1, GET 2" with computations! Use whichever method is easier, get both, *at least approximately*.
- This is a way of obtaining prior for objective Bayesian inference.
- This is a route to extend Bayesian optimality beyond the iid setup, especially to high dimensional cases.

Why should the twain (nearly) meet?

- Coherence and decision theoretic optimality of bootstrap prediction, *at least approximately*.
- Frequentist properties of Bayesian predictors.
- "BUY 1, GET 2" with computations! Use whichever method is easier, get both, *at least approximately*.
- This is a way of obtaining prior for objective Bayesian inference.
- This is a route to extend Bayesian optimality beyond the iid setup, especially to high dimensional cases.

Why should the twain (nearly) meet?

- Coherence and decision theoretic optimality of bootstrap prediction, *at least approximately*.
- Frequentist properties of Bayesian predictors.
- "BUY 1, GET 2" with computations! Use whichever method is easier, get both, *at least approximately*.
- This is a way of obtaining prior for objective Bayesian inference.
- This is a route to extend Bayesian optimality beyond the iid setup, especially to high dimensional cases.

Why should the twain (nearly) meet?

- Coherence and decision theoretic optimality of bootstrap prediction, *at least approximately*.
- Frequentist properties of Bayesian predictors.
- "BUY 1, GET 2" with computations! Use whichever method is easier, get both, *at least approximately*.
- This is a way of obtaining prior for objective Bayesian inference.
- This is a route to extend Bayesian optimality beyond the iid setup, especially to high dimensional cases.

Why should the twain (nearly) meet?

- Coherence and decision theoretic optimality of bootstrap prediction, *at least approximately*.
- Frequentist properties of Bayesian predictors.
- "BUY 1, GET 2" with computations! Use whichever method is easier, get both, *at least approximately*.
- This is a way of obtaining prior for objective Bayesian inference.
- This is a route to extend Bayesian optimality beyond the iid setup, especially to high dimensional cases.

The matching of Bayesian and bootstrap prediction

Predictors:

Bayesian predictor: $g_{nh} = \int \pi_n(\theta_n; Y_n, t) h_n(t; Y_n) dt$.
Bootstrap predictor: $g_n^* = \int \pi_n(\theta_n; Y_n, t) l_n^*(t | Y_n) dt$.

Theorem

If the first few moments of bootstrap distribution and posterior of ξ nearly match, then g_n^* and g_{nh} are nearly equivalent as predictors of π_{n0}, in the sense

$$A_\beta(\pi_{n0}, g_n^*) - A_\beta(\pi_{n0}, g_{nh}) \approx 0$$

for almost every data sample Y_n for every fixed β.

- By ≈ 0 we mean $\left(o(n^{-\gamma/2})\right)$, for our choice of γ.
- This forms a bridge between frequentist asymptotics (in bootstrap) and Bayesian asymptotics.

Example continued (Laird and Louis, 1987)

Example

Let $Y_i|\theta_i \sim N(\theta_i, 1)$ and θ_i's are i.i.d. $N(0, A)$.
Y_i's are i.i.d. $N(0, 1 + A = \xi)$. Let $\xi_n = \frac{1}{n}\sum Y_i^2$.

- Bootstrap version (I_n^*): $\xi_n^* = \frac{\sum Y_i^2}{n^2} V$, where $V \sim \chi_n^2$.
- Bayesian version (h_n): Consider the prior $\xi \sim$ inverse Gamma (a, d). The posterior distribution is inverse Gamma $(a + \sum Y_i^2, d + n)$.
- For Bayesian and bootstrap predictors to be $o(n^{-2+\epsilon})$ close for almost all data sample Y_n, the first four moments of the posterior and bootstrap distribution have to be $o(n^{-4})$ close.
- This involves matching χ^2 and inverse Gamma moments closely.

Laird and Louis example continued

Matching moments

Matching moments of $\frac{\sum Y_i^2}{n^2}\chi_n^2$ and $IG(a + \sum Y_i^2, d + n)$.

- That is, solve for (a, d) such that

Mean:	$\frac{1}{n}\sum Y_i^2$	$\frac{a+\sum Y_i^2}{d+n-2}$
Variance:	$\frac{2}{n^3}\left(\sum Y_i^2\right)^2$	$\frac{2(a+\sum Y_i^2)^2}{(d+n-2)^2(d+n-4)}$
Skewness:	...	...
Kurtosis:	...	...

- Finding a (a, d) so that these moments match seems impossible!
- Inverse Gamma posterior and χ^2 bootstrap moments differ systematically by $O(n^{-1})$.

Matching moments

Posterior and bootstrap moments differ by $O(n^{-1})$. Change either the prior, or the bootstrap.

- Solution 1: **Modify the parametric bootstrap scheme**: Perturb the bootstrap distribution by slightly changing its mean and variance.
- For example, define the parametric bootstrap predictive density as

$$g_n^*(\theta_n; Y_n) = \int \pi_n(\theta_n; Y_n, (1 + n^{-1}u)t) l_n(t|Y_n) dt \lambda(u) du.$$

for a suitable function $\lambda(\cdot)$.

Laird and Louis example continued: Two alternative solutions

- Solution 2: **Use a "bootstrap moment matching" prior**: Find a prior such that the posterior moments match with bootstrap moments.
- Try the class of scale mixture of inverse Gamma's?

$$\sum p_i IG(a_i, d)$$

Laird and Louis example continued: Two alternative solutions

Theorem

There exists a scale mixture of three $(k+1)$ inverse Gamma, whose posterior will match the first four $(2k)$ moments of the bootstrap.

- There is $\approx$ 150 years of math literature on finding distribution with given moments (that we have not read, yet) ...
- Chebyshev (1850s), Markov (1880s), Steiltjes (1890s), Hamburger (1920s), Hausdorff (1920s), Nevanlinna (1920s), Hardy (1920s), Havilan (1930s), ...
- Algorithms for this is in Devroye (1986), Hörmann *et al* (2004).

Matching moments

Works for the Laird Louis example.

- **Question 1:** Is there always such a match possible?
- Tentative answer is yes.
- **Question 2:** Can we find a match in scale mixtures of nice priors?