Are Bayesian Inferences Weak for Wasserman’s Example?

Longhai Li
longhai@math.usask.ca

Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon, Saskatchewan, S7N 5E6 Canada

Quebec City, SSC 2010
25 May 2010
In the textbook by Wasserman (2004), under a Section titled “Strengths and Weaknesses of Bayesian Inference” (pages 186-189), he used a simple example to “demonstrate” the weaknesses of Bayesian inferences in high-dimensional and nonparametric problems.

I have carried out a simple Bayesian analysis for this example, obtaining a simple Bayes estimate, and used a frequentist method — mean square error, to compare with Horwitz-Thompson he suggested to use for this problem.
Wasserman’s Example

The model in the example may be appropriate for sampling survey problems. Observation \( Y_i \) is modeled by a mixture distribution of a huge number, \( B \), of Bernoulli distributions, parameterized by \( \theta_b \), for \( b = 1, \ldots, B \):

\[
X_i \sim \text{Uniform}(1, \ldots, B),
\]

\[
Y_i | X_i \sim \text{Bernoulli}(\theta_{X_i}).
\]

In practice it is possible that some of these \( Y_i \) are unobserved, for example when those sampled individuals refuse to respond. Let \( \xi_b \) denote the probability that the individual indexed by \( b \) will respond to the question. Then, given \( X_i \), the distribution of \( R_i \) is

\[
R_i | X_i \sim \text{Bernoulli}(\xi_{X_i}).
\]

In addition, we assume that \( R_i \) and \( Y_i \) are independently distributed given \( X_i \).

Interested Parameter:

\[
\varphi = \frac{1}{B} \sum_{b=1}^{B} \theta_b = P(Y_i = 1).
\]
The likelihood function of $\theta$ and $\xi$ based on the data $D$ is the product of the joint distributions of either $(X_i, R_i = 1, Y_i)$ or $(X_i, R_i = 0)$, for $i = 1, \ldots, n$:

$$L(\theta, \xi ; D) = \frac{1}{B^n} \prod_{i=1}^{n} \xi^{R_i} X_i (1 - \xi X_i)^{1-R_i} \prod_{\{i : R_i=1\}} \theta^{Y_i} X_i (1 - \theta X_i)^{1-Y_i} \quad (5)$$

Wasserman’s argument based on this likelihood function is rephrased as follows:

The above likelihood function is relevant to at most $n$ different $\theta_b$. Therefore, when $B$ is greatly larger than $n$, the likelihood function contains information of only a tiny fraction of $\theta$. The posterior distribution of $\theta$ is almost equal to the whatever prior distribution, therefore cannot lead to a good inference for the interested parameter, $\varphi$. 

Horwitz-Thompson Estimator

Wasserman suggested the following estimator for $\varphi$:

$$\hat{\varphi}_{HT} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i Y_i}{\xi X_i},$$

(6)

where, when $R_i = 0$, $Y_i$ is imaginary and can be assigned arbitrarily. This estimator treats both observed $Y_i = 0$, and sets missing $Y_i$ to 0, and count $1/\xi X_i$ times each observed $Y_i = 1$. Note that he assumes that the parameter $\xi$ is known.

One can easily show that this estimate has mean $\varphi$, by iterative expectation formula. This estimator is also consistent since its MSE converges to 0 as $n \to \infty$:

$$\text{MSE}(\hat{\varphi}_{HT}) = \frac{1}{n} \text{Var} \left( \frac{R_i Y_i}{\xi X_i} \right)$$

(7)

$$= \frac{1}{n} \left( E \left( \frac{R_i Y_i}{\xi^2 X_i} \right) - \varphi^2 \right)$$

(8)

$$= \frac{1}{n} \left( \frac{1}{B} \sum_{b=1}^{B} \frac{\theta_b}{\xi_b} - \varphi^2 \right).$$

(9)
A Bayesian Analysis
Prior Specification

I will assume that \( \theta \) and \( \xi \) are independent given some hyperparameters.

The prior for \( \theta \):

\[
\begin{align*}
\theta_1, \ldots, \theta_B | \alpha_T, f & \sim \text{Beta}(\theta | \alpha_T f, \alpha_T (1 - f)), \\
f, \alpha_T & \sim \text{Beta}(f | \alpha_F, \alpha_F) \times \pi_T(\alpha_T).
\end{align*}
\]

Some properties of the above prior:

\[
E(\theta_b) = f \\
\text{Var}(\theta_b) = \frac{f(1 - f)}{\alpha_T + 1} \\
\theta_b | f, \alpha_T \rightarrow^d \text{Bernoulli}(f), \quad \text{as } \alpha_T \rightarrow \infty
\]

The prior for \( \xi \) is denoted by \( \pi_\xi(\xi) \). We don’t need to specify it as it will be unrelated to the posterior of \( \theta \) once we assume that \( \theta \) and \( \xi \) are independent. Similar for \( \pi_T(\alpha_T) \).
Posterior Analysis

When $B$ is very large, $\varphi$ is very close to $f$ from LLN. I will turn to find the posterior distribution of $f$ given $D$.

I will start with **joint distribution of data and parameters**:

$$P(D, \theta, \xi, f, \alpha_T | \alpha_F)$$

$$\propto \prod_{\{i|R_i=1\}} \theta_{X_i}^{Y_i} (1 - \theta_{X_i})^{1-Y_i} \times$$

$$B \prod_{b=1}^{B} \left[ \text{Beta}(\theta_b | \alpha_T f, \alpha_T (1 - f)) \right] \times \text{Beta}(f | \alpha_F, \alpha_F) \pi_T(\alpha_T) \times$$

$$\prod_{i=1}^{n} \theta_{X_i}^{R_i} (1 - \theta_{X_i})^{1-R_i} \times \pi_{\xi}(\xi)$$

Let $\mathcal{I}_X = \{X_1, \ldots, X_n\}$. **Since $B$ is greatly larger than $n$, I will assume that all $X_i$'s are distinct** for finding an approximate estimate of the posterior of $f$. 
Posterior Analysis (Cont’d)

I will next integrate $\theta$ away from the above joint distribution:

(1) Those $\theta_b$ for $b \notin \mathcal{I}_X$ can be integrated away from their prior, leading to 1.

(2) Those $\theta_b$ for $b \in \mathcal{I}_X$ can be integrated away too, leading to a Bernoulli distribution for $Y_i$:

$$\int_0^1 \theta_b^{Y_i b} (1 - \theta_b)^{1-Y_i b} \text{Beta}(\theta_b | \alpha_T f, \alpha_T (1 - f)) \, d\theta_b$$

$$= f^{Y_i b} (1 - f)^{1-Y_i b}.$$

We can also integrate $\xi$ away, leading to an expression unrelated to $f$.

After integrating away $\xi$ and $\theta$, we obtain the joint distribution of data and $f$:

$$P(D, f | \alpha_F) = c \times \prod_{\{i: R_i = 1\}} f^{Y_i} (1 - f)^{1-Y_i} \times \text{Beta}(f | \alpha_F, \alpha_F), \quad (15)$$
Bayes Estimator

The posterior distribution of $f$ is therefore a Beta distribution:

$$P(f \mid \mathcal{D}, \alpha_F) = \text{Beta}(f \mid n_1 + \alpha_F, n_0 + \alpha_F), \quad (16)$$

where

$$n_1 = \sum_{\{i: R_i = 1\}} Y_i = \sum_{i=1}^{n} R_i Y_i, \quad (17)$$

$$n_0 = \sum_{\{i: R_i = 1\}} (1 - Y_i) = \sum_{i=1}^{n} R_i - n_1. \quad (18)$$

The best guess for $f$ that minimizes the expected square loss is the mean of the posterior distribution of $f$, which leads to the Bayes estimator for $f$:

$$\hat{\phi}_{BS} = \frac{n_1 + \alpha_F}{n_0 + n_1 + 2 \alpha_F}. \quad (19)$$

I will compare $\hat{\phi}_{HT}$ and $\hat{\phi}_{BS}$ with criterion of mean square error.
End of Bayesian Analysis
In the first comparison, I set $\xi$ all equal to $\delta$. For each $\delta$, a set of $\theta$ are drawn from a transformed normal random numbers:

$$\theta_b \sim \Phi(Z_b), \quad Z_b \sim N(0, 0.5^2), \text{for } b = 1, \ldots, B$$

(20)
Comparison of MSEs (1)
In the second comparison, I generated pairs of $\theta$ and $\xi$ from transformed multivariate normal numbers with different correlations. The MSE with similar $\xi$ and similar correlations (generated from the same parameters) are plotted in a graph against the true values of $\varphi$:
Comparison of MSEs (2)

\( \lambda = 0.04, \ \rho = -0.69 \)

\( \lambda = 0.04, \ \rho = 0 \)

\( \lambda = 0.04, \ \rho = 0.69 \)

\( \lambda = 0.14, \ \rho = -0.76 \)

\( \lambda = 0.14, \ \rho = 0 \)

\( \lambda = 0.14, \ \rho = 0.76 \)

\( \lambda = 0.5, \ \rho = -0.8 \)

\( \lambda = 0.5, \ \rho = 0 \)

\( \lambda = 0.5, \ \rho = 0.8 \)

\( \lambda = 0.86, \ \rho = -0.76 \)

\( \lambda = 0.86, \ \rho = 0 \)

\( \lambda = 0.86, \ \rho = 0.76 \)

\( \lambda = 0.96, \ \rho = -0.69 \)

\( \lambda = 0.96, \ \rho = 0 \)

\( \lambda = 0.96, \ \rho = 0.69 \)
I generated three pairs of nearly independent $\theta$ and $\xi$ with high $\varphi$ values. With each pair, I simulated 5000 values of two estimators, and drew their histograms:
Histograms of the Simulated Estimators

$\lambda = 0.09, \varphi = 0.91, \rho = -0.01$

Histogram of 5000 Horwitz–Thompson samples

Histogram of 5000 Bayes samples

$\lambda = 0.5, \varphi = 0.91, \rho = 0$

Histogram of 5000 Horwitz–Thompson samples

Histogram of 5000 Bayes samples

$\lambda = 0.91, \varphi = 0.91, \rho = 0$

Histogram of 5000 Horwitz–Thompson samples

Histogram of 5000 Bayes samples
From my comparisons, the simple Bayes estimator isn’t weak for Wasserman’s example. Indeed, it is stronger than Horwitz-Thompson estimator for most parameter configurations.

Indeed, Bayesian inferences have been applied successfully to many high-dimensional and nonparametric problems. Appropriate Bayesian inferences can avoid overfitting problem of MLE, easily model data with complex structure, naturally use prior knowledge to improve inference, and automatically consider uncertainty in inference. They therefore show superiority in many “hard” problems.
Thank You!

Questions and Comments?