Are Bayesian Inferences Weak for Wasserman's Example?

Longhai Li*

(first version: 5 June 2009, this version: 16 December 2009)

Abstract: An example was given in the textbook *All of Statistics* (Wasserman, 2004, pages 186-188) for arguing that, in the problems with a great many parameters Bayesian inferences are weak, because they rely heavily on the likelihood function that captures information of only a tiny fraction of the total parameters. Alternatively he suggested non-Bayesian Horwitz-Thompson estimator, which cannot be obtained from a likelihood-based approaches, including Bayesian approaches. He argued that Horwitz-Thompson estimator is good since it is unbiased and consistent. In this paper, I compared the mean square errors of Horwitz-Thompson estimator with a Bayes estimator at a wide range of parameter configurations. I also simulated these two estimators to visualize them directly. From these comparisons, I conclude that the simple Bayes estimator works better than Horwitz-Thompson estimator for most parameter configurations. Hence Bayesian inferences are not weak for this example.

Keywords: critique of Bayesian inference, Wasserman's example, Horwitz-Thompson estimator, Bayes estimator, decision theory

*Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, Saskatchewan, S7N5E6, CANADA. Email: longhai@math.usask.ca. Web: http://math.usask.ca/~longhai.

1 Introduction

In the textbook by Wasserman (2004), under a Section titled "Strengths and Weaknesses of Bayesian Inference" (pages 186-189), he used an example, which was simplified from Robins and Ritov (1997), to support his comments on the weaknesses of Bayesian inferences:

> The moral of the story is this. Bayesian methods are tied to the likelihood function. But in high dimensional (and nonparametric) problems, the likelihood may not yield accurate inferences. ... Bayesians are slaves to the likelihood function. When the likelihood goes awry, so will Bayesian inference. ... Generally, Bayesian methods run into problems when the parameter space is high dimensional.

Sims (2006) has analyzed the example in details and shown that for this example Bayesian inferences are not weak. However, in his analysis some changes were made to the model, which doesn't change the essence of the example, but makes the argument less convincing. In this paper, I will focus on the very example without any modification, and show that a Bayes estimator is not weaker, actually stronger, than the suggested non-Bayesian Horwitz-Thompson estimator for most parameter configurations.

The above comments on Bayesian inferences have also been refuted silently by numerous successful applications of Bayesian inferences in high dimensional and nonparametric models, such as Bayesian neural network models (see eg. Neal, 1996), Gaussian process regression and classification models (see eg. Rasmussen and Williams, 2006), Dirichlet process mixture models (see eg. MacEachern and Mueller, 1998; Jain and Neal, 2004), and Bayesian logistic regression models with high-order interactions (Li and Neal, 2008). However, many students and young researchers do not have chances of seeing the aforementioned real but complicated Bayesian works, therefore such a simple example in a textbook by a prominent statistician may still be very influential. The purpose of this paper is to reduce such influence, which I think inappropriate.

I will start with describing the example, then propose an estimator derived from a Bayesian approach for the interested parameter, and use simulations to compare it with Horwitz-Thompson estimator, based on the frequentist criterion — mean square error.

2 The example

I will first describe the example presented in Wasserman (2004). In this example, observation Y_i is modeled by a mixture distribution of a *huge* number, B, of Bernoulli distributions,

parameterized by θ_b, for $b = 1, \ldots, B$. Let X_i denote the component label of Y_i. This mixture model can be expressed as:

$$X_i \sim \text{Uniform}(1, \ldots, B), \tag{1}$$

$$Y_i \,|\, X_i \sim \text{Bernoulli}(\theta_{X_i}). \tag{2}$$

The above model is appropriate for sampling survey problems. The B is the size of a population (eg. residents of a country), which are indexed by integers $1, \ldots, B$. We randomly draw an individual (with replacement) from the population, called a surveyee, and then write down the surveyee's index, denoted by X_i. The surveyee is then asked a question with only two choices: 0/1. The answer of the surveyee is denoted by Y_i. Differently from conventional models for such problems, we assume that the surveyee answers the two-choice question randomly, with probability θ_{X_i} for 1. This randomness models that the surveyee's answer may also depend on some factors related to the survey environment, such as the surveyee's mood at the moment.

In practice it is possible that some of these Y_i are unobserved, for example when those sampled individuals refuse to respond. Let's use a binary random variable R_i to record whether Y_i is observed or not. The probability of R_i equal to 1 depends on the value of X_i, ie, it is a property associated with each individual. Let ξ_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

$$R_i \,|\, X_i \sim \text{Bernoulli}(\xi_{X_i}). \tag{3}$$

In addition, we assume that R_i and Y_i are independently distributed given X_i. This equivalently assumes that the environmental randomnesses related to whether the surveyee chooses to answer the question and how he/she answers the question are different. Suppose we have surveyed n individuals. The data on ith surveyee are either $(X_i, R_i = 1, Y_i)$ or $(X_i, R_i = 0)$, with Y_i missing when $R_i = 0$. I will denote these data on n surveyees collectively by $\mathcal{D}$.

I write the model parameters collectively as $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_B)$, and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_B)$. The likelihood function of $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ based on the data $\mathcal{D}$ is the product of the joint distributions of either $(X_i, R_i = 1, Y_i)$ or $(X_i, R_i = 0)$, for $i = 1, \ldots, n$:

$$L(\boldsymbol{\theta}, \boldsymbol{\xi}\,; \mathcal{D}) = \frac{1}{B^n} \prod_{i=1}^{n} \xi_{X_i}^{R_i} (1 - \xi_{X_i})^{1-R_i} \prod_{\{i:\ R_i = 1\}} \theta_{X_i}^{Y_i} (1 - \theta_{X_i})^{1-Y_i} \tag{4}$$

The parameter we are interested in estimating is the average of $\theta_1, \ldots, \theta_B$, the unconditional

probability $P(Y_i = 1)$, denoted by φ:

$$\varphi = \frac{1}{B} \sum_{b=1}^{B} \theta_b. \tag{5}$$

The likelihood function (4) contains information of at most n of unique θ_b's. Wasserman assumed that B is greatly larger than n. This is realistic in sampling survey problems where B is the size of population while n is a small number of surveyees. As B is greatly larger than n, the likelihood function contains information of only a tiny fraction of $\boldsymbol{\theta}$. Wasserman (2004) therefore argued that the posterior distribution of $\boldsymbol{\theta}$ is almost equal to the whatever prior distribution, therefore cannot lead to a good inference for the interested parameter, φ. It is true that, based on the likelihood function, we cannot obtain much information of a particular θ_b, because it is associated with very few cases, mostly only 1. However, we can infer fairly well the interested parameter φ, because it is only a summary of the total parameters. We can understand this by assuming that the θ_b can take only 0 or 1, that is, simplifying to the conventional models for sampling survey problems, where each surveyee is thought of having a firm opinion of the question asked, unaffected by survey environment. In such problems, we are confident of the method of estimating the population mean with the average when B is greatly larger than n. In this paper I will show that the method based on the average of those Y_i with $R_i = 1$ is also reasonable when $\boldsymbol{\theta}$ can take values between 0 and 1.

3 A Bayesian method for estimating the parameter

Taking a Bayesian approach, we first need to assign a prior for the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$. We can reasonably assume (θ_b, ξ_b) is independent of each other for different b given some hyperparameters.

I choose the prior for $\boldsymbol{\theta}$ as follows:

$$\theta_1, \ldots, \theta_B \,|\, \alpha_T, f \overset{\text{IID}}{\sim} \text{Beta}(\theta \,|\, \alpha_T f,\, \alpha_T (1 - f)), \tag{6}$$

where $\text{Beta}(\theta \,|\, a, b)$ is the probability density function with parameters a and b. In the above prior distribution, the mean of $\boldsymbol{\theta}$ is f, and α_T controls the width of the distribution of $\boldsymbol{\theta}$. When B is huge, the actual average of $\boldsymbol{\theta}$ — φ — is very close to the hyperparameter f, as justified by the Laws of Large Numbers. I therefore turn to infer the single parameter f. I will find a Bayes estimator for f, and then look at its performance in estimating φ.

We can similarly specify the prior for ξ. However if we assume $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are independent, the prior of ξ is irrelevant to the posterior of f. I therefore leave it a general form, denoted by $\pi_\xi(\xi)$. Note that in real problems, a more appropriate prior may be that $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are correlated. For example, in some political or religious surveys, individuals with high values of θ_b (eg. a minority group favoring a political or religious point-view) may be more likely to answer the question. Such correlations can be modeled with a joint prior distribution for θ_b and ξ_b. This is one of advantages of Bayesian inferences. Here I assume that they are independent, such that I can derive a simple Bayes estimator (see (15)) to compare with Horwitz-Thompson estimator suggested by Wasserman (2004). Otherwise, I have to use some numerical methods, such as Markov chain Monte Carlo (MCMC) (see eg Neal, 1993), to infer the parameters. *Note that, however, the Bayes estimator derived from independent priors will be compared in terms of mean square error with correlated* $\boldsymbol{\theta}$ *and* $\boldsymbol{\xi}$. Therefore this independence assumption doesn't undermine the results of this paper.

We don't have good knowledge to fix f and $\alpha_{\scriptscriptstyle T}$, for which we need to assign a higher level prior distribution. I assign f and $\alpha_{\scriptscriptstyle T}$ the following distributions:

$$(f, \alpha_{\scriptscriptstyle T}) \sim \text{Beta}(f \,|\, \alpha_{\scriptscriptstyle F}, \alpha_{\scriptscriptstyle F}) \times \pi_{\scriptscriptstyle T}(\alpha_{\scriptscriptstyle T}), \tag{7}$$

with $\alpha_{\scriptscriptstyle F}$ fixed at some value. The $\alpha_{\scriptscriptstyle F}$ controls the concentration of f around $1/2$. A natural choice of $\alpha_{\scriptscriptstyle F}$ is 1, which implies that f is uniformly distributed over $(0, 1)$. If we choose a smaller value, the posterior of f depends more on the data, possibly closer to 0 or 1. We also need to give $\alpha_{\scriptscriptstyle T}$ a prior, such as an Inverse-Gamma or log-normal distribution (Gelman et al., 2004). However, as we will see, in the problems where B is greatly larger than n, the posterior of f is approximately independent of the value of $\alpha_{\scriptscriptstyle T}$. I therefore leave its prior a general form, as it is irrelevant to the posterior of f.

The joint distribution of the data $\mathcal{D}$, all model parameters and hyperparameters is written as:

$$\begin{aligned} & P(\mathcal{D}, \boldsymbol{\theta}, \boldsymbol{\xi}, f, \alpha_{\scriptscriptstyle T} \,|\, \alpha_{\scriptscriptstyle F}) \\ = \;& L(\boldsymbol{\theta}, \boldsymbol{\xi}; \mathcal{D}) \times \prod_{b=1}^{B} \left[\text{Beta}(\theta_b \,|\, \alpha_{\scriptscriptstyle T}\, f,\, \alpha_{\scriptscriptstyle T}(1-f))\, \pi_\xi(\xi_b)\right] \times \text{Beta}(f \,|\, \alpha_{\scriptscriptstyle F}, \alpha_{\scriptscriptstyle F}) \times \pi_{\scriptscriptstyle T}(\alpha_{\scriptscriptstyle T}). \end{aligned} \tag{8}$$

Since we are interested only in the posterior of f, I first integrate $\boldsymbol{\theta}$ away from the above joint distribution (8). Based on the assumption that B is greatly larger than n, I assume that all X_i, for $i = 1, \ldots, n$, are different, ie, none of the individuals were surveyed twice. (Note that in order to have all different X_i, we can also modify the sampling procedure

from with-replacement to without-replacement, with the distribution of $(X_1, \ldots, X_n)$ in (1) replaced with a uniform distribution over all possible combinations of choosing n items from B item. This modification changes only the factor $1/B^n$, without changing the likelihood function of $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$. In this paper I don't rely on this modification in order to keep the problem considered by Wasserman (2004) intact and therefore don't undermine the results of this paper.) Let $\mathcal{I}_X = \{X_1, \ldots, X_n\}$. The parameters $\{\theta_b : b \notin \mathcal{I}_X\}$ can be integrated away easily, as the integrands are just their prior probability density functions. For θ_b, with $b \in \mathcal{I}_X$, the likelihood term is associated with only *one* observation, whose index is denoted by $i(b)$, as we assume that no two of X_i are the same, we can therefore analytically integrate this θ_b away, as shown in details as follows:

$$\begin{aligned}
&\int_0^1 \theta_b^{Y_{i(b)}} \, (1-\theta_b)^{1-Y_{i(b)}} \, \mathrm{Beta}(\theta_b \,|\, \alpha_T f, \alpha_T(1-f)\,) \, d\,\theta_b \\
&= \frac{\Gamma(\alpha_T)}{\Gamma(\alpha_T\, f)\Gamma(\alpha_T\,(1-f))} \, \frac{\Gamma(\alpha_T\, f + Y_{i(b)}) \; \Gamma(\alpha_T(1-f) + (1-Y_{i(b)})}{\Gamma(\alpha_T + 1)} \qquad (9) \\
&= f^{Y_{i(b)}} (1-f)^{1-Y_{i(b)}}. \qquad (10)
\end{aligned}$$

That is, with θ_b integrated out, $Y_{i(b)}$ is a Bernoulli random variable with parameter f, whose distribution is unrelated to α_T. Note that there isn't such a simple expression when there are more than one observations associated with θ_b. The $\boldsymbol{\xi}$ and α_T can also be integrated away, resulting in an expression without f. Integrating $\boldsymbol{\theta}$, $\boldsymbol{\xi}$ and α_T out gives the joint distribution of data $\mathcal{D}$ and model parameters f:

$$P(\mathcal{D}, f \,|\, \alpha_F) = c \times \prod_{\{i:R_i=1\}} f^{Y_i}(1-f)^{1-Y_i} \times \mathrm{Beta}(f \,|\, \alpha_F, \alpha_F), \qquad (11)$$

where c is a factor without f. The expression (11), with c omitted, is the joint distribution of a standard Bernoulli-Beta model (Gelman et al., 2004). One can readily see that the posterior distribution of f is

$$P(f \,|\, \mathcal{D}, \alpha_F) = \mathrm{Beta}(f \,|\, n_1 + \alpha_F, n_0 + \alpha_F), \qquad (12)$$

where

$$\begin{aligned}
n_1 &= \sum_{\{i:R_i=1\}} Y_i = \sum_{i=1}^{n} R_i \, Y_i, \qquad (13) \\
n_0 &= \sum_{\{i:R_i=1\}} (1-Y_i) = \sum_{i=1}^{n} R_i - n_1. \qquad (14)
\end{aligned}$$

If we want to minimize the square error in guessing f, the best estimate is the mean of the posterior distribution (Schervish, 1995), given by:

$$\hat{\varphi}_{BS} = \frac{n_1 + \alpha_F}{n_0 + n_1 + 2\,\alpha_F}. \tag{15}$$

I denote the above fraction as $\hat{\varphi}_{BS}$ as I will use it to estimate φ. $\hat{\varphi}_{BS}$ is strongly related to the fraction of 1 in observed data Y_i, with slight modification by α_F, which avoids extreme conclusion when n_1 or n_0 is nearly 0. I will call $\hat{\varphi}_{BS}$ *Bayes estimator*.

4 Comparing with Horwitz-Thompson estimator

Wasserman (2004) suggested the following Horwitz-Thompson estimator for estimating φ:

$$\hat{\varphi}_{HT} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i\, Y_i}{\xi_{X_i}}, \tag{16}$$

where, when $R_i = 0$, as a statistic of data, Y_i is imaginary and can be assigned arbitrarily, but as a random variable, it is real, and is distributed as defined by (1) and (2). This estimator treats both observed $Y_i = 0$, and Y_i being missing as 0, and count each observed $Y_i = 1$ $1/\xi_{X_i}$ (which is assumed to be known) times. One can easily show that this estimate has mean φ, by iterative expectation formula. This estimator is also consistent, as shown below.

To compare the performance of $\hat{\varphi}_{BS}$ and $\hat{\varphi}_{HT}$, I will take a frequentist approach — comparing their mean square errors (MSE). Note that although the Bayes estimator is derived by assuming some form of prior over the parameters, the property of its mean square error is unrelated to the choice of prior.

The mean square error of $\hat{\varphi}_{HT}$ is equal to its variance:

$$\begin{aligned}
\mathrm{MSE}(\hat{\varphi}_{HT}) &= \frac{1}{n} \mathrm{Var}\left(\frac{R_i Y_i}{\xi_{X_i}}\right) && (17)\\
&= \frac{1}{n}\left(E\left(\frac{R_i Y_i}{\xi_{X_i}^2}\right) - \varphi^2\right) && (18)\\
&= \frac{1}{n}\left(\frac{1}{B}\sum_{b=1}^{B}\frac{\theta_b}{\xi_b} - \varphi^2\right). && (19)
\end{aligned}$$

From (19), we can see that $\hat{\varphi}_{HT}$ is consistent, because its MSE will converge to 0, as $n \to +\infty$.

However, this doesn't mean that it works well when n is finite, especially when n is greatly smaller than B, as assumed by Wasserman (2004) to argue against Bayesian inferences. To look at this, let's exam $\hat{\varphi}_{HT}$ in a special case when $\boldsymbol{\xi}$ are all the same, equal to δ, then the expression in (19) is simplified as:

$$\mathrm{MSE}(\hat{\varphi}_{HT}) = \frac{1}{n}(\varphi/\delta - \varphi^2). \tag{20}$$

We can see that with n and φ fixed when $\delta \to 0$, $\mathrm{MSE}(\hat{\varphi}_{HT})$ converges to $+\infty$. From this special case we can see clearly that $\hat{\varphi}_{BS}$ will work very poorly when most of $\boldsymbol{\xi}$ are small. In contrast, when $\delta \to 0$, $\sum_{i=1}^{n} R_i$ converges to 0, therefore $\hat{\varphi}_{BS}$ converges to 1/2, with consequence of that the mean square error converging to $(\varphi - 1/2)^2 \leq 1/4$.

We may not find a simple expression of the MSE of $\hat{\varphi}_{BS}$. I therefore used Monte Carlo simulations to estimate it. The simulations were implemented in R language (Team, 2008), with R function `sim_bayes` given in Section Appendix. In all of the following comparisons in this section, I set the number of simulations of $\hat{\varphi}_{BS}$ equal to 1000, and α_F in $\hat{\varphi}_{BS}$ equal to 0.1. In this paper I only show experiment results with the number of sample size, n, set to 100, and the population size, B, set to 100000. The results are similar to other choices of small n and large B.

The MSEs of $\hat{\varphi}_{BS}$ and $\hat{\varphi}_{HT}$ were compared at a wide variety of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ converted from multivariate normal random numbers with the standard normal cumulative function Φ. More specifically, (ξ_b, θ_b), for $b = 1, \ldots, B$, were generated independently from the distribution defined as follows:

$$\begin{aligned} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} &\sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & r\,\sigma_1\,\sigma_2 \\ r\,\sigma_1\,\sigma_2 & \sigma_2^2 \end{pmatrix} \right) \\ \xi &= \Phi(X_1), \quad \theta = \Phi(X_2). \end{aligned} \tag{21}$$

When μ_1 is larger, the distribution of ξ is more skewed to 1, otherwise more skewed to 0. It is similar for θ. The value of r is strongly related to the correlation of ξ and θ. Three such pairs of random numbers of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ drawn as above with different values of r are plotted in Figure 1, with parameters indicated in the titles.

In the first comparison, I set all of $\boldsymbol{\xi}$ to a common value δ, which takes 15 values evenly spaced between 0.02 and 0.98. Then for each δ, 20 sets of $\boldsymbol{\theta}$ were generated with the method (21) ($\boldsymbol{\xi}$ were discarded), using 20 values of μ_2 evenly spaced from -2 to 2, and σ_1 and σ_2 fixed at 0.5. Note that since $\boldsymbol{\xi}$ are all the same, r doesn't matter here. The MSE of $\hat{\varphi}_{HT}$ can

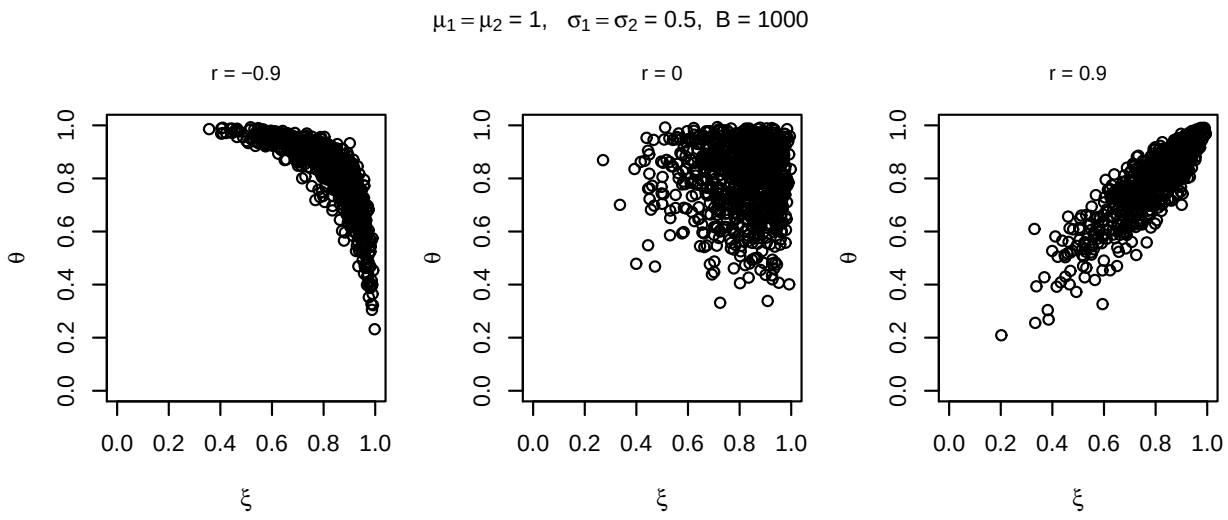

Figure 1: Three pairs of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ drawn with different values of r.

be computed directly with (20), and the MSE of $\hat{\varphi}_{BS}$ is approximated by simulating 1000 samples and computing the average of square errors in guessing φ (see R function `mse_bs` given in Section Appendix). The MSEs are plotted against the average of $\boldsymbol{\theta}$, namely φ, as shown in Figure 2, where I used gray and dashed lines to display 95% confidence intervals of the MSE of $\hat{\varphi}_{BS}$, computed with the Central Limit Theorem.

From Figure 2, we can see clearly that in terms of MSE, $\hat{\varphi}_{BS}$ works better than $\hat{\varphi}_{HT}$ for most parameter configurations, with the gap wider when δ is smaller or φ is larger. When δ is smaller than 0.5, the MSEs of $\hat{\varphi}_{HT}$ keep increasing as φ is approaching to 1. This is very unusual, as when φ is closer to 0 or 1, a reasonable estimator should guess φ more accurately due to reduced uncertainty. This is just the case of $\hat{\varphi}_{BS}$, whose MSEs peak at $1/2$ symmetrically. When δ is closer to 1, the performance of these two estimators becomes similar, which is not surprising, as when all of $\boldsymbol{\xi}$ are 1, these two estimators are almost the same, except that $\hat{\varphi}_{BS}$ is modified by α_F. When both φ and δ are very small, we see that MSEs of $\hat{\varphi}_{HT}$ are lower than that of $\hat{\varphi}_{BS}$, as in such situations, the bias in $\hat{\varphi}_{BS}$ becomes dominating in its MSE, as is usual for biased statistic with smaller MSE.

In the second comparison, I generated pairs of possibly correlated $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ with equations (21). Fixing $\sigma_1 = \sigma_2 = 0.5$, I generated one pair of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ with each combination of μ_1 (related to $\boldsymbol{\xi}$) in set $\{-2, -1.2, 0, 1.2, 2\}$, r in set $\{-0.85, 0, 0.85\}$, and μ_2 (related to $\boldsymbol{\theta}$) in a set of 20 values evenly spaced between -2 and 2. The MSEs of $\hat{\varphi}_{BS}$ and $\hat{\varphi}_{HT}$ with the same μ_1 and r are plotted along against the values of φ, as shown in Figure 3. The titles of plots in Figure 3 show the average of values of λ (λ = the average of $\boldsymbol{\xi}$), and the average of correlations between $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ in the plot, respectively denoted by $\bar{\lambda}$ and $\bar{\rho}$. Similar to the

previous cases where $\boldsymbol{\xi}$ are fixed, from Figure 3, we can conclude that $\hat{\varphi}_{BS}$ works better than $\hat{\varphi}_{HT}$ for most parameter configurations. When $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ are more correlated positively the differences of MSEs of $\hat{\varphi}_{HT}$ to $\hat{\varphi}_{BS}$ become smaller. Particularly, $\hat{\varphi}_{HT}$ has slightly lower MSE than $\hat{\varphi}_{BS}$ when λ is around $1/2$ and φ is less than $1/2$ (the middle three plots in the rightest column). However, when φ is greater than 0.5, $\hat{\varphi}_{HT}$ is still inferior than $\hat{\varphi}_{BS}$. The MSEs of $\hat{\varphi}_{BS}$ are nearly symmetrical about $1/2$. For $\hat{\varphi}_{HT}$, when λ is small the MSEs increase as φ approaches to 1.

To look directly at the difference of these two estimators, I generated Monte Carlo samples of them at some particular parameter configurations. I generated three pairs of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ with $\mu_1 = -1.5, 0, 1.5$ respectively, and fixing $\mu_2 = 1.5$, $r = 0$, $\sigma_1 = \sigma_2 = 0.5$. Here I want to look at the two estimators when φ is high, the situations where $\hat{\varphi}_{BS}$ works much better than $\hat{\varphi}_{HT}$ from previous investigations with MSE. I generated 5000 samples for each estimator for each pair of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. The histograms of these samples are displayed in Figure 4. The actual values of λ, φ, and correlation, ρ, between $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ are indicated in the titles of plots. We can see that the samples of $\hat{\varphi}_{BS}$ are more concentrated around the true φ than those of $\hat{\varphi}_{HT}$. This explains why the MSE of $\hat{\varphi}_{BS}$ is smaller than $\hat{\varphi}_{HT}$. In addition, the histograms of $\hat{\varphi}_{HT}$ reveal that there is a fairly large probability that $\hat{\varphi}_{HT}$ exceeds 1, the largest possible value of φ. This clearly shows that $\hat{\varphi}_{HT}$ isn't a good estimator for φ. In contrast, $\hat{\varphi}_{BS}$ would never estimate φ with a value greater than 1, as seen from its expression (15). $\hat{\varphi}_{HT}$ is unbiased but its large variance makes it worse than $\hat{\varphi}_{BS}$ in terms of mean square error.

5 Closing remarks

Wasserman (2004) used two examples to support his comments that likelihood-based inferences, including Bayesian inferences, are weak for high dimensional and nonparametric models. For the first example, the comparisons of this paper have shown clearly that a simple Bayes estimator, $\hat{\varphi}_{BS}$, works better than the suggested non-Bayesian estimator, $\hat{\varphi}_{HT}$, for most parameter configurations. Hence Bayesian inferences are not weak for this example. The second example is more easily found erroneous (see Sims, 2006). Hence these examples cannot be used to argue that Bayesian inferences are weak for high-dimensional or non-parametric models.

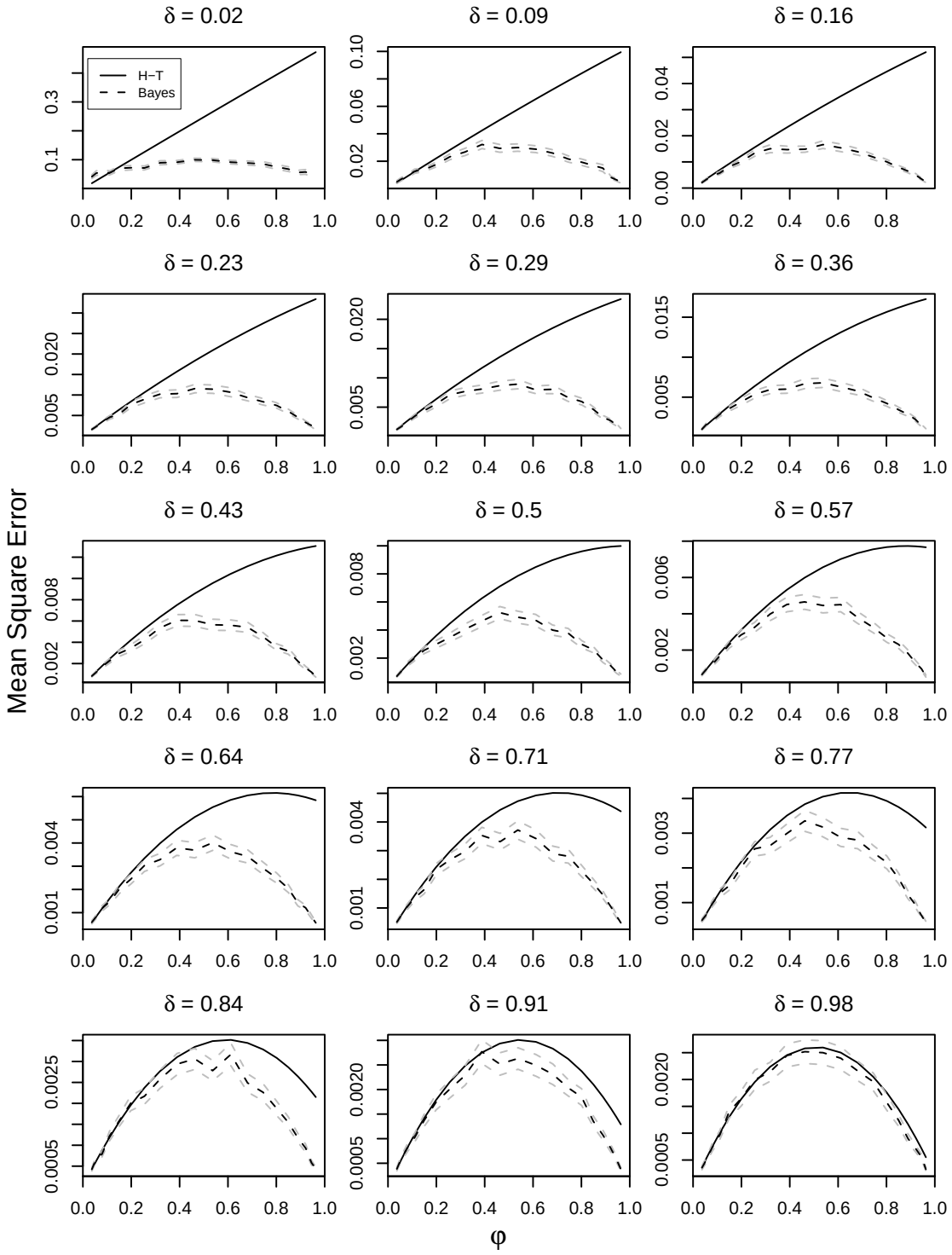

Figure 2: Comparison of mean square errors of $\hat{\varphi}_{BS}$ and $\hat{\varphi}_{HT}$, when $\boldsymbol{\xi}$ are fixed at δ, and $\boldsymbol{\theta}$ are randomly drawn with equations (21). The gray and dashed lines show 95% confidence intervals in the Monte Carlo estimates of MSE of $\hat{\varphi}_{BS}$. In all plots, x-axis is φ, and y-axis is MSE.

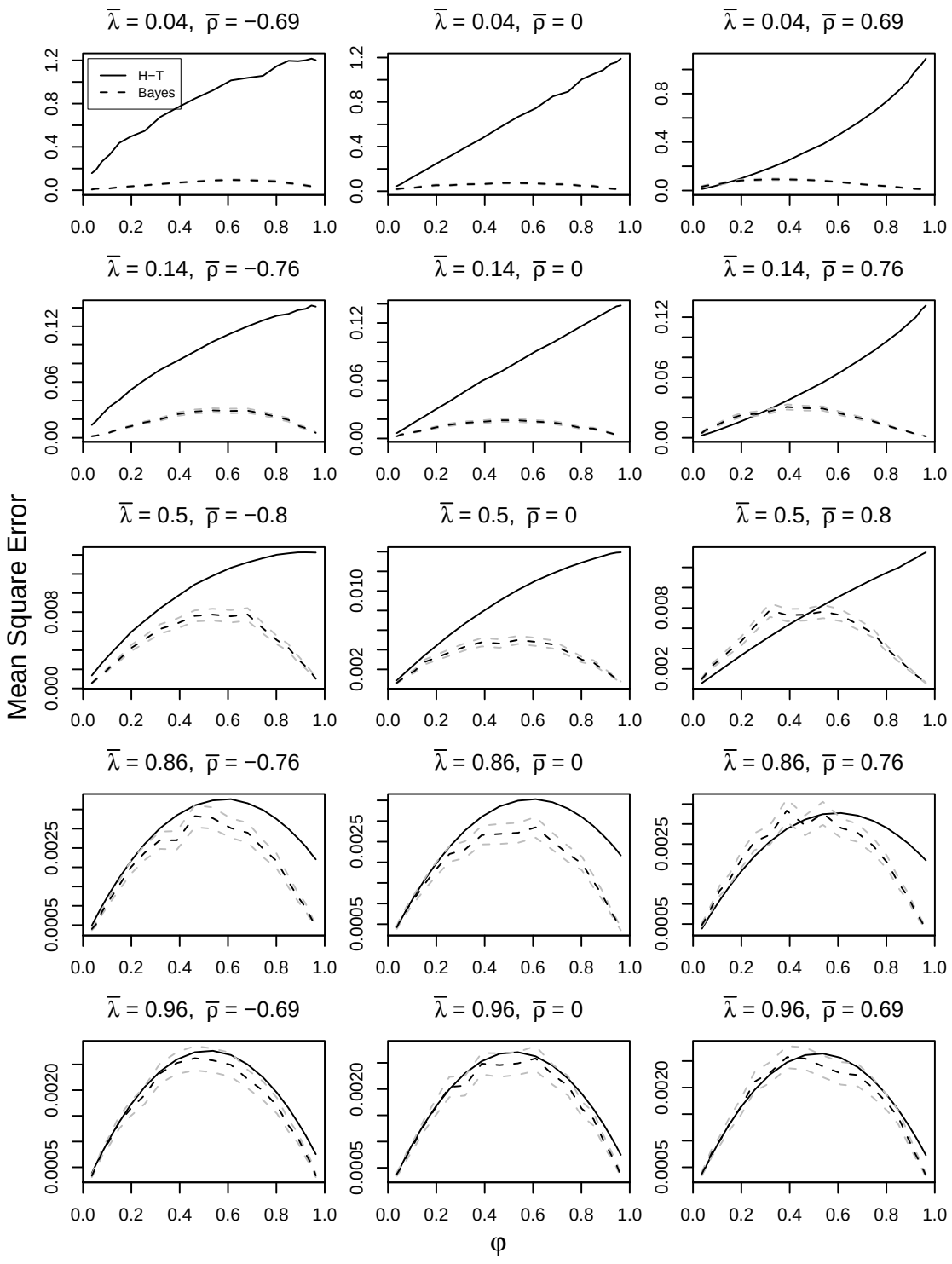

Figure 3: Comparison of mean square errors of $\hat{\varphi}_{BS}$ and $\hat{\varphi}_{HT}$, when $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are randomly drawn with equations (21). The gray and dashed lines show 95% confidence intervals in the Monte Carlo estimates of MSE of $\hat{\varphi}_{BS}$. In all plots, x-axis is φ, and y-axis is MSE. In the titles, $\bar{\lambda}$ is the average of the averages of $\boldsymbol{\xi}$ in all 20 data sets with different values of $\boldsymbol{\theta}$, and $\bar{\rho}$ is the average of the correlations of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ in all 20 data sets.

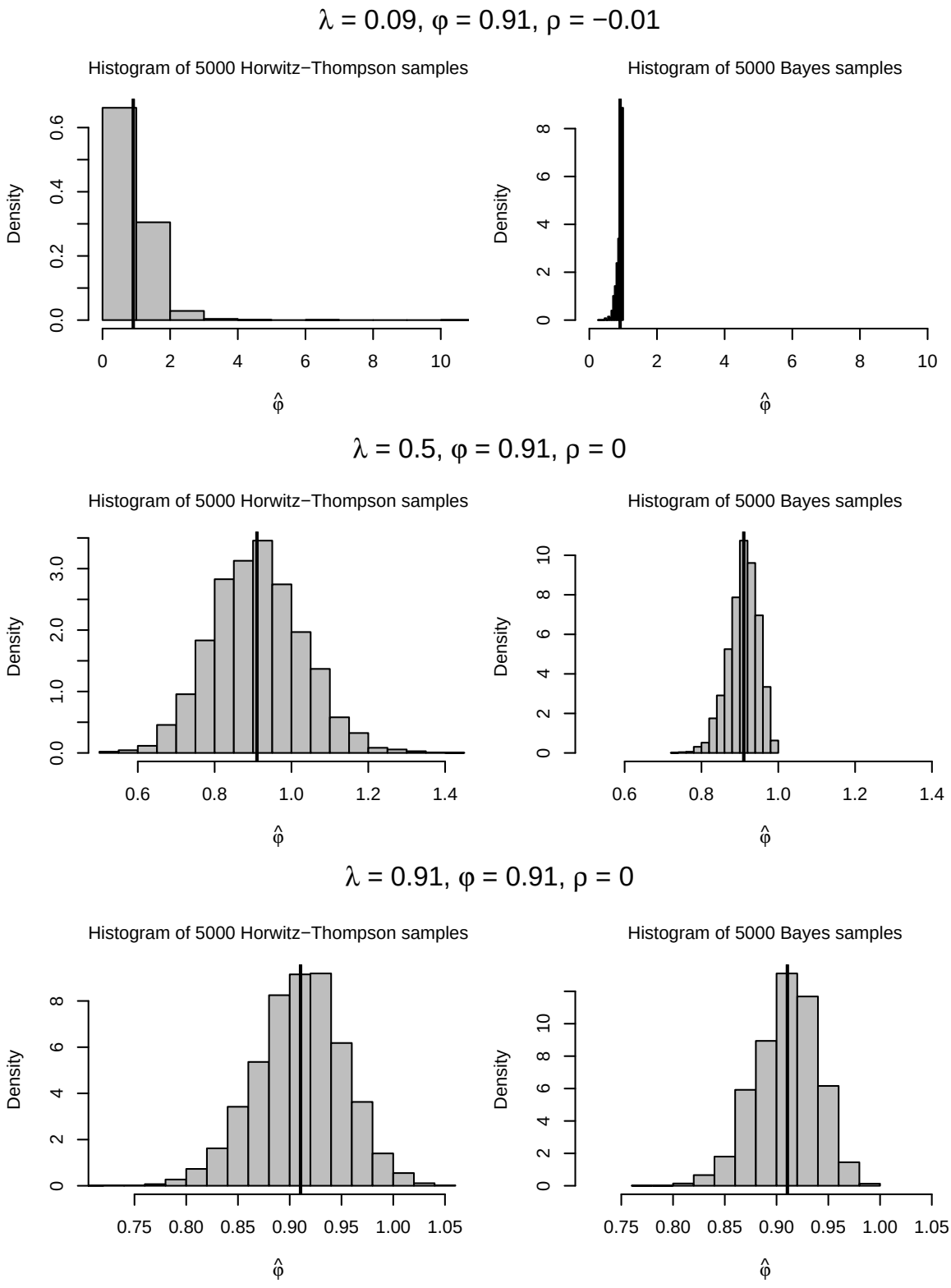

Figure 4: Histogram of simulated samples of $\hat{\varphi}_{HT}$ and $\hat{\varphi}_{BS}$. The black vertical lines indicate the true value of φ. In the titles, λ is the average of $\boldsymbol{\xi}$, φ is the average of $\boldsymbol{\theta}$, and ρ is the correlation between them.

Appendix R functions for simulations

```
## simulating horwitz-thompson estimator
## no_sim  --- number of simulations of phi
## n       --- number of observations in estimating phi
## theta   --- the given theta values, a vector of length B
## xi      --- the given xi values, a vector of length B
sim_horwitz_thompson <- function (no_sim, n, theta, xi)
{
   B <- length (theta)
   if(B != length (xi)) stop ("theta and xi don't match")

   one_sim <- function ()
   {
      X <- sample (1:B, n, replace = T)
      ## draw Bernoulli random numbers  with probabilities xi[X]
      R <- (runif(n) < xi[X]) * 1
      Y <- (runif(n) < theta[X]) * 1
      ## compute Horwitz-Thompson estimator
      mean (R * Y / xi[X])
   }

   replicate ( no_sim, one_sim() )
}

## simulating Bayes estimator
## no_sim  --- number of simulations of phi
## n       --- number of observations in estimating phi
## theta   --- the given theta values, a vector of length B
## xi      --- the given xi values, a vector of length B
## alpha   --- parameter of Beta prior for 'phi'
sim_bayes <- function(no_sim, n, theta, xi, alpha)
{
   B <- length(theta)
   if(B != length(xi)) stop("theta and xi don't match")

   one_sim <- function()
   {
      X <- sample(1:B, n, replace = T)
```

```
      ## draw Bernoulli random numbers  with probabilities xi[X]
      R <- (runif(n) < xi[X]) * 1
      Y <- (runif(n) < theta[X]) * 1

      ## compute Bayes estimator
      (sum (Y[R==1]) + alpha) / (sum (R == 1) + 2 * alpha)
   }

   replicate (no_sim, one_sim())
}

## compute the mean square error (mse) of Bayes estimator for a vector of 'phi'
## arguments:
## phi          ---   a vector of averages of 'theta'
## lambda       ---   average of 'xi', a scalar
## alpha        ---   parameter of Beta prior for 'phi'
## n            ---   sample size in estimator
## value:
## a vector of mse for each value in 'phi', with 'lambda' fixed at a value
mse_bs <- function(theta, xi, alpha = 0.1, n, no_sim = 1000)
{
   square_errors <- (sim_bayes (no_sim,n,theta,xi,alpha) - mean(theta) ) ^2
   mse <- mean( square_errors )
   sd <- sd (square_errors) / sqrt (no_sim)
   list (mse = mse, sd = sd)
}
```

References

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Texts in Statistical Science, Chapman and Hall/CRC. 5, 6

Jain, S. and Neal, R. M. (2004), "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model," *Journal of Computational and Graphical Statistics*, 13, 158–182. 2

Li, L. and Neal, R. M. (2008), "Compressing Parameters in Bayesian High-order Models with Application to Logistic Sequence Models," *Bayesian Analysis*, 3, 793–822. 2

MacEachern, S. N. and Mueller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–238. 2

Neal, R. M. (1993), "Probabilistic Inference using Markov Chain Monte Carlo Methods," Tech. rep., Dept. of Computer Science, University of Toronto. 5

— (1996), *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics No. 118, New York: Springer-Verlag. 2

Rasmussen, C. E. and Williams, C. K. I. (2006), *Gaussian processes for machine learning*, Springer. 2

Robins, J. M. and Ritov, Y. (1997), "Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models," *Statistics in Medicine*, 16, 285–319. 2

Schervish, M. J. (1995), *Theory of Statistics*, Springer Series in Statistics, Springer. 7

Sims, C. A. (2006), "On An Example Of Larry Wasserman," online manuscript, available from `http://sims.princeton.edu/yftp/WassermanExmpl/WassermanComment.pdf`. 2, 10

Team, R. D. C. (2008), *R: A Language and Environment for Statistical Computing*, Vienna, Austria, iSBN 3-900051-07-0. 8

Wasserman, L. (2004), *All of statistics: a concise course in statistical inference*, Springer. 1, 2, 4, 5, 6, 7, 8, 10

Acknowledgements

This work was supported by Natural Sciences and Engineering Research Council of Canada. I thank Weixin Yao for providing helpful comments on an earlier draft of this paper.