Bayesian Estimators for Robins-Ritov's Problem

Stefan Harmeling
School of Informatics
University of Edinburgh
5 Forrest Hill
Edinburgh EH1 2QL, Scotland
stefan.harmeling@ed.ac.uk

Marc Toussaint
TU Berlin
Franklinstr. 28/29, FR 6-9
10587 Berlin, Germany
mtoussai@cs.tu-berlin.de

Both authors have equally contributed to this work.

Informatics Research Report EDI-INF-RR-????

SCHOOL *of* INFORMATICS
Institute of Adaptive and Neural Computation

October 2007

Abstract : Bayesian or likelihood-based approaches to data analysis became very popular in the field of Machine Learning. However, there exist theoretical results which question the general applicability of such approaches; among those a result by Robins and Ritov which introduce a specific example for which they prove that a likelihood-based estimator will fail (i.e. it does for certain cases not converge to a true parameter estimate, even given infinite data). In this paper we consider various approaches to formulate likelihood-based estimators in this example, basically by considering various extensions of the presumed generative model of the data. We can derive estimators which are very similar to the classical Horvitz-Thompson and which also account for a priori knowledge of an observation probability function.

Keywords : Robins Ritov Problem, Likelihood-based Inference, Bayesian Inference, Horvitz-Thompson Estimator, Hierarchical Model

1 Introduction

Robins and Ritov [1997] begin their paper as follows:

> "In the analysis of data obtained by stratified random sampling, 'likelihoodist' and Bayesian statisticians often claim that inference concerning the population mean should be the same regardless of whether the stratum-specific randomization (selection) probabilities are or are not known to the data analyst."

In this paper we explore some 'likelihoodist' approaches to derive estimators for such kinds of problems and also investigate them empirically. We do not at all intent to prove the theorems on the problems with Bayesian and likelihood-based (LB) approaches wrong—on the contrary we support them empirically in a special case. However, there are alternative approaches to formulate LB estimators, basically by considering alternative latent generative processes for the data at hand. In fact, for a certain assumed generative process we can derive an LB estimator that is very similar to the classical Horvitz-Thompson (HT) estimator. From a Bayesian point of view one might then inversely argue that the HT estimator is "as if implicitly assuming such a generative process".

The outline of this paper is as follows. In the next section we investigate a simplified version of the problem without hidden observations. In this scenario one can already conclude that in the LB approach any estimator that tries to avoid ignoring the data completely, must assume dependencies between certain variables. This is realized by including an additional hyperparameter in the assumed generative process. This discussion solves the "weaknesses" of Bayesian inference brought forward in Wasserman [2005] in the context of Robins-Ritov's problem. Section 3 then considers the case in which some of the data can not be observed. A straight-forward LB estimator performs well in cases in which the observation probability ξ is not correlated to the mean θ of the observation. Here, however, we get to the core of the actual discussion of Robins and Ritov [1997]: The straight-forward LB estimator ignores the observation probabilities and is biased if in fact ξ and θ are correlated. Robins-Ritov correctly point out that the HT estimator still works even in this correlated case if ξ is known—and they proof that, in their framework, a likelihood based estimator will fail to converge correctly by constructing such a correlated case. In section 4 we reason about the case that ξ and θ might be dependent, and conclude that we need a non-factorized joint prior, and we show that this dependency *does not drop out of the likelihood* in the framework with the hyperparameter μ. The LB estimator we derive for a certain joint prior $P(\xi, \theta)$ is very similar to the HT estimator, except for an empirically grounded normalization instead of normalizing by the number of data points. Experiments show that the LB estimator converges considerably faster (the variance decreases faster). Finally, in section 5, we consider a continuous domain and Gaussian Processes as a prior over θ's. Interestingly, the prior smoothness of θ as given by the kernel function has a very intuitive effect on the estimator, which is analogous to observing multiple observations for one X in the discrete case.

2 No transfer without hyperparameters

Let us first consider a preliminary problem in which all data is observed, but which pinpoints already one crucial aspect, namely whether one can generalize from observed sites to unobserved sites. The next section will address the problem as presented in Robins and Ritov [1997] with similar notations as in Wasserman [2005].

Problem 1 (discrete X, no missing data) *The observed data D consists of n pairs (X_i, Y_i) with $1 \leq i \leq n$ which are sampled i.i.d. as follows:*

$$X_i \sim \text{uniform on} \{1, \ldots, C\} \tag{1}$$

$$Y_i \sim \mathcal{N}(\theta_{X_i}, 1) \tag{2}$$

with $C \gg n$, and with the unknown vector $\theta = (\theta_1, \theta_2, \ldots, \theta_C)$. The task is to find an estimator for

$$\psi = \frac{1}{C} \sum_{j=1}^{C} \theta_j \ . \tag{3}$$

Let us introduce some notations. We interpret $X = j$ as randomly choosing a site j from which we obtain an observation $Y \sim \mathcal{N}(\theta_j, 1)$. The data set D of size n will include samples from only J different sites, some of which might be sampled repeatedly. W.l.o.g. we assume that the domain of X is sorted such that $\{1, \ldots, J\}$ are the sites we actually have observations for, while $\{J+1, \ldots, C\}$ have not been sampled. Let $n_j := \#\{i : X_i = j\}$ be the number of observations we have for site j. Then we use the notation $\boldsymbol{Y}_j$ for the n_j-dimensional vector containing all the observations we received form site j.

A simple unbiased estimator to solve this problem is

$$\hat{\psi} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{4}$$

In constrast, a naive likelihood-based approach is to compute the MLE estimates $\hat{\theta}_j^{\mathrm{MLE}}$ for all $1 \leq j \leq C$ and then simply propose $\hat{\psi} = 1/C \ \sum_j \hat{\theta}_j^{\mathrm{MLE}}$ as an estimator for ψ. However, as Wasserman [2005] pointed out, this approach has a severe problem: after examining the likelihood,

$$P(D|\theta) = \prod_{i=1}^{n} \frac{1}{C} \, \mathcal{N}_{Y_i}(\theta_{X_i}, 1) \ , \tag{5}$$

(where $\mathcal{N}_z(\mu, \sigma^2)$ denotes the (possibly multivariate) probability density of z of the Gaussian distribution with mean μ and variance σ^2), we see that we can not compute an MLE estimate $\hat{\theta}_j^{\mathrm{MLE}}$ for unobserved sites j because θ_j does not appear in the likelihood. The fact that we only observed a small fraction of the domain ($n \ll C$) corrupts the approach. Also retreating to MAP estimates $\hat{\theta}^{\mathrm{MAP}}$ does not help, because for unobserved j the posterior is equal to the prior, i.e. $P(\theta_j \big| D) = P(\theta_j)$. Thus for very large C, the estimate via $\hat{\psi} = 1/C \ \sum_j \hat{\theta}_j^{\mathrm{MAP}}$ converges against the mean of the prior $P(\theta_j)$, which means that the data is ignored.

For this naive MAP approach we have to conclude [in agreement with Wasserman, 2005]:

1. Most posteriors of θ_j given the data are equal to the prior.
2. The posterior of θ given the data completely factorizes. Even given some data, every θ_j is unrelated to every other $\theta_{j'}$ for $j' \neq j$. Thus what we learn from the data about θ_j at site j does not transfer to $\theta_{j'}$ at a non-observed site $j' \neq j$.
3. It seems that the estimator in Eq. (4) *presumes* that samples are not produced by completely independent sites, but rather, that these sites share some latent property — only on the basis of this presumption it is possible to transfer what one has learned about one site to another. This is a significantly different view on the generative process in Problem 1.

This last point suggests the following approach: from a Bayesian point of view, when we think that transfer from observed sites to unobserved sites is possible, then this must *explicitly* be accounted for in terms of the presumed generative process. In our case, we include a hyperparameter μ in the generative model (Fig. 1(b)) that explicitly reflects our belief that we can learn about unobserved sites from the data:

$$\theta_j \sim \mathcal{N}(\mu, 1) \ . \tag{6}$$

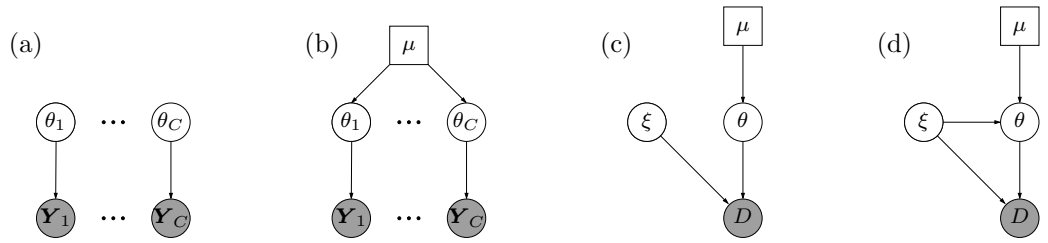

Figure 1: (a) Generative model when θ are considered completely independent. (b) Generative model when a hyperparameter μ is assumed. The indexing of the X-domain is chosen such that for $j \leq J$ we observe n_j data samples Y_{jk}, which are combined in the n_j-dimensional vector $\boldsymbol{Y}_j$. (Unobserved sites are not displayed. (c) Generative model in the case of non-uniform observation probabilities ξ. (d) The case when ξ and θ are assumed dependent.

2.1 Bayesian and MLE-based estimators

We will consider two possible approaches to derive estimators, both of which are likelihood-based in the sense that $P(D|\mu)$ plays the crucial role:

1. In *Empirical Bayesian Analysis* we first use the data to learn an MLE estimate $\hat{\mu}^{\text{MLE}} = \text{argmax}_\mu P(D|\mu)$ of the hyperparameter. Then we derive the desired estimator from the generative model given the MLE parameter $\hat{\mu}^{\text{MLE}}$ by integrating over θ,

$$\hat{\psi} = \int \frac{1}{C} \sum_{j=1}^{C} \theta_j \; P(\theta|\hat{\mu}^{\text{MLE}}) \; d\theta = \frac{1}{C} \sum_{j=1}^{C} \int \theta_j \prod_{j=1}^{C} P(\theta_j|\hat{\mu}^{\text{MLE}}) \; d\theta \tag{7}$$

$$= \frac{1}{C} \sum_{j=1}^{C} \int \theta_j \; \mathcal{N}_{\theta_j}(\hat{\mu}^{\text{MLE}}, 1) \; d\theta_j = \hat{\mu}^{\text{MLE}} \tag{8}$$

2. In *Fully Bayesian Analysis*, we presume a prior over the hyperparameter, e.g. $P(\mu) = \mathcal{N}_\mu(0, \sigma)$, and use the data to compute a posterior $P(\mu|D) = \mathcal{N}_\mu(\hat{\mu}^{\text{BAY}}, \sigma') \propto P(D|\mu) \; P(\mu)$. Then we can derive an estimator by integrating over θ and μ,

$$\begin{aligned}
\hat{\psi} &= \int\int \frac{1}{C} \sum_{j=1}^{C} \theta_j \; P(\theta|\mu) \; P(\mu|D) \; d\theta \; d\mu \\
&= \frac{1}{C} \sum_{j=1}^{C} \int\int \theta_j \; \mathcal{N}_{\theta_j}(\mu, 1) \; \mathcal{N}_\mu(\hat{\mu}^{\text{BAY}}, \sigma') \; d\theta_j \; d\mu \\
&= \frac{1}{C} \sum_{j=1}^{C} \int \theta_j \; \mathcal{N}_{\theta_j}(\hat{\mu}^{\text{BAY}}, 1 + \sigma') \; d\theta_j \; = \hat{\mu}^{\text{BAY}}
\end{aligned} \tag{9}$$

   which is the posterior mean of μ.

We conclude this section by deriving these two estimators for Problem 1.

Theorem 1 *Given data as generated in Problem 1 and the generative model (6) for the θ_j, the MLE estimate of the hyperparameter μ is*

$$\hat{\mu}^{MLE} = \frac{\sum_{i=1}^{n} Y_i/(n_{X_i}+1)}{\sum_{i=1}^{n} 1/(n_{X_i}+1)} \quad \overset{C\to\infty}{\longrightarrow} \quad \frac{1}{n} \sum_{i=1}^{n} Y_i \qquad \text{(in prob.)} \tag{10}$$

and converges for $C \to \infty$ against the estimator in Eq. (4).

Proof Remember that $\boldsymbol{Y}_j$ denotes a vector whose entries are those Y_i for which $X_i = j$. Let us calculate the likelihood of the data:

$$
\begin{aligned}
P(D|\mu) &= \int P(D|\theta)\ P(\theta|\mu)\ d\theta = \frac{1}{C^n}\int \prod_{j=1}^{J} P(\boldsymbol{Y}_j|\theta_j)\ P(\theta_j|\mu)\ d\theta \\
&= \frac{1}{C^n}\prod_{j=1}^{J}\int \mathcal{N}_{\boldsymbol{Y}_j}(\mathbf{1}\theta_j, I)\ \mathcal{N}_{\theta_j}(\mu, 1)\ d\theta_j \\
&\propto \frac{1}{C^n}\prod_{j=1}^{J}\int \mathcal{N}_{\boldsymbol{Y}_{j\perp}}(0, I)\ \mathcal{N}_{\theta_j}(\langle \boldsymbol{Y}_j\rangle, 1/n_j)\ \mathcal{N}_{\theta_j}(\mu, 1)\ d\theta_j \\
&= \frac{1}{C^n}\prod_{j=1}^{J}\mathcal{N}_{\boldsymbol{Y}_{j\perp}}(0, I)\ \int \mathcal{N}_{\theta_j}\Big(\frac{1}{n_j+1}(n_j\langle \boldsymbol{Y}_j\rangle + \mu), \frac{1}{n_j+1}\Big)\ \mathcal{N}_{\langle \boldsymbol{Y}_j\rangle}\Big(\mu, 1+\frac{1}{n_j}\Big)\ d\theta_j \\
&= \frac{1}{C^n}\prod_{j=1}^{J}\mathcal{N}_{\boldsymbol{Y}_{j\perp}}(0, I)\ \mathcal{N}_{\langle \boldsymbol{Y}_j\rangle}\Big(\mu, 1+\frac{1}{n_j}\Big) \qquad (11)
\end{aligned}
$$

In the third line we used the orthogonal decomposition of a Gaussian into a component parallel to the n_j-dimensional vector $\mathbf{1} = (1, 1, .., 1)^T$ and orthogonal to it. With $\langle \boldsymbol{y}\rangle = \mathbf{1}^T\boldsymbol{y}/n_j$ define $\boldsymbol{y}_\perp := \boldsymbol{y} - \mathbf{1}\langle \boldsymbol{y}\rangle$ with the property $\boldsymbol{y}_\perp^T\mathbf{1} = 0$. Note that

$$
(\boldsymbol{y} - \mathbf{1}\theta)^2 = (\boldsymbol{y}_\perp + \mathbf{1}(\langle \boldsymbol{y}\rangle - \theta))^T\ (\boldsymbol{y}_\perp + \mathbf{1}(\langle \boldsymbol{y}\rangle - \theta)) = \boldsymbol{y}_\perp^2 + n_j(\langle \boldsymbol{y}\rangle - \theta)^2 \qquad (12)
$$

allows us to decompose the Gaussian in the second line of Eq. (11):

$$
\mathcal{N}_{\boldsymbol{Y}_j}(\mathbf{1}\theta_j, I) \propto \mathcal{N}_{\boldsymbol{Y}_{j\perp}}(0, I)\ \mathcal{N}_{\theta_j}(\langle \boldsymbol{Y}_j\rangle, 1/n_j) \qquad (13)
$$

In the fourth line we used the product rule for Gaussians which is

$$
\mathcal{N}_x(a, A)\ \mathcal{N}_x(b, B) = \mathcal{N}_x(c, C)\ \mathcal{N}_a(b, A + B) \qquad (14)
$$

with $C = (A^{-1} + B^{-1})^{-1}$ and $c = C(A^{-1}a + B^{-1}b)$. The terms in the log-likelihood which contain μ are:

$$
\log P(D|\mu) = \cdots + \sum_{j=1}^{J}\log \mathcal{N}_{\langle \boldsymbol{Y}_j\rangle}(\mu, 1+\frac{1}{n_j}) = \cdots + \sum_{j=1}^{J}\Big[-\frac{1}{2}\frac{(\langle \boldsymbol{Y}_j\rangle - \mu)^2}{1+1/n_j}\Big] \qquad (15)
$$

$$
\partial_\mu \log P(D|\mu) \propto \sum_{j=1}^{J}\frac{\langle \boldsymbol{Y}_j\rangle - \mu}{1+1/n_j} = 0\ . \qquad (16)
$$

which implies

$$
\hat{\mu}^{\text{MLE}} = \frac{\sum_{j=1}^{J}\langle \boldsymbol{Y}_j\rangle/(1+1/n_j)}{\sum_{j=1}^{J}1/(1+1/n_j)} = \frac{\sum_{i=1}^{n}Y_i/(n_{X_i}+1)}{\sum_{i=1}^{n}1/(n_{X_i}+1)}\ . \qquad (17)
$$

For $C \to \infty$, every site is visited at most once, i.e. $n_j = 1$ for all $j = 1, \ldots, J$ and thus $\hat{\mu}^{\text{MLE}} \to \frac{1}{n}\ \sum_{i=1}^{n}Y_i$. ■

Theorem 2 *Given data according to Problem 1, the generative model (6) for the θ_j, and a hyperparameter prior $P(\mu) = \mathcal{N}_\mu(0, \sigma)$, the posterior mean is*

$$
\hat{\mu}^{BAY} = \frac{\sum_{i=1}^{n}Y_i/(n_{X_i}+1)}{1/\sigma + \sum_{i=1}^{n}1/(n_{X_i}+1)} \quad \overset{C\to\infty}{\longrightarrow} \quad \frac{1}{2/\sigma + n}\ \sum_{i=1}^{n}Y_i \qquad \text{(in prob.)} \qquad (18)
$$

Proof Given the likelihood $P(D|\mu)$ as derived in the last line of equation (11), we have

$$P(\mu|D) \propto P(D|\mu)\ P(\mu) \propto \frac{1}{C^n}\Big[\prod_{j=1}^{J}\mathcal{N}_{\boldsymbol{Y}_{j\perp}}(0,I)\Big]\ \Big[\mathcal{N}_\mu(0,\sigma)\ \prod_{j=1}^{J}\mathcal{N}_\mu\Big(\langle\boldsymbol{Y}_j\rangle, 1+\frac{1}{n_j}\Big)\Big]$$

$$= \frac{1}{C^n}\Big[\prod_{j=1}^{J}\mathcal{N}_{\boldsymbol{Y}_{j\perp}}(0,I)\Big]\ \mathcal{N}_\mu(\langle\mu\rangle,\sigma')\ \Big[\prod_{j=1}^{J}\mathcal{N}_{\langle\boldsymbol{Y}_j\rangle}(\textit{indep. of }\mu)\Big]\ ,$$

$$\text{with } \frac{1}{\sigma'} = \frac{1}{\sigma} + \sum_{j=1}^{J}\frac{1}{1+1/n_j} = \frac{1}{\sigma} + \sum_{i=1}^{n}\frac{1}{n_{X_i}+1}$$

$$\text{and } \langle\mu\rangle = \sigma'\sum_{j=1}^{J}\frac{\langle\boldsymbol{Y}_j\rangle}{1+1/n_j} = \sigma'\sum_{i=1}^{n}\frac{Y_i}{n_{X_i}+1}\ . \tag{19}$$

In the second line, we used again the product rule for Gaussians,

$$\prod_j \mathcal{N}_x(a_j,A_j) = \mathcal{N}_x(c,C)\ \prod_j \mathcal{N}_{a_j}(\cdots)\ , \quad \frac{1}{C} = \sum_j\frac{1}{A_j}\ , \quad c = C\sum_j\frac{a_j}{A_j}\ , \tag{20}$$

where the dots are some terms independent of x. The normalization constraint of $P(\mu|D)$ then leads to $P(\mu|D) = \mathcal{N}_\mu(\langle\mu\rangle,\sigma')$ and $\hat{\mu}^{\text{BAY}} = \langle\mu\rangle$. ■

3 Randomly missing data and independent ξ and θ

The following problem is formulated as in Wasserman [2005] with the exception that we define Y_i to be Gaussian rather than Bernoulli variables. Robins and Ritov [1997] also considered Y_i to be Gaussian, but furthermore a continuous domain for X, which we address in section 5.

Problem 2 *The observed data D consists of n tuples (X_i, R_i, Y_i) with $1 \le i \le n$ which are sampled i.i.d. as follows:*

$$X_i \sim \text{uniform on}\{1,\ldots,C\} \tag{21}$$

$$R_i \sim \text{Bernoulli}(\xi_{X_i}) \tag{22}$$

$$Y_i \sim \begin{cases} \mathcal{N}(\theta_{X_i},1) & \text{if } R_i = 1 \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

with the unknown vector $\theta = (\theta_1,\theta_2,\ldots,\theta_C)$, but with known vector ξ bounded away from 0 and 1, (i.e. $0 < \delta \le \xi_j \le 1-\delta < 1$ for some small δ), and with $C \gg n$. The task is to find an estimator for

$$\psi = \frac{1}{C}\ \sum_{j=1}^{C}\theta_j\ . \tag{24}$$

One classical approach is the Horvitz-Thompson (HT) estimator which is

$$\hat{\psi}^{\text{HT}} = \frac{1}{n}\ \sum_{i=1}^{n}\frac{R_iY_i}{\xi_{X_i}}\ . \tag{25}$$

The idea of the HT estimator is to compensate for the probability of observing data at site $X = j$ by dividing with ξ_j. With other words, it amplifies observed data, if it was unlikely to observe it.

In order to derive a likelihood-based estimator we again assume $\theta_j \sim \mathcal{N}(\mu,1)$ with the hyper-parameter μ (Fig. 1(c)). Notationwise, we again assume that the indices $j = 1,\ldots,J,J+1,\ldots,C$ are properly sorted as before. However, now we define $n_j := \#\{i : X_i = j \text{ and } R_i = 1\}$.

Theorem 3 *Given data from Problem 2 and presuming the generative model with hyperparameter μ as in Fig. 1(c), the MLE estimator for μ is the same as for Problem 1, i.e., it ignores the observation probabilities ξ:*

$$\hat{\mu}^{MLE} = \frac{\sum_{i=1}^{n} R_i \; Y_i/(n_{X_i}+1)}{\sum_{i=1}^{n} R_i/(n_{X_i}+1)} \quad \stackrel{C\to\infty}{\longrightarrow} \quad \frac{1}{\sum_{i=1}^{n} R_i} \; \sum_{i=1}^{n} R_i \; Y_i \tag{26}$$

with exactly the same estimator as in Theorem 1. Note that for clarity we included a factor R_i in the enumerator, which is redundant given the convention $R_i = 0 \Rightarrow Y_i = 0$. Given the prior $P(\mu) = \mathcal{N}(0, \sigma)$, also the posterior $P(\mu|D)$ is the same as before, with posterior mean

$$\hat{\mu}^{BAY} = \frac{\sum_{i=1}^{n} R_i \; Y_i/(n_{X_i}+1)}{1/\sigma + \sum_{i=1}^{n} R_i/(n_{X_i}+1)} \; . \quad \stackrel{C\to\infty}{\longrightarrow} \quad \frac{1}{2/\sigma + \sum_{i=1}^{n} R_i} \; \sum_{i=1}^{n} R_i \; Y_i \; . \tag{27}$$

Proof Instead of equation (11), we now have

$$\begin{aligned} P(D|\mu) &= \int P(D|\theta) \; P(\theta|\mu) \; d\theta = \int \prod_{i=1}^{n} \frac{1}{C} \; \xi_{X_i}^{R_i} (1-\xi_{X_i})^{1-R_i} \; P(Y_i \big| X_i, \theta)^{R_i} \; P(\theta \big| \mu) \; d\theta \\ &= \frac{1}{C^n} \Big[\prod_{i=1}^{n} \xi_{X_i}^{R_i} (1-\xi_{X_i})^{1-R_i} \Big] \Big[\prod_{j=1}^{J} \int P(\boldsymbol{Y}_j \big| \theta_j) \; P(\theta_j \big| \mu) \; d\theta_j \Big] \end{aligned} \tag{28}$$

In the first line, the terms $P(Y_i \big| X_i, \theta)^{R_i}$ drop out for non-observed data, i.e. for $R_i = 0$. Hence, in the second line we can rearrange the data indexing, again grouping data from one site to a joint Gaussian vector $\boldsymbol{Y}_j$. As a result, the likelihood perfectly factorizes, the first factor is independent of μ, and $\hat{\mu}^{\text{MLE}}$ is the same as in theorem 1 when considering only observed data. ■

That is, the likelihood-based estimators ignore the a priori knowledge of the observation probabilities ξ. However, it does implement a "reweighting of the data" in the estimator depending on the multiplicities n_j of observing the same site.

We performed simple experiments to compare the Horvitz-Thompson estimator and our likelihood-based estimator from Theorem 3:

Experiment 1 (unrelated θ and ξ) *To simulate the limit $C \to \infty$ we assume that all visited data sites X_i are different. We generate a data set by first sampling $\theta_i \sim \mathcal{N}(\mu, 1)$ and further $\xi_i \sim \mathcal{U}([.1, .9])$ for each observed site i. Note that θ and ξ are completely unrelated. After that we sample R_i and Y_i according to the model in Problem 2. We compute the values of both estimators (25) and (26) for $1 \le n \le 10^5$ data points. We repeat this for 20 data sets. Fig. 2 displays the results.*

The experiment suggests that both, the LB estimator (26) as well as the Horvitz-Thompson (25) work fine. However, as can be seen from the logarithmic plots of the variance, the variance of the LB estimator is much smaller than that of the HT estimator. Furthermore, the experiments suggest that the variance for the HT estimator increases with larger $\mu \neq 0$.

4 Likelihood-based estimators for Robins Ritov's proof case: ξ and θ dependent

Let us now get to the core of Robins and Ritov [1997]. The authors consider uniform unbiasedness of an estimator. This means that the estimator has to be unbiased for *every possible choice of θ and ξ*. In the experiment we performed above, though, we chose ξ and θ independently and thus it was very unlikely that we ended up with an accidentally correlated ξ and θ, e.g., where θ tends to be large whenever also ξ is (or inversely).

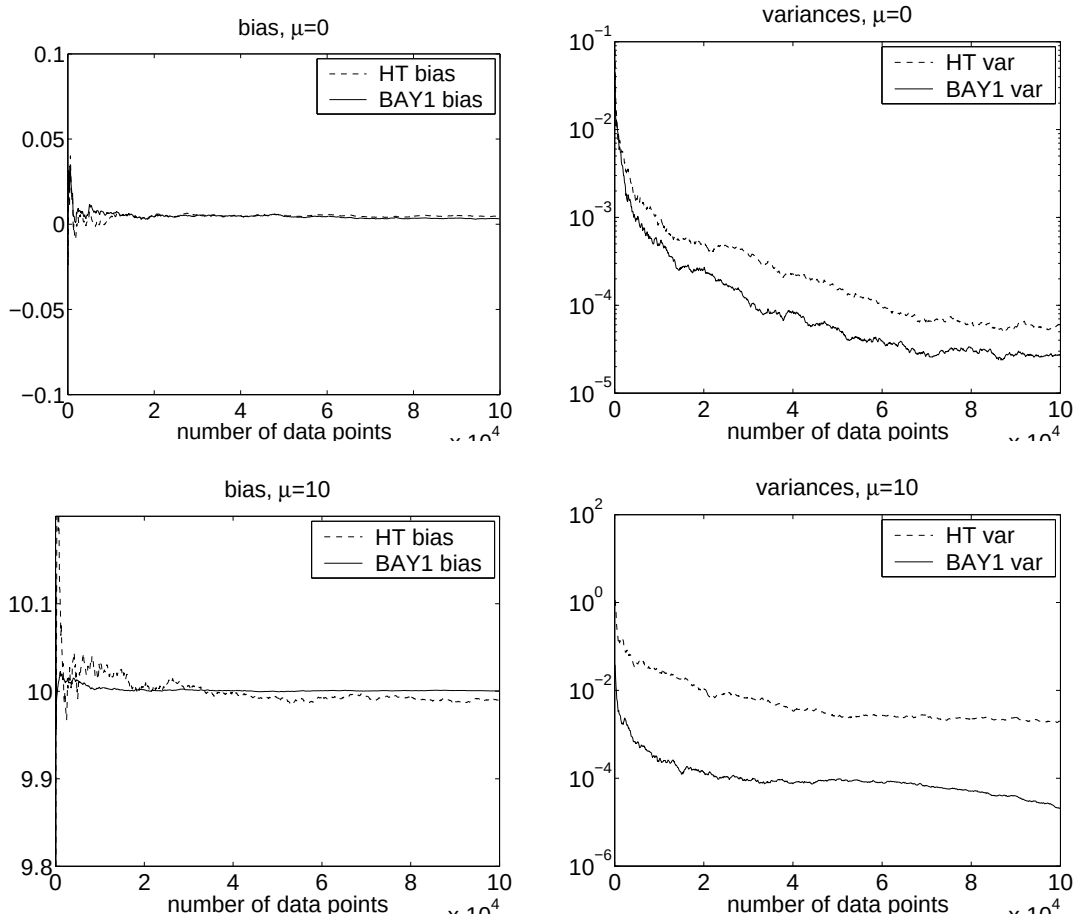

Figure 2: Experiment 1; bias and variance of the estimators (25) and (26) from 20 runs, for $\mu = 0$ and $\mu = 10$.

Theorem 3 in Robins and Ritov [1997] proves that a likelihood-based estimator in their framework cannot be uniformly unbiased. The proof is by constructing a specific ξ and θ for which the estimator will be biased. This specific choice of ξ and θ used in the proof is very interesting and instructive for us: They consider ξ and θ correlated such that at half the sites $\xi = .25 + d$ and $\theta = .5 + a$ and at the other half of the sites $\xi = .25 - d$ and $\theta = .5 - a$.

In fact, experiment 2 will show that such cases will mislead the LB estimator (26) and lead it to converge to a heavily biased estimate, while the HT estimator is indeed unbiased. However, the example motivates us to explicitly consider a dependency between ξ and θ that will enter the likelihood (Fig. 1(d)).

Theorem 4 *Assuming the prior of θ depends on ξ in the form $\theta_j \sim \mathcal{N}_{\theta_j}(\mu, \alpha\xi_j)$ for some $\alpha > 0$, and conditioning the model on a known ξ, then the ξ enters the MLE estimator for μ as follows:*

$$\hat{\mu}^{MLE} = \frac{\sum_{i=1}^{n} R_i Y_i/(\alpha\xi_{X_i} n_{X_i} + 1)}{\sum_{i=1}^{n} R_i/(\alpha\xi_{X_i} n_{X_i} + 1)} \quad \overset{C,\alpha\to\infty}{\longrightarrow} \quad \frac{\sum_{i=1}^{n} R_i Y_i/\xi_{X_i}}{\sum_{i=1}^{n} R_i/\xi_{X_i}} \ , \tag{29}$$

and with the prior $P(\mu) = \mathcal{N}(0, \sigma)$, the posterior mean is

$$\hat{\mu}^{BAY} = \frac{\sum_{i=1}^{n} R_i Y_i/(\alpha\xi_{X_i} n_{X_i} + 1)}{1/\sigma + \sum_{i=1}^{n} R_i/(\alpha\xi_{X_i} n_{X_i} + 1)} \quad \overset{C,\alpha\to\infty}{\longrightarrow} \quad \frac{\sum_{i=1}^{n} R_i Y_i/\xi_{X_i}}{2/\sigma + \sum_{i=1}^{n} R_i/\xi_{X_i}} \ , \tag{30}$$

Proof Starting from equation (28) and adding the dependency of θ_j on ξ_j we have

$$
\begin{aligned}
P(D|\mu) &= \frac{1}{C^n}\Big[\prod_{i=1}^{n}\xi_{X_i}^{R_i}(1-\xi_{X_i})^{1-R_i}\Big]\Big[\prod_{j=1}^{J}\int P(\boldsymbol{Y}_j|\theta_j)\; P(\theta_j|\mu,\xi_j)\; d\theta_j\Big] \\
&= \frac{1}{C^n}\Big[\textit{indep. of } \mu\Big]\;\prod_{j=1}^{J}\int \mathcal{N}_{\boldsymbol{Y}_j}(\mathbf{1}\theta_j,1)\;\mathcal{N}_{\theta_j}(\mu,\alpha\xi_j)\; d\theta_j \\
&\propto \prod_{j=1}^{J}\int \mathcal{N}_{\boldsymbol{Y}_{j,\perp}}(0,I)\;\mathcal{N}_{\theta_j}(\langle\boldsymbol{Y}_j\rangle,1/n_j)\;\mathcal{N}_{\theta_j}(\mu,\alpha\xi_j)\; d\theta_j \\
&= \prod_{j=1}^{J}\mathcal{N}_{\boldsymbol{Y}_{j,\perp}}(0,I)\;\int \mathcal{N}_{\theta_j}(\cdots)\;\mathcal{N}_{\langle\boldsymbol{Y}_j\rangle}(\mu,\alpha\xi_j+1/n_j)\; d\theta_j \\
&= \prod_{j=1}^{J}\mathcal{N}_{\boldsymbol{Y}_{j\perp}}(0,I)\;\mathcal{N}_{\langle\boldsymbol{Y}_j\rangle}(\mu,\alpha\xi_j+1/n_j) \qquad (31)
\end{aligned}
$$

MLE estimator for μ:

$$
\log P(D|\mu) = \cdots + \sum_{j=1}^{J}\log\mathcal{N}_{\langle\boldsymbol{Y}_j\rangle}(\mu,\alpha\xi_j+1/n_j) = \cdots + \sum_{j=1}^{J}\Big[-\frac{1}{2}\;\frac{(\langle\boldsymbol{Y}_j\rangle-\mu)^2}{\alpha\xi_j+1/n_j}\Big] \qquad (32)
$$

$$
\partial_\mu \log P(D|\mu) = \sum_{j=1}^{J}\frac{\langle\boldsymbol{Y}_j\rangle-\mu}{\alpha\xi_j+1/n_j} = 0 \qquad (33)
$$

from which follows

$$
\hat{\mu}^{\text{MLE}} = \frac{\sum_{j=1}^{J}\langle\boldsymbol{Y}_j\rangle/(\alpha\xi_j+1/n_j)}{\sum_{j=1}^{J}1/(\alpha\xi_j+1/n_j)} = \frac{\sum_{i=1}^{n}Y_iR_i/(\alpha\xi_{X_i}n_{X_i}+1)}{\sum_{i=1}^{n}R_i/(\alpha\xi_{X_i}n_{X_i}+1)}\;. \qquad (34)
$$

The posterior mean $\hat{\mu}^{\text{BAY}}$ follows directly from the likelihood (31), as in the proof to Theorem 2, with

$$
\begin{aligned}
P(\mu|D) &= \mathcal{N}(\hat{\mu}^{\text{BAY}},\sigma')\;, \quad \frac{1}{\sigma'} = \frac{1}{\sigma}+\sum_{j=1}^{J}\frac{1}{\alpha\xi_j+1/n_j} = \frac{1}{\sigma}+\sum_{i=1}^{n}\frac{R_i}{1+\alpha\xi_jn_{X_i}} \\
\hat{\mu}^{\text{BAY}} &= s'\sum_{j=1}^{J}\frac{\langle\boldsymbol{Y}_j\rangle}{\alpha\xi_j+1/n_j} = s'\sum_{i=1}^{n}\frac{R_i\;Y_i}{1+\alpha\xi_jn_j} \qquad (35)
\end{aligned}
$$

■

Experiment 2 (ξ and θ correlated) *Again, all sites X_i are different. We generate a data set by first sampling $\theta_i \sim \mathcal{N}(\mu,1)$ and then* dependently *choosing $\xi_i = 1/2 + (\theta_i - \mu)$ (capped to the interval $[.1,.9]$). We compute the three estimators (25), (26), and (29) for $1 \le n \le 10^5$ data points. We repeat this for 20 data sets. Fig. 3(a&b) display the bias $\hat{\psi}-\mu$ over the 20 data sets and the variance of the estimators with increasing number of data points.*

The estimator (29) is indeed very similar to the HT estimator; for $\alpha\to\infty$ they only differ by the normalization constant (replacing $1/n$ by $1/\sum_{i=1}^{n}R_i/\xi_{X_i}$). For $\mu=0$, the experiments show that their estimates are almost identical. However, for larger μ the LB estimator (29) again has considerably lower variance. The experiments also confirm that the LB estimator (26) without assumed ξ and θ dependency is heavily biased.

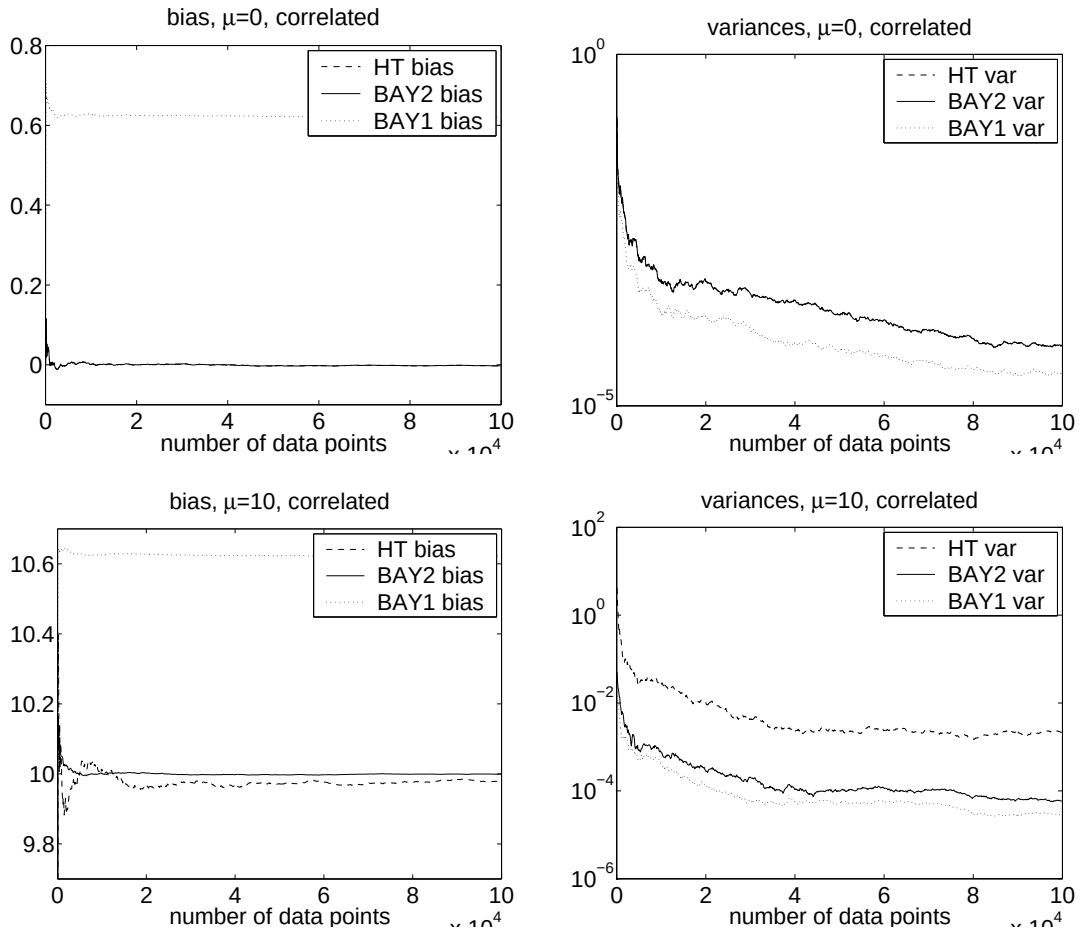

Figure 3: Experiment 2; bias and variance of the estimators (25), (29) (BAY2), and (26) (BAY1) from 20 runs, for $\mu = 0$ and $\mu = 10$. For $\mu = 0$ the HT estimator and our likelihood-based estimator (29) are very similar (hardly distinguishable in these averages).

5 Gaussian Process priors in the continuous case

Problem 3 (continuous X, some missing data) *The data D consists of n tuples (X_i, R_i, Y_i) with $1 \leq i \leq n$ which are sampled i.i.d. as follows:*

$$X_i \sim \mathcal{U}([0,1]^k) \tag{36}$$

$$R_i \sim Bernoulli(\xi(X_i)) \tag{37}$$

$$Y_i \sim \begin{cases} \mathcal{N}(\theta(X_i), 1) & \text{if } R_i = 1 \\ 0 & \text{otherwise} \end{cases} \tag{38}$$

where $\xi : [0,1]^k \to [\delta, 1-\delta]$ for small $\delta > 0$ is a function that provides for each site X a success probability $\xi(X)$, and with the unknown function $\theta : [0,1]^k \to \mathbb{R}$. Find an estimator for $\psi = \int_{[0,1]^k} \theta(x)\, dx$.

Theorem 5 *Assuming a Gaussian process prior $\theta \sim \mathrm{GP}(\mu, k)$ for the function θ, with a constant mean function $\mu \in \mathbb{R}$ and kernel function k, the MLE estimator for μ is*

$$\hat{\mu}^{MLE} = \frac{\mathbf{1}^T\ (I + K_{\boldsymbol{X}})^{-1}\ \boldsymbol{Y}}{\mathbf{1}^T\ (I + K_{\boldsymbol{X}})^{-1}\ \mathbf{1}} \tag{39}$$

where $K_{\boldsymbol{X}}$ is the Gram matrix at the observed data sites with entries $K_{i'i} = k(X_{i'}, X_i)$, and $\boldsymbol{Y}$ is the $(\sum_{i=1}^n R_i)$-dimensional vector of all observations, i.e. it contains all entries Y_i such that $R_i = 1$.

Proof Let $\theta_{\boldsymbol{X}}$ be the vector of the values $\theta(X_i)$ for all i with $R_i = 1$, i.e. the evaluation of θ at the observed sites. Note that, given the GP prior, the distribution of $\theta_{\boldsymbol{X}}$ is a joint Gaussian, $\theta_{\boldsymbol{X}} \sim \mathcal{N}(\mu\mathbf{1}, K_{\boldsymbol{X}})$ with constant mean $\mu\mathbf{1}$ and the Gram matrix as covariance matrix.

$$\begin{aligned} P(D\big|\mu) &= \int P(D|\theta)\ P(\theta|\mu)\ d\theta = \int \prod_{i=1}^n \xi(X_i)^{R_i}(1-\xi(X_i))^{1-R_i}\ P(Y_i\big|X_i,\theta)^{R_i}\ P(\theta\big|\mu)\ d\theta \\ &\propto \int \mathcal{N}_{\boldsymbol{Y}}(\theta_{\boldsymbol{X}}, I)\ \mathcal{N}_{\theta_{\boldsymbol{X}}}(\mu\mathbf{1}, K_{\boldsymbol{X}})\ d\theta_{\boldsymbol{X}} = \int \mathcal{N}_{\theta_{\boldsymbol{X}}}(c, C)\ \mathcal{N}_{\boldsymbol{Y}}(\mu\mathbf{1}, I + K_{\boldsymbol{X}})\ d\theta_{\boldsymbol{X}} \\ &= \mathcal{N}_{\boldsymbol{Y}}(\mu\mathbf{1}, I + K_{\boldsymbol{X}}) \end{aligned} \tag{40}$$

with c and C appropriately chosen according to the product rule in Eq. (14). This implies the MLE estimator for μ:

$$\hat{\mu}^{\text{MLE}} = \frac{\mathbf{1}^T\ (I + K_{\boldsymbol{X}})^{-1}\ \boldsymbol{Y}}{\mathbf{1}^T\ (I + K_{\boldsymbol{X}})^{-1}\ \mathbf{1}}. \tag{41}$$

■

Remark 6 *For the kernel function $k(x', x) = \alpha\xi_x\delta_{x'x}$ we get*

$$\hat{\mu}^{MLE} = \frac{\sum_{i=1}^n R_i Y_i/(1+\alpha\xi_{X_i})}{\sum_{i=1}^n R_i/(1+\alpha\xi_{X_i})}\ , \tag{42}$$

which is exactly the same as the estimator (29) in the discrete case for $C \to \infty$ (when each site is observed only once).

It is interesting to see that here the choice of the kernel function plays the central role. Assuming smoothness of the function θ enters the estimator in a way analogous to observing the same point multiple times, as in the estimator (26). Indeed, smoothness means that what has been learned about one site can be transferred to another. Assuming a ξ-dependent variance term $\alpha\xi_x\delta_{x'x}$ in the kernel retrieves the estimator we proposed to handle the case of dependent ξ and θ.

6 Conclusion

We investigated several approaches to derive likelihood-based estimators in the example of Robins-Ritov, which are summarized in Table 1. The estimators are based on additional assumptions on the latent generative model. In particular, we considered a hyperparameter μ, which allows the likelihood-based approach to transfer what can be learned from observations at some sites to unobserved sites. And we considered a dependence of the ξ and θ in terms of an explicit prior $P(\theta|\xi, \mu)$, which allows us to define a likelihood-based estimator which depends on and accounts for the a priori knowledge of the "stratum-specific randomization probabilities" ξ. The likelihood-based estimators we derive from this approach are very similar to the HT estimator. Experiments show that they are unbiased also for correlated ξ and θ, but appear to have lower variance than the HT estimator.

Acknowledgements

S.H. is grateful to the European Community for being supported by a European Community Marie Curie Fellowship (MEIF-CT-2005-025578). M.T. is grateful to the German Research Foundation (DFG) for the Emmy Noether fellowship TO 409/1-1.

	frequentist	Emprical Bayesian	Fully Bayesian
no missing data (Pr. 1)	(Eq. (4)) $\frac{1}{n}\sum_{i=1}^{n} Y_i$	(Th. 1) $\frac{1}{n}\sum_{i=1}^{n} Y_i$	(Th. 2) $\frac{1}{2/\sigma + n}\sum_{i=1}^{n} Y_i$
missing data (Pr. 2) indep. ξ and θ	(HT) $\frac{\sum_{i=1}^{n} R_i Y_i/\xi_{X_i}}{n}$	(Th. 3) $\frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i}$	(Th. 3) $\frac{\sum_{i=1}^{n} R_i Y_i}{2/\sigma + \sum_{i=1}^{n} R_i}$
missing data (Pr. 2) dep. ξ and θ	(HT) $\frac{\sum_{i=1}^{n} R_i Y_i/\xi_{X_i}}{n}$	(Th. 4) $\frac{\sum_{i=1}^{n} R_i Y_i/\xi_{X_i}}{\sum_{i=1}^{n} R_i/\xi_{X_i}}$	(Th. 4) $\frac{\sum_{i=1}^{n} R_i Y_i/\xi_{X_i}}{2/\sigma + \sum_{i=1}^{n} R_i/\xi_{X_i}}$
missing data & continuous (Pr. 3)	(HT) $\frac{\sum_{i=1}^{n} R_i Y_i/\xi(X_i)}{n}$	(Th. 5) $\frac{\mathbf{1}^T (I + K_{\boldsymbol{X}})^{-1} \boldsymbol{Y}}{\mathbf{1}^T (I + K_{\boldsymbol{X}})^{-1} \mathbf{1}}$	

Table 1: Overview of the derived estimators (for the case $C \to \infty$).

References

James M. Robins and Ya'acov Ritov. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16:285–319, 1997.

Larry Wasserman. *All of Statistics*. Springer, 2005.