ROBINS-WASSERMAN, ROUND N

CHRISTOPHER A. SIMS

It may be worthwhile to build up the discussion of this example from the simplest case, showing how Bayesian methods lead to useful solutions at each stage. As we complicate the model, pitfalls for Bayesian inference will arise: naively extending Bayesian priors that seem reasonable for simple versions of the model to more complicated ones can inadvertently imply prior certainty or near-certainty about exactly the parameter that we are trying to estimate (and thus presumably are uncertain about). The lesson here is that in complex problems, prior distributions should always be assessed for their implications about data behavior and parameters of interest, to be sure that they are not putting too much probability on data behavior we actually believe highly unlikely or too little probability on parameter values or data behavior we actually consider plausible.

Very similar examples, making similar points, appear in the paper by Harmeling and Toussaint (2007), which I read only after much of this comment was written. Some of the initial examples below are simpler than in the Harmeling and Toussaint paper, and they did not explore the nature of possible slow convergence under independent priors on $\theta()$ in the continuously-distributed-X case.

I. A SIMPLE CASE

For most of this discussion we will be preserving the idea that the observable data (X_i, R_iY_i, R_i) are i.i.d., indexed by i, that the marginal distribution of X_i is known, that Y_i is always 0 or 1, with $P[Y_i = 1 \mid X_i, R_i] =$

Date: October 8, 2012.

This work [partially] supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370. Any opinions, findings and conclusions or recomendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF)

$\theta(X_i)$, that $\theta()$ is an unknown function, that R_i is always zero or one, and that $P[R_i = 1 \mid X_i] = \pi(X_i)$, with $\pi()$ known. What will change between simpler and more complex versions of the model is the nature of the X random variable and of the $\pi()$ function.

Suppose that X_i can take on just two values, M or F. We have collected data by sending out an interviewer to interview students at random and to ask them whether they intend to vote for the Democrat. The interviewer was to spend the first two days interviewing men, the second two days interviewing women, but got sick after day three. So women are half as likely to be in the sample as men, the actual number included is random, and not exactly half the number of men. We assume that men and women are in equal numbers overall in our target population, and want to estimate the overall proportion of likely Democratic voters. Y_i is then 1 if the respondent says he or she will vote Democratic. $\theta(M)$ is the proportion of men who intend to vote Democratic, $\theta(F)$ is the proportion of women.

The obvious thing to do here is estimate the overall proportion of Democratic voters as the proportion of such men in the sample of men, plus the proportion of such women in the sample of women, divided by two. The pdf for a random draw from the interviewed sample is

$$\begin{aligned} \frac{2}{3}\theta(M)^{Y_i}(1 = \theta(M))^{1-Y_i} \quad & \text{if } X_i = M \\ \frac{1}{3}\theta(F)^{Y_i}(1 = \theta(F))^{1-Y_i} \quad & \text{if } X_i = F\,. \end{aligned}$$

The full sample likelihood is therefore

$$\left(\frac{2}{3}\right)^{N_{M_1}+N_{M0}} \theta(M)^{N_{M1}}(1-\theta(M))^{N_{M2}} \left(\frac{1}{3}\right)^{N_{F_1}+N_{F0}} \theta(F)^{N_{F1}}(1-\theta(F))^{N_{F2}}\,,$$

Where, e.g., N_{F0} is the number of women not declaring themselves likely to vote Democratic ($Y_i = 0$), and the other N's in the expression are defined similarly. If we have independent uniform priors on $\theta(M)$ and $\theta(F)$, this likelihood is proportional to our posterior distribution. The maximum likelihood estimators of the two parameters are

$$\hat{\theta}(M) = \frac{N_{M1}}{N_{M1}+N_{M2}}\,, \qquad \hat{\theta}(F) = \frac{N_{F1}}{N_{F1}+N_{F2}}$$

So the MLE reproduces what I've called the "obvious thing to do". Note that the 2/3, and 1/3 terms in the likelihood are just scale factors, so the

posterior distribution is unaffected by them. The fact that we know that women are underrepresented in the sample is not relevant, since we know how many women were actually included and can use that knowledge directly.

The posterior mean here is not quite the same as the MLE. From the fact that the posterior on parameter is in the form of a Beta density, we know it is

$$\hat{\hat{\theta}}(M) = \frac{N_{M1+1}}{N_{M1} + N_{M2} + 2}\,, \qquad \hat{\theta}(F) = \frac{N_{F1} + 1}{N_{F1} + N_{F2} + 2}$$

For large samples, this is very close to the MLE. For very small samples, it "shrinks" toward the prior mean, which was .5 for each parameter. For example, if the sample contained just two men and one woman, all of whom said they would vote Democratic, the posterior mean for the overall proportion of Democrats would be $(3/4 + 2/3)/2 = 17/24 =$, while the MLE would be 1.0.

Note that we have no analogue to the R_i variable in this discussion. Suppose that instead of getting sick the fourth day, the interviewer that day just lost the results, even though the count of people interviewed (all female, of course), is known. Now we can think of the sample as including the fourth-day interviewees. For them, $R_i = 0$, while for the others $R_i = 1$. For men, $P[R_i = 1 \mid X_i = M] = \pi(M) = 1$, while for women $\pi(F) = .5$. We can write the single-observation pdf for this expanded sample as

$$\begin{aligned} \frac{1}{2}\theta(M)^{Y_i}(1 = \theta(M))^{1-Y_i} \quad & \text{if } X_i = M \\ \frac{1}{4}\theta(F)^{Y_i}(1 = \theta(F))^{1-Y_i} \quad & \text{if } X_i = F \text{ and } R_i = 1 \\ \frac{1}{4} \quad & \text{if } R_i = 0\,, \end{aligned}$$

which leads to the full-sample likelihood

$$\left(\frac{1}{2}\right)^{N_{M_1}+N_{M0}} \theta(M)^{N_{M1}}(1-\theta(M))^{N_{M2}} \left(\frac{1}{4}\right)^{N_{F_1}+N_{F0}} \theta(F)^{N_{F1}}(1-\theta(F))^{N_{F2}} \left(\frac{1}{4}\right)^{N_0}\,,$$

where N_0 is the number of observations with Y observations lost. But this likelihood differs from the previous one, as a function of θ, only through a scale factor. So it leads to the same MLE and posterior mean. This reflects

the fact, here intuitively obvious, that the interviewees for whom data have been lost are irrelevant to inference about the two θ parameters.

The Horwitz-Thomson estimator could be applied here. It would tell us to estimate the overall proportion of Democrats as

$$\frac{\sum_{X_i=M} Y_i}{N} + \frac{\sum_{X_i=F} Y_i}{.5N}.$$

This differs from the MLE in that it replaces $N_{M1} + N_{M0}$ in the denominator of the estimator of $\theta(M)$ by an estimator of it: $N\pi(M)P[X_i = M]$. This does provide a consistent estimator, but if we expected, based on sample size, $\pi(M)$, and $P[X - i = M] = .5$, to see 40 men in the sample, but in fact we see 36, dividing the number of Democrats in the sample by 36 gives a more accurate estimator than dividing by 40. This is particularly easy to see if the number actually in the sample exceeds the number expected, since then the Horwitz-Thomson estimator can exceed one.

II. More X values

Suppose that instead of X being M or F, it is the home state of the student. Now there are 50 possible values of X instead of two. We might know the ratio of numbers of students from each state to population of the state and want to treat that as $\pi(X_i)$. (Of course students aren't really a representative sample of any state's population, but we'll ignore that.) Now, even if the sample size is a few hundred, there will be small numbers of observations for each state, and possibly some states with no observations. If the sample size is around 50, there will be many states with no observations. We have 50 unknown $\theta(X)$ parameters. If we continue assuming that our prior beliefs about these 50 parameters make them independent $U(0,1)$ variables, the posterior density does not depend at all on any of the $\theta(X)$'s for states that have not been observed. So the MLE is not defined. The posterior mean is well defined, because the posterior density is defined over the whole θ space. But for each state that is not observed, the posterior mean and the prior mean are the same, i.e. .5. And the shrinkage toward .5 of the posterior mean will be proportionately much larger than in our M/F example, because there are in any case only a few observations on each state.

In fact, if we think of the number of possible X values as extremely large relative to sample size, we can see that the posterior mean is almost

the same as the prior mean, regardless of what the observed data are. In Wasserman's original textbook example he noted these facts and argued that this meant Bayesian inference, and indeed any likelihood-based inference, degenerated and became useless in this situation with extremely large numbers of X values. He also noted that the Horwitz-Thomson estimator still gives a usable value in this situation.

But the problem with Bayesian inference here is only that we have been uncritical in maintaining the i.i.d. $U(0,1)$ prior on the $\theta(X)$ values even as the count M of X values increased. The i.i.d. $U(0,1)$ assumption implies our prior mean for the parameter of interest, $\psi = \sum_X \theta(X)P[X]$ is .5 (since the prior mean of of each $\theta(X)$ is .5 and $P[X]$ is just $1/N$ by assumption.) But the prior variance of ψ is $1/(12 \cdot N)$, because ψ is, in our prior, an average of N $U(0,1)$ variables. So with $M = 50$, our prior asserts that before we see the data, we think ψ to be .5 with a standard deviation of about .04. With $M = 1000$, the prior standard deviation becomes approximately .01. But ψ is the unknown that interests us. It makes no sense to try to use data to estimate it if we are a priori nearly certain of its value.

We need to find a way to specify a prior that reflects our substantial uncertainty about ψ, while also, like the i.i.d. prior, representing the idea that we don't have different beliefs about different $\theta(X)$'s. This is straightforward if we introduce the idea of an exchangeable prior distribution. Here, this could be, for example, a prior that specified that for each X value, $E[\theta(X) \mid X] = \psi]$. An example would be a Beta$(2\psi - 1, 1 - 2\psi)$ distribution, i.i.d. across X values, except for the fact that we condition on the unknown parameter ψ. With a $U(0,1)$ prior on ψ, the posterior is proportional to

$$\prod_{i=1}^{N} \theta(X_i)^{N_{11i}}(1-\theta(X_i))^{N_{10i}}\theta(X_i)^{2\psi-1}(1-\theta(X_i))^{1-2\psi}\,.$$

The i in this expression runs over the possible value of X, not over the observation numbers. For X values for which there are no observations on Y ($N_{11i} = N_{10i} = 0$), there is only the Beta pdf for $\theta(X)$, which integrates to 1. We can treat these $\theta(X)$ values as nuisance parameters and integrate them out. They do not affect the posterior on ψ. For the X values with non-zero numbers of observations, we again get a Beta posterior, whose mean is close to the naive $N_{11j}/(N_{11j} + N_{10j})$ estimator if the X value has

many Y observations, but shrink toward ψ, the (unknown) prior mean, for X's with few observations.

This is all sensible, unless few or none of the X values repeat in the sample. This might be the expected outcome if X had millions of potential values, all with very small probabilities. When there is only one observation per X value. the marginal posterior on ψ specializes to being proportional to

$$\psi^{N_{11}}(1-\psi)^{N_{10}},$$

i.e. to a Beta$(N_{11}+1, N_{10}+1)$ form. This will concentrate, if $N_{11}+N_{10}$ is large, but still with only one observation per X value, on $\psi = N_{11}/(N_{11}+N_{10}$. That this expression is a biased estimator of ψ is precisely the selection bias problem. We get exactly this posterior, in fact, for any assumed distribution for $\theta(X)$ that has mean ψ and is i.i.d. across X values for given ψ.

We have arrived at this biased estimator because, again, we have let a reasonable low-dimensional prior expand unthinkingly to high dimensions. By asserting that the mean of $\theta(X)$ is ψ for all X, with $\pi()$ in our available information set, we have asserted that $E[\theta(X) \mid \pi(X)] \equiv \psi$. But this assumes away selection bias. Despite this assumption, if we have large numbers of observations for each X value, we have, as we have already seen, a sensible estimator that converges to the truth. But in a situation where we have many observations, but very small numbers of observations for each X value, assuming away selection bias leads to Bayesian posteriors that will sharply concentrate on false values if the independence assumption is incorrect.

But we knew not only that we were estimating ψ, so that our prior should not imply high a priori certainty about ψ, but also that selection bias is realistically expected to be present, so that assuming it away in the prior is a mistake. Selection bias arises when $E[\theta(X) \mid \pi(X)] = \hat{\theta}(\pi(X))$ varies with $\pi(X)$, but we nonetheless treat some observations with differing $\theta(\pi(X))$ values as if they had the same values of θ. We recognize selection bias, therefore, by treating the $\hat{\theta}()$ function as an unknown parameter. Exactly how we model the $\theta(\pi)$ function should depend on details of the application, but one way to generate a "conservative" model for it is to maximize the entropy of the distribution of the observed data

conditional on the parameters defining known moments. Here the problem is[1]

$$\begin{aligned}\max_{\theta()} \sum_X \big[& -p_{11}(\pi(X),\theta(X)) \log(p_{11}(\pi(X),\theta(X)) \\ & - p_{10}(\pi(X),\theta(X)) \log(p_{10}(\pi(X),\theta(X)) \\ & - (1-\pi(X)) \log(1-\pi(X)) \quad \big] \\ \text{subject to} \quad & \frac{1}{M}\sum_X \theta(X) = \psi\,,\end{aligned} \tag{1}$$

where p_{ij} is the probability of $R_i = 1, S_i = j$ and M is the number of distinct X values. Taking first order conditions and solving gives us the result that

$$\theta(x) = \frac{1}{1+\exp\left(\frac{\lambda}{\pi(x)}\right)}\,. \tag{2}$$

The λ in this expression is the Lagrange multiplier on the constraint that the integral of θ over x be equal to ψ. The value of λ therefore determines ψ. With this parameterization, the likelihood for a sample of size N, with S_{11}, S_{10} and S_0 being the set of sample values with $R_i = 1, Y_i = 1$, $R_i = 1, Y_i = 0$, and $R_i = 0$, respectively, becomes

$$\prod_{S_{11}} \frac{\pi(X_i)}{1+\exp\left(\frac{\lambda}{\pi(X_i)}\right)} \prod_{S_{10}} \frac{\pi(X_i)\exp(\frac{\lambda}{\pi(X_i)}}{1+\exp\left(\frac{\lambda}{\pi(X_i)}\right)} \prod_{S_0} (1-\pi(X_i))\,. \tag{3}$$

Forming the log likelihood and taking first order conditions for a maximum with respect to λ gives us

$$\sum_i \frac{\theta(X_i)R_i}{\pi(X_i)} = \sum_i \frac{Y_i R_i}{\pi(X_i)}\,. \tag{4}$$

The right-hand side is of course just N times the Horwitz-Thompson estimator. The left-hand side is N times a sample average of an i.i.d. random variable whose expectation is $\psi = (1/M)\sum_X \hat{\theta}(X)$. So the MLE chooses λ to make an estimate of ψ based on the sampled values of $\hat{\theta}(\pi(X))$ match the Horwitz-Thompson estimator. Since the right-hand-side can exceed one, the MLE with some probability sets $\lambda = \infty$ and therefore $\psi = 1$, but

[1] In my earlier note on Wasserman's textbook example, I applied nearly the same principle, but started from an assumption that we would treat the observables as Y_i/π_i only. Here we derive the focus on Y_i/π_i, rather than assuming it.

the MLE of ψ never exceeds one. It is not exactly the Horwitz-Thompson estimator, because the left-hand-side of (4) is equal to $(1/M)\sum_X \hat{\theta}(X)$ only in expectation, not exactly, but it will clearly be close to Horwitz-Thompson in large samples. A Bayesian posterior mean will clearly also be close to Horwitz-Thompson in large samples.

So we have arrived at a Bayesian estimator that improves on Horwitz Thompson (since it is admissible) without any "frequentist pursuit". All that was necessary was that we recognize that we are concerned about selectivity bias and that therefore we want our prior to make dependence between π and θ have non-negligible prior probability.

III. CONTINUOUSLY DISTRIBUTED X

So far, we have considered only the case of discretely distributed X, and this does not really confront the Robins-Ritov-Wasserman asymptotic theorems.

Robins, Ritov and Wasserman claim that Bayesian methods hit special difficulties *because* they insist on recognizing that knowledge that the true $\theta()$ is a measurable function on X-space provides useful information. If $\theta()$ is a stochastic process on X-space that is with probability one a measurable function of X, it is impossible that $\theta(X), \theta(X')$ be independent for all pairs X, X' in X-space. That is because measurable functions are "almost continuous". They can have lots of discontinuities, but they must be well approximated by continuous functions over X-sets of large measure. The methods proposed above, for discrete X treat draws of $\theta(X_i)$ as independent, conditional on a small number of parameters, regardless of how close together in X-space the draws may be.

Of course one possibility is to continue to postulate an exact functional relationship between θ and π, as we did above in the entropy-maximizing model. This would keep $\theta(X), \theta(X') \mid \pi(X)\pi(X')$ trivially independent, since there would be no variation in the θ's conditional on the π's. More generally, we could model $\theta() \mid \pi()$ as having a conditional mean given π of $\hat{\theta}(\pi(X_i))$ and flexibly parameterize the $\hat{\theta}()$ function.

But the theorems Wasserman and Robins invoke to suggest problems with Bayesian inference seem to require independence of the prior distributions of $\theta()$ and π. As should be clear by this point, there is no reason Bayesian inference should involve a prior making θ and π independent if estimation of ψ and concern about selection bias are prominent issues.

But it may nonetheless be worthwhile to display an example of Bayesian estimation in an infinite dimensional space, with π and θ a priori independent, so we can see the pathology that the asymptotic theory suggests will be present.

This is much easier if we switch to a different example, that Robins and Wasserman have used in an earlier unpublished paper that is available on the internet. Harmeling and Toussaint discuss this same model. It is a standard nonlinear, non-parametric regression model:

$$Y_i = \theta(X_i) + \varepsilon_i\,, \qquad \varepsilon_i \mid X_i \sim N(0, \sigma^2) \tag{5}$$

$$X_i \text{ has pdf } \pi(X), \text{ with support } [0,1]\,. \tag{6}$$

It is assumed that the observations on Y_i and X_i are jointly i.i.d. across i.

A standard approach here might be kernel estimation of θ. A Bayesian approach might postulate a Gaussian stochastic process as a prior on $\theta()$. Harmeling and Toussaint suggest just such an approach. They give explicit results only for the case where the cross-covariance function between θ and π has a Dirac delta function component at 0, dependent on π. This is interesting, but such a cross-covariance function corresponds to a generalized stochastic process, which does not have measurable time paths. They display formulas for more general Gaussian processes, but don't explore the behavior of the resulting estimators. So we will follow Wahba (1990) and postulate that $\theta()$ is distributed as the sample path of a Wiener process on $[0,1]$.[2] Wiener process time paths are quite non-smooth, being nowhere differentiable, but they are continuous and integrable. They can approximate a wide variety of paths quite well. But paths for a standard Wiener process are very unlikely to show rapid oscillations within the $[0,1]$ interval. If in fact $\theta()$ shows such rapid oscillations, Bayesian estimation based on the assumption that $\theta()$ is a draw from a Wiener process, like frequentist kernel estimates of θ, will not capture the oscillations until the sample size has become quite large. The Bayesian estimates behave much like kernel estimates, except that they in

[2] Examples of such a Bayesian approach to nonlinear regression are explored in more detail in my 2000 paper.

effect adapt bandwidth of the kernel to the local density of sample values of X. And of course kernel estimates will have bandwidth shrinking toward zero at some rate. If θ oscillates rapidly, and $\pi()$ also oscillates rapidly in a way that makes it correlate with $\theta()$, estimates of $\psi = \int \theta(x)dx$ will be biased until effective bandwidth has become small enough to track the oscillations.

Consider what happens when $\theta()$ oscillates between the values 2 and 1 over 32 equal-length subintervals of $[0, 1]$, while $\pi()$ also oscillates, between the values .1 and .9, over those same 32 subintervals. So long as the effective bandwidth is greater than about .03, the Bayesian or kernel estimates will be biased, because they average together Y values that have different θ's and different π weights. Figures 1-3 display the true θ function (in red) and the Bayesian posterior mean estimates of it using as prior on θ a standard Wiener process with variance 10 at $x = 0$ and 11 at $x = 1$, and assuming $\sigma^2 = .21$. With only 100 observations, as in Figure 1, most of the subintervals have very few observations, especially those in the odd-numbered positions. The estimate of $\theta()$ is therefore far too low at every point in $[0, 1]$. With 1,000 observations, as in Figure 2, the oscillatory nature of θ is becoming apparent, but the bias is still great. With 10,000 observations, as in Figure 3, the estimate $\hat{\theta}()$ is fairly accurate, so the downward bias in ψ is small. It should be intuitively clear that, no matter what stochastic process I use as my prior on $\theta()$, if the prior is independent of π, we can for any sample size M specify measurable $\pi(), \theta()$ pairs that will make bias large for all sample sizes less than M.

Here as in the previous examples, a Horwitz-Thompson style estimator, the sample mean of $Y_i/\pi(X_i)$, is unbiased and consistent for ψ, so long as $E[Y_i/X_i]$ exists.

But for a practicing Bayesian, or a frequentist who approaches this problem with kernel estimation, this is not devastating news. Using a prior that makes θ and π independent amounts to asserting that selection bias is likely to be small, and if that is true, and if the true θ is as smooth as a Wiener process sample path, the Bayesian approach or a kernel estimate will be much more efficient than Horwitz-Thompson. There is nothing "conservative" about reporting, because of use of inefficient tools of inference, that there is no evidence for (say) a large value of ψ when in fact sensible beliefs about the smoothness of $\theta()$ imply that there is in fact such evidence. Furthermore a Bayesian, or a sensible frequentist using

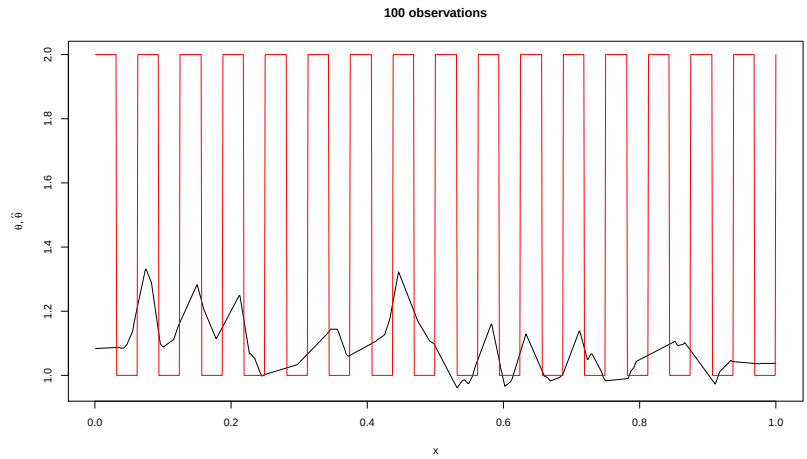

FIGURE 1

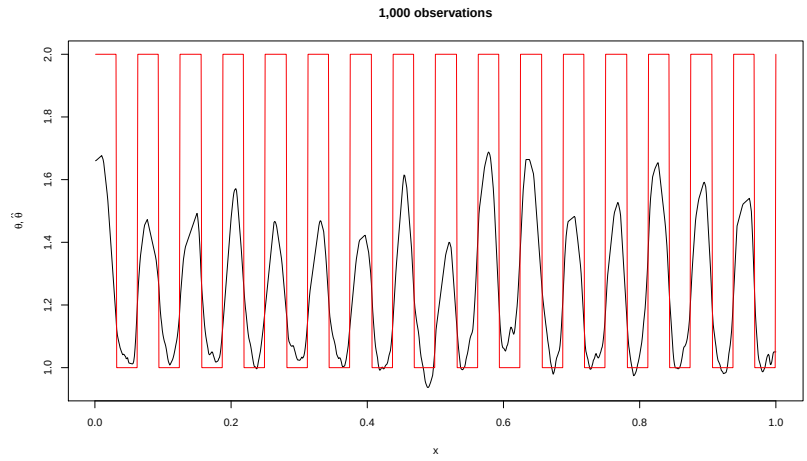

FIGURE 2

kernel methods, who saw Figure 2, and who knew the π function, would surely realize that selection bias probably was strongly present. There is no logical difficulty for Bayesian inference in the idea that observations might lead one to generalize the prior. It can be thought of as part of an implicit Bayesian sieve — we first use a simpler prior, but check the odds on it relative to one that allows additional free parameters, expanding to the more general prior if the odds ratio suggests that's necessary.

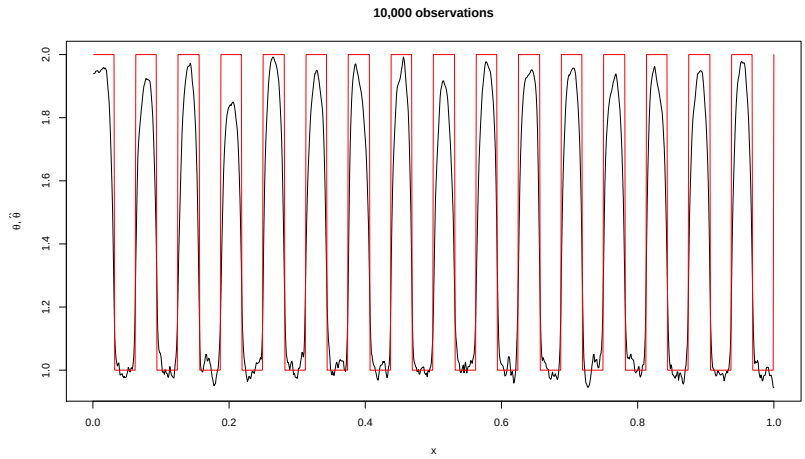

FIGURE 3

While this example does represent estimation of an unknown parameter $\theta()$ in an infinite-dimensional space, it is a space of functions on $[0,1]$. Robins and Wasserman in some of their discussion invoke the possibility of $\theta()$ lying in a space of functions on $[0,1]^d$, with d, say 100,000. Since $[0,1]^d$ and $[0,1]$ can be mapped onto each other by a map measurable in both directions, the size of d should not have an impact on the underlying mathematics. It does make a difference, though, if one believes that the dimensions should in some sense all be treated symmetrically. When we think about arguments justifying a claim that the residual in a linear model is independent of the explanatory variable X, one common consideration is that the residual is influenced by a large number of similar sources of disturbance, no one of which is very much correlated with X. Similarly in these examples, if we believe θ probably depends on a large number of X variables, no one of which has any special relation to $\pi(X)$, this might justify a conclusion that selection bias is very unlikely in the problem at hand. In that case, exploiting the extra efficiency that is available from this assumption makes perfect sense, and using Horwitz-Thompson would be a pointless sacrifice of efficiency.

But as should be clear by this point, in any situation where Horwitz-Thompson looks attractive, we must believe that selection bias is at least a realistic possibility. That means, if X takes values in $[0,1]^{100,000}$, that we nonetheless do not think it likely that $\theta(X)$ and $\pi(X)$ are unrelated, so a

prior that treats all dimensions of X space symmetrically makes no sense. There are many possible approaches to specifying a prior that reflects such beliefs, and they should be adapted to what we know about the particular data set and the particular uses contemplated for the analysis. For example, we might use a prior in which $E[\theta(X) \mid \pi(X)] = \hat{\theta}(\pi(X))$ is an unknown smooth function of the known $\pi(X)$, with $\theta(X) - \hat{\theta}(X)$ a Gaussian process with $\text{Cov}(\theta(x), \theta(x') \mid \pi()) = 0$ unless $|x - x'| < \delta$, with δ extremely small. This will lead to something close to kernel smoothing over $\pi(X)$ values, so long as sample sizes are modest enough that few if any observed X values are within δ of one another. So in our Figures 1-3 examples of oscillating π and θ, where $\pi(X)$ takes on only two values, it would lead to averaging Y_i values for $\pi(X_i) = .1$, averaging Y_i values for $\pi(X_i) = .9$, and taking the unweighted average of these two numbers. This would provide unbiased results even in small samples.

Another approach would be to order the dimensions of X with those elements of the X vector most likely to be important for both π and θ early in the ordering, and then make $\theta(X)$ a linear function of X with coefficients on low-numbered dimensions of X having larger prior variances than those on high-numbered dimensions. This would make $\theta(X)$ and $\pi(X)$ likely to be correlated, despite a large d.

IV. Conclusion

Robins and Wasserman have presented not a single case where likelihood based inference, or Bayesian inference in particular, leads one astray. They have presented examples where naive approaches to specifying priors on infinite dimensional spaces can unintentionally imply dogmatic beliefs about parameters of interest. Such examples are interesting and instructive, but they are not cases where a Bayesian approach to inference fails to give good results.

References

Harmeling, S., and M. Toussaint (2007): "Bayesian Estimators for Robins-RitovâĂŹs Problem," Discussion paper, School of Informatics, University of Edinburgh, http://eprints.pascal-network.org/archive/00003871/01/harmeling-toussaint-07-ritov.pdf.

SIMS, C. A. (2000): "Using a Likelihood Perspective to Sharpen Econometric Discourse: Three Examples," *Journal of Econometrics*, 95(2), 443–462, http://www.princeton.edu/~sims/.

WAHBA, G. (1990): *Spline Models for Observational Data*, vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia.

DEPARTMENT OF ECONOMICS, PRINCETON UNIVERSITY
E-mail address: sims@princeton.edu