Robins has pointed out that I mistook the intended interpretation of
the blog example. The high-dimensional $X_i$ random variable is observed,
not unobserved. In that case the original form of the likelihood is correct,
but the issues in formulating a prior on the high-dimensional parameter
space remain similar.

We know the selection probabilities $\pi(X_i)$ for each observation because
we know the function $\pi()$. We do not know the function $\theta()$, characteriz-
ing the probabilities of $Y_i = 1$ given $X_i$. In some applications, we might
think it reasonable to assume that the two random variables $\pi(X_i)$ and
$\theta(X_i)$ are independent, that is the selection mechanism is independent of
$E[Y \mid X]$. As should be intuitively clear, that makes the problem much
easier, as we can simply throw away the $R_i = 0$ observations and work
with the remaining sample as an i.i.d. sample with the same distribution
of $\theta(X_i)$ as an unselected sample. The problem is still high-dimensional,
but in a straightforward way. Let’s consider that case first.

I. THE PROBLEM WITH NO SELECTION BIAS

Our prior will be a probability distribution over functions $\theta()$ defined
on $X$, i.e. a stochastic process with $X$ as index set. $X$ is a $d$-dimensional
unit cube, with $d$ large. One kind of prior, which might make sense in
some applications, would postulate that $\theta()$ is drawn from a highly non-
smooth stochastic process with the property that $E[\theta(X)] = \hat{\psi}$ for all $X$.
To make this more concrete, consider the following Bayesian sieve setup.
We postulate a sequence of finite-parameter processes for $\theta()$, with the
$k$'th having $\theta()$ constant within each of $k^d$ cells in $X$-space, defined by
partitioning $X$-space into $d$-cubes with side $1/k$. Say $d$ is 10. We put zero
probability on $k < 100$, and, say, probability $2^{k-99}$ on $k$'s 100 or larger.
Within each of these spaces, $\theta_{jk}$, the value of $\theta()$ within the $j$'th cell of the
$k$'th partition, is drawn from a distribution on $[0, 1]$ with mean $\hat{\psi}$. (Hence-
forth we’ll drop the $k$ subscript, as we deal mainly with inference for a
single $k$, discussing inference for $k$ later and informally.) $\hat{\psi}$ itself is drawn
from a $U(0, 1)$ distribution. The number of cells in the $k$'th partition being
$k^d$, or $10^{20}$ for the $k = 100$, $d = 10$ case, the chance that a sample of size $N$
will have two or more $X_i$ draws in the same cell is less than $N/(10^{20} - 1)$,
so there is less than 1/100 probability of our sample containing $X$’s from
the same cell unless the sample size exceeds $10^{18}$. It is useful to think of
the parameter space within the $k$'th partition as $\{\hat{\psi}, \tilde{\theta}_j\}$, where $\tilde{\theta}_j$ is the
value of $\theta - \hat{\psi}$ within the $j'$th cell. The likelihood function times prior is now
\[
\prod_{i=1}^{N}(\hat{\psi} + \hat{\theta}_{j(i)})^{Y_i}(1 - \hat{\theta}_{j(i)})^{1-Y_i}g(\hat{\theta}_{j(i)}) \prod_{m \notin \{j(\ell), \ell = 1, \ldots, N\}} g(\hat{\theta}_m),
\] where $j(i)$ is the index of the cell in which $X_i$ falls. So long as we in fact have a sample with no $X$'s in the same cell, it is easy to integrate out the $\hat{\theta}_j$ terms here. Those for cells, indexed by $m$, that do not appear in the sample, just disappear, as we are only integrating their pdf's. Those for $\theta_{j(i)}$'s, since each of them appears just once, in either the form $\hat{\psi} + \hat{\theta}_{j(i)}$ or $1 - \hat{\psi} - \hat{\theta}_{j(i)}$, also integrate out easily, because the integrals are expected values and by construction $\hat{\theta}_{j(i)}$ has mean zero for all $i$. So the posterior on $\hat{\psi}$, once the nuisance parameters $\hat{\theta}_j$ are integrated out, is
\[
\hat{\psi}^n(1 - \hat{\psi})^m,
\] where $n$ is the number of $Y = 1$ observations in the observed sample and $m$ is the number of $Y = 0$ observations. The posterior mean of $\hat{\psi}$ is $n/(n + m)$, and this is a good estimate of
\[
\psi = \int \theta(x)dx = \frac{1}{k^d} \sum_{j=1}^{k^d} \theta_j,
\] because according to our prior $\psi$ is the average of a very large sum of i.i.d, bounded random variables with the same mean. Different forms of distribution for $\hat{\theta}_j$ would imply different distributions for $\psi \mid \hat{\psi}$, but with sample size $N$ much less than the $10^{20}$ cells, the $\hat{\psi} - \psi$ will be negligibly small compared to posterior uncertainty about $\hat{\psi}$ itself.

What about considering higher values of $k$? So long as there are no repeated cells in the sample for $k \geq 100$, the posterior has the same form for all $k$, so inference does not depend on $k$. Once we reach sample sizes with sizable numbers of repeats for cells, we would in principle want to compute posterior probabilities across $k$. This is a lot of work, though, and might not be worthwhile. The “$k^d$ cubes” models give the same results as would any model that implies independence, or near-independence, of $\theta(X_i)$ values that are not extremely close to each other, conditioning on their common, unknown mean. Though it is nice to think of the sequence of $k^d$-cell spaces with $k$ growing, because using that sieve would yield consistency and precise inference on a fairly big class of $\theta()$'s, both smooth...
and non-smooth, a serious empirical researcher who thought it quite possible that in her sample quite a few of the $X_i$’s are close enough that the corresponding $\theta(X_i)$’s are a priori dependent, would want to think more carefully about how to model local smoothness.

Furthermore, though there is indeed possible information about $\psi$ in $X$, in this (independence of $\theta$ from $\pi$) case, the gain from using this information is likely to be modest. We have an i.i.d. sample of $Y$, after all, whose expected value is $\psi$. What do we gain from using a smoothness prior to model $\theta()$? Even in the best case, where we have a single model for $\theta()$ with a small number of unknown parameters, the rate of convergence of our estimate of our estimate of $\psi$ will remain $1/\sqrt{N}$. Consider the case where $X$ just takes on two values, for men and women. We believe $\theta$ is different for men and women, we believe we know the frequency of men and women in the population, and we have a random sample of the population. Using information on $X$ would let us estimate $\theta$ separately for men and women, then weight the estimates together using the population frequencies of the two sexes, rather than the frequencies in our random sample. This would produce sharper inference for $\psi$, if we were sure of our population frequencies, but the loss from ignoring $\theta$ differences across sexes would not be great in moderately large samples.

With $X$ fairly high-dimensional and expected to be smooth, we could of course start our sieve with a sequence of logit or probit models that successively expand the number of dimensions of $X$ allowed to influence $\theta$, and then continue to expand by adding higher order interaction terms. This is the approach most applied researchers would no doubt take if the $X$’s were interpretable variables and there was reason to take a strong smoothness assumption seriously.

II. WITH SELECTION BIAS

We would not have bothered to specify $\pi()$ and stipulate that it is known if we thought it were irrelevant. It makes no sense, therefore, to use a stochastic process on $X$ as as a distribution for $\theta$ if it implies the random variables $\theta(X)$ and $\pi(X)$ are independent. Of course it may be true that we have little idea of what the nature of this dependence might be. We might, for example, think it equally likely that $E[\theta(X) \mid \pi(X)]$ is increasing in $\pi(X)$ or decreasing in $\pi(X)$. That does not mean that we
believe them independent. It means we need to include unknown parameter in our prior that control the form of the dependence.

One approach is to use a sequence of parameter spaces in each of which we construct a partition of \([0, 1]\) and model \(\theta(X)\) for \(X\)'s that land in the \(j\)th bin for \(\pi(X)\) just as we modeled \(\theta\)'s on the whole \(X\)-space in the previous section. That is, we model \(\theta(X)\) for \(X\) in the \(j\)'th bin of \(\pi\) values as a highly non-smooth function of \(X\) with conditional mean given \(\pi(X)\) in the \(j\)'th bin \(\hat{\phi}_j\). Everything else is as in the previous section, except carried out \(\pi\)-bin by \(\pi\)-bin. Since we are unlikely to know how rapidly \(E[\theta(X) \mid \pi(X)]\) varies with \(\pi\), we would probably want to examine posterior weights on several levels of \(\pi\)-partitions in even modest sized samples. On the other hand, because by construction \(\pi\) varies little within bins, selection bias must become small when the bins get small, even if \(E[\theta(X) \mid \pi(X)]\) varies a lot within bins. Finer bins, after a point, mainly increase precision of estimates, rather than reducing bias.

As in the previous section, we might also decide to consider models with \(\theta(\cdot)\) a smooth function of \(X\) within bins. For example, we could make \(\theta\) a logit or probit function of \(X\), with an expanding sequence of dimensions of \(X\) included, then polynomial terms, as suggested in the previous section. Here it would be natural to have one of the variables in the model be \(\pi(X)\) itself. Probably this would require numerical analysis of the posterior, but in modest-sized, well-behaved models posterior probability would probably concentrate on low-dimensional models.

### III. Conclusion

Here again, as in the case of the model with \(X\) unobservable, a straightforward Bayesian approach to the problem yields practically usable estimators that would deal with selection bias and probably have good frequentist properties — though again, I have not checked frequentist properties or used them to motivate the Bayesian prior.

As in the cases I discussed in rounds 1 and 2, it is possible, because this is high-dimensional problem, to stumble into specifying a prior whose finite-dimensional sections seem plausible, but that inadvertently is dogmatic about important aspects of the problem. For example, if we made our non-smooth models of \(\theta(X)\) over fine partitions of \(X\) space simply assert i.i.d. distributions of \(\theta\) across bins, without the dependence implied
by the $\hat{\psi}$ parameter, we would implicitly be asserting that we know $\psi$ with high precision before looking at any data.

Department of Economics, Princeton University

E-mail address: sims@princeton.edu