HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2016 June 01.

Published in final edited form as:

Biometrics. 2015 June ; 71(2): 296–299. doi:10.1111/biom.12273.

On Bayesian estimation of marginal structural models

James M. Robins[1], **Miguel A. Hernán**[1], and **Larry Wasserman**[2]

[1]Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

[2]Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Saarela et al. are concerned with integrating propensity scores into a Bayesian framework. Some of us have previously written (Robins and Ritov, 1997; Robins and Wasserman, 2000; http://normaldeviate.wordpress.com/2012/08/28/robins-and-wasserman-respond-to-a-nobel-prize-winner/; posted 28 Aug 2012, accessed 1 Oct 2014) about this topic, every time making much the same argument. Here we present a simplified version that captures the main points.

A simple setting

Though our argument applies to the complex observational data considered by Saarela et al, it is easier to understand it in the simpler setting of a double-blind, placebo-controlled randomized clinical trial of a non-time-varying treatment and under complete compliance. In the spirit of the authors, we assume the trial subjects are representative of a much larger population and the trial results will guide treatment decisions in the population.

Let $\mathbf{V} = \{Z_i, X_i; Y_i; i = 1, \ldots, n\}$ denote the data on the n trial subjects, where Z_i is the binary treatment arm indicator, Y_i is the binary outcome, and X_i is a high-dimensional vector of baseline covariates. The randomization probabilities $pr[Z = 1|X]$ are chosen by a randomizer. By de Finetti's theorem (e.g., Bernardo and Smith, 1994), a Bayesian can write the marginal density $p(\mathbf{V})$ of V

$$p(\mathbf{V}) = \int_{\phi,\gamma} p(\boldsymbol{Z}, \mathbf{X}, \boldsymbol{Y}; \phi, \gamma) p(\phi, \gamma) d\mu(\phi) d\mu(\gamma),$$

$$p(\boldsymbol{Z}, \mathbf{X}, \boldsymbol{Y}; \phi, \gamma) = \mathscr{L}_1(\phi)\mathscr{L}_2(\gamma), \mathscr{L}_1(\phi) = \prod_{i=1}^{n} L_{1i}(\phi), \mathscr{L}_2(\gamma) = \prod_{i=1}^{n} L_{2i}(\gamma),$$

where $L_1(\varphi) = f(Y|Z, X; \varphi_1) f(X; \varphi_2)$and $L_2(\gamma) = f(Z|X, \gamma)$. We have already integrated out the authors' unmeasured frailty U.

The propensity score $e(X; \gamma^\dagger) = pr[Z = 1|X; \gamma^\dagger]$ is known to the randomizer by design, but let us provisionally assume that our Bayesian does not know it so he treats γ as random. [We assume there exist true values $(\varphi^\dagger, \gamma^\dagger)$ of (φ, γ) but, even if not, our argument, slightly modified, is still valid.]

Like the authors, we take our goal to be the estimation of the counterfactual probabilities $\theta^{\dagger}=\left(\theta_0^{\dagger},\theta_1^{\dagger}\right)$, where $\theta_z^{\dagger}=pr(Y_z=1)$, and Y_z is a subject's counterfactual response under treatment level z. Randomization implies that $\theta_z^{\dagger}$ is identified and equals

$$\theta_z^{\dagger}=\int pr\left[Y=1|Z=z,X=x;\phi_1^{\dagger}\right]f\left(x;\phi_2^{\dagger}\right)dx$$

Why Bayesian inference must ignore the propensity score

Bayesian logic is rigidly defined: given a likelihood and a prior, one turns the Bayesian crank to obtain a posterior. There is no wiggle room. A fact concisely summarized in the slogan "There is no Bayes but Bayes." Because the parameter θ of interest is a functional of the parameters φ, the posterior for θ is completely determined by the posterior of φ. If φ and γ are a priori independent, the posterior of φ is obtained from the $\mathscr{L}_1$ (φ) factor of the observed likelihood and the prior p (φ) for φ.

Therefore, Bayesian inference concerning θ cannot be a function of the propensity score e (X; $\gamma^{\dagger}$) because the Bayesian's posterior for φ–and thus for φ–does not depend on γ. Saarela et al. assume φ and γ are a priori independent and yet argue that inverse probability weighting by a function of the propensity score e (X; $\gamma^{\dagger}$) can be given a Bayesian interpretation. In light of the above, their arguments cannot be valid.

Why propensity scores should not be ignored

Why do the authors, as Bayesians, work so hard to include propensity scores in their inference when, according to Bayes, they are irrelevant? Our guess is that the authors recognize that an analysis–Bayesian or otherwise–that ignores a known propensity score can go seriously wrong because one's prior knowledge of pr [Y = 1|Z = z, X] is meager when X is high-dimensional.

Specifically, consider any estimator $\hat{\theta}$ of $\theta^{\dagger}$ that does not depend on the known propensity score. Robins and Ritov (1997) prove that $\hat{\theta}$ cannot be uniformly consistent for $\theta^{\dagger}$ over the large infinite dimensional model $\mathscr{M}$ that includes any laws b_z (X) = pr [Y = 1|Z = z, X], any density f (x) for X, and any propensity function e (X) = pr [Z = 1|X] bounded away from 0 and 1. The practical implication of this theorem is that, whenever e (X; $\gamma^{\dagger}$) is a complex function of our high dimensional X and the (infinite-dimensional) parameters γ and φ are a priori independent, the posterior for θ will fail to concentrate around the true value of $\theta^{\dagger}$ as n goes to infinity because any model we specify for f (Y|Z, X; φ_1) is almost certainly incorrect (imposing smoothness will not really help). This practical implication is obvious; the Robins and Ritov theorem serves as a mathematical formalization.

In contrast, estimators that use the known randomization probabilities, like the Horvitz-Thompson (1952) estimator of $\theta_z^{\dagger}$, can be uniformly $n^{1/2}$-consistent over $\mathscr{M}$. The deficiencies of the Horvitz-Thompson estimator–it may exceed 1, it ignores data on X except for the one-dimensional summary e (X; $\gamma^{\dagger}$), and it can be very inefficient–can be remedied by using an improved version: the so-called *locally semiparametric efficient regression*

estimator (Scharfstein et al., 1999). In observational studies, this estimator is doubly robust when the unknown $e(X; \gamma^{\dagger})$ is replaced by an estimate. More efficient doubly robust estimators are reviewed by Rotnitzky et al (2012).

When the priors are dependent

Our argument relies on the authors' assumption that φ and γ are a priori independent. This assumption is often reasonable, as shown in the Appendix. However, when φ and γ are a priori dependent–which implies that the posterior for θ will depend on the propensity score $e(X; \gamma)$–two new issues arise.

First, in observational studies with $\gamma^{\dagger}$ unknown, the posterior for γ will depend on the data through the φ part of the likelihood. The authors find this troubling since this procedure fails to "retain the balancing property of propensity scores." But again true Bayesians cannot have it both ways. The parameters φ and γ are either a priori independent or they are not. If one wants to use dependent priors to make the posterior for θ to depend on the propensity score, then one must accept that the posterior for the propensity score will depend on the φ part of the likelihood.

The above is not only a philosophical issue concerning schools of inference. It implies that true Bayesian inference based on finite-dimensional working models will generally fail to be doubly robust since misspecification of either the outcome or propensity model will bleed into the estimation of the parameters of the other correct model. As the authors discuss in their supplemental material, this lack of double robustness confronted both McCandless et al (2010) and Zigler et al (2013) who proposed approaches to prevent the bleeding. But, as useful as the approaches may be, they cannot be truly Bayesian.

Second, even in a randomized trial with known propensity score, simply making φ and γ dependent a priori does not imply that the posterior for θ will concentrate around the truth. The dependent prior still has to be carefully engineered for that to happen. As an example we can construct a locally semiparametric efficient Bayes estimator $\hat{\theta}_{Bayes}$ as follows. We assume that, conditional on the known $\gamma^{\dagger}$ and k given functions $w_{m,z}(x)$, $pr(Y = 1|Z = z, X = x; \varphi_{1,z})$ is a finite-dimensional parametric function $\text{expit}\left\{\sum_{m=1}^{k} \eta_{m,z} w_{m,z}(x)\right\}$ with $w_{k,z}(x) = 1/pr(Z = z|X = x; \gamma^{\dagger})$. Then, if we put smooth or non-informative priors over the parameters $\varphi_{1z} = (\eta_{1,z}, \ldots, \eta_{k,z})$, the Bayes estimator $\hat{\theta}_{Bayes}$ will be asymptotically equivalent to the frequentist locally semiparametric efficient estimator cited earlier and thus be $n^{1/2}$-consistent. Thus, by using carefully tuned dependent priors, we have obtained a Bayes estimator that has good frequentist behavior by mimicking a locally semiparametric efficient frequentist estimator.

But this is a Pyrrhic victory. If we need to engineer the dependent prior just to mimic a frequentist answer, is it really Bayesian inference? We call Bayesian inference which is carefully manipulated to force an answer with good frequentist behavior, *frequentist pursuit*. There is nothing wrong with it. But if you want to be Bayesian, then accept that, in this example, your posterior will fail to concentrate around the true value.

Conclusion

Our arguments above may have left readers thinking "why bother? If you want good frequentist properties, just use a frequentist estimator rather than embarking on a frequentist pursuit." Indeed, it might appear that we are arguing that the Bayesian machinery should be reserved for implementing subjective Bayes inference that maps prior beliefs to posterior beliefs via the likelihood function, without regard for the frequentist properties of the resulting estimators. While we do believe that investigation of this mapping through Bayesian sensitivity analysis and/or robust Bayes is important and extremely useful, we also believe that the Bayesian approach can play other important roles, even when one is interested in good frequentist properties. We consider three cases.

First, Bayesian logic and machinery may sometimes lead to procedures with provably better frequentist operating characteristics than their current competitors, even asymptotically. An example is the conditional predictive and partial posterior predictive p-values of Bayarri and Berger (2000).

Second, when modelling complex phenomena (particularly in small and moderate samples), there may be Bayesian approaches that are rather straightforward to motivate and implement even when there is no good frequentist alternative, so the Bayes estimator is the best, or perhaps the only, frequentist game in town.

Third, to improve decision making under uncertainty, one can adopt a Bayes-frequentist compromise (Robins 2004, Sec 5.2) that combines honest subjective Bayesian inference with good frequentist behavior even when, as above, the model is so large and the likelihood function so complex that standard (uncompromised) Bayes procedures have poor frequentist performance. It follows immediately from our earlier arguments that such a compromise requires that our subjective Bayesian decision maker is only allowed to observe a specified vector function of X (depending on $e(X; \gamma^{\dagger})$) but not X itself. In this way one can circumvent the problem referred to by Robert (http://xianblog.wordpress.com/2013/01/17/robbins-and-wasserman; posted 17 Jan 2013, accessed 01 Oct 2014) as the *curse of marginalization*: "the classical Bayesian approach is an holistic system that cannot remove information to process a subset of the original problem."

Acknowledgments

This work was partly funded by NIH grant P01 CA134294.

Appendix: Example of a priori independence of the propensity score

Suppose a health insurance company needs to estimate the fraction θ of its patient population that will have a myocardial infarction (MI, $Y = 1$) in the next year, so as to determine the need for cardiac unit beds. They have 300 potential risk factors $X = (X_1, \ldots, X_{300})$ measured on each member. A general epidemiologist had earlier studied risk factors for MI by following 5000 patients for a year. Because MI was a rare event, he oversampled subjects whose X, in his opinion, indicated a higher conditional probability $b(x) = E[Y|X =$

x] of $Y = 1$. Hence, with Z the inclusion indicator, the sampling fraction $e(x) = pr(Z = 1|X = x)$ was a known but complex function.

The world's leading heart expert, our Bayesian, was hired to estimate $\theta = \int b(x)\, p(x)\, dx$, where $p(x)$ is the marginal density of x, based on the study data (**X,Z,ZY**). As world's expert, his beliefs about the risk function $b(\cdot)$ would not change upon learning the propensity score function $e(\cdot)$, as $e(\cdot)$ only reflected a nonexpert's beliefs. Hence the functions $b(\cdot)$ and $e(\cdot)$ are a priori independent. [Nonetheless, he would believe with high probability that the random variables $b(X)$ and $e(X)$ were positively correlated, knowing that the epidemiologist had read the expert literature on risk factors for MI.]

Robins and Ritov (1997) showed that once any Bayesian, cardiac expert or not, thoroughly queries the epidemiologist who selected $e(\cdot)$ about his reasoned opinions concerning $b(\cdot)$ (but not about $e(\cdot)$), the Bayesian will then have independent priors. The idea is that once you are satisfied that you have learned from the epidemiologist all he knows about $b(\cdot)$ that you did not, you will have an updated prior for $b(\cdot)$. Your updated prior for $b(\cdot)$ cannot then change if you subsequently are told $e(\cdot)$. Hence, we could take as many Bayesians as you please and arrange it so all had $b(\cdot)$ and $e(\cdot)$ a priori independent. This last argument is quite general and applies to many settings.

References

Bayarri MJ, Berger JO. P-values for composite null models. Journal of the American Statistical Association. 2000; 95:1127–1142. Rejoinder, pp 1168–1170.

Bernardo, JM.; Smith, AFM. Bayesian Theory. Chichester: Wiley; 1994.

Horvitz DG, Thompson DJ. A Generalization of Sampling Without Replacement From a Finite Universe. Journal of the American Statistical Association. 1952; 47:663–685.

Robins, JM. Optimal structural nested models for optimal sequential decisions. In: Lin, DY.; Heagerty, P., editors. Proceedings of the Second Seattle Symposium on Biostatistics. New York: Springer; 2004.

Robins JM, Ritov Y. "Toward A Curse Of Dimensionality Appropriate (CODA) Asymptotic Theory For Semi-parametric Models". Statistics in Medicine. 1997; 16(3):285–319. [PubMed: 9004398]

Robins JM, Wasserman L. Conditioning, likelihood, and coherence: a review of some foundational concepts. Journal of the American Statistical Association. 2000; 95:1340–1346.

Rotnitzky A, Lei Q, Sued M, Robins JM. Improved double-robust estimation in missing data and causal inference models. Biometrika. 2012; 99:439–456. [PubMed: 23843666]

Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association. 1999:1096–1120.