Abstract Title Page

Not included in page count.

Title: Bayesian Model Averaging for Propensity Score Analysis

Authors and Affiliations:

David Kaplan and Jianshen Chen
Department of Educational Psychology
University of Wisconsin - Madison

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

In a classic paper, Rosenbaum and Rubin (1983) proposed propensity score analysis as a practical tool for reducing selection bias through balancing on measured covariates. Since then, a variety of propensity score techniques have been developed for both the estimation and the application of the propensity score. Models for estimating the propensity score equation have included parametric logit regression with chosen interaction and polynomial terms (e.g., Dehejia & Wahba, 1999; Hirano & Imbens, 2001), and generalized boosting modeling (McCaffrey, Ridgeway, & Morral, 2004). Methods for estimating the treatment effect while accounting for the propensity score include stratification, weighting, matching, and regression adjustment. Rubin (1985) argued that a Bayesian approach to propensity score analysis should be of great interest to the applied Bayesian analyst, and yet propensity score estimation within the Bayesian framework was not addressed until relatively recently. Hoshino (2008) developed a quasi-Bayesian estimation method for general parametric models, such as latent variable models, and developed a Markov chain Monte Carlo (MCMC) algorithm to estimate the propensity score. McCandless, Gustafson, and Austin (2009) provided a practical Bayesian approach to propensity score stratification, estimating the propensity score and the treatment effect and sampling from the joint posterior distribution of model parameters via an MCMC algorithm. The marginal posterior probability of the treatment effect can then be obtained based on the joint posterior distribution. Similar to the McCandless et al. (2009)'s study, An (2010) presented a Bayesian approach that jointly models both the propensity score equation and outcome equation at the same time and extended this one-step Bayesian approach to propensity score regression and single nearest neighbor matching methods.

A consequence of the Bayesian joint modeling procedure utilized by McCandless et al. (2009) and An (2010) is that the propensity score estimates may be affected by the outcome variable that are observed after treatment assignment, and thus result in biased propensity score estimation. This is especially problematic if the relationship between the outcome and the propensity score is misspecified (McCandless, Douglas, Evans, & Smeeth, 2010). To solve this problem, McCandless et al. (2010) utilized an approximate Bayesian technique introduced by Lunn, Best, Spiegelhalter, Graham, and Neuenschwander (2009) for preventing undesirable feedback between propensity score model and outcome model components. Specifically, McCandless et al. (2010) included the posterior distribution of the propensity score parameters as covariate input in the outcome model so that the flow of information between the propensity score and the outcome is restricted. This so-called sequential Bayesian propensity score analysis yields treatment effect estimates that are comparable to estimates obtained from frequentist propensity score analysis. Nevertheless, as McCandless et al. (2010) point out, their method is only approximately Bayesian and also encounters the difficulty that the Markov chain is not guaranteed to converge.

In order to maintain a fully Bayesian specification while overcoming the conceptual and practical difficulties of the joint modeling methods of McCandless et al. (2009) and An (2010), a two-step Bayesian propensity score approach was recently developed by Kaplan and Chen (2012) that can incorporate prior information on the model parameters of both the propensity score equation and outcome model equation. Consistent with Bayesian theory (see e.g., de Finetti, 1974), specifying prior distributions on the model parameters is a natural way to quantify uncertainty - here in both the propensity score and outcome equations.

Kaplan and Chen (2012) conducted three simulation studies as well as a small case study comparing frequentist propensity score analysis with the two-step Bayesian alternative focusing on the estimated treatment effect and variance estimates. The effect of different sample sizes, true treatment effect and choice of priors on treatment e effect ect and variance estimates were also evaluated. Consistent with Bayesian theory, Kaplan and Chen (2012)'s findings showed that lower prior precision of treatment effect is desirable when no prior information is available in order to obtain estimates similar to frequentist results but with more accurate intervals; or higher prior precision is preferable when accurate prior information regarding treatment effect is attainable in order to obtain more precise treatment effect estimates. For the case of small sample size, the Bayesian approach shows slight superiority in the estimation of the treatment effect compared to the frequentist counterpart.

A further study of the covariate balance properties of the Kaplan and Chen (2012) approach was given in Chen and Kaplan. Their results of a case study revealed that both Bayesian and frequentist propensity score approaches substantially reduced initial imbalance and performance on covariate balance was similar in regard to the standardized mean/proportion differences and variance ratios in the treatment group and control group. Similar performance was also found with respect to 95% bootstrap intervals and posterior probability intervals. That is, although the frequentist propensity score approach provided slightly better covariate balance for the propensity score stratification and weighting methods, the two-step Bayesian approach offered slightly better covariate balance under optimal full matching method. Results of Chen and Kaplan's simulation study indicated similar findings. In addition, the Bayesian propensity score approach with informative priors showed equivalent balance performance compared to the Bayesian approach with noninformative priors, indicating that the specification of the prior distribution did not greatly influence the balance properties of the two-step Bayesian approach. The optimal full matching method, on average, offered the best covariate balance compared to stratification and weighting methods for both Bayesian and frequentist propensity score approaches. Chen and Kaplan also found that the two-step Bayesian approach under optimal full matching with highly informative priors provided, on average, the smallest standardized mean/proportion difference and variance ratio of the covariates between the treatment and control groups.

Chen and Kaplan argued that a benefit of conducting Bayesian propensity score analysis is that one can obtain a distribution of estimated propensity scores and thus a distribution of corresponding balance indices (e.g. Cohen's d and variance ratio) so that the variation in balance indices can be studied in addition to the point estimates to assist in balance checking. Good balance is achieved if both the point estimates and the posterior probability intervals of the balance indices fall into the desirable range.

Purpose / Objective / Research Question / Focus of Study:
Description of the focus of the research.

The Bayesian propensity score approaches described in the preceding paragraphs all assume that the propensity score model itself is, in some sense, fixed. Quoting Hoeting et al. (1999):

> "[S]tandard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-con_dent inferences and decisions that are more risky than one thinks they are."(Hoeting, Madigan, Raftery, and Volinsky,1999,pg. 382)

Thus, we argue that it is incorrect to treat the propensity score equation as fixed. Rather, as a model for treatment selection, it is reasonable to assume that many possible models could have been chosen. Therefore, a full accounting of uncertainty in propensity score analysis should also address model uncertainty, and thus the purpose of this paper is to explore Bayesian model averaging in the propensity score context.

Setting:
NA

Population / Participants / Subjects:
NA

Intervention / Program / Practice:
NA

Significance / Novelty of study:
Previous research on Bayesian propensity score analysis did not take into account model uncertainty. In this regard, an internally consistent Bayesian framework for model building and estimation must also account for model uncertainty. The significance of the current study is that it directly addresses the problem of uncertainty in propensity score models via the method of Bayesian model averaging (BMA).

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

Details about the model are given in Appendix B.

Usefulness / Applicability of Method:
The usefulness of the proposed method is that it provides the investigator a way to incorporate prior knowledge regarding the relationship between the covariates and treatment selection (via the Kaplan and Chen, 2012 approach) while at the same time acknowledging model uncertainty via Bayesian model averaging. In addition, we provide a fully Bayesian MCMC methodology to obtain propensity score and treatment effect estimates, as well as R code to conduct such an analysis.

Research Design:

Description of the research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

Our research design will utilize a combination of simulation studies and real data analysis. The simulation study will examine the choice of parameter and model priors. The real data example will examine a model relating full v. part day kindergarten attendance on achievement outcomes for first grade student using the ECLS-K.

Data Collection and Analysis:

Description of the methods for collecting and analyzing data.
(May not be applicable for Methods submissions)
NA

Findings / Results:

Description of the main findings with specific details.

The R code has been written. Preliminary findings suggest that the fully MCMC algorithm for Bayesian model averaging with in the PSA framework provides accurate expected a posteriori estimates of the treatment effect.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

A fully Bayesian approach to propensity score analysis must account for both parameter uncertainty and model uncertainty. Previous Bayesian approaches only examined parameter uncertainty. We find that by accounting for model uncertainty via Bayesian model averaging of the propensity score equation provides very good estimates of the propensity score, which in turn, provides very good estimates of the treatment effect. A central issue in all Bayesian analyses is the elicitation of priors. In the context of Bayesian PSA, priors the propensity score equation parameters can be difficult to elicit. We are presently examining a variety of different so-called "objective priors" for this purpose - including Jeffreys' priors and maximum entropy priors. The results of this work will not be part

References

An, W. H. (2010). Bayesian propensity score estimators: incorporating uncertainties in propensity scores into causal inference*. Sociological Methodology, 40* , 151-189.

de Finetti, B. (1974). *Theory of probability, vols. 1 and 2.* New York: John Wiley and Sons.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94* , 1053-1062.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology, 2* , 259-278.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science, 14* , 382-417.

Hoshino, T. (2008). A Bayesian propensity score adjustment for latent variable modeling and MCMC algorithm. *Computational Statistics & Data Analysis, 52* , 1413-1429.

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., & Neuenschwander, B. (2009). Combining MCMC with "sequential" PKPD modelling. Journal of *Pharmacokinetics and Pharmacodynamics*, *36* , 19-38.

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainly in graphical models using Occam's window*. Journal of the American Statistical Association*, *89* , 1535-1546.

Martin, A. D., Quinn, K. M., & Park, J. H. (2010, May 10). Markov chain Monte Carlo (MCMC) package. http://mcmcpack.wustl.edu/.

McCa_rey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9* , 403-425.

McCandless, L. C., Douglas, I. J., Evans, S. J., & Smeeth, L. (2010). Cutting feedback in Bayesian regression adjustment for the propensity score. *The International Journal of Biostatistics, 6* , Article 16.

McCandless, L. C., Gustafson, P., & Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine, 28* , 94-112.

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org (ISBN 3-900051-07-0)

Raftery, A. E., Hoeting, J., Volinsky, C., Painter, I., & Yeung, K. Y. (2009, September 18). Bayesian model averaging (BMA), version 3.12. http://www2.research.att.com/ volinsky/bma.html.

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models*. Journal of the American Statistical Association, 92* , 179-191.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70* , 41-55.

Rubin, D. B. (1985). The use of propensity scores in applied Bayesian inference. *Bayesian Statistics, 2* , 463-472.

Appendix

A recent paper by x & y (2012) advanced a two-step approach to Bayesian propensity score analysis that was found to quite accurately estimate the treatment effect, while at the same time preventing undesirable feedback between the propensity score model and that outcome model components characteristic of other Bayesian propensity score approaches. We apply Bayesian model averaging to the x & y (2012) model, and describe that model in this section.

Specification of the Two-Step Bayesian PSA model

In the x & y (2012) two-step Bayesian propensity score approach (hereafter, BPSA), the propensity score model specified as the following logit model.

$$Log\left(\frac{e(x)}{1-e(x)}\right) = \alpha + \beta' x, \tag{1}$$

where α is the intercept, β refers to the slope and x represents a design matrix of chosen covariates. For for this step, x & y (2012) used the R package *MCMClogit* to sample from the posterior distribution of α and β using a random walk Metropolis algorithm. After the propensity score estimates are obtained, a Bayesian outcome model is fit in the second step to estimate the treatment effect via various propensity score methods such as stratification, weighting, matching and regression adjustment.

To illustrate the x & y (2012) approach, consider a posterior sampling procedure of a chosen Bayesian logit model with 1000 iterations and a thinning interval of 1. Then for each observation, there will be $m = 1000$ propensity score estimates $\hat{e}(x)$ calculated using propensity score model parameters α and β as follows,

$$\hat{e}(x) = \frac{exp(\alpha + \beta' x)}{1 + exp(\alpha + \beta' x)}. \tag{2}$$

Based on each estimated propensity score, there will be $J = 1000$ treatment estimates generated from posterior distribution of γ ($i = 1, \ldots, m$, $j = 1, \ldots, J$), where γ is the treatment effect. x & y (2012) then provide the following treatment effect estimator,

$$E(\gamma \mid x, y, Z) = m^{-1} J^{-1} \sum_{i=1}^{m} \sum_{j=1}^{J} \gamma_j(\eta_i), \tag{3}$$

where $J^{-1}\sum_{j=1}^{J}\gamma_j(\eta_i)$ is the posterior sample mean of γ in the Bayesian outcome model based on the i^{th} set of propensity scores η_i. This posterior sample mean is then averaged over m sets of propensity scores. The posterior variance of γ is then based on the total variance formula,

$$Var(\gamma \mid x, y, Z) = m^{-1}\sum_{i=1}^{m}\sigma^2_{\gamma(\eta_i)} + (m-1)^{-1}\sum_{i=1}^{m}\{\mu_{\gamma(\eta_i)} - m^{-1}\sum_{i=1}^{m}\mu_{\gamma(\eta_i)}\}^2, \quad (4)$$

where

$$\sigma^2_{\gamma(\eta_i)} = (J-1)^{-1}\sum_{j=1}^{J}[\{\gamma_j(\eta_i) - J^{-1}\sum_{j=1}^{J}\gamma_j(\eta_i)\}]^2, \quad (5)$$

is the posterior sample variance of γ in the Bayesian outcome model under the i^{th} set of propensity scores and

$$\mu_{\gamma(\eta_i)} = J^{-1}\sum_{j=1}^{J}\gamma_j(\eta_i), \quad (6)$$

is the posterior sample mean of γ in the same Bayesian outcome model. Notice that two sources of variation are present in equation (4). The first source of variation is the average of the posterior variances of γ across the posterior samples of propensity scores, represented by the first part of the right hand side of equation (4), and the second source of variation comes from the variance of the posterior means of γ obtained across the posterior samples of propensity scores, estimated by the second part of the right of hand side of equation (4) x & y (2012).

Bayesian Model Averaging

Consider a quantity of interest such as a future observation or a parameter. Following the notation given in Madigan & Raftery (1994), we will denote this quantity as Δ. Next, consider a set of competing models M_k, $k = 1, 2, \ldots, K$ that are not necessarily nested. The posterior distribution of Δ given data y can be written as

$$p(\Delta|y) = \sum_{k=1}^{K} p(\Delta|M_k)p(M_k|y). \quad (7)$$

where $p(M_k|y)$ is the posterior probability of model M_k written as

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{\sum_{l=1}^{K} p(y|M_l)p(M_l)}, \qquad l \neq k. \tag{8}$$

The interesting feature of equation (8) is that $p(M_k)$ will likely be different for different models. The term $p(y|M_k)$ can be expressed as an integrated likelihood

$$p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k, \tag{9}$$

where $p(\theta_k|M_k)$ is the prior density of θ_k under model M_k (Raftery, 1997). Thus, BMA provides an approach for combining models specified by researchers. The advantage of BMA has been discussed in Madigan & Raftery (1994) who showed that BMA provides better predictive performance than that of a single model. Given that the propensity score is the predicted probability of treatment assignment given a set of covariates, we hypothesize that BMA should provide better prediction of treatment assignment than a single propensity score equation.

Computational Considerations

As pointed out by Hoeting (1999), BMA is difficult to implement. In particular, they note that the number of terms in equation (7) can be quite large, the corresponding integrals are hard to compute (though possibly less so with the advent of MCMC), the specification of $p(M_k)$ may not be straightforward, and choosing the class of models to average over is also challenging. The problem of reducing the overall number of models that one could incorporate in the summation of equation (7) has led to a solution based on the so-called *leaps and bounds* algorithm.

For our paper, we propose a fully Bayesian MCMC methodology. In the first step, we use the R program *BMA* to select models (covariates) with certain cumulative posterior probability. In the second step, we use the Bayesian logit program *MCMClogit* to obtain the posterior distribution of propensity scores for each selected model. In the third step, we sample from the posterior distribution of PS in each model with posterior probabilities as weights to obtain final posterior distribution of the propensity score. This is then used in the outcome model via weighting, stratification, or matching.