Journal of Statistical Planning and Inference ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A Bayesian nonparametric causal model

George Karabatsos [a,*], Stephen G. Walker [b]

[a] University of Illinois-Chicago, 1040 W. Harrison St. (MC 147), Chicago, IL 60607, United States
[b] University of Kent, United Kingdom

ARTICLE INFO

Article history:
Received 19 March 2009
Received in revised form
28 October 2011
Accepted 31 October 2011

Keywords:
Bayesian nonparametrics
Causal inference
Observational studies

ABSTRACT

Typically, in the practice of causal inference from observational studies, a parametric model is assumed for the joint population density of potential outcomes and treatment assignments, and possibly this is accompanied by the assumption of no hidden bias. However, both assumptions are questionable for real data, the accuracy of causal inference is compromised when the data violates either assumption, and the parametric assumption precludes capturing a more general range of density shapes (e.g., heavier tail behavior and possible multi-modalities). We introduce a flexible, Bayesian nonparametric causal model to provide more accurate causal inferences. The model makes use of a stick-breaking prior, which has the flexibility to capture any multi-modalities, skewness and heavier tail behavior in this joint population density, while accounting for hidden bias. We prove the asymptotic consistency of the posterior distribution of the model, and illustrate our causal model through the analysis of small and large observational data sets.

1. Introduction

The potential outcomes framework is fundamental to causal inference in the medical and social sciences. The framework was introduced by Neyman (1923/1990) and it was further elaborated by Rubin during the 1970s (e.g., Rubin, 1978). In causal inference, the main objective is to infer causal effects, with each causal effect defined as the difference between the outcome of a subject had s/he been exposed to the active treatment, and the outcome had the same subject been exposed to the control treatment. However, causal inference often needs to be performed from an observational study instead of from a randomized experiment (e.g., Rosenbaum, 2002a), because it is not feasible to assign treatments. Thus in these situations, the treatment assignment probabilities are unknown. To obtain accurate causal inferences, these probabilities need to be accurately estimated given all relevant pre-treatment covariates, so that it becomes possible to compare the outcomes of subjects receiving different treatments.

Let us elaborate by providing a brief review of causal inference under the potential outcomes framework. For simplicity and with no loss of generality, consider the case of two treatments; the control treatment ($w=0$) and the active treatment ($w=1$). Then the framework refers to a given population of N subjects who received treatments, $(w_1,\ldots,w_N)$, vectors of observed pre-treatment covariates describing the subjects, $(\mathbf{x}_1,\ldots,\mathbf{x}_N)$, the subjects potential outcomes under the control treatment, $(y_1^{(0)},\ldots,y_N^{(0)})$, and the potential outcomes under active treatment, $(y_1^{(1)},\ldots,y_N^{(1)})$, where for every subject i, both potential outcomes $(y_i^{(0)},y_i^{(1)})$ are viewed as occurring in a common point in time. This notation for the framework is possible under stable unit-treatment values assumption (SUTVA), which is often adopted. SUTVA refers to the assumption

* Corresponding author.
E-mail address: georgek@uic.edu (G. Karabatsos).

0378-3758/$ - see front matter © 2011 Published by Elsevier B.V.
doi:10.1016/j.jspi.2011.10.013

that, for each unit i, the potential outcomes $(y_i^{(0)},y_i^{(1)})$ are unaffected by the treatments received by all the other subjects (e.g., Rubin, 1990). A causal effect refers to a comparison of potential outcomes under treatment and potential outcomes under control, on a common set of subjects. A commonly used summary of treatment effect is the average causal effect in the population, defined by the population expectation $E(y^{(1)}-y^{(0)})=E(y^{(1)})-E(y^{(0)})$. The fundamental problem facing causal inference (Rubin, 1978; Holland, 1986) is that, for each unit i, it is impossible to compare the potential outcomes $y_i^{(0)}$ and $y_i^{(1)}$ for each subject to infer causal effects, because it is only possible to expose each subject to one of the treatments, and thus, only one of the potential outcomes is observable from each subject. Also, while a sample from the population can be used to estimate $E(y^{(1)}|w=1)$ and $E(y^{(0)}|w=0)$, these expectations do not necessarily coincide with the marginal expectations $E(y^{(1)})$ and $E(y^{(0)})$ that define the average causal effect.

However, the assumption of strongly ignorable treatment assignments, and propensity scores, provide keys to solving the fundamental problem of causal inference in the context of an observational study. Rosenbaum and Rubin (1983a) proved that when treatment assignments are strongly ignorable given the observed covariates $\mathbf{x}$, then given any value of the propensity score, $e_1(\mathbf{x})=\mathrm{pr}(w=1|\mathbf{x})$ (with $e_0(\mathbf{x})=1-e_1(\mathbf{x})$), the causal effect can be defined by $E\{y^{(1)}|w=1,e_1(\mathbf{x})\}-E\{y^{(0)}|w=0,e_1(\mathbf{x})\}$, and the average causal effect can be defined by $E_{e_1(\mathbf{x})}[E\{y^{(1)}|w=1,e_1(\mathbf{x})\}-E\{y^{(0)}|w=0,e_1(\mathbf{x})\}]$, where $E_{e_1(\mathbf{x})}$ is the expectation with respect to the population distribution of $e_1(\mathbf{x})$. Treatment assignments are said to be strongly ignorable when they are probabilistic and conditionally independent of the outcomes given $\mathbf{x}$, that is, when $0<\mathrm{pr}(w=1|\mathbf{x})<1$ and $(y^{(0)},y^{(1)})\perp w|\mathbf{x}$ for all $\mathbf{x}$, and as a consequence, the joint probability of treatment assignments can in general be defined by $\prod_{i=1}^{N} e_{w_i}(\mathbf{x}_i)$ (e.g., Rubin, 2007). Rosenbaum and Rubin's (1983a) theoretical results justify the use of different types of causal models for observational studies. They include models that stratify subjects either by propensity scores estimated via logistic regression (Rosenbaum and Rubin, 1984), models that weigh observations by the inverse of estimated propensity scores (e.g., Rosenbaum, 1987; Robins et al., 2000),models based on matching pairs of subjects on their estimated propensity scores, and covariance adjustment models which regress observed outcomes on w and the estimated propensity score. Also, while it is possible to regress outcomes on both w and $\mathbf{x}$, this approach may yield paradoxical results in causal inference (Holland and Rubin, 1983; Lord, 1967). For reviews of the voluminous literature on methods of causal inference, see for example, Rosenbaum (2002a); Gelman and Meng (2004), and Rubin (2006). However, hidden bias is present in many observational studies because, for example, the investigator does not know all the important covariates. When there is hidden bias, treatment assignments are not strongly ignorable given the observed covariates $\mathbf{x}$, but instead, they are strongly ignorable given the covariates $(\mathbf{x},\mathbf{u})$, where $\mathbf{u}$ is a set of unobserved pre-treatment covariates. Among distinct pairs of subjects satisfying $\mathbf{x}_i=\mathbf{x}_j$, an observational study is free from hidden bias whenever $e_1(\mathbf{x}_i)=e_1(\mathbf{x}_j)$ holds for all such pairs, and hidden bias is present in the study when there is at least one inequality among these pairs (Rosenbaum, 2002a).

Typically, in the practice of causal modeling, a parametric assumption is made for the joint population density $f(y,w|\mathbf{x})$, and often, it is also assumed that there is no hidden bias. Both assumptions are questionable, and when either assumption is inconsistent with real data, the average causal effect estimate can be quite misleading, and its confidence (or posterior credible) interval can be falsely precise (concerning hidden bias, see for example, Drake, 1993; McCandless et al., 2007). In particular, the parametric assumption precludes the capturing a more general range of density shapes, such as heavier tail behavior and possible multi-modalities in the joint density. Also, in the past, hidden bias has often been addressed through a sensitivity analysis. This involves expanding the causal model by including a hidden bias parameter that governs the distribution of the unobserved covariates, which typically results in an unidentified model. Under the frequentist inference approach, sensitivity analysis is conducted by plugging in a range of values of the bias parameter into the causal model, and the rationale is that if the average causal effect estimate does not lead to a change in the qualitative conclusions about the effect (e.g., the conclusion of a positive average causal effect), then the number of interpretations of the observational data is reduced and the causal conclusions become more defensible (e.g., Cornfield et al., 1959; Rosenbaum and Rubin, 1983b; Lin et al., 1998; Rosenbaum, 2002a). Unfortunately, this approach faces a number of interpretability and statistical issues (e.g., Robins, 2002; Greenland, 2005). In a fully Bayesian approach to sensitivity analysis, a prior is specified on all parameters of the (nonidentifiable) causal model, including the bias parameter (e.g., Greenland, 2005; McCandless et al., 2007, 2008a, 2008b). However, because of the nonidentifiability, this prior distribution needs to accurately capture information about the unknown sampling distribution of all model parameters, in order to ensure consistency of the posterior mean estimate of average causal effect, and to ensure that the credible interval has a nominal large sample coverage probability (Gustafson, 2005; McCandless et al., 2007). This prior information can be difficult to elicit in many observational studies.

To address these open problems we introduce a Bayesian nonparametric causal model. The model makes use of a stick-breaking prior for the joint distribution, having the flexibility to capture any multi-modalities, skewness and heavier tail behavior, while accounting for hidden bias. The stick-breaking prior is a general type of nonparametric prior which includes the Dirichlet process as a special case. Using an identifiable model, we account for hidden bias by capturing heterogeneity in subjects caused by unobserved covariates, as is done in the frailty approach to survival analysis based on proportional hazards (e.g., Duchateau and Janssen, 2008).

We review the stick-breaking prior distribution in the next section, and in Section 3 we introduce our Bayesian nonparametric causal model. In that section we discuss the features and assumptions of the model, and discuss methods of posterior inference using a Gibbs sampler. Technical details are given in appendices. In Section 4 we illustrate the Bayesian nonparametric causal model through the analysis of small and large observational data sets. With Section 5 we conclude with a proof of the posterior consistency of our Bayesian nonparametric causal model.

Please cite this article as: Karabatsos, G., Walker, S.G., A Bayesian nonparametric causal model. Journal of Statistical Planning and Inference (2011), doi:10.1016/j.jspi.2011.10.013

2. Stick breaking prior

The stick-breaking prior is a general type of Bayesian nonparametric prior that can be specified to support the entire space of random distributions G. Under such a prior, denoted by SB($\mathbf{a},\mathbf{b},H$), a random distribution G is a stochastic process constructed as follows (e.g., Müller and Quintana, 2004). Let $\boldsymbol{\theta}_j \sim H$ independently and $v_j \sim \text{Beta}(a_j,b_j)$ independently for $j=1,2,\ldots$, where H is a distribution with density h defined with respect to Lebesgue measure, with $\mathbf{a}=(a_1,a_2,\ldots,a_j,\ldots)'$ and $\mathbf{b}=(b_1,b_2,\ldots,b_j,\ldots)'$ vectors of positive-valued parameters. Then a random distribution is constructed by $G(\cdot)=\sum_{j=1}^{\infty}\omega_j\delta_{\theta_j}(\cdot)$, where $w_1=v_1$ and $\omega_j=v_j\prod_{k<j}(1-v_k)$ for $j>1$, where $H(\cdot)=\text{E}\{G(\cdot)\}$ defines the prior mean, and $\delta_{\boldsymbol{\theta}}$ denotes the degenerate distribution with point-mass at $\boldsymbol{\theta}$.

This construction is called a "stick-breaking" procedure because at each stage j, we randomly break what is left of a stick of unit length, and assign the length of this break to the current ω_j value. Also, it is obvious from this construction that a stick-breaking prior supports discrete distributions with probability 1. The stick-breaking prior includes other important nonparametric priors as special cases. For example, the two-parameter Poisson–Dirichlet process (Pitman, 1996) is obtained by taking $a_j=1-a$ and $b_j=b+ja$ for $0\le a<1$ and $b>-a$. As another example, taking $a=0$ and $b=\alpha$ results in the Dirichlet process (Sethuraman, 1994), where α is a precision parameter representing the prior degree of belief in H.

3. Bayesian nonparametric causal model

For a sample set of data $\{(w_i,\mathbf{x}_i,y_i)\}_{i=1}^n$, and binary treatments ($w=0,1$), the Bayesian nonparametric causal model is a nonparametric mixture model for the joint density of potential outcomes and treatment assignments. For $i=1,\ldots,n$, this model is defined by

$$f(y_i,w_i|\mathbf{x}_i,G)=\int f(y_i|\beta_0+\beta_T w_i+\beta e_1(\mathbf{x}_i;\lambda_0,\boldsymbol{\lambda}),\sigma^2)e_{w_i}(\mathbf{x}_i;\lambda_0,\boldsymbol{\lambda})\,\text{d}G(\beta_0,\beta_T,\lambda_0)$$

$$G\sim \text{SB}(\mathbf{a},\mathbf{b},H=\text{Normal}(\boldsymbol{\mu},\mathbf{T})).$$

In words, the joint density of potential outcomes and treatment assignments, $f(y,w|\mathbf{x},G)$, is modeled by conditional densities $f(y|\eta,\sigma^2)$ and $e_1(\mathbf{x};\lambda_0,\boldsymbol{\lambda})$, where $f(y|\eta,\sigma^2)$ has mean and variance (η,σ^2), and is defined with respect to Lebesgue measure or counting measure (as appropriate), and $e_1(\mathbf{x};\lambda_0,\boldsymbol{\lambda})$ is the propensity score defined by a logit model $e_1(\mathbf{x};\lambda_0,\boldsymbol{\lambda})=[1+\exp\{-(\lambda_0+\boldsymbol{\lambda}'\mathbf{x})\}]^{-1}$. Adopting the standard framework for generalized linear models (GLMs; McCullagh and Nelder, 1989), the $f(y|\eta_i,\sigma^2)$ $(i=1,\ldots,n)$ can be specified by either normal $(y|\eta_i,\sigma^2)$, poisson$(y|\exp\eta_i)$, or binomial$(y|n_i,\{1+\exp(-\eta_i)\}^{-1})$ densities. Specifically, $f(y|\eta,\sigma^2)$ depends on a random intercept (β_0), a random causal effect (β_T), and on covariance adjustment on the propensity score via the term $\beta e_1(\mathbf{x};\lambda_0,\boldsymbol{\lambda})$,while the propensity score $e_1(\mathbf{x};\lambda_0,\boldsymbol{\lambda})$ depends on a random intercept (λ_0). The mixing distribution G of the random parameters $(\beta_0,\beta_T,\lambda_0)$ is modeled nonparametrically by a stick-breaking prior with parameters $(\mathbf{a},\mathbf{b},H)$, and we specify H (the prior mean of G) as the trivariate normal distribution with mean and covariance parameters $(\boldsymbol{\mu},\mathbf{T})$, with $\boldsymbol{\mu}=(\mu_{\beta_0},\mu_{\beta_T},\mu_{\lambda_0})$. Also, to facilitate working with the nonparametric mixture model, we introduce latent $(\beta_{0i},\beta_{Ti},\lambda_{0i})$ associated with (y_i,w_i), and then the joint density of the (y_i,w_i) given $(\beta_{0i},\beta_{Ti},\lambda_{0i})$ can be written as

$$\prod_{i=1}^{n} f(y_i|\beta_{0i}+\beta_{Ti}w_i+\beta e_1(\mathbf{x}_i;\lambda_{0i},\boldsymbol{\lambda}),\sigma^2)e_{w_i}(\mathbf{x}_i;\lambda_{0i},\boldsymbol{\lambda}),$$

with $(\beta_{0i},\beta_{Ti},\lambda_{0i})|G\overset{iid}{\sim}G$ and $G\sim\text{SB}(\mathbf{a},\mathbf{b},H=\text{Normal}(\boldsymbol{\mu},\mathbf{T}))$. After marginalizing over $(\beta_{0i},\beta_{Ti},\lambda_{0i})$, the resulting joint density of the (y_i,w_i) is $\prod_{i=1}^n f(y_i,w_i|\mathbf{x}_i,G)$. To complete the Bayesian model specification, we specify a Normal (μ_p,σ_p^2) prior on β, a p-variate Normal $(\mu_{\boldsymbol{\lambda}},\Sigma_{\boldsymbol{\lambda}})$ prior on $\boldsymbol{\lambda}$, a trivariate Normal $(\boldsymbol{\mu}_0,\mathbf{T}_0)$ hyperprior on $\boldsymbol{\mu}$, an InverseWishart $(v_0,\mathbf{R}_0)$ hyperprior on $\mathbf{T}$, and when the error variance is treated as an unknown, a Gamma $(a_1/2,a_2/2)$ prior for σ^{-2}. Finally, using standard arguments of probability theory regarding Bayes' theorem, a sample set of data $\{(w_i,\mathbf{x}_i,y_i)\}_{i=1}^n$ updates the joint prior distribution to yield a posterior distribution of $(\beta_0,\beta_T,G,\beta,\sigma^2,\lambda_0,\boldsymbol{\lambda},\boldsymbol{\mu},\mathbf{T})$ given the data.

By flexible nonparametric modeling of G, the mixing distribution of the random $(\beta_0,\beta_T,\lambda_0)$, the model provides an approach to causal inference, which can capture multi-modalities, skewness and heavier tail behavior in the joint population distribution of potential outcomes and treatment assignments, while accounting for hidden bias. As cogently argued by Rubin (1985, p. 469), for accurate causal inference, not only is it important to accurately model the treatment assignments (via propensity scores), but it is also important to accurately model the outcomes. While in our model we assume that the pre-treatment covariates $(\mathbf{x}_i)$ only influence the outcomes (y_i) outcome via the propensity score, as is done in McCandless et al. (2009), we provide a flexible model for the outcomes through nonparametric mixture modeling of the distribution of random intercepts and causal effects (β_{0i},β_{Ti}). (Though, in principle, the outcomes can be modeled to also depend directly on covariates.) Also, our causal model assumes independence between observed pre-treatment covariates $(\mathbf{x}_i)$ and random intercepts (λ_{0i}). This is a similar modeling technique and assumption used for the frailty approach to survival models based on proportional hazards (e.g., Duchateau and Janssen, 2008). Importantly, by modeling the random intercepts λ_{0i} in our causal model, we account for hidden bias by capturing heterogeneity in subjects caused by unobserved covariates.

Please cite this article as: Karabatsos, G., Walker, S.G., A Bayesian nonparametric causal model. Journal of Statistical Planning and Inference (2011), doi:10.1016/j.jspi.2011.10.013

Of course, treatments may be multi-valued in a given observational study. That is, there may be more than two treatments, $w=0,1,2,\ldots,W$, including one control treatment ($w=0$), or possibly more. It turns out that the Bayesian nonparametric causal model can be easily extended for multi-valued treatments, using Lechner's (2001) theoretical results which build on Rosenbaum and Rubin's (1983a) proofs for strongly ignorable binary treatments. We provide details in Appendix A. Interestingly, multi-valued treatments can also be implemented to weaken the SUTVA, in order to model situations where the outcome of a given subject is influenced by the treatments received by other subjects (see Hong and Raudenbush, 2006).

Also, a Gibbs sampler was developed to sample the full posterior distribution of the parameters of our Bayesian nonparametric causal model, by iterative sampling of full conditional posterior distributions. Appendix B contains all the technical details of the Gibbs sampler, where it is shown that latent variable methods are used to sample from many of these full conditionals. A MATLAB program was written to perform the Gibbs sampling, and it can be obtained through correspondence with the first author.

4. Illustrations

In the following subsections, we illustrate the Bayesian nonparametric causal model on a small and large data set, respectively. In each application we specified noninformative priors for most of the parameters of the model. We specified a Normal(0,10) prior for β, a trivariate Normal ($\mathbf{0}$,diag(10)) prior for $\boldsymbol{\lambda}$ and for $\boldsymbol{\mu}$, an InverseWishart (1,diag(1)) hyper–prior on $\mathbf{T}$; a stick-breaking prior with parameters $a_j=1-a$ and $b_j=b+ja$, with $a=.25$ and $b=1$, and a Gamma (1, 1) prior for σ^{-2} (where appropriate). Also, in each application, we applied the Gibbs sampler to generate 15,000 samples. The last 10,000 samples were used for posterior inference, as plots indicated that the Gibbs samples converged to samples from the target posterior distribution before 5000th Gibbs sample.

4.1. DNA data

In an observational study, Zhao et al. (2000) conducted an investigation to determine whether occupational exposure to the chemical 1,3-butadiene causes a specific alteration of the human DNA adduct *N*-1-(2,3,4-trihydroxybutyl)-adenine (*N*-1-THB-Ade). This chemical is used to produce a variety of polymers. Zhao et al. (2000) presented data from a chemical operation in the Czech Republic, to compare measurements of *N*-1-THB-Ade in 15 males who worked with the chemical, against those measurements of 11 male controls who worked in the heat production unit. Obviously, such a comparison needs to be made in the context of an observational study, instead of in a randomized experiment. Given the potential danger of this chemical, it would be unethical to randomly assign each subject into one of the treatment conditions. The full data set is presented in Table 1, and for each subject, consists of a measurement of *N*-1-THB-Ade in units of adducts per 10^9 nucleotides, along with observations of three pre-treatment covariates: age, indicator of smoker, and number of cigarettes smoked per day (CigDay). The 1,3-butadiene chemical is found in cigarette smoke, and more than half of the exposed workers were smokers.

Zhao et al. (2000) analyzed the data using the ordinary Wilcoxon's rank sum hypothesis test of zero causal effect, while ignoring the pretreatment covariates. However, this test assumes that the subjects were randomly assigned to treatments, which is very questionable. In response, Rosenbaum (2002b) analyzed the data using Wilcoxon's rank sum hypothesis test after adjusting for the propensity scores, with these scores estimated as a function of the three covariates. According to both of these hypothesis tests, it was concluded that there were significantly higher levels of *N*-1-THB-Ade among exposed workers compared to controls, indicating a positive causal effect. However, both tests assumed no hidden bias.

We analyzed the data set with the Bayesian nonparametric causal model, which provides a flexible model for the outcomes and treatment assignments, while accounting for hidden bias. As done in Rosenbaum (2002b), we analyze the natural log of the *N*-1-THB-Ade outcome measurements, and furthermore, we specify a normal model for these measurements. Also the treatment assignments were assume to depend on the three pre-treatment covariates and unobserved covariates. Table 2 presents summaries of the posterior distributions of all model parameters, including a summary of the posterior predictive distribution of the random vector $(\beta_0,\beta_T,\lambda_0)$ based on the mixing distribution G modeled by the (nonparametric) stick-breaking prior. The posterior mean of the average causal effect is 1.03 with 95% credible interval $(-1.49,3.30)$, so there does not seem to be a significant effect of the 1,3-butadiene treatment. Also, the posterior mean estimate and 95% posterior credible interval of the causal effect for each subject is presented in the last column of Table 1, and we see that for each subject the interval contains zero. Table 2 shows that age was the observed covariate having the most influence on the treatment assignments, with its coefficient having posterior mean .67. Furthermore, the posterior mean of the variance parameter τ_{λ_0} was 6.81, indicating the presence of hidden bias, that is, heterogeneity in subject-level intercepts (λ_{0i}) as a result of unobserved covariates. Finally, among the 26 subjects, the number of distinct parameter vectors $(\beta_0,\beta_T,\lambda_0)$ typically ranged from 1 to 7, according to the 95% credible interval of the number in the posterior distribution.

So it is concluded that exposure to 1,3-butadiene does not produce a significant causal effect, in contrast to the previous findings of Rosenbaum (2002b) and Zhao et al. (2000). This difference, at least in part, can be explained by the fact that the Bayesian nonparametric causal model fully accounts for uncertainty due to hidden bias and random error, unlike the Wilcoxon tests.

Please cite this article as: Karabatsos, G., Walker, S.G., A Bayesian nonparametric causal model. Journal of Statistical Planning and Inference (2011), doi:10.1016/j.jspi.2011.10.013

Table 1
Data from Zhao et al. (2000), on human DNA adducts for workers exposed to 1,3-butadiene and controls (*N*-1-THB-Ade is measured in adducts per 109 nucleotides). The last column displays posterior estimate of the causal effect, for each subject.

Treatment group	Age	Smoker	Cigarettes per day	*N*-1-THB-Ade	Posterior mean causal effect (95%)
Exposed	57	Y	15	0.3	1.03 (−1.35,3.15)
	50	Y	20	0.5	1.03 (−1.37,3.24)
	28	Y	15	1.0	1.03 (−1.35,3.14)
	59	Y	40	0.8	1.05 (−1.33,3.21)
	23	Y	20	1.0	1.03 (−1.42,3.14)
	49	Y	15	12.5	1.07 (−1.35,3.37)
	49	Y	2	0.3	1.04 (−1.35,3.29)
	24	Y	5	4.3	1.07 (−1.30,3.37)
	45	N	0	1.5	1.06 (−1.33,3.13)
	48	N	0	0.1	1.02 (−1.37,3.18)
	38	N	0	0.3	1.03 (−1.38,3.14)
	44	N	0	18.0	1.08 (−1.34,3.36)
	43	N	0	25.0	1.08 (−1.33,3.38)
	44	N	0	0.3	1.02 (−1.35,3.20)
	57	N	0	1.3	1.04 (−1.37,3.23)
Control	36	Y	10	0.1	1.06 (−1.33,3.29)
	20	Y	20	0.1	1.04 (−1.37,3.33)
	31	Y	10	2.3	1.03 (−1.41,3.22)
	50	Y	25	3.5	0.99 (−1.55,3.30)
	31	N	0	0.1	1.06 (−1.34,3.30)
	54	N	0	0.1	1.05 (−1.31,3.29)
	54	N	0	1.8	1.03 (−1.38,3.28)
	55	N	0	0.5	1.02 (−1.35,3.20)
	44	N	0	0.1	1.06 (−1.33,3.30)
	49	N	0	0.2	1.03 (−1.33,3.31)
	51	N	0	0.1	1.06 (−1.30,3.31)

Table 2
Posterior estimates of model parameters, for the DNA data.

Parameter	Posterior mean	95%	Parameter	Posterior mean	95%
β	.77	(−3.27,4.67)	μ_{β_T}	.97	(−2.16,3.65)
λ_{smoker}	.01	(−0.04,0.05)	μ_{β_0}	−1.16	(−4.17,2.38)
λ_{age}	.67	(−.32,1.55)	μ_{λ_0}	−.45	(−3.17,2.73)
λ_{CigDay}	.00	(−.03,.04)	τ_{β_T}	17.25	(.16,33.34)
σ^2	3.36	(1.51,9.86)	τ_{β_0}	11.22	(.16,29.31)
			τ_{λ_0}	6.81	(.15,32.63)
β_T	1.03	(−1.49,3.30)	τ_{β_T,β_0}	−8.87	(−11.15,7.64)
β_0	−1.26	(−4.21,2.18)	τ_{β_T,λ_0}	1.30	(−9.80,9.79)
λ_0	−.49	(−2.60,2.24)	τ_{β_0,λ_0}	−1.03	(−9.25,8.62)

4.2. Chicago public school data

We illustrate the Bayesian nonparametric causal model on a large observational data set of interest in educational research. Since the enactment of the No Child Left Behind (NCLB) Act in school year 2002–2003, public schools are held accountable for ensuring that all schoolchildren attain at least minimum proficiency on state-level examinations in reading, math, and science. Here, we analyze available data of 208,286 students from 576 Chicago public schools, who were either in grades 3, 4, 5, 6, 7, 8, or 11 during school year 2006–2007. Treating the 576 schools as the "subjects", the objective of this analysis is to infer the causal effect of a low income neighborhood on the percentage of students in the school with exam scores attaining at least the minimum proficiency in the three subject areas. A school is said to be in a low income neighborhood (the active treatment) when at least 60% of its students are from low-income families, otherwise, the school does not belong in a low income neighborhood (the control treatment). We modeled the percentage outcomes by a binomial distribution for the number of students in a school attaining minimum proficiency out of a total number of students taking the examination. Also, we modeled treatment assignments by nine school-level covariates: an indicator of whether the school has at least 95% highly-qualified teachers (HQT), an indicator of high school (HS), an indicator of elementary school (ELEM) (with zeroes for both HS and ELEM indicating a charter school), an indicator of the school having high minority status (HIMIN=1 if the school has at least 60% minority students, 0 otherwise), the log

Please cite this article as: Karabatsos, G., Walker, S.G., A Bayesian nonparametric causal model. Journal of Statistical Planning and Inference (2011), doi:10.1016/j.jspi.2011.10.013

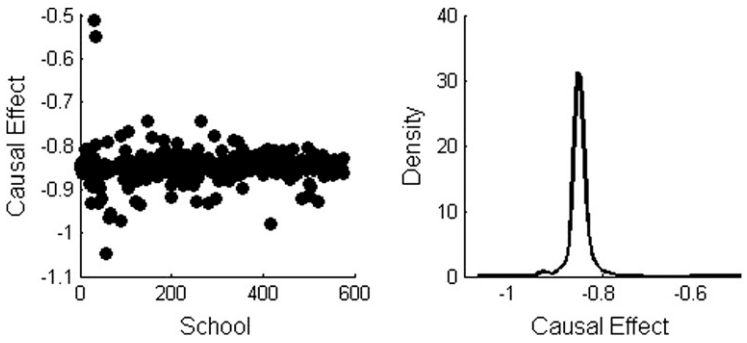

Fig. 1. Left panel: posterior mean estimate of the causal effect, for each of the 576 schools. Right panel: the posterior mean estimate of the density of school causal effects.

Table 3
Posterior estimates of model parameters, for the Chicago public school data.

Parameter	Posterior mean	95%	Parameter	Posterior mean	95%
β	3.42	(3.29,3.55)	β_T	−.85	(−.97,−.73)
λ_{HQT}	.04	(.03,.13)	β_0	−1.33	(−1.51,−1.13)
λ_{HS}	−7.43	(−7.81,−6.98)	λ_0	.45	(.13,.84)
λ_{ELEM}	−6.10	(−6.37,−5.74)	μ_{β_T}	−.85	(−.89,−.80)
λ_{HIMIN}	−1.23	(−1.34,−1.14)	μ_{β_0}	−1.33	(−1.49,−1.19)
λ_{logSIZE}	−.07	(−.09,−.06)	μ_{λ_0}	.46	(.19,.81)
λ_{ATT}	9.77	(9.56,10.06)	τ_{β_T}	.02	(.01,.05)
λ_{MOB}	−3.58	(−3.77,−3.41)	τ_{β_0}	.02	(.01,.05)
λ_{TRU}	−.78	(−.91,−.65)	τ_{λ_0}	.03	(.01,.05)
λ_{PAR}	.31	(.28,.35)	τ_{β_T,β_0}	−.006	(−.02,.002)
			τ_{β_T,λ_0}	−.003	(−.015,.01)
			τ_{β_0,λ_0}	−.002	(−.014,.01)

number of students in the school (logSIZE), the attendance rate of the school (ATT), enrollment change of the school from beginning to the end of the school year (MOB), the school truancy rate (TRU), and the rate of parental involvement (PAR) measured by the proportion of students with parents who regularly communicated with the school's teachers.

According to samples from the posterior predictive distribution of G, the average population causal effect is given by −.85, with 95% posterior credible interval of (−.97,−.73). Thus, overall, there is a significant negative effect of the low income neighborhood treatment on student achievement, compared to the non-low income neighborhood (control) treatment. The left panel of Fig. 1 presents the posterior mean estimates of the low income neighborhood causal effect for each of the 576 Chicago public schools. We see that the causal effects vary from one school to another, and while most of the schools have a causal effect around −.85, a couple of schools have a causal effect of −.5, and five schools have a causal effect of around −1. Not presented in the figure is our finding that, for each school, the 95% posterior credible interval of the causal effect is different from zero. The right panel of Fig. 1 presents the posterior mean estimate of the population density of causal effects. A close inspection reveals that this density is multimodal and somewhat skewed.

Table 3 presents the posterior summaries of all the model parameters, including a summary of the posterior predictive distribution of the random vector $(\beta_0,\beta_T,\lambda_0)$. From the table, we see, for instance, that the most influential covariates for the propensity score are attendance rate, high-school status, and elementary school status. Also, among the 576 schools, the number of distinct (latent) parameter vectors $(\beta_{0i},\beta_{Ti},\lambda_{0i})$ ranged from 36 to 187, according to the 95% credible interval of the number in the posterior distribution. Finally, the posterior mean of the variance parameter τ_{λ_0} was .03, indicating the presence of some hidden bias, as manifested by the heterogeneity of school-level intercepts (λ_{0i}) caused by unobserved covariates.

5. Consistency

Here we discuss Bayesian consistency for the model

$$f(y,w|\mathbf{x},\phi,G) = \int f(y,w|\mathbf{x},\phi,\theta)\, \mathrm{d}G(\theta)$$

and we will write $z=(w,y)$ and $\xi=(\phi,G)$. There are prior distributions assigned to ϕ and G; specifically G will have a Dirichlet process prior. We will first discuss weak consistency using ideas from Schwartz (1965)and Walker (2003).

Please cite this article as: Karabatsos, G., Walker, S.G., A Bayesian nonparametric causal model. Journal of Statistical Planning and Inference (2011), doi:10.1016/j.jspi.2011.10.013

So let us assume that ξ_0 is the data generating parameter and let

$$A_\epsilon = \left\{ \xi : \left| \iint g_j(z,\mathbf{x}) m(\mathrm{d}\mathbf{x}) \{ f(\mathrm{d}z|\mathbf{x},\xi) - f(\mathrm{d}z|\mathbf{x},\xi_0) \} \right| < \epsilon,\ j = 1,\ldots,M \right\},$$

for bounded continuous functions g_j, be a weak neighborhood of ξ_0. Here m is taken to be the distribution of the $\mathbf{x}_i$. We will use Π to denote the prior on ξ and assume that

$$\Pi\left(\xi : d(\xi,\xi_0) = \int m(\mathrm{d}\mathbf{x}) d_K\{ f(\cdot|\mathbf{x},\xi), f(\cdot|\mathbf{x},\xi_0) \} < \delta \right) > 0,$$

for all $\delta > 0$, where d_K denotes the Kullback–Leibler divergence between probability density functions. This condition means that ξ_0 belongs to the Kullback–Leibler support of the prior.

Now let

$$I_n = \int R_n(\xi) \Pi\,(\mathrm{d}\xi),$$

where

$$R_n(\xi) = \prod_{i=1}^{n} f(z_i|\mathbf{x}_i,\xi)/f(z_i|\mathbf{x}_i,\xi_0),$$

which can be written as

$$I_n = \int \exp\,[-n\{ d_n(\xi,\xi_0) \}] \Pi\,(\mathrm{d}\xi),$$

where

$$d_n(\xi,\xi_0) = n^{-1} \sum_{i=1}^{n} \log\,\{ f(z_i|\mathbf{x}_i,\xi_0)/f(z_i|\mathbf{x}_i,\xi) \}.$$

So, for some $c > 0$,

$$\limsup_n e^{nc} I_n > \int_{d_K(\xi,\xi_0) < \rho} \left\{ \limsup_n \exp\,[n\{c - d_n(\xi,\xi_0)\}] \right\} \Pi\,(\mathrm{d}\xi).$$

Now $d_n(\xi,\xi_0) \to d(\xi,\xi_0)$ a.s. and so $\limsup_{\{n: d_K(\xi,\xi_0) < \rho\}} \exp\,[n\{c - d_n(\xi,\xi_0)\}] = +\infty$ for all $c > \rho$. This is a standard argument (Schwartz, 1965) and hence $I_n > e^{-nc}$ a.s. for all large n for any $c > 0$. This follows since we can make ρ, and hence c, as small as we wish.

We should now consider the numerator for $\Pi_n(A_\epsilon^c) = \Pi(A_\epsilon^c | (z_i,\mathbf{x}_i)_{i=1}^n)$ in our bid to show that this posterior probability goes to 0 a.s. for all $\epsilon > 0$. Hence, consider

$$I_{n\epsilon} = \int_{A_\epsilon^c} R_n(\xi) \Pi\,(\mathrm{d}\xi)$$

and note that

$$A_\epsilon^c = \left(\bigcup_{j=1}^{M} A_j^+ \right) \bigcup \left(\bigcup_{j=1}^{M} A_j^- \right),$$

where

$$A_j^+ = \left\{ \xi : \int m(\mathrm{d}\mathbf{x}) \left[\int g_j(z,\mathbf{x}) \{ f(\mathrm{d}z|\mathbf{x},\xi) - f(\mathrm{d}z|\mathbf{x},\xi_0) \} \right] > \epsilon \right\},$$

$$A_j^- = \left\{ \xi : \int m(\mathrm{d}\mathbf{x}) \left[\int g_j(z,\mathbf{x}) \{ f(\mathrm{d}z|\mathbf{x},\xi) - f(\mathrm{d}z|\mathbf{x},\xi_0) \} \right] < -\epsilon \right\}.$$

Following Walker (2003) we introduce the "predictive" density

$$f_{nj+}(z,\mathbf{x}) = \int f(z,\mathbf{x}|\xi) \Pi_{n\,A_j^+}\,(\mathrm{d}\xi),$$

where $\Pi_{n,A}$ is the posterior restricted and normalized to the set A. A similar definition also applies to f_{nj-}. It is immediately obvious that, for example, $f_{nj+} \in A_j^+$ and hence if $f_{n\cdot\cdot}$ is not in a weak neighborhood of $f(z,\mathbf{x}|\xi_0)$ then it is neither in an L_1, and hence Hellinger, neighborhood of $f(z,\mathbf{x}|\xi_0)$. Therefore, for all n, $d_H\{ f(\cdot|\xi_0), f_{nj\cdot}(\cdot) \} > \gamma$ for some $\gamma > 0$, where d_H denotes the Hellinger distance between probability density functions.

It is now proven in Walker (2003, Theorem 1) that $\Pi_n(A_j^-) \to 0$ a.s. for every $j = 1,\ldots,M$, and hence $\Pi_n(A_\epsilon^c) \to 0$ a.s. Hence we have established weak consistency.

Please cite this article as: Karabatsos, G., Walker, S.G., A Bayesian nonparametric causal model. Journal of Statistical Planning and Inference (2011), doi:10.1016/j.jspi.2011.10.013

For stronger forms of consistency, such as for sets of the type

$$A_\epsilon = \left\{ \xi : d(\xi,\xi_0) = \int m(\mathrm{d}\mathbf{x}) d_H\{f(\cdot|\mathbf{x},\xi), f(\cdot|\mathbf{x},\xi_0)\} < \epsilon \right\},$$

further assumptions need to be made. The denominator for $\Pi_n(A_\epsilon^c)$ is dealt with in the same way as for weak consistency, but the numerator needs more attention.

Following Walker et al. (2005, Section 3.2) we can imply strong consistency from weak consistency when $f(z|\mathbf{x},\phi,\theta)$ is bounded and continuous in θ for all $(z,\mathbf{x},\phi)$. While this covers the Poisson and Binomial cases, it does not cover the Normal case, except when the variance is bounded away from 0, which is a recourse to guarantee strong consistency.

Weak consistency for the posterior distribution of G and ϕ, assuming the model is correct, holds. The term "model is correct" refers to the fact that ξ_0 belongs to the Kullback–Leibler support of the prior. This does not of course imply anything stronger about estimating consistently any particular individual parameters.

Acknowledgments

We thank Editor Narayanaswamy Balakrishnan and anonymous referees for comments that helped improve the presentation of manuscript.

Appendix A. The Bayesian nonparametric causal model for multi-valued treatments

Lechner (2001) proved that when treatment assignments are strongly ignorable given $\mathbf{x}$, then for any given value of the vector of propensity scores, $\mathbf{e}(\mathbf{x}) = (e_1(\mathbf{x}), \ldots, e_W(\mathbf{x}))'$ (where $0 < e_w(\mathbf{x}) = \mathrm{pr}(w|\mathbf{x}) < 1$, $w = 0, 1, 2, \ldots, W$, and $e_0(\mathbf{x}) = 1 - \sum_{w=1}^{W} e_w(\mathbf{x})$), the causal effect can be defined by $\mathrm{E}\{y^{(j)}|w=j,\mathbf{e}(\mathbf{x})\} - \mathrm{E}\{y^{(k)}|w=k,\mathbf{e}(\mathbf{x})\}$, and the average causal effect can be defined by $\mathrm{E}_{\mathbf{e}(\mathbf{x})}[\mathrm{E}\{y^{(j)}|w=j,\mathbf{e}(\mathbf{x})\} - \mathrm{E}\{y^{(k)}|w=k,\mathbf{e}(\mathbf{x})\}]$, for every distinct pair of treatments $j,k \in \{0, 1, \ldots, W\}$. Using these results, the Bayesian nonparametric causal model for binary treatments can be extended to multi-valued treatments by replacing $\beta_T w$ with $\boldsymbol{\beta}_T' \mathbf{w}$, replacing $\beta e_1(\lambda_0 + \boldsymbol{\lambda}'\mathbf{x})$ with $\boldsymbol{\beta}' \mathbf{e}(\mathbf{x}; \boldsymbol{\lambda}_0, \boldsymbol{\lambda})$, where $\mathbf{w}$ is a W-dimensional 0–1 column vector indicating the treatment received by a subject (with the control treatment indicated by a vector of zeroes), $\boldsymbol{\beta}_T$ and $\boldsymbol{\beta}$ are each W-dimensional column vectors, and the vector of propensity scores $\mathbf{e}(\mathbf{x}_i; \boldsymbol{\lambda}_{0i}, \boldsymbol{\lambda})$ is specified by a mixed multinomial logit model

$$e_w(\mathbf{x}_i; \boldsymbol{\lambda}_{0i}, \boldsymbol{\lambda}) = \frac{\exp(\lambda_{0wi} + \boldsymbol{\lambda}'\mathbf{x}_i)}{1 + \sum_{t=1}^{W} \exp(\lambda_{0ti} + \boldsymbol{\lambda}'\mathbf{x}_i)}, \quad w = 1, \ldots, W,$$

where the random vector $\boldsymbol{\lambda}_{0i} = (\lambda_{0ti} | t = 1, \ldots, W)'$ provides a way to account for hidden bias. Obviously, for two treatments (w=0,1), this multinomial regression model reduces to the logit model for binary treatments. Then to complete the Bayesian model specification, a prior distribution is specified for $(\beta, \sigma^2, \boldsymbol{\lambda}, G, \boldsymbol{\mu}, \mathbf{T})$, where G is the random distribution of $(\beta_0, \boldsymbol{\beta}_T, \boldsymbol{\lambda}_0)$.

Appendix B. Gibbs sampling methods

The Gibbs sampling algorithm for the random distribution G of the latent parameters $(\beta_{0i}, \beta_{Ti}, \lambda_{0i})$ $(i = 1, \ldots, n)$ relies on the introduction of strategic latent variables, as first described by Walker (2007) and Kalli et al. (2010). For the case of binary treatments, the starting form of the model is

$$f_G(y,w|\mathbf{x}) = \int f(y,w|\beta_0,\beta_T,\beta,\sigma^2,\lambda_0,\boldsymbol{\lambda},\mathbf{x})\, \mathrm{d}G(\beta_0,\beta_T,\lambda_0),$$

with $f(y,w|\beta_0,\beta_T,\beta,\sigma^2,\lambda_0,\boldsymbol{\lambda},\mathbf{x}) = f(y|\beta_0 + \beta_T w + \beta e(\mathbf{x};\lambda_0,\boldsymbol{\lambda}),\sigma^2) e_w(\mathbf{x}_i;\lambda_0,\boldsymbol{\lambda})$,

where G is a stick-breaking process so that

$$f_G(y,w|\mathbf{x}) = \sum_{j=1}^{\infty} \omega_j f(y,w|\beta_{0j},\beta_{Tj},\beta,\sigma^2,\lambda_{0j},\boldsymbol{\lambda},\mathbf{x}).$$

We first introduce the latent variable u such that

$$f_G(y,w,u|\mathbf{x}) = \sum_{j=1}^{\infty} \mathbf{1}(u < \omega_j) f(y,w|\beta_{0j},\beta_{Tj},\beta,\sigma^2,\lambda_{0j},\boldsymbol{\lambda},\mathbf{x})$$

and the importance here is that the number of ω_j greater than u is finite. We now introduce another latent variable k which picks out the component from which y comes, with

$$f_G(y,w,u,k|\mathbf{x}) = \mathbf{1}(u < \omega_k) f(y,w|\beta_{0k},\beta_{Tk},\beta,\sigma^2,\lambda_{0k},\boldsymbol{\lambda},\mathbf{x}),$$

with $\mathbf{1}(\cdot)$ the indicator function. Hence the likelihood function based on n observations is given by

$$\prod_{i=1}^{n} \mathbf{1}(u_i < \omega_{k_i}) f(y_i,w_i|\beta_{0k_i},\beta_{Tk_i},\beta,\sigma^2,\lambda_{0k_i},\boldsymbol{\lambda},\mathbf{x}_i).$$

Please cite this article as: Karabatsos, G., Walker, S.G., A Bayesian nonparametric causal model. Journal of Statistical Planning and Inference (2011), doi:10.1016/j.jspi.2011.10.013

The Gibbs sampler can now be described. Starting with the k_i, we sample (v,u) by sampling (v) then $(u|v)$. Hence, the conditional for v_j; for $j=1,\ldots,M$, where $M=\max_i d_i$, is beta (a_j+n_j,b_j+m_j), where $n_j=\sum_i \mathbf{1}(d_i=j)$ and $m_j=\sum_i \mathbf{1}(d_i>j)$. We then sample $u_i\sim \text{Uniform}(0,\omega_{k_i})$. Before discussing sampling the $(\beta_{0j},\beta_{Tj},\lambda_{0j})$ and how many more of the v_j we need to sample, let us consider the sampling of the d_i. Now we have $\text{pr}(d_i=j)\propto \mathbf{1}(\omega_j>u_i)f(y_i,w_i|\beta_{0j},\beta_{Tj},\beta,\sigma^2,\lambda_{0j},\boldsymbol{\lambda},\mathbf{x}_i)$. Hence we need to sample as many v_js until we are sure we have all the ω_j which are greater than u_i. We will know we have all of these when we have N_i such that $\sum_{j=1}^{N_i}\omega_j>1-u_i$. Hence, we keep sampling $v_{M+1},\ldots,V_N$ and also sample $(\beta_{01},\beta_{T1},\lambda_{01}),\ldots,(\beta_{0N},\beta_{TN},\lambda_{0N})$, where $N=\max_i N_i$. For $j<M+1$ we have

$$(\beta_{0j},\beta_{Tj},\lambda_{0j})\propto \prod_{d_i=j} f(y_i,w_i|\beta_{0j},\beta_{Tj},\beta,\sigma^2,\lambda_{0j},\boldsymbol{\lambda},\mathbf{x}_i)h(\beta_{0j},\beta_{Tj},\lambda_{0j})$$

and this can be sampled via a random-walk Metropolis–Hastings algorithm. For $j>M$, $(\beta_{0j},\beta_{Tj},\lambda_{0j})\propto h(\beta_{0j},\beta_{Tj},\lambda_{0j})$ which can be sampled directly. The new M will be less than the old N and hence the Metropolis–Hastings algorithm can always be implemented. Also, in order to obtain samples from the predictive distribution of $(\beta_0,\beta_T,\lambda_0)$, we can, at each iteration of the Gibbs sampler, sample from the random G. In particular, we sample a v from the uniform distribution on $(0,1)$ and take $(\beta_0,\beta_T,\lambda_0)$ to be $(\beta_{0j},\beta_{Tj},\lambda_{0j})$ if $\sum_{l=1}^{j-1}\omega_l<v<\sum_{l=1}^{j}\omega_l$, otherwise if $v>\sum_{l=1}^{N}\omega_l$ we take $(\beta_0,\beta_T,\lambda_0)$ to be distributed as $h(\beta_0,\beta_T,\lambda_0)$.

We now explain how we sample from the full conditional posterior distributions of the remaining parameters, $(\beta,\sigma^2,\boldsymbol{\lambda},\boldsymbol{\mu},\mathbf{T})$. We Gibbs sample $(\beta,\boldsymbol{\lambda})$ using the approach by Damien et al. (1999). Writing the likelihood of each observation as $L(y_i,w_i|\beta,\boldsymbol{\lambda},rest)=f(y_i,w_i|\beta_{0k_i},\beta_{Tk_i},\beta,\sigma^2,\lambda_{0k_i},\boldsymbol{\lambda},\mathbf{x}_i)$, the full conditional posterior density of β is proportional to normal $(\beta|\mu_P,\sigma_P^2)\mathbf{1}(\beta\in A_\beta)$, where $A_\beta=\{\beta: L(y_i,w_i|\beta,\boldsymbol{\lambda},rest)>u_i,i=1,\ldots,n\}$, where each u_i is a Uniform $(0,L(y_i,w_i|\beta,\boldsymbol{\lambda},rest))$ draw. Likewise, the full conditional posterior density of each λ_l (with $\boldsymbol{\lambda}=(\lambda_l|l=1,\ldots,p)'$) is proportional to Normal $(\boldsymbol{\lambda}|\mu_\lambda,\Sigma_\lambda)$ $\mathbf{1}(\lambda_l\in A_{\lambda_l})$, where $A_{\lambda_l}=\{\lambda_l: L(y_i,w_i|\beta,\boldsymbol{\lambda},rest)>u_i,i=1,\ldots,n\}$, and where each u_i is a Uniform $(0,L(y_i,w_i|\beta,\boldsymbol{\lambda},rest))$ draw. At each stage of the Gibbs sampler, each of the sets A_β and A_{λ_l} can be identified using Neal (2003) "stepping-out" or "doubling" algorithm. Also, based on standard results on Bayesian inference of multivariate normal distributions (Evans, 1965), given $(\beta_{0k_i},\beta_{Tk_i},\lambda_{0k_i})$, $i=1,\ldots,n$, the full conditional posterior of $\boldsymbol{\mu}$ given $\mathbf{T}$ is a trivariate

$$\text{Normal}((n_c\mathbf{T}^{-1}+\mathbf{T}_0^{-1})^{-1}(n_c\mathbf{T}^{-1}\boldsymbol{\mu}_c+(\mathbf{T}_0^{-1}\boldsymbol{\mu}_0)),(n_c\mathbf{T}^{-1}+\mathbf{T}_0^{-1})^{-1})$$

distribution, and a sample of $\mathbf{T}$ from its full conditional is given by $\mathbf{T}^{-1}\sim$ Wishart $(v_0+(p-1)+n_c,(\mathbf{R}_0+\mathbf{U})^{-1})$, with $\mathbf{U}=\sum_c((\beta_{0c},\beta_{Tc},\lambda_{0c})-\boldsymbol{\mu})((\beta_{0c},\beta_{Tc},\lambda_{0c})-\boldsymbol{\mu})'$, and n_c is the number of distinct $(\beta_{0k_i},\beta_{Tk_i},\lambda_{0k_i})$ $(i=1,\ldots,n)$ with $\boldsymbol{\mu}_c$ denoting their average. Finally, when the outcomes y_i are modeled as normal with unknown variance σ^2, the full conditional posterior of σ^{-2} is a Gamma $((a_1+n)/2,((a_2+\mathbf{r}_i'\mathbf{r}_i)/2)^{-1})$ distribution, where $\mathbf{r}_i=(y_i-(\beta_{0k_i}+\beta_{Tk_i}w_i+\beta e(\lambda_{0i}+\boldsymbol{\lambda}'\mathbf{x}_i)|i=1,\ldots,n)$ (e.g., Kleinman and Ibrahim, 1998).

Finally, the same sampling procedures described above are easily extended to the model with more than two treatments, $w=0,1,\ldots,W$. The only difference being that we would replace $(\beta_T,\beta,\lambda_0)$ with $(\boldsymbol{\beta}_T,\boldsymbol{\beta},\boldsymbol{\lambda}_0)$ and replace $e_1(\mathbf{x};\lambda_0,\boldsymbol{\lambda}')$ with $\mathbf{e}(\mathbf{x};\boldsymbol{\lambda}_0,\boldsymbol{\lambda})$, with the vector $\mathbf{e}(\mathbf{x};\boldsymbol{\lambda}_0,\boldsymbol{\lambda})$ defined by the random-intercept multinomial–logit model (Appendix A) having likelihood $e_{w_i}(\mathbf{x}_i;\lambda_{0wi},\boldsymbol{\lambda})$ (for $i=1,\ldots,n$).

References

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., Wynder, E., 1959. Smoking and lung cancer: recent evidence and a discussion of some questions. Journal of the National Cancer Institute 22, 173–203.

Damien, P., Wakefield, J., Walker, S., 1999. Gibbs sampling for Bayesian nonconjugate and hierarchical models using auxiliary variables. Journal of the Royal Statistical Society, Series B 61, 331–344.

Drake, C., 1993. Effects of misspecification of the propensity score on estimators of treatment effects. Biometrics 49, 1231–1236.

Duchateau, L., Janssen, P., 2008. The Frailty Model. Springer, New York.

Evans, I., 1965. Bayesian estimation of parameters of a multivariate normal distribution. Journal of the Royal Statistical Society, Series B 27, 279–283.

Gelman, A., Meng, X.-L., 2004. Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. Wiley, New York.

Greenland, S., 2005. Multiple-bias modelling for analysis of observational data. Journal of the Royal Statistical Society, Series A 168, 267–306.

Gustafson, P., 2005. On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. Statistical Science 20, 111–140.

Holland, P., 1986. Statistics and causal inference. Journal of the American Statistical Association 81, 945–960.

Holland, P., Rubin, D., 1983. On Lord's paradox. In: Wainer, H., Messick, S. (Eds.), Principles of Modern Psychological Measurement, Lawrence Erlbaum Associates, Hillsdale.

Hong, G., Raudenbush, S., 2006. Evaluating kindergarten retention policy: a case study of causal inference for multilevel observational data. Journal of the American Statistical Association 101, 901–910.

Kalli, M., Griffin, J., Walker, S., 2010. Slice sampling mixture models. Statistics and Computing 21, 93–105.

Kleinman, K., Ibrahim, J., 1998. A semiparametric Bayesian approach to the random effects model. Biometrics 54, 921–938.

Lechner, M., 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In: Pfeiffer, F., Lechner, M. (Eds.), Econometric Evaluation of Labour Market Policies, Physica-Verlag, Heidelberg, pp. 43–58.

Lin, D., Psaty, B., Kronmal, R., 1998. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics 54, 948–963.

Lord, F., 1967. A paradox in the interpretation of group comparisons. Psychological Bulletin 685, 304–305.

McCandless, L., Gustafson, P., Austin, P., 2009. Bayesian propensity score analysis for observational data. Statistics in Medicine 28, 94–112.

McCandless, L., Gustafson, P., Levy, A., 2007. Bayesian sensitivity analysis for unmeasured confounding in observational studies. Statistics in Medicine 26, 2331–2347.

McCandless, L., Gustafson, P., Levy, A., 2008a. A sensitivity analysis using information about measured confounders yielded improved uncertainty assessments for unmeasured confounding. Journal of Clinical Epidemiology 61, 247–255.

Please cite this article as: Karabatsos, G., Walker, S.G., A Bayesian nonparametric causal model. Journal of Statistical Planning and Inference (2011), doi:10.1016/j.jspi.2011.10.013

McCandless, L., Richardson, S., Best, N., 2008b. Adjustment for Unmeasured Confounding Using Propensity Scores. Department of Epidemiology and Public Health, Imperial College London, UK.
McCullagh, P., Nelder, J., 1989. Generalized Linear Models, second ed. Chapman and Hall, London.
Müller, P., Quintana, F., 2004. Nonparametric Bayesian data analysis. Statistical Science 19, 95–110.
Neal, R., 2003. Slice sampling (with discussion). Annals of Statistics 31, 705–767.
Neyman, J., 1923/1990. On the application of probability theory to agricultural experiments: essay on principles, section 9. Annals of Agricultural Science Translated in Statistical Science 5, 465–472.
Pitman, J., 1996. Some developments of the Blackwell–MacQueen urn scheme. In: Ferguson, T., Shapeley, L., MacQueen, J. (Eds.), Statistics, Probability and Game Theory. Papers in Honor of David Blackwell, Institute of Mathematical Sciences, Hayward, CA, pp. 245–268.
Robins, J., 2002. Comments on 'Covariance adjustment in randomized experiments and observational studies' by P.R. Rosenbaum. Statistical Science 17, 309–321.
Robins, J., Hernán, M., Brumback, B., 2000. Marginal structural models and causal inference in epidemiology. Epidemiology 11, 550–560.
Rosenbaum, P., 1987. Model-based direct adjustment. Journal of the American Statistical Association 82, 387–394.
Rosenbaum, P., 2002a. Observational Studies, second ed. Springer-Verlag, New York.
Rosenbaum, P., 2002b. Covariance adjustment in randomized experiments and observational studies (with discussion). Statistical Science 17, 286–327.
Rosenbaum, P., Rubin, D., 1983a. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.
Rosenbaum, P., Rubin, D., 1983b. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. Journal of the Royal Statistical Society, Series B 45, 212–218.
Rosenbaum, P., Rubin, D., 1984. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 79, 516–524.
Rubin, D., 1978. Bayesian inference for causal effects: the role of randomization. Annals of Statistics 6, 34–58.
Rubin, D., 1985. The use of propensity scores in applied Bayesian inference. In: Bernardo, J., De Groot, M., Lindley, D., Smith, A. (Eds.), Bayesian Statistics, Elsevier Science Publishers, North Holland, pp. 463–472.
Rubin, D., 1990. Neyman (1923) and causal inference in experiments and observational studies. Statistical Science 5, 472–480.
Rubin, D., 2006. Matched Sampling for Causal Effects. Cambridge University Press, Cambridge.
Rubin, D., 2007. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Statistics in Medicine 26, 20–36.
Schwartz, L., 1965. On Bayes procedures. Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete 4, 10–26.
Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Statistica Sinica 4, 639–650.
Walker, S., 2003. On sufficient conditions for Bayesian consistency. Biometrika 90, 482–488.
Walker, S., 2007. Sampling the Dirichlet mixture model with slices. Communications in Statistics: Simulation and Computation 36, 45–54.
Walker, S., Lijoi, A., Prünster, I., 2005. Data tracking and the understanding of Bayesian consistency. Biometrika 92, 765–778.
Zhao, C., Vodicka, P., Sram, R., Hemminki, K., 2000. Human DNA adducts of 1,3-butadiene, an important environmental carcinogen. Carcinogenesis 21, 107–111.

Please cite this article as: Karabatsos, G., Walker, S.G., A Bayesian nonparametric causal model. Journal of Statistical Planning and Inference (2011), doi:10.1016/j.jspi.2011.10.013