

# Adjustment uncertainty in effect estimation

BY CIPRIAN M. CRAINICEANU, FRANCESCA DOMINICI

*Department of Biostatistics, Johns Hopkins University, 615 North Wolfe Street, Baltimore, Maryland 21205, U.S.A.*

ccrainic@jhsp.edu fdominic@jhsp.edu

AND GIOVANNI PARMIGIANI

*Department of Oncology, Johns Hopkins University, 550 North Broadway, Baltimore, Maryland 21205, U.S.A.*

gp@jhu.edu

## SUMMARY

Often there is substantial uncertainty in the selection of confounders when estimating the association between an exposure and health. We define this type of uncertainty as ‘adjustment uncertainty’. We propose a general statistical framework for handling adjustment uncertainty in exposure effect estimation for a large number of confounders, we describe a specific implementation, and we develop associated visualization tools. Theoretical results and simulation studies show that the proposed method provides consistent estimators of the exposure effect and its variance. We also show that, when the goal is to estimate an exposure effect accounting for adjustment uncertainty, Bayesian model averaging with posterior model probabilities approximated using information criteria can fail to estimate the exposure effect and can over- or underestimate its variance. We compare our approach to Bayesian model averaging using time series data on levels of fine particulate matter and mortality.

*Some key words:* Adjustment uncertainty; Air pollution; Bayesian model averaging.

## 1. INTRODUCTION

In epidemiology, regression models are generally used to estimate the effect of an exposure  $X$  on a response  $Y$ , while controlling for confounders chosen from a large set of potential confounders  $U$ . Examples include both controlled and observational studies in nearly all fields of scientific investigation. We refer to the uncertainty in the selection of confounders and in their functional relationship with the outcome as ‘adjustment uncertainty’. There are important differences between adjustment uncertainty and model uncertainty. In adjustment uncertainty, the estimation of the adjusted exposure effect, such as the regression coefficient of  $X$  in a model for  $Y$ , is the goal of the analysis. Additional covariates  $U$  are included to account for potential confounding. In model uncertainty, all predictors ( $X, U$ ) are equally important, and their inclusion in the regression model is evaluated based on measures of prediction performance, rather than adequacy in controlling for confounders of  $X$ .

In this paper we propose a two-stage statistical approach for adjustment uncertainty in exposure effect estimation, a specific implementation, and associated visualization tools. In the first stage we identify a sequence of exposure models,  $\alpha_{m,X}$ , containing the best set of  $m$  predictors of

the exposure  $X$ ,  $U_{\alpha_{m,X}}$ , among the potential confounders  $U$  for every  $m = 1, \dots, M$ , where  $M$  is the total number of potential confounders. For each  $m$  we plot the exposure effect estimate for the model with  $Y$  as outcome,  $X$  as exposure, and  $U_{\alpha_{m,X}}$  as confounders. We also plot the deviance difference between models  $\alpha_{m-1,X}$  and  $\alpha_{m,X}$  to visualize the exposure prediction change. The set of confounders is identified using stabilization of the exposure effect estimates and deviance differences. In the second stage we identify a sequence of outcome models, that is regression models for  $Y$ , having  $X$  as exposure and the set of confounders identified in the first stage. The outcome model space is explored using the same technique to identify additional covariates that are good predictors of  $Y$ .

A nice feature of our methodology is its practical visualization. For example, we produce plots of point estimates and confidence intervals of the exposure effect as a function of different levels of confounding adjustment. Moreover, the proposed methodology provides a sound statistical platform for identifying confounders from a large set of potential confounders and quantifying their effect on exposure effect estimates.

The methodological development of this paper is motivated by time series studies of health effects of air pollution exposure. Time series studies of air pollution and health carried out around the world, such as Bell et al. (2004), Dominici et al. (2000), Katsouyanni et al. (2001) and Samet et al. (2000), provide important epidemiological evidence for regulatory purposes and are at the centre of an intense national debate in the U.S.A.; see for example Kaiser (2005) and National Research Council (1998, 1999, 2001). Therefore, properly accounting for adjustment uncertainty is of fundamental importance in this context. In this paper we show how our proposed methodology provides asymptotically unbiased estimators of the effect of air pollution exposure on mortality. Our results also show that IC-based Bayesian model averaging, defined as Bayesian model averaging with posterior model probabilities approximated using BIC and AIC, or using Zellner's priors, produces biased estimators of the exposure effect and its associated variability. We compare our methodology with IC-based Bayesian model averaging because in air-pollution research, weighted averaging of model-specific coefficients using posterior probabilities derived from information criteria as weights has been advocated as a way of handling adjustment uncertainty in the estimation of effects, for example by Clyde (2000) and Koop & Tole (2004).

The R function STEADy, simulations and analyses for this paper are available at [www.biostat.jhsph.edu/~ccrainic/webpage/software/STEADy.zip](http://www.biostat.jhsph.edu/~ccrainic/webpage/software/STEADy.zip).

## 2. STATISTICAL APPROACHES FOR ADJUSTMENT UNCERTAINTY

### 2.1. Likelihood models and estimands

We are interested in the association between the exposure  $X$  and outcome  $Y$ , adjusted by the effect of a large set of potential confounders  $U = (U_1, \dots, U_M)$ . We allow the covariates  $Z = (Z_1, \dots, Z_K)$  always to enter into the model because of their scientific importance.

Consider a model space of the form

$$Y_i = \beta^\alpha X_i + \sum_{k=1}^K \gamma_k^\alpha Z_{ik} + \sum_{m=1}^M \alpha_m \delta_m^\alpha U_{im} + \epsilon_i^\alpha \quad (i = 1, \dots, n),$$

where  $n$  is the sample size,  $K$  is the number of covariates that are always included in the model, and  $M$  is the number of potential confounders. Here  $\alpha = (\alpha_1, \dots, \alpha_M)^\top \in \{0, 1\}^M$  with  $\alpha_i = 1$  if and only if the  $i$ th covariate is included in the model. We refer to the vector  $\alpha$  as the model

and use the notation  $\alpha \subseteq \alpha'$  if the model  $\alpha$  is nested within model  $\alpha'$ . The interpretation of the coefficient  $\beta^\alpha$  may vary with  $\alpha$ .

We assume that the  $M$  potential confounders include all the relevant confounders necessary for identifying the true exposure effect. In our notation this means that the true model  $\alpha^*$  belongs to the model space and we denote by  $\beta^* = \beta^{\alpha^*}$  the corresponding true exposure effect. In addition, we assume that all models nesting  $\alpha^*$  estimate the same true effect, so  $\beta^\alpha = \beta^*$  whenever  $\alpha^* \subseteq \alpha$ .

We define adjustment uncertainty in effect estimation as estimation of  $\beta^*$  when  $\alpha^*$  is unknown. Identifying  $\alpha^*$  is difficult, but identifying a reasonable  $\alpha_0$  such that  $\alpha^* \subseteq \alpha_0$  is easier and will be the focus of the estimation procedure described in the next section. If  $\alpha^* \subseteq \alpha$  then we can provide unbiased estimators  $\hat{\beta}^\alpha$  of  $\beta^*$ . Typically  $\text{var}(\hat{\beta}^\alpha) \geq \text{var}(\hat{\beta}^*)$  and confidence intervals based on model  $\alpha$  will tend to be conservative.

In contrast, any model  $\alpha$  such that  $\alpha^* \not\subseteq \alpha$  will contain an incorrect estimand of the exposure effect and will likely provide invalid inferences from either a frequentist or a Bayesian standpoint. For example, a Bayesian model averaging estimate of  $\beta^*$  is obtained by specifying a prior  $p(\alpha) p(\beta^\alpha, \gamma^\alpha, \delta^\alpha | \alpha)$  and calculating

$$\tilde{\beta}^* = \sum_{\alpha: \alpha^* \subseteq \alpha} E(\beta^\alpha | \alpha, D) p(\alpha | D) + \sum_{\alpha: \alpha^* \not\subseteq \alpha} E(\beta^\alpha | \alpha, D) p(\alpha | D),$$

where  $D$  denotes the data and  $p(\alpha | D)$  is the posterior probability of model  $\alpha$ . In § 6 we illustrate how this can lead to seriously misleading results as commonly used priors may lead to large posterior weights on models that include an incorrect estimand.

Motivated by these considerations, we define and address two problems related to adjustment uncertainty. First, we identify a reasonable  $\alpha_0$  such that  $\alpha^* \subseteq \alpha_0$ . Secondly, we develop a computationally tractable framework that describes the relationship between inference about  $\beta^\alpha$  and the model space structure.

## 2.2. Structured estimation under adjustment uncertainty

In this section we introduce an algorithm for exploration of the exposure and outcome space to ultimately identify a reasonable model  $\alpha_0$  such that  $\alpha^* \subseteq \alpha_0$ . Our approach focuses on the joint model for outcome and exposure  $[Y, X | Z, U] = [Y | X, Z, U][X | Z, U]$  and can be described in two steps. In the first step, we focus on the exposure model space  $\mathcal{M}_X$ , identify strong predictors of the exposure  $X$  and include them in the outcome model for  $Y$ . Here the set of predictors includes all covariates  $Z$  and any subset of the potential confounders  $U = (U_1, \dots, U_M)$ .

In § 2.1, we introduced  $\alpha = (\alpha_1, \dots, \alpha_M)^\top \in \{0, 1\}^M$  with  $\alpha_i = 1$  if and only if  $U_i$  is in the model. Here we also denote by  $U_\alpha$  the subset of potential confounders of  $U$  selected by the indicator vector  $\alpha$ . We divide  $\mathcal{M}_X$  into  $M + 1$  subsets, or orbits, corresponding to models with a fixed number of confounders  $U$ . For example, the  $m$ th orbit is the set of all regression models  $\alpha$  with the property  $\sum_{i=1}^M \alpha_i = m$ . Within orbit  $m$  we select the model  $\alpha_{m,X}$  that maximizes the likelihood. While computationally intensive, finding the maximum likelihood model within each orbit avoids the typical problems of greedy algorithms such as forward selection or projection-pursuit. Details of the likelihood maximization are provided in Appendix A. The set of  $M + 1$  regression models  $\alpha_{0,X}, \dots, \alpha_{M,X}$  will be called the dominant model class. Once the dominant model class is determined, it is useful to display the deviance differences  $D(\alpha_{m-1,X}) - D(\alpha_{m,X})$ , which show the change in deviance between adjacent orbits and indicate what combinations of variables are good predictors of  $X$  and therefore are likely confounders. The set of orbits in which the deviance difference function becomes small identifies an appropriate range for the required dimension of the exposure model. Showing point estimates and confidence intervals

corresponding to the outcome models  $[Y | X, Z, U_{\alpha_{0,X}}], \dots, [Y | X, Z, U_{\alpha_{M,X}}]$  is a useful visualization tool for detecting confounding.

In the second step we identify strong predictors of outcome that have not been identified in the first step. Using the same procedure as in that step, we partition the outcome models into  $L + 1$  orbits. Here  $L \leq M$  because the number of potential confounders is usually reduced during the first phase of the procedure. Denote by  $\alpha_{l,Y}$  the model that maximizes the likelihood of the outcome models on the  $l$ th orbit, and call  $\alpha_{0,Y}, \dots, \alpha_{L,Y}$  the dominant model class for the outcome models. The plot of deviance differences  $D(\alpha_{l-1,Y}) - D(\alpha_{l,Y})$  indicates which covariates, in addition to  $X$  and  $Z$ , are predictive of  $Y$ . As before, we identify the region in which the deviance difference function becomes small. We also display the point estimates and confidence intervals for the parameter of  $X$  corresponding to the outcome models  $[Y | X, Z, U_{\alpha_{0,Y}}], \dots, [Y | X, Z, U_{\alpha_{L,Y}}]$ . This plot illustrates the amount of variation in the exposure effect estimate when adjusting for  $Z$  and a subset of  $U$ . The point estimate and confidence interval of the exposure effect are obtained after both the difference deviance function for outcome models and the parameter estimate have stabilized. Stabilization is evidence that  $\alpha^*$  was reached.

In summary, the algorithm consists of the following steps.

*Step 1.* Obtain the dominant model class  $\text{DMC}_X = \alpha_{0,X}, \dots, \alpha_{M,X}$  of exposure models.

*Step 2.* Plot deviance differences for the exposure models  $\alpha_{0,X}, \dots, \alpha_{M,X}$ .

*Step 3.* Plot point estimates and confidence intervals of the exposure effect under the outcome models  $[Y | X, Z, U_{\alpha_{0,X}}], \dots, [Y | X, Z, U_{\alpha_{M,X}}]$ .

*Step 4.* Identify  $U_{P,X}$  such that  $D(\alpha_{m-1,X}) - D(\alpha_{m,X})$  is small for every  $m \geq P$ .

*Step 5.* Define  $Z = Z \cup U_{P,X}$ ,  $U = U \setminus U_{P,X}$ ,  $L = M - P$ .

*Step 6.* Obtain the dominant model class  $\text{DMC}_Y = \alpha_{0,Y}, \dots, \alpha_{L,Y}$  of outcome models.

*Step 7.* Plot deviance differences for the outcome models  $\alpha_{0,Y}, \dots, \alpha_{L,Y}$ .

*Step 8.* Plot point estimates and confidence intervals for the exposure parameter in the sequence of models  $\alpha_{0,Y}, \dots, \alpha_{L,Y}$ .

*Step 9.* Identify the region where  $D(\alpha_{l-1,Y}) - D(\alpha_{l,Y})$  becomes small.

*Step 10.* Among models that provide similar exposure effect estimates identify the one with smallest exposure effect variance.

### 3. THEORETICAL RESULTS FOR ADJUSTMENT UNCERTAINTY

In this section we investigate the theoretical properties of the adjustment uncertainty estimator and compare it with other weighted estimators, such as the IC-based Bayesian model averaging estimator for which the weights are the posterior model probabilities. Even though Bayesian model averaging has been proposed to incorporate model uncertainty in prediction (Draper, 1995; Madigan & Raftery, 1994; Madigan et al., 1996; Hoeting et al., 1999), it has also been advocated (Raftery, 1994; Clyde, 2000; Koop & Tole, 2004) as a procedure that incorporates the model selection uncertainty into exposure effect estimation.

We first introduce some notation. Let  $\alpha$  be a general model and let  $\alpha_E$ ,  $\alpha^*$  and  $\alpha_F$  be the model with no additional  $U$  covariate, the true model and the full model containing all the confounders, respectively. Let  $U = R \cup I$ , where  $R$  is the set of confounders for  $\alpha^*$  and  $I$  is the set of the remaining confounders. Let  $\hat{\beta}_n^\alpha$  be the maximum likelihood estimator of  $\beta^\alpha$ , the exposure effect under model  $\alpha$ . We assume that, under  $\alpha^*$ ,  $\hat{\beta}_n^\alpha \rightarrow \beta^\alpha$  almost surely. For a

set of weights  $W_n = \{w_n(\alpha)\}$  with  $\sum_{\alpha} w_n(\alpha) = 1$  and  $w_n(\alpha) \geq 0$ , define the weighted estimator  $\hat{\beta}_{W_n} = \sum_{\alpha} w_n(\alpha) \hat{\beta}_n^{\alpha}$ . Since  $W_n$  is in a compact set it contains at least one convergent subsequence and all  $\hat{\beta}_n^{\alpha}$  sequences are convergent. To ensure convergence of the weighted estimator  $\hat{\beta}_{W_n}$ , we will assume that  $W_n$  itself converges to a limit,  $W = \{w(\alpha)\}_{\alpha}$  say. Let  $\hat{\beta}_n^{\text{AU}} = \hat{\beta}_n(\hat{\alpha}_n^{\text{AU}})$ , where  $\hat{\alpha}_n^{\text{AU}} = \arg \min\{\hat{v}_n(\alpha) : |\hat{\beta}_n^{\alpha} - \hat{\beta}_n^{\alpha_F}| < \epsilon/h_n\}$ ,  $\hat{v}_n(\alpha) = \text{var}_{\alpha}(\hat{\beta}_n^{\alpha})$ ,  $h_n/n^{1/2} \rightarrow 0$  when  $n \rightarrow \infty$ , and  $\epsilon > 0$  is a constant.

**THEOREM 1.** *Under typical regularity assumptions the following hold:*

- (i)  $\hat{\beta}_n^{\text{AU}}$  is asymptotically unbiased;
- (ii) the asymptotic bias of  $\hat{\beta}_{W_n}$  is  $\text{bias}\{\hat{\beta}_{W_n}\} = \sum_{\alpha \in I} w(\alpha) \beta^{\alpha} - \{\sum_{\alpha \in I} w(\alpha)\} \beta^*$ ;
- (iii) if  $v(\alpha) = \lim_{n \rightarrow \infty} \{n \hat{v}_n(\alpha)\}$  then  $\limsup_{n \rightarrow \infty} \{n \hat{v}_n(\hat{\alpha}_n^{\text{AU}})\} \leq v(\alpha_F)$ . If there exists an  $\alpha_0$  such that  $\{\alpha^* \subset \alpha_0, v(\alpha_0) < v(\alpha_F)\}$  then the inequality is strict.

The proof of the theorem is in Appendix B. With a slight abuse of notation, here  $\alpha \in I$  denotes models that do not contain all variables in  $R$ . If  $w(\alpha) = 0$  for every  $\alpha \in I$  then the weighted estimator is asymptotically unbiased. However, if this is not true the asymptotic unbiasedness could only be the result of a lucky choice of weights. Indeed, the asymptotic bias is zero if and only if  $\beta^* = \sum_{\alpha \in I} w(\alpha) \beta^{\alpha} / \sum_{\alpha \in I} w(\alpha)$ , which can only happen by accident when  $\beta^*$  and the set  $I$  are unknown. Moreover, if  $\beta^{\alpha} > \beta^*$  for every  $\alpha \in I$  then all weighted estimators with positive weights, will be biased.

The minimum variance of the adjustment uncertainty estimator is ensured by definition in finite samples. Theorem 1 shows that any subsequence of adjustment uncertainty estimators for which the limit of  $n \hat{v}_n(\alpha)$  exists is asymptotically unbiased with asymptotic variance smaller than the asymptotic variance of the full model. In (iii) we also provide a sufficient condition under which the dominance is strict.

A popular choice is to use the Bayesian Information Criterion (BIC) weights

$$w_n(\alpha) = \exp\{-\text{BIC}_n(\alpha)/2\} / \left[ \sum_{\alpha'} \exp\{-\text{BIC}_n(\alpha')/2\} \right],$$

where  $\text{BIC}_n(\alpha) = -2 \log\{\rho_n(\alpha)\} + p_{\alpha} \log(n)$ , and  $\rho_n(\alpha)$  and  $p_{\alpha}$  are the maximum likelihood and the number of parameters of model  $\alpha$ , respectively. We examine now the properties of BIC as a method for adjustment uncertainty.

**THEOREM 2.** *Consider a normal family of models as described in Section 2.1 with the usual regularity conditions (Shao, 1998, Chapter 4) for existence, consistency and asymptotic normality of the maximum likelihood estimators. Assume that there exists  $S \subset R$  with the property that  $\delta_m^{\alpha^*} = c_m n^{-1/2}$  for every  $U_m \in S$ , where  $\alpha^*$  is the true model. Then  $\lim_{n \rightarrow \infty} w_n(\alpha^{**})/w_n(\alpha) = \infty$  for every model  $\alpha$  such that  $\alpha^* \subset \alpha$ , where  $\alpha^{**}$  is the model obtained from  $\alpha^*$  by deleting the true confounders from  $S$ , and  $w_n(\cdot)$  are the BIC weights.*

The proof is in Appendix B. This result applies when there exists a set of confounders,  $S$ , with effects on the outcome that are local to zero in the true model,  $\alpha^*$ . In contrast, standard asymptotic results consider scenarios where none of the confounders have effects that are local to zero. However, these results are not consistent with the evidence provided by our simulation in § 4, suggesting that BIC could place too much mass on models that do not contain all the confounders, even when the sample sizes are very large. If all confounders have local to zero effects on the outcome, that is  $S = R$ , then the model without confounding adjustment has, asymptotically, infinitely more mass than any model that estimates  $\beta^*$ . Moreover, if the exposure,  $X$ , is not forced

into models and  $X \in S$  then the model that estimates zero exposure effect has, asymptotically, infinitely more mass than any model that estimates  $\beta^*$ . The assumption in Theorem 2 is the local asymptotic assumption and is widely used for confidence-interval estimation.

The finite-sample behaviour of the BIC when the sample size is large and the signal is small-to-moderate can be illustrated using a simple example. Suppose that only two nested models,  $\alpha_0 \subset \alpha_1$ , are available with  $\alpha_1$  having just one extra parameter. The BIC selects model  $\alpha_0$  if and only if  $\text{BIC}_n(\alpha_0) < \text{BIC}_n(\alpha_1)$ , which is equivalent to  $2 \log\{\rho(\alpha_1)/\rho(\alpha_0)\} < \log(n)$ . The left-hand side of this inequality is the likelihood-ratio statistic for testing  $\alpha_0$  versus  $\alpha_1$ . In this case selection using the BIC is equivalent to using the likelihood-ratio test with a critical value equal to  $\log(n)$ . This may be reasonable when  $n$  is small or moderate. For example,  $\log(n)$  varies between 3.9 and 6.2 when  $n$  varies between 50 and 500. These values are comparable to the 0.95 and 0.99 quantiles of a  $\chi^2_1$  distribution, which are 3.84 and 6.63 respectively. More dramatic effects occur when the sample size is large with the level of the test dropping to 0.002 for  $n = 15\,000$  and to 0.0005 for  $n = 150\,000$ . Thus, using BIC in large datasets might ignore even very strong signals and lead to wrong conclusions. This discussion is of particular interest for studies of air pollution effects on mortality, which routinely use over 10 000 observations per city.

#### 4. SIMULATION STUDY

##### 4.1. Data generating process

We describe a simulation study to compare the adjustment uncertainty algorithm with the BIC-based Bayesian model averaging approach for estimating the exposure effect. We consider a relatively simple data-generating mechanism to allow a transparent comparison across methods, and yet capture important features of effect estimation in air pollution research and other areas. We generate counts from a model  $Y_i \sim \text{Po}(\mu_i)$ , where

$$\log(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 U_{2i} + \beta_3 U_{3i} \quad (i = 1, \dots, 1000), \quad (1)$$

and  $(X_{1i}, U_{2i}, U_{3i})$  are independent normal vectors with mean zero, variance 1 and all covariances zero except for  $\text{cov}(X_{1i}, U_{2i}) = \rho$ . Here  $X_1$  is the exposure variable,  $U_2$  is a confounder that predicts both  $X_1$  and  $Y$ , and  $U_3$  predicts  $Y$  but is independent of the exposure. To define a correctly adjusted effect of  $X_1$ , the confounder  $U_2$  must be included in the model. The set  $U$  of potential confounders includes  $U_2$ ,  $U_3$  and 50 additional independent  $N(0, 1)$  random variables. In this example there is no  $Z$  variable. The model space  $\mathcal{M}$  has  $2^{52}$  models if we include the exposure variable by default, and  $2^{53}$  if we allow it to be included or excluded.

We explore two simulation scenarios: the first draws a new set of  $X$ 's,  $U$ 's and  $Y$ 's for each new dataset while the second draws a new set of  $Y$ 's only and keeps the same set of  $X$ 's and  $U$ 's. We simulated 100 datasets from model (1) under each scenario. We set the parameter  $\rho$  to 0, corresponding to no confounding, or to 0.7 corresponding to a moderate confounding. The coefficients are set at  $\beta_1 = \beta_2 = \beta_3 = 0.1$  and  $\beta_0 = 0$ .

##### 4.2. Results when the exposure is always in the model

For each simulated dataset,  $\beta_1$  was estimated based on our adjustment uncertainty algorithm and using IC-based Bayesian model averaging over the outcome dominant model class. Posterior model probabilities were approximated using three methods: BIC (BIC-D), the likelihood ratio test (LRT-D) and AIC (AIC-D). These criteria correspond to a penalty of  $\log(1000) = 6.9$ , 4 and 2, respectively, for each additional variable. For BIC we also report results based on a full stochastic exploration algorithm, see Appendix A for details, of the model space (BIC-SE). Since Bayesian model averaging approaches are based on prediction and therefore on outcome models, we first implement our algorithm on the outcome's dominant model class only (AU-OD), thus allowing for

Table 1. Means and standard errors of  $\beta_1$  estimates over 100 datasets for  $\rho = 0.7$ . The true parameter is  $\beta_1 = 0.1$ . Results for random  $U$ 's correspond to the scenario where both covariates and outcomes were simulated. Results for fixed  $U$ 's correspond to the scenario where only the outcome is simulated

| Scenario |                     | AU-OD  | BIC-D  | BIC-SE | LRT-D  | AIC-D  |
|----------|---------------------|--------|--------|--------|--------|--------|
| Random   | $M(\hat{\beta}_1)$  | 0.1020 | 0.1330 | 0.1310 | 0.1230 | 0.1130 |
| Random   | $SE(\hat{\beta}_1)$ | 0.0048 | 0.0057 | 0.0055 | 0.0059 | 0.0055 |
| Fixed    | $M(\hat{\beta}_1)$  | 0.1020 | 0.1350 | 0.1240 | 0.1210 | 0.1110 |
| Fixed    | $SE(\hat{\beta}_1)$ | 0.0040 | 0.0048 | 0.0054 | 0.0040 | 0.0045 |

AU-OD, Adjustment uncertainty using the outcome model class; BIC-D, LRT-D, AIC-D, Bayesian model averaging using the outcome model class and penalties 6.9, 4 and 2, respectively; BIC-SE, Bayesian model averaging using stochastic search with the BIC penalty, 6.9.

a more direct comparison between the two approaches. The preferred application of our algorithm has two stages and involves a preliminary search for confounders in the exposure model.

When  $\rho = 0$ , that is in the absence of confounding, all methods produced very similar estimates of  $\beta_1$ , these results are not reported. Table 1 summarizes the average,  $M(\hat{\beta}_1)$ , and standard error,  $SE(\hat{\beta}_1)$ , of the  $\beta_1$  estimates over 100 datasets for  $\rho = 0.7$  and for the two simulation scenarios described above, with  $X$  and  $U$  fixed and  $X$  and  $U$  random. Under AU-OD we report estimates of  $\beta_1$  after they stabilize, which occurs by the 20th orbit. In both scenarios, AU-OD and IC-based Bayesian model averaging produce very different results with AU-OD outperforming IC-based Bayesian model averaging in terms of both bias and variance. The bias of Bayesian model averaging with BIC-D and BIC-SE is such that the  $(1-10^{-6})\%$  approximate confidence interval based on 100 simulations does not contain the true value of the parameter. When Bayesian model averaging with AIC-D is used the 95% confidence interval does not contain the true value of the parameter.

Figure 1 illustrates why the two approaches perform so differently. Figure 1 (b) shows the estimates of  $\beta_1$  obtained under AU-OD plotted against each of the 52 models considered, for each of 3 datasets and for  $\rho = 0.7$ . The noticeable drop in point estimates occurs when  $U_2$  enters the model. Adjustment uncertainty chooses the parameter estimate after stabilization of the deviance difference function and stabilization of the  $\beta_1$  estimates.

Figure 1 (a) shows deviance differences plotted against models under AU-OD. The horizontal lines are placed at the BIC and AIC penalties. Results are shown for the same three datasets simulated from model (1) with  $\rho = 0.7$ . An interesting feature of Fig. 1 (a) and (b) is that the large changes in the parameter estimates do not correspond to changes in the outcome likelihood models. This is the major reason for the failure of outcome model strategies to provide unbiased exposure model estimators. These plots also highlight the severity of the BIC penalty. Except for very few initial models, deviance differences are always smaller than the penalty and therefore they favour the more parsimonious models.

An inspection of the list of variables that enter in every model revealed that the first model includes only  $U_3$  in addition to  $X_1$  but does not include  $U_2$ . This happens because the deviance gain of including  $U_2$  is typically in the range [3, 8], and BIC assigns higher probability to the model including only  $U_3$ . In this model, the estimate of  $\beta_1$  is artificially inflated because  $U_2$  is correlated with  $X_1$  and not included into the model. Thus, the BIC-D estimate is a weighted average of estimators, some appropriately adjusted when  $U_2$  is in the model and some biased when  $U_2$  is excluded. Bayesian model averaging with posterior model probabilities based on information criteria with smaller penalties has similar limitations, though in this example they tend to do slightly better.

Next we consider the complete, two-stage adjustment uncertainty procedure. Figure 1 (c) shows deviance differences plotted against models for exposure's dominant class (AU-ED). The

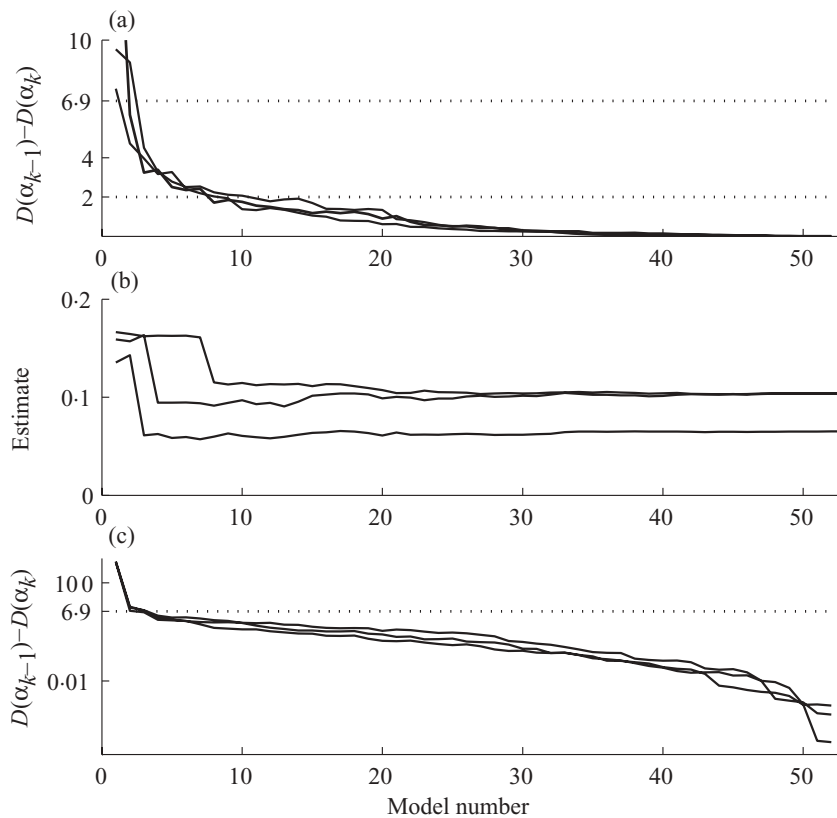


Fig. 1. Results for three datasets simulated from model (1) with  $\rho = 0.7$  with random  $X$ s and  $U$ s. (a) Deviance differences between the outcome dominant models on orbits  $k - 1$  and  $k$ . The dotted lines are placed at the BIC and AIC penalties  $\log(1000) = 6.91$  and 2, respectively. (b) Estimates of  $\beta_1$  for each of the 3 simulated datasets all plotted against the model number. (c) Deviance differences, on log scale, between the exposure dominant model on orbit  $k - 1$  and on orbit  $k$ . The dotted line is placed at 6.9. The line at 2 was omitted because it was too close to the one at 6.9.

horizontal lines are placed at the BIC penalty. When  $\rho = 0.7$ , we found that AU-ED identifies  $U_2$  as an important predictor of  $X_1$ . We then forced  $U_2$  into the outcome model and applied AU-OD: here we found that the  $\beta_1$ 's estimates stabilize on about the 5th orbit. The two-step procedure, AU-ED followed by AU-OD, produced slightly better results than AU-OD alone.

Figure 2 compares the standard errors of the estimates of  $\beta_1$  for 100 independent datasets obtained using AU-OD, on the  $X$ -axis, versus BIC-D, on the  $Y$ -axis, when  $\rho = 0.7$ , Fig. 2 (a), and  $\rho = 0$ , Fig. 2 (b). When  $\rho = 0.7$ , BIC-D can produce up to 35% wider or narrower confidence intervals than AU-OD. Small standard errors for BIC-D correspond to models that do not include the key confounder  $U_2$ . In these models, part of the signal generated by  $U_2$  is captured by the parameter estimate of  $X_1$ . Since  $X_1$  and  $U_2$  are positively correlated,  $\hat{\beta}_1$  is biased upwards and its statistical uncertainty is understated.

A different mechanism operates when BIC-D provides larger standard errors. As a result of the large BIC penalty, Bayesian model averaging assigns sizeable probability both to models that contain and do not contain  $U_2$ . In these cases, the  $\beta_1$ 's standard error obtained from Bayesian model averaging is artificially inflated because  $\beta_1$  is estimated by averaging model-specific estimates that include and do not include  $U_2$ . Our adjustment uncertainty procedure avoids these

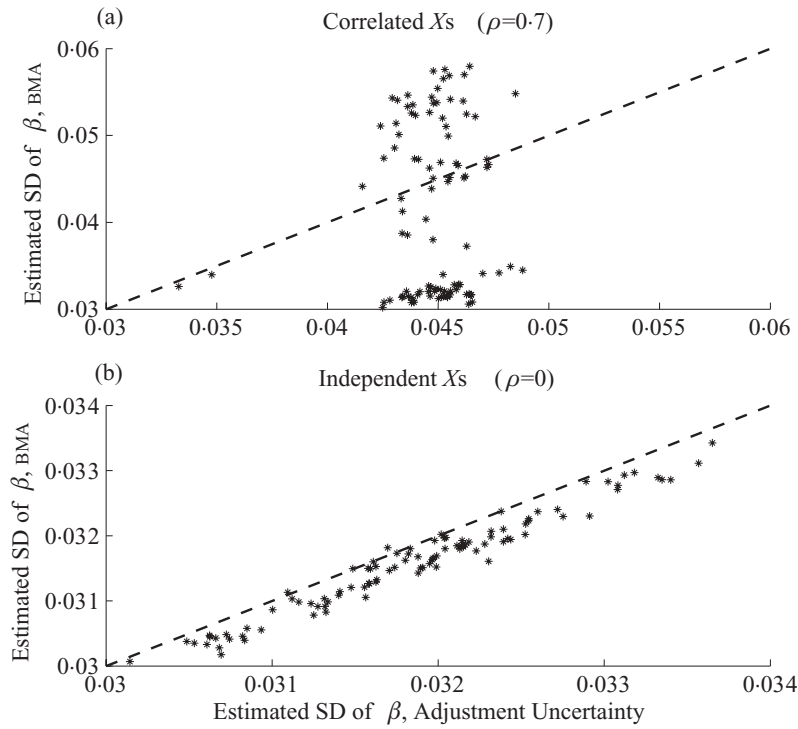


Fig. 2. Estimated standard errors of  $\hat{\beta}_1$  for 100 simulated datasets using adjustment uncertainty,  $X$ -axis, and BIC-based Bayesian model averaging (BMA),  $Y$ -axis. Data were simulated from Model (1) with (a)  $\rho = 0.7$  and (b)  $\rho = 0$  using random  $X$ s and  $U$ s.

problems because it relies on stabilization of point estimates and confidence intervals to ensure that all potential confounders are included in the model.

In summary this simulation study illustrates that, in the presence of confounding, BIC-based Bayesian model averaging produces a biased estimator of the parameter of interest and that the associated 95% confidence interval may be as much as 35% wider or narrower than the interval produced by adjustment uncertainty, which has approximately correct coverage.

Finally, when  $\rho = 0$ , BIC-based Bayesian model averaging and adjustment uncertainty produce almost identical estimators of  $\hat{\beta}_1$  and associated standard errors.

#### 4.3. Results when the exposure is not always in the model

In the previous section we compared adjustment uncertainty and IC-based Bayesian model averaging when the exposure  $X$  is always included in the model. While we consider that this is the right approach, it is often argued in practice that the exposure should not be forced into the model. We now investigate the effect of not forcing  $X$  into all models on exposure effect estimation.

We simulated 100 independent datasets from model (1) with  $\rho = 0.7$  and applied Bayesian model averaging implemented by BIC-D without forcing  $X_1$  into the model. Figure 3 (a) shows the sorted posterior probability that  $X_1$  is included in the model against the 100 simulated datasets. Figure 3 (b) shows the histogram of the 100 estimates of  $\beta_1$ . The true parameter is  $\beta_1 = 0.1$  and the  $t$ -test for  $\beta_1 = 0$  has  $p$ -values ranging from 0.1 to 0.01. For 25 out of the 100 datasets, the posterior probability that  $X_1$  is included in the model is less than 0.1. Moreover,  $M(\hat{\beta}_1)$  decreases by more than 30%, from 0.133 when  $X_1$  is forced into the model to 0.089 when  $X_1$  is not.

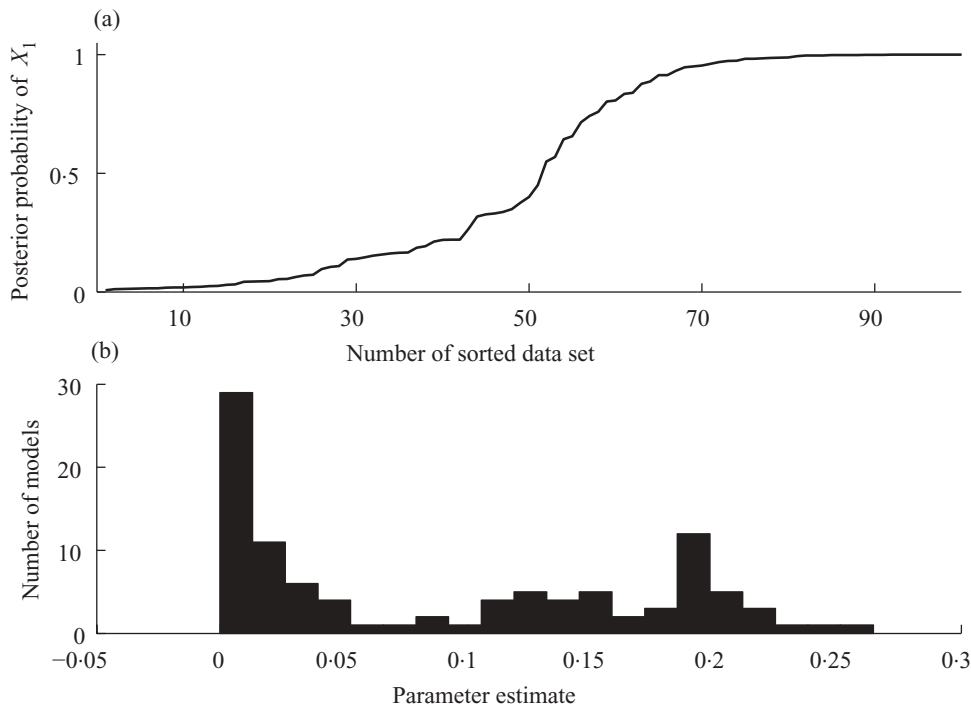


Fig. 3. Bayesian model averaging when the exposure is not forced into the model. (a) Ordered posterior probabilities that  $X_1$  is included in the model for each of the 100 datasets. Posterior probabilities are calculated using the BIC approximation. (b) Histogram of the  $\beta_1$  estimates across the 100 datasets. Data were simulated for  $\rho = 0.7$  with random  $X$ s and  $U$ s with the true value of the parameter  $\beta_1 = 0.1$ .

The standard error  $SE(\hat{\beta}_1)$  increases 46% from 0.0057 to 0.0083. While the central value of the histogram is equal to 0.089 and is close to the true effect,  $\beta_1 = 0.1$ , the shape of the distribution of the  $M(\hat{\beta}_1)$  is multimodal with little probability mass around 0.1. In fact, there is roughly a 50% chance of underestimating the true effect by more than 50% and a 32% chance of overestimating the true effect by more than 50%.

## 5. LEVELS OF PM<sub>10</sub> AND ALL-CAUSE MORTALITY: DETROIT 1987–2000

Estimating the health effects of air pollution poses several methodological challenges related to detecting and quantifying weak signals in large datasets. Commonly used overdispersed Poisson regression models for time series analyses of air pollution and health, such as in Burnett et al. (2001), Dominici et al. (2000), Dominici et al. (2004) and Toulomi et al. (1997), can include hundreds of covariates and be applied to datasets with thousands of observations.

In this section we illustrate our proposed methodology with time series data on air pollution and health. We use time series data for Detroit for the period 1987–2000 to estimate the percent increase in mortality associated with a  $10 \mu\text{g}/\text{m}^3$  increase in particulate matter with diameter smaller than  $10 \mu\text{m}$ , PM<sub>10</sub>, accounting for adjustment uncertainty. This dataset is part of the ongoing National Morbidity Mortality Air Pollution Study which includes time series data for the period 1987–2000 for the 100 largest cities; see Dominici et al. (2003) and Peng et al. (2006). Related R-software and data are available at [www.ihapss.jhsph.edu](http://www.ihapss.jhsph.edu).

We assume the following model for the number of daily deaths  $Y_t \sim \text{Po}(\mu_t, v_t)$ , where  $v_t = \phi\mu_t$ ,  $\mu_t = E(Y_t | \mu_t, \phi)$  is the expected number of deaths on day  $t$ , and  $\phi$  is the overdispersion parameter. We model  $\mu_t$  as

$$\begin{aligned}
\log(\mu_t) = & \text{PM}_{10,t-1} + \text{Dow} + \text{Age}_2 + \text{Age}_3 + O_{3,t-1} \\
& + \text{ns}(T_t, \text{df}_T) + \text{ns}(T_{t,1-3}, \text{df}_T) + \text{ns}(D_t, \text{df}_D) + \text{ns}(D_{t,1-3}, \text{df}_D) \\
& + \text{ns}(t, \text{df}_t) + \text{ns}(t, 4) \times \text{Age}_2 + \text{ns}(t, 4) \times \text{Age}_3,
\end{aligned}$$

where  $\text{PM}_{10,t-1}$  denotes particulate matter of a diameter smaller than 10  $\mu\text{m}$  on day  $t - 1$ , Dow is a categorical variable indicating the day of the week,  $\text{Age}_2$  and  $\text{Age}_3$  are indicator variables of the age groups 65–74 and  $\geq 75$ , and  $O_{3,t-1}$  is the ambient ozone level on day  $t - 1$ . In addition  $T_{t,1-3} = (T_t + T_{t-1} + T_{t-2})/3$  and  $D_{t,1-3} = (D_t + D_{t-1} + D_{t-2})/3$  denote three-day averages of past temperature and dew point temperature, respectively. We include  $\text{ns}(T_t, \text{df}_T)$ ,  $\text{ns}(T_{t,1-3}, \text{df}_T)$ ,  $\text{ns}(D_t, \text{df}_D)$ , and  $\text{ns}(D_{t,1-3}, \text{df}_D)$  to adjust for the potential nonlinear confounding effects of temperature and dew point, where  $\text{ns}(\cdot, \text{df})$  denotes a natural cubic spline with  $\text{df}$  degrees of freedom. We include  $\text{ns}(t, \text{df}_t)$  to adjust for seasonal variations and long term trends in mortality rates due to unmeasured confounders such as influenza epidemics. Finally, we add the interaction terms  $\text{ns}(t, 4) \times \text{Age}_2$  and  $\text{ns}(t, 4) \times \text{Age}_3$  to allow these seasonal variations to be different across age groups. This model specification has been extensively used and discussed in previous analyses of these data; see for example [Dominici et al. \(2000\)](#) and [Kelsall et al. \(1997\)](#). The basic model can be defined by equation (2) with  $\text{df}_T = \text{df}_D = 6$  and  $\text{df}_t = 8$  per year for a total of 112 degrees of freedom over 14 years.

To apply both adjustment uncertainty and Bayesian model averaging estimation, we need to define the full model which includes a large and comprehensive set of all the potential confounders. We assume that the smooth functions of time, temperature and dew point temperature in the full model have degrees of freedom that are twice as large as those specified in the basic model, i.e.  $\text{df}_T = \text{df}_D = 12$ ,  $\text{df}_t = 16$  per year. The first step is to identify good predictors of  $\text{PM}_{10,t-1}$  using the model space,

$$\text{PM}_{10,t-1} = \text{confounders} + \epsilon_t, \quad (2)$$

where the  $\epsilon_t$  are independent  $N(0, \sigma_\epsilon^2)$  random variables. The total number of confounders in this model is 289. As a result of singularities in the design matrix we eliminated 55 spline bases. We obtain the exposure's dominant model class  $\text{DMC}_X = (U_{0,X}, \dots, U_{234,X})$  for model (2) using the stochastic search algorithm described in Appendix A. We used 500 iterations per orbit for a total of  $500 \times 234 = 117\,000$  iterations with an average computation time 1.62 seconds/iteration on a PC with 3.4 GHz CPU and 3.4 Gb RAM.

Figure 4(a) shows the deviance differences  $D(X | Z, U_{m,X}) - D(X | Z, U_{m+1,X})$ . Figure 4(b) shows the point estimates, and the 95% confidence intervals for the exposure effect for  $m = 0, \dots, 234$ . The horizontal dotted line is placed at 4. The deviance differences between neighbouring orbits become negligible starting roughly at about orbit 132, which closely agrees with the stabilization of the exposure effect point estimate. Our implementation provides the full list of confounders in the dominant model corresponding to each orbit. For example, orbit 132 includes  $O_3$  and Dow, but does not include the indicators for age categories. It also includes some, but not all, bases of the natural cubic splines. Interestingly,  $O_3$  had a  $p$ -value of 0.62 in the full outcome model and would be discarded by any model selection or averaging procedure focusing on the outcome models.

In the second stage we include in the outcome model the 132 covariates identified above. The outcome model space generated by the remaining  $234 - 132 = 102$  potential confounders is then explored. We used 500 iterations per orbit.

Figure 5(a) shows the deviance differences for neighbouring orbits in the outcome model dominant class. The horizontal dotted line is placed at 4. Figure 5(b) shows the point estimates and

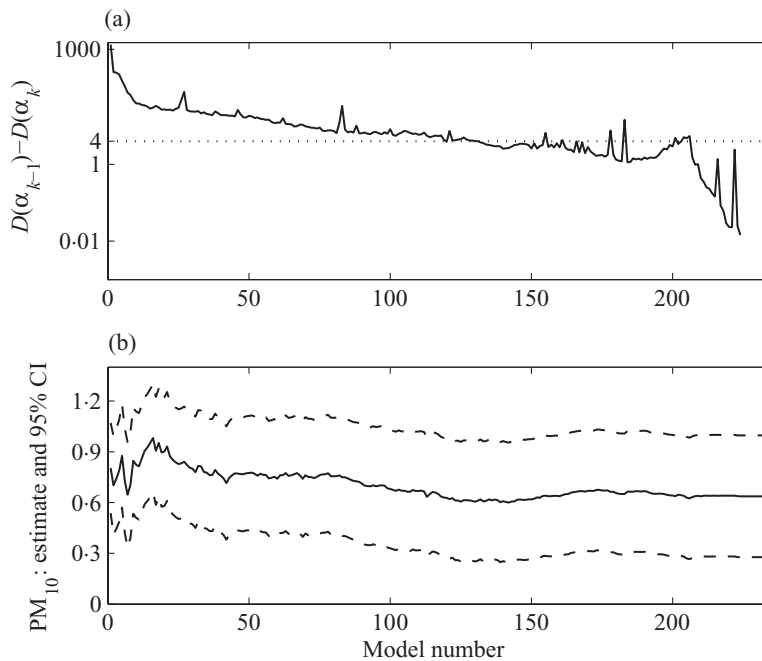


Fig. 4. Detroit 1987–2000 data. (a) Deviance differences between the exposure dominant models on orbits  $k - 1$  and  $k$  versus model number with complexity increasing from left to right. The horizontal dotted line is placed at 4. (b) Estimated percentage increase  $\times 10^{-3}$  in all-cause mortality associated with a  $10\mu\text{g}/\text{m}^3$  of with 95% confidence intervals corresponding to each exposure dominant model.

95% confidence intervals of the exposure effect. Figure 5 (a) shows that, among the first 15 models, approximately, the deviance differences are large. This suggests that, in addition to the 132 variables forced into the model, there are 15 to 20 more variables that may be predictive of mortality. However, Fig. 5 (b) shows remarkable stability of point estimates across all models in the dominant model class, suggesting that none of the 102 additional variables is necessary to obtain an unbiased parameter estimate. This indicates that selecting good predictors of the  $\text{PM}_{10}$  provides almost complete confounding adjustment. These results are consistent with the results in Dominici et al. (2004), where interest focused on estimating the exposure effect while adjusting for a smooth confounding function.

Another interesting feature of these results is that the standard errors of the exposure effect estimates increase by only 2% from orbit 0 to orbit 102. These models show that  $10\mu\text{g}/\text{m}^3$  increase in  $\text{PM}_{10}$  at lag 1 is associated with a 0.64 percent increase in all-cause mortality with 95% confidence interval equal to (0.30, 0.98). The confidence intervals are based on the dominant model for orbit 20 in the outcome dominant model class. Under the basic and the full models, the point estimates and 95% confidence intervals of the  $\text{PM}_{10,t-1}$  coefficient multiplied by 1000,  $\beta$ , are equal to 0.66, with interval (0.30, 1.02), and 0.64, with interval (0.28, 1.00), respectively. These estimates indicate the percent increase in all-cause mortality for  $10\mu\text{g}/\text{m}^3$  increase in previous day  $\text{PM}_{10}$ .

## 6. DISCUSSION

We now review the strengths and limitations of our approach. Our estimation method explores both the exposure and outcome model spaces and identifies a set of models that provide a consensus estimate of the exposure effect. Our approach is intuitive because it is designed to

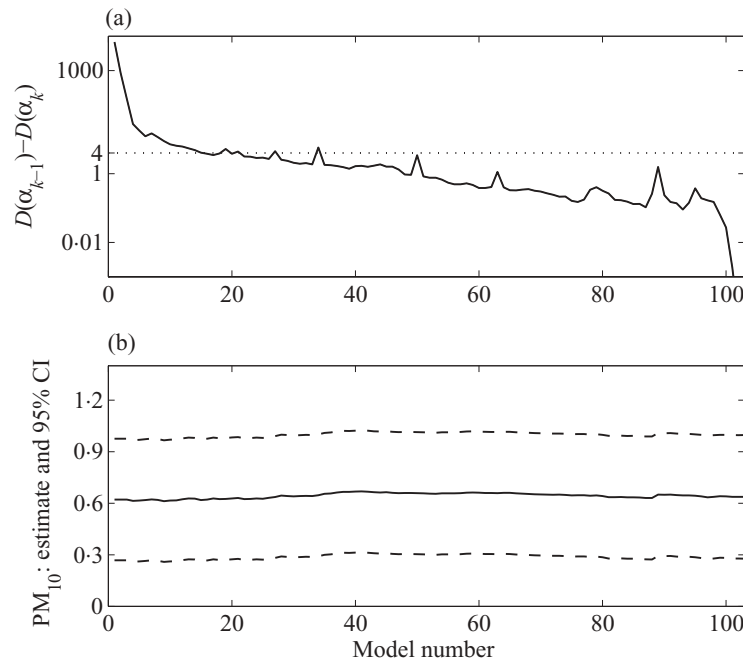


Fig. 5. Detroit 1987–2000 data. (a) Deviance differences between the outcome dominant models on orbits  $k-1$  and  $k$  versus model number with complexity increasing from left to right. The horizontal dotted line is placed at 4. (b) Estimated percentage increase  $\times 10^{-3}$  in all-cause mortality associated with a  $10 \mu\text{g}/\text{m}^3$  of  $PM_{10t-1}$  with 95% confidence intervals corresponding to each exposure dominant model.

answer directly the question ‘What is the effect of  $X$  on  $Y$ ?’, by dividing it into  $M+1$  simpler questions of the type ‘What is the effect of  $X$  on  $Y$  when we control for the  $m \in \{0, \dots, M\}$  covariates,  $U$ ?’ We achieve a faster stabilization of the exposure effect estimate by identifying important predictors of the exposure and forcing them in the outcome models instead of exploring only the outcome space directly.

Scientific knowledge can easily be incorporated in the algorithm. For example, known confounders can be forced into all models or be revalidated during the exploration of the exposure model space. In many applications, confounders are partitioned into scientifically relevant groups, such as biological, environmental or socio-economic factors. Our method can focus first on the biological risk factors, identify the important confounders and iterate the procedure with the other groups.

A limitation of our methodology is the need to specify a candidate set of predictors that includes all necessary confounders. In some circumstances, such specification might be challenging. However, the choice of a large set of potential predictors is a problem that our method shares with many methods that account for model uncertainty, including Bayesian model averaging. Scientific knowledge and data availability usually guide this choice while external studies and additional data may validate this choice and suggest alternative sets. Another limitation of the algorithm is that stabilization may be reached at a model that is not scientifically convincing. However in this circumstance, because the estimate is a consensus over a set of models, we can report our results under any scientifically meaningful model in the consensus set.

Our results highlight serious limitations of IC-based Bayesian model averaging estimators. Bayesian model averaging for parameter estimation can assign a considerable probability mass to models that include only a subset of the necessary confounders, thus biasing the exposure effect estimation. In addition, IC-based Bayesian model averaging can either over- or underestimate the statistical variance of the exposure effect estimate. Overestimation can occur when very different exposure effect estimates are averaged across models that provide different confounder adjustments. Underestimation can occur when a considerable probability mass is assigned to models that do not include important confounders and thus produce biased estimates of the exposure effect with small standard errors. Finally, the implementation of Bayesian model averaging depends on the calculation of posterior model probabilities, which depend on the prior distribution. It is well known, see for example [Lindley \(1957\)](#), that diffuse priors favour the most parsimonious model and informative priors would need to be specified. However, in time series studies of air pollution and health with a large number of correlated covariates, specifying informative and biologically meaningful priors is very challenging.

Currently, accounting for model and adjustment uncertainty when estimating health risk associated with exposure to air pollution is at the centre of a heated policy debate; see for example [Samet et al. \(2000\)](#), [Koop & Tole \(2004\)](#), [Moolgavkar \(2005\)](#) and [Thomas et al. \(2007\)](#). Recently, Bayesian model averaging has been applied in time series analysis of air pollution and health to account for model uncertainty. [Koop & Tole \(2004\)](#) applied Bayesian model averaging to daily time series data for the period 1992–1997 using the daily number of deaths as outcome and daily levels of particulate matter as exposure for Toronto, Canada. They used a linear model and prior model probabilities calibrated so that the posterior model probabilities are similar to the ones obtained by use of the BIC approximation. They concluded that ‘standard deviations for air pollution-mortality impacts become very large when model uncertainty is incorporated into the analysis. The authors argue that Bayesian model averaging should be used instead of expert-based model selection.

Our paper indicates several reasons for strong scepticism about the [Koop & Tole \(2004\)](#) analyses. First, their approach produces biased estimators of the exposure effect. One source of bias is that exposure variables were not forced into the outcome models and the estimated exposure effects were averaged with zero; see § 4.3 in this paper. A second source of bias is that estimated exposure effects were averaged across models that did not include all the relevant confounders. Moreover, their priors on coefficients are so diffuse that even variables with strong effects are likely to be excluded with high probability; see definitions B4 and B5 in their Appendix B. To be specific, the authors used Zellner’s g-priors [Zellner \(1983\)](#), with the degree of diffusion controlled by  $C = \max(M^2, n) \geq n$ . When  $C = n$  this is essentially equivalent to using a BIC approximation of the posterior model probabilities. However, [Koop & Tole \(2004\)](#) used  $M^2 \gg n$ , which can lead to ignoring even stronger signals. Finally, the authors used linear models to describe the distribution of low-count time series data, an inappropriate choice for describing the sampling variability.

#### ACKNOWLEDGEMENT

Funding for Ciprian Crainiceanu was provided by the Johns Hopkins Faculty Innovation Award. Funding for Francesca Dominici was provided by a grant from the U.S. National Institute for Environmental Health Sciences and by the U.S. National Institute of Health Sciences Center in Urban Environmental Health.

## APPENDIX A

## Technical details for model space search

Likelihood maximization for models in  $\mathcal{M}$  is usually fast and robust for large datasets and the dominant model in the orbit  $m$ ,  $\alpha_m$ , can be obtained using the following hybrid algorithm that contains an initial deterministic, and a second, stochastic, phase. The deterministic phase starts with a model,  $\alpha_0$  say, in the  $(m-1)$ th orbit  $O_{m-1}$  and searches for the one additional covariate that provides the largest increase in the likelihood. The model obtained by attaching the newly identified covariate to the model from orbit  $m-1$  is used to initiate the stochastic search. The stochastic search selects at random one covariate in  $U$  that is already in the model,  $U_i$  say, and one that is not,  $U_j$  say, and constructs a new model,  $\alpha_1$  say, by replacing  $U_i$  by  $U_j$ . The new model becomes the current model with probability  $p_{0 \rightarrow 1} = \min(1, L_1/L_0)$ , where  $L_0$  and  $L_1$  are the maximum likelihoods of the models  $\alpha_0$  and  $\alpha_1$  respectively. Otherwise a new pair of covariates is simulated and the procedure is iterated. This algorithm is fast because at every step it requires at most one likelihood calculation for each model proposed to be visited. By keeping records of the maximum likelihood for each model previously proposed, the likelihood of a new model is often known and the model does not need to be refitted.

## APPENDIX B

## Proofs

*Proof of Theorem 1.* (i)  $|\hat{\beta}_n^{AU} - \beta^*| \leq |\hat{\beta}_n^{AU} - \hat{\beta}_n^{\alpha_F}| + |\hat{\beta}_n^{\alpha_F} - \beta^*| < \epsilon/h_n + |\hat{\beta}_n^{\alpha_F} - \beta^*|$ . This shows that  $\hat{\beta}_n^{AU}$  is consistent.

(ii) Note that  $\hat{\beta}_{W_n} \rightarrow \sum_{\alpha} w(\alpha)\beta^{\alpha} = \sum_{\alpha \in I} w(\alpha)\beta^{\alpha} + \sum_{\alpha \in I^c} w(\alpha)\beta^{\alpha}$  almost surely under  $\alpha^*$ . If we write  $\beta^* = \sum_{\alpha \in I} w(\alpha)\beta^* + \sum_{\alpha \in I^c} w(\alpha)\beta^*$ , the asymptotic bias of  $\hat{\beta}_{W_n}$  is  $\sum_{\alpha \in I} w(\alpha)\beta^{\alpha} - \sum_{\alpha \in I} w(\alpha)\beta^*$ .

(iii) By construction,  $\hat{v}_n(\hat{\alpha}_n^{AU}) \leq \hat{v}_n(\alpha_F)$ , which implies that  $\limsup_{n \rightarrow \infty} \{n\hat{v}_n(\hat{\alpha}_n^{AU})\} \leq \limsup_{n \rightarrow \infty} \{n\hat{v}_n(\alpha_F)\} = v(\alpha_F)$ . Under standard regularity assumptions  $h_n(\hat{\beta}_n^{\alpha_0} - \beta^*) \rightarrow 0$  almost surely. Let  $A^h$  be the convergence set and denote by  $A^{\alpha_0}$  and  $A^{\alpha_F}$  the convergence sets for  $n\hat{v}_n(\alpha_0)$  and  $n\hat{v}_n(\alpha_F)$  respectively.

For every  $\omega \in A^h$  there exists  $N_{\omega, \epsilon}$ , such that, for every  $n \geq N_{\omega, \epsilon}$ , we have  $h_n|\hat{\beta}_n^{\alpha_0} - \beta^*| < \epsilon$ . Thus, for every  $n \geq N_{\omega, \epsilon}$ ,  $\alpha_0$  is one of the models in the set over which the minimum is taken in  $\hat{\alpha}_n^{AU} = \arg \min\{\hat{v}_n(\alpha) : |\hat{\beta}_n^{\alpha} - \hat{\beta}_n^{\alpha_F}| < \epsilon/h_n\}$ . Thus,  $n\hat{v}_n(\hat{\alpha}_n^{AU}) \leq n\hat{v}_n(\alpha_0)$  and, for every  $\omega \in A^h \cap A^{\alpha_0} \cap A^{\alpha_F}$ ,  $\limsup_{n \rightarrow \infty} \{n\hat{v}_n(\hat{\alpha}_n^{AU})\} \leq v(\alpha_0) < v(\alpha_F)$ .  $\square$

*Proof of Theorem 2.* We show first that for every model  $\alpha$  such that  $\alpha^* \subset \alpha$  we have  $\lim_{n \rightarrow \infty} w_n(\alpha^*)/w_n(\alpha) = \infty$ . Indeed,  $w_n(\alpha^*)/w_n(\alpha) = \exp[\log\{\rho(\alpha^*)/\rho(\alpha)\} + (p_{\alpha} - p_{\alpha^*})\log(n)/2]$ , where  $\rho(\alpha)$  and  $p_{\alpha}$  denote the maximum likelihood and the number of parameters for model  $\alpha$ , respectively. The result follows because  $-2\log\{\rho(\alpha^*)/\rho(\alpha)\}$  converges weakly to a  $\chi^2$  distribution with  $p_{\alpha} - p_{\alpha^*}$  degrees of freedom under the true model,  $\alpha^*$ .

We now show that  $\lim_{n \rightarrow \infty} w_n(\alpha^{**})/w_n(\alpha^*) = \infty$  under the true model  $\alpha^*$ . Denote by  $\theta_1$  the true parameter vector under  $\alpha^*$  and construct  $\theta_0$ , the vector obtained from  $\theta_1$  by setting to zero the entries corresponding to variables in  $S$ . Then we can write  $\theta_1 = \theta_0 + n^{-1/2}e_1$ , where  $e_1$  is a vector with all entries equal to zero with the exception of positions corresponding to variables in  $S$ . For the variable  $U_m$  in  $S$  its corresponding entry in  $e_1$  is  $c_m$ . Note that

$$w_n(\alpha^{**})/w_n(\alpha^*) \geq \exp[\log\{\rho(\theta_0, \hat{\sigma}_{\epsilon})/\rho(\alpha^*)\} + (p_{\alpha^*} - p_{\alpha^{**}})\log(n)/2],$$

where  $\rho(\theta_0, \hat{\sigma}_{\epsilon})$  is the likelihood of model  $\alpha^*$  evaluated at  $\theta_0$  and  $\hat{\sigma}_{\epsilon}$ , the maximum likelihood estimator of  $\sigma_{\epsilon}$  under model  $\alpha^*$ . It is enough to prove that  $\log\{\rho(\theta_0, \hat{\sigma}_{\epsilon})/\rho(\alpha^*)\}$  converges weakly to a random variable. Consider the following matrix format of the model  $\alpha^*$ ,  $Y = \Lambda\lambda + U\delta + \epsilon$ , where  $\lambda$  contains all parameters that do not depend on  $n$  and  $\delta$  contains all the parameters corresponding to variables in  $S$ . Let  $W = [\Lambda \mid U]$  and  $\delta = cn^{-1/2}$ , where  $c$  is a fixed parameter vector. Denote by  $\hat{\theta}_n$  the maximum likelihood

estimator of  $\theta_1$  under model  $\alpha^*$  and note that  $\rho(\alpha^*) = \rho(\hat{\theta}_n, \hat{\sigma}_\epsilon)$ ,  $\theta_1 = (\lambda, cn^{-1/2})$  and  $\hat{\theta}_n = (\hat{\lambda}, \hat{c}n^{-1/2})$ . Then

$$\log\{\rho(\theta_0, \hat{\sigma}_\epsilon)/\rho(\hat{\theta}_n, \hat{\sigma}_\epsilon)\} = -0.5(\theta_0 - \hat{\theta}_n)'(W^t W)(\theta_0 - \hat{\theta}_n)/\hat{\sigma}_\epsilon^2,$$

and

$$n^{1/2}(\hat{\theta}_1 - \theta_1) = n^{1/2}(W^t W)^{-1}W^t(Y - \Lambda\lambda - Ucn^{-1/2}),$$

which is distributed as  $N\{0, n\sigma_\epsilon^2(W^t W)^{-1}\}$ . Under the standard assumption that  $n(W^t W)^{-1} \rightarrow \Sigma^{-1}$  it follows that  $n^{1/2}(\hat{\theta}_1 - \theta_1)$  converges weakly to  $N(0, \sigma_\epsilon^2 \Sigma^{-1})$ . Because  $\theta_0 - \hat{\theta}_n = n^{-1/2}\{n^{1/2}(\theta_1 - \hat{\theta}_n) - e_1\}$  it follows that, under  $\alpha^*$ ,  $\log\{\rho(\theta_0, \hat{\sigma}_\epsilon)/\rho(\hat{\theta}_n, \hat{\sigma}_\epsilon)\}$  converges weakly to  $-0.5(X - \Sigma^{1/2}e_1/\sigma_\epsilon)'(X - \Sigma^{1/2}e_1/\sigma_\epsilon)$ . Here  $\Sigma^{1/2}$  is the unique positive square root of  $\Sigma$ ,  $X \sim N(0, I_{p^*})$  and  $p^*$  is the number of parameters in the true model,  $\alpha^*$ .  $\square$

## REFERENCES

- BELL, M., McDERMOTT, A., ZEGER, S., SAMET, J. & DOMINICI, F. (2004). Ozone and mortality in 95 U.S. cities from 1987 to 2000. *J. Am. Med. Assoc.* **292**, 2372–8.
- BURNETT, R., MA, R., JERRETT, M., GOLDBERG, M., CAKMAK, S., POPE, A. & KREWSKI, D. (2001). The spatial association between community air pollution and mortality: a new method of analyzing correlated geographic cohort data. *Envir. Health Perspect.* **109**, 375–80.
- CLYDE, M. (2000). Model uncertainty and health effects studies for particulate matter. *Environmetrics* **11**, 745–63.
- DOMINICI, F., McDERMOTT, A., DANIELS, M., ZEGER, S. L. & SAMET, J. M. (2003). *A Special Report to the Health Effects Institute on the Revised Analyses of the NMMAPS II Data*. Cambridge, MA: The Health Effects Institute.
- DOMINICI, F., McDERMOTT, A. & HASTIE, T. (2004). Improved semi-parametric time series models of air pollution and mortality. *J. Am. Statist. Assoc.* **99**, 938–48.
- DOMINICI, F., SAMET, J. M. & ZEGER, S. L. (2000). Combining evidence on air pollution and daily mortality from the twenty largest U.S. cities: A hierarchical modeling strategy (with Discussion). *J. R. Statist. Soc. A* **163**, 263–302.
- DRAPER, D. (1995). Assessment and propagation of model uncertainty (with Discussion). *J. R. Statist. Soc. B* **57**, 45–97.
- HOETING, J., MADIGAN, D., RAFTERY, A. & VOLINSKY, C. (1999). Bayesian model averaging: a tutorial. *Statist. Sci.* **14**, 382–417.
- KAISER, J. (2005). Mounting evidence indicts fine-particles pollution. *Science* **307**, 1858–61.
- KATSOUYANNI, K., TOULOMI, G., SAMOLI, E., GRYPARIS, A., LeTERTRE, A., MONOPOLIS, Y., ROSSI, G., ZMIROU, D., BALLESTER, F., BOUMGHAR, A. & ANDERSON, H. R. (2001). Confounding and effect modification in the short-term effects of ambient particles on total mortality: Results from 29 European cities within the APHEA2 project. *Epidemiology* **12**, 521–31.
- KELSALL, J., SAMET, J. M. & ZEGER, S. L. (1997). Air pollution, and mortality in Philadelphia, 1974–1988. *Am. J. Epidemiol.* **146**, 750–62.
- KOOP, G. & TOLE, L. (2004). Measuring the health effects of air pollution: to what extent can we really say that people are dying of bad air. *J. Envir. Econ. Manag.* **47**, 30–54.
- LINDLEY, D. (1957). A statistical paradox. *Biometrika* **44**, 187–92.
- MADIGAN, D. & RAFTERY, A. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Statist. Assoc.* **89**, 1535–46.
- MADIGAN, D., RAFTERY, A., VOLINSKY, C. & HOETING, J. (1996). Bayesian model averaging. In *Integrating Multiple Learned Models, (IMLM-96)*, Ed. P. Chan, S. Stolfo and D. Wolpert, pp. 77–83.
- MOOLGAVKAR, S. (2005). A review and critique of the EPA's rationale for a fine particle standard. *Regulat. Toxicol. Pharmacol.* **42**, 123–44.
- NATIONAL RESEARCH COUNCIL (1998). *Research Priorities for Airborne Particulate Matter. Part I. Immediate Priorities and a Long-Range Research Portfolio*. Washington, DC: National Academy Press.
- NATIONAL RESEARCH COUNCIL (1999). *Research Priorities for Airborne Particulate Matter. Part II. Evaluating Research Progress and Updating the Portfolio*. Washington, DC: National Academy Press.
- NATIONAL RESEARCH COUNCIL (2001). *Research Priorities for Airborne Particulate Matter. Part III. Early Research Progress*. Washington, DC: National Academy Press.
- PENG, R., DOMINICI, F. & LOUIS, T. A. (2006). Model choice in multi-site time series studies of air pollution and mortality (with Comments). *J. R. Statist. Soc. A* **169**, 179–203.
- RAFTERY, A. (1994). Bayesian Model Selection in Social Research (with Comments). *Sociol. Methodol.* **25**, 111–63.

- SAMET, J. M., ZEGER, S. L., DOMINICI, F., CURRIERO, F., COURSAK, I., DOCKERY, D., SCHWARTZ, J. & ZANOBBETTI, A. (2000). *The National Morbidity, Mortality, and Air Pollution Study (HEI Project No. 96–7): Morbidity and Mortality from Air Pollution in the United States*. Cambridge, MA: Health Effects Institute.
- SHAO, J. (1998). *Mathematical Statistics*. New York: Springer.
- THOMAS, D. C., JERRETT, M., KUENZLI, N., LOUIS, T., DOMINICI, F., ZEGER, S., SCHWARZ, J., BURNETT, R. T., KREWSKI, D. & BATES, D. (2007). Bayesian model averaging in time-series studies of air pollution and mortality. *J. Toxicol. Envir. Health* **70**, 311–5.
- TOULOMI, G., KATSOUYANNI, K., ZMIROU, D. & SCHWARTZ, J. (1997). Short-term effects of ambient oxidant exposure on mortality: A combined analysis within the APHEA project. *Am. J. Epidemiol.* **146**, 177–83.
- ZELLNER, A. (1983). Applications of Bayesian analysis and econometrics. *Statistician* **32**, 23–34.

[Received September 2006. Revised September 2007]