Effect Estimation in Presence of Confounding

**Goal:**
to understand the effect of $X$ on $Y$
for fixed levels of the confounder(s) $U$. 
$E\{y_i\} = \beta^\alpha x_i + \sum_{m=1}^{M} \alpha_m \gamma_m^\alpha u_{im}$

- $u$’s are candidate confounders, including:
  - measured candidate confounders
  - basis elements to describe functions of candidate confounders
- $\alpha$’s are binary indicators of inclusion of the corresponding candidate confounder or basis element
Selecting Confounders and Estimating Effects

\[ E\{y_i\} = \beta^\alpha x_i + \sum_{m=1}^{M} \alpha_m \gamma_m^\alpha u_{im} \]

- **Standard Approach**
  Select a Model (i.e. fix or estimate \( \alpha \))
  Conditionally on \( \alpha \), estimate \( \beta \)

- **Sensitivity Analysis**
  Present \( \beta \)'s obtained over a range of selections of \( \alpha \)
  Select \( \alpha \) so that sensitivity around it is low

- **Accounting for Uncertainty on Selection Process**
  Treat \( \alpha \) as an unknown parameter; integrate it out.
Sensitivity Analysis: Effect Stabilization

Sensitivity of the national average lag effect of PM$_{10}$ on mortality to different statistical models to adjust for confounding (NMMAPS 1987-2000)

Using information at the short and long (trend and seasonality) time scales

weak moderate strong

Using information only at the very short time scales

% increase in mortality for 10 µg/m$^3$ increase in PM$_{10}$

Different statistical models to adjust for confounding

Peng Dominici Louis JRSSC 2006
Asymptotic Bias and Model Size

\[ y_t = \beta x_t + f(t) + \epsilon_t \]
\[ x_t = g(t) + \eta_t \]
\[ f(t) = \sum_{l=1}^{r} h_l(t) \gamma_l^f \]
\[ g(t) = \sum_{l=1}^{q} h_l(t) \gamma_l^g \]

- Modeling \( f(t) \) including fewer than \( \min(r, q) \) leads to estimating the incorrect parameter.
- Modeling \( f(t) \) with sufficient degrees of freedom to represent the relationship between \( x_t \) and \( t \) leads to an asymptotically unbiased estimate of \( \beta \), even if \( r \neq q \)

*Speckman JRSSB 1988*

*Dominici McDermott & Hastie JASA, 2004*
"STEADy" algorithm

Stage 1  Identify predictors of the exposure $X$ by modeling $[X|U]$

Stage 2  Include those predictors in the estimation of $\beta$.
          Identify predictors of Y conditionally to $X$ and to the confounders identified in stage 1

- **Visualization:**
  graph the estimates of the effect and its c.i. as a function of model size. For each model size, select the best model by maximum likelihood.

- **Estimation:**
  Set criteria for model size based on effect stabilization.

*Crainiceanu Dominici Parmigiani Bka, 2008*
Figure 4: Results for Detroit 1987–2000. Top panel: deviance differences between the exposure dominant models on orbit $k$ and on orbit $k + 1$ plotted against model numbers with complexity increasing from left to right. The horizontal red lines are placed at the BIC and AIC penalties $\log(7464) = 8.92$ and 2, respectively. Bottom graph: Estimated percent increase in all cause mortality associated with a $10 \mu g/m^3$ of $PM_{10}$ with 95% confidence intervals corresponding to each exposure dominant model.
Measuring the health effects of air pollution: to what extent can we really say that people are dying from bad air?

Gary Koop\textsuperscript{a,}\textsuperscript{*} and Lise Tole\textsuperscript{b}

\textsuperscript{a}Department of Economics, University of Leicester, Leicester LE1 7RH, UK
\textsuperscript{b}Department of Geography, University of Leicester, Leicester LE1 7RH, UK

Received 3 July 2002; revised 21 January 2003
The basic idea behind Bayesian model averaging can be summarized as follows: Suppose a researcher is entertaining $R$ possible models, denoted by $M_1, \ldots, M_R$, in order to learn about a parameter of interest, $\theta$ (e.g. the effect of a pollutant on health). For the models in this paper, it is straightforward to use the data to calculate the probability that a model is a correct one. That is, $p(M_r|Data)$ can be calculated for $r = 1, \ldots, R$. It is also straightforward to calculate a point estimate of $\theta$ in every model. We take the posterior mean, $E(\theta|Data, M_r)$, as this point estimate. According to the rules of conditional expectation, it follows that:

$$E(\theta|Data) = \sum_{r=1}^{R} p(M_r|Data)E(\theta|Data, M_r). \quad (2.1)$$

In words, the overall point estimate of $\theta$ is the weighted average of the point estimates in every model. The weights in the weighted average are the posterior model probabilities, $p(M_r|Data)$ for $r = 1, \ldots, R$. This same logic applies to functions of $\theta$. For instance, we can use:
Standard Implementations:

\[ p\{\beta|\text{Data}\} = p\{\beta|\alpha, \text{Data}\}p\{\alpha|\text{Data}\} \]

- **Default Conjugate**
  - \(p(\alpha)\) is uniform
  - \(p(\beta, \gamma|\alpha)\) is conjugate and vague
  - \(p(\alpha|\text{Data})\) is obtained by analytically integrating out the \(\beta, \gamma\).

- **BIC-based**
  - \(p(\alpha|\text{Data})\) is approximated using normalized BIC scores

*Raftery 1997*
Example 1

\[ Y_t = 0.1 \, X_t + 0.1 \, U_{1t} + 0.1 \, U_{2t} + \epsilon_t \]

- \( T = 1000 \).
- \((X_i, U_{1i}, U_{2i})\) are normal vectors with mean zero and a covariance matrix,

\[
\begin{bmatrix}
1 & 0.7 & 0 \\
0.7 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix}
\]

- The set of potential confounders \( U \) includes \( U_1, U_2 \) as well as 49 additional independent \( N(0, 1) \) random variables.
**Example 1: Simulation Results**

<table>
<thead>
<tr>
<th></th>
<th>True model</th>
<th>STEADy</th>
<th>BMA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>X in</td>
<td></td>
</tr>
<tr>
<td>$\text{M}(\hat{\beta})$</td>
<td>0.108</td>
<td>0.107</td>
<td>0.147</td>
</tr>
<tr>
<td>$\text{SE}(\hat{\beta})$</td>
<td>0.045</td>
<td>0.045</td>
<td>0.044</td>
</tr>
<tr>
<td>SE of $\hat{\beta}$ across data sets</td>
<td>0.045</td>
<td>0.045</td>
<td>0.050</td>
</tr>
<tr>
<td>$\text{MSE}(\hat{\beta})$</td>
<td>0.002</td>
<td>0.002</td>
<td>0.005</td>
</tr>
<tr>
<td>Coverage rate of 50% C.I./P.I.</td>
<td>14/25</td>
<td>14/25</td>
<td>7/25</td>
</tr>
</tbody>
</table>

**Table:** The target coverage rate is 12.5/25.
BMA with X in: Prediction versus Effect Estimation

does not make sense and does not match example
Example 1: Simulation Results, continued

<table>
<thead>
<tr>
<th></th>
<th>True model</th>
<th>STEADy</th>
<th>BMA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>X in</td>
<td>X in or out</td>
</tr>
<tr>
<td>$M(\hat{\beta})$</td>
<td>0.108</td>
<td>0.107</td>
<td>0.147</td>
</tr>
<tr>
<td>$SE(\hat{\beta})$</td>
<td>0.045</td>
<td>0.045</td>
<td>0.044</td>
</tr>
<tr>
<td>SE of $\hat{\beta}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>across data sets</td>
<td>0.045</td>
<td>0.045</td>
<td>0.050</td>
</tr>
<tr>
<td>$MSE(\hat{\beta})$</td>
<td>0.002</td>
<td>0.002</td>
<td>0.005</td>
</tr>
<tr>
<td>Coverage rate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>of 50% C.I./P.I.</td>
<td>14/25</td>
<td>14/25</td>
<td>7/25</td>
</tr>
</tbody>
</table>

Table: The target coverage rate is 12.5/25.
BMA with X in or out: ugly posterior
Comparing Effect Estimates Across Models

- In effect estimation, when a true confounder is added or removed from the regression model, the interpretation of the exposure coefficient changes.

- When a model includes all true confounders and one adds an additional variable that is not associated with $X$ or that is not associated with $X$ nor $Y$, the interpretation of the exposure coefficient does not change.

- In prediction, the predicted quantities typically maintain the same interpretation across models.
- **MINIMAL MODEL** $\alpha^*$
  smallest model that includes all the confounders.

- **TRUE EFFECT** $\beta^* = \beta^{\alpha^*}$

- All models that contain the predictors in the minimal model will provide estimates of the exposure effect that are also interpretable as estimate of $\beta^*$

  If $\alpha \subseteq \alpha^*$ then $\beta^\alpha = \beta^*$
\[
\sum_{\alpha^Y} P(\beta|\alpha^Y, Y)P(\alpha^Y|Y) = \\
= \sum_{\alpha^Y \supseteq \alpha_*^Y} P(\beta_*|\alpha^Y, Y)P(\alpha^Y|Y) + \sum_{\alpha^Y \not\supseteq \alpha_*^Y} P(\beta|\alpha^Y, Y)P(\alpha^Y|Y).
\]
\[
\sum_{\alpha^Y} P(\beta | \alpha^Y, Y) P(\alpha^Y | Y) = \\
\sum_{\alpha^Y} P(\beta^{\alpha^Y} | \alpha^Y, Y) P(\alpha^Y | Y) = \\
= \sum_{\alpha^Y \supseteq \alpha_*^Y} P(\beta_* | \alpha^Y, Y) P(\alpha^Y | Y) + \sum_{\alpha^Y \nsubseteq \alpha_*^Y} P(\beta^{\alpha^Y} | \alpha^Y, Y) P(\alpha^Y | Y).
\]
\[ \sum_{\alpha^Y} P(\beta|\alpha^Y, Y) P(\alpha^Y|Y) = \]

\[ \sum_{\alpha^Y} P(\beta^*|\alpha^Y, Y) P(\alpha^Y|Y) = \]

\[ = \sum_{\alpha^Y \supseteq \alpha^*_Y} P(\beta^*|\alpha^Y, Y) P(\alpha^Y|Y) + \sum_{\alpha^Y \nsubseteq \alpha^*_Y} P(\beta^*|\alpha^Y, Y) P(\alpha^Y|Y). \]
BAC (Bayesian Adjustment for Confounding)

\[ E\{X_t\} = \sum_{m=1}^{M} \alpha^X_m \delta^X_m U_{tm} \]

\[ E\{Y_t\} = \beta^Y \alpha^X_t + \sum_{m=1}^{M} \alpha^Y_m \delta^Y_m U_{tm} \]

- specify a prior distribution \( p(\alpha^Y, \alpha^X) \)
- assume prior dependence \( p(\alpha^Y | \alpha^X) \)

Wang, Parmigiani, Dominici, Biometrics 2012
Comparison of estimates of $\beta$ from four methods along with the true model in the first simulation scenario. The target coverage rate is 12.5/25.

<table>
<thead>
<tr>
<th></th>
<th>True model</th>
<th>BAC</th>
<th>STEADy</th>
<th>BMA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$X$ in</td>
</tr>
<tr>
<td>$M(\hat{\beta})$</td>
<td>0.108</td>
<td>0.108</td>
<td>0.107</td>
<td>0.147</td>
</tr>
<tr>
<td>$SE(\hat{\beta})$</td>
<td>0.045</td>
<td>0.045</td>
<td>0.045</td>
<td>0.044</td>
</tr>
<tr>
<td>SE of $\hat{\beta}$ across data sets</td>
<td>0.045</td>
<td>0.045</td>
<td>0.045</td>
<td>0.050</td>
</tr>
<tr>
<td>$MSE(\hat{\beta})$</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.005</td>
</tr>
<tr>
<td>Coverage rate of 50% C.I.</td>
<td>14/25</td>
<td>15/25</td>
<td>14/25</td>
<td>7/25</td>
</tr>
</tbody>
</table>
prior distribution on $\alpha^Y|\alpha^X$ and on $\alpha^X|\alpha^Y$ such that

$$\frac{P(\alpha^Y_m=1|\alpha^X_m=1)}{P(\alpha^Y_m=0|\alpha^X_m=1)} = \omega \quad \frac{P(\alpha^Y_m=1|\alpha^X_m=0)}{P(\alpha^Y_m=0|\alpha^X_m=0)} = 1$$

where $1 < \omega < \infty$

- we assign high probabilities for predictors selected by the exposure model to be included in the outcome model
- we assign low probabilities for predictors not selected by the outcome model to be included in the exposure model
To implement the Markov chain Monte Carlo (MCMC) algorithm, we make the following assumptions:

A1: $(\beta^{\alpha^Y}, X)$ are independent of $\alpha^Y$ given $(\alpha^X, \tilde{Y})$, where $\tilde{Y} = Y - \beta^{\alpha^Y} X$

A2: $X$ is independent of $\alpha^Y$ given $\alpha^X$

A3: $(\beta^{\alpha^Y}, Y)$ are independent of $\alpha^X$ given $(\alpha^Y, X)$

A4: $\tilde{Y}$ is independent of $\alpha^X$ given $\alpha^Y$
Why we think BAC works

- Our goal is to estimate $\beta^*$

$$P(\beta^* | D) = \sum_{\alpha^Y} p(\beta^* | \alpha^Y, D)p(\alpha^Y | D),$$

- When $\omega$ is large, the prior $p(\alpha^Y | \alpha^X)$ leads to $p(\alpha^Y | D)$ such that:
  - is close to one for models $\alpha^Y \supseteq \alpha^*$
  - is close to zero for models $\alpha^Y \nsubseteq \alpha^*$

- When $\alpha^Y \supseteq \alpha^*$ then $\beta^{\alpha^Y} = \beta^*$

- Therefore, approximately,

$$P(\beta^* | D) = \sum_{\alpha^Y} P(\beta^{\alpha^Y} | \alpha^Y, D)P(\alpha^Y | D),$$

(1)

where $P(\beta^{\alpha^Y} | \alpha^Y, D)$ can be directly estimated from observed data.
How Traditional BMA Can Fail

\[ \sum_{\alpha^Y} P(\beta^{\alpha^Y}|\alpha^Y, Y)P(\alpha^Y|Y) = \]

\[ \sum_{\alpha^Y \supseteq \alpha^*} P(\beta^{\alpha^*}|\alpha^Y, Y)P(\alpha^Y|Y) + \]

\[ \sum_{\alpha^Y \nsubseteq \alpha^*} P(\beta^{\alpha^Y}|\alpha^Y, Y)P(\alpha^Y|Y) \]

- The **second term** averages across models that do not include \( \alpha^* \), and therefore do not estimate the same effect.

- A common practice in traditional implementations of BMA is to use a uniform prior. This choice might lead to large weights to models in the second term of the equation.
We also fit a two-stage BAC (TBAC) where we cut the feedback from $\alpha^Y$ to $\alpha^X$.

In stage one of TBAC, we specify a uniform prior on $\alpha^X$, a conditional prior on $\alpha^Y|\alpha^X$ we calculate

$$
P(\alpha^X|X) \propto P(X|\alpha^X)P(\alpha^X)$$
$$P(\alpha^Y|X) = \sum_{\alpha^X} P(\alpha^Y|\alpha^X)P(\alpha^X|X),$$

In stage two of TBAC, we use $P(\alpha^Y|X)$ as prior on $\alpha^Y$ to estimate $P(\alpha^Y, \beta_\alpha^Y|D)$.

TBAC can be considered as a Bayesian model averaging method on the outcome model with an informative prior $P(\alpha^Y|X)$ obtained from stage one.

This prior is the key difference between TBAC and traditional BMA, in which a flat uniform prior on the outcome model is typically assumed.
Example 2

\[ Y_t = 0.1 X_t + 0.1 U_{1t} + \ldots + 0.1 U_{14t} + \epsilon_t, \]  

(2)

- \( T = 1000 \)
- \( (X_i, U_{1i}, \ldots, U_{7i}) \) are independent normal vectors with mean zero and a covariance matrix, \( \Sigma = (\sigma_{kl})_{8 \times 8} \), corresponding to an AR(1) process. \( \sigma_{kk} = 1, \sigma_{kl} = \rho^{|k-l|}, 1 \leq k, l \leq 8. \)
  \( \rho = 0.7 \)
- The rest of the confounders \( U_{8i}, \ldots, U_{14i} \) follows \( N(0, 1) \) distribution and are independent of the other variables in the model.
- The set of potential confounders also includes 43 additional independent \( N(0, 1) \) random variables.
### Example 3: Simulation Results

<table>
<thead>
<tr>
<th></th>
<th>True Model</th>
<th>BAC</th>
<th>STEADy</th>
<th>FBMA</th>
<th>NBMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{M}(\hat{\beta})$</td>
<td>0.115</td>
<td>0.123</td>
<td>0.113</td>
<td>0.204</td>
<td>0.170</td>
</tr>
<tr>
<td>$\text{SE}(\hat{\beta})$</td>
<td>0.050</td>
<td>0.051</td>
<td>0.051</td>
<td>0.055</td>
<td>0.069</td>
</tr>
<tr>
<td>$\text{SE of } \hat{\beta}$ across data sets</td>
<td>0.043</td>
<td>0.043</td>
<td>0.044</td>
<td>0.055</td>
<td>0.088</td>
</tr>
<tr>
<td>$\text{MSE}(\hat{\beta})$</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.014</td>
<td>0.012</td>
</tr>
<tr>
<td>Coverage rate of 50% C.I.</td>
<td>14/25</td>
<td>12/25</td>
<td>14/25</td>
<td>3/25</td>
<td>7/25</td>
</tr>
</tbody>
</table>

**Table:** Comparison of estimation of $\beta$ from the four methods along with the gold standard in the second simulation scenario. The target coverage rate is 12.5/25.
<table>
<thead>
<tr>
<th>Method</th>
<th>$\hat{\beta}$</th>
<th>SE($\hat{\beta}$)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model</td>
<td>0.291</td>
<td>0.092</td>
<td>(0.110, 0.471)</td>
</tr>
<tr>
<td>BAC $\omega = \infty$</td>
<td>0.226</td>
<td>0.081</td>
<td>(0.067, 0.385)</td>
</tr>
<tr>
<td></td>
<td>$\omega = 10$</td>
<td>0.217</td>
<td>(0.060, 0.371)</td>
</tr>
<tr>
<td></td>
<td>$\omega = 4$</td>
<td>0.186</td>
<td>(0.019, 0.351)</td>
</tr>
<tr>
<td></td>
<td>$\omega = 2$</td>
<td>0.155</td>
<td>(0.007, 0.317)</td>
</tr>
<tr>
<td>TBAC $\omega = \infty$</td>
<td>0.229</td>
<td>0.083</td>
<td>(0.071, 0.403)</td>
</tr>
<tr>
<td></td>
<td>$\omega = 10$</td>
<td>0.216</td>
<td>(0.071, 0.367)</td>
</tr>
<tr>
<td></td>
<td>$\omega = 4$</td>
<td>0.190</td>
<td>(0.035, 0.347)</td>
</tr>
<tr>
<td></td>
<td>$\omega = 2$</td>
<td>0.155</td>
<td>(0.010, 0.313)</td>
</tr>
<tr>
<td>STEADy</td>
<td>0.221</td>
<td>0.089</td>
<td>(0.045, 0.396)</td>
</tr>
<tr>
<td>FBMA</td>
<td>0.140</td>
<td>0.077</td>
<td>(−0.008, 0.298)</td>
</tr>
<tr>
<td>NBMA</td>
<td>0.007</td>
<td>0.033</td>
<td>(0.000, 0.131)</td>
</tr>
<tr>
<td>Stepwise</td>
<td>0.106</td>
<td>0.066</td>
<td>(−0.023, 0.234)</td>
</tr>
</tbody>
</table>

**Table:** Comparison of estimates of $PM_{2.5}$ effect on CVD hospitalization rate based on BAC, TBAC, STEADy, FBMA, NBMA, stepwise, and the full model.
Model selection affects effect estimation: uncertainty in model selection should be taken into consideration

Methods (priors) that work well for predictions can fail in effect estimation

Considering both the outcome and the exposure model helps improve our ability to focus on the *same* estimand across models
Thanks and References