Causal inference under mis-specification:
adjustment based on the propensity score

Widemberg S. Nobre
DME – UFRJ
11 December 2023

Seminário de Probabilidade

Jointly with David Stephens, Erica Moodie, and Alexandra Schmidt
Suppose a binary exposure denoted by $Z$ and assume that the observed outcome data are generated according to the structural model

$$Y_i = X_{0i} \xi + Z_i \tau + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad (1)$$

where for $p$-dimensional parameter $\xi$ the term $X_{0i} \xi$ defines the true treatment-free mean model.

The goal is to estimate $\tau$ under the experimental model design in (1).
Suppose that the available data are derived from a observational design with $X_i$ representing a set of confounders

- In the observational data generating process, $X_i$ affects the generation of $Y_i$ and $Z_i$, simultaneously, for each $i$

Consider that the following semi-parametric model is adjusted

$$Y_i = h(X_i; \phi) + Z_i \tau + \epsilon_i$$  \hspace{1cm} (2)

where $h(X_i; \phi)$ represents how do we perform confounding adjustment in this particular linear case
An estimate $\tau$ solves

$$
\sum_{i=1}^{n} (Z_i - b(X_i))(Y_i - \tau Z_i) = 0
$$

where $b(X_i) = P(Z_i = 1 \mid X_i)$ is the propensity score. In a frequentist setting, $b(X_i)$ is replaced with $b(X_i; \hat{\gamma})$, where $\hat{\gamma}$ is the solution of

$$
\sum_{i=1}^{n} X_i^\top (Z_i - b(X_i; \gamma)) = 0_p
$$
Frequentist solution

The estimator

\[ \hat{\hat{\tau}} = \sum_{i=1}^{n} \frac{(Z_i - b(X_i; \hat{\gamma})) Y_i}{(Z_i - b(X_i; \hat{\gamma}))Z_i} \]

is consistent if the model \( b(X_i; \hat{\gamma}) \) is correctly specified.

An equivalent result is obtained based on the OLS estimator when performing propensity score regression

\[ Y_i = b(X_i; \hat{\gamma})\phi + Z_i\tau + \epsilon_i \]
Bayesian Inference under Exchangeability

– Joint probability model

$$f_{X,Z,Y}(x, z, y) = f_X(x) f_{Z|X}(z|x) f_{Y|Z,X}(y|z, x)$$

– de Finetti’s representation

$$p_X(x_{1:n}) = \int \prod_{i=1}^{n} f_X(x_i; \eta) \pi_0(\eta) d\eta,$$

$$p_{Z|X}(z_{1:n}|x_{1:n}) = \int \prod_{i=1}^{n} f_{Z|X}(z_i|x_i; \gamma) \pi_0(\gamma) d\gamma,$$

$$p_{Y|X,Z}(y_{1:n}|x_{1:n}, z_{1:n}) = \int \prod_{i=1}^{n} f_{Y|X,Z}(y_i|x_i, z_i; \beta) \pi_0(\beta) d\beta.$$

(3)
Bayesian solution

• **Implication 1:** considering a parametric representation $f_X(x) \equiv f_X(x; \eta)$, $f_{Z|X}(z|x) \equiv f_{Z|X}(z|x; \gamma)$ and $f_{Y|Z,X}(y|z, x) \equiv f_{Y|Z,X}(y|z, x; \beta)$, the triples $(y_i, z_i, x_i)$ are independent given $\varphi = (\eta, \gamma, \beta)$

• **Implication 2:** after specify a prior model for $\varphi$, by standard assumptions, the posterior distribution of $\varphi$ converges to the degenerate point $\varphi_0 = (\eta_0, \gamma_0, \beta_0)$ with

$$f_{X,Z,Y}(x, z, y) \equiv f_{X,Z,Y}(x, z, y; \varphi_0) = f_X(x; \eta_0)f_{Z|X}(z|x; \gamma_0)f_{Y|Z,X}(y|z, x; \beta_0)$$

corresponding to the true (presuming) data generating model

• **Implication 3:** when performing regression with propensity score adjustment, the proposed model does not match

$$f_X(x; \eta_0)f_{Z|X}(z|x; \gamma_0)f_{Y|Z,X}(y|z, x; \beta_0)$$
Bayesian solution

Different proposed solutions for the problem

▶ Joint Bayesian propensity score model: Inference is based on the joint likelihood (McCandless et al, 2009):

\[
\ell(\gamma, \beta) = \prod_{i=1}^{n} f_{Z|X}(z_i|x_i, \gamma)f_{Y|Z,X,\varepsilon}(y_i|z_i, x_i, e(x_i; \gamma), \beta).
\] (4)

▶ Two-step cutting feedback

▶ Two-step plug-in
Bayesian solution

Different proposed solutions for the problem

- Joint Bayesian propensity score model: Inference is based on the joint likelihood (McCandless et al, 2009):

$$
\ell(\gamma, \beta) = \prod_{i=1}^{n} f_{Z|X}(z_i|x_i, \gamma) f_{Y|Z,X,E}(y_i|z_i, x_i, e(x_i; \gamma), \beta).
$$

- Two-step cutting feedback

- Two-step plug-in

What should we do?
The joint specification results in structural bias for any sample size.

Two-step cutting feedback results in measurement error-like bias:

\[ b_i^{(l)} \doteq b_i + \hat{b}(x_i; \gamma_0)(\gamma^{(l)} - \gamma_0) = b_i + u_i^{(l)}(x_i) \]

Two-step plug-in is the best answer, although there is an issue involving coverage rates.
Consider the following data generating mechanism with Normal outcome and binary treatment models.

Suppose the outcome model is specified as

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \tau Z_i + \epsilon_i \]  

with \( \tau = 5 \) and \( (\beta_0, \beta_1, \beta_2, \beta_3) = (3, -2, 10, 6) \), and \( \epsilon_i \sim Normal(0, 1) \).
In the treatment assignment model, suppose that we have 

$$Z_i | X_i = x_i; \gamma_0 \sim \text{Bernoulli}(p_i), \text{ with logit}(p_i) = \gamma_{00} + \gamma_{01}x_{i1} + \gamma_{02}x_{i2} + \gamma_{03}x_{i3},$$

for $\gamma_0 = (2, -2, -2, 1)^\top$.

Confounders are simulated from a trivariate normal distribution with mean $(2, -1, 0.5)^\top$ and $\text{Cov}(X_j, X_k) = 0.8^{|j-k|}$ for $j, k = 1, 2, 3$. 

---

**Illustrative Example**
The propensity score regression model is implemented by first fitting a Bayesian model for $Z$ given $X$, obtaining the predicted values $\hat{b}_i = \hat{\gamma}_0 + \hat{\gamma}_1 x_{i1} + \hat{\gamma}_2 x_{i2} + \hat{\gamma}_3 x_{i3}$, and then fitting the regression model

$$\mathbb{E}[Y|X = x, Z = z, B = \hat{b}; \beta, \phi, \tau] = \beta_0 + \phi \hat{b} + \tau z$$ (6)

which is mis-specified in its treatment-free component, but correctly specified in terms of the treatment-effect component.
**Table 1**: Frequentist properties of Bayesian estimators: $\sqrt{n}$ times the standard deviation, and coverage (Cov.) of 95% interval, in 2000 replicate samples using the exact regression model (Exact), a two-step propensity score regression model (PSR).

| $n$   | $\sqrt{n} \times \text{s.d.}$ Exact | Cov.   | $\sqrt{n} \times \text{s.d.}$ PSR | Cov.   |
|-------|-------------------------------------|--------|----------------------------------|--------|
| 200   | 2.623                               | 95.12  | 4.075                            | 81.64  |
| 500   | 2.589                               | 94.92  | 4.032                            | 81.27  |
| 1000  | 2.569                               | 95.38  | 3.985                            | 81.34  |
| 2000  | 2.589                               | 95.35  | 3.981                            | 81.27  |
The mis-specification renders poorly coverage rates

Two main goals of the paper:
  ▶ Justify the two-step plug-in approach as fully Bayesian procedure, ie., a Bayesian inference that uses probabilistic arguments and prior-to-posterior updating using Bayes Theorem.
  ▶ Correct the coverage rates due to model mis-specification.
Suppose that data are generated according to some likelihood model $f_O(.; \theta_0)$ which we cannot and do not need to specify correctly.

The Bayes estimate is a function of the observed data that minimizes the Bayes risk, or the posterior expected loss for some loss function $\ell(t, \theta) : \Theta \times \Theta \rightarrow \mathbb{R}^+$, that is

$$\hat{\theta} = \arg\min_{t \in \Theta} \mathbb{E}_{\pi_n}[\ell(t, \theta)] = \arg\min_{t \in \Theta} \int \ell(t, \theta) \pi_n(\theta) \, d\theta.$$
If the loss function can be written

\[ \ell(t, \theta) = \int u(s, t)f_\theta(s; \theta) \, ds = E_{f_\theta}[u(S, t); \theta] \]  \tag{7} 

for some function \( u(s, t) : \mathcal{X} \times \Theta \to \mathbb{R}^+ \), then the estimation problem can be rewritten

\[ \hat{\theta} = \arg \min_{t \in \Theta} \int u(s, t) \left\{ \int f_\theta(s; \theta)\pi_n(\theta) \, d\theta \right\} \, ds = \arg \min_{t \in \Theta} E_{p_n}[u(S, t)] \]  \tag{8} 

where \( p_n(s) \) is the posterior predictive distribution implied by the Bayesian specification.
Bayesian decision-theoretic inference

For example, if, for $t \in \Theta$, $u(s, t) = -\log f_O(s; t)$, then we have that

$$
\hat{\theta} = \arg \max_{t \in \Theta} \int \left\{ \int \log f_O(s; t) f_O(s; \theta) \, ds \right\} \pi_n(\theta) \, d\theta.
$$

(9)

In particular, assuming $f_O(s; t) \equiv \text{Normal}(t, 1)$, the calculation becomes

$$
\arg \min_{t \in \Theta} \int \int (s - t)^2 \phi(s - \theta) \, ds \pi_n(\theta) \, d\theta = \int \left\{ \int s \phi(s - \theta) \, ds \right\} \pi_n(\theta) \, d\theta
$$

$$
= \int \theta \pi_n(\theta) \, d\theta
$$
Suppose that, while assuming the data are generated by $f_O$, we wish to perform inference in an alternative model with density $f$ with support $\mathcal{X}$, parameterized by $\vartheta \in \Theta'$.

The decision theoretic framework can still be followed defining a loss function $\ell(t', \theta) : \Theta' \times \Theta \rightarrow \mathbb{R}^+$ as

$$
\ell(t', \theta) = \mathcal{K}(f_O(\cdot; \theta), f(\cdot; t')) = \int \log \left( \frac{f_O(s; \theta)}{f(s; t')} \right) f_O(s; \theta) \, ds = \mathbb{E}_{f_O}[u_\theta(S, t'); \theta]
$$

where $u_\theta(s, t') = \log (f_O(s; \theta)/f(s; t'))$. 
By arguments equivalent to those leading to (9), we have that

\[ \hat{\vartheta} = \arg \max_{t' \in \Theta'} \int \left\{ \int \log f(s; t') f_O(s; \theta) \, ds \right\} \pi_n(\theta) \, d\theta, \]  

(10)

where the maximization over \( t' \) may not depend on \( \theta \).

Therefore, if there is a standard method to sample \( \theta \) from its posterior distribution, we may convert it to obtain a sample from \( \vartheta \) as

\[ \vartheta^{(l)} = \arg \max_{t' \in \Theta'} \int \log f(s; t') f_O(s; \theta^{(l)}) \, ds \]  

(11)

Monte Carlo methods can be used to perform the above integration.
Posterior samples of $\vartheta$ through

$$
\vartheta = \arg \max_{t' \in \Theta} \sum_{i=1}^{n} \omega_i \log f(o_i; t')
$$

where $\omega = (\omega_1, \ldots, \omega_n) \sim \text{Dirichlet}(1, 1, \ldots, 1)$.  

A posterior sample formed by repeatedly sampling the Dirichlet weights to yield $\omega^{(1)}, \ldots, \omega^{(L)}$, with subsequent transformations to yield $\vartheta^{(1)}, \ldots, \vartheta^{(L)}$ is an exact sample from the posterior distribution for $\vartheta$. 

The Bayesian Bootstrap
Bayesian inference for the structured causal model

Mis-specified model

$$y = z\tau + b(x)\phi + \epsilon_i$$

The Bayesian inference procedure with loss function

$$u((y, z, x); \tau, \phi) = (y - z\tau - b(x)\phi)^2$$

yields to $$\pi_n(\tau)$$ concentrated at right value as $$n$$ grows.

If $$b(x)$$ is unknown, then the following loss function can be assumed

$$u((y, z, x); \tau, \phi, \gamma) = -\log f_{Y|X,Z}(y|x, z; \phi, \tau, \gamma^{opt}) - \log f_{Z|X}(z|x; \gamma)$$

where $$\gamma^{opt} = \arg\max_t \int \log f_{Z|X}(z|x; \gamma) \ dF_0(z|x)$$
If $b(x)$ is known, the Bayesian Bootstrap yields a inference procedure that relies on

$$(\tau, \phi) = \arg\min_{t_1, t_2} \sum_{i=1}^{n} \omega_i (y_i - z_i t_1 - b(x_i) t_2)^2$$

This proposed solution is inspired in the frequentist theory, and aims to correct the under coverage associated with model mis-specification.
Bayesian inference for the structured causal model

Table 2: Frequentist properties of Bayesian estimators: $\sqrt{n}$ times the standard deviation, and coverage (Cov.) of 95% interval, in 2000 replicate samples using the exact regression model (Exact), a two-step propensity score regression model (PSR), a PSR with frequentist bootstrap, and a PSR with Bayesian bootstrap.

| $n$  | Exact $\sqrt{n} \times$ s.d. Cov. | PSR $\sqrt{n} \times$ s.d. Cov. | Boot PSR $\sqrt{n} \times$ s.d. Cov. | Bayesian Boot. $\sqrt{n} \times$ s.d. Cov. |
|------|------------------------------------|---------------------------------|--------------------------------------|------------------------------------------|
| 200  | 2.623 95.12                       | 4.075 81.64                    | 3.924 95.60                         | 3.958 94.30                              |
| 500  | 2.589 94.92                       | 4.032 81.27                    | 3.955 94.60                         | 3.913 94.10                              |
| 1000 | 2.569 95.38                       | 3.985 81.34                    | 3.974 94.60                         | 3.890 94.75                              |
| 2000 | 2.589 95.35                       | 3.981 81.27                    | 3.929 94.65                         | 3.925 94.65                              |
Further Simulation Studies

In the data generating mechanism assumes $p = 3$ confounders, with $x = (x_1, x_2, x_3) \sim Normal((-1, 2, 0.5), \Sigma)$, with $\Sigma_{ij} = \text{Cov}(X_i, X_j) = 0.8^{|i-j|}$, for $i, j = 1, 2, 3$, and simulate a continuous treatment $Z_i$ and continuous outcome $Y_i$ from Normal distributions with unit variance and means

$$\mu_Z = 1 - x_1 + x_2 + 2x_3 - x_1x_2 + 2x_2x_3,$$

$$\mu_Y = 1 + 5z + x_1 + x_2 + x_3 + 5x_2x_3.$$

respectively. For each sample size, we generate 1000 datasets under the above scheme. For the exposure model, we fit the mean model $\mu_Z = \tilde{x}\gamma$, where the linear predictor is based on row vector $\tilde{x} = (1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3)$, using linear regression.
Further Simulation Studies

- **‘Unadjusted (UN)’:** unadjusted for confounding;
  
  \[
  \text{UN} : \quad \beta_0 + \tau Z \\
  \text{UN-ext} : \quad \beta_0 + x_1 \beta_1 + x_2 \beta_2 + x_3 \beta_3 + \tau Z
  \]

- **‘Joint (JT)’:** the joint model from equation (4);
  
  \[
  \text{JT} : \quad \beta_0 + \phi \bar{x} \gamma + \tau Z \\
  \text{JT-ext} : \quad \beta_0 + x_1 \beta_1 + x_2 \beta_2 + x_3 \beta_3 + \phi \bar{x} \gamma + \tau Z
  \]
Further Simulation Studies

▶ ‘Cutting feedback (CF)’: the cut feedback approach

\[
\text{CF} : \quad \beta_0 + \phi \tilde{b} + \tau z \\
\text{CF-ext} : \quad \beta_0 + x_1 \beta_1 + x_2 \beta_2 + x_3 \beta_3 + \tau z + \phi \tilde{b}
\]

▶ ‘Two-step (2S)’:

\[
\text{2S} : \quad \beta_0 + \phi \tilde{b} + \tau z \\
\text{2S-ext} : \quad \beta_0 + x_1 \beta_1 + x_2 \beta_2 + x_3 \beta_3 + \phi \tilde{b} + \tau z
\]

▶ ‘Correct’: a correct specification of the linear regression model.
Table 3: Summary of the conventional Bayesian estimates of $\tau$ under a normal exposure. The rows correspond to mean bias of the point estimates of the posterior 95% credible intervals of $\tau$.

| Outcome | $n$       |
|---------|-----------|
|         | 200 | 500 | 1000 | 2000 |
| UN      | 2.084 | 2.092 | 2.093 | 2.089 |
| UN-ext  | 2.401 | 2.448 | 2.444 | 2.444 |
| JT      | -0.355 | -0.345 | -0.344 | -0.345 |
| JT-ext  | -0.092 | -0.088 | -0.089 | -0.090 |
| CF      | 0.059 | 0.027 | 0.013 | 0.006 |
| CF-ext  | 0.045 | 0.021 | 0.011 | 0.005 |
| 2S      | -0.002 | 0.001 | 0.001 | 0.000 |
| 2S-ext  | -0.002 | 0.001 | 0.001 | 0.000 |
| Correct | -0.002 | 0.001 | -0.001 | 0.000 |
Table 4: Summary of the conventional Bayesian estimates of $\tau$ under a normal exposure. The rows correspond to the RMSE of the posterior 95% credible intervals of $\tau$.

| Outcome | $n$ | 200  | 500  | 1000 | 2000 |
|---------|-----|------|------|------|------|
| UN      |     | 2.086| 0.093| 2.093| 2.089|
| UN-ext  |     | 2.416| 2.454| 2.447| 2.445|
| JT      |     | 0.365| 0.349| 0.346| 0.346|
| JT-ext  |     | 0.117| 0.100| 0.095| 0.093|
| CF      |     | 0.092| 0.054| 0.035| 0.024|
| CF-ext  |     | 0.084| 0.051| 0.034| 0.023|
| 2S      |     | 0.071| 0.047| 0.033| 0.023|
| 2S-ext  |     | 0.071| 0.047| 0.033| 0.023|
| Correct |     | 0.056| 0.036| 0.025| 0.018|
Table 5: Summary of the conventional Bayesian estimates of $\tau$ under a normal exposure. The rows correspond to the coverage rates of the posterior 95% credible intervals of $\tau$.

| Coverage | Outcome | 200 | 500 | 1000 | 2000 |
|----------|---------|-----|-----|------|------|
|          | UN      | 0.0 | 0.0 | 0.0  | 0.0  |
|          | UN-ext  | 0.0 | 0.0 | 0.0  | 0.0  |
|          | JT      | 0.1 | 0.0 | 0.0  | 0.0  |
|          | JT-ext  | 75.0| 49.7| 19.8 | 2.1  |
|          | CF      | 100.0| 100.0| 100.0| 100.0|
|          | CF-ext  | 100.0| 100.0| 100.0| 100.0|
|          | 2S      | 100.0| 100.0| 100.0| 100.0|
|          | 2S-ext  | 100.0| 100.0| 100.0| 100.0|
|          | Correct | 94.1 | 94.5 | 94.1 | 94.0 |
Table 6: Summary of the estimates of $\tau$ under a normal exposure using the Bayesian bootstrap in the outcome model, and different approaches to the propensity score model parameters posterior: True indicates the true value of $\gamma$ is used; Parametric indicates a parametric Normal model is used; Linked (LBB) indicates that common Dirichlet weights were used in the two model components.

| Coverage | Outcome     | $\pi_n(\gamma)$ | 200  | 500  | 1000 | 2000 |
|----------|-------------|------------------|------|------|------|------|
|          | PS          | True             | 94.2 | 94.0 | 95.0 | 96.0 |
|          | PS-ext      | True             | 93.1 | 92.8 | 94.1 | 94.8 |
|          | CF          | Parametric       | 100.0| 100.0| 100.0| 100.0|
|          | CF-ext      | Parametric       | 100.0| 100.0| 100.0| 100.0|
|          | 2S          | Parametric       | 100.0| 100.0| 100.0| 100.0|
|          | 2S-ext      | Parametric       | 100.0| 100.0| 100.0| 100.0|
|          | 2S          | Linked BB        | 94.2 | 92.8 | 94.7 | 94.1 |
|          | 2S-ext      | Linked BB        | 94.2 | 92.8 | 94.7 | 94.1 |

$n$
Discussion

- Justified the use of two-step plug-in approach as fully Bayesian inference procedure
- Proposed approach has good Bayesian and frequentist properties
- A future avenue of research is to address mis-specification under dependent data
Muito obrigado pela atenção