Bayesian Treatment Effect Estimation with Many Potential Confounders

Carlos M. Carvalho (UT Austin)
P. Richard Hahn (Chicago Booth)
David Puelz (UT Austin)

Insper, March 2016

Everyone knows...

...that unmeasured confounders can lead to biased estimates of regression coefficients (omitted variable bias)

Suppose we're interested in the **treatment effect** of dietary kale intake.

And want to know how effective it is at lowering cholesterol, which is our **outcome variable**.

Unfortunately, we have only observational data (i.e., not a randomized study).

Kale intake predicts exercise

Our bad luck, only gym-rats seem to eat much kale. And exercise is known to lower cholesterol: the “direct” effect is **confounded**.

$$Y_i = \beta_0 + \alpha D_i + \varepsilon_i,$$

Because $\mathrm{cov}(D_i, \varepsilon_i) \neq 0$, we can write

$$Y_i = \beta_0 + \alpha D_i + \omega D_i + \tilde{\varepsilon}.$$

Since $\mathrm{cov}(D_i, \tilde{\varepsilon}_i) = 0$, we mis-estimate α as $\alpha + \omega$.

We must "adjust" for weekly exercise

The good news is, we can **control** for weekly exercise, X_i, by including it in the regression:

$$Y_i = \beta_0 + \alpha D_i + \beta X_i + \varepsilon_i.$$

This "clears out" the confounding: conditional on X_i, $\text{cov}(D_i, \varepsilon_i) = 0$ and we're good to go.

But what if we don't know what we need to control for?

Everyone knows...

...that shrinkage priors (e.g., point-mass priors) allow us to "safely" include many covariates in a regression (even more than our sample size!)

We have lots of theory backing this up:

- Stein type results on admissibility (yay ridge regression!)
- Oracle type results
- Intuition concerning bias-variance trade-offs

So we should control for as many things as possible and use our favorite shrinkage prior, right?

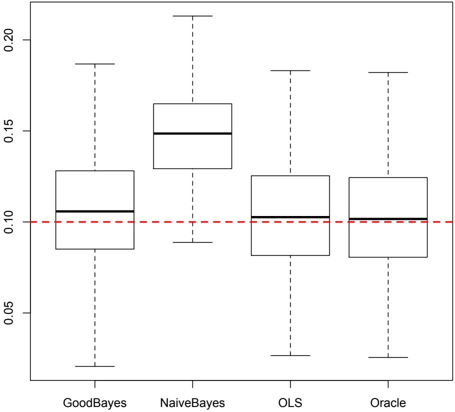
The obvious approach

$$Y_i = \beta_0 + \alpha D_i + \beta X_i + \varepsilon_i.$$

- a flat prior on the treatment effect: $(\alpha, \sigma_\varepsilon^2) \propto 1/\sigma_\varepsilon$,
- shrinkage prior on β (e.g., a horseshoe prior).

And we're off to the races!

oops



It turns out that this "obvious" approach is really bad at getting reasonable estimates of the treatment effect α.

What happened?

Consider the **selection equation**:

$$D = X\gamma + \epsilon.$$

By substitution we can write the **response equation** as

$$\begin{aligned} Y_i &= \alpha(X_i\gamma + \epsilon) + X_i\beta + \varepsilon_i, \\ &= \alpha(X_i\gamma + \epsilon) + X_i\Delta + [\varepsilon_i + X_i(\beta - \Delta)]. \end{aligned}$$

For $\gamma \neq 0$, biasing β towards zero biases $\text{cov}(D, \varepsilon)$ away from zero!

Recap so far

Adjusting for confounding is fundamentally different than estimating a best linear predictor.

Shrinkage priors want to “explain” (i.e. predict) Y using a small number of large magnitude coefficients.

The “obvious” model is indifferent if one of those coefficients happens to be α — we bias towards **mis-identification**.

Shrinkage priors BIAS the treatment effect coefficient!

Previous work

Here are some notable references on this...

- **Wang, Parmigiani, Dominici (2012), "Bayesian adjustment for confounding" (BAC)**
- Propensity scores: Zigler and Dominici (2014), Weihua An (2010)
- Lasso-based: Belloni, Chernozhukov and Hansen (2015)
- Instrumental variables: Hahn and Lopes, Hansen and Kozbur (2014), Chernozhukov, Hansen and Spindler (2015)

Our solution has the virtue of being relatively straightforward.

The typical parametrization

$$\begin{aligned}
\text{Selection Eq.:} \quad & D = \mathbf{X}^t\gamma + \epsilon, & \epsilon \sim N(0, \sigma_\epsilon^2), \\
\text{Response Eq.:} \quad & Y = \alpha D + \mathbf{X}^t\beta + \nu, & \nu \sim N(0, \sigma_\nu^2).
\end{aligned}$$

These equations correspond to the factorization of the joint distribution

$$f(Y, D \mid \gamma, \beta, \sigma_\epsilon, \sigma_\nu) = f(Y \mid D, \beta, \sigma_\epsilon) f(D \mid \gamma, \sigma_\nu).$$

This factorization implies a complete separation of the parameter sets: independent priors on the regression parameters

$$\pi(\beta, \gamma, \alpha) = \pi(\beta)\pi(\gamma)\pi(\alpha)$$

imply that only the response equation is used in estimating β and α.

Our reparametrization: a latent error approach

We reparametrize as

$$\begin{pmatrix} \alpha \\ \beta + \alpha\gamma \\ \gamma \end{pmatrix} \rightarrow \begin{pmatrix} \alpha \\ \beta_d \\ \beta_c \end{pmatrix}.$$

which gives the new equations

$$\begin{aligned} \text{Selection Eq.:} \quad & D = \mathbf{X}^t \beta_c + \epsilon, & \epsilon \sim N(0, \sigma_\epsilon^2), \\ \text{Response Eq.:} \quad & Y = \alpha(D - \mathbf{X}^t \beta_c) + \mathbf{X}^t \beta_d + \nu, & \nu \sim N(0, \sigma_\nu^2). \end{aligned}$$

We can now shrink β_d and β_c with impunity!

Control functions

Our re-parametrization generalizes, and falls under the category of an approach called "control functions".

$$
\begin{aligned}
D_i &= g(\mathbf{X}_i) + \epsilon_i, \\
Y_i &= f(D_i, \mathbf{X}_i) + \nu_i
\end{aligned}
$$

To isolate the causal component of $f(D, \mathbf{X})$, we rewrite it as $f(D - g(\mathbf{X}), \mathbf{X})$.

A special case assumes additive separability of f:

$$
\begin{aligned}
D_i &= g(\mathbf{X}_i) + \epsilon_i, \\
Y_i &= f_1(D_i - g(\mathbf{X}_i)) + f_2(\mathbf{X}_i) + \nu_i.
\end{aligned}
$$

Simulation study

$$
\begin{aligned}
\text{Selection Eq.:} \quad & D = \mathbf{X}^t\beta_c + \epsilon, & \epsilon \sim N(0, \sigma_\epsilon^2), \\
\text{Response Eq.:} \quad & Y = \alpha(D - \mathbf{X}^t\beta_c) + \mathbf{X}^t\beta_d + \nu, & \nu \sim N(0, \sigma_\nu^2).
\end{aligned}
$$

Set $var(D) = var(Y) = 1$ and center and scale the columns of $\mathbf{X}$.

Define the ℓ_2 norms of the confounding and direct effects as $\rho^2 = \|\beta_c\|_2^2$ and $\phi^2 = \|\beta_d\|_2^2$ so that

$$
\begin{aligned}
var(D) &= \rho^2 + \sigma_\epsilon^2 \\
var(Y) &= \kappa^2 + \phi^2 + \sigma_\nu^2,
\end{aligned}
$$

with $\sigma_\epsilon^2 = 1 - \rho^2$ and $\sigma_\nu^2 = 1 - \alpha^2(1-\rho^2) - \phi^2$ and $\kappa^2 = \alpha^2(1-\rho^2)$.

ρ^2		Bias	Coverage	I.L.	MSE
0.1	New Approach	-0.0032	0.943	0.2357	0.0037
	OLS	-0.0016	0.951	0.2477	0.004
	Naive Regularization	-0.0112	0.895	0.2089	0.0037
	Oracle OLS	0.0023	0.946	0.2173	0.0031
0.3	New Approach	-0.0047	0.95	0.2751	0.0047
	OLS	-0.0018	0.951	0.2808	0.0052
	Naive Regularization	-0.0355	0.848	0.2293	0.0057
	Oracle OLS	0.0026	0.946	0.2464	0.004
0.5	New Approach	-3e-04	0.963	0.3345	0.0066
	OLS	-0.0022	0.951	0.3323	0.0072
	Naive Regularization	-0.0768	0.746	0.2631	0.012
	Oracle OLS	0.0031	0.946	0.2915	0.0056
0.7	New Approach	0.0084	0.964	0.4374	0.0113
	OLS	0.0024	0.944	0.4303	0.0123
	Naive Regularization	-0.1559	0.543	0.3292	0.0346
	Oracle OLS	0.004	0.946	0.3764	0.0093
0.9	New Approach	-0.004	0.972	0.7403	0.0292
	OLS	0.0045	0.954	0.7469	0.0351
	Naive Regularization	-0.4482	0.231	0.4779	0.2391
	Oracle OLS	0.0069	0.946	0.6519	0.0278

Table: $\mathbf{n} = \mathbf{100}, \mathbf{p} = \mathbf{30}, \mathbf{k} = \mathbf{3}$. $\kappa^2 = 0.05$. $\phi^2 = 0.7$. $\sigma_\nu^2 = 0.25$.

ρ^2		Bias	Coverage	I.L.	MSE
0.1	New Approach	0.0082	0.918	0.3632	0.0105
	OLS	-0.0017	0.944	0.4785	0.0144
	Naive Regularization	-0.0068	0.835	0.2957	0.0097
	Oracle OLS	-0.001	0.952	0.3235	0.0065
0.3	New Approach	-1e-04	0.94	0.4203	0.0128
	OLS	-0.002	0.944	0.5425	0.0186
	Naive Regularization	-0.035	0.837	0.3191	0.0126
	Oracle OLS	-0.0011	0.952	0.3668	0.0084
0.5	New Approach	-0.0047	0.93	0.5183	0.0196
	OLS	-0.0023	0.944	0.6419	0.026
	Naive Regularization	-0.0869	0.738	0.3555	0.0222
	Oracle OLS	-0.0014	0.952	0.434	0.0117
0.7	New Approach	0.0056	0.937	0.6926	0.0341
	OLS	0.0046	0.934	0.8204	0.0478
	Naive Regularization	-0.189	0.539	0.4033	0.0565
	Oracle OLS	-0.0018	0.952	0.5604	0.0195
0.9	New Approach	-0.0772	0.959	1.1572	0.0804
	OLS	-0.0156	0.931	1.4347	0.1402
	Naive Regularization	-0.5419	0.102	0.4868	0.3297
	Oracle OLS	-0.003	0.952	0.9706	0.0585

Table: $\mathbf{n = 50, p = 30, k = 3}$. $\kappa^2 = 0.05$. $\phi^2 = 0.7$. $\sigma_\nu^2 = 0.25$.

Empirical example: Levitt abortion reanalysis

According to "Freakonomics":

- unwanted children are more likely to grow up to be criminals,
- therefore legalized abortion, which leads to fewer unwanted children, leads to lower levels of crime in society.

To investigate, they conduct three analyses, one each for three different types of crime: violent crime, property crime, and murders.

Donohue III and Levitt data

Y is per capita crime rates (violent crime, property crime, and murders) by state, from 1985–1997, and D, is the "effective" abortion rate.

The control variables, $\mathbf{X}$, are:

- prisoners per capita (log),
- police per capita (log),
- state unemployment rate,
- state income per capita (log),
- percent of population below the poverty line,
- generosity of AFDC (lagged by fifteen years),
- concealed weapons law,
- beer consumption per capita.

Including state and year dummy variables brings the total number of control variables to $p = 66$ (with $n = 624$).

Replication

	Property Crime		Violent Crime		Murder	
	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
OLS	-0.110	-0.072	-0.171	-0.090	-0.221	-0.040
Our way	-0.113	-0.073	-0.182	-0.098	-0.222	-0.039
naive	-0.075	-0.010	0.079	0.301	-0.186	0.085

An augmented control set

Our expanded model includes the following additional control variables:

- interactions between the original eight controls and year,
- interactions between the original eight controls and year squared,
- interactions between state effects and year,
- interactions between state effects and year squared.

When allowing for this degree of flexibility, estimation becomes quite challenging, with just $n = 624$ observations and $p = 176$ control variables.

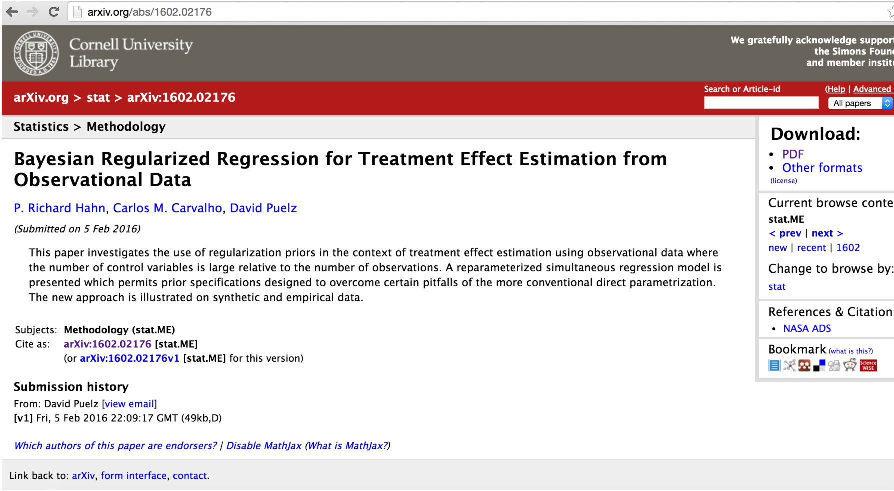
Augmented analysis results

	Property Crime		Violent Crime		Murder	
	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
OLS	-0.226	0.019	-0.374	0.336	-0.125	1.763
Our way	-0.038	0.014	-0.114	0.053	-0.081	0.279
naive	0.007	0.129	0.011	0.412	-0.227	0.116

Recap

- Social scientists want to draw causal conclusions from observational data.
- This can only be done if sufficient control variables are included.
- If too many control variables are included, statistical properties suffer.
- Regularization is known to improve statistical estimation, but if employed naively, regularization actually makes causal inference worse!
- Our new parametrization fixes this flaw.

Done…



arxiv.org/abs/1602.02176

Cornell University Library

We gratefully acknowledge support
the Simons Foun
and member instit

arXiv.org > stat > arXiv:1602.02176

Search or Article-id (Help | Advanced

All papers

Statistics > Methodology

Bayesian Regularized Regression for Treatment Effect Estimation from Observational Data

P. Richard Hahn, Carlos M. Carvalho, David Puelz

(Submitted on 5 Feb 2016)

This paper investigates the use of regularization priors in the context of treatment effect estimation using observational data where the number of control variables is large relative to the number of observations. A reparameterized simultaneous regression model is presented which permits prior specifications designed to overcome certain pitfalls of the more conventional direct parametrization. The new approach is illustrated on synthetic and empirical data.

Subjects: **Methodology (stat.ME)**
Cite as: **arXiv:1602.02176 [stat.ME]**
(or **arXiv:1602.02176v1 [stat.ME]** for this version)

Submission history

From: David Puelz [view email]
[v1] Fri, 5 Feb 2016 22:09:17 GMT (49kb,D)

Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)

Link back to: arXiv, form interface, contact.

Download:

- PDF
- Other formats

(license)

Current browse conte
stat.ME
< prev | next >
new | recent | 1602

Change to browse by:
stat

References & Citation
- NASA ADS

Bookmark (what is this?)