Bayesian regularized regression
for treatment effect estimation
with many potential confounders

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

BYU, Feb 2016
Outline

- A problem motivated from social science.
- Fixing a flaw with naive regularization.
- Simulation results.
- An applied example from Freakonomics.
Deconfounding

It is well-known that unmeasured confounders can lead to biased estimates of regression coefficients.

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 \( \text{cov}(D_i, \varepsilon_i) \neq 0 \), we can write

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

Since \( \text{cov}(D_i, \tilde{\varepsilon}_i) = 0 \), we mis-estimate \( \alpha \) 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?
So what is wrong with just including anything we can think of in our regression?

The problem is that we typically have finite data. The more things we add, the more variable (untrustworthy) our estimator will be.

How do people deal with this in practice?
Let’s Take the Con out of Econometrics

By Edward E. Leamer

Econometricians would like to project the image of agricultural experimenters who divide a farm into a set of smaller plots of land and who select randomly the level of fertilizer to be used on each plot. If some plots are assigned a certain amount of fertilizer while others are assigned none, then the difference between the mean yield of the fertilized plots and the mean yield of the unfertilized plots is a measure of the effect of fertilizer on agricultural yields. The econometrician’s humble job is only to determine if that difference is large enough to suggest a real effect of fertilizer, or is so small that it is more likely due to random variation.

One should not jump to the conclusion that there is necessarily a substantive difference between drawing inferences from experimental as opposed to nonexperimental data. The images I have drawn are deliberately prejudicial. First, we had the experimental scientist with hair neatly combed, wide eyes peering out of horn-rimmed glasses, a white coat, and an electronic calculator for generating the random assignment of fertilizer treatment to plots of land. This seems to contrast sharply with the nonexperimental farmer with overalls, unkempt hair, and bird droppings on his boots. Another image, drawn by Orcutt, is even more damaging:
Regularization

It is well-known 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
- bias-variance trade-off intuitions
- “bet on sparsity” ideas (oracle results)
Stein “paradox”

An estimator $\hat{\theta}_1$ is said to dominate another estimator $\hat{\theta}_2$ if, for all values of (the true) $\theta$, the risk (expected loss) of $\hat{\theta}_1$ is lower than, or equal to, the risk of $\hat{\theta}_2$, with strict inequality for some $\theta$.

An estimator admissible if no other estimator dominates it, otherwise it is inadmissible.

A famous result of Stein says: the maximum likelihood estimator of the mean of a multivariate Gaussian distribution is inadmissible under mean squared error.

(For $p > 2$ or for stochastic predictors.)
The intuition behind Stein’s result is that a “shrunk” estimator — a convex combination of the MLE and some constant estimator — dominates the MLE because its risk is lower precisely when \( \theta \) is equal to that constant. (Perversely, the constant used doesn’t matter!)

This intuition can be extended to the more general idea that risk performance can be improved by judiciously adding bias. In the case of MSE we have

\[
E\{(\theta - \hat{\theta})^2\} = \text{bias}^2 + \text{variance}.
\]

This can be extended to other losses. The key idea is stability. Mukherjee et al. *Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization* (2006).
In a Bayesian framework, our estimator naturally exhibits shrinkage, via the role of the prior distribution.

Under MSE, the Bayes estimator is the posterior mean. In the normal means problem, this leads to the so-called “ridge” estimator, which is precisely the form of the estimator suggested by Stein.

(See Efron's book *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction* for more connections.)

Modern shrinkage priors try to improve on the ridge estimator by designing sensible “shrinkage profiles”.
The horseshoe prior

For scalar observations $y_j \sim N(\theta_j, 1)$ and prior $\theta_j \sim N(0, \lambda^2_j)$, the posterior mean of $\theta_j$ may be expressed as

$$E(\theta_j \mid y_j) = \{1 - E(\kappa_j \mid y_j)\} y_j,$$

where $\kappa_j = 1/(1 + \lambda^2_j)$. 

[Graph showing comparison of different priors and horseshoe distribution]
The horseshoe prior

The horseshoe prior has a hierarchical representation as

$$\pi(\beta_j \mid \lambda) = N(0, \lambda^2 \lambda_j^2),$$

$$\lambda_j \overset{iid}{\sim} C^+(0,1).$$

Further, it admits a closed form approximation via the inequality

$$\frac{K}{2} \log (1 + 4/\theta^2) < \pi_{HS}(\theta) < K \log (1 + 2/\theta^2).$$

We will use this representation for our computations later.
Theorem 4. Suppose the true sampling model \( p_{\theta_0} \) is \( y_j \sim N(\theta_0, \sigma^2) \). Then:

1. For \( \hat{p}_n \) under the horseshoe prior, the optimal rate of convergence of \( R_n \) when \( \theta_0 = 0 \) is \( R_n = O\left(n^{-1}(\log n - b \log \log n)\right) \), where \( b \) is a constant. When \( \theta_0 \neq 0 \), the optimal rate is \( R_n = O(n^{-1} \log n) \).

2. Suppose \( p(\theta) \) is any other density that is continuous, bounded above, and strictly positive on a neighbourhood of the true value \( \theta_0 \). For \( \hat{p}_n \) under \( p(\theta) \), the optimal rate of convergence of \( R_n \), regardless of \( \theta_0 \), is \( R_n = O(n^{-1} \log n) \).

Proof. See the Appendix. \( \square \)
Naive regularization for deconfounding regressions

\[ Y_i = \beta_0 + \alpha D_i + \beta X_i + \varepsilon_i. \]

- a flat prior on the treatment effect: \((\alpha, \sigma_\varepsilon^2) \propto 1/\sigma_\varepsilon,\)

- shrinkage prior on \(\beta\) (e.g., the horseshoe prior).

- Now “turn the Bayesian crank”…
Something goes wrong

It turns out that this “obvious” approach is really bad at getting reasonable estimates of the treatment effect $\alpha$. 
Bad bias versus good bias

Assume that:

\[ D_i = X_i^t \gamma + \epsilon_i. \]

Now substitute a shrunk estimate, \( \beta - \Delta \), in place of the true (unknown) \( \beta \) vector:

\[ Y_i = \alpha D_i + X_i^t (\beta - \Delta) + [\nu_i + X_i^t \Delta]. \]

This implies that \( \nu_i \) is taken to be \( \nu_i + X_i \Delta \), which gives

\[ \text{cov}(\nu_i + X_i^t \Delta, X^t \gamma + \epsilon) \neq 0. \]

Biasing \( \beta \) towards zero biases \( \text{cov}(D, \epsilon) \) 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 $\alpha$ — we bias towards mis-identification.

Shrinkage priors bias the treatment effect coefficient dramatically!
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)


Our solution has the virtue of being relatively straightforward.
The typical parametrization

Selection Eq.: \[ D = X^t \gamma + \epsilon, \quad \epsilon \sim N(0, \sigma^2_\epsilon), \]
Response Eq.: \[ Y = \alpha D + X^t \beta + \nu, \quad \nu \sim N(0, \sigma^2_\nu). \]

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 \( \beta \) and \( \alpha \).
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

Selection Eq.: \( D = X^t \beta_c + \epsilon, \quad \epsilon \sim N(0, \sigma^2_\epsilon), \)

Response Eq.: \( Y = \alpha(D - X^t \beta_c) + X^t \beta_d + \nu, \quad \nu \sim N(0, \sigma^2_\nu). \)

We can now shrink \( \beta_d \) and \( \beta_c \) with impunity!
Control functions

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

\[ D_i = g(X_i) + \epsilon_i, \]
\[ Y_i = f(D_i, X_i) + \nu_i \]

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

A special case assumes additive separability of \( f \):

\[ D_i = g(X_i) + \epsilon_i, \]
\[ Y_i = f_1(D_i - g(X_i)) + f_2(X_i) + \nu_i. \]
Simulation study

Selection Eq.: \[ D = X^t \beta_c + \epsilon, \quad \epsilon \sim N(0, \sigma_\epsilon^2), \]

Response Eq.: \[ Y = \alpha(D - X^t \beta_c) + X^t \beta_d + \nu, \quad \nu \sim N(0, \sigma_\nu^2). \]

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

Define the \( \ell_2 \) norms of the confounding and direct effects as \( \rho^2 = \|\beta_c\|_2^2 \)
and \( \phi^2 = \|\beta_d\|_2^2 \) so that

\[
\text{var}(D) = \rho^2 + \sigma_\epsilon^2 \\
\text{var}(Y) = \kappa^2 + \phi^2 + \sigma_\nu^2,
\]

with \( \sigma_\epsilon^2 = 1 - \rho^2 \) and \( \sigma_\nu^2 = 1 - \alpha^2(1 - \rho^2) - \phi^2 \) and \( \kappa^2 = \alpha^2(1 - \rho^2) \).
<table>
<thead>
<tr>
<th>$\rho^2$</th>
<th></th>
<th>Bias</th>
<th>Coverage</th>
<th>I.L.</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>New Approach</td>
<td>-0.0032</td>
<td>0.943</td>
<td>0.2357</td>
<td>0.0037</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-0.0016</td>
<td>0.951</td>
<td>0.2477</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.0112</td>
<td>0.895</td>
<td>0.2089</td>
<td>0.0037</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>0.0023</td>
<td>0.946</td>
<td>0.2173</td>
<td>0.0031</td>
</tr>
<tr>
<td>0.3</td>
<td>New Approach</td>
<td>-0.0047</td>
<td>0.95</td>
<td>0.2751</td>
<td>0.0047</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-0.0018</td>
<td>0.951</td>
<td>0.2808</td>
<td>0.0052</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.0355</td>
<td>0.848</td>
<td>0.2293</td>
<td>0.0057</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>0.0026</td>
<td>0.946</td>
<td>0.2464</td>
<td>0.004</td>
</tr>
<tr>
<td>0.5</td>
<td>New Approach</td>
<td>-3e-04</td>
<td>0.963</td>
<td>0.3345</td>
<td>0.0066</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-0.0022</td>
<td>0.951</td>
<td>0.3323</td>
<td>0.0072</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.0768</td>
<td>0.746</td>
<td>0.2631</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>0.0031</td>
<td>0.946</td>
<td>0.2915</td>
<td>0.0056</td>
</tr>
<tr>
<td>0.7</td>
<td>New Approach</td>
<td>0.0084</td>
<td>0.964</td>
<td>0.4374</td>
<td>0.0113</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>0.0024</td>
<td>0.944</td>
<td>0.4303</td>
<td>0.0123</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.1559</td>
<td>0.543</td>
<td>0.3292</td>
<td>0.0346</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>0.004</td>
<td>0.946</td>
<td>0.3764</td>
<td>0.0093</td>
</tr>
<tr>
<td>0.9</td>
<td>New Approach</td>
<td>-0.004</td>
<td>0.972</td>
<td>0.7403</td>
<td>0.0292</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>0.0045</td>
<td>0.954</td>
<td>0.7469</td>
<td>0.0351</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.4482</td>
<td>0.231</td>
<td>0.4779</td>
<td>0.2391</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>0.0069</td>
<td>0.946</td>
<td>0.6519</td>
<td>0.0278</td>
</tr>
</tbody>
</table>

Table: $n = 100$, $p = 30$, $k = 3$. $\kappa^2 = 0.05$. $\phi^2 = 0.7$. $\sigma^2_\nu = 0.25$. 
<table>
<thead>
<tr>
<th>$\rho^2$</th>
<th>New Approach</th>
<th>OLS</th>
<th>Naive Regularization</th>
<th>Oracle OLS</th>
<th>Bias</th>
<th>Coverage</th>
<th>I.L.</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.0082</td>
<td>-0.0017</td>
<td>-0.0068</td>
<td>-0.001</td>
<td>0.1</td>
<td>0.918</td>
<td>0.3632</td>
<td>0.0105</td>
</tr>
<tr>
<td>0.3</td>
<td>-1e-04</td>
<td>-0.002</td>
<td>-0.035</td>
<td>-0.0011</td>
<td>0.3</td>
<td>0.94</td>
<td>0.4203</td>
<td>0.0128</td>
</tr>
<tr>
<td>0.5</td>
<td>-0.0047</td>
<td>-0.0023</td>
<td>-0.0869</td>
<td>-0.0014</td>
<td>0.5</td>
<td>0.93</td>
<td>0.5183</td>
<td>0.0196</td>
</tr>
<tr>
<td>0.7</td>
<td>0.0056</td>
<td>0.0046</td>
<td>-0.189</td>
<td>-0.0018</td>
<td>0.7</td>
<td>0.937</td>
<td>0.6926</td>
<td>0.0341</td>
</tr>
<tr>
<td>0.9</td>
<td>-0.0772</td>
<td>-0.0156</td>
<td>-0.5419</td>
<td>-0.003</td>
<td>0.9</td>
<td>0.959</td>
<td>1.1572</td>
<td>0.0804</td>
</tr>
</tbody>
</table>

Table: $n = 50, p = 30, k = 3$. $\kappa^2 = 0.05$. $\phi^2 = 0.7$. $\sigma^2_\nu = 0.25$. 
<table>
<thead>
<tr>
<th>$\rho^2$</th>
<th></th>
<th>Bias</th>
<th>Coverage</th>
<th>I.L.</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>New Approach</td>
<td>-0.0053</td>
<td>0.93</td>
<td>0.4137</td>
<td>0.0126</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-0.0031</td>
<td>0.951</td>
<td>0.4699</td>
<td>0.0145</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.027</td>
<td>0.434</td>
<td>0.1528</td>
<td>0.0178</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>0.0044</td>
<td>0.946</td>
<td>0.4123</td>
<td>0.0111</td>
</tr>
<tr>
<td>0.3</td>
<td>New Approach</td>
<td>-0.0101</td>
<td>0.933</td>
<td>0.472</td>
<td>0.0176</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-0.0035</td>
<td>0.951</td>
<td>0.5329</td>
<td>0.0186</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.075</td>
<td>0.373</td>
<td>0.1625</td>
<td>0.0256</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>0.005</td>
<td>0.946</td>
<td>0.4675</td>
<td>0.0143</td>
</tr>
<tr>
<td>0.5</td>
<td>New Approach</td>
<td>-0.001</td>
<td>0.933</td>
<td>0.5751</td>
<td>0.0245</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-0.0041</td>
<td>0.951</td>
<td>0.6305</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.1407</td>
<td>0.304</td>
<td>0.1751</td>
<td>0.0411</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>0.0059</td>
<td>0.946</td>
<td>0.5532</td>
<td>0.02</td>
</tr>
<tr>
<td>0.7</td>
<td>New Approach</td>
<td>0.0044</td>
<td>0.95</td>
<td>0.7509</td>
<td>0.0368</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-0.0049</td>
<td>0.953</td>
<td>0.8156</td>
<td>0.0394</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.265</td>
<td>0.134</td>
<td>0.1801</td>
<td>0.0918</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>0.0076</td>
<td>0.946</td>
<td>0.7141</td>
<td>0.0333</td>
</tr>
<tr>
<td>0.9</td>
<td>New Approach</td>
<td>-0.01</td>
<td>0.939</td>
<td>1.2784</td>
<td>0.1131</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-0.0022</td>
<td>0.942</td>
<td>1.416</td>
<td>0.1345</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.6114</td>
<td>0.002</td>
<td>0.1841</td>
<td>0.3983</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>0.0132</td>
<td>0.946</td>
<td>1.2369</td>
<td>0.0999</td>
</tr>
</tbody>
</table>

Table: $n = 100, p = 30, k = 3$. $\kappa^2 = 0.05$. $\phi^2 = 0.05$. $\sigma^2_{\nu} = 0.9$. 
<table>
<thead>
<tr>
<th>$\rho^2$</th>
<th>Method</th>
<th>Bias</th>
<th>Coverage</th>
<th>I.L.</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>New Approach</td>
<td>0.0021</td>
<td>0.919</td>
<td>0.6073</td>
<td>0.0306</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-0.0119</td>
<td>0.93</td>
<td>0.8888</td>
<td>0.0528</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.0291</td>
<td>0.443</td>
<td>0.2207</td>
<td>0.0352</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>-0.0019</td>
<td>0.952</td>
<td>0.6138</td>
<td>0.0234</td>
</tr>
<tr>
<td>0.3</td>
<td>New Approach</td>
<td>-0.0033</td>
<td>0.909</td>
<td>0.6918</td>
<td>0.0421</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-0.0038</td>
<td>0.944</td>
<td>1.0294</td>
<td>0.0668</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.0651</td>
<td>0.402</td>
<td>0.237</td>
<td>0.0428</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>-0.0022</td>
<td>0.952</td>
<td>0.696</td>
<td>0.0301</td>
</tr>
<tr>
<td>0.5</td>
<td>New Approach</td>
<td>-0.011</td>
<td>0.894</td>
<td>0.8191</td>
<td>0.064</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>0.0071</td>
<td>0.927</td>
<td>1.2041</td>
<td>0.103</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.1354</td>
<td>0.349</td>
<td>0.233</td>
<td>0.0577</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>-0.0026</td>
<td>0.952</td>
<td>0.8235</td>
<td>0.0421</td>
</tr>
<tr>
<td>0.7</td>
<td>New Approach</td>
<td>-0.028</td>
<td>0.904</td>
<td>1.0842</td>
<td>0.105</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-8e-04</td>
<td>0.938</td>
<td>1.5533</td>
<td>0.1603</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.2752</td>
<td>0.217</td>
<td>0.2474</td>
<td>0.1163</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>-0.0033</td>
<td>0.952</td>
<td>1.0632</td>
<td>0.0702</td>
</tr>
<tr>
<td>0.9</td>
<td>New Approach</td>
<td>-0.1078</td>
<td>0.948</td>
<td>1.8128</td>
<td>0.2303</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>0.0291</td>
<td>0.942</td>
<td>2.6708</td>
<td>0.4893</td>
</tr>
<tr>
<td></td>
<td>Naive Regularization</td>
<td>-0.6045</td>
<td>0.015</td>
<td>0.2576</td>
<td>0.4096</td>
</tr>
<tr>
<td></td>
<td>Oracle OLS</td>
<td>-0.0058</td>
<td>0.952</td>
<td>1.8415</td>
<td>0.2106</td>
</tr>
</tbody>
</table>

Table: $n = 50, p = 30, k = 3$. $\kappa^2 = 0.05$. $\phi^2 = 0.05$. $\sigma^2_{\nu} = 0.9$. 
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, $X$, 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

<table>
<thead>
<tr>
<th></th>
<th>Property Crime</th>
<th>Violent Crime</th>
<th>Murder</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.5%</td>
<td>2.5%</td>
<td>2.5%</td>
</tr>
<tr>
<td></td>
<td>97.5%</td>
<td>97.5%</td>
<td>97.5%</td>
</tr>
<tr>
<td>OLS</td>
<td>-0.110</td>
<td>-0.171</td>
<td>-0.221</td>
</tr>
<tr>
<td></td>
<td>-0.072</td>
<td>-0.090</td>
<td>-0.040</td>
</tr>
<tr>
<td>Our way</td>
<td>-0.113</td>
<td>-0.182</td>
<td>-0.222</td>
</tr>
<tr>
<td>naive</td>
<td>-0.073</td>
<td>-0.098</td>
<td>-0.039</td>
</tr>
<tr>
<td></td>
<td>-0.075</td>
<td>0.079</td>
<td>-0.186</td>
</tr>
<tr>
<td></td>
<td>-0.010</td>
<td>0.301</td>
<td>0.085</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>Property Crime</th>
<th>Violent Crime</th>
<th>Murder</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.5%</td>
<td>97.5%</td>
<td>2.5%</td>
</tr>
<tr>
<td>OLS</td>
<td>-0.226</td>
<td>0.019</td>
<td>-0.374</td>
</tr>
<tr>
<td>Our way</td>
<td>-0.038</td>
<td>0.014</td>
<td>-0.114</td>
</tr>
<tr>
<td>naive</td>
<td>0.007</td>
<td>0.129</td>
<td>0.011</td>
</tr>
</tbody>
</table>
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.
Thanks for listening.
Computing posterior means under the horseshoe prior is nontrivial.

Typically a Markov chain Monte Carlo method called a Gibbs sampler is used. However, this approach scales badly as the number of predictors gets large.

We introduced a new method based on the idea of an elliptical slice sampler.
1. Choose a starting value $x_0$ for which $f(x_0) > 0$.
2. Sample a $y$ value uniformly between 0 and $f(x_0)$.
3. Draw a horizontal line across the curve at this $y$ position.
4. Sample a point $(x, y)$ from the line segments within the curve.
5. Repeat from step 2 using the new $x$ value.
In our application, we want to sample from the posterior

\[ \pi(\theta \mid X, Y) \propto f(Y \mid X, \theta)\pi(\theta) \]

where \( \pi(\theta) \) is a horseshoe prior.

Our strategy is going to be to take advantage of the fact that \( f(Y \mid X, \theta) \) is a Gaussian likelihood, so has level sets that are ellipses.

This means that we can travel along that ellipse and not alter the likelihood contribution to the posterior, so we need only consider the contribution of the prior \( \pi(\theta) \).
Elliptical slice sampler

This idea is due to Murray et al. "Elliptical slice sampling" (2009).
Let $\hat{\beta} = (X^t X)^{-1} X^t \tilde{Y}$ denote the least squares solution and for an initial value of $\beta$, define $\Delta := \beta - \hat{\beta}$. We then sample $\beta$ as a vector, according to the following algorithm.

1. Draw $\zeta \sim N(0, \sigma^2 (XX^t)^{-1})$.
2. For $\nu \sim \text{Uniform}(0, 1)$ define $\ell := \log (\pi(\beta)) + \log (\nu)$.
3. Draw angle $\varphi \sim \text{Uniform}(0, 2\pi)$; set $\text{lower} \leftarrow \varphi - 2\pi$ and $\text{upper} \leftarrow \varphi$.
4. Set $\Delta' \leftarrow \Delta \cos \varphi + \zeta \sin \varphi$ and $\beta' \leftarrow \hat{\beta} + \Delta'$.
5. while $\log (\pi(\beta')) < \ell$
   5.1 if $\varphi < 0$, set $\text{lower} \leftarrow \varphi$, else set $\text{upper} \leftarrow \varphi$.
   5.2 Draw angle $\varphi \sim \text{Uniform}(\text{lower}, \text{upper})$
   5.3 Update $\Delta' \leftarrow \Delta \cos \varphi + \zeta \sin \varphi$ and $\beta' \leftarrow \hat{\beta} + \Delta'$.
6. Set $\Delta \leftarrow \Delta'$ and $\beta \leftarrow \hat{\beta} + \Delta'$.
Remarks

This sampler draws from the posterior for $\beta$ under a flat prior and then adjust these draws according to the prior distribution, rotating the vector along level curves of the posterior (equivalently, likelihood) until an improvement is reached with respect to the prior.

We use $\pi(\beta \mid \nu) = \prod_{j=1}^{p} \log \left\{1 + 4/(\beta_j/\nu)^2\right\}/\nu$ and we sample the hyperparameter $\nu$ (the so-called “global” scale parameter) within our Gibbs sampler as a random walk Metropolis update.
Our treatment effect regression model

Here we must jointly sample \((\alpha, \beta^t, \gamma^t)\) under flat priors and transform to \(\beta_c\) and \(\beta_d\) before prior evaluation. For initial values of the parameter, this gives:

1. Draw \(\zeta_1 \sim N(0, \sigma^2_{\nu}(XX^t)^{-1})\) and \(\zeta_2 \sim N(0, \sigma^2_{\epsilon}(XX^t)^{-1})\) and defining \(\zeta^t = (\zeta_1^t, \zeta_2^t)\).

2. For \(\nu \sim \text{Uniform}(0, 1)\) define \(\ell := \log (\pi(\beta_c)) + \log (\pi(\beta_d)) + \log (\nu)\).

3. Draw angle \(\varphi \sim \text{Uniform}(0, 2\pi)\); set \(\text{lower} \leftarrow \varphi - 2\pi\) and \(\text{upper} \leftarrow \varphi\).

4. Set \(\Delta' \leftarrow \Delta \cos \varphi + \zeta \sin \varphi\), \(\alpha' \leftarrow \hat{\alpha} + \Delta'_1\), \(\beta' \leftarrow \hat{\beta} + \Delta'_{2:(p+1)}\), \(\gamma' \leftarrow \hat{\gamma} + \Delta'_{(p+2):(2p+1)}\); with \(\beta'_c = \gamma'\) and \(\beta'_d = \alpha'\gamma' + \beta'\).

5. \textbf{while} \(\log (\pi(\beta'_c)) + \log (\pi(\beta'_d)) < \ell\)
   
   \hspace{1em} 5.1 \textbf{if} \(\varphi < 0\), set \(\text{lower} \leftarrow \varphi\), \textbf{else} set \(\text{upper} \leftarrow \varphi\).
   
   \hspace{1em} 5.2 Draw angle \(\varphi \sim \text{Uniform}(\text{lower}, \text{upper})\)
   
   \hspace{1em} 5.3 Update \(\Delta' \leftarrow \Delta \cos \varphi + \zeta \sin \varphi\) and \(\alpha' \leftarrow \hat{\alpha} + \Delta'_1\), \(\beta' \leftarrow \hat{\beta} + \Delta'_{2:(p+1)}\), \(\gamma' \leftarrow \hat{\gamma} + \Delta'_{(p+2):(2p+1)}\); with \(\beta'_c = \gamma'\) and \(\beta'_d = \alpha'\gamma' + \beta'\).

6. Set \(\Delta \leftarrow \Delta'\) and \(\alpha \leftarrow \hat{\alpha} + \Delta'_1\), \(\beta \leftarrow \hat{\beta} + \Delta'_{2:(p+1)}\), \(\gamma \leftarrow \hat{\gamma} + \Delta'_{(p+2):(2p+1)}\); with \(\beta_c = \gamma\) and \(\beta_d = \alpha\gamma + \beta\).