Manual to accompany MATLAB package for Bayesian Shrinkage Approach for Treatment Effect Estimation

Dimitris Korobilis Kenichi Shimizu

University of Glasgow

November 3, 2021

Contents

1 Introduction

This notes manual accompanies the monograph "Bayesian Approaches to Shrinkage and Sparse Estimation: A guide for applied econometricians" by Korobilis and Shimizu (2021) and the associated MATLAB code.

The directories in the file BayesHDTE.zip, are:

HDCofounding	Hierarchical Bayes version of Antonelli et al. (2019)
NaiveOutcomeLasso	Naive outcome Lasso
DoublePostSelection	Double post selection of Belloni et al. (2014b)
NaiveBayes	Bayesian Naive Regularization
genData	Generate data assuming some sparsity
PenalizedToolBox	Matlab Toolbox of McIlhagga (2016)

2 Literature on selection among many controls

Let y_i denote the outcome and T_i be the treatment variable (for now, assume it is binary). The set of cofounders x_i is high-dimensional. The parameter of interest is the treatment effect α in the model below:

$$y_i = \beta_0 + \alpha T_i + x_i'\beta + \epsilon_i \tag{1}$$

2.1 Naive outcome Lasso and naive post selection

A *naive outcome Lasso* approach would be to apply Lasso to the equation above, excluding α from the L1 penalty. Any control variable that is highly correlated with T_i but weakly with y_i tends to drop out of the selection because adding such control variable does not increase predictive power for y_i so much.

A *naive post selection* approach would be to fit the naive outcome Lasso in the first stage and then regress y_i on T_i as well as on the selected covariates to estimate and do inference about the treatment effect. If some covariates with non-zero relationship with the outcome are not selected in the first stage, this leads to omitted-variable bias in the second stage regression.

Belloni et al. (2014a) explain that there are two problems with the naive approaches. First, it ignores to describe the relationship between T_i and x_i, which is a key to understand the omitted variable bias. Second, the model (1) is not representing a prediction rule for y_i given T_i and x_i, and hence it is not adequate to apply high-dimensional methods such as Lasso directly.

Code **NaiveOutcomeLasso.m** estimates both naive outcome Lasso and naive post selection.

2.2 Double post selection of Belloni et al. (2014b)

Belloni et al. (2014b) propose to work with the following system of two reduced form equations:

$$y_i = x_i'\beta + \epsilon_i$$
$$T_i = x_i'\psi + \upsilon_i$$

As both equations above represent predictive relationships, one can apply high-dimensional methods directly. In the first stage, they estimate Lasso on each equation above. In the second stage, they estimate the treatment effect by running a regression of the outcome on the treatment and the union of the chosen control variables from the two Lassos. This helps one can to guard against omitted variable biases.

Code **DoublePostSelection.m** estimates the Double post selection.

2.3 High-dimensional cofounding adjustment of Antonelli et al. (2019)

Antonelli et al. (2019) propose a spike-and-slab Lasso prior approach to the problem that the naive approach of just using shrinkage methods on the response equation has: the coefficient on a control variable that is highly correlated with T_i but weakly with y_i tends to be shrunk to zero. Their proposed framework is

$$\begin{aligned}
y_i|T_i, x_i, \alpha, \beta, \sigma^2 &\sim N\left(\beta_0 + \alpha T_i + x_i'\beta, \sigma^2\right) \\
p(\beta|\gamma, \sigma^2) &= \prod_{j=1}^{p} \gamma_j \psi_1(\beta_j; \lambda_1, \sigma^2) + (1-\gamma_j)\psi_0(\beta_j; \lambda_0, \sigma^2) \\
p(\gamma|\theta) &= \prod_{j=1}^{p} \theta^{\omega_j \gamma_j} \left(1-\theta^{\omega_j}\right)^{1-\gamma_j} \\
p(\theta|a, b) &\sim \text{Beta}(a, b) \\
p(\sigma^2|c, d) &\sim \text{InvGamma}(c, d) \\
\beta_0, \alpha &\sim N(0, K)
\end{aligned}$$

where $\psi_0(\beta_j; \lambda_0, \sigma^2) = \frac{\lambda_0}{2\sigma} e^{-\lambda_0|\beta_j|/\sigma}$ and $\psi_1(\beta_j; \lambda_1, \sigma^2) = \frac{\lambda_1}{2\sigma} e^{-\lambda_1|\beta_j|/\sigma}$. They fix $a = 1$ and $b = 0.1p$. λ_1 is fixed to be a small value, say 0.1, so that the prior variance in the slab component is high enough to be uninformative.

The hyperparameter λ_0 is chosen via empirical Bayes. A new feature that they introduce is the weights ω_j which are tuning parameters that they use to prioritize variables to have $\gamma_j = 1$ if they are associated with the treatment. Specifically, they first fit the standard Lasso on the model

for predicting T given X. For x_j with non-zero coefficient from the Lasso, they set $\omega_j = \delta$ for some $\delta \in (0,1)$. For other variables, $\omega_j = 1$. On the one hand, a smaller value of δ leads to higher inclusion probability and hence more protection against the omitted variable bias. On the other hand, one needs to ensure a small enough inclusion probability for an unimportant variable in the outcome model (that is x_j with $\beta_j = 0$).

The conditional probability that x_j belongs to the slab component is

$$p^*_{\omega_j}(\beta_j|\theta,\lambda_0,\sigma^2) = P\left(\gamma_j|\beta_j,\lambda_0,\theta,\sigma^2,\omega_j\right) = \frac{\psi_1(\beta_j;\lambda_1,\sigma^2)\theta^{\omega_j}}{\psi_1(\beta_j;\lambda_1,\sigma^2)\theta^{\omega_j} + \psi_0(\beta_j;\lambda_0,\sigma^2)(1-\theta^{\omega_j})}$$

They first run a Gibbs sampler with $\omega_j = 1$ for all j and then plug in posterior means for the unknown coefficients in the above inclusion probability. The authors then choose $\delta \in (0,1)$ as the smallest value of ω_j such that $p^*_{\omega_j}(0|\theta,\lambda_0,\sigma^2)$ is less than 0.1.

2.4 Bayesian naive regularization

A Bayesian counterpart of the naive outcome Lasso would be to consider a shrinkage or variable selection prior on β and a uninformative prior on α in (1). The hierarchical model under a generic prior on β can be summarized as

$$\begin{aligned}
y_i|\beta_0,\alpha,\beta,\sigma^2,T_i,x_i &\sim N\left(\beta_0 + \alpha T_i + x_i'\beta, \sigma^2\right)\\
\beta_0,\alpha &\sim N(0,\sigma^2 K)\\
\beta|Q_\beta &\sim N(0,\sigma^2\text{diag}\left(Q_\beta\right))\\
Q_\beta &\sim \pi(Q_\beta)\\
\sigma^2 &\sim \text{InvGamma}(c,d)
\end{aligned}$$

where $Q_\beta = \left(\tau_1^2,\ldots,\tau_p^2\right)$ collects the prior variances for β. Its prior distribution $\pi(Q_\beta)$ depends on the choice of the variable selection prior. By defining $\beta^* = (\beta_0,\alpha,\beta)'$, $y^* = (y_1,\ldots,y_n)'$, $X^* = (x_1^*,\ldots,x_n^*)'$, and $\epsilon = (\epsilon_1,\ldots,\epsilon_n)'$ with $x_i^{*'} = (1,T_i,x_i')$, we can write the outcome model in a matrix form as

$$y^* = X^*\beta^* + \epsilon^*$$

Note that conditioning on X^*, the joint distribution decomposes as

$$p\left(y,\beta_0,\alpha,\beta,Q_\beta,\sigma^2\right) = p\left(y|\beta_0,\alpha,\beta,\sigma^2\right)p\left(\beta|Q_\beta\right)p(\beta_0)p(\alpha)p(\sigma^2)p(Q_\beta)$$

Hence, the updating rule for Q_β will remain the same as in the case of a general high-dimensional linear regression. We summarize the Gibbs sampler below. See **??** for derivations.

1. The conditional posterior for β^* is normal

$$\beta^*|\bullet \sim N\left(A^{-1}X^{*'}y^*, \sigma^2 A^{-1}\right)$$

   where $A = X^{*'}X^* + D^{-1}$ with D is a diagonal matrix with diagonal (K, K, Q_β)

2. Sample σ^2 from the full conditional (in case of an inverse Gamma prior):

$$\sigma^2|\bullet \sim \text{InvGamma}\left(c + \frac{n}{2} + \frac{p}{2}, d + \frac{1}{2}||y - \beta_0 - \alpha T - X\beta|| + \frac{1}{2}\beta' D^{-1}\beta\right)$$

3. Update Q_β

The last step depends on the specific choice of the prior.

Code **NaiveBayes.m** estimates the naive Bayes approach.

2.5 Hierarchical Bayes version of Antonelli et al. (2019)

In their two-step procedure, the authors use the spike and slab Lasso prior for the slope parameters. Here we generalize their approach to an arbitrary spike and slab prior. This hierarchical model can be summarized as

$$\begin{aligned}
y_i|\beta_0, \alpha, \beta, \sigma^2, T_i, x_i &\sim N\left(\beta_0 + \alpha T_i + x_i'\beta, \sigma^2\right) \\
\beta_0, \alpha|\sigma^2 &\sim N(0, \sigma^2 K) \\
\beta_j|\gamma_j, \sigma^2 &\sim \gamma_j N(\beta_j; 0, \sigma^2\tau_{1j}^2) + (1-\gamma_j)N(\beta_j; 0, \sigma^2\tau_{0j}^2) \\
\tau_{0j}^2, \tau_{1j}^2 &\sim \pi(\tau_{0j}^2, \tau_{1j}^2) \\
\sigma^2|c, d &\sim \text{InvGamma}(c, d) \\
\gamma_j|\theta &\sim \theta^{\omega_j\gamma_j}\left(1-\theta^{\omega_j}\right)^{1-\gamma_j} d\theta \\
\theta|a, b &\sim \text{Beta}(a, b)
\end{aligned}$$

where $\pi(\tau_{0j}^2, \tau_{1j}^2)$ depends on the specified prior.

In the first step, we fit Lasso on the model of T_i on x_i in order to decide an active covariate set. We will set $\omega_j = \delta \in (0, 1)$ if x_j belongs to the active set and $\omega_j = 1$ otherwise. The probability

that x_j belongs to the slab component given $\theta, \tau_{0j}^2, \tau_{1j}^2$ is

$$p^*_{\omega_j}(\beta_j|\theta, \tau_{0j}^2, \tau_{1j}^2, \sigma^2) = \frac{N(\beta_j; 0, \sigma^2\tau_{1j}^2)\theta^{\omega_j}}{N(\beta_j; 0, \sigma^2\tau_{1j}^2)\theta^{\omega_j} + N(\beta_j; 0, \sigma^2\tau_{0j}^2)(1-\theta^{\omega_j})}$$

We first run the Gibbs sampler with $\omega_j = 1$ for all j. Then, the value of $\delta \in (0,1)$ [1] is chosen as the smallest value of ω_j such that $p^*_{\omega_j}(0|\theta, \tau_{0j}^2, \tau_{1j}^2, \sigma^2)$ is below 10%, where for the unknown parameter values we use posterior means based on the initial Gibbs sampling.

A Gibbs sampler is summarized below

1. Sample β^* from the full conditional:

$$\beta^*|\bullet \sim N\left(A^{-1}X^{*'}y^*, A^{-1}\right)$$

where $A = X^{*'}X^* + D^{-1}$ with D is a diagonal matrix with diagonal $\left(K, K, \{\gamma_j\tau_{1j}^2 + (1-\gamma_j)\tau_{0j}^2\}_{j=1:p}\right)$

2. Sample σ^2 from the full conditional:

$$\sigma^2|\bullet \sim \text{InvGamma}\left(c + \frac{n}{2} + \frac{p}{2}, d + \frac{1}{2}||y - \beta_0 - \alpha T - X\beta|| + \frac{1}{2}\beta' D^{-1}\beta\right)$$

3. Sample γ_j from Bernoulli distribution with the mean parameter

$$\frac{N(\beta_j; 0, \sigma^2\tau_{1j}^2)\theta^{\omega_j}}{N(\beta_j; 0, \sigma^2\tau_{1j}^2)\theta^{\omega_j} + N(\beta_j; 0, \sigma^2\tau_{0j}^2)(1-\theta^{\omega_j})}$$

4. Sample θ based on a Metropolis-Hastings algorithm

$$p(\theta|\bullet) \propto \theta^{a+\sum_{j=1}^p \omega_j\gamma_j}(1-\theta)^b \prod_{j=1}^p (1-\theta^{\omega_j})^{(1-\gamma_j)}$$

5. Update τ_{0j}^2, τ_{1j}^2

The priors we will use

- Narisetty & He (2014): τ_{0j}^2, τ_{1j}^2 fixed
- Student-t distribution instead on normal
- Laplace distribution

[1] We could allow δ_j to vary for each covariate j.

  - The prior on τ_{1j}^2 is an exponential distribution and τ_{0j}^2 is proportional to a small number times τ_{1j}^2
  - The priors on τ_{0j}^2, τ_{1j}^2 are both exponential distributions

- Horseshoe prior on the slab component

Code **HDCofounding.m** estimates the Double post selection.

3 Simulation setup

3.1 Methods to compare

(1) Naive Lasso, (2) Naive post selection, (3) Double post selection, (4) Naive Bayes, and (5) High-dimensional cofounder adjustment approach with various spike-and-slab priors.

3.2 Data generating process

The model used to generate data is as follows. For simplicity, we should focus on continuous treatments.

$$y_i = \beta_0 + \alpha T_i + x_i'\beta + \epsilon_i$$
$$T_i = x_i'\psi + v_i$$

where $\beta_0 = 0$ and $\alpha = 1$. We let $\beta_j = c_y\bar{\beta}_j$ and $\psi_j = c_t\bar{\psi}_j$ and choose the constants c_y and c_t in order to achieve desired level of signal-to-noise ratios. We set $(n, p) = (100, 200)$.

Relevant variables x_{ij} are categorized into four types: (a) strong cofounders that are strongly correlated with both T_i and y_i, (b) weak cofounders that are strongly correlated with T_i but weakly with y_i, (c) instruments that are strongly correlated with T_i but uncorrelated with y_i, and (d) strong predictors that are strongly correlated with y_i but uncorrelated with T_i.

For (a), we set $\bar{\psi}_j = 1$ if j is odd (-1 if even) and $\bar{\beta}_j = 1$ if j is odd (-1 if even), for (b), $\bar{\psi}_j = 1$ if j is odd (-1 if even) and $\bar{\beta}_j = 0.3$ if j is odd (-0.3 if even), for (c) $\bar{\psi}_j = 1$ if j is odd (-1 if even) and $\bar{\beta}_j = 0$, and for (d) $\bar{\psi}_j = 0$ and $\bar{\beta}_j = 1$ if j is odd (-1 if even). For variables that do not belong to none of the four groups, we set $\bar{\psi}_j = 0$ and $\bar{\beta}_j$ i.i.d. $\sim N(0, 0.1^2)$.

3.2.1 Correlation among covariates

The covariates are either uncorrelated or correlated. In the first case, we take $x_{ij} \sim$ i.i.d. $N(0, \sigma_x^2)$ for all j and i. In the second case, the covariates are spatially correlated and generated as x_i i.i.d. $\sim$

$N(0, \Sigma)$ for all i with $\Sigma_{kj} = \rho^{|j-k|}$, where ρ determines how strongly the covariates are correlated. We fix $\sigma_x^2 = 1$ and $\rho = 0.5$.

3.2.2 Sparcity

Given $q \in (0, 1)$, we let $\lfloor 100 \times q \times p \rfloor$ be the percentage of the relevant covariates which fall into the four types defined above.

3.2.3 Error variance

We follow Belloni et al. (2014b) for defining the error variances. For homoskedastic errors, let $v_i \sim$ i.i.d. $N(0, \sigma_t^2)$ and $\epsilon_i \sim$ i.i.d. $N(0, \sigma_y^2)$ with $\sigma_y^2 = \sigma_t^2 = 1$. For heteroskedastic errors, let $v_i \sim$ i.i.d. $N(0, \sigma_t^2(x_i))$ and $\epsilon_i \sim N(0, \sigma_y^2(T_i, x_i))$ where $\sigma_t(x_i) = \sqrt{\frac{\left(1+x_i'\theta\right)^2}{\mathbb{E}_n\left(1+x_i'\theta\right)^2}}$ and $\sigma_y(T_i, x_i) = \sqrt{\frac{\left(1+\beta_0+\alpha T_i+x_i'\theta\right)^2}{\mathbb{E}_n\left(1+\beta_0+\alpha T_i+x_i'\theta\right)^2}}$, where $\mathbb{E}_n$ denotes the empirical expectation.

3.2.4 Signal-to-noise ratio

In a general linear regression $y_i = x_i'\beta + \epsilon_i$, the signal-to-noise ratio (SNR) is defined as $SNR = \frac{||\Sigma_X^{1/2}\beta||^2}{\sigma^2}$ where σ^2 is the error variance and Σ_X is a $p \times p$ covariance matrix of x_i. $||\Sigma_X^{1/2}\beta||^2 = \beta'\Sigma_X\beta$ measures the overall signal strength, where $||\cdot||$ is the ℓ^2-norm. A related quantity is R^2_{pop}, the population value of R^2, defined as $\frac{SNR}{1+SNR}$. If $\beta_j = c\tilde{\beta}_j$ for $j = 1, ..., p$ for some scalar c, then $c = \sqrt{\frac{\sigma^2}{\tilde{\beta}'\Sigma_X\tilde{\beta}}\frac{R^2_{pop}}{1-R^2_{pop}}}$. Hence, we can choose c to achieve a desired value of R^2_{pop} or SNR.

Now, let R_t^2 and R_y^2 be some pre-specified population values of R^2 for the treatment and the outcome equations, respectively. We follow Belloni et al. (2014b) to compute the constants c_t and c_y. For homeskedastic case, we can compute the constants as follows.

1. For given (n, p, Σ_X), generate X and for a given q, define $\bar{\psi}_j$ and $\bar{\beta}_j$ for $j = 1, \ldots, p$.
2. Given R_t^2, compute $c_t = \sqrt{\frac{\sigma_t^2}{\bar{\psi}'\Sigma_X\bar{\psi}}\frac{R_t^2}{1-R_t^2}}$.
3. Given R_y^2, compute $c_y = \sqrt{\frac{\sigma_y^2}{\bar{\beta}'\Sigma_X\bar{\beta}}\frac{R_y^2}{1-R_y^2}}$.

For heteroskedastic case, also use the above formulas for computing the constants as if v_i and ϵ_i homoskedastic following Belloni et al. (2014b).

We choose values of (q, R_t^2, R_y^2) to define various data-generating-processes. Consider the following four scenarios.

1. uncorrelated predictors AND homoscedasticity
2. correlated predictors AND homoscedasticity
3. uncorrelated predictors AND heteroskedasticity
4. correlated predictors AND heteroskedasticity

References

Antonelli, J., Parmigiani, G., & Dominici, F. (2019). High-dimensional confounding adjustment using continuous spike and slab priors. *Bayesian Analysis*, *14*(3), 805.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29–50.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650.

McIlhagga, W. H. (2016). penalized: A matlab toolbox for fitting generalized linear models with penalties.

Narisetty, N. N., & He, X. (2014, 04). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, *42*(2), 789-817. Retrieved from https://doi.org/10.1214/14-AOS1207 doi: 10.1214/14-AOS1207