Regularized Regression with Change-points: Introducing Hidden Markov Bayesian Bridge Model

Jong Hee Park*
Seoul National University
jongheepark@snu.ac.kr

Soichiro Yamauchi†
Harvard University
syamauchi@g.harvard.edu

March 30, 2019

Abstract

Recent innovations in regularization methods offer an important breakthrough to regression analysis with many predictors. However, existing regularization methods commonly assume that the level of sparsity or shrinkage does not change over time. This assumption is problematic because regularizing time-varying parameters toward zero can lead to erroneous inferential results. In this paper, we present a statistical method that allows both regularization and estimation of parameter changes in high dimensional data. The proposed method, which we call hidden Markov Bayesian bridge model (HMBB), uses the Bayesian bridge model for parameter regularization and a hidden Markov model to estimate parameter changes. Simulation studies show that HMBB outperforms other regularization methods in recovering time-varying parameters as well as time-constant parameters in various settings. We apply HMBB to the estimation of the effect of U.S. food aid on civil conflicts and report new findings.

Keywords: Bayesian bridge, change-point, hidden Markov model, regularization

*Professor, Department of Political Science and International Relations, Seoul National University
†Ph.D. student, Department of Government, Harvard University.

1 Introduction

In social science data analysis, researchers are usually interested in estimating the effect of time-varying covariates ($\mathbf{X}$) on a response variable ($\boldsymbol{y}$) in the presence of many time-varying confounding variables ($\mathbf{Z}$). Time series cross-national data in political science is an example where $\mathbf{X}$ and $\mathbf{Z}$ are time-varying covariates of country-specific factors and $\boldsymbol{y}$ is a response vector observed at the country level. In this setup, an important challenge is two-fold; (1) identifying time-varying effects of $\mathbf{X}$ on $\boldsymbol{y}$ while (2) properly controlling for time-varying effects of confounding variables ($\mathbf{Z}$) on $\boldsymbol{y}$.

We consider this challenge as *the change-point problem in high-dimensional regression analysis*. It is a change-point problem as we need to examine temporal heterogeneity of parameters. It is also a high-dimensional problem because design matrices of subset data (e.g. a segment of $(\mathbf{X}, \mathbf{Z})$ pertaining to the first regime) could have larger p than n, where p is the number of predictors and n is the number of observations.

Recent innovations in regularization methods offer an important breakthrough to high-dimensional regression analysis. However, existing regularization methods commonly assume that the level of sparsity or shrinkage does not change over time and hence applying these methods to time series data with change-points can lead to erroneous inferential results.

Figure 1 illustrates the change-point problem in regularized regression analysis. We generate 100 time series observations with a single break. The number of predictor is 50, among which 40 predictors are zero. Values of 10 nonzero coefficients are randomly drawn in each regime from a uniform distribution $\mathcal{U}(-3, 3)$. Colored dots of Regime 1 and Regime 2 indicate the ground truth. We fit seven popular regularization methods (Lasso, Elastic Net, Ridge, adaptive lasso, fused lasso, Bayesian lasso, and horseshoe) in addition to the ordinary least squares (OLS) method, which is shown as benchmark.

The results are striking. RMSEs of the regularization methods are as large as that of OLS and, even worse, many large time-varying signals are forced to be zero by most regularization methods. We highlight two types of inferential fallacy here. First, for coefficient changes into opposite directions, estimates of regularization methods suffer from attenuation bias. Second, for coefficient changes with same signs, estimates of regularization methods underestimate regime-changing parameters.

In this paper, we propose a statistical method to address the above problems. We take a fully Bayesian approach to the change-point problem in high-dimensional regression analysis and present a method that allows both regularization and estimation of breaks in high dimensional regression analysis. The proposed method uses Polson, Scott and Windle (2014)'s Bayesian bridge model for efficient parameter regularization in a high-dimensional regression model. The Bayesian bridge model provides two advantages over other modeling choices. The first advantage is that the Bayesian bridge model has the property of avoiding the overshrinkage of large coefficients, cor-

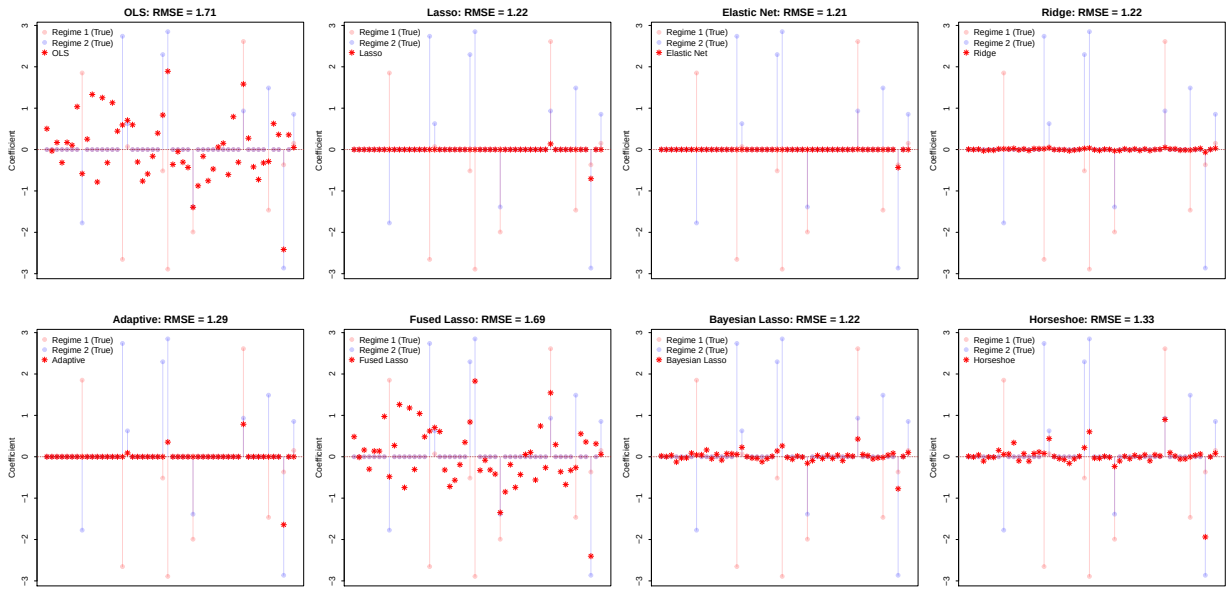

Figure 1: *Change-point Problem in Regularized Regression Analysis: True parameter values are displayed by transparent dots (•) and vertical lines. Estimates are marked by red asterisks (∗). We generate a synthetic time series data set with* $T = 100$ *and the number of predictors is 50. The number of non-sparse predictors is 10. A single break is planted in the mid-point (*$t = 50$*). Regime-specific non-sparse parameters (40 for each regime) are generated from a uniform distribution* $\mathcal{U}(-3, 3)$ *RMSE is the root mean squared error of estimated coefficients: RMSE* $= \sqrt{p^{-1}\sum_{j=1}^{p}(\widehat{\beta}_j - \beta_j^{true})^2}$.

responding to the oracle property in classical regularization estimators (Polson and Scott, 2010). The second advantage of the Bayesian bridge model is computational efficiency due to low correlations global and local shrinkage parameters. As Polson, Scott and Windle (2014) noted, Bayesian inference of global-local shrinkage prior models such as Bayesian lasso and horseshoe prior models suffer from high autocorrelations in MCMC draws as global and local shrinkage parameters are highly correlated with each other. We call the proposed method the hidden Markov Bayesian bridge model (HMBB).

1.1 Related Works

Many statistical methods have been developed to address the first problem of identifying time-varying effects (e.g. Quandt, 1958; Chow, 1960; Chernoff and Zacks, 1964; Hamilton, 1989; Andrews, 1993; Barry and Hartigan, 1993; Bai and Perron, 1998; Chib, 1998). However, most of these methods work under a simple regression model with only a constant or a few covariates. A large number of covariates poses a serious computational challenge in these change-point models as the number of parameters increases multiplicatively with the number of breaks.

There have been some attempts to solve the change-point problem in regularized regression analysis. First, the fused lasso is one of the most well known examples (Tibshirani et al., 2004;

Bleakley and Vert, 2011; Tang and Song, 2016; Qian and Su, 2016). The fused lasso applies the method of fusion either across parameters in the nearest neighbor ($\sum_{j=2}^{p} |\beta_j - \beta_{j-1}|$) for classification and pattern recognition or across time ($\sum_{t=2}^{T} |\beta_t - \beta_{t-1}|$) for break detection. Thus, a search of jumps in parameter values for all parameters across time is computationally demanding and inefficient as the number of parameters increases. Moreover, the idea of parameter "fusion" does not provide measures of uncertainty on break points and break numbers.

Second, there has been a surge of high-dimensional change-point detection methods in frequentist approaches (e.g. Frick, Munk and Sieling, 2014; Chan, Yau and Zhang, 2014; Lee, Seo and Shin, 2016; Lee et al., 2017). Most of these methods focus on simple cases of high-dimensional change-point problems. By simple cases, we mean the case in which covariates with regime-changing coefficients are known or the case in which the number of covariates with regime-changing coefficients are small. However, except some rare cases, most social science researchers do not have clear knowledge about the number of breaks, the timing of breaks, the scope of time-varying covariate effects, and the range of covariates (and their interactions) that need to be included in a regression model.

1.2 Plan of the Paper

After presenting our model in Section 2, we examine the performance of HMBB in various high-dimensional data settings (Section 3). We check the performance of HMBB in high-dimensional data under various sparsity settings. Simulation studies show that HMBB outperforms other regularization methods in recovering time-varying parameters as well as time-constant parameters. Then, Section 4 revisit two studies in economics and political science. First, we illustrate how HMBB can be used to estimate heterogeneous causal effects in the framework of the two-stage lease square (2SLS) regression analysis using Nunn and Qian (2014)'s study of the effect of US food aid on civil conflicts. We show how HMBB improves their original analysis by identifying parameter heterogeneity in the first stage equation. Second, we discuss how HMBB can be used to identify time-varying strong signals from many covariates using HMBB and the decoupled shrinkage and selection (DSS) method proposed by Hahn and Carvalho (2015). We revisit Alvarez, Garrett and Lange (1991)'s study on the role of left parties on economic growth.

The proposed method is available via the open-source software BridgeChange in R environment.

2 The Proposed Methodology

2.1 The Setup

Polson, Scott and Windle (2014) present a fully Bayesian treatment of the bridge model. The key intuition of their Bayesian treatment lies in constructing joint priors for β_j and local shrinkage

parameters (λ_j) using Lévy processes.[1] To introduce the Bayesian bridge model, we write the basic setup for Bayesian regularized regression models as a scale mixture of normal structure:

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n) \\
\beta_j|\tau^2, \lambda_j^2 &\sim \mathcal{N}(0, \tau^2\lambda_j^2) \\
\lambda_j^2 &\sim p(\lambda_j^2) \\
\tau^2 &\sim p(\tau^2) \\
\sigma^2 &\sim p(\sigma^2)
\end{aligned}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$ where n is the number of observations and p is the number of predictors. In this setup, the global shrink parameter τ^2 controls the overall sparsity of the model and local shrinkage parameters λ_j identify signals. What distinguishes different Bayesian regularization models is the choice of a prior distribution of $\boldsymbol{\beta}$ and its hyperparameter (τ^2). For example, the use of a double exponential prior for $\boldsymbol{\beta}$ leads to Bayesian lasso (Park and Casella, 2008) and the horseshoe prior model uses a normal prior for $\boldsymbol{\beta}$ and an inverted-beta distribution for λ_j^2 (Carvalho, Polson and Scott, 2010).

The prior distribution of $\boldsymbol{\beta}$ for the Bayesian bridge model is a product of independent exponential power priors:

$$p(\boldsymbol{\beta}|\tau, \alpha) \propto \prod_{j=1}^{p} \exp(-|\beta_j/\tau|^\alpha) \quad \tau = \nu^{-1/\alpha}. \tag{2.1}$$

Let $p(\lambda_j)$ be the density of $2S_{\alpha/2}$ where S_α is the Lévy alpha-stable distribution. Then, using Lévy processes and scale mixtures of normal representation discussed in Polson and Scott (2012), a joint prior distribution of regression parameter $\boldsymbol{\beta}$ and local shrinkage parameter $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_j)$ are represented as follows:

$$p(\boldsymbol{\beta}, \Lambda|\tau, \alpha) \quad \propto \quad \prod_{j=1}^{p} \exp\left(-\frac{\beta_j^2}{2\tau^2}\lambda_j\right) p(\lambda_j). \tag{2.2}$$

Prior distributions of the remaining parameters (σ^2, α, τ) are defined as follows:

$$\begin{aligned}
\sigma^2 &\sim \text{Inverse-Gamma}\left(\frac{a_0}{2}, \frac{b_0}{2}\right) \\
\alpha &\sim \text{Uniform}\,(0, 1) \\
\nu &\sim \text{Gamma}\,(c_0, d_0)
\end{aligned}$$

where $\tau = \nu^{-1/\alpha}$.

[1] According to Polson and Scott (2012), "all totally monotone penalty functions that vanish at zero correspond to priors that can be represented in terms of a subordinator" (Polson and Scott, 2012, 292).

Then, the posterior distribution of the Bayesian bridge linear regression model is

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2, \Lambda, \alpha, \nu | \boldsymbol{y}, \mathbf{X}) &\propto p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \Lambda|\tau, \alpha) p(\sigma^2) p(\alpha) p(\nu) \\ &\propto \exp\Big[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \Big] \prod_{j=1}^{p} \exp\left(-\frac{\beta_j^2}{2\tau^2}\lambda_j \right) p(\lambda_j) \\ &\quad \times \left(\frac{1}{\sigma^2} \right)^{\frac{a_0}{2}+1} \exp\left(-\frac{b_0}{2\sigma^2} \right) \nu^{c_0 - 1} \exp(-d_0 \nu). \end{aligned} \tag{2.3}$$

2.2 The Hidden Markov Bayesian Bridge Model

We utilize a hidden Markov model (HMM) to detect major structural breaks in times series data in various forms. As shown by many authors (Baum et al., 1970; Chib, 1998; Robert, Ryden and Titterington, 2000; Cappe, Moulines and Ryden, 2005; Scott, James and Sugar, 2005; Frühwirth-Schnatter, 2006; Teh et al., 2006), HMM efficiently detects change-points in various regression models using the conditional independence of data given hidden states.

Let $\mathbf{S}$ denote a vector of hidden state variables where s_t is an integer-valued hidden state variable at t

$$\mathbf{S} = \{(s_1, \ldots, s_n) : s_t \in \{1, \ldots, M\}, t = 1, \ldots, n\},$$

and $\mathbf{P}$ as a forward moving $M \times M$ transition matrix where $\mathbf{p}_i$ is the ith row of $\mathbf{P}$ and M is the total number of hidden states. Then, the data density of HMBB can be written as follows:

$$\begin{aligned} \prod_{t=1}^{n} p(y_t | \mathbf{x}_t, \boldsymbol{\beta}, \sigma^2, \Lambda, \alpha, \nu) &= \int p(y_1 | s_1, \mathbf{x}_1, \boldsymbol{\beta}_1, \sigma_1^2, \Lambda_1, \alpha_1, \nu_1) \\ &\quad \times \prod_{t=2}^{n} \sum_{m=1}^{M} p(y_t | \mathbf{x}_t, s_t, \boldsymbol{\beta}_{s_t}, \sigma_{s_t}^2, \Lambda_{s_t}, \alpha_{s_t}, \nu_{s_t}) \\ &\quad \times \Pr(s_t = m | s_{t-1}, \mathbf{Y}_{t-1}, \mathbf{X}_{t-1}, \boldsymbol{\beta}, \sigma^2, \Lambda, \alpha, \nu) d\mathbf{S} \end{aligned}$$

where $\mathbf{Y}_{t-1}$ and $\mathbf{X}_{t-1}$ indicate all the observed data up to $t-1$. The state transition is defined as a forward-moving first-order discrete Markov process:

$$\begin{aligned} s_t | \mathbf{P}, \boldsymbol{\pi} &\sim \text{Markov}(\mathbf{P}, \boldsymbol{\pi}) \\ \underbrace{\mathbf{P}}_{M \times M} &= (\mathbf{p}_1, \ldots, \mathbf{p}_M) \\ \mathbf{p}_i &\sim \text{Dirichlet}(\alpha_{i,1}, \ldots, \alpha_{i,M}) \text{ for all } i < M. \end{aligned}$$

To illustrate the model introduced above, we discuss a case with one change-point. Suppose that we know the location of the structural break (i.e., a vector $\mathbf{S}$ is known). Then, we can write

a posterior density as

$$
\begin{aligned}
&p(\lambda_j)p(\sigma^2)p(\alpha)p(\nu)\prod_{t=1}^{n}\prod_{m=1}^{2}\left\{\exp\left(-\frac{1}{2\sigma_m^2}(y_t-\mathbf{x}_t^\top\boldsymbol{\beta}_m)^2\right)\prod_{j=1}^{p}\exp\left(-\frac{\beta_{m,j}^2}{2\tau_m^2}\lambda_{m,j}\right)\right\}^{\mathbf{1}\{s_t=m\}}\\
&=p(\lambda_j)p(\sigma^2)p(\alpha)p(\nu)\prod_{1\le t\le t^\star}\left\{\exp\left(-\frac{1}{2\sigma_1^2}(y_t-\mathbf{x}_t^\top\boldsymbol{\beta}_1)^2\right)\prod_{j=1}^{p}\exp\left(-\frac{\beta_{1,j}^2}{2\tau_1^2}\lambda_{1,j}\right)\right\}\\
&\quad\times\prod_{t^\star<t'\le n}\left\{\exp\left(-\frac{1}{2\sigma_2^2}(y_{t'}-\mathbf{x}_{t'}^\top\boldsymbol{\beta}_2)^2\right)\prod_{j=1}^{p}\exp\left(-\frac{\beta_{2,j}^2}{2\tau_2^2}\lambda_{2,j}\right)\right\}
\end{aligned}
$$

where $t^\star = \underset{t:s_t=1}{\arg\max}\, s_t$. (Note that $\mathbf{S}$ is ordered, so $s_t = 1$ for all $1 \le t \le t^\star$). The above posterior density illustrates that if we were to know the change point location(s) a priori, it would be equivalent to fit two separate regression models with shrinkage prior to the data before and after the break, thus enabling time-varying shrinkage.

However, we usually do not have prior knowledge about $\mathbf{S}$, if not the number of breaks. Instead, the proposed model recovers $\mathbf{S}$ using Chib (1998)'s algorithm together with other model parameters such as regression coefficients and shrinkage parameters.

2.3 Posterior Computation

We discuss the sampling algorithm of HMBB, highlighting three major modifications from Polson and Scott (2012). For notational simplicity, we denote segmented data corresponding to state m as $\boldsymbol{y}_m$ and $\mathbf{X}_m$. Also, n_m denotes the number of observations pertaining to state m.

- Sampling $p(\boldsymbol{\beta}|\alpha, \boldsymbol{\Lambda}, \sigma^2, \tau, \mathbf{P}, \mathbf{S}, \boldsymbol{y})$

  The posterior of $\boldsymbol{\beta}$ follows the multivariate normal distribution, which is given by

$$
\boldsymbol{\beta}_m|\sigma^2, \lambda_m, \alpha_m, \tau, \mathbf{P}, \mathbf{S}, \boldsymbol{y}_m \sim \mathcal{N}_p\left(\frac{\mathbf{V}\mathbf{X}_m'\boldsymbol{y}_m}{\sigma_m^2}, \mathbf{V} = \left(\mathbf{X}_m'\mathbf{X}_m + \frac{\sigma_m^2}{\tau^2}\lambda_m\mathbf{I}\right)^{-1}\right). \tag{2.4}
$$

  1. When $n \gg p$, we can directly sample $\boldsymbol{\beta}_m$ from Equation 2.4.
  2. When $n \le p$, inverting the covariance matrix $(p \times p)$ is very expensive, which costs roughly $O(p^3)$. Instead, we use the singular value decomposition (SVD) of the design matrix: $\mathbf{X}_m = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ where $\mathbf{U} \in O_{n_m \times n_m}$, $\mathbf{V} \in O_{p \times n_m}$ and $\mathbf{D} = \text{diag}(d_1, \ldots, d_{n_m})$. Let $\boldsymbol{\lambda}_m = \text{diag}(\{\lambda_{k,m}\sigma_m^2/\tau_m^2\}_{k=1}^p)$ be a diagonal matrix of penalty parameters. We further define $\overline{\mathbf{D}} = [\mathbf{D}|\mathbf{0}_{n_m \times (p-n_m)}]$ and $\overline{\mathbf{V}} = [\mathbf{V}|\mathbf{0}_{p \times (p-n_m)}]$ as augmented matrices.

     When the design matrix is not (column) full-rank, which is the case with $p > n$, this operation allows $\overline{\mathbf{D}}$ to have dimension $n_m \times p$. This is crucial since $\boldsymbol{\lambda}_m$ has dimension

of $p \times p$. Then,

$$\begin{aligned}(\mathbf{X}_m^\top \mathbf{X}_m + \boldsymbol{\lambda}_m) &= \mathbf{V}\mathbf{D}^\top \mathbf{U}^\top \mathbf{U}\mathbf{D}\mathbf{V}^\top + \boldsymbol{\lambda}_m \\ &= \mathbf{V}\mathbf{D}^\top \mathbf{D}\mathbf{V}^\top + \boldsymbol{\lambda}_m \\ &= \overline{\mathbf{V}}(\overline{\mathbf{D}}^\top \overline{\mathbf{D}} + \boldsymbol{\lambda}_m)\overline{\mathbf{V}}^\top.\end{aligned}$$

First, the posterior variance is given by

$$\boldsymbol{\Sigma}_m = (\overline{\mathbf{V}}(\overline{\mathbf{D}}^\top \overline{\mathbf{D}} + \boldsymbol{\lambda}_m)\overline{\mathbf{V}}^\top)^{-1} = \overline{\mathbf{V}}(\overline{\mathbf{D}}^\top \overline{\mathbf{D}} + \boldsymbol{\lambda}_m)^{-1}\overline{\mathbf{V}}^\top.$$

This quantity is easy to compute since $(\overline{\mathbf{D}}^\top \overline{\mathbf{D}} + \boldsymbol{\lambda}_m)$ is a diagonal matrix of the form

$$\overline{\mathbf{D}}^\top \overline{\mathbf{D}} + \boldsymbol{\lambda}_m = \text{diag}(\{d_k^2 + \lambda_{k,m}\sigma_m^2/\tau_m^2\}_{k=1}^{n_m}, \{\lambda_{k',m}\sigma_m^2/\tau_m^2\}_{k'=n_m+1}^{p}).$$

Second, the posterior mean $\boldsymbol{\mu}_m$ is then given by

$$\begin{aligned}\boldsymbol{\mu}_m &= \boldsymbol{\Sigma}_m \mathbf{X}_m^\top \boldsymbol{y}_m/\sigma_m^2 \\ &= \overline{\mathbf{V}}(\overline{\mathbf{D}}^\top \overline{\mathbf{D}} + \boldsymbol{\lambda}_m)^{-1}\overline{\mathbf{D}}^\top \mathbf{U}^\top \boldsymbol{y}_m/\sigma_m^2.\end{aligned}$$

Then, using $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$, the sampling is given by

(a) Let $\mathbf{A} = \overline{\mathbf{V}}(\overline{\mathbf{D}}^\top \overline{\mathbf{D}} + \boldsymbol{\lambda}_m)^{-1/2}$.

(b) Draw z_k from standard normal for $k = 1, \ldots, p$.

(c) Update $\boldsymbol{\beta}_m$ by $\boldsymbol{\beta}_m \leftarrow \boldsymbol{\mu}_m + \boldsymbol{A}\boldsymbol{z}$

(d) Repeat the above steps for all $m = \{1, \ldots, M\}$

- Sampling $p(\alpha|\boldsymbol{\Lambda}, \boldsymbol{\beta}, \sigma^2, \tau, \mathbf{P}, \mathbf{S}, \boldsymbol{y})$

  It is well known that when $0 < \alpha \leq 1$, the classical bridge estimator has the variable selection feature (Murphy, 2012). We believe that Polson, Scott and Windle (2014) use the support of $0 < \alpha \leq 1$ to set the bridge model between two polar cases of the lasso ($\alpha = 1$) and the subset selection method ($\alpha = 0$) in classical statistics. However, the support of $0 < \alpha \leq 1$ does not guarantee the variable selection feature in Bayesian framework because the posterior of regression parameters depends on the complicated order statistics as shown by Polson and Scott (2010). Generally speaking, one-group model of Bayesian regularization methods cannot guarantee the variable selection feature (Polson and Scott, 2010; Hahn and Carvalho, 2015). Monte Carlo experiments also show that the constraint of $0 < \alpha \leq 1$ often produces larger RMSEs than Bayesian bridge estimates with $0 < \alpha \leq 2$.

  Polson, Scott and Windle (2014) suggest a random-walk Metropolis Hastings (MH) sampler.

However, a random-walk MH sampler can produce highly correlated draws, slowing down the mixing of the Markov chain. We use a Griddy Gibbs sampler (Tanner, 1996) for the sampling of α because α is univariate and its support is bounded by $(0, 2]$. For a Griddy Gibbs sampler, we evaluate the unnormalized density of the conditional posterior at points between 0.1 and 2.0 at an interval of 0.02. Then, we draw α_m from the points on the grid with probability proportional to the empirical density.

- Sampling $p(\tau|\mathbf{\Lambda}, \boldsymbol{\beta}, \alpha, \sigma^2, \mathbf{P}, \mathbf{S}, \boldsymbol{y})$.

  Sample ν first and then transform ν to τ.

$$\nu_m \sim \text{Gamma}(c, d)$$
$$\tau_m = \nu^{-\frac{1}{\alpha_m}}$$

  where $c = c_0 + p/\alpha_m$ and $d = d_0 + \sum_{j=1}^{p} |\beta_{j,m}|^{\alpha_m}$.

- Sampling $p(\beta_0|\mathbf{\Lambda}, \boldsymbol{\beta}, \alpha, \tau, \sigma^2, \mathbf{P}, \mathbf{S}, \boldsymbol{y})$

  We separately estimate the intercepts for each regime. Since all the data are centered, this estimate does not affect updates for other parameters. But posterior samples for the intercept are useful for making prediction on the original scale.

$$\beta_{0m} \leftarrow \overline{\boldsymbol{y}}_m - \overline{\mathbf{X}}_m^\top \boldsymbol{\beta}_m$$

  where

$$\overline{\boldsymbol{y}}_m = \frac{\sum_{t=1}^{n} \mathbf{1}\{s_t = m\} y_t}{\sum_{t=1}^{n} \mathbf{1}\{s_t = m\}}, \quad \text{and} \quad \overline{\mathbf{X}}_{m,j} = \frac{\sum_{t=1}^{n} \mathbf{1}\{s_t = m\} \mathbf{X}_{m,tj}}{\sum_{t=1}^{n} \mathbf{1}\{s_t = m\}}. \tag{2.5}$$

- Sampling $\mathbf{S}|\mathbf{\Lambda}, \boldsymbol{\beta}, \alpha, \tau, \sigma^2, \mathbf{P}, \boldsymbol{y}$ Sample $\mathbf{S}$ recursively using Chib (1998)'s algorithm. Using Bayes' Theorem, Chib (1998) shows that

$$p(s_t|\mathbf{S}^{t+1}, \mathbf{\Theta}) \quad \propto \quad \underbrace{p(s_t|\mathbf{\Theta}, \mathbf{Y}_{1:t}, \mathbf{X}_{1:t})}_{\text{State probabilities given all data up to } t} \quad \overbrace{p(s_{t+1}|s_t, \mathbf{\Theta})}^{\text{Transition probability at } t} .$$

  The second part on the right hand side is a one-step ahead transition probability at t, which can be obtained from a sampled transition matrix ($\mathbf{P}$). The first part on the right hand side is state probabilities given all data, which can be simulated via a forward-filtering-backward-sampling algorithm as shown in Chib (1998).

- Sampling from $\mathbf{P}|\mathbf{\Lambda}, \boldsymbol{\beta}, \alpha, \tau, \sigma^2, \mathbf{S}, \boldsymbol{y}$

$$p_{kk} \sim \mathcal{B}eta(a_0 + j_{k,k} - 1, b_0 + j_{k,k+1})$$

where p_{kk} is the probability of staying when the state is k, and $j_{k,k}$ is the number of jumps from state k to k, and $j_{k,k+1}$ is the number of jumps from state k to $k+1$.

3 Simulation Studies

3.1 Simulation Design

In this section, we conduct a series of Monte Carlo simulations to test the performance of the proposed models in high-dimensional regression with change-points when $n \leq p$. To save space, we only report test results from high dimensional data with a change-point. Additional simulation results without change-points, which illustrates our implementation against Lasso, elastic net and ridge, are reported in the supplementary material (SI Section 3).

Following Donoho (2005) and Donoho and Stodden (2006), simulated data vary by two dimensions: the level of underdeterminedness ($\delta = n/p$) and the level of sparsity ($\rho = k/n$) where n is the number of observations and k is the number of non-sparse predictors. To make interpretation simple, we fix the number of predictors (p) at 200 and vary n from 10 to 200, and k from 1 to 200 so that both the level of underdeterminedness ($\delta = n/p$) and the sparsity level ($\rho = k/n$) take 50 equidistance points on the interval $[0.1, 1]$.

Then, we use an underlying model of $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, x_{ij} \sim N(0,1), \boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, 4^2\mathbf{I}_n)$ by varying δ and ρ. The change point is set at the mid point, $\lfloor n/2 \rfloor$, and coefficients are drawn independently for each regime. Based on the value of k, regression coefficients are set as $\boldsymbol{\beta}_{1:k} \sim \text{Uniform}(0, 50)$ and $\boldsymbol{\beta}_{k+1:p} = \mathbf{0}$.[2] We create 50^2 unique pairs of (δ, ρ) and for each pair (δ, ρ) and simulate 20 datasets from the same underlying model. In total, the number of simulated data sets is $50^2 \times 20 = 50,000$.

Since there exists no comparable method that implements change-point analysis of regularization methods, we develop two hybrid lasso estimates as benchmark:

- Lasso (Estimate): Two step Lasso estimates using the estimated break point. In the first step, a break detection is done using the lasso residuals and the HMM. In the second stage, the lasso method is applied to subset data, respectively.
- Lasso (Oracle): Separate Lasso estimates using the true break point

For HMBB estimates, we set the correct number of break, but the location of the break point is determined by HMBB. The point of comparison is to see (1) whether a HMBB with an unknown break point outperforms a two-step approach of Lasso (Estimate) and (2) how closely HMBB performs against Lasso (Oracle) that uses the ground truth knowledge about a break point.

[2] We also consider a correlated design matrix where $\text{Cov}(X_{ij}, X_{ij'}) = \Sigma_{j,j'} = \eta$ for $j \neq j'$ and $\Sigma_{jj} = 1$. We draw $\mathbf{X}_i$ from a multivariate normal distribution with mean zero and covariance Σ, $\mathbf{X}_i \sim \text{MVN}(\mathbf{0}, \Sigma)$. We take $\eta = 0.3, 0.7$. Results for simulations based on correlated design matrix are reported in SI Section 3.4 for a change-point case and in SI Section 3.3 for a no change-point case.

We evaluate performance of different regularization methods using the criteria summarized in Table 1. First, Prediction Loss is related with the persistency or risk consistency (e.g., see Greenshtein and Ritov, 2004) – one of the oracle properties that high-dimensional regression estimator wishes to satisfy. Second, Normalized Estimation Loss captures parameter consistency. Achieving high performance on Normalized Estimation Loss usually requires stronger assumptions than those for the prediction loss. Last, Cross-validation Loss checks out-of-sample predictive accuracy. We conduct a 2-fold cross-validation prediction to compute the cross-validation loss.

Table 1: Simulation Performance Criteria

Metric	Formula	Property
Prediction Loss	$\mathcal{L}_{\text{pred}}(\widehat{\boldsymbol{\beta}};\boldsymbol{\beta}^\star) = \frac{1}{n}\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^\star\|^2$	in-sample model fit
Normalized Estimation Loss	$\mathcal{L}_2(\widehat{\boldsymbol{\beta}};\boldsymbol{\beta}^\star) = \frac{\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^\star\|^2}{\|\boldsymbol{\beta}^\star\|^2}$	parameter consistency.
Cross-validation Loss	$\mathcal{L}_{\text{CV}}(\widehat{\boldsymbol{y}};\boldsymbol{y}^\star) = \frac{1}{\lvert\mathcal{I}^c\rvert}\sum_{t\in\mathcal{I}^c}(y_t - \mathbf{X}_t^\top\widehat{\boldsymbol{\beta}})^2$	out-of-sample predictive accuracy

3.2 Simulation Results

Table 2: Average Estimation Loss from High Dimensional Data with Change-points. The reported numbers are from simulations of 50,000 data sets. True data has one break. MCMC simulation for HMBB is 100 and burn-in is 100.

	Prediction Loss	
Method	Mean	SD
HMBB	0.35	0.11
Lasso (Estimate)	0.16	0.06
Lasso (Oracle)	0.15	0.05
	Normalized Estimation Loss	
Method	Mean	SD
HMBB	0.10	0.00
Lasso (Estimate)	0.12	0.01
Lasso (Oracle)	0.09	0.02
	Cross-validation Loss	
Method	Mean	SD
HMBB	0.45	0.11
Lasso (Estimate)	0.53	0.14
Lasso (Oracle)	0.46	0.16

Table 2 summarizes the results of the simulation for a single change-point case. HMBB produces slightly larger values of prediction loss than Lasso (Estimate) and Lasso (Oracle). Panel (A) in

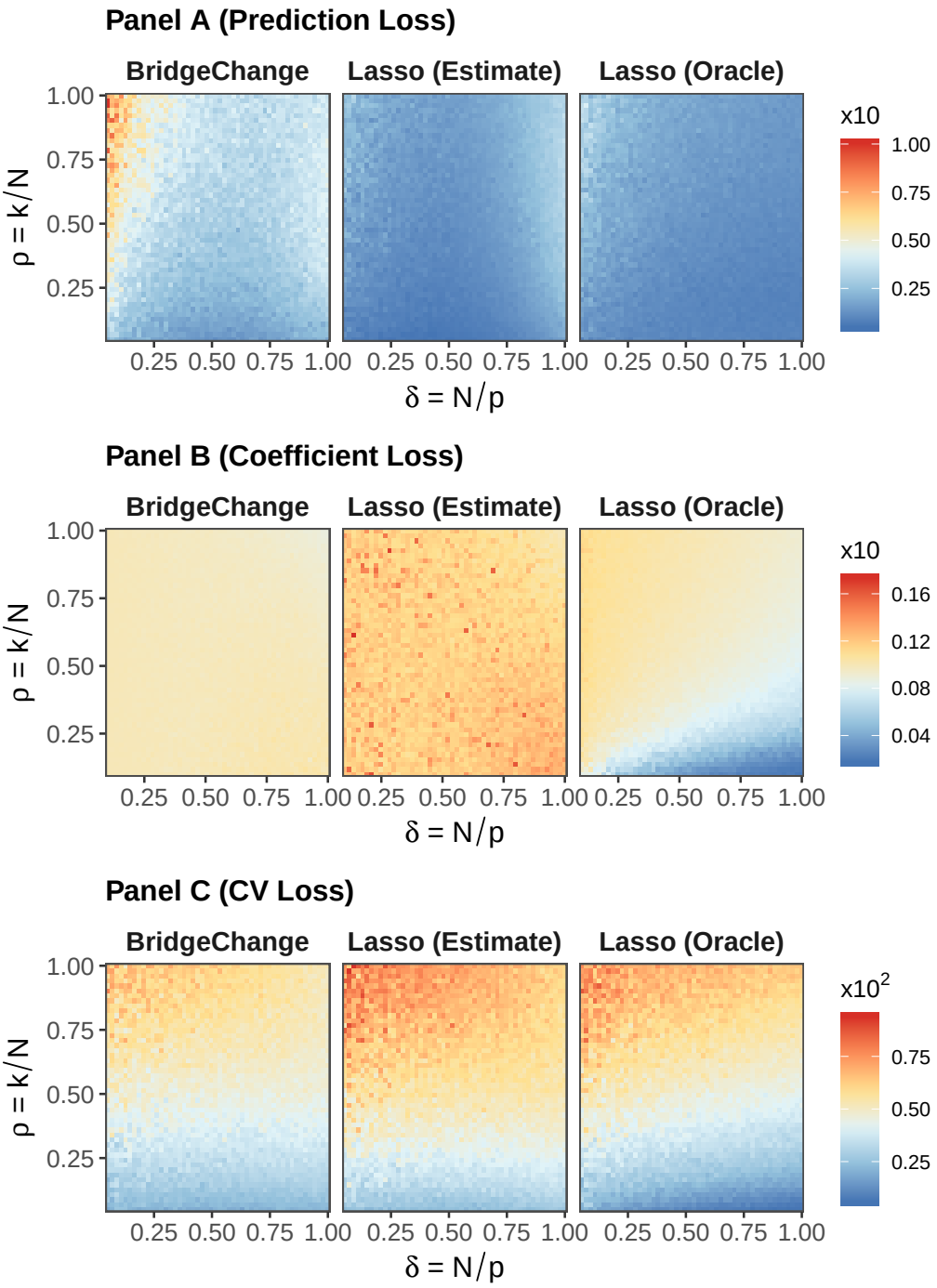

Figure 2: Results of Simulation Studies using Univariate Time Series Data with One Change-point. **Panel (A)**: Prediction Loss, $\mathcal{L}_{\text{pred}}(\widehat{\boldsymbol{\beta}}; \boldsymbol{\beta}^{\text{true}})$. **Panel (B)**: Normalized Estimation Loss, $\mathcal{L}_2(\widehat{\boldsymbol{\beta}}; \boldsymbol{\beta}^{\text{true}}) = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\text{true}}\|_2 / \|\boldsymbol{\beta}^{\text{true}}\|_2$. **Panel (C)**: Cross-validation Loss, $\mathcal{L}_{\text{CV}}(\widehat{\boldsymbol{y}}^{\text{test}}; \boldsymbol{y}^{\text{test}})$. We fix $p = 200$ and vary α and ρ between 0.05 and 1. Thus, each cell in the graph represents a data with (N, p, k). We simulate 25 data sets from each (N, p, k) and take the median error.

Figure 2 shows that the source of the problem is when data is highly small and the number of non-sparse signal is close to the number of observations. However, for the normalized estimation

loss and the cross-validation loss, HMBB outperforms Lasso (Estimate). To our surprise, HMBB slightly outperforms Lasso (Oracle) in the cross-validation loss. Panel (C) of Figure 2 implies that Lasso (Oracle) predicts out-of-sample data relatively well when k/n is small. However, as k/n becomes larger, the overfitting of Lasso (Oracle) produces poor predictive values of out-of-sample data compared to HMBB. This result can be interpreted in two ways. First, it implies that by taking a known break point as fixed, Lasso (Oracle) produces overfitting. Second, this result can be connected with the theoretical threshold of penalized regression models with ℓ_1 norm. According to Donoho and Stodden (2006), "there is a breakdown point for standard model selection schemes, such that model selection only works well below a certain critical complexity level" (Donoho and Stodden, 2006, 1).

4 Applications

4.1 Estimating Heterogenous Causal Effects

Nunn and Qian (2014) examine the effect of food aid on civil conflicts using the instrumental variable design. The authors exploit two exogenous variations in this study. Specifically, the first is exogenous time variation in US wheat production, which is driven by weather conditions in the US. The second is cross-sectional variations in a country's likelihood of being a US food aid recipient. Then, they use the interaction of last year's wheat production and the frequency of a country's US food aid receipt as an instrument, which is denoted by z_{irt}.

The original study reports 2SLS estimates. The first and second stage equations used in Nunn and Qian (2014) are given by

$$y_{irt} = \beta d_{irt} + \mathbf{x}_{irt}\Gamma + \varphi_{rt} + \psi_{ir} + \nu_{irt} \tag{4.1}$$

$$d_{irt} = \alpha z_{irt} + \mathbf{x}_{irt}\Gamma + \varphi_{rt} + \psi_{ir} + \epsilon_{irt}. \tag{4.2}$$

Here i denotes a country, r denotes a region, and t is a year. φ_{rt} and ψ_{ir} are fixed-effects at the region-year and country-year levels, respectively. d_{itr} is the endogenous variable (the quantity of wheat aid) and $\mathbf{x}_{itr}$ includes a set of exogenous control variable. The total number of observations is 4,089 covering 125 non-OECD countries during the 36 years, 1971-2006.

There are two potential methodological concerns in Nunn and Qian (2014)'s analysis. The first concern is the high-dimensionality of the data. While country-level observations vary from 5 to 36 years, the number of control variables including all dummy indicators is as large as 352. The second concern is temporal heterogeneity in the first-stage equation parameters (Equation 4.2). If the effects of the (included and excluded) instruments vary across time, causal inference using the 2SLS must take into account these parameter heterogeneity in Equation 4.2.

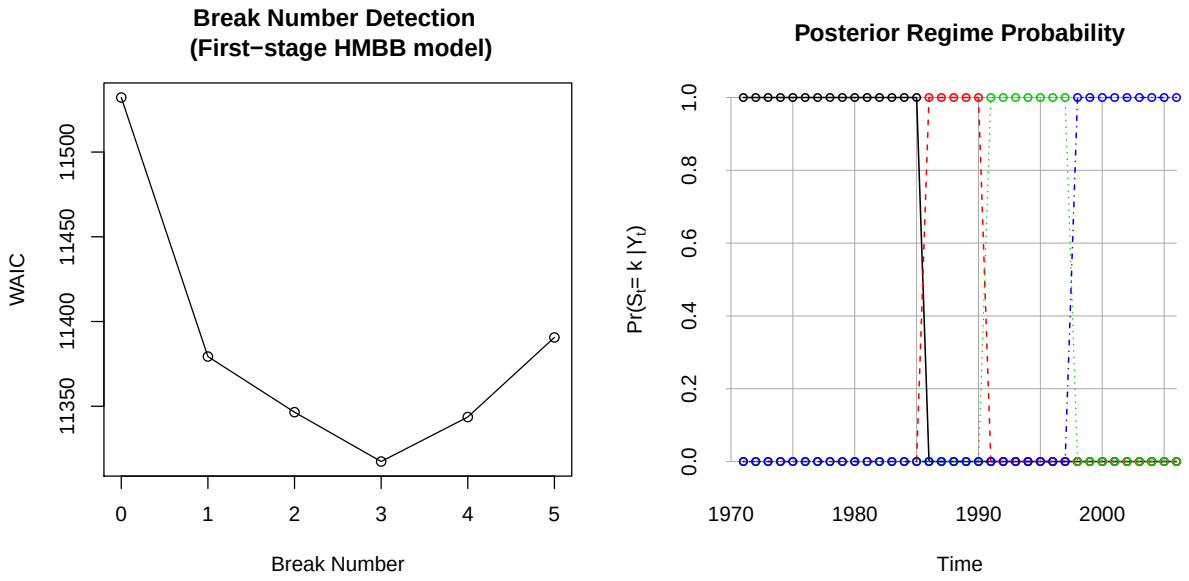

Figure 3: Break Detection in the First Stage Regression

We revisit the first-stage estimation of the original analysis using HMBB. The left panel of Figure 3 shows WAIC scores for six HMBBs with a varying number of breaks up to five. Adding breaks improves the predictive accuracy of the model until three breaks (i.e. four regimes) and then it deteriorates the predictive accuracy. The right panel shows that estimated break points are 1986, 1991, and 1998.

The left panel of Figure 4 shows the time-varying effects of the instrument on the endogenous variable (α in Equation 4.2). The blue dots show the pooled estimate using double machine learning method and red dots indicate regime-specific estimates using HMBB. We can see that the effect of the instrument on the treatment variable shifts dramatically over time, showing the largest effect between 1987 and 1991, followed by the period of 1971-1985. After 1992, which corresponds to the post-Cold War period, the effect of the instrument on the endogenous variable diminishes significantly toward 0.

Then, how do these regime-changing effects of the instrument affect causal effects of food aid on civil conflicts (β)? To answer this question, we first partitioned data based on the four identified regimes and then apply the debiasing method proposed for the IV regression with high-dimensional covariates by Chernozhukov, Hansen and Spindler (2015) and Chernozhukov et al. (2018). This "double machine learning" (DML) method, which is implemented in hdm package in R, guarantees that point estimates are not biased due to regularization and returns proper confidence intervals.[3]

The blue dots in the right panel of Figure 4 indicate the original pooled 2SLS estimate and red dots are regime-specific β's (DML after HMBB). It clearly shows that the pooled causal estimate

[3]We report the results from this method on the entire dataset in SI Section 8.

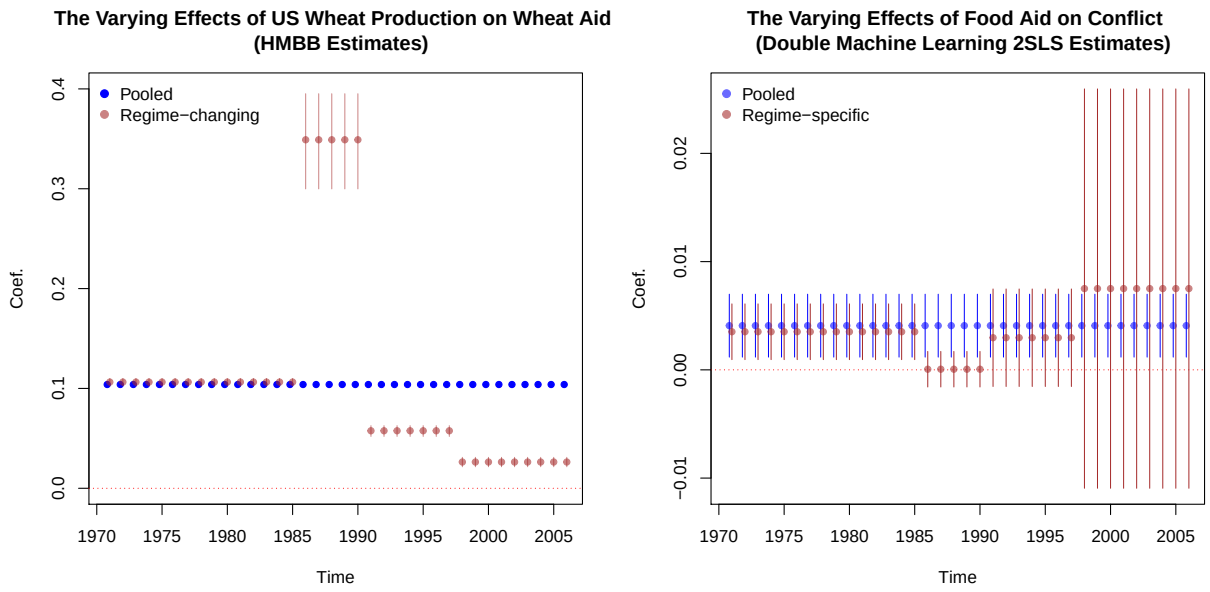

Figure 4: Regime-changing α and β in the first- and second-stage equations in Equation 4.1 and Equation 4.2, respectively: Vertical bars indicate 95% credible intervals (left) and 95% confidence intervals (right).

of 0.004 represents only two regimes (1971-1985 and 1992-1997). The causal effect of food aid on civil conflicts is close to 0 during 1987-1991, which corresponds to the period of the largest partial correlation between the instrument and the endogenous variable as shown in the left panel of Figure 4. If we take the size of the partial correlation as reflecting the strength of the instrument, there is strong evidence to suspect the existence of sizable causal effects of food aid on intrastate conflicts.

It is notable to see the post-1991 causal effect estimates of food aid on intrastate conflicts include 0 in their 95% confidence intervals. The break point of 1991 coincides with many important international events that affect US food aid and intrastate conflicts at the same time such as the revolutions within Eastern European countries in 1989, the dissolution of the Soviet Union, and the end of the Cold War in 1991. That is, the diminishing effect of food aid on intrastate conflicts after 1991 is likely to be caused by systematic changes in the international system and corresponding changes in the US foreign policy toward developing countries.[4]

[4]This possibility is also noted by Nunn and Qian (2014) and they addressed this issue by interacting "food aid and a Cold War indicator variable"(Nunn and Qian, 2014, 1662). They found the coefficient of the interaction term to be "negative, moderate in magnitude, but statistically insignificant" (Nunn and Qian, 2014, 1662) and did not further investigate the influence of the Cold War. First, the lack of statistical significance does not mean anything toward the tested hypothesis. Second, the single interaction term of the Cold War dummy with the instrument does not suffice to check time-varying causal effects. The effects of the Cold War are much wider and other nuisance parameters must have different associations with the endogenous variable and the dependent variable during and after the Cold War. In this regard, the room for dynamic misspecification is huge for the "negative, moderate in magnitude, but statistically insignificant" results.

Table 3: *Estimates of α and β in the first- and second-stage equations in Equation 4.1 and Equation 4.2, respectively. Est. refers to point estimates, 95%CI Low (High) refers to lower (upper) bound of 95% confidence intervals.*

Data	Parameter	Est.	95%CI Low	95% CI High
Pooled	α	0.104	0.103	0.105
Regime 1	α	0.106	0.103	0.109
Regime 2	α	0.350	0.296	0.407
Regime 3	α	0.058	0.054	0.065
Regime 4	α	0.026	0.023	0.031
Pooled	β	0.004	0.001	0.007
Regime 1	β	0.004	0.001	0.006
Regime 2	β	0.000	-0.002	0.002
Regime 3	β	0.003	-0.002	0.007
Regime 4	β	0.008	-0.011	0.026

4.2 Changepoint Analysis and Variable Selection

Alvarez, Garrett and Lange (1991) study the effect of labor party government on economic growth using a time series cross-national data covering 16 OECD countries for the period of 1970 - 1984. The key finding of Alvarez, Garrett and Lange (1991) was the positive interaction effect of the left-party government size with the centralized labor. This study is one of the most important findings in comparative politics, producing many subsequent studies (Alvarez, Garrett and Lange, 1991; Beck, Katz and Alvarez, 1993; Beck and Katz, 1995; Western, 1998). Because of the short time series (15 years), they included only one interaction term among 6 predictors and did not examine time-varying effects.[5] The data cover 16 OECD countries for the period of 1970 – 1984. The dependent variable is the annual growth rate and independent variables cover political economic covariates of economic growth.[6]

We examine the full interaction model with 21 $(6 + \binom{6}{2})$ predictors that include all pairwise interactions.[7] Our goal is to *select* important time-varying predictors of economic growth out of many possible predictors via the posterior summarization of HMBB using the DSS method (Hahn

[5]The data are obtained from `pcse` package in R.

[6]The independent variables are as follows:

1. `lagg1`: The lagged growth rate
2. `opengdp`: weighted OECD demand measured by OECD growth rates
3. `openex`: weighted OECD export
4. `openimp`: weighted OECD import
5. `leftc`: The cabinet composition of left-leaning parties
6. `central`: The degree of labor organization encompassment is measured by summing standardized scores for the density and centralization of union movements in each of the countries

[7]We demean the data by year to remove year fixed-effects. Country-wise demeaning is not feasible due to the time invariant covariate (`central`).

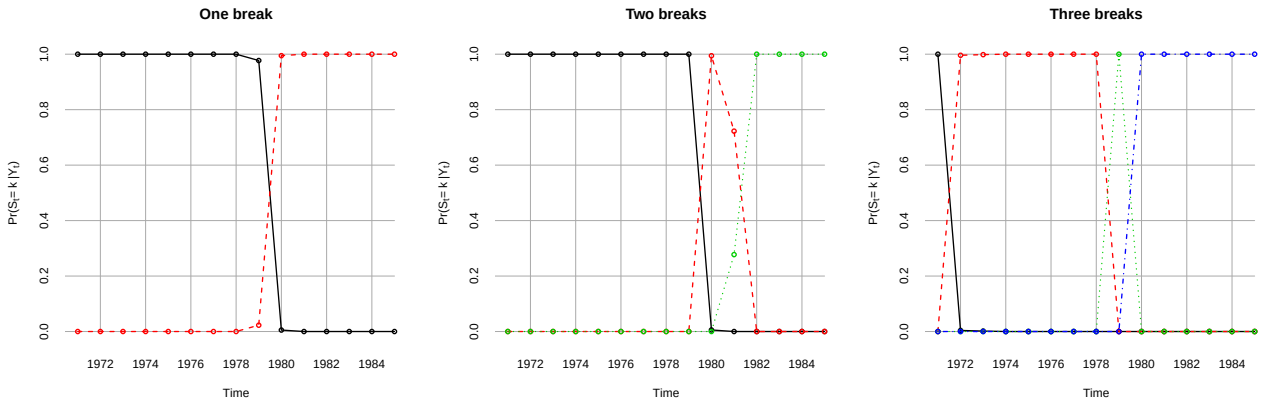

Figure 5: *Transition of Latent States by the Number of Breaks: The year of 1979 is consistently detected as a break point. Adding more than one break produces singleton states. The fully interacted Alvarez, Garrett and Lange (1991) model is used for HMBB analysis.*

and Carvalho, 2015). The DSS method is a hybrid Bayesian method to choose strong signals from noisy data by optimizing the posterior mean with regard to an ℓ_0 function. Its goal is to find a compromise between prediction accuracy and inferential parsimony by considering variable selection as a problem of posterior summarization. Following Hahn and Carvalho (2015), we write the DSS loss function for HMBB estimates at regime m

$$\mathcal{L}(\boldsymbol{\gamma}_m) = \arg\min_{\boldsymbol{\gamma}_m} \overbrace{\|\mathbf{X}_m\boldsymbol{\beta}_m^* - \mathbf{X}_m\boldsymbol{\gamma}_m\|_2^2}^{\text{squared prediction loss}} + \overbrace{\lambda\|\gamma_m\|_0}^{\text{parsimony penalty}} \tag{4.3}$$

where $\mathbf{X}_m\boldsymbol{\beta}_m^*$ is the fitted value of HMBB at regime m. Then, we use the adaptive lasso method to find $\boldsymbol{\gamma}_m$ (Zou, 2006).

First, the break number diagnostics using WAIC or the approximate log marginal likelihoods indicate a strong sign of a single break in 1979. Adding more than one break produces singleton states (*i.e.* latent states with only one observation) as shown in Figure 5. That is, models with more than one break do not improve our substantive knowledge much from the one we obtain from the one break model.

Next, we examine the post-selected predictors for each regime. The left panel of Figure 6 shows DSS estimates of all 21 regression parameters. It is clear that most coefficients show dramatic shifts toward zero after 1979. The right panel of Figure 6 zooms in left-party government-related parameters, which are one of the key explanatory variables of Alvarez, Garrett and Lange (1991). Strikingly, effects of left-party government-related parameters disappear after 1979. That is, direct and indirect effects of government partisanship (measured by the cabinet composition of left-leaning parties) existed only up until 1979. Thus, we can conclude that Alvarez, Garrett and Lange (1991)'s original claim on the conditional effect of government partisanship on economic growth does not

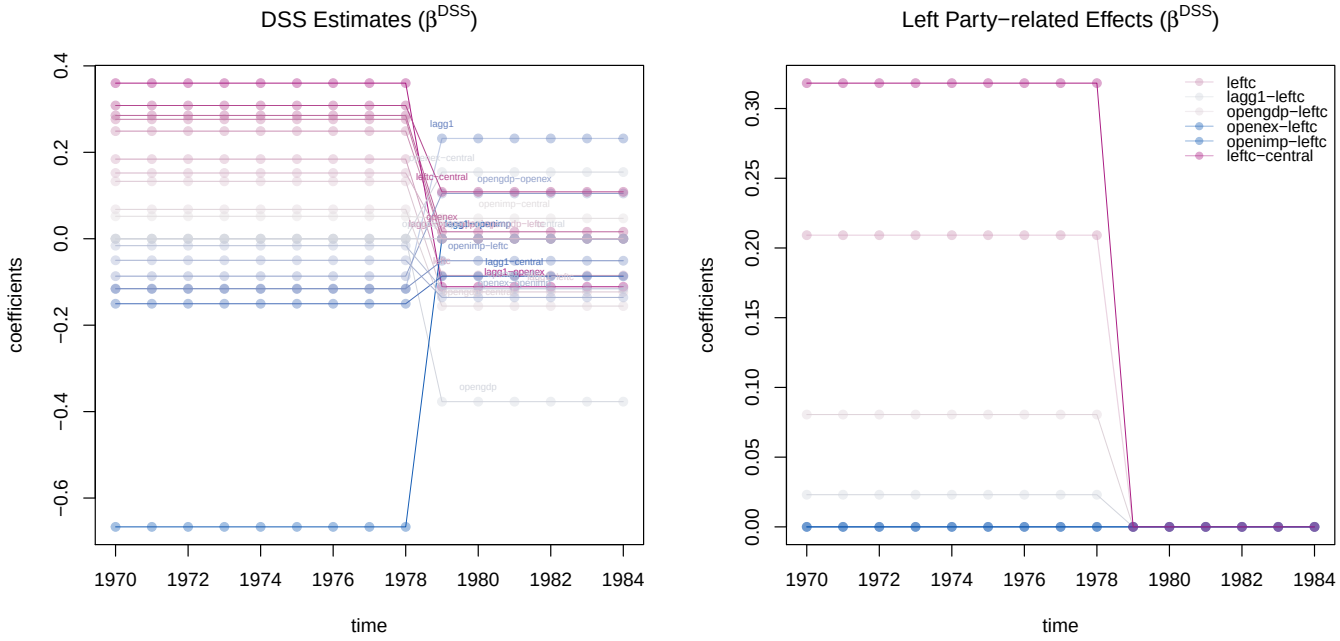

Figure 6: *A Latent Regime Change and Regime-changing Covariate Effects in the Fully Interacted Alvarez, Garrett and Lange (1991) Model*

hold after 1979.[8]

5 Concluding Remarks

In this paper, we proposed a model-based approach to the change-point problem in regularized regression analysis using Polson, Scott and Windle (2014)'s Bayesian bridge model and Chib (1998)'s non-ergodic hidden Markov model. We presented HMBB as a principled statistical method for regularization and estimation of parameter changes in high dimensional regression analysis.

Our simulation studies show that our modified Bayesian bridge model outperforms other regularization methods such as the lasso, elastic net, and ridge models in various high-dimensional data settings. In particular, when the underlying data generating model of high-dimensional data involves temporal heterogeneity in parameter values, HMBB successfully identifies break points and parameter changes while efficiently handling a large number of predictors. We also showed how HMBB can be used to uncover time-varying nonlinear effects by applying HMBB to the nonparametric regression model.

Our replication of Nunn and Qian (2014)'s study of the effect of US food aid on civil conflicts showed that HMBB can help researchers identify heterogeneous causal effects in the framework

[8] Although this paper does not aim to explain why the effects disappear after 1979, we conjecture that the break can be explained by a combination of factors such as the second oil shock, lingering effects of stagflation, and the rise of right-party governments in OECD countries at the end of 1970s.

of the instrumental variable regression analysis in large, long panel data ($N = 125, T = 5 \sim 36, p = 352$). HMBB helps researchers uncover unknown parameter heterogeneity in the first-stage equation while accounting for numerous nuisance parameters. The replication of Alvarez, Garrett and Lange (1991) how HMBB can detect a shift in regression parameters of an expanded model. The original data set has small data (16 countries and 15 time series observations) and a small number of covariates (6 covariates), which is a typical regression setup for social scientists. HMBB allows researchers to fit a parameter-expanded model (in our case, a fully interacted model with 21 predictors) while accounting for time-varying effects of these expanded predictors. Although our two examples are panel data in social sciences, HMBB can be used for the analysis of any type of high dimensional longitudinal (or time series) data.

In our ongoing work, we extend HMBB into discrete response panel models. Another fruitful application is to implement a changepoint detection algorithm using hidden Markov models in two different Bayesian regularization methods. The current framework is based on the one-group approach to parameter regularization and variable selection (Polson and Scott, 2010). Thus, it would be straightforward to extend HMBB to the Bayesian lasso (Park and Casella, 2008; Hans, 2009), the horseshoe prior model (Carvalho, Polson and Scott, 2010), and Dirichlet-Laplace prior (Bhattacharya et al., 2015). One drawback of the one-group approach is the lack of variable selection property. In this sense, it would be fruitful to extend a changepoint detection algorithm to the two-group approach such as the spike-and-slab prior models (Mitchell and Beauchamp, 1988; George and McCulloch, 1993). However, these two-group models suffer from large computational costs. In this context, recent innovations in the two-group approach such as the spike-and-slab lasso prior model (Rocková and George, 2018) and the neuronized prior model (Shin and Liu, N.d.) can lay down an efficient framework to compromise variable selection and changepoint detection within a reasonable computation cost.

References

Alvarez, R Michael, Geoffrey Garrett and Peter Lange. 1991. "Government partisanship, labor organization, and macroeconomic performance." *American Political Science Review* 85(2):539–556.

Andrews, Donald W.K. 1993. "Tests for Parameter Instability and Structural Change With Unknown Change Point." *Econometrica* 61(4):821–856.

Bai, Jushuan and Pierre Perron. 1998. "Estimating and Testing Linear Models with Multiple Structural Changes." *Econometrica* 66(1):47–78.

Barry, Daniel and J.A. Hartigan. 1993. "A Bayesian Analysis of Change Point Problems." *Journal of the American Statistical Association* 88(421):309–319.

Baum, Leonard E., Ted Petrie, George Soules and Norman Weiss. 1970. "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains." *The Annals of Mathematical Statistics* 41(1):164–171.

Beck, Nathaniel and Jonathan Katz. 1995. "What to Do (and Not to Do) with Time-Series Cross-Sectional Data." 89:634–647.

Beck, Nathaniel, Jonathan N. Katz and R Michael Alvarez. 1993. "Government Partisanship, Labor Organization, and Macroeconomic Per- formance: A Corrigendum." *American Political Science Review* 87:945–948.

Bhattacharya, Anirban, Debdeep Pati, Natesh S Pillai and David B Dunson. 2015. "Dirichlet–Laplace priors for optimal shrinkage." *Journal of the American Statistical Association* 110(512):1479–1490.

Bleakley, Kevin and Jean-Philippe Vert. 2011. "The group fused Lasso for multiple change-point detection." https://arxiv.org/pdf/1106.4199.pdf.

Cappe, Oliver, Eric Moulines and Tobias Ryden. 2005. *Inference in Hidden Markov Models*. Springer-Verlag.

Carvalho, Carlos M, Nicholas G Polson and James G Scott. 2010. "The horseshoe estimator for sparse signals." *Biometrika* 97:465–480.

Chan, Ngai Hang, Chun Yip Yau and Rong-Mao Zhang. 2014. "Group LASSO for Structural Break Time Series." *Journal of the American Statistical Association* 109(506):590–599.

Chernoff, Herman and Shelemyahu Zacks. 1964. "Estimating the Current Mean of a Normal Distribution Which is Subject to Changes in Time." *Annals of Mathematical Statistics* 35:999–1018.

Chernozhukov, Victor, Christian Hansen and Martin Spindler. 2015. "Post-selection and post-regularization inference in linear models with many controls and instruments." *American Economic Revieww: Paper & Proceedings* 105(5):486–90.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* 21(1):C1–C68.

Chib, Siddhartha. 1998. "Estimation and Comparison of Multiple Change-Point Models." *Journal of Econometrics* 86(2):221–241.

Chow, Gregory C. 1960. "Tests of Equality Between Sets of Coefficients in Two Linear Regressions." *Econometrica* 28(3):591–605.

Donoho, David. 2005. "High-Dimensional Centrally Symmetric Polytopes with Neighborliness Proportional to Dimension." *Discrete and Computational Geometry* 35(4):617–652.

Donoho, David and Victoria Stodden. 2006. Breakdown Point of Model Selection When the Number of Variables Exceeds the Number of Observations. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*. pp. 1916–1921.

Frick, Klaus, Axel Munk and Hannes Sieling. 2014. "Multiscale Change Point Inference." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(3):495–580.

Frühwirth-Schnatter, Sylvia. 2006. *Finite Mixture and Markov Switching Models*. Heidelberg: Springer Verlag.

George, Edward I. and Robert E. McCulloch. 1993. "Variable selection via Gibbs sampling." *Journal of the American Statistical Association* 88(423):881–889.

Greenshtein, Eitan and Ya'Acov Ritov. 2004. "Persistence in High-dimensional Linear Predictor Selection and the Virtue of Overparametrization." *Bernoulli* 10(6):971–988.

Hahn, P. Richard and Carlos M. Carvalho. 2015. "Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective." *Journal of the American Statistical Association* 110(509):435–448.

Hamilton, James D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica* 57(2):357–384.

Hans, Chris. 2009. "Bayesian lasso regression." *Biometrika* 96:835–845.

Lee, Sokbae, Myung Hwan Seo and Youngki Shin. 2016. "The Lasso for High Dimensional Regression with a Possible Change Point." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78(1):193–210.

Lee, Sokbae, Yuan Liao, Myung Hwan Seo and Youngki Shin. 2017. "Oracle Estimation of a Change Point in High Dimensional Quantile Regression." *Journal of the American Statistical Association* (just-accepted).

Mitchell, T. J. and J. J. Beauchamp. 1988. "Bayesian Variable Selection in Linear Regression." *Journal of the American Statistical Association* 83(404):1023–1032.
URL: *http://dx.doi.org/10.2307/2290129*

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. Boston, M.A.: The MIT Press.

Nunn, Nathan and Nancy Qian. 2014. "U.S. Food Aid and Civil Conflict." *American Economic Review* 104(6):1630–1666.

Park, Trevor and George Casella. 2008. "The bayesian lasso." *Journal of the American Statistical Association* 103(482):681–686.

Polson, Nicholas G and James G Scott. 2010. "Shrink globally, act locally: sparse Bayesian regularization and prediction." *Bayesian Statistics* 9:501–538.

Polson, Nicholas G and James G Scott. 2012. "Local shrinkage rules, Lévy processes, and regularized regression." *Journal of the Royal Statistical Society (Series B)* 74:287–311.

Polson, Nicholas G., James G. Scott and Jesse Windle. 2014. "The Bayesian Bridge." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(4):713 – 733.

Qian, Junhui and Liangjun Su. 2016. "Shrinkage Estimation of Regression Models with Multiple Structural Changes." *Econometric Theory* 32(6):1376–1433.

Quandt, Richard E. 1958. "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes." *Journal of the American Statistical Association* 53(284):873–880.

Robert, Christian P., Tobias Ryden and D. M. Titterington. 2000. "Bayesian Inference in Hidden Markov Models through the Reversible Jump Markov Chain Monte Carlo Method." *Journal of the Royal Statistical Society, Ser. B* 62(1):57–75.

Rocková, Veronika and Edward I. George. 2018. "The Spike-and-Slab LASSO." *Journal of the American Statistical Association* 113(521):431–444.
URL: *https://doi.org/10.1080/01621459.2016.1260469*

Scott, Steven L., Gareth M. James and Catherine A. Sugar. 2005. "Hidden Markov Models for Longitudinal Comparisons." *Journal of the American Statistical Association* 100(470):359–369.

Shin, Minsuk and Jun S. Liu. N.d. "Neuronized Priors for Bayesian Sparse Linear Regression." arXiv:1810.00141.

Tang, Lu and Peter X.K. Song. 2016. "Fused Lasso Approach in Regression Coefficients Clustering – Learning Parameter Heterogeneity in Data Integration." *Journal of Machine Learning Research* 17:1–23.

Tanner, Martin A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Springer.

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal and David M. Blei. 2006. "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association* 101(476):1566–1581.

Tibshirani, Robert, Michael Saunders, Saharon Rosset, Ji Zhu and Keith Knight. 2004. "Sparsity and smoothness via the fused lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(1):91–108.

Western, Bruce. 1998. "Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach." *American Journal of Political Science* 42(4):1233–1259.

Zou, Hui. 2006. "The Adaptive Lasso and Its Oracle Properti." *Journal of the American Statistical Association* 101(476):1418–1429.