Bayesian Estimation of High-dimensional and Time-varying Parameters in Longitudinal Data Analysis

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

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

Abstract

In social science, regularized regression methods have been successfully applied to analyze the data with many covariates and interaction terms, possibly larger than the sample size. Yet, structural changes in longitudinal data pose a serious challenge to the application of existing regularization methods because time-varying effects can be over-regularized. Motivated by two existing studies from social science where the exact timing and scope of structural changes is unknown, we develop a hidden Markov Bayesian bridge model that investigates parametric change-points in high-dimensional longitudinal data. The proposed method successfully uncovers change-points in the relationship between government partisanship and economic growth and the relationship between food aid and the civil war on-set.

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

[*] Professor, Department of Political Science and International Relations, Seoul National University

[†] Ph.D. student, Department of Government, Harvard University.

1 Introduction

Longitudinal data analysis is one of most common types of data analysis in social sciences. The main quantity of interest in longitudinal analysis is effects of a few predictors (e.g. the amount of foreign aid receipt) on a response variable (e.g. the frequency of civil wars) in the presence of many confounding variables (e.g. other international, political, social, economic, and cultural factors). The long list of control variables in longitudinal analysis plays an important role as statistical control for bias reduction of causal parameters or candidate sets for multiplicative (or interaction) relationship checks. For either purpose, regularization methods have a strong appeal to researchers of longitudinal data analysis with many predictors.

However, it is not straightforward to apply existing regularization methods to longitudinal data analysis because existing regularization methods firmly rely on the assumption that *the level of sparsity (or shrinkage) does not change over time.* Applying regularization methods that ignore the possibility of parametric shifts can lead to erroneous inferential results by forcing time-varying non-zero coefficients to zero. We call it the *change-point regularization problem.*

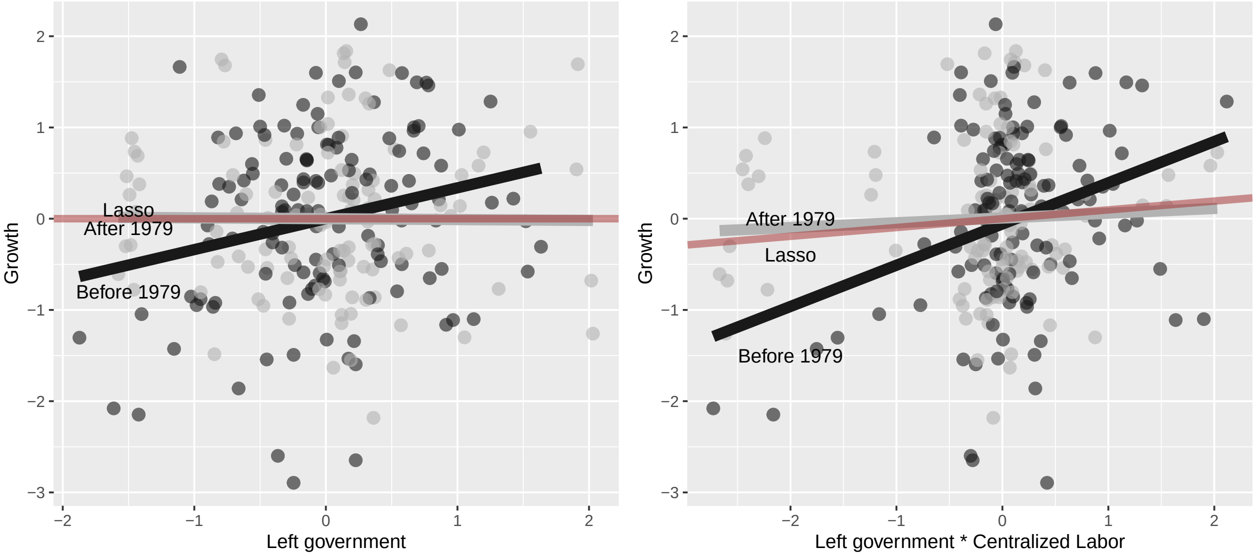

Figure 1: The Regularization Bias in the Politics of Growth Model: The model includes all pairwise interactions of Alvarez, Garrett and Lange (1991): $n = 240, p = 21$. We use 1979 as a known break point and fit the same model to each sub-sample. Thick lines indicate regression slopes of the left party government size (left) and its interaction with labor union centralization (right) from the pre-1979 data (Before 1979), the post-1979 data (After 1979), and the lasso estimate from the entire data (Lasso).

The change-point regularization problem arises when we consider a possibility of parameter break in longitudinal analysis with many predictors. The politics of growth debate in political science is a good example. Alvarez, Garrett and Lange (1991) examined a longitudinal data set of 16 OECD countries over 1970-1984 to see how government partisanship affects economic growth. They found that centralized labor organizations were “conducive to better economic performance when the Left was politically powerful" while weaker union movements "had desirable consequences for growth and inflation when governments were dominated by rightist parties" (Alvarez, Garrett and Lange, 1991, 551).

For the sake of illustration, we assume the first year of the Thatcher government as a known break point in the politics of growth model although our proposed method does not need this assumption. The left panel of Figure 1 shows changing effects of the left party government size on economic growth and the right panel shows changing effects of the left party government size with labor union centralization on economic growth. The positive effects of the left party government size and its interaction with labor union centralization almost disappear after 1979. The conditional effect of government partisanship reported by Alvarez, Garrett and Lange (1991) is limited to the pre-1979 samples. If we estimate parameters of the politics of growth model using lasso (Tibshirani, 1996), the strong pre-1979 signals are compromised with the weak post-1979 signals due to the pooling and then the pooled weak signals are further shrunk toward zero by the penalty of the lasso. As a result, the lasso fails to capture the conditional effect of the left party government size as shown in Figure 1.

Tibshirani et al. (2004) noted this problem and warned that parameters "that can be ordered in some meaningful way" (Tibshirani et al., 2004, 91) need special attention in regularization. To address the ordered parameter problem, Tibshirani et al. (2004) proposed a method of parameter fusion, which has been successfully applied to problems in classification and pattern recognition where adjacent parameters imply structured associations. We apply the method of parameter fusion to the simulated data set and reports the results in Figure 2.[1] The left panel shows the ground truth where the x-axis indicates time and the y-axis indicates the true parameter value. The middle panel shows the fused lasso estimates of unknown break points.

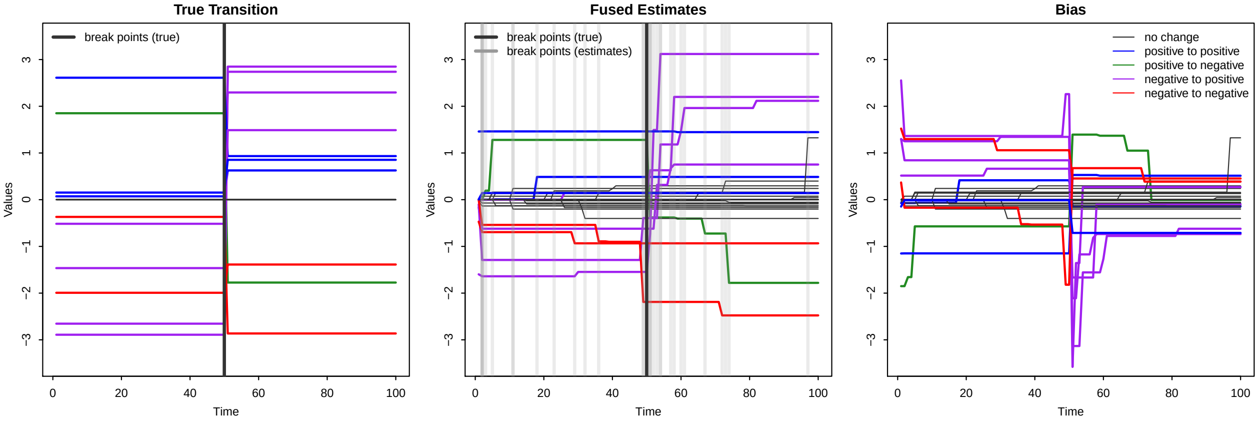

Figure 2: Change-point Analysis using the Fused Lasso: The dark vertical line in the middle is the timing of the true break and grey vertical lines are estimated break points by sharp jumps in parameters. Different colors indicate different dynamic patterns (e.g. no change, positive to negative, negative to positive, positive to positive, and negative to negative). Implementation details are discussed in the supplementary information.

[1] See the appendix for the implementation details.

Several problems can be pointed out in the use of parameter fusion to identify change-points. First, there are too many "jumps" and they do not show a coherent pattern around the true break point. Second, these many jumps generate large biases as shown in the right-most panel. Third, researchers need to transform the original T (the number of time units) by p (the number of predictors) design matrix into a T by $p \times T$ matrix by multiplying a low triangular time dummy matrix (100×5000 matrix in this example). The augmented $p \times (T - 1)$ dimensionality puts an additional burden to the change-point analysis that needs to make inference from sub-sample data.

Recently, 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., 2018). 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, the case in which the number of covariates with regime-changing coefficients are small, or the case in which the number of breaks is known and limited to one. However, except in some rare cases, researchers do not have clear knowledge about the number of breaks, the timing of breaks, or the scope of time-varying covariate effects. Thus, there is a strong demand for a general statistical framework for high-dimension regression analysis of longitudinal data with multiple change-points.

In this paper, our goal is to propose a fully Bayesian approach to the change-point regularization problem so that all the uncertainties in the change-point regularization problem (e.g. the number, location, and scope of break) can be properly incorporated in the inferential process (Gelman et al., 2004; Kyung et al., 2010; Bhadra et al., 2017). Our strategy is to combine a highly efficient Bayesian regularization model with a hidden Markov model (HMM). Polson, Scott and Windle (2014)'s Bayesian bridge model has the property of avoiding the overshrinkage of large coefficients, corresponding to the oracle property in classical regularization estimators while allowing efficient sampling of global and local shrinkage parameters. We use HMM to identify multiple change-points of regression parameters in various settings following important work of Chib (1998), Scott, James and Sugar (2005), Frühwirth-Schnatter (2006), Chib and Kang (2013), and Ko, Chong and Ghosh (2015). We call the proposed method the hidden Markov Bayesian bridge model (HMBB). Our proposed method can be easily applied to most regression models with many predictors and inferential outcomes are easy to interpret. The proposed method is available via the open-source software `BridgeChange` in R environment (`https://github.com/soichiroy/BridgeChange`).

2 The Change-point Regularization Problem in Longitudinal Data Analysis

Figure 3 illustrates the change-point regularization problem in longitudinal data analysis. Researchers are interested in understanding how p numbers of predictors are associated with the response variable. Here, p is quite large compared to the size of sample data. Researchers suspect that the relationships between p numbers of predictors and the response variable are not time-homogeneous: predictors have time-varying associations with the response variable. Even though the original data has larger n than p, the sub-sample data (e.g. data belonging to Regime 1) could have more predictors than observations, posing a serious computational challenge. Then, researchers need to regularize parameters within each sub-sample while detecting parameter breaks from the entire data. We present two social science studies that show the ubiquity of the change-point regularization problem in longitudinal data analysis.

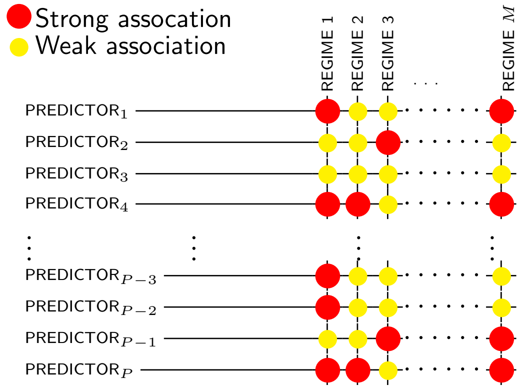

Figure 3: Change-point Regularization Problem in Longitudinal Analysis: The number of parameters (p) is fixed while the sample size is shorter in each sub-sample (or regime) as the number of breaks ($M - 1$) increases. Predictors have time-heterogeneous relationships (different dot colors) with the response variable over time.

2.1 The Politics of Growth

After examining a longitudinal data set of 16 OECD countries over 1970-1984, Alvarez, Garrett and Lange (1991) concluded that centralized labor organizations could be beneficial to economic growth when they were accompanied by the left party government while decentralized labor organizations were better for the economy when they were sided with right party governments. Since then, the conditional effect of centralized labor and government partisanship on growth has received continuing attention in comparative political economy literature (Alvarez, Garrett and Lange, 1991; Beck, Katz and Alvarez, 1993; Beck and Katz, 1995; Western, 1998; Scruggs, 2001; Rueda, 2008).

Despite the intense attention to Alvarez, Garrett and Lange (1991)'s analysis, there are still many issues that have not been fully addressed by previous studies. First, we do not know whether and when the conditional effect of centralized labor and government partisanship on growth changed during the sample period. There were many important game-changing events such as oil shocks (1973, 1978, 1980), stagflation, and conservative reforms in the U.K, U.S. and Australia during the sample period. We do not know whether and how these game-changing events affected the conditional effect under investigation. Also, we do not know how these game-changing events affect relationships between other control covariates and the response variable. Mis-specifying any of these time-varying relationships could endanger the entire statistical inference.

Second, we do now know how many breaks will be needed to properly address effects of the above-mentioned game-changing events. Also, we do not know the duration of these effects. If some of the effects are too short-lived (say 2-4 years), the corresponding design matrix is likely to have rank-deficiency or the large n small p problem, endangering the statistical analysis of the entire sample.

2.2 Food Aid as the Cause of Civil War

Another example of the change-point regularization problem in longitudinal analysis is Nunn and Qian (2014)'s study of food aid effect on the onset of intrastate conflicts. Nunn and Qian (2014) examined the causal effect of food aid (d) on intrastate conflicts (y) using an instrument variable (z). The instrument was found from the interaction of last year's wheat production and the frequency of a country's US food aid receipt. The total number of observations is 4,089 covering 125 non-OECD countries during the 36 years, 1971-2006. The number of control variables is very large ($p = 352$) but regularization methods were not considered.

Although there were several important international events, most notably the end of the Cold War, that may directly or indirectly affect effects of covariates on the endogenous variable and the response variable, Nunn and Qian (2014) did not fully investigate the *scope* of parameter changes. Instead, Nunn and Qian (2014) used a Cold War dummy variable and rejected the possibility of time-varying causal effects.[2]

As in the case of Alvarez, Garrett and Lange (1991), there are many issues in Nunn and Qian (2014)'s analysis. First, how accurate is β given the tumultuous international history during 1971-2006? If we write Nunn and Qian (2014)'s model using time-varying subscripts (s_t indicating state dependence),

$$y_{it} = \overbrace{\beta_{s_t} d_{it}}^{\text{time-varying causal effect}} + \overbrace{\sum_{k=1}^{p} x_{k,it}\delta_{k,s_t}}^{\text{time-varying nuisances}} + \overbrace{\nu_i + \xi_t}^{\text{fixed-effects}} + \varepsilon_{it} \tag{2.1}$$

$$d_{it} = \overbrace{\alpha_{s_t} z_{it}}^{\text{time-varying instrument effect}} + \overbrace{\sum_{k=1}^{p} x_{k,it}\gamma_{k,s_t}}^{\text{time-varying nuisances}} + \overbrace{\nu_i + \xi_t}^{\text{fixed-effects}} + \epsilon_{it}. \tag{2.2}$$

The causal effect of food aid on intrastate conflicts (β) was estimated by the two-stage least squares (TSLS) method. Here an important assumption is "constant parameters [α_{s_t} in Equation 2.2 for the relation between z_i and d_i" (Angrist, Imbens and Rubin, 1996, 450). Then, β can be considered a weighted average of per-unit treatment effect (Angrist and Imbens, 1995, 435). However, if α_{s_t} varies significantly over time, the change-point analysis of Equation 2.2 can be useful i) to check the monotonicity assumption (Angrist, Imbens and Rubin, 1996, Assumption 5) and ii) to show "causally relevant" time periods ($i.e.$ time periods under which the treatment status of subjects is homogeneously affected by the instrument).

Second, the large number of nuisance parameters may have time-heterogenous effects. Omitting these time-heterogenous relationships can affect the consistent estimation of β.

[2] Nunn and Qian (2014) reported that the interaction term with the Cold War dummy variable was "negative, moderate in magnitude, but statistically insignificant" (Nunn and Qian, 2014, 1662).

How can we properly estimate the large number of time-heterogenous nuisance parameters when they may have multiple breaks?

The politics of growth model and the food aid-civil war model clearly demonstrate that the change-point regularization problem can arise in most longitudinal data analysis with a large number of predictors. However, researchers rarely check the possibility of parameter breaks beyond a period dummy variable because sub-sample data may not be long enough to fit their full model with many predictors. Therefore, the change-point problem almost always comes with the regularization problem. This is why we need a statistical method to address the change-point regularization problem in longitudinal analysis.

3 The Proposed Method

In this section, we introduce HMBB for change-point estimation and parameter regularization in longitudinal data analysis. In Section 3.1, we discuss the Bridge estimator and the Bayesian bridge model. Then, in Section 3.2, we explain the model structure and sampling algorithm of HMBB.

3.1 The Setup

The Bridge estimator can be motivated by the following penalized likelihood formulation,

$$\widehat{\boldsymbol{\beta}}_{\text{bridge}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2} \sum_{t=1}^{n} (y_t - \mathbf{x}_t^\top \boldsymbol{\beta})^2 + \nu \sum_{j=1}^{p} |\beta_j|^\alpha \right\} \tag{3.1}$$

where $0 < \alpha \leq 2$ (Frank and Friedman, 1993; Fu, 1998). The popular lasso estimator and ridge regression are obtained as special cases of this estimator when $\alpha = 1$ and $\alpha = 2$, respectively. When $0 < \alpha \leq 1$, the bridge regression has the variable selection feature (Frank and Friedman, 1993; Huang, Horowitz and Ma, 2008). Because of this generality, the Equation 3.1 has been receiving an increasing attention in the statistical literature (Fan and Li, 2001; Liu et al., 2007; Huang, Horowitz and Ma, 2008; Huang et al., 2009; Armagan, 2009).

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

$$p(\boldsymbol{\beta}|\tau, \alpha) \propto \prod_{j=1}^{p} \exp(-|\beta_j/\tau|^\alpha) \quad \tau = \nu^{-1/\alpha}. \tag{3.2}$$

Using Lévy processes and scale mixtures of normal representation, 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:

$$p(\boldsymbol{\beta}, \Lambda|\tau, \alpha) \;\; \propto \;\; \prod_{j=1}^{p} \exp\left(-\frac{\beta_j^2}{2\tau^2}\lambda_j\right) p(\lambda_j) \tag{3.3}$$

where $p(\lambda_j)$ is the density of $2S_{\alpha/2}$ and S_α is the Lévy alpha-stable distribution.

The posterior distribution of Polson, Scott and Windle (2014)'s Bayesian bridge model is

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2, \Lambda, \alpha, \nu|\boldsymbol{y}, \mathbf{X}) \;\; &\propto \;\; p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \Lambda|\tau, \alpha)p(\sigma^2)p(\alpha)p(\nu) \\ &\propto \;\; \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\right] \prod_{j=1}^{p} \exp\left(-\frac{\beta_j^2}{2\tau^2}\lambda_j\right) p(\lambda_j) \\ &\quad\;\; \times \left(\frac{1}{\sigma^2}\right)^{\frac{a_0}{2}+1} \exp\left(-\frac{b_0}{2\sigma^2}\right) \nu^{c_0-1} \exp(-d_0\nu). \end{aligned} \tag{3.4}$$

Polson, Scott and Windle (2014)'s Bayesian bridge model provides two important advantages for high dimensional longitudinal analysis. First, the Bayesian bridge model avoids over-shrinking large coefficients. Second, Polson, Scott and Windle (2014)'s sampling scheme provides an efficient blocking between global (τ, α) and local shrinkage parameters (Λ).

3.2 The Hidden Markov Bayesian Bridge Model

We modify Polson, Scott and Windle (2014)'s Bayesian bridge model for high dimensional longitudinal analysis. The first challenge is the larger number of covariates than the number of observations. In social science data analysis, $n \leq p$ is not uncommon. Researchers include many control variables for statistical control while the number of observations is small. Also, researchers are interested in uncovering interaction effects among many covariates. The second challenge is time-varying effects. Longitudinal (or time series) data poses a unique opportunity to examine temporal heterogeneity of existing theories, which can be captured by change-points in model parameters.

3.2.1 Sampling of Regression Coefficients using SVD

When $n \leq p$, inverting the covariance matrix $(p \times p)$ is very expensive (roughly $O(p^3)$), which applies to most Bayesian models including the Bayesian bridge model. In order to resolve this, 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 $\mathbf{D} = \text{diag}(d_1, \dots, d_{n_m})$. 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 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, this operation allows $\overline{\mathbf{D}}$ to have dimension $n_m \times p$, which is crucial since $\boldsymbol{\lambda}_m$ has dimension of $p \times p$. Then,

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

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

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

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 of $\boldsymbol{\beta}_m$ can be done in the following order:

1. Let $\mathbf{A} = \overline{\mathbf{V}}(\overline{\mathbf{D}}^\top \overline{\mathbf{D}} + \boldsymbol{\lambda}_m)^{-1/2}$.
2. Draw z_k from standard normal for $k = 1, \ldots, p$.
3. Update $\boldsymbol{\beta}_m$ by $\boldsymbol{\beta}_m \leftarrow \boldsymbol{\mu}_m + \boldsymbol{A}\boldsymbol{z}$
4. Repeat the above steps for all $m = \{1, \ldots, M\}$.

3.2.2 Hidden Markov Embedding

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:

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

where $\mathbf{Y}_{t-1}$ and $\mathbf{X}_{t-1}$ indicate all the observed data up to $t-1$ and subscripts s_t in model parameters (e.g. $\boldsymbol{\beta}, \sigma^2, \Lambda, \alpha, \nu$) indicate hidden states they belong to. The state transition is defined as a forward-moving first-order discrete Markov process with an initial probability of $\boldsymbol{\pi}$:

$$\begin{aligned}
s_t|\mathbf{P}, \boldsymbol{\pi} &\sim \text{Markov}(\mathbf{P}, \boldsymbol{\pi}), \boldsymbol{\pi} = (1, 0, \dots, 0) \\
\underbrace{\mathbf{P}}_{M \times M} &= (\mathbf{p}_1, \dots, \mathbf{p}_M) \\
\mathbf{p}_i &\sim \text{Dirichlet}(\alpha_{i,1}, \dots, \alpha_{i,M}) \text{ for all } i < M.
\end{aligned}$$

To illustrate the model introduced above, we discuss a case with one change-point (and generalize it into a multiple change-point case later). 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

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

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.

3.2.3 Further Modeling Issues

In the bridge model, α determines the shape of shrinkage constraint. Polson, Scott and Windle (2014) suggest a random-walk Metropolis Hastings sampler for the sampling of α from the support of $0 < \alpha \le 1$. However, unlike classical statistics, the support of $0 < \alpha \le 1$ does not guarantee the variable selection feature in Bayesian framework because the posterior mean is non-sparse with probability zero (Hahn and Carvalho, 2015). Moreover, 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 from the support of $0 < \alpha \le 2$.

The Bayesian bridge model is based on the shrinkage approach to the high dimensional problem. However, researchers of longitudinal analysis often face the variable selection problem in high dimensional data. One example is the multiplicative (or interaction) relationship check between interested causal variables and many candidate predictors. For this purpose, we use Hahn and Carvalho (2015)'s decoupled shrinkage and selection (DSS) method in the context of HMBB. The goal of DSS 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 estimates at regime m as

$$\mathcal{L}(\boldsymbol{\gamma}_m) = \arg\min_{\boldsymbol{\gamma}_m} \overbrace{\|\mathbf{X}_m\boldsymbol{\beta}_m^* - \mathbf{X}_m\boldsymbol{\gamma}_m\|_2^2}^{\text{squared prediction loss}} + \overbrace{\lambda\|\gamma_m\|_0}^{\text{parsimony penalty}} \tag{3.5}$$

where $\mathbf{X}_m\boldsymbol{\beta}_m^*$ is the fitted value of HMBB at regime m. We can use the adaptive lasso method (Zou, 2006) or other equivalent methods to find $\boldsymbol{\gamma}_m$.[3]

4 Simulation Studies

In this section, we conduct a series of Monte Carlo simulations to test the performance of HMBB in various data settings. Following Donoho (2005) and Donoho and Stodden (2006), simulated data varies 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 $[0.1, 1]$.

We create $50 \times 50 = 2,500$ 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 $50^2 \times 20 = 50,000$. The entire test results are reported both in a numerical summary table and in the format of "phase diagrams" used by Donoho (2005) and Donoho and Stodden (2006). 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.[4]

Since there exists no comparable regularization method that implements change-point analysis, we develop two hybrid lasso estimates as benchmark (Lasso (Estimate) and Lasso (Oracle)). Lasso (Estimate) first identifies a break point from the residual break test, which only uses two parameters (the mean and the variance), and regularizes parameters for each regime in two steps. Lasso (Oracle) regularizes regime-specific parameters without break estimation. In contrast, HMBB estimates a break point and regularizes regime-specific parameters at the same time.

[3] We explain how to check model validity of HMBB using popular Bayesian model comparison methods in the appendix.

[4] To space space, we explain the simulation setup and test criteria in the appendix.

Table 1: Simulation Performance Criteria

Metric	Formula	Property
Prediction Loss	$\mathcal{L}_{\text{pred}}(\widehat{\boldsymbol{\beta}};\boldsymbol{\beta}^{\star}) = \frac{1}{n}\lVert\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^{\star}\rVert^2$	in-sample model fit
Normalized Estimation Loss	$\mathcal{L}_2(\widehat{\boldsymbol{\beta}};\boldsymbol{\beta}^{\star}) = \frac{\lVert\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{\star}\rVert^2}{\lVert\boldsymbol{\beta}^{\star}\rVert^2}$	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

Table 2: Summary of Single Change-point Case: The reported numbers are averaged from 50, 000 simulated data sets. Data has one break. MCMC simulation for HMBB is 100 and burn-in is 100.

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

Table 2 shows the numerical summary of the simulations. When we compare three methods by the average size of prediction loss, HMBB performs worse than Lasso (Estimate) and Lasso (Oracle). In the normalized estimation loss and cross-validation loss, HMBB performs better than Lasso (Estimate) and as good as Lasso (Oracle).

Panel (A) in Figure 4 shows that the poor performance of HMBB occurs mostly within the small area around the left corner where n is highly smaller than p. This is due to the poor performance of HMBB in break detection. Except this area, HMBB shows comparable performances with Lasso (Estimate) and Lasso (Oracle). Panel (B) of Figure 4 clearly demonstrates that HMBB shows a more stable performance than Lasso (Estimate) across various levels of underdeterminedness and sparsity. Lasso (Oracle) does very well in the lower-right corner where k/n is small and n/p is large.

The most surprising result was found in the cross-validation loss test, which is shown in Panel (C). HMBB slightly outperforms both Lasso (Estimate) and Lasso (Oracle). Also, the performance of HMBB is more stable than the others. The vertical separation of the cross-validation loss is stronger in lasso-based methods, which is what Donoho and Stodden (2006) refers to as the theoretical threshold of ℓ_1-based methods. While Lasso (Oracle) predicts out-of-sample data relatively well when k/n is small (i.e. when the sparsity level is high), Lasso (Oracle) does poorly as k/n becomes larger (*i.e.* data becomes less sparse). HMBB's out-of-sample predictive accuracy is slightly better than Lasso (Oracle) when k/n approaches to 1.

It should be stressed that the simulation test assumes that weak signals are exactly zero, which favors lasso-based sparsity methods in design. Also, in the case of the single break test, HMBB conducts dual tasks of break detection and parameter regularization in one step while lasso-based competing models use two-step approaches. In spite of these difficulties, HMBB performs quite well regardless of whether the underlying data has a break or not.

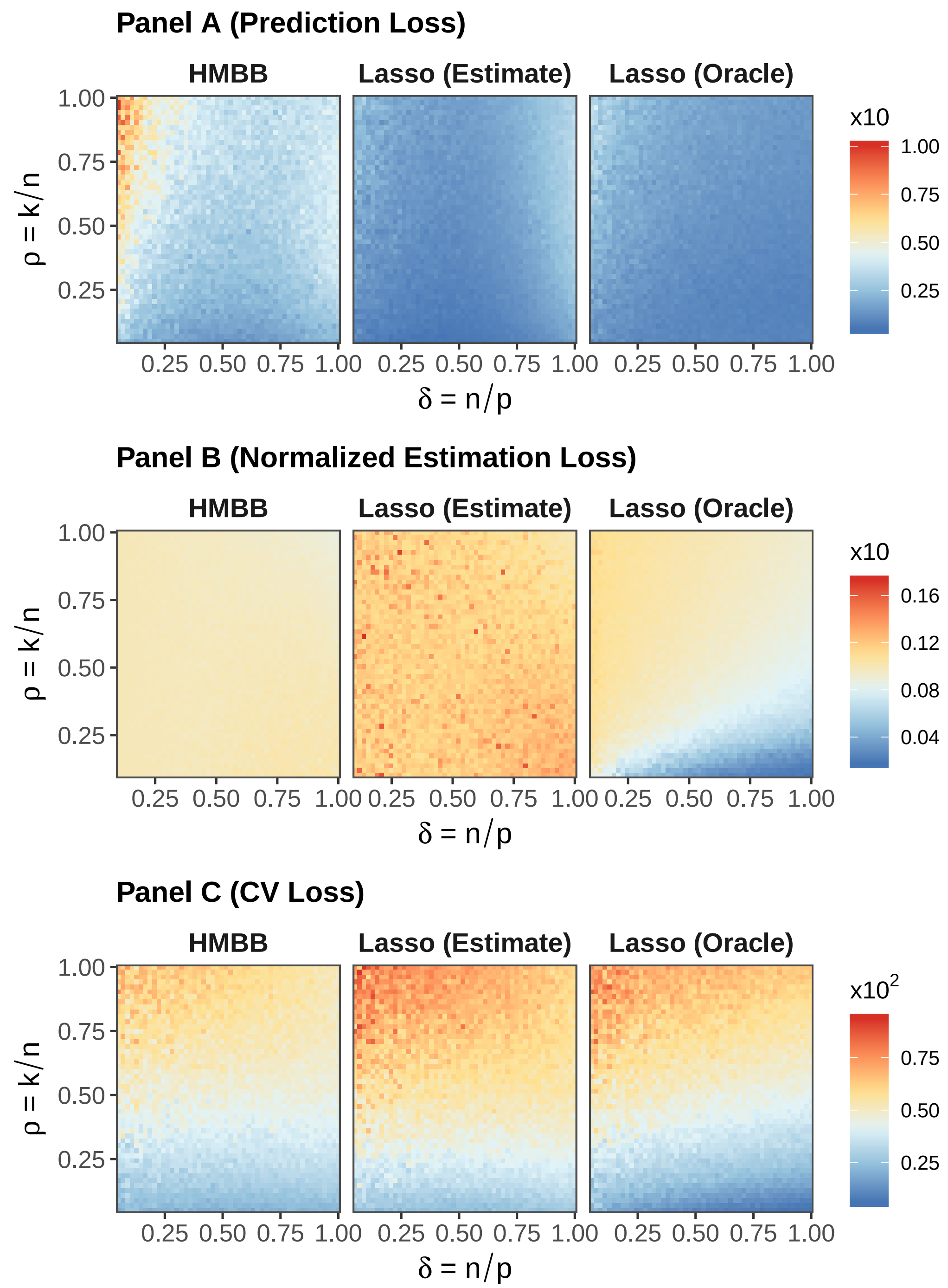

Figure 4: Results of Simulation Studies using Univariate Time Series Data with One Change-point. **Panel (A)**: Prediction Loss, $\mathcal{L}_{\text{pred}}(\widehat{\boldsymbol{\beta}};\boldsymbol{\beta}^{\text{true}})$. **Panel (B)**: Normalized Estimation Loss, $\mathcal{L}_2(\widehat{\boldsymbol{\beta}};\boldsymbol{\beta}^{\text{true}})$. **Panel (C)**: Cross-validation Loss, $\mathcal{L}_{\text{CV}}(\widehat{\boldsymbol{y}};\boldsymbol{y}^{\star})$. We fix $p = 200$ and vary α and ρ between 0.1 and 1. Thus, each cell in the graph represents a data set with (n, p, k). We simulate 20 data sets from each (n, p, k) and take the median error.

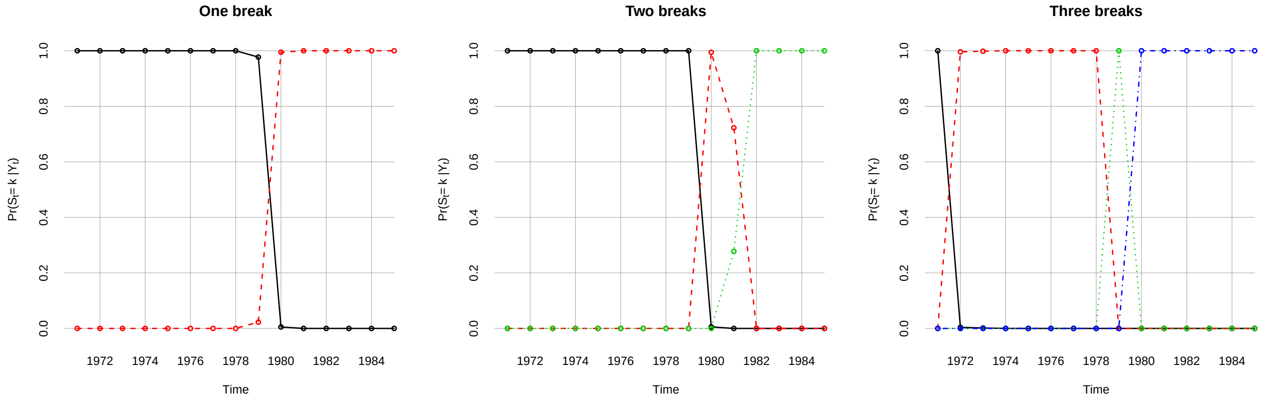

Figure 5: *Transition of Latent States by the Number of Breaks: The year of 1979 is consistently detected as a break point. Adding more than one break produces singleton states. The fully interacted Alvarez, Garrett and Lange (1991) model is used for HMBB analysis.*

More impressively, HMBB consistently shows the best performance in the cross-validation loss, which captures the out-of-sample predictive accuracy. Also, when the break point is well identified, HMBB shows robust performance in all performance criteria.[5]

5 Applications

In this section, we illustrate the utility of HMBB in the two examples we discussed in Section 2.

5.1 Uncovering Time-Varying Interaction Effects

In Alvarez, Garrett and Lange (1991), the dependent variable is the annual growth rate and independent variables are six covariates of economic growth: lagged growth rate (`lagg1`), weighted OECD demand (`opengdp`), weighted OECD export (`openex`), weighted OECD import (`openimp`), cabinet composition of left-leaning parties (`leftc`), and the degree of labor organization encompassment (`central`). We examine time-varying effects of the full interaction model with 21 $(= 6 + \binom{6}{2})$ predictors.[6]

The left panel of Figure 6 shows DSS-HMBB estimates of 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, positive 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

[5] Additional simulation results using correlated data are available in the supplementary information.

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

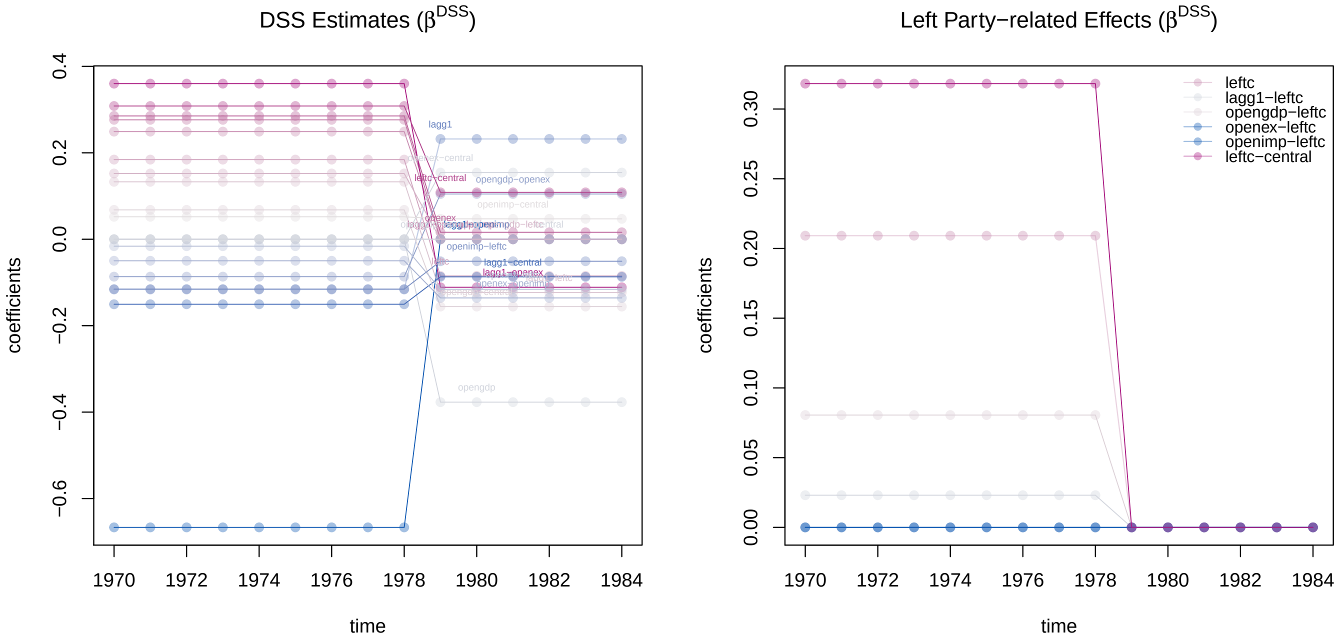

Figure 6: *A Latent Regime Change and Regime-changing Covariate Effects in the Fully Interacted Alvarez, Garrett and Lange (1991) Model. Reported coefficients are from the DSS-HMBB method.*

(1991)'s original claim on the conditional effect of government partisanship on economic growth does not hold after 1979.[7]

5.2 Uncovering Time-Varying Causal Effects

In Nunn and Qian (2014)'s study, our concern is time-varying effects of the instrument on the endogenous variable. As discussed above, the interpretation of β as the *causal* estimand requires constant α in Equation 2.2. Using HMBB, we identify hidden regimes of Equation 2.2 within which the response to the instrument (α) is homogenous. Then, we can estimate corresponding βs that capture time-varying causal effects of food aid on intrastate conflicts within each regime.

The left panel of Figure 7 shows WAIC scores of panel HMBBs using the model specification of Equation 2.2 with a varying number of breaks from zero to five. The predictive accuracy of the model improves by allowing breaks in model parameters of Equation 2.2 until three breaks and then it deteriorates afterward. The expected values of the three estimated break points are 1986, 1991, and 1998 (in the right panel).

The left panel of Figure 8 shows the time-varying relationships between the instrument and the endogenous variable (α_{s_t} in Equation 2.2). Blue dots indicate the pooled estimate using the double machine learning method and red dots indicate HMBB estimates. The association between the instrument and the endogenous variable is positive across time,

[7] 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.

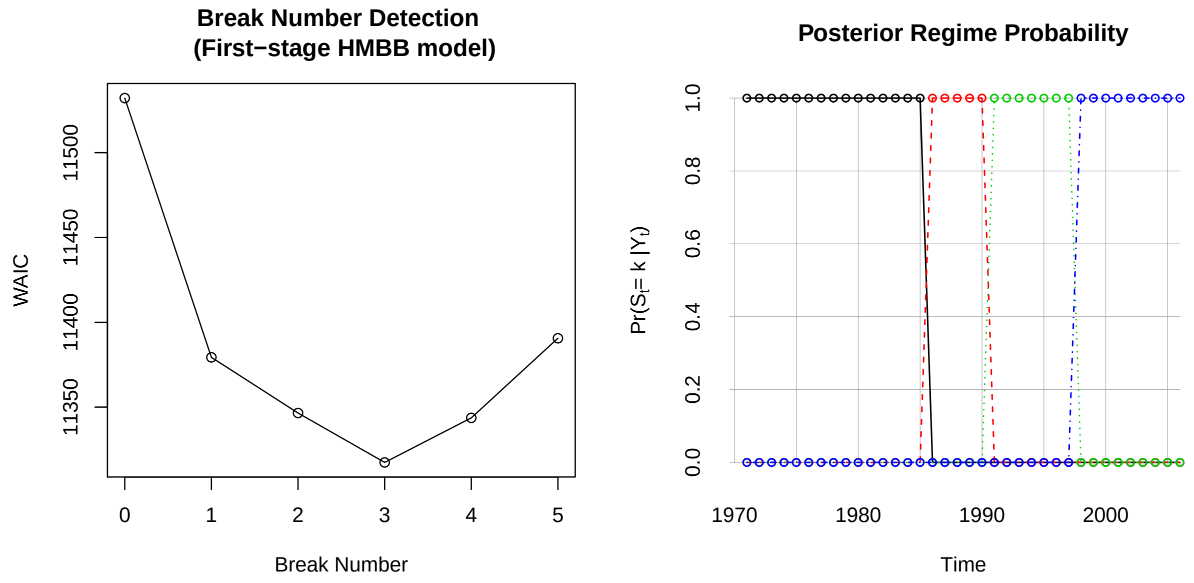

Figure 7: Detecting Parametric Breaks in Equation 2.2: The left panel shows WAIC scores of panel HMBBs using the model specification of Equation 2.2 with a varying number of breaks from zero to five. The right panel shows posterior probabilities of hidden regimes from 1 to 4.

indicating that the monotonicity assumption (Angrist, Imbens and Rubin, 1996, Assumption 5) is satisfied: that is, an increase in wheat production at year $t - 1$ does not decrease U.S. food aid at year t. However, the magnitude of the association varies a lot over time, having the strongest association between 1987 and 1991, followed by the period of 1971-1985. After 1992, the association between the instrument and the endogenous variable diminishes significantly.

Then, how do the time-varying associations between the instrument and the endogenous variable affect the causal effect of food aid on intrastate conflicts (β_{s_t})? To answer this question, we first partitioned data based on the four estimated regimes and then employ the "double machine learning" method to estimate the regime-specific estimate of β (DML-HMBB).[8]

The blue dot in the right panel of Figure 8 indicates the pooled DML estimate of β (0.004107 with standard error of 0.001279)[9] and red dots indicate DML-HMBB estimates of β_{s_t}. To our surprise, the DML-HMBB estimate of β is close to zero when the partial correlation between the instrument and the endogenous variable is largest (1986-1990). That is, *the causal effect food aid on intrastate conflicts is close to zero in the subset of data within which the instrument is strongest.* Nunn and Qian (2014)'s finding of the positive

[8] Nunn and Qian (2014)'s unbalanced panel data has 352 predictors from 125 countries. The number of time units varies from 5 to 36 years. Thus, if the duration of any hidden state is shorter than 5 or 6, the "large p small n" problem arises and we need to regularize parameters. However, the direct application of regularization methods to TSLS generates a bias by suppressing the causal parameter with all the nuisance parameters. To avoid this bias, Chernozhukov et al. (2017) proposed the "double machine learning" (DML) method that utilizes the Neyman orthogonalization of residuals.

[9] The non-DML estimate reported by Nunn and Qian (2014) is 0.00254 with standard error of 0.00088.

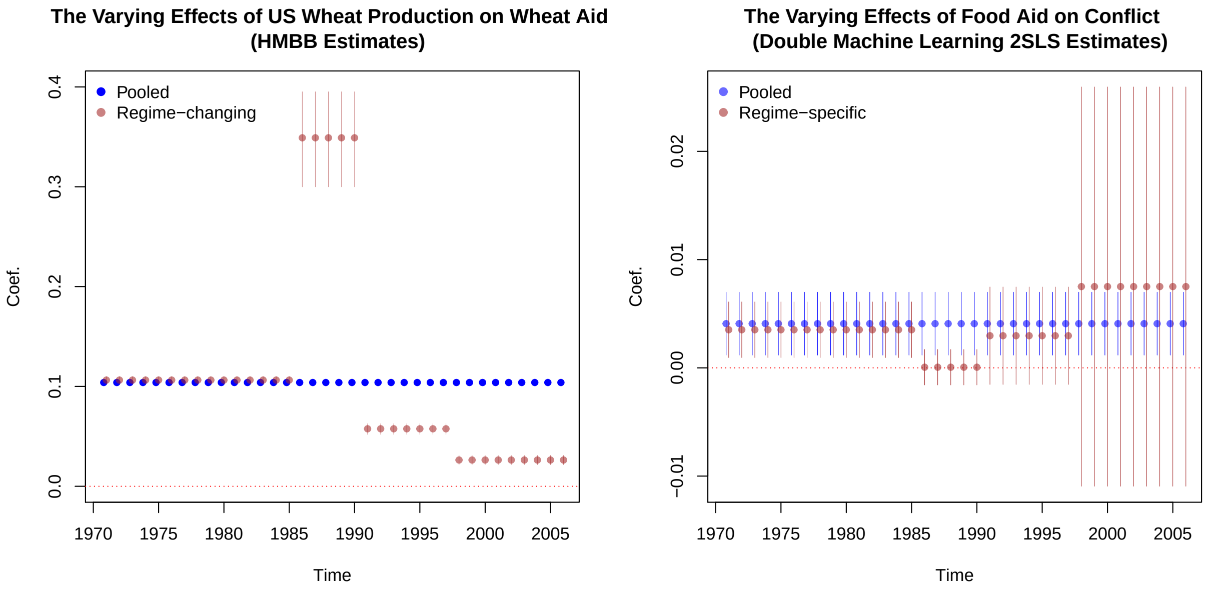

Figure 8: Regime-changing α and β in the first- and second-stage equations in Equation 2.1 and Equation 2.2, respectively: Vertical bars indicate 95% credible intervals (left) and 95% confidence intervals (right).

and statistically significant effect of food aid on intrastate conflicts holds only up to 1985.

6 Concluding Remarks

Recent innovations in regularization methods bypass applied researchers of longitudinal data analysis because most regularization methods do not specifically address the question of time-varying covariate effects in high-dimensional regression models. In this paper, we showed that the change-point regularization problem is ubiquitous in longitudinal data analysis with a large number of predictors. It is not difficult to adapt existing regularization methods for high-dimensional regression analysis using longitudinal data and this paper belongs to one of the adaptations. However, we highlight that our proposed method has several advantages: (1) an efficient sampling algorithm for parameter regularization and change-point detection, (2) robust performance against various types of high-dimensional data, and (3) an easy applicability to conventional regression models.

As we have shown in the politics of growth example and the food aid-civil war example, social science researchers rarely venture for the possibility of parameter breaks beyond a period dummy variable because sub-sample data may not be long enough to fit their full model with many predictors. Our proposed method can address this concern directly and provide effective solutions to the change-point regularization problem in longitudinal data analysis. Using our proposed method, applied researchers in social sciences can make a better statistical inference of time-heterogenous interaction effects and time-heterogenous causal effects from longitudinal data.

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.

Angrist, Joshua D. and Guido W. Imbens. 1995. “Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity.” *Journal of the American Statistical Association* 90(430):431–442.

Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin. 1996. “Identification of Causal Effects Using Instrumental Variables.” *Journal of the American Statistical Association* 91(434):444–455.

Armagan, Artin. 2009. “Variational Bridge Regression.” *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics* 5:17–24.

Beck, Nathaniel and Jonathan Katz. 1995. “What to Do (and Not to Do) with Time-Series Cross-Sectional Data.” *American Political Science Review* 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.

Bhadra, Anindya, Jyotishka Datta, Nicholas G. Polson and Brandon T. Willard. 2017. “Lasso Meets Horseshoe.” arXiv:1706.10179.

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.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen and Whitney Newey. 2017. “Double/Debiased/Neyman Machine Learning of Treatment Effects.” *American Economic Review* 107(5):261–265.

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

Chib, Siddhartha and Kyu Ho Kang. 2013. “Change Points in Affine Arbitrage-free Term Structure Models.” *Journal of Financial Econometrics* 11(2):302–334.

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.

Fan, Jianqing and Runze Li. 2001. "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties." *Journal of the American Statistical Association* 96(456):1348–1360.

Frank, Ildiko E. and Jerome H. Friedman. 1993. "A statistical view of some chemometrics regression tools." *Technometrics* 35:109—148.

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.

Fu, Wenjiang J. 1998. "Penalized Regressions: The Bridge versus the Lasso." *Journal of Computational and Graphical Statistics* 7(3):397–416.

Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin. 2004. *Bayesian Data Analysis.* 2nd ed. New York: Chapman and Hall.

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.

Huang, Jian, Joel L Horowitz and Shuangge Ma. 2008. "Asymptotic Properties of Bridge Estimators in Sparse High-dimensional Regression Models." *The Annals of Statistics* pp. 587–613.

Huang, Jian, Shuange Ma, Huiliange Xie and Cun-Hui Zhang. 2009. "A group bridge approach for variable selection." *Biometrika* 96(2):339–355.

Ko, Stanley I. M., Terence T. L. Chong and Pulak Ghosh. 2015. "Dirichlet Process Hidden Markov Multiple Change-point Model." *Bayesian Analysis* 10(2):275 – 296.

Kyung, Minjung, Jeff Gill, Malay Ghosh and George Casella. 2010. "Penalized Regression, Standard Errors, and Bayesian Lassos." *Bayesian Analysis* 5(2):369–412.

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. 2018. "Oracle Estimation of a Change Point in High Dimensional Quantile Regression." *Journal of the American Statistical Association* 113(523):1184–1194.

Liu, Yufeng, Hao Helen Zhang, Cheolwoo Park and Jeongyoun Ahn. 2007. "Support Vector Machines with Adaptive Lq Penalty." *Computational Statistics & Data Analysis* 51(12):6380 – 6394.

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

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.

Rueda, David. 2008. "Left Government, Policy, and Corporatism: Explaining the Influence of Partisanship on Inequality." *World Politics* 60(3):349–389.

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.

Scruggs, Lyle. 2001. "The Politics of Growth Revisited." *Journal of Politics* 63(1):120–140.

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

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* pp. 267–288.

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 Properties." *Journal of the American Statistical Association* 101(476):1418–1429.