The Lasso for High-Dimensional Regression with a Possible Change in Parameters

Myung Hwan Seo[1]

[1]London School of Economics

Joint with Sokbae Lee & Youngki Shin

August 06, 2012

Yonsei University

Threshold Models in Economics

- A threshold model offers applied researchers a simple yet useful framework to model nonlinear relationships by splitting the data into subsamples with a change-point due to a covariate threshold.
- There are a large number of applications in economics.
  - For example, Durlauf and Johnson (1995) argued that *cross-country growth models with multiple equilibria* could exhibit threshold effects, and
  - Khan and Senhadji (2001) examined the existence of threshold effects in the *relationship between inflation and growth.*
  - In addition, *racial segregation* (Card, Mas, and Rothstein, 2009) as well as *financial contagion* (Pesaran and Pick, 2007) can be modeled as a threshold effect.

Covariate Selection in Threshold Models

- Typically, the choice of the threshold variable is well motivated in applied work (e.g. *initial per capita output and/or the initial adult literacy rate* in Durlauf and Johnson (1995), and *the minority share* in a neighborhood in Card, Mas, and Rothstein (2009)), but selection of other covariates is subject to applied researchers' discretion.
- However, covariate selection is important in identifying threshold effects since a piece of evidence favoring threshold effects with a particular set of covariates could be overturned by a linear model with a broader set of regressors.

Goal of This Paper

- In this paper, we develop a method for estimating a linear regression model with a possible change-point due to a covariate threshold, while selecting relevant regressors from a set of many potential covariates.
- In particular, we consider the ℓ_1 penalized least squares (Lasso) estimator of parameters, including the unknown threshold parameter, in a sparse high-dimensional threshold model when the number of possible covariates can be much larger than the sample size.
- This project is in progress.

Literature

- The Lasso (Least Absolute Shrinkage and Selection Operator) and related methods in sparse high-dimensional settings have received much attention in statistics and has gained some interests in economics as well. For example, see Belloni and Chenozhukov (2011) and Fan, Lv, and Qi (2011) for latest reviews aimed at economics audience.
- Although there is a large literature on Lasso type methods and also equally a large literature on change points, sample splitting, and threshold models, there does not seem to be any paper yet that considers a model in the next slide.
- Our theory builds on Bickel, Ritov, and Tsybakov (2009), which cleverly exploits the sparsity of the model.

Lasso Estimation

Estimation Problem

- Let $\{(Y_i, X_i, Q_i) : i = 1, \ldots, n\}$ be a sample of independent observations such that

$$Y_i = X_i'\beta_0 + X_i'\delta_0 1\{Q_i < \tau_0\} + U_i, \quad i = 1, \ldots, n, \qquad (1)$$

  where for each i, X_i is an $M \times 1$ deterministic vector, Q_i is a deterministic scalar, U_i follows $N(0, \sigma^2)$, and $1\{\cdot\}$ denotes the indicator function.
- We are particularly interested in the case of large M, i.e. high-dimensional model.
- The inference problem here is to estimate unknown parameters $(\beta_0, \delta_0, \tau_0) \in \mathbb{R}^{2M+1}$.

Notation

- For an L-dimensional vector a, let $|a|_p$ denote the ℓ_p norm of a, and $|J|$ denote the cardinality of J, where $J(a) = \{j \in \{1, \ldots, L\} : a_j \neq 0\}$.
- In addition, let $\mathcal{M}(a)$ denote the number of nonzero elements of a. Then,

$$\mathcal{M}(a) = \sum_{j=1}^{L} 1\{a_j \neq 0\} = |J(a)|.$$

  The value $\mathcal{M}(\alpha_0)$ characterizes the sparsity of the model (1).
- For any n-dimensional vector $W = (W_1, \ldots, W_n)'$, define the empirical norm as

$$\|W\|_n := \left(n^{-1} \sum_{i=1}^{n} W_i^2 \right)^{1/2}.$$

Notation (cont.)

- Let $\mathbf{X}_i(\tau)$ denote the $(2M \times 1)$ vector such that $\mathbf{X}_i(\tau) = (X_i', X_i' 1\{Q_i < \tau\})'$ and let $\mathbf{X}(\tau)$ denote the $(n \times 2M)$ matrix whose i-th row is $\mathbf{X}_i(\tau)'$.
- Let $\alpha_0 = (\beta_0', \delta_0')'$.
- Then (1) can be written as

$$Y_i = \mathbf{X}_i(\tau_0)'\alpha_0 + U_i, \quad i = 1, \ldots, n.$$

Notation (cont.)

- Let $\mathbf{y} \equiv (Y_1, \ldots, Y_n)'$. For any fixed τ, consider the residual sum of squares

$$S_n(\alpha, \tau) = n^{-1} \sum_{i=1}^{n} \left(Y_i - X_i'\beta - X_i'\delta 1\{Q_i < \tau\} \right)^2$$
$$= \|\mathbf{y} - \mathbf{X}(\tau)\alpha\|_n^2 ,$$

  where $\alpha = (\beta', \delta')'$.

- Indicating by the superscript $^{(j)}$ the j-th element of a vector or the j-th column of a matrix, define the following $(2M \times 2M)$ diagonal matrix:

$$\mathbf{D}(\tau) := \text{diag} \left\{ \left\| \mathbf{X}(\tau)^{(j)} \right\|_n , \quad j = 1, ..., 2M \right\} .$$

- We make the following notational convention, that is, $\widehat{\mathbf{D}} = \mathbf{D}(\widehat{\tau})$ and $\mathbf{D} = \mathbf{D}(\tau_0)$, and similarly, $\widehat{S}_n = S_n(\widehat{\alpha}, \widehat{\tau})$ and $S_n = S_n(\alpha_0, \gamma_0)$, etc.

LASSO Estimation

- For each fixed τ, define the LASSO solution $\widehat{\alpha}(\tau)$ by

$$\widehat{\alpha}(\tau) := \text{argmin}_{\alpha \in \mathbb{R}^{2M}} \left\{ S_n(\alpha, \tau) + \lambda \left| \mathbf{D}(\tau)\alpha \right|_1 \right\},$$

  where λ is a tuning parameter that depends on n.
- We now estimate τ_0 by

$$\widehat{\tau} := \text{argmin}_{\tau \in \mathbb{T} \subset \mathbb{R}} \left\{ S_n(\widehat{\alpha}(\tau), \tau) + \lambda \left| \mathbf{D}(\tau)\widehat{\alpha}(\tau) \right|_1 \right\},$$

  where $\mathbb{T} := [t_0, t_1]$ is a parameter space for τ_0.
- In fact, for any finite n, $\widehat{\tau}$ is given by an interval and we simply define the maximum of the interval as our estimator.
- Then the estimator of α_0 is defined as $\widehat{\alpha} := \widehat{\alpha}(\widehat{\tau})$.

Empirical Illustrations

Empirical Growth Models

- We consider the following model specification:

$$gr_i = \beta_0 + \beta_1 lgdp60_i + X_i'\beta_2 + 1\{Q_i < \tau\}(\delta_0 + \delta_1 lgdp60_i + X_i'\delta_2) + U_i$$

  - The variable X_i is a vector of additional covariates related to education, market efficiency, political stability, market openness, demographic characteristics etc..
  - The threshold variable Q_i is either real GDP per capita or the adult literacy rate in the initial year, 1960.
- We include as many covariates (X) as possible, which would mitigate the omitted variable bias.
- The main interest would be to find that both β_1 and δ_1 are negative and where the appropriate threshold τ lies.

Empirical Growth Models (cont.)

- For comparison, we also estimate the model without the threshold effect.
- Since we use two different samples depending on the threshold variables, we estimate four different models in total. (Will be explained in the next slide.)
- Recall that our method is also robust to the case where there is no threshold effect.

Data

- We use the Barro and Lee (1994)'s dataset from 1960 to 1985. The literacy rate of each country in 1960 comes from Durlauf and Johnsen (1995).
- Depending on the selection of the threshold variable, we have 80 observations ($Q = GDP$) or 70 observations ($Q = literacy$) available in the sample.
- Also, the number of regressors are $M = 45$ (the GDP sample) and $M = 46$ (the literacy sample), respectively.
- Since we have $2M$ regressors in the threshold model, this is a high-dimensional model and we cannot estimate the model using the standard least squares method.

Selection of λ

- Theory, which will be shown later, suggests that the regularization parameter λ is $\lambda := A\sigma\sqrt{\frac{\log(3M)}{nr_n}}$.
- The parameters M, n, and r_n are determined from data.
- We set $A = 2.8$.
- Regarding σ, we first calculate its upper bound by the unconditional sample standard deviation of gr. Next, we decrease it dividing the upper bound with some constants like $25, 50, \ldots, 200$.

Estimation Results

Selected Variables by Threshold LASSO with $Q = gdp60$

Regularization Parameter	Common Effect (β)	Lower Regime Effect (δ)
λ_{max}	None	None
$\lambda_{max}/25$	ls_k, *lbmp*	None
$\lambda_{max}/50$	ls_k, *lbmp*	None
$\lambda_{max}/75$	ls_k, *gcon/gdp*, *lbmp*	*syrm60*
$\lambda_{max}/100$	ls_k, *gcon/gdp*, *wartime*, *lbmp*,	*syrm60*, *seccm60*, *wartime*
$\lambda_{max}/125$	ls_k, *lfert*, *gcon/gdp*, *wartime*, *lbmp*	*seccm60*, *wartime*
$\lambda_{max}/150$	*lgdp60*, ls_k, *lfert*, *gcon/gdp*, *wartime*, *lbmp*	*hyrm60*, *seccm60*, *tot*
$\lambda_{max}/175$	*lgdp60*, ls_k, *prim60*, *llife*, *lfert*, *gcon/gdp*, *wartime*, *lbmp*	*hyrm60*, *nom60*, *seccm60*, *wartime*, *tot*
$\lambda_{max}/200$	*lgdp60*, ls_k, *hyrm60*, *prim60* *llife*, *lfert*, *gcon/gdp*, *wartime* *lbmp*	*hyrm60*, *nom60*, *seccm60*, *wartime*, *tot*, *lgdp60*$\times$*hyrf60*

ls_k: capital savings, *lbmp*: black market premium, *gcon*: gov. con., *lfert*: fertility, *hyrm*: average higher schooling years, *secm*: percentage of secondary schooling attained.

Estimation Results (cont.)

Estimates of *lgdp60*

Regularization Parameter	LASSO	Post LASSO (Least Squares)	Threshold
$\lambda_{max}/150$	-0.0003	-0.0129^{***} (0.0032)	2798
$\lambda_{max}/175$	-0.0026	-0.0163^{***} (0.0034)	2898
$\lambda_{max}/200$	-0.0042	-0.0158^{***} (0.0034)	2898

*** significance at 1% level.

Estimation Results (cont.)

Selected Variables by Threshold LASSO with $Q = lr$

Regularization Parameter	Common Effect (β)	Lower Regime Effect (δ)
λ_{max}	None	None
$\lambda_{max}/25$	*ls_k, gcon/gdp, lbmp*	None
$\lambda_{max}/50$	*ls_k, gcon/gdp, wartime, lbmp*	*prim60, tot*
$\lambda_{max}/75$	*ls_k, gcon/gdp, wartime, lbmp*	*syrm60, wartime*
$\lambda_{max}/100$	*lgdp60, ls_k, lfert, gcon/gdp, wartime, lbmp*	*pricm60, seccm60*
$\lambda_{max}/125$	*lgdp60, ls_k, hyrm60, prim60, llife, lfert, gcon/gdp, wartime lbmp, tot*	*pricm60, seccm60, lgdp60×hyrf60*
$\lambda_{max}/150$	*lgdp60, ls_k, hyrm60, prim60, pricm60, llife, lfert, gcon/gdp, wartime, lbmp, tot*	*pricm60, seccm60, lgdp60×hyrf60*
$\lambda_{max}/175$	*lgdp60, ls_k, hyrm60, prim60, pricm60, llife, lfert, gcon/gdp, wartime, lbmp, tot*	*pricm60, seccm60, lgdp60×hyrf60*
$\lambda_{max}/200$	*lgdp60, ls_k, hyrm60, prim60, pricm60, llife, lfert, gcon/gdp, wartime, lbmp, tot*	*pricm60, seccm60, lgdp60×hyrf60*

Estimation Results (cont.)

Estimates of *lr*

Regularization Parameter	LASSO	Post LASSO (Least Squares)	Threshold
$\lambda_{max}/100$	-0.0007	-0.0106^{***} (0.0028)	82
$\lambda_{max}/125$	-0.0041	-0.0174^{***} (0.0031)	82
$\lambda_{max}/150$	-0.0063	-0.0172^{***} (0.0030)	82
$\lambda_{max}/175$	-0.0078	-0.0172^{***} (0.0030)	82
$\lambda_{max}/200$	-0.0090	-0.0172^{***} (0.0030)	82

*** significance at 1% level.

Estimation Results (cont.)

- The initial GDP (*lgdp60*) has no threshold effect.
- The threshold estimates are slightly above the mean value of *gdp60* and much above the mean value of *lr*.
- The estimates for *lgdp60* are all negative and very significant.
- Education related variables are more important to the lower regime countries and have threshold effects.
- In the linear LASSO models, we get the similar results for the estimate of *lgdp60*. However, it does not show the different effects of the education variables. Furthermore, when we use the same magnitude of decreasing steps, the number of included variables jumps from 3 to 25 at the first step, $\lambda_{max}/25$.

Analysis of the LASSO Estimator

Additional Notation

- Define
  - $f_{(\alpha,\tau)}(x,q) := x'\beta + x'\delta 1\{q < \tau\}$,
  - $f_0(x,q) := x'\beta_0 + x'\delta_0 1\{q < \tau_0\}$, and
  - $\widehat{f}(x,q) := x'\widehat{\beta} + x'\widehat{\delta} 1\{q < \widehat{\tau}\}$.
- Define

$$V_{1j} := \left(n\sigma \left\|X^{(j)}\right\|_n\right)^{-1} \sum_{i=1}^{n} U_i X_i^{(j)},$$

$$V_{2j}(\tau) := \left(n\sigma \left\|X^{(j)}(\tau)\right\|_n\right)^{-1} \sum_{i=1}^{n} U_i X_i^{(j)} 1\{Q_i < \tau\},$$

Additional Notation (cont.)

- Define the events

$$\mathbb{A} := \bigcap_{j=1}^{M} \{2|V_{1j}| \leq \mu\lambda/\sigma\},$$

$$\mathbb{B} := \bigcap_{j=1}^{M} \left\{2 \sup_{\tau \in \mathbb{T}} |V_{2j}(\tau)| \leq \mu\lambda/\sigma\right\},$$

  for a positive constant $\mu < 1$.

- Also define $J_0 := J(\alpha_0)$ and $R_n := R_n(\alpha_0, \tau_0)$, where

$$R_n(\alpha, \tau) := 2n^{-1} \sum_{i=1}^{n} U_i X_i' \delta \{1(Q_i < \widehat{\tau}) - 1(Q_i < \tau)\}.$$

A Consistency of the Lasso

- Define $X_{\max} := \max(\mathbf{D})$ and $X_{\min} := \min(\mathbf{D}(t_0))$.
- Also, let $\alpha_{\max}$ denote the maximum value that all the elements of α can take in absolute value.

Lemma (Consistency of the Lasso)

Conditional on the event $\mathbb{A} \bigcap \mathbb{B}$, we have

$$\begin{aligned} & \left\| \widehat{f} - f_0 \right\|_n^2 + (1-\mu)\lambda \left| \widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0) \right|_1 \\ \leq \quad & 6\lambda X_{\max} \alpha_{\max} \mathcal{M}(\alpha_0) + 2\mu\lambda X_{\max} \left| \delta_0 \right|_1 . \end{aligned}$$

This lemma implies that

$$\left\| \widehat{f} - f_0 \right\|_n \lesssim \sqrt{\lambda \mathcal{M}(\alpha_0)},$$

thereby establishing a consistency of the Lasso, provided that $\lambda \mathcal{M}(\alpha_0) \to 0$ and the probability of $\mathbb{A} \bigcap \mathbb{B}$ tends to one asymptotically.

Oracle Inequalities of the Lasso

- We now give oracle inequalities of the Lasso.

Lemma (Oracle Inequalities of the Lasso)

Assume that Assumption 3 holds with $\kappa = \kappa(s, \frac{1+\mu+L_1}{1-\mu})$ *for* $\mu < 1$ *and* $\mathcal{M}(\alpha_0) \leq s \leq M$. *Also let Assumption 4 or 5 hold. Then conditional on the event* $\mathbb{A} \bigcap \mathbb{B}$, *we have*

$$\left\| \widehat{f} - f_0 \right\|_n^2 \leq \frac{(2 + L_1)^2 X_{\max}^2 L_2}{\kappa^2} \lambda^2 \mathcal{M}(\alpha_0).$$

and

$$|\widehat{\alpha} - \alpha_0|_1 \leq \frac{(2 + L_1)^2 X_{\max}^2 L_2}{(1 - \mu) X_{\min} \kappa^2} \lambda \mathcal{M}(\alpha_0).$$

Oracle Inequalities of the Lasso (cont.)

- The first oracle inequality implies that

$$\left\| \widehat{f} - f_0 \right\|_n \lesssim \lambda \sqrt{\mathcal{M}(\alpha_0)},$$

  which gives a faster rate of convergence than the previous lemma such that

$$\left\| \widehat{f} - f_0 \right\|_n \lesssim \sqrt{\lambda} \sqrt{\mathcal{M}(\alpha_0)}.$$

- Also, it gives that

$$|\widehat{\alpha} - \alpha_0|_1 \lesssim \lambda \mathcal{M}(\alpha_0).$$

Restricted Eigenvalue (RE) Assumption

- Assumption (Restricted Eigenvalue (RE) (s, c_0))

  For some integer s such that $1 \leq s \leq 2M$ and a positive number c_0, the following condition holds:

$$\kappa(s, c_0) := \min_{\substack{J_0 \subseteq \{1,\ldots,2M\}, \\ |J_0| \leq s}} \min_{\substack{\gamma \neq 0, \\ \left|\gamma_{J_0^c}\right|_1 \leq c_0 \left|\gamma_{J_0}\right|_1}} \frac{|\mathbf{X}(\tau_0)\gamma|_2}{\sqrt{n}|\gamma_{J_0}|_2} > 0.$$

  - This is just a restatement of restricted eigenvalue assumption of Bickel, Ritov, and Tsybakov (2009) when τ_0 were known.
  - It is clearly weaker than the p.d. of $n^{-1}\mathbf{X}(\tau_0)'\mathbf{X}(\tau_0)$ but it can be shown that the square matrix of any $2s$-dimensional submatrix of $\mathbf{X}(\tau_0)$ is p.d.
  - This ensures the uniqueness of the sparse representation and is useful to obtain Oracle inequalities.

Oracle Conditions

- Assumption (Oracle Condition I)

tion I *For some positive constants L_1, either of the following conditions holds:*

$$\left\| f_{(\alpha_0,\widehat{\tau})} - f_0 \right\|_n^2 \leq L_1 \lambda \left| \hat{\mathbf{D}} \left(\hat{\alpha} - \alpha_0 \right)_{J_0} \right|_1, \tag{2}$$

$$\lambda \left(\left| \widehat{\mathbf{D}}^{1/2} \alpha_0 \right|_1 - \left| \mathbf{D}^{1/2} \alpha_0 \right|_1 \right) + R_n \leq L_1 \lambda \left| \hat{\mathbf{D}}^{1/2} \left(\hat{\alpha} - \alpha_0 \right)_{J_0} \right|_1, \tag{3}$$

- This assumption is a high-level assumption that is useful to obtain an oracle inequality.
- If $\delta_0 = 0$, the LHS is zero and thus they are trivially satisfied with $L_1 = 0$.
- the RHS depends on $\hat{\alpha} - \alpha_0$ while the LHS relies on $\hat{\tau} - \tau_0$. Thus, these conditions impose restrictions on the relative convergence of $\hat{\alpha}$ and $\hat{\tau}$ to their true values, respectively.

Oracle Conditions

Assumption (Oracle Condition II)

For some positive constant $L_2 < \infty$, the following condition holds:

$$\left\| f_{(\widehat{\alpha},\tau_0)} - f_0 \right\|_n^2 \leq L_2 \left\| \widehat{f} - f_0 \right\|_n^2. \tag{4}$$

- If $|\hat{\tau} - \tau_0|$ is of the same order as $|\hat{\alpha} - \alpha_0|$, then this should hold with large probabilities.
- If $\delta_0 = 0$, it would hold as well provided that $\hat{\beta}$ and $\hat{\delta}$ converge at the same rate.

Sparsity of the Lasso

- We now provide an oracle inequality regarding the sparsity of the Lasso estimator $\widehat{\alpha}$.

Lemma (Sparsity of the Lasso)

Assume that the RE assumption with $\kappa = \kappa(\mathcal{M}(\alpha_0), \frac{1+\mu}{1-\mu})$ *for* $\mu < 1$ *and the oracle condition assumption hold. Assume further that the largest eigenvalue of* $\mathbf{X}(\tau)'\mathbf{X}(\tau)/n$ *is bounded uniformly in* $\tau \in \mathbb{T}$ *by* $\phi_{\max}$. *Then conditional on the event* $\mathbb{A} \bigcap \mathbb{B}$, *we have*

$$\mathcal{M}(\widehat{\alpha}) \leq \frac{16\phi_{\max} L_2}{(1-\mu)^2 (1-L_1)^2 \kappa^2} \frac{X_{\max}^2}{X_{\min}^2} \mathcal{M}(\alpha_0).$$

Probability of $\mathbb{A} \cap \mathbb{B}$

- We now establish conditions under which $\mathbb{A} \cap \mathbb{B}$ has probability close to one with a suitable choice of λ.
- Define

$$r_n = \min_{1 \le j \le M} \frac{\left\| X^{(j)}(t_0) \right\|_n^2}{\left\| X^{(j)} \right\|_n^2},$$

where $X^{(j)}(\tau) \equiv (X_1^{(j)} 1\{Q_1 < \tau\}, \ldots, X_n^{(j)} 1\{Q_n < \tau\})'$ as before.

Lemma (Probability of $\mathbb{A} \cap \mathbb{B}$)

Let Φ denote the cdf function of the standard normal. Then,

$$\mathbb{P}\{\mathbb{A} \cap \mathbb{B}\} \ge 1 - 6M\Phi\left(-\frac{\mu\sqrt{nr_n}}{2\sigma}\lambda\right).$$

Theorem (Oracle Inequalities)

Assume that the same conditions as above hold. Let $\hat{\alpha}$ *be obtained with*

$$\lambda = A\sigma\left(\frac{\log 3M}{nr_n}\right)^{1/2}$$

and $A > 2\sqrt{2}/\mu$ *and* $\mu < 1$. *Then, with probability at least* $1 - (3M)^{1-A^2\mu^2/8}$, *we have*

$$\begin{aligned}
\left\|\widehat{f} - f_0\right\|_n &\leq \frac{2A\sigma\sqrt{L_2}}{(1-L_1)\kappa}\sqrt{\frac{\mathcal{M}(\alpha_0)\log 3M}{nr_n}}X_{\max}, \\
|\widehat{\alpha} - \alpha_0|_1 &\leq \frac{4A\sigma L_2}{(1-L_1)(1-\mu)\kappa^2}\sqrt{\frac{\log 3M}{nr_n}}\mathcal{M}(\alpha_0)\frac{X_{\max}^2}{X_{\min}},
\end{aligned}$$

and

$$\mathcal{M}(\widehat{\alpha}) \leq \frac{16\phi_{\max}L_2}{(1-\mu)^2(1-L_1)^2\kappa^2}\frac{X_{\max}^2}{X_{\min}^2}\mathcal{M}(\alpha_0).$$

The probability in the main theorem is computed by bounding $2\Phi(x)$ by $\exp\left(-x^2/2\right)$ as in BRT.

Selection of λ

- It is necessary to choose σ.
- Since σ is unknown, we may employ iteration.
  - First, we set $\sigma = \left(n^{-1} \sum_{i=1}^{n} \left(Y_i - n^{-1} \sum_{i=1}^{n} Y_i \right)^2 \right)^{1/2}$ and compute $\widehat{S}_n \left(\hat{\alpha}, \hat{\tau} \right)$.
  - Second, let $\sigma = \sqrt{\widehat{S}_n}$ and estimate (α, τ). This is reasonable since the sample variance of Y is an upper bound for σ^2.
  - Third, let $\sigma = \sqrt{RSS}$ and estimate (α, τ).
  - Fourth, continue till converge.

Discontinuous Threshold Model

Discontinuous Threshold Model

We add the following conditions to characterize the discontinuous model.

Assumption (Identification under Sparsity)

For some $s \geq \mathcal{M}(\alpha_0)$, *and for any* $\alpha \in \{\alpha : \mathcal{M}(\alpha) \leq s\}$ *and* τ *such that* $(\alpha, \tau) \neq (\alpha_0, \tau_0)$ *and* $|\tau - \tau_0| \geq \min_{i \neq j} |Q_i - Q_j|$,

$$\left\| f_{(\alpha,\tau)} - f_{(\alpha_0,\tau_0)} \right\|_n \neq 0.$$

Assumption (Discontinuity of Regression)

For a given $s \geq \mathcal{M}(\alpha_0)$, *and for any* η *and* τ *s.t.* $|\tau - \tau_0| \geq \eta > \min_{i \neq j} |Q_i - Q_j|$ *and* $\alpha \in \{\alpha : \mathcal{M}(\alpha) \leq s, \}$, *there exists a* $c > 0$ *such that*

$$\left\| f_{(\alpha,\tau)} - f_{(\alpha_0,\tau_0)} \right\|_n^2 \geq c\eta > 0.$$

Remark on Discontinuity Assumption

- The classical threshold regression model with discontinuity satisfies the condition under a fairly general condition that Q has a density that is continuous and positive everywhere.
- To justify it, the paper works out a simple case where the only regressor is the constant 1,

$$\begin{aligned}
& \mathrm{E}\left(y_i - f_i\left(\alpha, \tau\right)\right)^2 - \mathrm{E}\left(y_i - f_i\left(\alpha_0, \tau_0\right)\right)^2 \\
= \; & \mathrm{E}\left(f_i\left(\alpha_0, \tau_0\right) - f_i\left(\alpha, \tau\right)\right)^2 \\
= \; & \left(\alpha_1 - \alpha_{10}\right)^2 P\left(Q < \tau \wedge \tau_0\right) + \left(\alpha_2 - \alpha_{20}\right)^2 P\left(Q \geq \tau \vee \tau_0\right) \\
& + \left(\alpha_2 - \alpha_{10}\right)^2 P\left(\tau \leq Q < \tau_0\right) + \left(\alpha_1 - \alpha_{20}\right)^2 P\left(\tau_0 \leq Q < \tau\right) \\
\geq \; & c\left|\tau - \tau_0\right|,
\end{aligned}$$

for some c, where $\alpha_1 = \beta + \delta$ and $\alpha_2 = \beta$, unless $|\alpha_2 - \alpha_{10}|$ is too small when $\tau < \tau_0$ or $|\alpha_1 - \alpha_{20}|$ is too small when $\tau > \tau_0$. However, when $|\alpha_2 - \alpha_{10}|$ is small, say smaller than ε, $|\alpha_2 - \alpha_{20}|$ is bounded above zero due to the discontinuity that $\alpha_{10} \neq \alpha_{20}$ and $P\left(Q \geq \tau \vee \tau_0\right) = P\left(Q \geq \tau_0\right)$ is also bounded above zero. This implies the inequality still holds. Since the same reasoning applies for the latter case, we can conclude our discontinuity assumption holds in the standard discontinuous threshold regression setup.

Additional Assumptions (cont.)

Assumption (Smoothness of Design)

For any $\eta > 0$, there exists C such that

$$\sup_j \sup_{|\tau-\tau_0|<\eta} \left| \frac{1}{n} \sum_{i=1}^{n} \left| X_i^{(j)} \right|^2 \left[1\left(Q_i < \tau_0\right) - 1\left(Q_i < \tau\right) \right] \right| \leq C\eta.$$

Now, introduce an event $\mathbb{D}$, which is defined as

$$\left\{ \sup_{|\tau-\tau_0|<\eta} \left| \frac{2}{n} \sum_{i=1}^{n} U_i X_i' \delta_0 \left[1\left(Q_i < \tau_0\right) - 1\left(Q_i < \tau\right) \right] \right| \leq \lambda\sqrt{\eta} : \eta_1 \leq \eta \leq \eta_2 \right.$$

where $\eta_1 = \frac{36 L_2}{c(1-\mu)\kappa^2} \frac{X_{\max}^3}{X_{\min}} \lambda^2 s$ and $\eta_2 = 5\lambda X_{\max} \alpha_{\max}\left(\mathcal{M}\left(\hat{\alpha}\right)\right)/c$.

Theorem (Oracle Inequalities for Discontiuous Threshold Model)

Under certain regularity conditions and

$$\lambda = A\sigma\left(\frac{\log 3M}{nr_n}\right)^{1/2}$$

with $A > 2\sqrt{2}/\mu$, *we have*

$$\begin{aligned}
\left\|\widehat{f} - f_0\right\|_n &\leq \frac{3A\sigma\sqrt{L_2}}{\kappa}\left(\frac{\log 3M}{nr_n}\right)^{1/2}\sqrt{s}X_{\max}, \\
\left|\widehat{\alpha} - \alpha_0\right|_1 &\leq \frac{9A\sigma L_2}{(1-\mu)\kappa^2}\frac{X_{\max}^2}{X_{\min}}\left(\frac{\log 3M}{nr_n}\right)^{1/2}s.
\end{aligned}$$

and

$$\begin{aligned}
\left|\hat{\tau} - \tau_0\right| &\leq \frac{36A^2\sigma^2 L_2}{c(1-\mu)\kappa^2}\frac{X_{\max}^3}{X_{\min}}\frac{\log 3M}{nr_n}s \\
\mathcal{M}(\hat{\alpha}) &\leq \frac{36\phi_{\max}L_2}{(1-\mu)^2\kappa^2}\frac{X_{\max}^2}{X_{\min}^2}s,
\end{aligned}$$

with probability at least $1 - (3M)^{1-A^2\mu^2/8} - 8(3M)^{-A^2h_n^2/8r_n}$, *provided that* $\lambda < c(1-\mu)X_{\min}^2\left(12X_{\max}C\left|\delta_0\right|_1\right)^{-1}$ *and* $\sqrt{\eta_2} \geq (2X_{\min})^{-1}C\left|\delta_0\right|_1\eta_2$.

Monte Carlo Simulation

Simulation Design

- The base model

$$Y_i = X_i'\beta_0 + X_i'\delta_0 1\{Q_i < \tau_0\} + U_i, \quad i = 1, \ldots, n,$$

  where X_i is a M-dimensional vector generated from $N(0, I)$, Q_i is a scalar generated from the uniform distribution on the interval of $(0, 1)$, and the error term U_i is generated from $N(0, 0.5^2)$.
- The threshold parameter is set as $\tau_0 = 0.3, 0.4$, or 0.5
- And $\beta_0 = (1, 0, 1, 0, \ldots, 0)$, and $\delta_0 = c \cdot (0, -1, 1, 0, \ldots, 0)$ where $c = 0$ or 1. Thus, there is no threshold effect when $c = 0$.
- The number of observations is set as $n = 200$.
- Finally, the dimension of X_i is set as $M = 50, 100$, and 200, so that the total number of regressors are 100, 200, and 400, respectively.

Estimation Algorithm

- A slight modification of the standard LASSO/LARS algorithm:
  - Given the regularization parameter λ, we estimate the model for each grid point of τ that spans over 71 equi-spaced points on the interval of $[0.15, 0.85]$.
- Next, choose $\widehat{\tau}$ by

$$\widehat{\tau} := \arg \min_{\tau \in \mathcal{T} \subset \mathbb{R}} \left\{ \widehat{S}\left(\widehat{\alpha}\left(\tau\right), \tau\right) + \lambda \left| D\left(\tau\right)^{1/2} \widehat{\alpha}\left(\tau\right) \right|_1 \right\}$$

  and $\widehat{\alpha} := \widehat{\alpha}(\widehat{\tau})$.
- The regularization parameter λ is chosen by

$$\lambda := A \times \sigma \sqrt{\frac{\log\left(3M\right)}{n r_n}}$$

  where $r_n = \min_j ||X^{(j)}(t_0)||_n^2 / ||X^{(j)}||_n^2$ and $\sigma = 0.5$ is assumed to be known.
- For the constant A, we use four different values such as $A = 2.8, 3.2, 3.6, 4.0$.

Comparison

- Least Squares when possible
- Oracle 1 - knows which variables are relevant
- Oracle 2 - knows the true threshold value τ_0 in addition.
- Criterion: mean-squared prediction error (PE), which is computed numerically for each sample as follows.
  - For each sample s and corresponding estimates $\widehat{\beta}_s$, $\widehat{\delta}_s$, and $\widehat{\tau}_s$, we generate a new data $\{Y_j, X_j, Q_j\}$ of 400 observations and calculate

$$\widehat{PE}_s = \frac{1}{400} \sum_{j=1}^{400} \left(g(x_j, q_j; \beta_0, \delta_0, \tau_0) - g(x_j, q_j; \widehat{\beta}_s, \widehat{\delta}_s, \widehat{\tau}_s) \right)^2$$

    where $g(x, q; \beta, \delta, \tau) = x'\beta + x'\delta 1\{q < \tau\}$. The mean, median, and standard deviation of the prediction error are calculated from the 400 replications, $\{\widehat{PE}_s\}_{s=1}^{400}$.
- Dependence: Σ has the form of $(\Sigma)_{i,j} = \rho^{|i-j|}$ for $i, j = 1, \ldots, M$, and $\rho = 0.1, 0.3$, and 0.5.

Table: Simulation Results ($M = 50$, $\tau_0 = 0.5$)

Jump Scale	Estimation Method	Regularization Parameter	Prediction Error			$\hat{\tau}$ Consistency	
			Mean	Median	SD	Bias	RMSE
$c = 1$	LS	None	0.285	0.276	0.074	-0.001	0.014
	LASSO	$A = 2.8$	0.041	0.030	0.035	-0.002	0.020
		$A = 3.2$	0.048	0.033	0.049	-0.001	0.029
		$A = 3.6$	0.067	0.037	0.086	0.007	0.059
		$A = 4.0$	0.095	0.050	0.120	0.024	0.088
	Oracle 1	None	0.013	0.006	0.019	-0.001	0.009
	Oracle 2	None	0.005	0.004	0.004	0.000	0.000
$c = 0$	LS	None	6.332	0.460	41.301	N/A	
	LASSO	$A = 2.8$	0.013	0.011	0.007		
		$A = 3.2$	0.014	0.012	0.008		
		$A = 3.6$	0.015	0.014	0.009		
		$A = 4.0$	0.017	0.016	0.010		
	Oracle 1	None	0.009	0.008	0.005		
	Oracle 2	None	0.005	0.004	0.004		

Note: M denotes the column size of X_i and τ denotes the threshold parameter. Oracle 1 & 2 are estimated by the LS when sparcity is known and when sparcity and τ_0 are known, respectively. All simulations are based on 400 replications of a sample with 200 observations.

Table: Simulation Results ($M = 50$, $\tau_0 = 0.3$ or 0.4, $c = 1$)

Threshold	Estimation Method	Regularization Parameter	Prediction Error			$\hat{\tau}$ Consistency	
			Mean	Median	SD	Bias	RMSE
$\tau = 0.3$	LS	None	2.559	0.511	16.292	0.008	0.021
	LASSO	$A = 2.8$	0.062	0.035	0.091	0.014	0.101
		$A = 3.2$	0.089	0.041	0.125	0.037	0.150
		$A = 3.6$	0.127	0.054	0.159	0.078	0.208
		$A = 4.0$	0.185	0.082	0.185	0.147	0.280
	Oracle 1	None	0.012	0.006	0.017	-0.001	0.008
	Oracle 2	None	0.005	0.004	0.004	0.000	0.000
$\tau = 0.4$	LS	None	0.317	0.304	0.095	-0.000	0.014
	LASSO	$A = 2.8$	0.052	0.034	0.063	-0.001	0.043
		$A = 3.2$	0.063	0.037	0.083	0.001	0.065
		$A = 3.6$	0.090	0.045	0.121	0.016	0.103
		$A = 4.0$	0.133	0.061	0.162	0.054	0.157
	Oracle 1	None	0.014	0.006	0.022	-0.002	0.009
	Oracle 2	None	0.005	0.004	0.004	0.000	0.000

Note: The results with $c = 0$, i.e. no threshold, are very similar to those in Table 1. The notation and estimation methods are explained in the footnote of Table 1.

Table: Simulation Results ($M = 100$, $\tau_0 = 0.5$)

Jump Scale	Estimation Method	Regularization Parameter	Prediction Error			$\hat{\tau}$ Consistency	
			Mean	Median	SD	Bias	RMSE
$c = 1$	LS	None		N/A		N/A	
	LASSO	$A = 2.8$	0.048	0.033	0.046	-0.003	0.026
		$A = 3.2$	0.056	0.037	0.059	-0.004	0.034
		$A = 3.6$	0.072	0.045	0.087	0.003	0.054
		$A = 4.0$	0.101	0.056	0.122	0.018	0.087
	Oracle 1	None	0.013	0.006	0.018	-0.002	0.009
	Oracle 2	None	0.005	0.005	0.004	0.000	0.000
$c = 0$	LS	None		N/A		N/A	
	LASSO	$A = 2.8$	0.016	0.015	0.008		
		$A = 3.2$	0.017	0.015	0.009		
		$A = 3.6$	0.019	0.017	0.010		
		$A = 4.0$	0.021	0.019	0.012		
	Oracle 1	None	0.009	0.009	0.005		
	Oracle 2	None	0.005	0.005	0.004		

Table: Simulation Results ($M = 100$, $\tau_0 = 0.3$ or 0.4, $c = 1$)

Threshold	Estimation Method	Regularization Parameter	Prediction Error			$\hat{\tau}$ Consistency	
			Mean	Median	SD	Bias	RMSE
$\tau = 0.3$	LS	None		N/A		N/A	
	LASSO	$A = 2.8$	0.060	0.038	0.074	0.004	0.081
		$A = 3.2$	0.084	0.043	0.115	0.026	0.138
		$A = 3.6$	0.121	0.051	0.154	0.069	0.201
		$A = 4.0$	0.175	0.067	0.186	0.135	0.274
	Oracle 1	None	0.015	0.006	0.021	-0.002	0.009
	Oracle 2	None	0.005	0.004	0.004	0.000	0.000
$\tau = 0.4$	LS	None		N/A		N/A	
	LASSO	$A = 2.8$	0.052	0.038	0.053	-0.002	0.042
		$A = 3.2$	0.064	0.044	0.069	-0.002	0.056
		$A = 3.6$	0.095	0.052	0.123	0.023	0.112
		$A = 4.0$	0.135	0.066	0.161	0.054	0.160
	Oracle 1	None	0.014	0.006	0.021	-0.002	0.008
	Oracle 2	None	0.005	0.005	0.004	0.000	0.000

Table: Simulation Results ($M = 200$, $\tau_0 = 0.5$)

Jump Scale	Estimation Method	Regularization Parameter	Prediction Error			$\hat{\tau}$ Consistency	
			Mean	Median	SD	Bias	RMSE
$c = 1$	LS	None		N/A		N/A	
	LASSO	$A = 2.8$	0.049	0.039	0.033	-0.004	0.016
		$A = 3.2$	0.058	0.043	0.050	-0.002	0.031
		$A = 3.6$	0.080	0.054	0.083	0.001	0.051
		$A = 4.0$	0.121	0.069	0.137	0.015	0.097
	Oracle 1	None	0.015	0.007	0.020	-0.002	0.009
	Oracle 2	None	0.006	0.005	0.004	0.000	0.000
$c = 0$	LS	None		N/A		N/A	
	LASSO	$A = 2.8$	0.019	0.018	0.010		
		$A = 3.2$	0.020	0.019	0.011		
		$A = 3.6$	0.023	0.021	0.012		
		$A = 4.0$	0.026	0.024	0.014		
	Oracle 1	None	0.010	0.009	0.005		
	Oracle 2	None	0.006	0.005	0.004		

Table: Simulation Results ($M = 200$, $\tau_0 = 0.3$ or 0.4, $c = 1$)

Threshold	Estimation Method	Regularization Parameter	Prediction Error			$\hat{\tau}$ Consistency	
			Mean	Median	SD	Bias	RMSE
$\tau = 0.3$	LS	None		N/A		N/A	
	LASSO	$A = 2.8$	0.067	0.044	0.071	-0.004	0.061
		$A = 3.2$	0.099	0.052	0.131	0.024	0.139
		$A = 3.6$	0.141	0.064	0.171	0.065	0.205
		$A = 4.0$	0.191	0.092	0.195	0.111	0.261
	Oracle 1	None	0.014	0.006	0.018	-0.001	0.009
	Oracle 2	None	0.005	0.004	0.004	0.000	0.000
$\tau = 0.4$	LS	None		N/A		N/A	
	LASSO	$A = 2.8$	0.054	0.043	0.039	-0.006	0.020
		$A = 3.2$	0.074	0.050	0.085	-0.003	0.062
		$A = 3.6$	0.104	0.062	0.128	0.013	0.103
		$A = 4.0$	0.155	0.082	0.175	0.040	0.160
	Oracle 1	None	0.016	0.007	0.022	-0.002	0.010
	Oracle 2	None	0.006	0.005	0.004	0.000	0.000

Table: Simulation Results ($M = 50$, $\tau_0 = 0.5$, $\rho = 0.1$)

Jump Scale	Estimation Method	Regularization Parameter	Prediction Error			$\hat{\tau}$ Consistency	
			Mean	Median	SD	Bias	RMSE
$c = 1$	LS	None	0.279	0.271	0.067	-0.001	0.013
	LASSO	$A = 2.8$	0.048	0.033	0.047	-0.002	0.029
		$A = 3.2$	0.059	0.037	0.068	0.003	0.047
		$A = 3.6$	0.089	0.045	0.118	0.020	0.094
		$A = 4.0$	0.121	0.060	0.141	0.036	0.121
	Oracle 1	None	0.012	0.006	0.016	-0.001	0.009
	Oracle 2	None	0.005	0.004	0.004	0.000	0.000
$c = 0$	LS	None	6.939	0.437	42.698	N/A	
	LASSO	$A = 2.8$	0.013	0.012	0.008		
		$A = 3.2$	0.014	0.012	0.009		
		$A = 3.6$	0.015	0.013	0.010		
		$A = 4.0$	0.017	0.015	0.011		
	Oracle 1	None	0.010	0.009	0.005		
	Oracle 2	None	0.005	0.004	0.004		

Table: Simulation Results ($M = 50$, $\tau_0 = 0.5$, $\rho = 0.3$)

Jump Scale	Estimation Method	Regularization Parameter	Prediction Error			$\hat{\tau}$ Consistency	
			Mean	Median	SD	Bias	RMSE
$c = 1$	LS	None	0.283	0.273	0.075	-0.001	0.018
	LASSO	$A = 2.8$	0.075	0.043	0.087	-0.002	0.086
		$A = 3.2$	0.108	0.059	0.115	0.014	0.129
		$A = 3.6$	0.160	0.099	0.137	0.041	0.177
		$A = 4.0$	0.208	0.181	0.143	0.062	0.217
	Oracle 1	None	0.013	0.006	0.017	-0.001	0.011
	Oracle 2	None	0.005	0.004	0.004	0.000	0.000
$c = 0$	LS	None	6.939	0.437	42.698	N/A	
	LASSO	$A = 2.8$	0.012	0.011	0.007		
		$A = 3.2$	0.013	0.011	0.008		
		$A = 3.6$	0.014	0.013	0.009		
		$A = 4.0$	0.016	0.014	0.010		
	Oracle 1	None	0.009	0.008	0.005		
	Oracle 2	None	0.005	0.004	0.004		

Table: Simulation Results ($M = 50$, $\tau_0 = 0.5$, $\rho = 0.5$)

Jump Scale	Estimation Method	Regularization Parameter	Prediction Error			$\hat{\tau}$ Consistency	
			Mean	Median	SD	Bias	RMSE
$c = 1$	LS	None	0.289	0.277	0.077	-0.003	0.023
	LASSO	$A = 2.8$	0.153	0.133	0.109	0.043	0.209
		$A = 3.2$	0.202	0.231	0.106	0.042	0.257
		$A = 3.6$	0.232	0.255	0.098	0.049	0.283
		$A = 4.0$	0.259	0.271	0.088	0.027	0.300
	Oracle 1	None	0.012	0.006	0.014	-0.002	0.012
	Oracle 2	None	0.005	0.004	0.004	0.000	0.000
$c = 0$	LS	None	6.939	0.437	42.698	N/A	
	LASSO	$A = 2.8$	0.011	0.010	0.007		
		$A = 3.2$	0.012	0.011	0.007		
		$A = 3.6$	0.013	0.012	0.008		
		$A = 4.0$	0.014	0.013	0.009		
	Oracle 1	None	0.009	0.008	0.005		
	Oracle 2	None	0.005	0.004	0.004		

Conclusion

Conclusion

1. We propose the Lasso for high-dimensional regression with a possible change-point.
2. The method was illustrated by the growth model with multiple equilibria.
3. We derive the Oracle inequalities (non-asymptotic) for the Lasso estimators and provide regularity conditions.
4. Main advantage of the proposed method is that it works without prior knowledge of the presence of change-point.
5. Numerical study demonstrates that it works well under various scenarios.