The lasso for high dimensional regression with a possible change point

Sokbae Lee,

*Seoul National University, Republic of Korea, and Institute for Fiscal Studies, London, UK*

Myung Hwan Seo

*London School of Economics and Political Science, UK, and Seoul National University, Republic of Korea*

and Youngki Shin

*University of Western Ontario, London, Canada*

[Received August 2013. Final revision October 2014]

Summary. We consider a high dimensional regression model with a possible change point due to a covariate threshold and develop the lasso estimator of regression coefficients as well as the threshold parameter. Our lasso estimator not only selects covariates but also selects a model between linear and threshold regression models. Under a sparsity assumption, we derive non-asymptotic oracle inequalities for both the prediction risk and the $l_1$-estimation loss for regression coefficients. Since the lasso estimator selects variables simultaneously, we show that oracle inequalities can be established without pretesting the existence of the threshold effect. Furthermore, we establish conditions under which the estimation error of the unknown threshold parameter can be bounded by a factor that is nearly $n^{-1}$ even when the number of regressors can be much larger than the sample size $n$. We illustrate the usefulness of our proposed estimation method via Monte Carlo simulations and an application to real data.

Keywords: Lasso; Oracle inequalities; Sample splitting; Sparsity; Threshold models

1. Introduction

The lasso and related methods have received rapidly increasing attention in statistics since the seminal work of Tibshirani (1996). For example, see Bühlmann and van de Geer (2011) as well as Fan and Lv (2010) and Tibshirani (2011) for a general overview and recent developments.

In this paper, we develop a method for estimating a high dimensional 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 propose the $l_1$ penalized least squares (lasso) estimator of parameters, including the unknown threshold parameter, and analyse its properties under a sparsity assumption when the number of possible covariates can be much larger than the sample size.

To be specific, let $\{(Y_i, X_i, Q_i) : i = 1, \ldots, n\}$ be a sample of independent observations such that
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 $\mathbf{1}\{\cdot\}$ denotes the indicator function. The scalar variable $Q_i$ is the threshold variable and $\tau_0$ is the unknown threshold parameter. Since $Q_i$ is a fixed variable in our set-up, expression (1.1) includes a regression model with a change point at unknown time (e.g. $Q_i = i/n$). In this paper, we focus on the fixed design for $\{(X_i, Q_i): i = 1, \ldots, n\}$ and independent normal errors $\{U_i: i = 1, \ldots, n\}$. This set-up has been extensively used in the literature (e.g. Bickel et al. (2009)).

A regression model such as model (1.1) offers applied researchers a simple yet useful framework to model non-linear relationships by splitting the data into subsamples. Empirical examples include cross-country growth models with multiple equilibria (Durlauf and Johnson, 1995), racial segregation (Card et al., 2008) and financial contagion (Pesaran and Pick, 2007), among many others. Typically, the choice of the threshold variable is well motivated in applied work (e.g. initial per capita output in Durlauf and Johnson (1995), and the minority share in a neighbourhood in Card et al. (2008)), but selection of other covariates is subject to applied researchers’ discretion.

However, covariate selection is important in identifying threshold effects (i.e. non-zero $\delta_0$) since a statistical model favouring threshold effects with a particular set of covariates could be overturned by a linear model with a broader set of regressors. Therefore, it seems natural to consider the lasso as a tool to estimate model (1.1).

The statistical problem that we consider is to estimate unknown parameters $(\beta_0, \delta_0, \tau_0) \in \mathbb{R}^{2M+1}$ when $M$ is much larger than $n$. For the classical set-up (estimation of parameters without covariate selection when $M$ is smaller than $n$), estimation of model (1.1) has been well studied (e.g. Tong (1990), Chan (1993) and Hansen (2000)). Also, a general method for testing threshold effects in regression (i.e. testing $H_0: \delta_0 = 0$ in model (1.1)) is available for the classical set-up (e.g. Lee et al. (2011)).

Although there are many references on lasso-type methods and also equally many on change points, sample splitting and threshold models, there seem to be only a handful of references that intersect both topics. Wu (2008) proposed an information-based criterion for carrying out change point analysis and variable selection simultaneously in linear models with a possible change point; however, the method proposed in Wu (2008) would be infeasible in a sparse high dimensional model. In change point models without covariates, Harchaoui and Lévy-Leduc (2008, 2010) proposed a method for estimating the location of change points in one-dimensional piecewise constant signals observed in white noise, using a penalized least square criterion with an $l_1$-type penalty. Zhang and Siegmund (2012) developed Bayes information criterion like criteria for determining the number of changes in the mean of multiple sequences of independent normal observations when the number of change points can increase with the sample size. Ciuperca (2014) considered a similar estimation problem to ours, but the corresponding analysis was restricted to the case when the number of potential covariates is small.

In this paper, we consider the lasso estimator of regression coefficients as well as the threshold parameter. Since the change point parameter $\tau_0$ does not enter additively in model (1.1), the resulting optimization problem in the lasso estimation is non-convex. We overcome this problem by comparing the values of standard lasso objective functions on a grid over the range of possible values of $\tau_0$.

Theoretical properties of the lasso and related methods for high dimensional data have been examined by Fan and Peng (2004), Bunea et al. (2007), Candès and Tao (2007), Huang et al. (2008a,b), Kim et al. (2008), Bickel et al. (2009) and Meinshausen and Yu (2009), among
many others. Most of the references consider quadratic objective functions and linear or nonparametric models with an additive mean 0 error. There has been recent interest in extending this framework to generalized linear models (e.g. van de Geer (2008) and Fan and Lv (2011)), to quantile regression models (e.g. Belloni and Chernozhukov (2011a), Bradic et al. (2011) and Wang et al. (2012)), and to hazards models (e.g. Bradic et al. (2012) and Lin and Lv (2013)). We contribute to this literature by considering a regression model with a possible change point and then deriving non-asymptotic oracle inequalities for both the prediction risk and the $l_1$ estimation loss for regression coefficients under a sparsity scenario.

Our theoretical results build on Bickel et al. (2009). Since the lasso estimator selects variables simultaneously, we show that oracle inequalities that are similar to those obtained in Bickel et al. (2009) can be established without pretesting the existence of the threshold effect. In particular, when there is no threshold effect ($\delta_0 = 0$), we prove oracle inequalities that are basically equivalent to those in Bickel et al. (2009). Furthermore, when $\delta_0 \neq 0$, we establish conditions under which the estimation error of the unknown threshold parameter can be bounded by a factor of nearly $n^{-1}$ when the number of regressors can be much larger than the sample size. To achieve this, we develop some sophisticated chaining arguments and provide sufficient regularity conditions under which we prove oracle inequalities. The superconsistency result of $\hat{\tau}$ we develop is well known when the number of covariates is small (see, for example, Chan (1993) and Seijo and Sen (2011a, b)). To the best of our knowledge, our paper is the first work that demonstrates the possibility of a nearly $n^{-1}$-bound in the context of sparse high dimensional regression models with a change point.

The remainder of this paper is as follows. In Section 2 we propose the lasso estimator, and in Section 3 we give a brief illustration of our proposed estimation method by using a real data example in economics. In Section 4 we establish the prediction consistency of our lasso estimator. In Section 5 we establish sparsity oracle inequalities in terms of both the prediction loss and the $l_1$-estimation loss for regression coefficients under a sparsity scenario.

1.1. Notation

We collect the notation that is used in the paper here. For $\{(Y_i, X_i, Q_i) : i = 1, \ldots, n\}$ following model (1.1), let $X_i(\tau)$ denote the $2M \times 1$ vector such that $X_i(\tau) = (X_i', X_i' 1 \{Q_i < \tau\})'$ and let $X(\tau)$ denote the $n \times 2M$ matrix whose $ith$ row is $X_i(\tau)'$. For an $L$-dimensional vector $a$, let $|a|_p$ denote the $l_p$-norm of $a$, and $|J(a)|$ denote the cardinality of $J(a)$, where $J(a) = \{ j \in \{1, \ldots, L\} : a_j \neq 0 \}$. In addition, let $\mathcal{M}(a)$ denote the number of non-zero elements of $a$, i.e. $\mathcal{M}(a) = \sum_{j=1}^L 1 \{a_j \neq 0\} = |J(a)|$. Let $a_j$ denote the vector in $\mathbb{R}^L$ that has the same co-ordinates as $a$ on $J$ and zero co-ordinates on the complement $J^c$ of $J$. For any $n$-dimensional vector $W = (W_1, \ldots, W_n)'$, define the empirical norm as $\|W\| := (n^{-1} \sum_{i=1}^n W_i^2)^{1/2}$. Let the superscript $(j)'$ denote the $j$th element of a vector or the $j$th column of a matrix depending on the context. Finally, 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 $\hat{f}(x, q) := x' \hat{\beta} + x' \hat{\delta} 1 \{q < \hat{\tau}\}$. Then, we define the prediction risk as

$$\|\hat{f} - f_0\| := \left[ \frac{1}{n} \sum_{i=1}^n \left( \hat{f}(X_i, Q_i) - f_0(X_i, Q_i) \right)^2 \right]^{1/2}.$$

2. Lasso estimation

Let $\alpha_0 = (\beta_0', \delta_0')'$. Then, using notation defined above, we can rewrite model (1.1) as
\[ Y_i = X_i (\tau_0) \beta_0 + U_i, \quad i = 1, \ldots, n. \] (2.1)

Let \( y \equiv (Y_1, \ldots, Y_n)' \). For any fixed \( \tau \in \mathbb{T} \), where \( \mathbb{T} \equiv [t_0, t_1] \) is a parameter space for \( \tau_0 \), consider the residual sum of squares
\[
S_n(\alpha, \tau) = n^{-1} \sum_{i=1}^{n} (Y_i - X_i (\beta - X_i \delta) 1 \{ Q_i < \tau \})^2
= \| y - X(\tau) \alpha \|^2_n,
\]
where \( \alpha = (\beta', \delta')' \).

We define the following \( 2M \times 2M \) diagonal matrix:
\[
D(\tau) := \text{diag}\{\| X^{(j)}(\tau) \|_n, j = 1, \ldots, 2M \}.
\]

For each fixed \( \tau \in \mathbb{T} \), define the lasso solution \( \hat{\alpha}(\tau) \) by
\[
\hat{\alpha}(\tau) := \arg \min_{\alpha \in \mathcal{A} \subset \mathbb{R}^{2M}} \{ S_n(\alpha, \tau) + \lambda |D(\tau)\alpha|_1 \},
\] (2.2)
where \( \lambda \) is a tuning parameter that depends on \( n \) and \( \mathcal{A} \) is a parameter space for \( \alpha_0 \).

It is important to note that the scale normalizing factor \( D(\tau) \) depends on \( \tau \) since different values of \( \tau \) generate different dictionaries \( X(\tau) \). To see this more clearly, define
\[
X^{(j)}(\tau) \equiv (X^{(j)}_1, \ldots, X^{(j)}_n)', \quad X^{(j)}(\tau) \equiv (X^{(j)}_1 1 \{ Q_1 < \tau \}, \ldots, X^{(j)}_n 1 \{ Q_n < \tau \}').
\] (2.3)

Then, for each \( \tau \in \mathbb{T} \) and for each \( j = 1, \ldots, M \), we have \( \| X^{(j)}(\tau) \|_n = \| X^{(j)} \|_n \) and \( \| X^{(M+j)}(\tau) \|_n = \| X^{(j)}(\tau) \|_n \). Using this notation, we rewrite the \( l_1 \)-penalty as
\[
\lambda |D(\tau)\alpha|_1 = \lambda \sum_{j=1}^{2M} \| X^{(j)}(\tau) \|_n |\alpha^{(j)}| = \lambda \sum_{j=1}^{M} \{ \| X^{(j)} \|_n |\alpha^{(j)}| + \| X^{(j)}(\tau) \|_n |\alpha^{(M+j)}| \}.
\]

Therefore, for each fixed \( \tau \in \mathbb{T} \), \( \hat{\alpha}(\tau) \) is the weighted lasso that uses a data-dependent \( l_1 \)-penalty to balance covariates adequately.

We now estimate \( \tau_0 \) by
\[
\hat{\tau} := \arg \min_{\tau \in \mathbb{T} \subset \mathbb{R}} [S_n \{ \hat{\alpha}(\tau), \tau \} + \lambda |D(\tau)\hat{\alpha}(\tau)|].
\] (2.4)

In fact, for any finite \( n \), \( \hat{\tau} \) is given by an interval and we simply define the maximum of the interval as our estimator. If we wrote the model by using \( 1 \{ Q_i > \tau \} \), then the convention would be the minimum of the interval being the estimator. Then the estimator of \( \alpha_0 \) is defined as \( \hat{\alpha} := \hat{\alpha}(\hat{\tau}) \). In fact, our proposed estimator of \( (\alpha, \tau) \) can be viewed as the one-step minimizer such that
\[
(\hat{\alpha}, \hat{\tau}) := \arg \min_{\alpha \in \mathcal{A} \subset \mathbb{R}^{2M}, \tau \in \mathbb{T} \subset \mathbb{R}} \{ S_n(\alpha, \tau) + \lambda |D(\tau)\alpha|_1 \}. \] (2.5)

It is worth noting that we penalize \( \beta_0 \) and \( \delta_0 \) in expression (2.5), where \( \delta_0 \) is the change of regression coefficients between two regimes. Model (1.1) can be written as
\[
Y_i = \begin{cases} 
X_i \beta_0 + U_i, & \text{if } Q_i \geq \tau_0, \\
X_i \beta_1 + U_i, & \text{if } Q_i < \tau_0,
\end{cases}
\] (2.6)
where $\beta_1 \equiv \beta_0 + \delta_0$. In view of model (2.6), alternatively, one might penalize $\beta_0$ and $\beta_1$ instead of $\beta_0$ and $\delta_0$. We opted to penalize $\delta_0$ in this paper since the case $\delta_0 = 0$ corresponds to the linear model. If $\delta = 0$, then this case amounts to selecting the linear model.

3. Empirical illustration

In this section, we apply the proposed lasso method to growth regression models in economics. The neoclassical growth model predicts that economic growth rates converge in the long run. This theory has been tested empirically by looking at the negative relationship between long-run growth rate and initial gross domestic product (GDP) given other covariates (see Barro and Sala-i-Martin (1995) and Durlauf et al. (2005) for literature reviews). Although empirical results confirmed the negative relationship between growth rate and initial GDP, there has been some criticism that the results depend heavily on the selection of covariates. Recently, Belloni and Chernozhukov (2011b) showed that lasso estimation can help to select the covariates in the linear growth regression model and that the lasso estimation results reconfirm the negative relationship between long-run growth rate and initial GDP.

We consider the growth regression model with a possible threshold. Durlauf and Johnson (1995) provided the theoretical background of the existence of multiple steady states and estimated the model with two possible threshold variables. They checked the robustness by adding other available covariates to the model, but it is not still free from the criticism of ad hoc variable selection. Our proposed lasso method might be a good alternative in this situation. Furthermore, as we shall show later, our method works well even if there is no threshold effect in the model. Therefore, one might expect more robust results from our approach.

The regression model that we consider has the form

$$gr_i = \beta_0 + \beta_1 lgdp60_i + X_i' \beta_2 + \mathbf{1}\{Q_i < \tau\}(\delta_0 + \delta_1 lgdp60_i + X_i' \delta_2) + \varepsilon_i, \quad \text{(3.1)}$$

where $gr_i$ is the annualized GDP growth rate of country $i$ from 1960 to 1985, $lgdp60_i$ is the log-GDP in 1960 and $Q_i$ is a possible threshold variable for which we use the initial GDP or the adult literacy rate in 1960 following Durlauf and Johnson (1995). Finally, $X_i$ is a vector of additional covariates related to education, market efficiency, political stability, market openness and demographic characteristics. In addition, $X_i$ contains cross-product terms between $lgdp60_i$ and education variables. Table 1 gives a list of all covariates used and a description of each variable. We include as many covariates as possible, which might mitigate the potential omitted variable bias. The data set mostly comes from Barro and Lee (1994), and the additional adult literacy rate is from Durlauf and Johnson (1995). Because of missing observations, we have 80 observations with 46 covariates (including a constant term) when $Q_i$ is the initial GDP ($n = 80$ and $M = 46$), and 70 observations with 47 covariates when $Q_i$ is the literacy rate ($n = 70$ and $M = 47$). It is worthwhile to note that the number of covariates in the threshold models is bigger than the number of observations ($2M > n$ in our notation). Thus, we cannot adopt the standard least squares method to estimate the threshold regression model.

Table 2 summarizes the model selection and estimation results when $Q_i$ is the initial GDP. In the on-line appendix F (see Table 4), we report additional empirical results with $Q_i$ being the literacy rate. To compare different model specifications, we also estimate a linear model, i.e. all $\delta$s are 0s in model (3.1), by standard lasso estimation. In each case, the regularization parameter $\lambda$ is chosen by the ‘leave-one-out’ cross-validation method. For the range $T$ of the threshold parameter, we consider an interval between the 10% and 90% sample quantiles for each threshold variable.

Main empirical findings are as follows. First, the marginal effect of $lgdp60_i$, which is given by
Table 1. List of variables

| Variable name | Description |
|---------------|-------------|
| **Dependent variable** | Annualized GDP growth rate in the period 1960–1985 |
| gr | Annualized GDP growth rate in the period 1960–1985 |
| **Threshold variables** | |
| gdp60 | Real GDP per capita in 1960 (1985 price) |
| lr | Adult literacy rate in 1960 |
| **Covariates** | |
| lgdp60 | Log-GDP per capita in 1960 (1985 price) |
| lr | Adult literacy rate in 1960 (only included when $Q = lr$) |
| $\ell_s$ | log(Investment/output) annualized over 1960–1985; a proxy for log(physical savings rate) |
| $\ell_{gr pop}$ | log(population growth rate) annualized over 1960–1985 |
| pyrm60 | log(average years of primary schooling) in the male population in 1960 |
| pyrf60 | log(average years of primary schooling) in the female population in 1960 |
| syrm60 | log(average years of secondary schooling) in the male population in 1960 |
| syrf60 | log(average years of secondary schooling) in the female population in 1960 |
| hyrm60 | log(average years of higher schooling) in the male population in 1960 |
| hyrf60 | log(average years of higher schooling) in the female population in 1960 |
| nom60 | Percentage of no schooling in the male population in 1960 |
| nof60 | Percentage of no schooling in the female population in 1960 |
| prim60 | Percentage of primary schooling attained in the male population in 1960 |
| prif60 | Percentage of primary schooling attained in the female population in 1960 |
| pricm60 | Percentage of primary schooling complete in the male population in 1960 |
| pricf60 | Percentage of primary schooling complete in the female population in 1960 |
| secm60 | Percentage of secondary schooling attained in the male population in 1960 |
| secf60 | Percentage of secondary schooling attained in the female population in 1960 |
| seccm60 | Percentage of secondary schooling complete in the male population in 1960 |
| seccf60 | Percentage of secondary schooling complete in the female population in 1960 |
| llife | log(life expectancy at age 0) averaged over 1960–1985 |
| lfert | log(fertility rate) averaged over 1960–1985 |
| edu/gdp | Government expenditure on education per GDP averaged over 1960–1985 |
| gcon/gdp | Government consumption expenditure net of defence and education per GDP averaged over 1960–1985 |
| revol | Number of revolutions per year over 1960–1984 |
| revcoup | Number of revolutions and coups per year over 1960–1984 |
| wardun | Dummy for countries that participated in at least one external war over 1960–1984 |
| wartime | Fraction of time over 1960–1985 involved in external war |
| lbmp | log(1 + black market premium averaged over 1960–1985) |
| tot | Term-of-trade shock |
| lgdp60 $\times$ ‘educ’ | Product of two covariates (interaction of lgdp60 and education variables from pyrm60 to seccf60); total 16 variables |

$$
\left. \frac{\partial gr_i}{\partial lgdp60_i} \right|_{\text{educ}_i} = \beta_1 + \text{educ}_i^9 \tilde{\beta}_2 + 1 \{ Q_i < \gamma \} (\delta_1 + \text{educ}_i^9 \tilde{\delta}_2),
$$

where educ$_i$ is a vector of education variables and $\tilde{\beta}_2$ and $\tilde{\delta}_2$ are subvectors of $\beta_2$ and $\delta_2$ corresponding to educ$_i$, is estimated to be negative for all the observed values of educ$_i$. This confirms the theory of the neoclassical growth model. Second, some non-zero coefficients of interaction terms between lgdp60 and various education variables show the existence of threshold effects in both threshold model specifications. This result implies that the growth convergence rates can vary according to different levels of the initial GDP or the adult literacy rate in 1960. Specifically,
Table 2. Model selection and estimation results with $Q = \text{gdp60}$

| Variable     | Value for the linear model | Values for the threshold model, $\hat{\tau} = 2898$ |
|--------------|----------------------------|-----------------------------------------------|
|              | $\hat{\beta}$ | $\hat{\delta}$ |                              |
| Constant     | -0.0923         | -0.0811         | —                              |
| lgdp60       | -0.0153         | -0.0120         | —                              |
| lgk          | 0.0333          | 0.0038          | —                              |
| lgr_pop      | 0.0018          | —                | —                              |
| pyrf60       | 0.0027          | —                | —                              |
| syrm60       | 0.0157          | —                | —                              |
| hyrm60       | 0.0122          | 0.0130          | —                              |
| hyrf60       | -0.0389         | —                | -0.0807                       |
| nom60        | —               | —                | 2.64 × 10^{-5}                |
| prim60       | -0.0004         | -0.0001         | —                              |
| pricm60      | 0.0006          | -1.73 × 10^{-4} | -0.35 × 10^{-4}               |
| pricf60      | -0.0006         | —                | —                              |
| secc60       | 0.0005          | —                | —                              |
| seccm60      | 0.0010          | —                | 0.0014                        |
| llife        | 0.0697          | 0.0523          | —                              |
| lfert        | -0.0136         | -0.0047         | —                              |
| edu/gdp      | -0.0189         | —                | —                              |
| gcon/gdp     | -0.0671         | -0.0542         | —                              |
| revol        | -0.0588         | —                | —                              |
| revcoup      | 0.0433          | —                | —                              |
| wardum       | -0.0043         | —                | -0.0022                       |
| wartime      | -0.0019         | -0.0143         | -0.0023                       |
| lbmp         | -0.0185         | -0.0174         | -0.0015                       |
| tot          | 0.0971          | —                | 0.0974                        |
| lgdp60 × pyrf60 | —               | -3.81 × 10^{-6} | —                              |
| lgdp60 × syrm60 | —               | —                | 0.0002                        |
| lgdp60 × hyrm60 | —               | —                | 0.0050                        |
| lgdp60 × hyrf60 | —               | -0.0003         | —                              |
| lgdp60 × nom60 | —               | —                | 8.26 × 10^{-6}               |
| lgdp60 × prim60 | -6.02 × 10^{-7} | —                | —                              |
| lgdp60 × prif60 | -3.47 × 10^{-6} | —                | -8.11 × 10^{-6}              |
| lgdp60 × pricf60 | -8.46 × 10^{-6} | —                | —                              |
| lgdp60 × seccm60 | -0.0001         | —                | —                              |
| lgdp60 × seccf60 | -0.0002         | -2.87 × 10^{-6} | —                              |

| $\lambda$ | 0.0004 | 0.0034 |
| $M(\hat{\delta})$ | 28 | 26 |
| Number of covariates | 46 | 92 |
| Number of observations | 80 | 80 |

†The regularization parameter $\lambda$ is chosen by the ‘leave-one-out’ cross-validation method. $M(\hat{\delta})$ denotes the number of covariates to be selected by the lasso estimator and a dash indicates that the regressor is not selected. Recall that $\hat{\beta}$ is the coefficient when $Q \geq \hat{\gamma}$ and that $\hat{\delta}$ is the change of the coefficient value when $Q < \hat{\gamma}$.

in both threshold models, we have $\hat{\delta}_1 = 0$, but some $\hat{\delta}_2$s are not 0. Thus, conditionally on other covariates, there are different technological diffusion effects according to the threshold point. For example, a developing country (lower $Q$) with a higher education level will converge faster perhaps by absorbing advanced technology more easily and more quickly. Finally, the lasso with the threshold model specification selects a more parsimonious model than that with the linear specification even though the former doubles the number of potential covariates.
4. Prediction consistency of lasso estimator

In this section, we consider the prediction consistency of the lasso estimator. We make the following assumptions.

Assumption 1.

(a) For the parameter space $\mathcal{A}$ for $\alpha_0$, any $\alpha \equiv (\alpha_1, \ldots, \alpha_{2M}) \in \mathcal{A} \subset \mathbb{R}^{2M}$, including $\alpha_0$, satisfies $\max_{j=1,\ldots,2M} |\alpha_j| \leq C_1$ for some constant $C_1 > 0$. In addition, $\tau_0 \in \mathbb{T} \equiv [t_0, t_1]$ that satisfies $\min_{i=1,\ldots,n} Q_i < t_0 < t_1 < \max_{i=1,\ldots,n} Q_i$.

(b) There are universal constants $C_2 > 0$ and $C_3 > 0$ such that $\|X^{(j)}(\tau)\|_n \leq C_2$ uniformly in $j$ and $\tau \in \mathbb{T}$, and $\|X^{(j)}(t_0)\|_n \geq C_3$ uniformly in $j$, where $j = 1, \ldots, 2M$.

(c) There is no $i \neq j$ such that $Q_i = Q_j$.

Assumption 1(a) imposes the boundedness for each component of the parameter vector. The first part of assumption 1(a) which implies that $|\alpha|_1 \leq 2C_1 M$ for any $\alpha \in \mathcal{A}$, seems to be weak, since the sparsity assumption implies that $|\alpha_0|_1$ is much smaller than $C_1 M$. Furthermore, in the literature on change point and threshold models, it is common to assume that the parameter space is compact. For example, see Seijo and Sen (2011a, b).

The lasso estimator in expression (2.5) can be computed without knowing the value of $C_1$, but $\mathbb{T} \equiv [t_0, t_1]$ must be specified. In practice, researchers tend to choose some strict subset of the range of observed values of the threshold variable. Assumption 1(b) imposes that each covariate is of the same magnitude uniformly over $\tau$. In view of the assumption that $\min_{i=1,\ldots,n} Q_i < t_0$, it is not stringent to assume that $\|X^{(j)}(t_0)\|_n$ is bounded away from zero.

Assumption 1(c) imposes that there is no tie among $Q_i$s. This is a convenient assumption such that we can always transform general $Q_i$ to $Q_i = i/n$ without loss of generality. This holds with probability 1 for the random-design case if $Q_i$ is continuously distributed.

Define

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

where $X^{(j)}$ and $X^{(j)}(\tau)$ are defined in expression (2.3). Assumption 1(b) implies that $r_n$ is bounded away from zero. In particular, we have that $1 \geq r_n \geq C_3/C_2 > 0$.

Recall that

$$\|\hat{f} - f_0\|_n := \left[ \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{f}(X_i, Q_i) - f_0(X_i, Q_i) \right\}^2 \right]^{1/2}, \quad (4.1)$$

where $\hat{f}(x, q) := x'\hat{\beta} + x'\hat{\delta} 1\{q < \hat{\tau}\}$ and $f_0(x, q) := x'\beta_0 + x'\delta_0 1\{q < \tau_0\}$. To establish theoretical results in the paper (in particular, oracle inequalities in Section 5), let $(\hat{\alpha}, \hat{\tau})$ be the lasso estimator defined by expression (2.5) with

$$\lambda = A\sigma \left( \frac{\log(3M)}{nr_n} \right)^{1/2}, \quad (4.2)$$

for a constant $A > 2\sqrt{2}/\mu$, where $\mu \in (0, 1)$ is a fixed constant. We now present the first theoretical result of this paper.

Theorem 1 (consistency of the lasso). Let assumption 1 hold. Let $\mu$ be a constant such that $0 < \mu < 1$, and let $(\hat{\alpha}, \hat{\tau})$ be the lasso estimator defined by expression (2.5) with $\lambda$ given by equation (4.2). Then, with probability at least $1 - (3M)^{1-A^2\mu^2/8}$, we have

$$\|\hat{f} - f_0\|_n \leq K_1 \{\lambda M(\alpha_0)\}^{1/2},$$
where \( K_1 \equiv \{ 2C_1 C_2 (3 + \mu) \}^{1/2} > 0 \).

The non-asymptotic upper bound on the prediction risk in theorem 1 can be translated easily into an asymptotic convergence. Theorem 1 implies the consistency of the lasso, provided that \( n \to \infty, M \to \infty \) and \( \lambda \mathcal{M}(\alpha_0) \to 0 \). Recall that \( \mathcal{M}(\alpha_0) \) represents the sparsity of model (2.1). In view of equation (4.2), the condition \( \lambda \mathcal{M}(\alpha_0) \to 0 \) requires that \( \mathcal{M}(\alpha_0) = o([nr_\delta / \log(3M)]^{1/2}) \). This implies that \( \mathcal{M}(\alpha_0) \) can increase with \( n \).

Remark 1. Note that the prediction error increases as \( A \) or \( \mu \) increases; however, the probability of correct recovery increases if \( A \) or \( \mu \) increases. Therefore, there is a trade-off between the prediction error and the probability of correct recovery.

5. Oracle inequalities

In this section, we establish finite sample sparsity oracle inequalities in terms of both the prediction loss and the \( l_1 \)-estimation loss for unknown parameters. First of all, we make the following assumption.

Assumption 2 (uniform restricted eigenvalue (URE) \( (s, c_0, S) \)). For some integer \( s \) such that \( 1 \leq s \leq 2M \), a positive number \( c_0 \) and some set \( S \subset \mathbb{R} \), the following condition holds:

\[
\kappa(s, c_0, S) := \min_{\tau \in S} \min_{\{ J_0 \in \{ 1, \ldots, 2M \} : |J_0| \leq s \}} \min_{|J_0| \leq s, |\gamma J_0| \leq c_0} \frac{|X(\tau) \gamma|_2}{\sqrt{n} |\gamma J_0|_2} > 0.
\]

If \( \tau_0 \) were known, then assumption 2 is just a restatement of the restricted eigenvalue assumption of Bickel et al. (2009) with \( S = \{ \tau_0 \} \). Bickel et al. (2009) provided sufficient conditions for the restricted eigenvalue condition. In addition, van de Geer and Bühlmann (2009) showed the relationships between the restricted eigenvalue condition and other conditions on the design matrix, and Raskutti et al. (2010) proved that restricted eigenvalue conditions hold with high probability for a large class of correlated Gaussian design matrices.

If \( \tau_0 \) is unknown as in our set-up, it seems necessary to assume that the restricted eigenvalue condition holds uniformly over \( \tau \). We consider separately two cases depending on whether \( \delta_0 = 0 \) or not. On the one hand, if \( \delta_0 = 0 \) so that \( \tau_0 \) is not identifiable, then we need to assume that the URE condition holds uniformly on the whole parameter space, \( \mathbb{T} \). On the other hand, if \( \delta_0 \neq 0 \) so that \( \tau_0 \) is identifiable, then it suffices to impose that the URE condition holds uniformly on a neighbourhood of \( \tau_0 \). In the on-line appendix A, we provide two types of sufficient conditions for assumption 2. One type is based on modifications of assumption 2 of Bickel et al. (2009) and the other type is in the same spirit as van de Geer and Bühlmann (2009), section 10.1. Using the second type of results, we verify primitive sufficient conditions for the URE condition in the context of our simulation designs. See the on-line appendix A for details.

The URE condition is useful for us to improve the result in theorem 1. Recall that, in theorem 1, the prediction risk is bounded by a factor of \( \{ \lambda \mathcal{M}(\alpha_0) \}^{1/2} \). This bound is too large to give us an oracle inequality. We shall show below that we can establish non-asymptotic oracle inequalities for the prediction risk as well as the \( l_1 \)-estimation loss, thanks to the URE condition.

The strength of the proposed lasso method is that it is not necessary to know or pretest whether \( \delta_0 = 0 \) or not. It is worth noting that we do not have to know whether there is a threshold in the model to establish oracle inequalities for the prediction risk and the \( l_1 \)-estimation loss for \( \alpha_0 \), although we divide our theoretical results into two cases below. This implies that we can
make prediction and estimate $\alpha_0$ precisely without knowing the presence of a threshold effect or without pretesting for it.

### 5.1. Case I: no threshold

We first consider the case that $\delta_0 = 0$. In other words, we estimate a threshold model via the lasso method, but the true model is simply a linear model $Y_i = X_i' \beta_0 + U_i$. This is an important case to consider in applications, because one may not be sure not only about covariates selection but also about the existence of the threshold in the model.

Let $\phi_{\text{max}}$ denote the supremum (over $\tau \in T$) of the largest eigenvalue of $X(\tau)'X(\tau)/n$. Then, by definition, the largest eigenvalue of $X(\tau)'X(\tau)/n$ is bounded uniformly in $\tau \in T$ by $\phi_{\text{max}}$. The following theorem gives oracle inequalities for the first case.

**Theorem 2.** Suppose that $\delta_0 = 0$. Let assumptions 1 and 2 hold with $\kappa = \kappa\{s, (1 + \mu)/(1 - \mu), T\}$ for $0 < \mu < 1$, and $M(\alpha_0) \leq s \leq M$. Let $\hat{\alpha}, \hat{\tau}$ be the lasso estimator defined by expression (2.5) with $\lambda$ given by expression (4.2). Then, with probability at least $1 - 3M/n$, we have

\[
\|f_{\hat{\alpha}, \hat{\tau}} - f_0\|_n \leq K_2 \frac{\sigma}{\kappa^2} \left( \frac{\log(3M)}{nr_n} s \right)^{1/2},
\]

\[
|\hat{\alpha} - \alpha_0|_1 \leq K_2 \frac{\sigma}{\kappa^2} \left( \frac{\log(3M)}{nr_n} s \right)^{1/2},
\]

\[
M(\hat{\alpha}) \leq K_2 \frac{\phi_{\text{max}}}{\kappa^2} s
\]

for some universal constant $K_2 > 0$.

To appreciate the usefulness of the inequalities derived above, it is worth comparing inequalities in theorem 2 with those in theorem 7.2 of Bickel et al. (2009). The latter corresponds to the case that $\delta_0 = 0$ is known *a priori* and $\lambda = 2A\sigma \log(M/n)^{1/2}$ in our notation. If we compare theorem 2 with theorem 7.2 of Bickel et al. (2009), we can see that the lasso estimator in model (2.5) gives qualitatively the same oracle inequalities as the lasso estimator in the linear model, even though our model is much more overparameterized in that $\delta$ and $\tau$ are added to $\beta$ as parameters to estimate.

Also, as in Bickel et al. (2009), there is no requirement on $\alpha_0$ such that the minimum value of non-zero components of $\alpha_0$ is bounded away from zero. In other words, there is no need to assume the minimum strength of the signals. Furthermore, $\alpha_0$ is well estimated here even if $\tau_0$ is not identifiable at all. Finally, note that the value of the constant $K_2$ is given in the proof of theorem 2 and that theorem 2 can be translated easily into asymptotic oracle results as well, since both $\kappa$ and $r_n$ are bounded away from zero by the URE condition and assumption 1 respectively.

### 5.2. Case II: fixed threshold

This subsection explores the case where the threshold effect is well identified and discontinuous. We begin with the following additional assumptions to reflect this.

**Assumption 3** (identifiability under sparsity and discontinuity of regression). For a given $s \geq M(\alpha_0)$, and for any $\eta$ and $\tau$ such that $|\tau - \tau_0| > \eta \geq \min_i |Q_i - \tau_0|$ and $\alpha \in \{\alpha : M(\alpha) \leq s\}$, there is a constant $c > 0$ such that

\[
\|f(\alpha, \tau) - f_0\|_n^2 > c\eta.
\]
Assumption 3 implies, among other things, that for some \( s \geq \mathcal{M}(\alpha_0) \), and for any \( \alpha \in \{ \alpha : \mathcal{M}(\alpha) \leq s \} \) and \( \tau \) such that \( (\alpha, \tau) \neq (\alpha_0, \tau_0) \),
\[
\| f(\alpha, \tau) - f_0 \|_n \neq 0. \tag{5.1}
\]
This condition can be regarded as identifiability of \( \tau_0 \). If \( \tau_0 \) were known, then a sufficient condition for the identifiability under the sparsity would be that \( \text{URE}(s, c_0, \{ \tau_0 \}) \) holds for some \( c_0 > 1 \). Thus, the main point in result (5.1) is that there is no sparse representation that is equivalent to \( f_0 \) when the sample is split by \( \tau \neq \tau_0 \). In fact, assumption 3 is stronger than just the identifiability of \( \tau_0 \) as it specifies the rate of deviation in \( f \) as \( \tau \) moves away from \( \tau_0 \), which in turn dictates the bound for the estimation error of \( \hat{\tau} \). We provide further discussions on assumption 3 in the on-line appendix B.

**Remark 2.** The restriction \( \eta \geq \min_i |Q_i - \tau_0| \) in assumption 3 is necessary since we consider the fixed design for both \( X_i \) and \( Q_i \). Throughout this section, we implicitly assume that the sample size \( n \) is sufficiently large such that \( \min_{i \neq j} |Q_i - Q_j| \) is very small, implying that the restriction \( \eta \geq \min_{i \neq j} |Q_i - Q_j| \) never binds in any of the inequalities below. This is typically true for the random-design case if \( Q_i \) is continuously distributed.

**Assumption 4** (smoothness of design). For any \( \eta > 0 \), there is a constant \( C < \infty \) such that
\[
\sup_j \sup_{|\tau - \tau_0| < \eta} \frac{1}{n} \sum_{i=1}^{n} |X_i^{(j)}|^2 I(Q_i < \tau_0) - I(Q_i < \tau) \leq C\eta.
\]

Assumption 4 has been assumed in the classical set-up with a fixed number of stochastic regressors to exclude cases like \( Q_i \) has a point mass at \( \tau_0 \) or \( \mathbb{E}(X_i | Q_i = \tau_0) \) is unbounded. In our set-up, assumption 4 amounts to a deterministic version of some smoothness assumption for the distribution of the threshold variable \( Q_i \). When \( (X_i, Q_i) \) is a random vector, it is satisfied under the standard assumption that \( Q_i \) is continuously distributed and \( \mathbb{E}(|X_i^{(j)}|^2 | Q_i = \tau) \) is continuous and bounded in a neighbourhood of \( \tau_0 \) for each \( j \).

To simplify the notation, in the following theorem, we assume without loss of generality that \( Q_i = i/n \). Then \( \mathbb{T} = [t_0, t_1] \subset (0, 1) \). In addition, let \( \eta_0 = \max[n^{-1}, K_1 \sqrt{\lambda \mathcal{M}(\alpha_0)}] \) where \( K_1 \) is the same constant in theorem 1.

**Assumption 5** (well-defined second moments). For any \( \eta \) such that \( 1/n \leq \eta \leq \eta_0 \), \( h_n^2(\eta) \) is bounded, where
\[
h_n^2(\eta) := \frac{1}{2n\eta} \max_{i \in \{1, [n(\tau_0 - \eta)]\}} (X_i^{(i)}\delta_0)^2
\]
and \([\cdot]\) denotes an integer part of any real number.

Assumption 5 assumes that \( h_n^2(\eta) \) is well defined for any \( \eta \) such that \( 1/n \leq \eta \leq \eta_0 \). Assumption 5 amounts to some weak regularity condition on the second moments of the fixed design. Assumption 3 implies that \( \delta_0 \neq 0 \) and that \( h_n^2(\eta) \) is bounded away from zero. Hence, assumptions 3 and 5 imply that \( h_n^2(\eta) \) is bounded and bounded away from zero.

To present the theorem below, it is necessary to make one additional technical assumption (see assumption 6 in the on-line appendix E). We opted not to show assumption 6 here, since we believe that this is just a sufficient condition that does not add much to our understanding of the main result. However, we would like to point out that assumption 6 can hold for all sufficiently large \( n \), provided that \( s\lambda|\delta_0|_1 \to 0 \), as \( n \to 0 \). See remark 4 in the on-line appendix E for details.

We now give the main result of this section.
Theorem 3. Suppose that assumptions 1 and 2 hold with $\mathcal{S} = \{|\tau - \tau_0| \leq \eta_0\}$, $\kappa = \kappa\{s, (2 + \mu)/(1 - \mu)\}, \mathcal{S}$ for $0 < \mu < 1$, and $\mathcal{M}(\alpha_0) \leq s \leq M$. Furthermore, assumptions 3, 4 and 5 hold and let $n$ be sufficiently large that assumption 6 in the on-line appendix E holds. Let $(\hat{\alpha}, \hat{\tau})$ be the lasso estimator defined by expression (2.5) with $\lambda$ given by expression (4.2). Then, with probability at least $1 - (3M)^{1-\alpha^2/8} - C_4(3M)^{-C_5/n}$ for some positive constants $C_4$ and $C_5$, we have

$$\|\hat{f} - f_0\|_n \leq K_3 \frac{\sigma}{\kappa} \left\{\log(3M)/nr_n\right\}^{1/2},$$

$$|\hat{\alpha} - \alpha_0| \leq K_3 \frac{\sigma}{\kappa^2} \left\{\log(3M)/nr_n\right\}^{1/2} s,$$

$$|\hat{\tau} - \tau_0| \leq K_3 \frac{\sigma^2}{\kappa^2} \log(3M)/nr_n,$$

$$\mathcal{M}(\hat{\alpha}) \leq K_3 \frac{\phi_{\max}}{\kappa^2} s$$

for some universal constant $K_3 > 0$.

Theorem 3 gives the same inequalities (up to constants) as those in theorem 2 for the prediction risk as well as the $l_1$-estimation loss for $\alpha_0$. It is important to note that $|\hat{\tau} - \tau_0|$ is bounded by a constant times $s \log(3M)/(nr_n)$, whereas $|\hat{\alpha} - \alpha_0|$ is bounded by a constant times $s \{\log(3M)/(nr_n)\}^{1/2}$. This can be viewed as a non-asymptotic version of the superconsistency of $\hat{\tau}$ to $\tau_0$. As noted at the end of Section 5.1, since both $\kappa$ and $r_n$ are bounded away from zero by the URE condition and assumption 1 respectively, theorem 3 implies asymptotic rate results immediately. The values of constants $C_4$, $C_5$ and $K_3$ are given in the proof of theorem 3.

The main contribution of this section is that we have extended the well-known superconsistency result of $\hat{\tau}$ when $M < n$ (see, for example, Chan (1993) and Seijo and Sen (2011a, b)) to the high dimensional set-up ($M \gg n$). In both cases, the main reason that we achieve the superconsistency for the threshold parameter is that the least squares objective function behaves locally linearly around the true threshold parameter value rather than locally quadratically, as in regular estimation problems. An interesting remaining research question is to investigate whether it would be possible to obtain the superconsistency result of $\hat{\tau}$ under a weaker condition, perhaps without a restricted eigenvalue condition.

6. Monte Carlo experiments

In this section we conduct some simulation studies and check the properties of the lasso estimator proposed. The baseline model is model (1.1), where $X_i$ is an $M$-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 to $\tau_0 = 0.3, 0.4, 0.5$ depending on the simulation design, and the coefficients are set to $\beta_0 = (1, 0, 1, 0, \ldots, 0)$, and $\delta_0 = c(0, -1, 1, 0, \ldots, 0)$ where $c = 0$ or $c = 1$. Note that there is no threshold effect when $c = 0$. The number of observations is set to $n = 200$. Finally, the dimension of $X_i$ in each design is set to $M = 50, 100, 200, 400$, so that the total numbers of regressors are 100, 200, 400 and 800 respectively. The range of $\tau$ is $\mathcal{T} = [0.15, 0.85]$.

We can estimate the parameters by the standard lasso–least angle regression algorithm of Efron et al. (2004) without much modification. Given a regularization parameter value $\lambda$, we estimate the model for each grid point of $\tau$ that spans over 71 equispaced points on $\mathcal{T}$. This
Table 3. Simulation results with $M = 50$†

| Jump scale: $c = 1$ | Estimation method | Constant for $\lambda$ | Prediction error | $\mathbb{E}[\mathcal{M} (\hat{\alpha})]$ | $\mathbb{E} |\hat{\alpha} - \alpha_0|$ | $\mathbb{E} |\hat{\tau} - \tau_0|$ |
|---------------------|--------------------|------------------------|------------------|----------------------------------------|-------------------------------|-------------------------------|
| $\tau_0 = 0.5$      | Least squares      | None                   | 0.285 0.276 0.074| 100.00 7.066 0.008                    |                               |                               |
|                     | Lasso              | $A = 2.8$              | 0.041 0.030 0.035| 12.94 0.466 0.010                    |                               |                               |
|                     |                    | $A = 3.2$              | 0.048 0.033 0.049| 10.14 0.438 0.013                    |                               |                               |
|                     |                    | $A = 3.6$              | 0.067 0.037 0.086| 8.44 0.457 0.024                     |                               |                               |
|                     |                    | $A = 4.0$              | 0.095 0.050 0.120| 7.34 0.508 0.040                     |                               |                               |
|                     | Oracle 1           | None                   | 0.013 0.006 0.019| 4.00 0.164 0.004                     |                               |                               |
|                     | Oracle 2           | None                   | 0.005 0.004 0.004| 4.00 0.163 0.000                     |                               |                               |
| $\tau_0 = 0.4$      | Least squares      | None                   | 0.317 0.304 0.095| 100.00 7.011 0.008                    |                               |                               |
|                     | Lasso              | $A = 2.8$              | 0.052 0.034 0.063| 13.15 0.509 0.016                    |                               |                               |
|                     |                    | $A = 3.2$              | 0.063 0.037 0.083| 10.42 0.489 0.023                    |                               |                               |
|                     |                    | $A = 3.6$              | 0.090 0.045 0.121| 8.70 0.535 0.042                     |                               |                               |
|                     |                    | $A = 4.0$              | 0.133 0.061 0.162| 7.68 0.634 0.078                     |                               |                               |
|                     | Oracle 1           | None                   | 0.014 0.006 0.022| 4.00 0.163 0.004                     |                               |                               |
|                     | Oracle 2           | None                   | 0.005 0.004 0.004| 4.00 0.163 0.000                     |                               |                               |
| $\tau_0 = 0.3$      | Least squares      | None                   | 2.559 0.511 16.292| 100.00 12.172 0.012                   |                               |                               |
|                     | Lasso              | $A = 2.8$              | 0.062 0.035 0.091| 13.45 0.602 0.030                    |                               |                               |
|                     |                    | $A = 3.2$              | 0.089 0.041 0.125| 10.85 0.633 0.056                    |                               |                               |
|                     |                    | $A = 3.6$              | 0.127 0.054 0.159| 9.33 0.743 0.099                     |                               |                               |
|                     |                    | $A = 4.0$              | 0.185 0.082 0.185| 8.43 0.919 0.168                     |                               |                               |
|                     | Oracle 1           | None                   | 0.012 0.006 0.017| 4.00 0.177 0.004                     |                               |                               |
|                     | Oracle 2           | None                   | 0.005 0.004 0.004| 4.00 0.176 0.000                     |                               |                               |

| Jump scale: $c = 0$ | Estimation method | Constant for $\lambda$ | Prediction error | $\mathbb{E}[\mathcal{M} (\hat{\alpha})]$ | $\mathbb{E} |\hat{\alpha} - \alpha_0|$ | $\mathbb{E} |\hat{\tau} - \tau_0|$ |
|---------------------|--------------------|------------------------|------------------|----------------------------------------|-------------------------------|-------------------------------|
| $\tau_0 = 0.5$      | Least squares      | None                   | 6.332 0.460 41.301| 100.00 20.936 0.000                   |                               |                               |
|                     | Lasso              | $A = 2.8$              | 0.013 0.011 0.007| 9.30 0.266 0.066                     |                               |                               |
|                     |                    | $A = 3.2$              | 0.014 0.012 0.008| 6.71 0.227 0.002                     |                               |                               |
|                     |                    | $A = 3.6$              | 0.015 0.014 0.009| 4.95 0.211 0.003                     |                               |                               |
|                     |                    | $A = 4.0$              | 0.017 0.016 0.010| 3.76 0.204 0.000                     |                               |                               |
|                     | Oracle 1 and       | None                   | 0.002 0.002 0.003| 2.00 0.054 0.000                     |                               |                               |
| Oracle 2            |                    |                        |                  |                                        |                               |                               |

†$M$ denotes the column size of $X_i$ and $\tau$ denotes the threshold parameter. Oracle 1 and oracle 2 are estimated by least squares when sparsity is known and when sparsity and $\tau_0$ are known respectively. All simulations are based on 400 replications of a sample with 200 observations.

‡Not applicable.

procedure can be conducted by using the standard linear lasso. Next, we plug in the estimated parameter $\hat{\alpha} (\tau) := (\tilde{\beta} (\tau), \tilde{\delta} (\tau) )$ for each $\tau$ into the objective function and choose $\hat{\tau}$ by expression (4.2). Finally, $\hat{\alpha}$ is estimated by $\hat{\alpha} (\hat{\tau})$. The regularization parameter $\lambda$ is chosen by expression (4.2) where $\sigma = 0.5$ is assumed to be known. For the constant $A$, we use four different values: $A = 2.8, 3.2, 3.6, 4.0$.

Table 3 and Figs 1 and 2 summarize these simulation results. To compare the performance of the lasso estimator, we also report the estimation results of the least squares estimation (‘least squares’) available only when $M = 50$ and two oracle models (oracle 1 and oracle 2). Oracle 1 assumes that the regressors with non-zero coefficients are known. In addition to that, oracle 2 assumes that the true threshold parameter $\tau_0$ is known. Thus, when $c \neq 0$, oracle 1 estimates $(\beta^{(1)}, \beta^{(3)}, \delta^{(2)}, \delta^{(3)})$ and $\tau$ by using least squares estimation whereas oracle 2 estimates only $(\beta^{(1)}, \beta^{(3)}, \delta^{(2)}, \delta^{(3)})$. When $c = 0$, both oracle 1 and oracle 2 estimate only $(\beta^{(1)}, \beta^{(3)})$. All results are based on 400 replications of each sample.
The reported mean-squared prediction error PE for each sample is calculated numerically as follows. For each sample \( s \), we have the estimates \( \hat{\beta}_s, \hat{\delta}_s \) and \( \hat{\tau}_s \). Given these estimates, we generate new data \( \{Y_j, X_j, Q_j\} \) of 400 observations and calculate the prediction error as

\[
\hat{PE}_s = \frac{1}{400} \sum_{j=1}^{400} \{f_0(x_j, q_j) - \hat{f}(x_j, q_j)\}^2.
\]  

The mean, median and standard deviation of the prediction error are calculated from the 400 replications, \( \{\hat{PE}_s\}_{s=1}^{400} \). We also report the mean of \( \mathcal{M}(\hat{\alpha}) \) and \( l_1 \)-errors for \( \alpha \) and \( \tau \). Table 3 reports the simulation results for \( M = 50 \). For simulation designs with \( M > 50 \), the least squares estimator is not available, and we summarize the same statistics only for the lasso estimator in Figs 1 and 2.

When \( M = 50 \), across all designs, the lasso estimator proposed performs better than the least squares estimator in terms of mean and median prediction errors, the mean of \( \mathcal{M}(\hat{\alpha}) \) and the
The performance of the lasso estimator becomes much better when there is no threshold effect, i.e. $c = 0$. This result confirms the robustness of the lasso estimator for whether or not there is a threshold effect. However, the least squares estimator performs better than the lasso estimator in terms of estimation of $\tau_0$ when $c = 1$, although the difference here is much smaller than the differences in prediction errors and the $l_1$-error for $\alpha$.

From Figs 1 and 2, we can reconfirm the robustness of the lasso estimator when $M = 100, 200, 400$. As predicted by the theory that was developed in previous sections, the prediction error and $l_1$-errors for $\alpha$ and $\tau$ increase slowly as $M$ increases. The graphs also show that the results are quite uniform across different regularization parameter values except $A = 4.0$.

In the on-line appendix F, we report additional simulation results, while allowing correlation between covariates. Specifically, the $M$-dimensional vector $X_i$ is generated from a multivariate normal $N(0, \Sigma)$ distribution with $\Sigma_{i,j} = \rho|i-j|$, where $(\Sigma)_{i,j}$ denotes the $(i,j)$ element of the $M \times M$ covariance matrix $\Sigma$ and $\rho = 0.3$. All other random variables are the same as above.

We obtained very similar results to those for the previous cases: the lasso outperforms the least squares estimator, and the prediction error, the mean of $\mathcal{M}(\hat{\alpha})$ and $l_1$-errors increase...
very slowly as $M$ increases. See further details in the on-line appendix F, which also reports satisfactory simulation results regarding frequencies of selecting true parameters when both $\rho = 0$ and $\rho = 0.3$.

In sum, the simulation results confirm the theoretical results that were developed earlier and show that the lasso estimator proposed will be useful for the high dimensional threshold regression model.

7. Conclusions

We have considered a high dimensional regression model with a possible change point due to a covariate threshold and have developed the lasso method. We have derived non-asymptotic oracle inequalities and have illustrated the usefulness of our proposed estimation method via simulations and a real data application.

We conclude this paper by providing some areas of future research. First, it would be interesting to extend other penalized estimators (e.g. the adaptive lasso of Zou (2006) and the smoothly clipped absolute deviation penalty of Fan and Li (2001)) to our set-up and to see whether we would be able to improve the performance of our estimation method. Second, an extension to multiple change points is also an important research topic. There has been some advance in this direction, especially regarding key issues like computational cost and the determination of the number of change points (see, for example, Harchaoui and Lévy-Leduc (2010) and Frick et al. (2014)). However, they are confined to a single regressor case, and the extension to a large number of regressors would be highly interesting. Finally, it would be also an interesting research topic to investigate the minimax lower bounds of the estimator proposed and its prediction risk like Raskutti et al. (2011, 2012) did in high dimensional linear regression set-ups.

Acknowledgements

We thank Marine Carrasco, Yuan Liao, Ya’acov Ritov, two referees and seminar participants at various places for their helpful comments. This work was supported by a National Research Foundation of Korea grant funded by the Korean Government (NRF-2012S1A5A8023573), the Institute of Economic Research of Seoul National University, by the European Research Council (ERC-2009-StG-240910-ROMETA) and by the Social Sciences and Humanities Research Council of Canada. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (www.sharcnet.ca) and Compute/Calcul Canada.

References

Barro, R. and Lee, J. (1994) Data set for a panel of 139 countries. Report. National Bureau of Economic Research, Cambridge. (Available from http://admin.nber.org/pub/barro.lee/.)
Barro, R. and Sala-i-Martin, X. (1995) Economic Growth. New York: McGraw-Hill.
Belloni, A. and Chernozhukov, V. (2011a) $l_1$-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.*, **39**, 82–130.
Belloni, A. and Chernozhukov, V. (2011b) High dimensional sparse econometric models: an introduction. In *Inverse Problems and High-dimensional Estimation* (eds P. Alquier, E. Gautier and G. Stoltz), pp. 121–156. Berlin: Springer.
Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–1732.
Bradic, J., Fan, J. and Jiang, J. (2012) Regularization for Cox’s proportional hazards model with NP-dimensionality. *Ann. Statist.*, **39**, 3092–3120.
Bradic, J., Fan, J. and Wang, W. (2011) Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *J. R. Statist. Soc. B*, **73**, 325–349.
Wu, Y. (2008) Simultaneous change point analysis and variable selection in a regression problem. *J. Multiv. Anal.*, 99, 2154–2171.

Zhang, N. R. and Siegmund, D. O. (2012) Model selection for high dimensional multi-sequence change-point problems. *Statist. Sin.*, 22, 1507–1538.

Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.*, 101, 1418–1429.

*Supporting information*

Additional 'supporting information' may be found in the on-line version of this article:

‘Online appendices’.