Thresholded Lasso: Variable Selection in High Dimensions

Laurent Callot[*] Mehmet Caner[†] Anders Bredahl Kock[‡]
and Juan Andres Riquelme[§]

February 9, 2015

Abstract

We propose a new estimator, the thresholded scaled Lasso, in high dimensional threshold regressions. First, we establish an upper bound on the ℓ_∞ estimation error of the scaled Lasso estimator of Lee et al. (2012). This is a non-trivial task as the literature on high-dimensional models has focused almost exclusively on ℓ_1 and ℓ_2 estimation errors. We show that this sup-norm bound can be used to distinguish between zero and non-zero coefficients at a much finer scale than would have been possible using classical oracle inequalities. Thus, our sup-norm bound is tailored to consistent variable selection via thresholding.

Our simulations show that thresholding the scaled Lasso yields substantial improvements in terms of variable selection. Finally, we use our estimator to shed further empirical light on the long running debate on the relationship between the level of debt (public and private) and GDP growth.

Keywords and phrases:Threshold model, sup-norm bound, thresholded scaled Lasso, oracle inequality, debt effect on gdp growth.
JEL classification: C13, C23, C26.

[*] VU University Amsterdam, Department of Econometrics and Operations Research, CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78) funded by the Danish National Research Foundation, and the Tinbergen Institute.

[†] North Carolina State University, Department of Economics, 4168 Nelson Hall, Raleigh, NC 27695. Email: mcaner@ncsu.edu

[‡] Aarhus University, Department of Economics and Business, and CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation. Fuglesangs Alle 4, 8210, Aarhus V Denmark. Email: akock@creates.au.dk

[§] North Carolina State University, Department of Economics,

1 Introduction

Threshold models have been heavily studied and used in the past twenty years or so. In econometrics the seminal articles by Hansen (1996) and Hansen (2000) showed that least squares estimation of threshold models is possible and feasible. These papers show how to test for the presence of a threshold and how to estimate the remaining parameters by least squares. Later, Caner and Hansen (2004) provided instrumental variable estimation of the threshold. These authors derived the limits for the threshold parameter in the reduced form as well as structural equations.

There have been many applications of threshold models in cross-section data. One of the most recent ones is the analysis of the public debt to GDP ratio in a threshold regression model by Caner et al. (2010). In the context of time series we refer to the articles by Caner and Hansen (2001), Seo (2006), Seo (2008), and Hansen and Seo (2002). In panel data, semi-parametrics, and least absolute deviation models, Hansen (1999), Linton and Seo (2007), Caner (2002), respectively, made contributions. For applications to stock markets and exchange rates we refer to Akdeniz et al. (2003) and Basci and Caner (2006). These authors argue that threshold model can contribute to reducing forecast errors.

To be precise, we shall study the model

$$Y_i = X_i'\beta_0 + X_i'\delta_0 1_{\{Q_i<\tau_0\}} + U_i, \qquad i = 1, ..., n \tag{1}$$

where $\beta_0, \delta_0 \in \mathbb{R}^m$ and τ_0 determines the location of the threshold/break. Q_i determines which regime we are in and could be the debt level in a growth regression or education in a wage regression. If $\delta_0 = 0$, there is no break and τ_0 is not identified. In that case the model is linear. In a very insightful recent paper Lee et al. (2012) proved finite sample oracle inequalities for the prediction and estimation error of the (scaled) Lasso applied to (1) in the case of fixed regressors and gaussian error terms. In their simulation section, they also extend their results to random regressors with Gaussian errors. Furthermore, they nicely showed that τ_0 exhibits the well known super efficiency phenomenon from low dimensional break point models even in the high-dimensional case. These authors also show that the scaled Lasso does not select too many irrelevant variables in the spirit of Bickel et al. (2009). However, their results are by no means trivial extensions of oracle inequalities for linear models as they show that the classical restricted eigenvalue condition must hold uniformly over the parameter space in threshold models. In addition, the probabilistic analysis is also much more refined than in the linear case.

The aim of this paper is to show that it is possible to consistently decide whether a break is present or not even in the high-dimensional change point model with random regressors. In other words, we show that it is possible to decide whether $\delta_0 = 0$ or if it possesses non-zero entries. To do so efficiently, we first establish an upper bound on the sup-norm convergence rate of the estimator $\hat{\delta}$ of δ_0 which is valid in even highly correlated designs. This is not an easy task as almost all previous work has focussed on establishing upper bounds on the ℓ_1 or ℓ_2 estimation error in the plain linear model. Exceptions are Lounici (2008) and van de Geer (2014) who provide sup-norm bounds in the high-dimensional linear model. To the best of our knowledge, we are the first to establish sup-norm bounds on the estimation error in a high-dimensional non-linear model. Our sup-norm bound is much smaller than the corresponding ℓ_1 and ℓ_2 bounds on the estimation error as it does *not* depend on the unknown number of non-zero coefficients s. Thus, our approach to break detection, which is based on thresholding, allows for a much finer distinction between zero and non-zero entries of δ_0. The result is that we can detect breaks which would be too small to detect if one thresholded based on classical ℓ_1 or ℓ_2 estimation error. In that sense, the sharp sup-norm bound is tailored to break detection in our context and we strengthen the result of selecting not too many irrelevant variables in the threshold model to selecting exactly the right ones with probability tending to one.

The debate regarding the impact of debt on GDP growth was recently reignited by the European public debt crisis as well the claim by Reinhart and Rogoff (2010) that public debt has a substantial negative effect on future GDP growth when the ratio of debt to GDP is over 90%. Following Reinhart and Rogoff (2010), several authors have econometrically investigated the presence of such a threshold. Of particular interest for us is the work of Cecchetti et al. (2012) who estimated threshold growth regressions using several measures of public and private debt as well as a set of standard controls. Using our thresholded Lasso estimator with the data of Cecchetti et al. (2012) we find robust evidence of a threshold in the effect of debt on future GDP growth. However, the effect of debt being above the threshold appears to be complex.

In Section 2, we recall the scaled Lasso estimator for threshold models of Lee et al. (2012). Section 3 establishes ℓ_∞ norm bounds for the estimation error of the scaled Lasso. This sup-norm bound is the basis for our new thresholded scaled Lasso estimator which is introduced in Section 4. Section 5 provides simulations supporting the selection consistency of our estimator. Section 6 reports the results of our growth regressions. All proofs are deferred to the appendix.

1.1 Notation

For any vector $x \in \mathbb{R}^k$ (for some $k \geq 1$), let $\|x\|_{\ell_1}$, $\|x\|_{\ell_2}$ and $\|x\|_{\ell_\infty}$ denote the ℓ_1, ℓ_2 and ℓ_∞ norms, respectively. Similarly, for any $m \times n$ matrix A, $\|A\|_{\ell_1}$, $\|A\|_{\ell_2}$ and $\|A\|_{\ell_\infty}$ denote the induced (operator) norms corresponding to the above three norms. They can be calculated as $\|A\|_{\ell_1} = \max_{1\leq j\leq n}\sum_{i=1}^m |A_{i,j}|$, $\|A\|_{\ell_2} = \sqrt{\phi_{\max}(A'A)}$ where $\phi_{\max}(\cdot)$ is the maximal eigenvalue, and $\|A\|_{\ell_\infty} = \max_{1\leq i\leq m}\sum_{j=1}^n |A_{i,j}|$, respectively. We will also need $\|A\|_\infty = \max_{i,j}|A_{i,j}|$ where the maximum extends over all entries of A. For real numbers a, b $a \vee b$ and $a \wedge b$ denote their maximum and minimum, respectively. Furthermore, the empirical norm of $y \in \mathbb{R}^n$ is given by $\|y\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n y_i^2}$.

We shall say that a real random variable Z is subgaussian if there exists positive constants A and B such that $P(|Z| > \tau) \leq Ae^{-Bt^2}$ for all $\tau > 0$. Z is said to be subexponential if there exists positive constants C and D such that $P(|Z| > \tau) \leq Ce^{-Dt}$ for all $\tau > 0$. For $x \in \mathbb{R}^k$, we will let $x^{(j)}$ denote its jth entry. Let "wpa1" denote with probability approaching one.

2 Scaled Lasso for Threshold Regression

Defining the $2m \times 1$ vectors $X_i(\tau) = \left(X_i', X_i'1_{\{Q_i<\tau\}}\right)'$ and $\alpha_0 = (\beta_0', \delta_0')'$ one can rewrite (1) as

$$Y_i = X_i(\tau_0)'\alpha_0 + U_i, \qquad i = 1, ..., n \tag{2}$$

where τ_0 is supposed to be an element of a parameter space $T = [t_0, t_1] \subset \mathbb{R}$ and α_0 is supposed to belong to a parameter space $\mathcal{A} \subset \mathbb{R}^{2m}$. This is exactly the model that Lee et al. (2012) studied in the case where m can be much larger than n. We shall be more specific about the probabilistic assumptions in Section 3.1. Let $J(\alpha_0) = \{j = 1, ..., 2m : \alpha_0 \neq 0\}$ be the indices of the non-zero coefficients with cardinality $|J(\alpha_0)|$. Denoting by $X(\tau)$ the $(n \times 2m)$ matrix whose rows are $X_i(\tau)'$, setting $Y = (Y_1, ..., Y_n)'$, and $U = (U_1, ..., U_n)$, (2) can be written more compactly as

$$Y = X(\tau_0)\alpha + U$$

Next, let $X^{(j)}(\tau)$ denote the jth column of $X(\tau)$ and define the $2m \times 2m$ diagonal matrix

$$D(\tau) = diag\{\|X^{(j)}(\tau)\|_n, j = 1, ..., 2m\}$$

Now set

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

where $\alpha = (\beta', \delta')' \in \mathcal{A}$ and define the scaled ℓ_1 penalty

$$\lambda\big\|D(\tau)\alpha\big\|_{\ell_1} = \lambda\sum_{j=1}^{2m}\|X^{(j)}(\tau)\|_n|\alpha_j|,$$

where λ is a tuning parameter about which we shall be explicit later. With this notation in place we define for each $\tau \in T$

$$\hat{\alpha}(\tau) = \operatorname*{argmin}_{\alpha\in A}\left\{S_n(\alpha,\tau) + 2\lambda\big\|D(\tau)\alpha\big\|_{\ell_1}\right\} \tag{3}$$

and

$$\hat{\tau} = \operatorname*{argmin}_{\tau\in T}\left\{S_n(\hat{\alpha}(\tau),\tau) + \lambda\big\|D(\tau)\hat{\alpha}(\tau)\big\|_{\ell_1}\right\}.$$

To be precise, $\hat{\tau}$ is an interval and in accordance with Lee et al. (2012) we define the maximum of the interval as the estimator $\hat{\tau}$. For every n, it suffices in practice to search over $Q_1, ..., Q_n$ as candidates for $\hat{\tau}$ as these are the points where $1_{\{Q_i<\tau\}}$, $i = 1, ..., n$ can change. Therefore, the estimator of (α_0, τ_0) is defined as $(\hat{\alpha}, \hat{\tau}) = (\hat{\alpha}(\hat{\tau}), \hat{\tau})$.

Assuming fixed regressors and and gaussian error terms Lee et al. (2012) established oracle inequalities for the prediction and ℓ_1 estimation error of the Lasso estimator $\hat{\alpha}$. When a break is present they also established upper bounds on the estimation error of $\hat{\tau}$. We contribute by establishing oracle inequalities in the sup-norm for this non-linear model and show that we can consistently detect breaks that are as small as $\sqrt{\frac{\log(m)}{n}}$.

3 Uniform Convergence Rate of the Scaled Lasso Estimator

In this section we establish upper bounds on the sup norm estimation error $\|\hat{\alpha} - \alpha_0\|_{\ell_\infty}$. As argued previously, and as will be made rigorous in Section 4, an upper bound $\|\hat{\delta} - \delta_0\|_{\ell_\infty}$ is what is really needed for break detection purposes. However, we shall actually establish a slightly stronger result here which also makes it possible to efficiently select variables from the first m columns of $X(\tau_0)$. This sup-norm bound is established separately for the case where no break is present and for the case where a break is present. Let X and $Z(\tau)$ denote the first and last m columns of $X(\tau)$ for $\tau \in T$, respectively, and define

$$r_n = \min_{1\le j\le m}\frac{\big\|Z^{(j)}(t_0)\big\|_n^2}{\big\|X^{(j)}\big\|_n^2}.$$

Note that under Assumption 1 below it follows by Lemma 3 in the appendix that r_n is bounded away from zero with probability tending to one. r_n is trivially never greater than one. Now define

$$\lambda = A\left(\frac{\log(3m)}{nr_n}\right)^{1/2} \tag{4}$$

as the tuning parameter for a constant $A \geq 0$. Assuming an i.i.d. sample we let $\Sigma(\tau) = E\left(X_1(\tau)X_1(\tau)'\right)$ denote the population covariance matrix of the covariates. In Lemma 1 below we give sufficient conditions for its inverse $\Theta(\tau)$ to exist as long as $\Sigma = E(X_1X_1')$ is invertible which is a standard assumption in regression models. Thus, the practical consequence is that the presence of indicator functions in the definition of $X_1(\tau)$ does not make it singular. Now we introduce the assumptions that our theorems rely on.

3.1 Assumptions

In this section we recall the assumptions used by Lee et al. (2012) in their Theorems 2 and 3 which are used as ingredients in the proofs of our Theorems 1 and 2. To be precise, we use the oracle inequalities for the ℓ_1 estimation errors of $\hat{\alpha}$ and $\hat{\tau}$ provided by Lee et al. (2012). We alter their assumptions slightly, as we are working in a random design as opposed to their fixed regressor design. However, Lee et al. (2012) have already argued how some of their assumptions could be valid in a random design and as a consequence we do note need to address these in detail.

Assumption 1. *Let $\{X_i, U_i, Q_i\}_{i=1}^n$ be an i.i.d. sample and let (X_1, U_1) be independent of Q_1. Furthermore, let Q_1 be uniformly distributed on $[0,1]$ and assume that all entries of X_1 and U_1 are subgaussian*[1] *with $\min_{1\leq j\leq m} E\left(X_1^{(j)^2}\right)$ bounded away from zero. (i) For the parameter space $\mathcal{A}$ for α_0, any $\alpha \equiv (\alpha_1, \cdots, \alpha_{2m}) \in \mathcal{A} \subset \mathbb{R}^{2m}$, including α_0, satisfies $\max_{1\leq j\leq 2m} |\alpha_j| \leq C_1$, for some constant $C_1 > 0$. In addition, $\tau_0 \in T = [t_0, t_1]$ with $0 < t_0 < t_1 < 1$. $(ii) \log(m)/n \to 0$.*

Assumption 1 is the one which has been altered the most compared to Lee et al. (2012) as the boundedness of certain norms of the covariates does no longer have to be assumed as this now follows directly from independence and subgaussianity of these. See Lemma 3 in the appendix for details. Furthermore, the absence of ties among the $Q_i,\ i = 1, ..., n$ (as required in Lee et al. (2012)) follows in an almost sure sense from these being uniformly (and thus continuously) distributed.

The assumption of the sample being i.i.d. can most likely be relaxed by exchanging the probabilisitic inequalities used in the appendix for ones allowing for weak dependences and/or heterogeneity. For convenience, we have also assumed that X_1 and Q_1 are independent. However, as the main contribution of this paper is to provide sup norm bounds for high-dimensional non-linear models as the first in the literature (to the best of our knowledge) we have chosen to keep the probabilistic framework simple in order not to suffocate the cardinal ideas in technicalities.

Assumption 2. *(Uniform Restricted Eigenvalue Condition). For some integer s such that $1 \leq s \leq 2m$, a positive number c_0 and some set $\mathcal{S} \subset \mathbb{R}$, the following condition holds wpa1*

$$\kappa(s, c_0, \mathcal{S}) = \min_{\tau\in\mathcal{S}} \min_{J_0\subset\{1,...,2m\},|J_0|\leq s} \min_{\gamma\neq 0,|\gamma_{J_0^c}|_1\leq c_0|\gamma_{J_0}|_1} \frac{|X(\tau)\gamma|_2}{n^{1/2}|\gamma_{J_0}|_2} > 0. \tag{5}$$

In the random design considered in this paper we require assumption 2 of Lee et al. (2012) above to be valid with probability tending to one. However, this is an unnecessarily high-level assumption as it can often be verified by assuming that $\Sigma(\tau)$ satisfies the uniform restricted eigenvalue condition (which it does in particular when it has full rank – as is in turns true under Assumption 1 if Σ has full rank as argued on page A4 in Lee et al. (2012)) and by showing that $\frac{1}{n}X'(\tau)X(\tau)$ is uniformly close to $\Sigma(\tau)$. Mimicking the arguments on pages A3-A6 in Lee et al. (2012) it can be shown that (5) above holds with probability tending to one under our Assumption 1 as long as Σ has full rank

[1] The notation suppresses that we are really dealing with a triangular array. Thus, more precisely, we assume uniform subgaussianity across the rows of this triangular array.

– a rather innocent assumption. Thus, Assumption 2 is almost automatic under Assumption 1 and we shall use this in the statements of Theorems 1 and 2 below.

For the next assumption, define $f_{\alpha,\tau}(x,q) = x'\beta + x'\delta 1_{\{q<\tau\}}$, and $f_0(x,q) = x'\beta_0 + x'\delta_0 1_{\{q<\tau_0\}}$ and let $m(\alpha)$ denote the number of non-zero elements of α.

Assumption 3. *(Identifiability under Sparsity and Discontinuity of Regression). For a given $s \geq |J(\alpha_0)|$, and for any η and τ such that $|\tau - \tau_0| > \eta \geq \min_i |Q_i - \tau_0|$, and $\alpha \in \{\alpha : m(\alpha) \leq s\}$ there exists a constant $c > 0$ such that, wpa1*

$$\|f_{\alpha,\tau} - f_0\|_n^2 > c\eta,$$

For this assumption Lee et al. (2012) (pages A7-A8) also provide sufficient conditions encompassing the assumptions made in Assumption 1 above.

Assumption 4. *(Smoothness of Design). For any $\eta > 0$, there exists a constant $C < \infty$ such that wpa1*

$$\sup_{1\leq j,k\leq m} \sup_{|\tau-\tau_0|<\eta} \frac{1}{n}\sum_{i=1}^{n} \left|X_i^{(j)} X_i^{(k)}\right| \left|1_{\{Q_i<\tau_0\}} - 1_{\{Q_i<\tau\}}\right| \leq C\eta.$$

Lee et al. (2012) argue that this is the case when the Q_i are continuously distributed and $E\left(\left|X_i^{(j)} X_i^{(k)}\right| |Q_i = \tau\right)$ is continuous and bounded in a neighborhood of τ_0 for all $1 \leq j,k \leq m$. Note however, that the outer supremum in Assumption 4 above is taken over all $1 \leq j,k \leq m$ as opposed to only $1 \leq j \leq m$ in Lee et al. (2012) as $\left|X_i^{(j)} X_i^{(k)}\right|$ has replaced ${X_i^{(j)}}^2$. This slight strengthening of the assumption is needed to establish an ℓ_∞ bound on the estimation error of $\hat{\alpha}$ in the case where a structural break is present (Theorem 2 below).

Assumption 5. *(Well defined second moments). For any η such that $1/n \leq \eta \leq \eta_0$, $h_n^2(\eta)$ is bounded where wpa1*

$$h_n^2(\eta) = \frac{1}{2n\eta} \sum_{i=\max\{1,[n(\tau_0-\eta)]\}}^{\min\{[n(\tau_0+\eta)],n\}} (X_i'\delta_0)^2,$$

where $[.]$ denotes the integer part of a real number.

Finally, we also need to impose the same technical regularity condition as Lee et al. (2012) which they denote Assumption 6 and present on page A23 of their paper. This assumption is satisfied asymptotically in our context when $s\left\|\delta_0\right\|_{\ell_1}\sqrt{\frac{\log(m)}{n}} \to 0$. Since $\max_{1\leq j\leq m}\delta_{0,j} \leq C_1$ by Assumption 1 above this is in turns true when $s\left|J(\delta_0)\right|\log(m)^{1/2}/\sqrt{n} \to 0$. The latter assumption will be assumed in Theorem 2 below (as we also need it for another purpose) and thus Assumption 6 in Lee et al. (2012) is automatic in our case.

3.2 sup-norm rate of convergence of $\hat{\alpha}$

We next turn to providing upper bounds on the ℓ_∞ estimation error of $\hat{\alpha}$. We distinguish between the case in which no break is present and the case in which a break is present.

Theorem 1. *Suppose that $\delta_0 = 0$ and let Assumptions 1 be satisfied. Furthermore, let $|J(\alpha)| \leq s$, assume that Σ has full rank and that $\Theta(\tau) = \Sigma^{-1}(\tau)$ satisfies $\sup_{\tau\in T}\left\|\Theta(\tau)\right\|_{\ell_\infty} < \infty$. Then, choosing λ as in (4) and assuming $s\sqrt{\frac{\log(mn)}{n}} \to 0$, one has*

$$\left\|\hat{\alpha} - \alpha_0\right\|_{\ell_\infty} = O_p\left(\sqrt{\frac{\log(m)}{n}}\right) = O_p(\lambda).$$

Thus, a fortiori, we also have $\left\|\hat{\delta}-\delta_0\right\|_{\ell_\infty} = O_p\left(\sqrt{\frac{\log(m)}{n}}\right) = O_p(\lambda)$.

Theorem 1 provides the stochastic order of the ℓ_∞ estimation error of $\hat{\alpha}$ for the case where no break is present. From Theorem 1 in Lee et al. (2012) (ignoring that their results are for non-random regressors) one can conclude that $\|\hat{\alpha}-\alpha_0\|_{\ell_1} = O_p\left(s\sqrt{\log(m)/n}\right)$. From this, one can of course also conclude that $\|\hat{\alpha}-\alpha_0\|_{\ell_\infty} \leq \|\hat{\alpha}-\alpha_0\|_{\ell_1} = O_p\left(s\sqrt{\log(m)/n}\right)$. However, our Theorem 1 shows that this rate is much too large as s may be almost as large as $O(\sqrt{n})$ without obstructing ℓ_1 norm consistency. Our much smaller bound will allow for more precise thresholding in Section 4.

We stress again that almost all research in high-dimensional models so far has focussed exclusively on providing upper bounds on the ℓ_1 and ℓ_2. ℓ_∞ bounds on the estimation error have been established for the Lasso in the plain linear regression model by Lounici (2008) and van de Geer (2014). However, to the best of our knowledge we are the first to establish sup-norm bounds for high-dimensional non-linear models, and certainly in the threshold model. As we shall see below, a sup-norm bound will yield much more precise variable selection results for the thresholded scaled Lasso than thresholding based on ℓ_1 or ℓ_2 bounds since the latter two are larger due to the presence of the unknown sparsity s. Next, consider the case where $\delta_0 \neq 0$, i.e. a break is present.

Theorem 2. *Suppose that* $\delta_0 \neq 0$ *and let Assumptions 1 and 3-5 be satisfied. Furthermore, let* $|J(\alpha)| \leq s$, *assume that* Σ *has full rank and that* $\left\|\Theta(\tau_0)\right\|_{\ell_\infty} < \infty$. *Then, choosing* λ *as in (4) and assuming* $s\left|J(\delta_0)\right|\sqrt{\frac{\log(m)}{n}} \to 0$, *one has*

$$\left\|\hat{\alpha}-\alpha_0\right\|_{\ell_\infty} = O_p\left(\sqrt{\frac{\log(m)}{n}}\right).$$

Thus, a fortiori, we also have $\left\|\hat{\delta}-\delta_0\right\|_{\ell_\infty} = O_p\left(\sqrt{\frac{\log(m)}{n}}\right) = O_p(\lambda)$.

The results of Theorem 2 are similar to those in Theorem 1 but the assumptions differ. First, $\left\|\Theta(\tau)\right\|_{\ell_\infty}$ only has to be bounded at τ_0 instead of uniformly over $T = [t_0, t_1]$ for $0 < t_0 < t_1 < 1$. Lemma 1 below shows that $\sup_{\tau\in T}\left\|\Theta(\tau)\right\|_{\ell_\infty} < \infty$ and $\left\|\Theta(\tau_0)\right\|_{\ell_\infty} < \infty$ in the equicorrelation design but of course with the former being no smaller than the latter. More importantly, requiring $s\left|J(\delta_0)\right|\log(m)^{1/2}/\sqrt{n} \to 0$ is in general more restrictive than requiring $s\sqrt{\frac{\log(mn)}{n}} \to 0$ as in Theorem 1. However, if the number of coefficient which break is bounded, i.e. $\left|J(\delta_0)\right| \leq B$ for an absolute constant B, then the rate requirement of Theorem 2 is actually slightly weaker than the one in Theorem 1.

The following Lemma shows that even when the covariates are highly correlated, Σ^{-1} exists and the assumptions $\sup_{\tau\in T}\left\|\Theta(\tau)\right\|_{\ell_\infty} < \infty$ and $\left\|\Theta(\tau_0)\right\|_{\ell_\infty} < \infty$ from Theorems 1 and 2, respectively, are satisfied. First, recall the definition of an equicorrelation design.

Definition 1. *We say that* Σ *is an equicorrelation matrix if*

$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

for some $-1 < \rho < 1$.

Lemma 1. *Let $\{X_i, U_i\}_{i=1}^n$ be an iid sample and assume that U_1 is uniformly distributed on $[0,1]$. Let $\Sigma = E(X_1 X_1')$ be an $m \times m$ equicorrelation matrix with $0 \leq \rho < 1$. Then Σ^{-1} exists and for all $\tau \in (0,1)$ one has $\left\|\Theta(\tau)\right\|_{\ell_\infty} \leq \frac{2}{(1-\tau)(1-\rho)}\left(2 \vee \frac{\tau+1}{\tau}\right)$. If, furthermore, $T = [t_0, t_1]$ for some $0 < t_0 < t_1 < 1$, then $\sup_{\tau \in T}\left\|\Theta(\tau)\right\|_{\ell_\infty}$ is bounded by a constant only depending on ρ.*

Lemma 1 states that $\left\|\Theta(\tau)\right\|_{\ell_\infty}$ is bounded for all $\tau \in (0,1)$ even when the correlation is arbitrarily close to, but different from, one. τ can not be zero or one since in that case $\Sigma(\tau)$ would be singular. From a modeling point of view this excludes breaks at the very endpoints of the sample which is a standard assumption in the literature.

4 Thresholded Scaled Lasso

In this section we utilize the ℓ_∞ bound established in Theorems 1 and 2 above to provide sharp thresholding results for the Scaled Lasso estimator. Recall that these theorems established that $\left\|\hat{\alpha} - \alpha_0\right\|_{\ell_\infty} \leq C\lambda$ with arbitrarily large probability, irrespective of whether a break is present or not, by choosing C sufficiently large. Before showing that the breaks can be revealed consistently we shall provide a slightly more general result stating that the truly zero coefficients can be distinguished from the non-zero ones. First, define the Thresholded Scaled Lasso estimator as

$$\tilde{\alpha}_j = \begin{cases} \hat{\alpha}_j & \text{if} \quad |\hat{\alpha}_j| \geq H \\ 0 & \text{if} \quad |\hat{\alpha}_j| < H \end{cases} \tag{6}$$

where H is the threshold determining whether a coefficient should be classified as zero or non-zero. In particular, we shall see that choosing $H = 2C\lambda$ results in consistent model selection. Here we stress once more that our threshold is much sharper than what would have been obtainable if we had directly used that $\left\|\hat{\alpha} - \alpha_0\right\|_{\ell_1} \leq Cs\lambda$ with probability tending to one from Lee et al. (2012). Thus, it is important to have an ℓ_∞ bound on the estimation error as this allows for a much finer distinction between the zero and the non-zero coefficients than would been possible from the usual ℓ_1 or ℓ_2 bounds. To be precise, let α_{0j} be a nonzero coefficient such that $|\alpha_{0j}|/\lambda \to \infty$ but $|\alpha_{0j}|/(s\lambda) \to 0$. Not that there may be a considerable wedge between $|\alpha_{0j}|/\lambda$ and $|\alpha_{0j}|/(s\lambda)$ as s can be almost as large as $\sqrt{n}$ such that this is a setting of practical relevance. Such an $\alpha_{0,j}$ will correctly be classified as non-zero when thresholding at the level λ (resulting from an ℓ_∞ bound) while it would wrongly be classified as zero when thresholding at the level $s\lambda$ (resulting from a plain ℓ_1 bound). This example underscores the importance of establishing ℓ_∞ bounds as in Theorems 1 and 2 prior to thresholding. Next, recall that $J(\alpha_0) = \{j = 1, ..., 2m : \alpha_{0j} \neq 0\}$ and define $J(\tilde{\alpha}) = \{j = 1, ..., 2m : \tilde{\alpha}_j \neq 0\}$. The following theorems establish the properties of the thresholded scaled Lasso and rely crucially on the ℓ_∞ bounds on the estimation error established in Theorems 1 and 2 above.

Theorem 3. *Let the assumptions of Theorems 1 and 2 be satisfied and assume that $\min_{j \in J(\alpha_0)} |\alpha_{0j}| > 3C\lambda$. Then, for all $\epsilon > 0$ there exists a C such that for $H = 2C\lambda = 2C\sqrt{\frac{\log(m)}{n}}$ one has $P\left(J(\tilde{\alpha}) = J(\alpha_0)\right) \geq 1 - \epsilon$.*

Theorem 3 states that consistent model selection is possible with the thresholded Lasso in the nonlinear break point regression model as long as the non-zero coefficients are at least of the order $\sqrt{\frac{\log(m)}{n}}$. This is considerably sharper than thresholding based on ℓ_1 estimation errors where consistent variable selection would require the non-zero coefficients to be at least of order $s\sqrt{\frac{\log(m)}{n}}$. The idea in the proof of Theorem 3 is similar to the one for the linear case in Lounici (2008).

Note that if one is only interested in finding out whether there is a break or not, i.e. whether δ_0 is non-zero or not, one can simply threshold $\tilde{\delta}$ only according to the rule in (6). Defining $J(\delta_0) = \{j = 1, ..., m : \delta_{0j} \neq 0\}$ and $J(\tilde{\delta}) = \{j = 1, ..., m : \tilde{\delta}_j \neq 0\}$ we have the following result on consistent break detection.

Theorem 4. *Let the assumptions of Theorems 1 and 2 be satisfied and assume that* $\min_{j \in J(\delta_0)} |\delta_{0j}| > 3C\lambda$. *Then, for all* $\epsilon > 0$ *there exists a* C *such that for* $H = 2C\lambda = 2C\sqrt{\frac{\log(m)}{n}}$ *one has* $P\left(J(\tilde{\delta}) = J(\delta_0)\right) \geq 1 - \epsilon$.

Break selection consistency is weaker than model selection consistency as it only requires classifying δ_0 correctly. However, it is still relevant as it answers the question whether a break is present or not. We discuss how to choose the threshold parameter C in practice in Section 5.

5 Simulations

In this section we report the results of a series of simulation experiments evaluating the finite sample properties of the thresholded scaled Lasso. We shall consider performance along the dimensions: increasing number of irrelevant variables, estimation in the absence of a threshold, increasing number of observations, scale of the parameters, and increasing number of non-zero variables.

The regressors are generated as $X_i \sim \mathcal{N}(0, I)$, the threshold variable $Q_i \sim \mathcal{U}[0, 1]$, and the innovations $U_i \sim \mathcal{N}(0, \sigma^2)$ where we set the residual variance $\sigma^2 = 0.25$, $i = 1, ..., n$. When the threshold parameter τ_0 is not explicitly stated it is set to $\tau_0 = 0.5$; we search for τ_0 over a grid from 0.15 to 0.85 by steps of 0.05. This grid is coarser than the grid used in Lee et al. (2012) which, in our experience, has a mild detrimental effect on the precision with which τ_0 is estimated but not on other measures of the quality of the estimator while substantially reducing computation time, thus allowing us to carry out more replications. We select the thresholding parameter C by BIC using a grid from 0.1 to 5, so that parameters smaller (in absolute value) than $\widehat{C}\widehat{\lambda}$ are set to zero by the thresholded scaled Lasso.

Every model is estimated with an intercept so that we estimate $2m + 1$ parameters, plus the threshold parameter τ_0. All the results reported below are based on 1000 replications. The simulation are carried with `R` (R Development Core Team, 2008) using the `glmnet` package of Friedman et al. (2010). The results (and those of the empirical application in section 6) can be replicated using `knitr` (Xie, 2014) and the supplementary material[2].

We report the following statistics, averaged across iterations.

- MSE: mean square prediction error.
- $|J(\hat{\alpha}) \cap J(\alpha_0)^c|$: number zero parameters incorrectly retained in the model.
- $|J(\alpha_0) \cap J(\hat{\alpha})^c|$: number of non-zero parameters excluded.
- Perfect Sel.: the share (in %) of iterations for which we have perfect model selection.
- $\|\hat{\alpha} - \alpha_0\|_1$: ℓ_1 estimation error for the parameters.
- $\|\hat{\alpha} - \alpha_0\|_\infty$: ℓ_∞ estimation error for the parameters.
- $|\hat{\tau} - \tau_0|$: absolute threshold parameter estimation error.

[2] Available at `https://github.com/lcallot/ttlas`

- C: selected (BIC) thresholding parameter.
- $\hat{\lambda}$: selected (BIC) penalty parameter.

	τ_0	MSE	$\lvert J(\hat{\alpha}) \cap J(\alpha_0)^c \rvert$	$\lvert J(\alpha_0) \cap J(\hat{\alpha})^c \rvert$	Perfect Sel	$\lVert \hat{\alpha} - \alpha_0 \rVert_1$	$\lVert \hat{\alpha} - \alpha_0 \rVert_\infty$	$\lvert \hat{\tau} - \tau_0 \rvert$	C	$\hat{\lambda}$
$m = 50$	0.3	1.20	4.50	0.06	3	3.39	0.85	0.27	-	0.06
		1.22	0.04	0.11	86	3.29	0.85	-	1.46	-
	0.4	1.42	5.23	0.07	1	3.68	0.94	0.22	-	0.05
		1.45	0.04	0.16	81	3.55	0.94	-	1.49	-
	0.5	1.55	5.72	0.07	0	3.99	1.00	0.18	-	0.05
		1.58	0.06	0.14	82	3.85	1.01	-	1.45	-
$m = 100$	0.3	1.34	5.59	0.05	1	3.99	0.95	0.25	-	0.07
		1.38	0.04	0.13	85	3.86	0.95	-	1.27	-
	0.4	1.56	6.26	0.08	0	4.29	1.03	0.22	-	0.07
		1.60	0.05	0.16	82	4.15	1.03	-	1.25	-
	0.5	1.77	7.27	0.12	0	4.77	1.10	0.19	-	0.07
		1.83	0.07	0.21	78	4.60	1.11	-	1.22	-
$m = 200$	0.3	1.57	7.06	0.10	0	4.65	1.06	0.25	-	0.09
		1.62	0.03	0.19	82	4.49	1.06	-	1.15	-
	0.4	1.80	8.10	0.12	0	5.04	1.14	0.22	-	0.09
		1.87	0.03	0.22	79	4.86	1.15	-	1.12	-
	0.5	2.22	9.20	0.26	0	5.82	1.27	0.18	-	0.09
		2.30	0.06	0.40	71	5.60	1.28	-	1.07	-
$m = 400$	0.3	1.73	8.81	0.15	0	5.38	1.16	0.26	-	0.10
		1.81	0.03	0.23	81	5.18	1.17	-	1.04	-
	0.4	2.16	9.35	0.33	0	6.17	1.30	0.22	-	0.12
		2.26	0.04	0.47	73	5.94	1.31	-	0.98	-
	0.5	2.84	9.81	0.66	0	7.26	1.46	0.19	-	0.13
		2.96	0.03	0.91	60	7.02	1.47	-	0.90	-

Table 1: Lasso (white background) and Thresholded Lasso (grey background). Increasing number of zero parameters and 3 locations of τ_0.

Table 1 contains the results of experiments where we consider 4 different dimensions for the parameter vectors and multiple locations for τ_0. The data is generated as follows:

- Sample size: $n = 200$, $\beta = [2, 2, 2, 2, 2, 0, ..., 0]$, $\delta = [2, -2, 2, -2, 2, 0, ..., 0]$.
- The length β and δ is $m = 50, 100, 200, 400$.

The most important finding in Table 1 is that across all settings the scaled Lasso almost never detects the true model while its thresholded version does so very often and rather consistent across the settings. As expected, the scaled Lasso does a good job at model screening in the sense that it retains all relevant variables in many instances. However, it often fails to exclude irrelevant variables. This is exactly where the thresholding sets in – it weeds out the falsely retained variables by the first step scaled Lasso. To illustrate this, consider the setting of $m = 400$ and $\tau_0 = 0.5$. Here the scaled Lasso includes almost ten irrelevant variables on average while its thresholded version includes as few as 0.03 irrelevant variables on average. Note also how the ℓ_∞ estimation error is much lower than the ℓ_1 counterpart confirming our theoretical results from Theorems 1 and 2, thus allowing for much sharper thresholding than usual. This important finding is confirmed in all of the other settings below.

Perfect model selection seems to be slightly easier for lower values of the threshold parameter τ_0. On the other hand, $\hat{\tau}$ becomes less precise as τ_0 is lowered. All other measures in general improve slightly when τ_0 is lowered. Increasing the dimension of the model, m, worsens most performance measures except for the estimation error of $\hat{\tau}$ which stays constant. Finally, in larger models more penalization is applied as can be seen from the larger choice of λ as m is increased.

Table 2 considers the case where no threshold effect is present, $\delta_0 = 0$, the exact data generating process is:

- Sample size: $n = 200$, $\beta = [2, 2, 2, 2, 2, 0, ..., 0]$, $\delta = [0, ..., 0]$.
- The length β and δ is $m = 50, 100, 200, 400$.

	MSE	$\lvert J(\hat{\alpha}) \cap J(\alpha_0)^c \rvert$	$\lvert J(\alpha_0) \cap J(\hat{\alpha})^c \rvert$	Perfect Sel	$\lVert \hat{\alpha} - \alpha_0 \rVert_1$	$\lVert \hat{\alpha} - \alpha_0 \rVert_\infty$	C	$\hat{\lambda}$
$m = 50$	0.29	1.56	0.00	23	0.60	0.16	-	0.07
	0.29	0.21	0.00	81	0.56	0.16	0.73	-
$m = 100$	0.30	1.56	0.00	23	0.65	0.17	-	0.08
	0.31	0.18	0.00	83	0.61	0.17	0.61	-
$m = 200$	0.31	1.45	0.00	27	0.70	0.18	-	0.09
	0.32	0.15	0.00	86	0.66	0.18	0.53	-
$m = 400$	0.32	1.44	0.00	27	0.74	0.19	-	0.10
	0.33	0.12	0.00	89	0.71	0.19	0.46	-

Table 2: Lasso (white background) and Thresholded Lasso (grey background). No threshold effect ($\delta = 0$), $n = 200$, 4 different length of the parameter vector.

The main finding of Table 2 is that almost all performance measures improve drastically compared to Table 1. This is the case in particular for large m as the performance is no longer worsened as m increases. Note, for example, that the MSE and ℓ_1 estimation error of $\hat{\alpha}$ are almost ten times lower for $m = 400$ than they were in Table 1. Most importantly for us, the perfect models selection percentage is now also stable across m.

In order to investigate the asymptotic properties of our procedure, Table 3 reveals the effect of increasing the sample size for two values of τ_0. The exact DGP is:

- Sample size: $n = 50, 100, 200, 500, 1000$.
- $\beta = [2, 2, 2, 2, 2, 0, ..., 0]$, $\delta = [2, -2, 2, -2, 2, 0, ..., 0]$.

		MSE	$\|J(\hat{\alpha})\cap J(\alpha_0)^c\|$	$\|J(\alpha_0)\cap J(\hat{\alpha})^c\|$	Perfect Sel	$\|\hat{\alpha}-\alpha_0\|_1$	$\|\hat{\alpha}-\alpha_0\|_\infty$	$\|\hat{\tau}-\tau_0\|$	C	$\hat{\lambda}$
$\tau_0 = 0.3$	$n = 50$	10.04	1.83	4.92	0	14.72	1.99	0.30	-	0.58
		10.64	0.29	5.51	0	14.66	1.99	-	0.67	-
	$n = 100$	3.34	7.22	1.09	0	7.92	1.51	0.27	-	0.15
		3.53	0.12	1.38	45	7.63	1.51	-	1.32	-
	$n = 200$	1.46	5.56	0.08	1	4.07	1.00	0.25	-	0.07
		1.50	0.04	0.16	82	3.95	1.00	-	1.25	-
	$n = 500$	0.76	3.31	0.01	6	2.27	0.64	0.17	-	0.04
		0.76	0.01	0.02	97	2.23	0.64	-	0.95	-
	$n = 1000$	0.50	2.62	0.00	10	1.51	0.45	0.06	-	0.03
		0.50	0.00	0.01	98	1.49	0.45	-	0.81	-
$\tau_0 = 0.5$	$n = 50$	8.98	1.81	4.84	0	14.56	2.00	0.21	-	0.48
		9.52	0.24	5.43	0	14.48	2.00	-	0.62	-
	$n = 100$	4.73	5.41	2.15	0	10.05	1.75	0.20	-	0.21
		4.94	0.12	2.62	23	9.84	1.75	-	1.00	-
	$n = 200$	1.83	7.41	0.12	0	4.83	1.14	0.18	-	0.07
		1.89	0.06	0.21	78	4.66	1.14	-	1.22	-
	$n = 500$	0.86	4.32	0.01	2	2.53	0.69	0.18	-	0.04
		0.87	0.01	0.04	96	2.48	0.69	-	0.96	-
	$n = 1000$	0.55	3.27	0.00	8	1.70	0.49	0.08	-	0.03
		0.55	0.01	0.01	98	1.67	0.49	-	0.80	-

Table 3: Lasso (white background) and Thresholded Lasso (grey background). Increasing sample size with $m = 100$ and 2 locations of τ_0.

As expected, the probability of correct model selection tends to one for the thresholded scaled Lasso. For the plain scaled Lasso, on the other hand, this probability reaches at most 11%. As seen already in Table 1, the problem that the scaled Lasso suffers from is false positives – it fails to exclude irrelevant variables even as the sample size increases. Finally, and as expected, the penalty applied (λ) decreases as n increases.

Table 4 considers different values of the non-zero coefficients to investigate the effect of the scale of these coefficients. The data is generated as:

- Sample size: $n = 100, 200$.
- $\beta = a[1, 1, 1, 1, 1, 0, ..., 0]$, $\delta = a[1, -1, 1, -1, 1, 0, ..., 0]$.
- $a = 0.3, 0.5, 1, 2$ is the scale of the non zero parameters.

		MSE	$\lvert J(\hat{\alpha}) \cap J(\alpha_0)^c \rvert$	$\lvert J(\alpha_0) \cap J(\hat{\alpha})^c \rvert$	Perfect Sel	$\lVert \hat{\alpha} - \alpha_0 \rVert_1$	$\lVert \hat{\alpha} - \alpha_0 \rVert_\infty$	$\lvert \hat{\tau} - \tau_0 \rvert$	C	$\hat{\lambda}$
$a = 0.3$	$n = 100$	0.50	0.41	5.50	0	2.27	0.30	0.28	-	0.15
		0.52	0.02	6.22	0	2.30	0.30	-	0.46	-
	$n = 200$	0.38	0.29	4.00	0	1.89	0.30	0.32	-	0.10
		0.39	0.01	4.56	0	1.91	0.30	-	0.44	-
	$n = 1000$	0.31	0.60	1.74	1	1.38	0.30	0.10	-	0.04
		0.31	0.00	2.21	5	1.38	0.30	-	0.51	-
$a = 0.5$	$n = 100$	0.75	0.57	4.49	0	3.43	0.50	0.25	-	0.15
		0.78	0.03	5.15	0	3.45	0.50	-	0.47	-
	$n = 200$	0.57	0.50	3.21	0	2.92	0.50	0.27	-	0.10
		0.58	0.01	3.95	0	2.93	0.50	-	0.48	-
	$n = 1000$	0.31	2.75	0.04	9	1.37	0.32	0.10	-	0.03
		0.31	0.00	0.06	94	1.35	0.32	-	0.75	-
$a = 1$	$n = 100$	1.87	1.12	3.52	0	6.31	1.00	0.22	-	0.18
		1.94	0.05	4.21	0	6.31	1.00	-	0.56	-
	$n = 200$	1.09	3.95	1.16	0	4.46	0.86	0.21	-	0.09
		1.12	0.04	1.54	39	4.39	0.86	-	0.88	-
	$n = 1000$	0.34	2.98	0.00	9	1.43	0.35	0.08	-	0.03
		0.34	0.00	0.01	99	1.41	0.35	-	0.83	-
$a = 2$	$n = 100$	4.68	5.32	2.12	0	10.01	1.76	0.20	-	0.21
		4.89	0.10	2.61	21	9.80	1.76	-	1.02	-
	$n = 200$	1.81	7.44	0.11	0	4.74	1.12	0.18	-	0.07
		1.87	0.05	0.21	78	4.57	1.12	-	1.23	-
	$n = 1000$	0.56	3.18	0.00	7	1.70	0.49	0.07	-	0.03
		0.56	0.01	0.01	98	1.68	0.49	-	0.79	-

Table 4: Lasso (white background) and Thresholded Lasso (grey background). Increasing parameter scale, 3 sample sizes, $\tau_0 = 0.5$.

When these are as small as 0.3 perfect model selection does not seem possible unless when $n = 1000$. On the other hand, the number of relevant variables excluded clearly decreases as n is increased. In general, no matter what the value of the non-zero coefficients are, all performance measures improve as n is increased, thus confirming the findings in Table 3. While variable selection is easier when the non-zero coefficients are well-separated from the zero ones, the MSE and estimation error of $\hat{\alpha}$ actually improve as the non-zero coefficients become smaller. The reason for this is that falsely classifying a non-zero coefficient as zero is less costly in terms of estimation error when this coefficient is already close to zero than when it is far from zero. On the other hand, $\hat{\tau}$ is estimated slightly more precisely as the non-zero coefficients become more separated from the zero ones.

		MSE	$\vert J(\hat{\alpha})\cap J(\alpha_0)^c\vert$	$\vert J(\alpha_0)\cap J(\hat{\alpha})^c\vert$	Perfect Sel	$\Vert\hat{\alpha}-\alpha_0\Vert_1$	$\Vert\hat{\alpha}-\alpha_0\Vert_\infty$	$\vert\hat{\tau}-\tau_0\vert$	C	$\hat{\lambda}$
$m_1 = 1$	$\tau_0 = 0.3$	0.34	0.10	0.00	90	0.36	0.25	0.25	-	0.09
		0.34	0.01	0.00	99	0.35	0.25	-	0.23	-
	$\tau_0 = 0.4$	0.35	0.09	0.00	91	0.36	0.26	0.22	-	0.09
		0.35	0.01	0.00	99	0.36	0.26	-	0.20	-
	$\tau_0 = 0.5$	0.36	0.10	0.00	90	0.37	0.27	0.23	-	0.09
		0.36	0.01	0.00	99	0.37	0.27	-	0.19	-
$m_1 = 5$	$\tau_0 = 0.3$	1.84	1.11	0.19	31	2.59	0.90	0.19	-	0.08
		1.86	0.01	0.28	78	2.57	0.91	-	0.68	-
	$\tau_0 = 0.4$	2.03	1.11	0.20	32	2.67	0.93	0.18	-	0.08
		2.05	0.02	0.30	76	2.64	0.93	-	0.63	-
	$\tau_0 = 0.5$	2.05	1.01	0.16	35	2.57	0.91	0.17	-	0.08
		2.06	0.02	0.27	79	2.55	0.91	-	0.61	-
$m_1 = 10$	$\tau_0 = 0.3$	5.08	2.85	0.81	5	6.54	1.36	0.19	-	0.08
		5.12	0.06	1.06	51	6.48	1.36	-	1.09	-
	$\tau_0 = 0.4$	4.84	2.68	0.66	7	6.17	1.28	0.18	-	0.08
		4.88	0.05	0.89	57	6.11	1.28	-	1.01	-
	$\tau_0 = 0.5$	5.05	2.56	0.65	7	6.10	1.25	0.18	-	0.08
		5.09	0.04	0.90	59	6.05	1.25	-	0.95	-
$m_1 = 25$	$\tau_0 = 0.3$	19.93	9.92	4.35	0	23.72	1.87	0.20	-	0.07
		20.27	0.31	5.56	10	23.45	1.88	-	2.45	-
	$\tau_0 = 0.4$	19.13	10.32	3.56	0	22.76	1.88	0.20	-	0.07
		19.48	0.30	4.63	11	22.46	1.88	-	2.31	-
	$\tau_0 = 0.5$	18.32	9.90	3.05	0	21.53	1.75	0.23	-	0.07
		18.62	0.30	3.94	19	21.25	1.75	-	2.09	-

Table 5: Lasso (white background) and Thresholded Lasso (grey background). Increasing number of non zero parameters (m_1), fixed number of zeros ($m_0 = 100$), and 3 locations of τ_0.

Finally, Table 5 investigates the effect of reducing the sparsity of the model, i.e. of increasing the number of non-zero coefficients.

- Sample size: $n = 200$.
- $\beta = [2, ..., 2, 0, ..., 0]$, $\delta = [2, ..., 2, 0, ..., 0]$.
- β and δ contain both $m_0 = 100$ parameters equal to zero.
- β and δ contain both $m_1 = 1, 5, 10, 50$ parameters equal to 2.
- The length β and δ is $m = m_0 + m_1$.

Irrespective of the value of τ_0, perfect model selection becomes harder as the number of relevant variables increases. As our theory is based on the assumption of sparsity, this is not a surprising finding. The MSE and estimation error of $\hat{\alpha}$ also increase by a lot while the estimation error of $\hat{\tau}$ is virtually unaffected by the number of relevant variables. Notice that the threshold parameter, C, increases drastically as the number of non-zero coefficients increases. The explanation for this is that thresholding seeks to avoid excluding one of the many relevant variables by setting the threshold higher as there are now more relevant variables at risk of being exluded.

6 Application

This application aims at investigating the presence of a threshold in the effect of debt on future GDP growth. The academic discussion regarding the impact of debt on growth, and the existence of a threshold above which debt becomes severely detrimental to future growth, has been reignited by Reinhart and Rogoff (2010) who provided evidence for the existence of such a threshold. The evidences presented by Reinhart and Rogoff (2010) have been challenged by Herndon et al. (2014), but others have put forth supportive evidences for this thesis, see among others Cecchetti et al. (2012); Caner et al. (2010); Baum et al. (2013).

6.1 Data

We use the data made available by Cecchetti et al. (2012)[3] which originates mainly from the IMF and OECD data bases. The data contains four measures of debt-to-GDP ratio for:

1. Government debt,
2. Corporate debt,
3. Private debt (corporate + household),
4. Total (non financial institutions) debt (private + government).

Notice that private and total debt are aggregate measures of debt.

The data of Cecchetti et al. (2012) also contains a measure of household debt that we drop as the series is incomplete. A set of control variables, composed of standard macroeconomic indicators, is also included in the data.

1. GDP: The logarithm of the *per capita* GDP.
2. Savings: Gross savings to GDP ratio.

[3] The original data is available at `http://www.bis.org/publ/work352.htm`, and can also be found in the replication material for this section.

3. ΔPop: Population growth.
4. School: Years spent in secondary education.
5. Open: Openness to trade, exports plus imports over GDP.
6. ΔCPI: Inflation.
7. Dep: Population dependency ratio.
8. LL: Ratio of liquid liabilities to GDP.
9. Crisis: An indicator for banking crisis in the subsequent 5 years. This is taken from Reinhart and Rogoff (2010).

The data is observed for 18 countries[4] from 1980 to 2009 at an annual frequency. We lose one observation at the start of the sample due to first differencing and five at the end of the sample due to computing the 5 years ahead average growth rate, so that the full sample is 1981-2004. The details on the construction of each variables can be found in Cecchetti et al. (2012).

6.2 Results

In order to evaluate the impact of debt on growth, as well as the potential presence of a threshold in this effect, we estimate a set of growth regressions. As in Cecchetti et al. (2012) our left hand side variable is the 5 years forward average rate of growth of per capita GDP. Even though our estimator is not a panel estimator we choose to pool the data so as to make our results comparable with those of Cecchetti et al. (2012) and benefit from a larger sample.

We report a first set of results focusing on the impact of government debt on future GDP growth in Table 6. We consider 3 different samples: 1981 to 2004 (full sample, 414 observations), 1990 to 2004 (252 observations), and a sample with no overlapping data (5 years[5], 90 observations). For the full sample we report results for models estimated with and without country specific dummies (noted FE in the tables). We do not report the estimated parameters associated with the country specific dummies.

We estimate the models including every control variable and a single debt measure, that is, 23 parameters to estimate (11 parameters in β,11 parameters in δ, and the threshold parameter τ) including the intercept and the thresholded intercept plus, in some instances, 17 country specific dummies. The country specific dummies are not penalized. The grid of threshold parameters goes from the 15^{th} to the 85^{th} centiles of the threshold variable by steps of 5 centiles. We select the thresholding parameter C by BIC using a grid from 0.1 to 5, so that parameters smaller (in absolute value) than $\widehat{C}\widehat{\lambda}$ are set to zero by the thresholded scaled Lasso.

Table 6 reports the estimated parameters for the 4 specifications of the model, all including government debt. The L and T in the header of the table indicates a scaled Lasso estimate ($\widehat{\beta}$, $\widehat{\delta}$) or thresholded scaled Lasso estimate ($\widetilde{\beta}$, $\widetilde{\delta}$). The upper panel of each table reports $\widehat{\beta}$ and $\widetilde{\beta}$, the middle panel $\widehat{\delta}$ and $\widetilde{\delta}$, and the lower panel give the values of $\widehat{\tau}$, $\widehat{\lambda}$, and $\widehat{C}$. Recall that the effect of the regressors when the threshold variable is below its threshold is given by $\widehat{\beta}+\widehat{\delta}$ ($\widetilde{\beta}+\widetilde{\delta}$) while the effect when the threshold variable is above its threshold is given by $\widehat{\beta}$ ($\widetilde{\beta}$) for the scaled Lasso (thresholded scaled Lasso).

[4] US, Japan, Germany, the United Kingdom, France, Italy, Canada, Australia, Austria, Belgium, Denmark, Finland, Greece, the Netherlands, Norway, Portugal, Spain, and Sweden.

[5] 1984,1989,1994,1999,2004.

	Threshold:	Government		Government		Government		Government	
		L	T	L	T	L	T	L	T
	intercept	42.43	42.43	79.611	79.611	86.416	86.416	136.988	136.988
	GDP	-3.643	-3.643	-7.419	-7.419	-7.495	-7.495	-11.621	-11.621
	Savings	-0.035	-0.035	0.033	0.033	0.02	0.02		
	ΔPop	-1.692	-1.692	-1.493	-1.493	-0.879	-0.879	-0.813	-0.813
	School	0.426	0.426	0.507	0.507	0.095	0.095	-0.082	-0.082
$\hat{\beta}$	Open	0.003		0.026		0.024	0.024	0.037	0.037
	ΔCPI	-0.061	-0.061	-0.056	-0.056	-0.157	-0.157	-0.252	-0.252
	Dep	-0.091	-0.091	-0.104	-0.104	-0.132	-0.132	-0.22	-0.22
	LL	-0.433	-0.433	0.33	0.33	0.574	0.574	0.631	0.631
	Crisis	-1.277	-1.277	-1.58	-1.58	-0.949	-0.949	-1.396	-1.396
	Government	-0.713	-0.713					-0.518	-0.518
	intercept	-12.167	-12.167	-1.504	-1.504				
	GDP								
	Savings	0.087	0.087	-0.037		-0.052	-0.052	0.008	
	ΔPop	1.563	1.563	0.42	0.42	0.222	0.222	0.61	0.61
	School	-0.077	-0.077			0.203	0.203	0.098	0.098
$\hat{\delta}$	Open	-0.006		0.007		0.012			
	ΔCPI								
	Dep	0.181	0.181			-0.035	-0.035		
	LL	0.827	0.827	0.909	0.909				
	Crisis	-0.459	-0.459	-0.294	-0.294	-1.338	-1.338		
	Government	1.762	1.762	1.471	1.471			-3.23	-3.23
	$\widehat{\tau}$	0.82	0.82	0.68	0.68	0.59	0.59	0.65	0.65
	$\widehat{\lambda}$	0.007	0.007	0.015	0.015	0.007	0.007	0.008	0.008
	$\widehat{C}$	-	0.1	-	0.3	-	0.1	-	0.1
	Sample	1981 - 2004		1981 - 2004		1990 - 2004		No overlap	
	FE	×		✓		✓		✓	

Table 6: 4 specifications with government debt included as threshold variable and regressor. Estimated parameters for the Lasso (L) and Thresholded Lasso (T). Empty cells are parameters set to zero, dashes indicate parameters not included in the model.

	Threshold:	Corporate		Private		Total	
		L	T	L	T	L	T
$\hat{\beta}$	intercept	140.097	140.097	126.236	126.236	134.725	134.725
	GDP	-11.642	-11.642	-10.616	-10.616	-11.396	-11.396
	Savings	-0.026	-0.026	-0.031	-0.031	-0.011	-0.011
	ΔPop	-1.063	-1.063			-0.995	-0.995
	School	-0.172	-0.172			-0.132	-0.132
	Open	0.053	0.053	0.041	0.041	0.047	0.047
	ΔCPI	-0.204	-0.204	-0.19	-0.19	-0.166	-0.166
	Dep	-0.242	-0.242	-0.191	-0.191	-0.235	-0.235
	LL	0.332	0.332	0.316	0.316	0.376	0.376
	Crisis	-0.96	-0.96	-0.319	-0.319	-0.943	-0.943
	Corporate	0.491	0.491	-	-	-	-
	Private	-	-	-0.968	-0.968	-	-
	Total	-	-	-	-	0.284	0.284
$\hat{\delta}$	intercept	8.261	8.261	2.301	2.301		
	GDP						
	Savings	-0.243	-0.243	0.022	0.022		
	ΔPop	-2.154	-2.154	-1.1	-1.1	2.387	2.387
	School	-0.29	-0.29	-0.33	-0.33	0.387	0.387
	Open			-0.007		0.063	0.063
	ΔCPI	-0.032	-0.032	-0.082	-0.082	0.777	0.777
	Dep					-0.192	-0.192
	LL	1.175	1.175	0.365	0.365		
	Crisis	-2.389	-2.389	-1.167	-1.167	-31.521	-31.521
	Corporate			-	-	-	-
	Private	-	-	0.563	0.563	-	-
	Total	-	-	-	-		
	$\widehat{\tau}$	0.69	0.69	1.62	1.62	2	2
	$\widehat{\lambda}$	0.001	0.001	0.005	0.005	0.002	0.002
	$\widehat{C}$	-	0.1	-	0.1	-	0.1
	Sample	1981 - 2004		1981 - 2004		1981 - 2004	
	FE	✓		✓		✓	

Table 7: Growth regressions with corporate, private, or total debt (see header) included both as threshold variable and as regressor. Estimated parameters, pooled data, Lasso (L) and Thresholded Lasso (T). Empty cells are parameters set to zero, dashes indicate parameters not included in the model.

A large fraction of $\widehat{\beta}$ is non zero, the Lasso drops a single variable twice, while $\widehat{\delta}$ is more sparse, the Lasso drops between 2 and 7 variables. The thresholding parameter $\widehat{C}$ is always chosen among the lowest values in the search grid, this nonetheless results in between 1 and 3 extra parameters being discarded compared to the scaled Lasso. A threshold ($\widehat{\tau}$) for the effect of government debt on growth is found at between 60% and 80% of GDP, consistent with the findings of Cecchetti et al. (2012); Reinhart and Rogoff (2010); Caner et al. (2010); Baum et al. (2013).

The level of GDP is found to have a negative effect on GDP per capita growth as predicted by the income convergence hypothesis, as do inflation, the dependency ratio, population growth, and crises. Considering the effect of both $\widehat{\beta}$ and $\widehat{\delta}$, our model indicates in most instances that government debt has a positive effect below the threshold and a negative effect, or no effect at all, above the debt threshold. *Ceteris paribus* a 10 percentage point increase in the government debt to GDP ratio, when it is above the threshold, is found to result in a decrease of the average 5 year growth rate between 0.07% and zero. Looking at this effect of high debt on future growth in isolation is overly restrictive though since there are large changes in the other parameters of the model when the debt threshold is crossed. This is in particular the case for financial variables. Interestingly, crises are found to have a more detrimental effect on growth for countries with a government debt ratio below the threshold and while liquid liabilities (LL) are beneficial to the future growth of a country with low debt this does not appear to be the case when debt is high.

Table 7 reports estimates for 3 other measures of debt in a model with country dummies and using the full sample, the same model used in the first two columns of Table 6. The sparsity pattern in Table 7 is comparable to that of Table 6 and some similarities are found between the estimated values. Again, the level of *per capita* GDP is found to have a negative impact on future growth, as are the dependency ratio, inflation, population growth, and financial crisis.

A threshold is always found and identified, 69% for corporate debt, 162% for private debt, and 200% for the total debt. The large value of the estimated thresholds for private and total debt can be explained by the fact that these are aggregate measures of debt and hence of a substantially larger magnitude than either corporate of government debts. The effect of corporate and total debt is found to be positive and not directly affected by the threshold whereas the effect of private debt is negative, and more so when private debt is high. As previously, financial crises are found to have a stronger negative impact on countries with low debt, though crises are detrimental to growth irrespective of the level of debt.

7 Conclusion

In this paper we considered high-dimensional threshold regressions and provided sup-norm oracle inequalities for the estimation error of the scaled Lasso of Lee et al. (2012). These results are non-trivial as most research has focused on either ℓ_1 or ℓ_2 oracle inequalities. The sup-norm bounds are shown to be crucial for exact variable selection by means of thresholding. To be precise, we can distinguish at a much finer scale between zero and non-zero coefficients than would have been possible if thresholding had been based on either ℓ_1 or ℓ_2 oracle inequalities.

We carry out simulations and show that the thresholded scaled Lasso performs well in model selection. Finally, we estimate a set of growth regression documenting the existence of a threshold in the amount of debt relative to GDP. Several parameters change when the threshold is crossed making the effect of high debt on future growth unclear.

Future work includes investigating the effect of multiple thresholds.

APPENDIX

The following result is needed in the proofs of Theorems 1 and 2. It is similar to Lemma 6 in Lee et al. (2012) but allows for random regressors and non-gaussian error terms.

Lemma 2. *Let Assumption 1 be satisfied. Then,*

$$\left\|\frac{1}{n}X'(\hat{\tau})U\right\|_{\ell_\infty}=O_p\left(\sqrt{\frac{\log(m)}{n}}\right)$$

Proof. First, note that $\left\|\frac{1}{n}X'(\hat{\tau})U\right\|_{\ell_\infty}\leq\sup_{\tau\in T}\left\|\frac{1}{n}X'(\tau)U\right\|_{\ell_\infty}$ such that it suffices to bound the right hand side. Let $\epsilon>0$ be arbitrary. By the independence of $(X_1,...,X_n,U_1,...,U_n)$ and $(Q_1,...,Q_n)$ one has for $j=1,...,m$,

$$\begin{aligned}P\left(\sup_{\tau\in T}\left|\frac{1}{n}\sum_{i=1}^{n}X_i^{(j)}U_i1_{\{Q_i<\tau\}}\right|>\epsilon\,\middle|\,(Q_1,...,Q_n)\right)&=P\left(\max_{1\leq k\leq n}\left|\frac{1}{n}\sum_{i=1}^{k}X_i^{(j)}U_i\right|>\epsilon\,\middle|\,(Q_1,...,Q_n)\right)\\&=P\left(\max_{1\leq k\leq n}\left|\frac{1}{n}\sum_{i=1}^{k}X_i^{(j)}U_i\right|>\epsilon\right)\qquad(7)\end{aligned}$$

almost surely, where the first equality used that conditional on $(Q_1,...,Q_n)$, $\left(1_{\{Q_1<\tau\}},...,1_{\{Q_n<\tau\}}\right)$ can only take n different values (and sorted $\{X_i,U_i,Q_i\}_{i=1}^n$ by $(Q_1,...,Q_n)$ in ascending order). The second equality used the independence $(X_1,...,X_n,U_1,...,U_n)$ and $(Q_1,...,Q_n)$. Next, by Corollary 4 in Montgomery-Smith (1993) there exists a universal constant $c>0$ such that

$$P\left(\max_{1\leq k\leq n}\left|\frac{1}{n}\sum_{i=1}^{k}X_i^{(j)}U_i\right|>\epsilon\right)\leq cP\left(\left|\sum_{i=1}^{n}X_i^{(j)}U_i\right|>\frac{\epsilon n}{c}\right)\qquad(8)$$

As $X_i^{(j)}U_i$ is subexponential (the product of two subgaussian variables is subexponential) for all $i=1,...,n$ and $j=1,...,m$, Corollary 5.17 in Vershynin (2012) yields

$$P\left(\left|\sum_{i=1}^{n}X_i^{(j)}U_i\right|>\frac{\epsilon n}{c}\right)\leq2\exp\left(-d\left[(\epsilon/K)^2\wedge(\epsilon/K)\right]n\right)\qquad(9)$$

where $d>0$ and $K=K(c)>0$ are absolute constants. Therefore, choosing $\epsilon=A\sqrt{\frac{\log(m)}{n}}$ for some $A\geq1$ yields

$$P\left(\left|\sum_{i=1}^{n}X_i^{(j)}U_i\right|>\frac{\epsilon n}{c}\right)\leq2\exp\left(-\frac{dA}{K^2\vee K}\left[\frac{\log(m)}{n}\wedge\sqrt{\frac{\log(m)}{n}}\right]n\right)\leq2\exp\left(-\frac{dA}{K^2\vee K}\log(m)\right)$$

where the second estimate used that $\log(m)/n\to0$ such that $\frac{\log(m)}{n}$ is smaller than its square root for n sufficiently large. Hence,

$$P\left(\sup_{\tau\in T}\left|\frac{1}{n}\sum_{i=1}^{n}X_i^{(j)}U_i1_{\{Q_i<\tau\}}\right|>\epsilon\,\middle|\,(Q_1,...,Q_n)\right)\leq2c\exp\left(-\frac{dA}{K^2\vee K}\log(m)\right)$$

for all $j=1,...,m$ almost surely. Taking expectations over $(Q_1,...,Q_n)$ yields

$$P\left(\sup_{\tau\in T}\left|\frac{1}{n}\sum_{i=1}^{n}X_i^{(j)}U_i1_{\{Q_i<\tau\}}\right|>\epsilon\right)\leq2c\exp\left(-\frac{dA}{K^2\vee K}\log(m)\right).\qquad(10)$$

Therefore, combining (9) and (10), a union bound over $2m$ terms yields

$$P\left(\sup_{\tau\in T}\left\|\frac{1}{n}X'(\tau)U\right\|_{\ell_\infty}>\epsilon\right)\leq 2m(1+c)\exp\left(-\frac{dA}{K^2\vee K}\log(m)\right).$$

Choosing A sufficiently large implies that $\sup_{\tau\in T}\left\|\frac{1}{n}X'(\tau)U\right\|_{\ell_\infty}=O_p\left(\sqrt{\frac{\log(m)}{n}}\right)$ using the definition of $\epsilon=A\sqrt{\log(m)/n}$. □

Lemma 3. *Let assumption 1 be satisfied. Then,* $\sup_{\tau\in T}\max_{1\leq j\leq 2m}\left\|X^{(j)}(\tau)\right\|_n=O_p(1)$ *and* $\min_{1\leq j\leq 2m}\left\|X^{(j)}(t_0)\right\|_n$ *is bounded away from zero wpa1.*

Proof. Consider the first claim and note that $\sup_{\tau\in T}\max_{1\leq j\leq 2m}\left\|X^{(j)}(\tau)\right\|_n=\max_{1\leq j\leq m}\left\|X^{(j)}(\tau)\right\|_n$. As $X_1^{(j)}$ is uniformly subgaussian in $j=1,...,m$ it also holds that $E\left(X_1^{(j)2}\right)$ is uniformly bounded (this follows by Lemma 2.2.1 in van der Vaart and Wellner (1996) and the inequalities at the bottom of page 95 in that reference). Thus, by the triangle inequality and subadditivity of $x\mapsto\sqrt{x}$,

$$\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^{(j)2}}\leq\sqrt{\frac{1}{n}\left|\sum_{i=1}^n\left(X_i^{(j)2}-EX_i^{(j)2}\right)\right|}+\sqrt{EX_1^{(j)2}}$$

and hence it suffices to bound $\sqrt{\frac{1}{n}\left|\sum_{i=1}^n\left(X_i^{(j)2}-EX_i^{(j)2}\right)\right|}$, or, equivalently, $\frac{1}{n}\left|\sum_{i=1}^n\left(X_i^{(j)2}-EX_i^{(j)2}\right)\right|$ uniformly in $j=1,...,n$ by a constant with probability tending to 1. As the $X_i^{(j)2}$ are uniformly subexponential (as they are a product of uniformly subgaussian random variables) in $j=1,...,m$ Corollary 5.17 in Vershynin (2012) implies that for any $\epsilon>0$ there exist constants $c,K>0$ (see Vershynin (2012) for the exact meaning of the constants) such that

$$P\left(\frac{1}{n}\left|\sum_{i=1}^n\left(X_i^{(j)2}-EX_i^{(j)2}\right)\right|>\epsilon\right)\leq 2\exp\left(-c\left[(\epsilon/K)^2\wedge(\epsilon/K)\right]n\right)$$

for all $j=1,...,m$. Now, choosing $\epsilon=K\vee K/c$, the union bound yields that

$$P\left(\max_{1\leq j\leq m}\frac{1}{n}\left|\sum_{i=1}^n\left(X_i^{(j)2}-EX_i^{(j)2}\right)\right|>\epsilon\right)\leq 2me^{-n}\to 0$$

as $\log(m)/n\to 0$. Thus, $K\vee K/c$ is large enough to be the sought constant.

Now turn to the second claim and observe $\min_{1\leq j\leq 2m}\left\|X^{(j)}(t_0)\right\|_n=\min_{m+1\leq j\leq 2m}\left\|X^{(j)}(t_0)\right\|_n$. Note that by Assumption 1,

$$\min_{1\leq j\leq m}E\left(X_1^{(j)2}1_{\{Q_1<t_0\}}\right)=\min_{1\leq j\leq m}E\left(X_1^{(j)2}\right)t_0=:r>0.$$

where the first equality used the independence of X_1 and Q_1 as well as that Q_1 us uniformly distributed on $[0,1]$. Therefore, it suffices to show that $\max_{1\leq j\leq m}\frac{1}{n}\left|\sum_{i=1}^n\left(X_i^{(j)2}1_{\{Q_i<t_0\}}-EX_i^{(j)2}1_{\{Q_i<t_0\}}\right)\right|\leq d\leq r/2$ with probability tending to one. As $X_1^{(j)2}1_{\{Q_1<t_0\}}$ is subexponential it follows once more from Corollary 5.17 in Vershynin (2012) that for $d=K\wedge r/2$

$$P\left(\frac{1}{n}\left|\sum_{i=1}^n\left(X_i^{(j)2}1_{\{Q_i<t_0\}}-EX_i^{(j)2}1_{\{Q_i<t_0\}}\right)\right|>d\right)\leq 2\exp\left(-c\left[(d/K)^2\wedge(d/K)\right]n\right)\leq 2e^{\frac{-cd^2}{K^2}n}$$

for $j = 1, ..., m$. Thus, by the union bound

$$P\left(\max_{1\leq j\leq m}\frac{1}{n}\Big|\sum_{i=1}^{n}\left(X_i^{(j)^2}1_{\{Q_i<t_0\}} - EX_i^{(j)^2}1_{\{Q_i<t_0\}}\right)\Big| \geq d\right) \leq 2me^{\frac{-cd^2}{K^2}n}$$

which tends to zero as $\frac{\log(m)}{n} \to 0$ by assumption 1. □

Proof of Theorem 1. Note first that when $\delta_0 = 0$, for any random variable V

$$Y_i = X_i'\beta_0 + U_i = X_i'\beta_0 + X_i'1_{\{Q_i<V\}}\delta_0 + U_i,$$

since

$$X_i'1_{\{Q_i<V\}}\delta_0 = 0.$$

In particular, this is true for $V = \hat{\tau}$. Next, since $\hat{\alpha} = (\hat{\beta}', \hat{\delta}')'$ satisfies the Karush-Kuhn-Tucker conditions for a minimum, one has

$$-\frac{1}{n}X'(\hat{\tau})\left(Y - X(\hat{\tau})\hat{\alpha}\right) + \lambda D(\hat{\tau})z(\hat{\tau}) = 0$$

where $\|z(\hat{\tau})\|_{\ell_\infty} \leq 1$ and $z(\hat{\tau})_j = sign(\hat{\alpha}_j)$ if $\hat{\alpha}_j \neq 0$ and. This can be rewritten as

$$\frac{1}{n}X'(\hat{\tau})X(\hat{\tau})\left(\hat{\alpha} - \alpha_0\right) = \frac{1}{n}X'(\hat{\tau})U_i - \lambda D(\hat{\tau})z(\hat{\tau}).$$

which is equivalent to

$$\Sigma(\hat{\tau})\left(\hat{\alpha} - \alpha_0\right) = \left(\Sigma(\hat{\tau}) - \frac{1}{n}X'(\hat{\tau})X(\hat{\tau})\right)\left(\hat{\alpha} - \alpha_0\right) + \frac{1}{n}X'(\hat{\tau})U - \lambda D(\hat{\tau})z(\hat{\tau}).$$

Next, $\Theta(\tau) = \Sigma(\tau)^{-1}$ exists for all $\tau \in T$ under Assumption 1 when Σ has full rank as argued in the discussion of Assumption 2. In fact, $\kappa = \kappa(s, 3, T) > 0$ with probability tending to one as is needed in order to invoke Theorem 2 of Lee et al. (2012) below. It follows that $\Sigma(\hat{\tau})$ is invertible with inverse $\Theta(\hat{\tau})$. Thus,

$$\hat{\alpha} - \alpha_0 = \Theta(\hat{\tau})\left(\Sigma(\hat{\tau}) - \frac{1}{n}X'(\hat{\tau})X(\hat{\tau})\right)\left(\hat{\alpha} - \alpha_0\right) + \Theta(\hat{\tau})\frac{1}{n}X'(\hat{\tau})U - \lambda\Theta(\hat{\tau})D(\hat{\tau})z(\hat{\tau}).$$

Now recall that for matrices A, B and a vector c of compatible dimensions, one has $\|ABc\|_{\ell_\infty} \leq \|A\|_{\ell_\infty}\|Bc\|_{\ell_\infty} \leq \|A\|_{\ell_\infty}\|B\|_\infty\|c\|_{\ell_1}$ (see, eg, Horn and Johnson (2013) Chapter 5). Using this,

$$\begin{aligned}\left\|\hat{\alpha} - \alpha_0\right\|_{\ell_\infty} \leq& \left\|\Theta(\hat{\tau})\right\|_{\ell_\infty}\left\|\left(\Sigma(\hat{\tau}) - \frac{1}{n}X'(\hat{\tau})X(\hat{\tau})\right)\right\|_\infty\left\|(\hat{\alpha} - \alpha_0)\right\|_{\ell_1}\\ &+\left\|\Theta(\hat{\tau})\right\|_{\ell_\infty}\left\|\frac{1}{n}X'(\hat{\tau})U\right\|_{\ell_\infty} + \lambda\left\|\Theta(\hat{\tau})\right\|_{\ell_\infty}\left\|D(\hat{\tau})\right\|_{\ell_\infty}\left\|z(\hat{\tau})\right\|_{\ell_\infty}\\ \leq& \sup_{\tau\in T}\left\|\Theta(\tau)\right\|_{\ell_\infty}\sup_{\tau\in T}\left\|\left(\Sigma(\tau) - \frac{1}{n}X'(\tau)X(\tau)\right)\right\|_\infty\left\|(\hat{\alpha} - \alpha_0)\right\|_{\ell_1}\\ &+\sup_{\tau\in T}\left\|\Theta(\tau)\right\|_{\ell_\infty}\left\|\frac{1}{n}X'(\hat{\tau})U\right\|_{\ell_\infty} + \lambda\sup_{\tau\in T}\left\|\Theta(\tau)\right\|_{\ell_\infty}\max_{1\leq j\leq m}\left\|X^{(j)}\right\|_n \qquad (11)\end{aligned}$$

where we have also used $\left\|z(\hat{\tau})\right\|_{\ell_\infty} \leq 1$. Next, note that $\sup_{\tau\in T}\left\|\Theta(\tau)\right\|_{\ell_\infty}$ is bounded by assumption. Furthermore, by Lemma 2, $\left\|\frac{1}{n}X'(\hat{\tau})U\right\|_{\ell_\infty} = O_p\left(\sqrt{\frac{\log(m)}{n}}\right)$ while $\max_{1\leq j\leq m}\left\|X^{(j)}\right\|_n = O_p(1)$

by Lemma 3. Finally, it follows by the arguments on page A6 and the last inequality before Appendix B in Lee et al. (2012) that $\sup_{\tau\in T}\left\|\left(\Sigma(\tau)-\frac{1}{n}X'(\tau)X(\tau)\right)\right\|_\infty = O_p\left(\sqrt{\frac{\log(mn)}{n}}\right)$ while $\left\|\hat{\alpha}-\alpha_0\right\|_{\ell_1} = O_p\left(s\sqrt{\frac{\log(m)}{n}}\right)$ by Theorem 2 in the same reference. Using this in (11) yields, with $\lambda = O\left(\sqrt{\log(m)/n}\right)$,

$$\left\|\hat{\alpha}-\alpha_0\right\|_{\ell_\infty} = O_p\left(\sqrt{\frac{\log(m)}{n}}\left(s\sqrt{\frac{\log(mn)}{n}}+2\right)\right) = O_p\left(\sqrt{\frac{\log(m)}{n}}\right)$$

as $s\sqrt{\frac{\log(mn)}{n}} \to 0$. □

Proof of Theorem 2. First, since $\hat{\alpha} = (\hat{\beta}', \hat{\delta}')'$ satisfies the Karush-Kuhn-Tucker conditions for a minimum, one has

$$-\frac{1}{n}X'(\hat{\tau})\left(Y - X(\hat{\tau})\hat{\alpha}\right) + \lambda D(\hat{\tau})z(\hat{\tau}) = 0$$

where $\|z(\hat{\tau})\|_{\ell_\infty} \leq 1$ and$z(\hat{\tau})_j = sign(\hat{\alpha}_j)$ if $\hat{\alpha}_j \neq 0$. This can be rewritten as

$$-\frac{1}{n}X'(\hat{\tau})\left(X(\tau_0)\alpha_0 - X(\hat{\tau})\hat{\alpha}\right) = \frac{1}{n}X'(\hat{\tau})U - \lambda D(\hat{\tau})z(\hat{\tau})$$

which is equivalent to

$$\frac{1}{n}X'(\hat{\tau})X(\hat{\tau})\left(\hat{\alpha}-\alpha_0\right) - \frac{1}{n}X'(\hat{\tau})\left(X(\tau_0)-X(\hat{\tau})\right)\alpha_0 = \frac{1}{n}X'(\hat{\tau})U - \lambda D(\hat{\tau})z(\hat{\tau}).$$

The above display can be rewritten as

$$\Sigma(\tau_0)\left(\hat{\alpha}-\alpha_0\right) - \frac{1}{n}X'(\hat{\tau})\left(X(\tau_0)-X(\hat{\tau})\right)\alpha_0 = \left(\Sigma(\tau_0) - \frac{1}{n}X'(\hat{\tau})X(\hat{\tau})\right)\left(\hat{\alpha}-\alpha_0\right) + \frac{1}{n}X'(\hat{\tau})U - \lambda D(\hat{\tau})z(\hat{\tau}).$$

Next $\Theta(\tau_0) = \Sigma(\tau_0)^{-1}$ exists under Assumption 1 by the discussion after Assumption 2 as Σ is assumed to exist. In fact, $\kappa = \kappa(s, 5, S) > 0$ where $S = \left\{|\tau-\tau_0| \leq \eta_0\right\}$ and $\eta_0 = n^{-1} \vee K_1\sqrt{s}\lambda$ [6] is satisfied with probability tending to one as is needed in order to invoke Theorem 3 of Lee et al. (2012) below (it is even satisfied when S is replaced by T). Thus, one may rewrite the above display as

$$\begin{aligned}\hat{\alpha}-\alpha_0 = {} & \Theta(\tau_0)\frac{1}{n}X'(\hat{\tau})\left(X(\tau_0)-X(\hat{\tau})\right)\alpha_0 + \Theta(\tau_0)\left(\Sigma(\tau_0) - \frac{1}{n}X'(\hat{\tau})X(\hat{\tau})\right)\left(\hat{\alpha}-\alpha_0\right)\\ & + \Theta(\tau_0)\frac{1}{n}X'(\hat{\tau})U - \lambda\Theta(\tau_0)D(\hat{\tau})z(\hat{\tau})\end{aligned}$$

such that arguments similar to those leading to (11) yield

$$\begin{aligned}\left\|\hat{\alpha}-\alpha_0\right\|_{\ell_\infty} \leq {} & \left\|\Theta(\tau_0)\right\|_{\ell_\infty}\left\|\frac{1}{n}X'(\hat{\tau})\left(X(\tau_0)-X(\hat{\tau})\right)\alpha_0\right\|_{\ell_\infty}\\ & + \left\|\Theta(\tau_0)\right\|_{\ell_\infty}\left\|\left(\Sigma(\tau_0) - \frac{1}{n}X'(\hat{\tau})X(\hat{\tau})\right)\right\|_\infty\left\|\hat{\alpha}-\alpha_0\right\|_{\ell_1}\\ & + \left\|\Theta(\tau_0)\right\|_{\ell_\infty}\left\|\frac{1}{n}X'(\hat{\tau})U\right\|_{\ell_\infty} + \lambda\left\|\Theta(\tau_0)\right\|_{\ell_\infty}\max_{1\leq j\leq n}\left\|X^{(j)}\right\|_n \qquad (12)\end{aligned}$$

[6] Here $K_1 = \sqrt{7C_1C_2}$ where C_2 is the constants proven to exist in Lemma 3 in the appendix ensuring that $\sup_{\tau\in T}\max_{1\leq j\leq 2m}\left\|X^{(j)}(\tau)\right\|_n \leq C_2$ with arbitrarily large probability (more precisely, for any $\epsilon > 0$ there exists a C_2 such that $\sup_{\tau\in T}\max_{1\leq j\leq 2m}\left\|X^{(j)}(\tau)\right\|_n \leq C_2$ with probability at least $1-\epsilon$).

where we used that $\left\|z(\hat{\tau})\right\|_{\ell_\infty} \leq 1$. First, note that $\left\|\Theta(\tau_0)\right\|_{\ell_\infty}$ is bounded by assumption. Next, denoting by $Z(\tau_0)$ and $Z(\hat{\tau})$ the last m columns of $X(\tau_0)$ and $X(\hat{\tau})$, respectively, one has

$$\left\|\frac{1}{n}X'(\hat{\tau})\left(X(\tau_0) - X(\hat{\tau})\right)\alpha_0\right\|_{\ell_\infty} = \left\|\frac{1}{n}X'(\hat{\tau})\left(Z(\tau_0) - Z(\hat{\tau})\right)\delta_0\right\|_{\ell_\infty} \tag{13}$$

By Theorem 3 in Lee et al. (2012) one has $|\hat{\tau} - \tau_0| = O_p\left(s\frac{\log(m)}{n}\right)$ such that the probability of $\mathcal{A} = \left\{|\hat{\tau} - \tau_0| \leq Ks\frac{\log(m)}{n}\right\}$ can be made arbitrarily large by choosing $K > 0$ sufficiently large. Thus, on $\mathcal{A}$,

$$\begin{aligned}\left\|\frac{1}{n}X'(\hat{\tau})\left(Z(\tau_0) - Z(\hat{\tau}))\right)\delta_0\right\|_{\ell_\infty} &\leq \sup_{1\leq j,k\leq m}\frac{1}{n}\sum_{i=1}^n\left|X_i^{(j)}X_i^{(k)}\right|\left|1_{\{Q_i<\tau_0\}} - 1_{\{Q_i<\hat{\tau}\}}\right|\left\|\delta_0\right\|_{\ell_1}\\ &\leq KC_1s\left|J(\delta_0)\right|\frac{\log(m)}{n}\end{aligned}$$

by Assumptions 1 and 4. As we have assumed that $s|J(\delta_0)|\log(m)^{1/2}/\sqrt{n} \to 0$, we have in particular that

$$\left\|\frac{1}{n}X'(\hat{\tau})\left(X(\tau_0) - X(\hat{\tau})\right)\alpha_0\right\|_{\ell_\infty} = O_p\left(\sqrt{\frac{\log(m)}{n}}\right).. \tag{14}$$

Next, note that

$$\left\|\left(\Sigma(\tau_0) - \frac{1}{n}X'(\hat{\tau})X(\hat{\tau})\right)\right\|_\infty \leq \left\|\left(\Sigma(\tau_0) - \frac{1}{n}X'(\tau_0)X(\tau_0)\right)\right\|_\infty + \left\|\frac{1}{n}\left(X'(\tau_0)X(\tau_0) - X'(\hat{\tau})X(\hat{\tau})\right)\right\|_\infty$$

First, by the subgaussianity of the covariates and the error terms Corollary 5.14 in Vershynin (2012) and a union bound yield that[7] $\left\|\left(\Sigma(\tau_0) - \frac{1}{n}X'(\tau_0)X(\tau_0)\right)\right\|_\infty = O_p\left(\sqrt{\frac{\log(m)}{n}}\right)$. Next, by arguments similar to the ones leading to (14), one also has

$$\left\|\frac{1}{n}\left(X'(\tau_0)X(\tau_0) - X'(\hat{\tau})X(\hat{\tau})\right)\right\|_\infty \leq \sup_{1\leq j,k\leq m}\frac{1}{n}\sum_{i=1}^n\left|X_i^{(j)}X_i^{(k)}\right|\left|1_{\{Q_i<\tau_0\}} - 1_{\{Q_i<\hat{\tau}\}}\right| \leq Ks\frac{\log(m)}{n}$$

on $\mathcal{A}$ by Assumption 4. Therefore, as $s\log(m)^{1/2}/\sqrt{n} \to 0$ (implied by our assumption $s|J(\delta_0)|\log(m)^{1/2}/\sqrt{n} \to 0$), we conclude that

$$\left\|\left(\Sigma(\tau_0) - \frac{1}{n}X'(\hat{\tau})X(\hat{\tau})\right)\right\|_\infty = O_p\left(\sqrt{\frac{\log(m)}{n}}\right) \tag{15}$$

Furthermore, by Lemma 2, $\left\|\frac{1}{n}X'(\hat{\tau})U\right\|_{\ell_\infty} = O_p\left(\sqrt{\frac{\log(m)}{n}}\right)$ and $\left\|\hat{\alpha} - \alpha_0\right\|_{\ell_1} = O_p\left(s\sqrt{\frac{\log(m)}{n}}\right)$ by Theorem 3 in Lee et al. (2012). Finally, $\max_{1\leq j\leq m}\left\|X^{(j)}\right\|_n = O_p(1)$ by Lemma 3 which in conjunction with (14) and (15) yields in (12)

$$\left\|\hat{\alpha} - \alpha_0\right\|_{\ell_\infty} = O_p\left(\sqrt{\frac{\log(m)}{n}}\right)$$

where have again used that $s\log(m)^{1/2}/\sqrt{n} \to 0$.

□

[7] Alternatively, the arguments on pages A4-A6 in Lee et al. (2012) yield a uniform (in τ) upper bound on $\left\|\left(\Sigma(\tau) - \frac{1}{n}X'(\tau)X(\tau)\right)\right\|_\infty$ of the order $O_p\left(\sqrt{\frac{\log(mn)}{n}}\right)$ which could also be used resulting in only slightly worse rates.

Proof of Lemma 1. First, note that

$$\Sigma(\tau) = \begin{pmatrix} \Sigma & \tau\Sigma \\ \tau\Sigma & \tau\Sigma \end{pmatrix}$$

such that by the formula for the inverse of a partitioned matrix with $\Theta = \Sigma^{-1}$

$$\Theta(\tau) = \Sigma^{-1}(\tau) = \begin{pmatrix} \frac{1}{1-\tau}\Sigma^{-1} & \frac{-1}{1-\tau}\Sigma^{-1} \\ \frac{-1}{1-\tau}\Sigma^{-1} & \frac{\tau}{\tau(\tau-1)}\Sigma^{-1} \end{pmatrix} = \frac{1}{1-\tau}\begin{pmatrix} 1 & -1 \\ -1 & \frac{1}{\tau} \end{pmatrix} \otimes \Theta. \tag{16}$$

Thus, it suffices to bound $\left\|\Sigma^{-1}\right\|_{\ell_\infty}$. To this end, note that $\Sigma = (1-\rho)I + \rho\iota\iota'$ where ι is a $m \times 1$ vector of ones. Thus, by the Sherman-Morrison-Woodbury formula, Σ^{-1} exists and equals

$$\Theta = \Sigma^{-1} = \frac{1}{1-\rho}\left(I - \frac{\rho\iota\iota'}{1-\rho+\rho m}\right)$$

which implies that (using $\rho/(1-\rho+\rho m) \leq 1$)

$$\left\|\Theta\right\|_{\ell_\infty} = \frac{1}{1-\rho}\left(1 - \frac{\rho}{1-\rho+\rho m} + \frac{\rho(m-1)}{1-\rho+\rho m}\right) = \frac{1}{1-\rho}\left(\frac{1-3\rho+2m\rho}{1-\rho+m\rho}\right) \leq \frac{2}{1-\rho}. \tag{17}$$

Thus, combining (16) and (17) yields the first claim of the lemma. The second claim follows trivially from the first. □

Proof of Theorem 3. We consider the zero and non-zero coefficients separately and show that both groups will be classified correctly. Note that by Theorems 1 and 2 for every $\epsilon > 0$ there exists a $C > 0$ such that $\|\hat{\alpha} - \alpha\| \leq C\lambda$ on a set $\mathcal{D}$ with probability at least $1-\epsilon$. The following arguments all take place on this set. Consider the truly zero coefficients first. To this end, let $j \in J(\alpha_0)^c$ and note that

$$\max_{j \in J(\alpha_0)^c} |\hat{\alpha}_j| \leq C\lambda < 2C\lambda = H$$

such that $\tilde{\alpha} = 0$ by the definition of the thresholded scaled Lasso.

Next, consider the non-zero coefficients. To this end, let $j \in J(\alpha_0)$ and note that

$$|\hat{\alpha}_j| \geq \min_{j \in J(\alpha_0)} |\alpha_j| - |\hat{\alpha}_j - \alpha_{j0}| \geq 3C\lambda - C\lambda = 2C\lambda = H$$

such that $|\tilde{\alpha}| = |\hat{\alpha}| \neq 0$ by the definition of the thresholded scaled Lasso and the assumption that $\min_{j \in J(\alpha_0)} |\alpha_j| > 3C\lambda$ □

Proof of Theorem 4. Proceeds exactly as the proof of Theorem 3. □

References

Akdeniz, L., A. Altay-Salih, and M. Caner (2003). Time varying betas help in asset pricing: threshold capm. *Studies in Nonlinear Dynamics and Econometrics*.

Basci, E. and M. Caner (2006). Are real exchange rates non-stationary or non-linear? evidence from a new threshold unit root test. *Studies in Nonlinear Dynamics and Econometrics*.

Baum, A., C. Checherita-Westphal, and P. Rother (2013). Debt and growth: New evidence for the euro area. *Journal of International Money and Finance 32*, 809–821.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 1705–1732.

Caner, M. (2002). A note on lad estimation of a threshold model. *Econometric Theory 18*, 800–814.

Caner, M. and B. E. Hansen (2001). Threshold autoregression with a unit root. *Econometrica 69*, 1555–1596.

Caner, M. and B. E. Hansen (2004). Iv estimation of threshold models. *Econometric Theory 20*, 813–843.

Caner, M., F. Koehler-Geib, and T. Grennes (2010). Finding the tipping point when sovereign debt turns bad. *Sovereign Debt and Financial Crisis*, 64–75.

Cecchetti, S. G., M. Mohanty, and F. Zampolli (2012). The real effects of debt. *Bank for International Settlements Working Paper No. 352.*.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica 68*, 575–603.

Hansen, Bruce, E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica 64*, 432–430.

Hansen, Bruce, E. (1999). Threshold effects in non dynamic panels, estimation, testing and inference. *Journal of Econometrics 93*, 345–368.

Hansen, Bruce, E. and B. Seo (2002). Testing for two-regime threshold cointegration in vector error correction models. *Journal of Econometrics 110*, 293–318.

Herndon, T., M. Ash, and R. Pollin (2014). Does high public debt consistently stifle economic growth? a critique of reinhart and rogoff. *Cambridge journal of economics 38*(2), 257–279.

Horn, R. and C. Johnson (2013). *Matrix Analysis.* Cambridge University Press.

Lee, S., M. H. Seo, and Y. Shin (2012). The lasso for high-dimensional regression with a possible change-point. *arXiv preprint arXiv:1209.4875*.

Linton, O. and M.-H. Seo (2007). A smoothed least squares estimator for threshold regression models. *Journal of Econometrics 141*, 704–735.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of Statistics 2*, 90–102.

Montgomery-Smith, S. (1993). Comparison of sums of independent identically distributed random variables. *Probability and Mathematical Statistics 14*, 281–285.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Reinhart, C. M. and K. S. Rogoff (2010). Growth in a time of debt. *American Economic Review 100*(2), 573–78.

Seo, M.-H. (2006). Bootstrap testing for the null of no cointegration in a threshold vector error correction model. *Journal of Econometrics 134*, 129–150.

Seo, M.-H. (2008). Unit root test in a threshold autoregression: asymptotic theory and residual-based block bootstrap. *Econometric Theory 24*, 1699–1716.

van de Geer, S. A. (2014). Statistical theory for high-dimensional models. *Lecture Notes*.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes.* Springer.

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices, in Compressed Sensing: Theory and Applications (edited by Y. Eldar and G. Kutyniok).* Cambridge University Press.

Xie, Y. (2014). knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, and R. D. Peng (Eds.), *Implementing Reproducible Computational Research.* Chapman and Hall/CRC. ISBN 978-1466561595.