Threshold Regression Estimation via Lasso, Elastic-Net, and Lad-Lasso:

A Simulation Study with Applications to Urban Traffic Data

by

Maria van Schaijik

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved May 2015 by the
Graduate Supervisory Committee:

Ioannis Kamarianakis, Co-Chair
Mark Reiser, Co-Chair
John Stufken

ARIZONA STATE UNIVERSITY

August 2015

ABSTRACT

Threshold regression is used to model regime switching dynamics where the effects of the explanatory variables in predicting the response variable depend on whether a certain threshold has been crossed. When regime-switching dynamics are present, new estimation problems arise related to estimating the value of the threshold. Conventional methods utilize an iterative search procedure, seeking to minimize the sum of squares criterion. However, when unnecessary variables are included in the model or certain variables drop out of the model depending on the regime, this method may have high variability. This paper proposes Lasso-type methods as an alternative to ordinary least squares. By incorporating an L_1 penalty term, Lasso methods perform variable selection, thus potentially reducing some of the variance in estimating the threshold parameter. This paper discusses the results of a study in which two different underlying model structures were simulated. The first is a regression model with correlated predictors, whereas the second is a self-exciting threshold autoregressive model. Finally the proposed Lasso-type methods are compared to conventional methods in an application to urban traffic data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Regime Switching Models, where linear dynamics depend on the level of a threshold variable have many applications in Economics and Finance. The idea is that certain predictor variables may play a different role in determining the value for the response variable depending on whether or not some threshold has been crossed. Often there is good reason, either empirical or theoretical, to believe that a certain variable contains a threshold. If this is the case, the first step in estimating the underlying model is to identify the location of the threshold. Once a threshold estimate is obtained, coefficients for predictor variables in each regime can be estimated.

Conventional methods for estimating the location of the threshold utilize ordinary least squares (OLS) regression in an iterative search procedure. These methods fit a multi-regime model for every potential threshold point with-in the threshold variable. Then the model with the smallest residual sum of squares is selected as the best non-linear model. In order to determine whether or not non-linear dynamics are present, the non-linear model must be compared to a linear model. This can be done by means of a log-likelihood ratio test, or by minimizing some information criterion, such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC).

There are a number of scenarios which can complicate threshold estimation. There may be a large number of potential predictor variables. These variables may be correlated, and some of these variables may drop out of the model depending on the regime. In these scenarios the conventional method for estimating the threshold may have high variability. We also consider self-exciting threshold auto regressive (SETAR) models, where an additional step of estimating the lag order p is necessary. This is usually done by first estimating the number of lags under a linear assumption. In addition to variability, the least squares method may incorrectly favor a linear model with additional predictors rather than selecting the true regime switching model.

Another problem with least squares regression is that this method is sensitive to outliers. In many applications, such as modeling stock market returns or traffic flow systems, outliers

are common, suggesting that the random disturbances may have heavy-tailed distributions. Thus there is also a need for robust threshold estimation methods.

We propose Lasso-type penalized estimation methods, which perform simultaneous variable selection and estimation, as an alternative method to OLS. Such methods could reduce some of the variance in estimating the threshold by eliminating insignificant variables and lags at each step in the iterative procedure. We consider traditional Lasso, and two variations: Elastic-net (E-net), which does well when variables are correlated, and Lad-Lasso. The latter combines least absolute deviation regression, which is robust to outliers, with penalized estimation. Thus it should perform well in the case of heavy tailed error distributions when there are also regime-dependent dynamics.

In this paper we will consider two different types of regime-switching models: one where the predictor variables are correlated and second a SETAR model. We will compare our estimation method to the conventional method in a simulation study. Our focus is on threshold estimation. We also evaluate how well Lasso, E-net, and Lad-lasso perform in selecting the correct variables or lags in each regime.

In application, when regime-switching dynamics are thought to be present, the number of regimes is usually unknown. However, a two-regime model is often sufficient to capture nonlinear dynamics. For the purposes of this paper, only two-regime models were considered.

Lasso has been applied to linear autoregressive models (Nardi and Rinaldo, 2011); in a very recent paper, Lee et al. (2014), demonstrated the benefits of Lasso for high-dimensional regression when there is the possibility of regime-switching behavior. The objective of this paper is to combine threshold estimation with penalized estimation.

The beginning of this paper, parts 1-4, provides an introduction to regime-switching and SETAR models. We discuss conventional methods for threshold estimation, model identification, and specifying lag order. We provide a brief overview of Lasso-type panelized estimation methods. In part five we present our alternative method for threshold estimation. We explain the motivation for employing these methods, and how we incorporate the Lasso-penalty. In part six we present the results of our simulation study. Finally we apply our

method and conventional methods to data from an urban traffic network. We conclude with a discussion of our results, and areas which require further exploration.

2 REGIME-SWITCHING MODELS

The relationship between the response variable and the predictor variables in a two-regime model are typically expressed as a piecewise function of the predictor variables and the noise, with the threshold parameter, γ.

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_{j1} + \epsilon_i \quad q_i \leq \gamma$$
$$y_i = \sum_{j=1}^{p} x_{ij}\beta_{j2} + \epsilon_i \quad q_i > \gamma$$

Which, for convenience, we re-parameterize as follows:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{j=1}^{p} x_{ij}\delta_j I(q_i > \gamma) + \epsilon_i$$

This expression allows us to think of q as a dummy variable. In matrix form we have:

$$Y = \boldsymbol{X\beta} + \boldsymbol{X\delta} \circ I(q_i > \gamma) + \boldsymbol{\epsilon}$$

where q_i is the i^{th} element of the threshold variable q, $\boldsymbol{\delta}$ is the vector containing the the change in intercept and change in slope coefficients, and $I(\cdot)$ represents a vector whose elements are one or zero and are determined by the indicator function.

From the re-parameterized model, it is evident that if the threshold, γ is known, we can employ the OLS criterion for estimation as we would in a linear scenario. However, if γ is unknown, reliable estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ depend on first having a good estimate of γ.

A common approach for estimating the threshold is through an iterative search procedure. It is usually assumed that the breakpoint exists within the observed threshold variable, $\boldsymbol{q}$, which may or may not belong to the set of covariates. The iterative procedure searches through a trimmed vector $\boldsymbol{qs}$, avoiding the extreme elements of $\boldsymbol{q}$, which could lead

to an improper model matrix. The residual sum of squares is minimized for each element in $\boldsymbol{qs}$. Finally the location of the breakpoint is selected by choosing the element of $\boldsymbol{qs}$ with the minimum residual sum of squares.

$$Q_i(\boldsymbol{\beta}) = ||Y - (\boldsymbol{X\beta} + \boldsymbol{X\delta} \circ I(\boldsymbol{q} > qs_i))||_2$$

$$\hat{\gamma} = argmin_{qs_i \in \boldsymbol{q}}(Q_i)$$

Regime-switching models, where the regime change is triggered by an observable variable, gained prominence during the 80s and 90s. Hansen (2011) provides an overview of the literature on threshold estimation for this class of models.

3 SETAR MODELS

A popular method of fitting time-series data is with an autoregressive (AR) model. In autoregressive models some of the variation in the current observations of a variable, y_t, can be explained by previous realizations (lags) of that variable. When fitting such an autoregressive model, the first step is to estimate the lag order p, the number of steps backward in time that contribute to predicting y at the current time, t.

Estimating p is done by minimizing an information criterion formulated as $IC(p)$=T$log(\hat{\sigma}(p))+c_T(p+1)$, where c_T is the penalty term specific to the information criterion chosen, and $\hat{\sigma}(p)$ is the mean of the squared residuals for the fitted model of order p. Once the lag order p has been estimated, the coefficients for each time lag can be estimated by ordinary least squares regression (OLS).

SETAR models have regime-switching dynamics, where the current regime is determined by the value of y at a certain lag, d. A two regime model can be expressed as follows:

$$y_t = \phi_{10} + \phi_{11} y_{t-1} + ... + \phi_{1p_1} y_{t-p_1} + \epsilon_t, \quad y_{t-d} \leq \gamma$$
$$y_t = \phi_{20} + \phi_{21} y_{t-1} + ... + \phi_{2p_2} y_{t-p_2} + \epsilon_t, \quad y_{t-d} > \gamma$$

Here there may be regime dependent lag orders, and gaps in significant lags. Again we can re-paramaterize the model in terms of ϕ and δ coefficients:

$$y_t = \phi_0 + \phi_k y_{t-k} + (\delta_0 + \delta_k y_{t-k}) I(y_{t-d} > \gamma) + \epsilon_t, \quad k = \max p_1, p_2$$

These models are sometimes denoted $AR(2, p_1, p_2)$, where the first parameter represents the number of regimes.

The same search procedure through the the threshold variable, here defined $q = y_{t-d}$, can be applied for estimating threshold-autoregressive models (Hansen, 1997). However, in the case of an $AR(2, p_1, p_2)$ model, there is an additional complication to the conventional estimation method for regime-switching dynamics, namely that the lag order must first be

estimated. This is usually done by by estimating p from an AR model. Thus when fitting SETAR models, conventional methods have two model selection stages, at the beginning where p is chosen and at the end when tests for linearity take place.

Pitarakis (2006) shows that estimating the correct lag order p, under regime-switching dynamics presents serious challenges, and errors in estimation can have a significant impact on subsequent tests for linearity. Since overfitting is usually preferred to under-fitting, the AIC criterion is often used at this stage. But if the lag order is over estimated, then at the second stage there will be unimportant lags included. Including extra lags will lead to larger variation in parameter estimation, which can lead to a failure to reject the null in subsequent tests for linearity.

To avoid the pitfalls related to testing for regime-switching dynamics with an incorrect order specification, Pitarakis proposed fitting all $p_{max}(p_{max}+1)$ possible models, where p_{max} is the maximum lag order considered. He recommends the BIC criterion for this method, which has good power without systematically pointing to non-linearity as the AIC criterion does. However, this method still does not consider the possibility of different lag orders in each regime, nor does it address the problem of gaps in significant lags.

4 LASSO-TYPE PENALIZED ESTIMATION METHODS

Lasso-type penalized estimation methods perform simultaneous model estimation and variable selection by minimizing the usual residual sum of squares (RSS), subject to a constraint on the size of the estimated coefficients. Lasso puts a penalty on the sum of the absolute value of the coefficients.

$$Q(\beta) = RSS + \lambda||\boldsymbol{\beta}||_1$$

Use of the L_1 norm allows the Lasso to shrink some coefficients to zero, effectively eliminating them from the model. Here λ is a tuning parameter: $\lambda = 0$ corresponds to least squares estimation, and the larger λ becomes, the more coefficients shrink to zero. λ values are typically selected through a k-fold cross validation procedure. While the penalty term in the Lasso criterion introduces some bias in the estimator, this is usually balanced by large improvements in the variance.

The Lasso method was first introduced by Tibshirani in 1996, as an alternative to stepwise procedures for model selection based on information criteria such as AIC and BIC. Since then, a number of variations on Lasso have been proposed. Among these are the Elastic-net and Lad-lasso, which can outperform Lasso under certain conditions.

One draw-back to Lasso is that it tends to select one predictor from a set of strongly correlated predictors. The Elastic-net (E-net) was proposed by Zou and Hastie (2004) to improve on the Lasso, acting "like a stretchable fishing net that retains 'all the big fish'." The Elastic-net performs its variable selection by incorporating both the L_1 and the L_2 penalty:

$$Q(\beta) = RSS + (1-\alpha)\lambda||\boldsymbol{\beta}||_2 + \alpha\lambda||\boldsymbol{\beta}||_1 \quad \alpha \in [0,1]$$

E-net gives non-zero coefficients to significant variables even when they are correlated.

A second weakness of Lasso is that large outliers tend to have an exaggerated impact on estimation due to the squared term in the objective function. Lad-Lasso was developed as a robust estimation method, by combining least absolute deviation regression with Lasso-type

penalized estimation.

$$Q(\beta) = ||Y - \boldsymbol{X}\boldsymbol{\beta}||_1 + \lambda||\boldsymbol{\beta}||_1$$

Gao and Huang (2010) discuss sparsity conditions and conditions on the structure of the model matrix needed for the Lad-lasso to be consistent in estimation and selection. They suggest either the AIC or BIC criterion for selecting the tuning parameter. However Lad estimation methods are computationally expensive, and a number of alternative choices for λ have been proposed. There is no consensus on the best choice for λ. For this paper we used the R package 'flare', which generates a default vector of λ values. We chose three values for λ based on this package.

5 LASSO METHODS FOR THRESHOLD ESTIMATION

The purpose of this paper is to determine whether introducing Lasso methods into the iterative search procedure, in place of OLS, will reduce the variance of $\hat{\gamma}$. There are two reasons why the L_1 penalty could potentially improve upon OLS. First, it would perform variable selection at each stage of the search procedure. We expect that this variable selection will reduce the variance of $\hat{\gamma}$ when we include variables that do not belong in the final model or when a certain variable is important in one regime but drops out in the second. Along similar lines, for $AR(2, p_1, p_2)$, the methods we propose do not require the initial step of selecting an appropriate lag order, and should do well when there are gaps in significant lags.

Secondly, we chose to consider Elastic-net because we were interested in methods that perform well when variables are correlated, and Lad-lasso because it is robust to outliers and heavy tailed distributions. Both of these conditions are prevalent in time-series applications.

Besides the potential advantages we examine in this paper, there are some additional motivations for considering Lasso methods. Tests for non-linearity are usually based on an information criterion as in Pitarakis (2006). Since such criteria place a penalty on the number of parameters included in the model, Lasso-methods may improve the power of these tests. Or such tests may not be necessary since, in theory, Lasso-methods should shrink the $\boldsymbol{\delta}$ coefficients to 0 when the underlying model is linear.

Applying Lasso-methods in place of OLS is straightforward. We simply incorporate the appropriate penalty into the search procedure. The following equations correspond to Lasso, E-net, and Lad-lasso respectively

$$Q_i(\beta) = ||Y - (\boldsymbol{X\beta} + \boldsymbol{X\delta} \circ I(q > q_i))||_2 + \lambda||(\boldsymbol{\beta}, \boldsymbol{\delta})||_1$$

$$Q_i(\beta) = ||Y - (\boldsymbol{X\beta} + \boldsymbol{X\delta} \circ I(q > q_i))||_2 + (1-\alpha)\lambda||(\boldsymbol{\beta}, \boldsymbol{\delta})||_1 + \alpha\lambda||(\boldsymbol{\beta}, \boldsymbol{\delta})||_1$$

$$Q_i(\beta) = ||Y - (\boldsymbol{X\beta} + \boldsymbol{X\delta} \circ I(q > q_i))||_1 + \lambda||(\boldsymbol{\beta}, \boldsymbol{\delta})||_1$$

Finally, the threshold is estimated by minimizing Q_i.

6 SIMULATIONS

To evaluate the performance of our method we conducted two different simulation studies and considered various scenarios for each. In the first study, the response variable was generated from correlated predictors; insignificant variables were also included in the model matrix. We investigate the effect of sample size, the location of the threshold, γ, and the distributions of the error terms.

In the second study, we generate data from an underlying SETAR model. Here again the methods were evaluated under different error term distributions. For Lad-lasso we used three different tuning parameters based on how a $\boldsymbol{\lambda}$ vector of length three would be generated in 'flare.' We report them as Lad1, Lad2, and Lad3. Lad3 corresponds to smallest value of the tuning parameter: $\lambda = \sqrt{\log(p)/n}$. We present our results below.

CORRELATED VARIABLES

For the correlated variables study we simulated sample sizes of 100, 200, and 400. We generated a matrix $\boldsymbol{X}$ of correlated variables, where the covariance of the i^{th} and j^{th} columns is $.8^{|i-j|}$, $i \neq j$, with a threshold variable, $q \sim N(0,1)$ that does not belong to the group of covariates. We consider two different threshold values, 0.25, and 0.75. Furthermore, we consider three alternative error distributions, Normal(0,0.5), Laplace(0,0.5), and Student's t on 3 degrees of freedom. For each scenario our vectors of coefficients were $\boldsymbol{\beta} = (0.5, 1, 1.5, 2, 0, 0, 0, 0, 0)^T$, and $\boldsymbol{\delta} = (0.5, -1, -1, 0, 0, 0, 0, 0)^T$. The first term in the coefficient vectors correspond to the intercept. Thus in the second regime the intercept increases, X_1 drops out of the model completely, and the coefficient for X_2 is reduced. 500 trials were run.

For each trial we applied the same iterative search procedure in the threshold variable, first with OLS, then via unpenalized, Lad regression (Median Regression), followed by the Lasso, Elastic-net ($\alpha = 0.5$) and the Lad-lasso criteria.

We present some summary statistics of our results in Tables 1-3, corresponding to the different error distributions. Each table presents the empirical bias, median absolute devi-

ation (MAD), interquartile range (IQR), and the 5^{th} and 95^{th} percentiles of the threshold estimate, for all five methods.

Under the Normal distribution no method clearly out-performed the other: Lad3, had the best results regarding bias, however of all the methods it also had the highest MAD; this improved with sample size. OLS did slightly worse than other methods in terms of bias however this also improved with sample size. This is somewhat surprising since we usually expect OLS methods to be unbiased but to have relatively high variance, and Lasso to make up for bias with smaller variance. We did not notice any difference in the relative performance of these methods, when the threshold was on the extreme (.75).

The story is similar for heavy tailed distributions. Though on the whole Median regression did better than OLS, it is not clear that penalized estimation methods can improve threshold estimation, at least of small sample sizes. We again notice less bias in the Lad3 estimator, but usually at the expense MAD and IQR. For Student errors, when $\gamma = .75$, robust methods, Median and Lad3, outperform other methods.

This simulation demonstrates the importance of sample size in selecting a value for λ in Lad-lasso. For $n = 100$, Median regression, which corresponds to $\lambda = 0$ performed better, but as n increased Lad3 began to outperform Median. For an interesting discussion on an appropriate choice of tuning parameter see Wang (2013). Wang recommends that the penalty should be $\lambda = \sqrt{2n\log(p)}$ for larger sample sizes, however he also notes that $if\, ||X_j||_1 < \lambda, then\, \hat{\beta}_j = 0$. Which means that if the tuning parameter is too large, even significant variables may be eliminated from the model. We found these penalties were too large for our estimation problem, which may be related to the sparsity of our model matrix when fitting models for extreme values of $\boldsymbol{qs}$.

Table 1: Normal(0,.5) Errors

		$\gamma = .25$					$\gamma = .75$				
		Bias	MAD	IQR	5%	95%	Bias	MAD	IQR	5%	95%
	OLS	-0.0189	0.0327	0.0432	0.1394	0.3063	-0.0303	0.0434	0.0677	0.5974	0.8310
	MEDIAN	-0.0210	0.0338	0.0440	0.1266	0.3091	-0.0319	0.0500	0.0693	0.5906	0.8409
$n = 100$	LASSO	-0.0200	0.0356	0.0467	0.1246	0.3140	-0.0322	0.0508	0.0704	0.5900	0.8540
	ENET	-0.0181	0.0378	0.0484	0.1394	0.3100	-0.0296	0.0506	0.0697	0.5945	0.8399
	LAD1	-0.8966	0.4452	0.7908	-1.2762	0.4737	-1.4443	0.3566	0.7345	-1.2762	0.4060
	LAD2	-0.6088	0.6387	0.8781	-1.1177	0.4228	-1.0724	0.6093	0.8158	-1.0886	0.7658
	LAD3	0.0043	0.0533	0.0734	0.1415	0.4157	-0.0182	0.0692	0.0940	0.5807	0.9126
	OLS	-0.0111	0.0119	0.0189	0.1930	0.2811	-0.0146	0.0226	0.0330	0.6815	0.7940
	MEDIAN	-0.0102	0.0134	0.0188	0.1965	0.2811	-0.0160	0.0216	0.0331	0.6814	0.7905
$n = 200$	LASSO	-0.0093	0.0133	0.0188	0.2011	0.2811	-0.0149	0.0237	0.0331	0.6769	0.7968
	ENET	-0.0101	0.0139	0.0193	0.1958	0.2854	-0.0155	0.0216	0.0340	0.6780	0.7922
	LAD1	-0.8454	0.3701	0.8193	-1.1652	0.6337	-1.3077	0.3670	0.9047	-1.1618	0.7629
	LAD2	-0.3879	0.5405	0.7968	-0.9680	0.5689	-0.7512	1.0452	1.3168	-0.9857	0.8479
	LAD3	-0.0025	0.0181	0.0243	0.1949	0.3228	-0.0068	0.0307	0.0414	0.6695	0.8312
	OLS	-0.0057	0.0072	0.0104	0.2239	0.2633	-0.0051	0.0089	0.0122	0.7159	0.7746
	MEDIAN	-0.0062	0.0081	0.0109	0.2228	0.2631	-0.0049	0.0093	0.0125	0.7146	0.7743
$n = 400$	LASSO	-0.0054	0.0085	0.0111	0.2220	0.2662	-0.0051	0.0095	0.0123	0.7138	0.7742
	ENET	-0.0046	0.0078	0.0104	0.2253	0.2671	-0.0051	0.0098	0.0129	0.7131	0.7735
	LAD1	-0.6569	0.7090	1.0317	-1.1163	0.8731	-1.2362	0.3228	0.9665	-1.1323	1.0045
	LAD2	-0.0963	0.0413	0.0717	-0.7453	0.4111	-0.1680	0.0288	0.0373	-0.7114	0.8388
	LAD3	-0.0029	0.0097	0.0126	0.2258	0.2704	0.0003	0.0131	0.0185	0.7155	0.7934

Table 2: Laplace(0,0.5) Errors

		$\gamma = .25$					$\gamma = .75$				
		Bias	MAD	IQR	5%	95%	Bias	MAD	IQR	5%	95%
$n = 100$	OLS	-0.0203	0.0463	0.0599	0.1152	0.3706	-0.0371	0.0584	0.0833	0.5175	0.8610
	MEDIAN	-0.0229	0.0446	0.0598	0.1095	0.3520	-0.0361	0.0580	0.0768	0.5147	0.8651
	LASSO	-0.0233	0.0550	0.0707	0.0874	0.3730	-0.0451	0.0745	0.1031	0.4718	0.8803
	ENET	-0.0219	0.0518	0.0693	0.0887	0.3683	-0.0436	0.0759	0.1028	0.4851	0.8879
	LAD1	-0.8967	0.4107	0.7729	-1.2811	0.5765	-1.4179	0.3929	0.7765	-1.2915	0.4912
	LAD2	-0.6133	0.6705	0.8996	-1.1109	0.5281	-1.0712	0.6146	0.8222	-1.1257	0.8075
	LAD3	0.0155	0.0656	0.0939	0.1198	0.5147	-0.0289	0.0806	0.1079	0.5055	0.9400
$n = 200$	OLS	-0.0102	0.0188	0.0233	0.1777	0.2923	-0.0171	0.0277	0.0344	0.6509	0.8023
	MEDIAN	-0.0082	0.0178	0.0233	0.1874	0.2959	-0.0148	0.0258	0.0317	0.6613	0.7997
	LASSO	-0.0110	0.0216	0.0278	0.1702	0.2989	-0.0158	0.0298	0.0385	0.6423	0.8082
	ENET	-0.0083	0.0224	0.0300	0.1755	0.3059	-0.0145	0.0296	0.0408	0.6601	0.8054
	LAD1	-0.8260	0.4265	0.8073	-1.1679	0.6671	-1.3742	0.3230	0.7077	-1.1717	0.7162
	LAD2	-0.5014	0.6021	0.8128	-1.0426	0.5223	-0.9354	0.6291	0.9453	-1.0221	0.8196
	LAD3	-0.0013	0.0228	0.0302	0.1808	0.3292	-0.0054	0.0360	0.0487	0.6541	0.8542
$n = 400$	OLS	-0.0048	0.0078	0.0103	0.2173	0.271	-0.0057	0.0110	0.0152	0.7102	0.777
	MEDIAN	-0.0056	0.0083	0.0104	0.2175	0.2681	-0.0055	0.0102	0.0141	0.7104	0.7779
	LASSO	-0.0049	0.0098	0.0132	0.2133	0.2766	-0.0059	0.0134	0.0174	0.7053	0.7821
	ENET	-0.0051	0.0103	0.0137	0.2137	0.2716	-0.0056	0.0121	0.0162	0.7039	0.7871
	LAD1	-0.6778	0.5907	0.9759	-1.127	0.9136	-1.1874	0.4597	1.0292	-1.1292	0.9566
	LAD2	-0.2408	0.1447	0.5136	-0.904	0.4697	-0.3459	0.0594	0.7824	-0.8843	0.8308
	LAD3	-0.0026	0.0111	0.0148	0.2195	0.2809	-0.0019	0.0144	0.0194	0.7116	0.8041

Table 3: Student(3) Errors

		$\gamma = .25$					$\gamma = .75$				
		Bias	MAD	IQR	5%	95%	Bias	MAD	IQR	5%	95%
$n = 100$	OLS	-0.0990	0.2060	0.2922	-0.9084	0.8275	-0.3742	0.3221	0.6293	-0.9589	1.0319
	MEDIAN	-0.0327	0.1324	0.1831	-0.3208	0.7094	-0.1754	0.1735	0.2556	-0.5253	1.0045
	LASSO	-0.0925	0.2309	0.3080	-0.7397	0.7718	-0.3086	0.2980	0.5146	-0.7995	1.0055
	ENET	-0.0725	0.2075	0.2882	-0.7104	0.8202	-0.3194	0.2750	0.5346	-0.8179	1.0054
	LAD1	-0.9072	0.3993	0.8056	-1.2652	0.5401	-1.4206	0.4061	0.7240	-1.2721	0.4669
	LAD2	-0.6770	0.5555	0.7363	-1.1244	0.6279	-1.1895	0.5394	0.7660	-1.1571	0.6730
	LAD3	0.0092	0.3123	0.4342	-0.7611	0.9677	-0.2308	0.2359	0.3591	-0.7316	1.0461
$n = 200$	OLS	-0.0334	0.0632	0.0871	-0.0716	0.4616	-0.1531	0.1056	0.1690	-0.4517	0.9501
	MEDIAN	0.0003	0.0457	0.0614	0.0904	0.4113	-0.0364	0.0735	0.1018	0.4651	0.9029
	LASSO	-0.0130	0.0851	0.1171	-0.0298	0.5257	-0.1154	0.1061	0.1727	-0.0013	0.9141
	ENET	-0.0210	0.0789	0.1061	-0.0765	0.4888	-0.1106	0.1086	0.1588	0.0266	0.9438
	LAD1	-0.7975	0.4776	0.8739	-1.1769	0.7247	-1.3808	0.3464	0.7926	-1.1885	0.5973
	LAD2	-0.6727	0.5396	0.7440	-1.0437	0.4908	-1.1883	0.5497	0.7581	-1.0760	0.4981
	LAD3	0.0557	0.0686	0.1020	0.1070	0.7199	-0.0464	0.0806	0.1077	0.3973	0.9688
$n = 400$	OLS	-0.0025	0.0263	0.0343	0.1471	0.3734	-0.0413	0.0426	0.0567	0.4801	0.8921
	MEDIAN	-0.0021	0.0193	0.0267	0.1700	0.3160	-0.0068	0.0274	0.0368	0.6503	0.8309
	LASSO	-0.0028	0.0328	0.0427	0.1522	0.3592	-0.0387	0.0447	0.0611	0.4807	0.8795
	ENET	-0.0012	0.0292	0.0405	0.1433	0.3637	-0.0377	0.0481	0.0631	0.5042	0.8866
	LAD1	-0.7338	0.4820	0.9371	-1.1353	0.8455	-1.2646	0.3258	0.9577	-1.1368	0.8597
	LAD2	-0.6306	0.5363	0.7585	-1.0310	0.5878	-1.1400	0.5072	0.6835	-1.0429	0.7207
	LAD3	0.0072	0.0260	0.0372	0.1736	0.3542	0.0013	0.0371	0.0512	0.6384	0.8808

SETAR

For the SETAR simulation study, we generated a time-series of 300 steps in each trial. Our threshold variable was the difference between lag 1 and lag 2 in absolute terms with $\gamma = 1$. There were two motivations for this choice of a threshold variable. First, it is not unreasonable to expect an increase in volatility to trigger a regime switch. Secondly, for simulation purposes, this choice helped to ensure that values reasonably close to the true threshold occurred within the the search vector. Our underlying model was as follows.

$$y_t = 0.4y_{t-1} - .5y_{t-2} + .3y_{t-3} + \epsilon_t, \qquad |y_{t-1} - y_{t-2}| \leq 1$$

$$y_t = 1 + .3y_{t-1} + .5y_{t-2} - .3y_{t-3} - .5y_{t-13} + \epsilon_t, \qquad |y_{t-1} - y_{t-2}| > 1$$

Thus in the first regime the lag order was 3, while in the second regime lag 2 and 3 drop out of the model, and lag 13 enters the model. This last choice was motivated by empirical findings based on monthly data where terms related to seasonality may be present in one regime but not in the other. Coefficients in each regime were chosen to satisfy conditions for stationarity. Here we also considered three random noise scenarios: Normal(0,0.5), Laplace(0,0.5), Student(3).

We used the same iterative search procedure through the threshold variable, minimizing the criterion for each method: OLS, Median, Lasso, Elastic-net($\alpha = 0.5$), and Lad1, Lad2, and Lad3. We ran 1,000 trials.

Figure 1 shows box-plots of $\hat{\gamma}$ for each of the methods under the three different error distributions. The red horizontal line represents the true breakpoint, γ. Figure 2 shows the same box-plots, but with the outliers removed for a better visual.

Figure 1: Boxplots SETAR

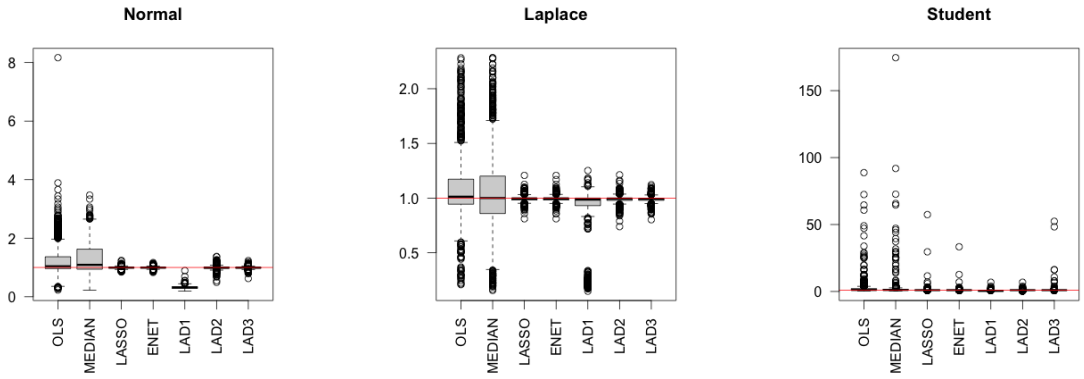
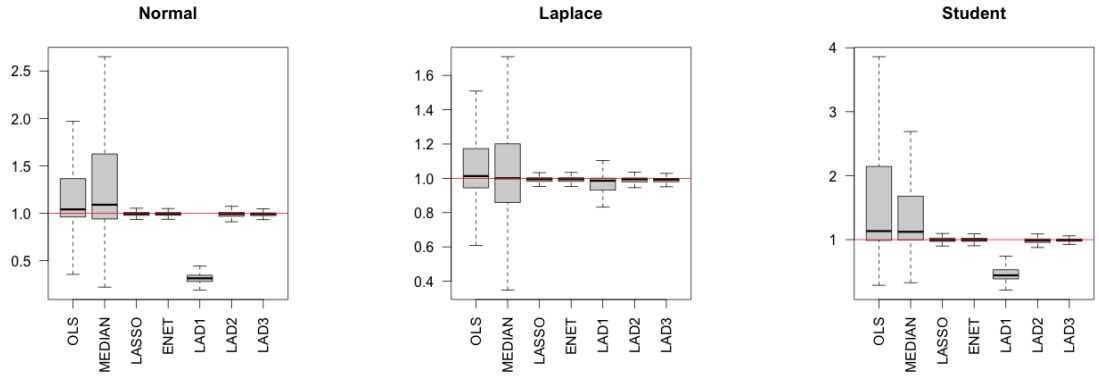


Figure 2: Boxplots SETAR: A Closer Look



In this simulation our results were much more striking. Lasso methods clearly out perform OLS and Median regression, in variance, and with the exception of Lad3, improve in terms of bias as well. We also observe, with the exception of the student errors that the empirical distribution of $\hat{\gamma}$ are symmetric and centered about γ.

Table 4 reports the empirical bias, MAD, IQR and the 5^{th} and 95^{th} percentiles for $\hat{\gamma}$, for each of the six methods. These summaries reflect the results observed from the box plots. Lasso and Elastic-net clearly outperformed, both in terms of bias and measures of spread, for Normal and Laplace. For Student(3), Lad2 and Lad3 had the best performance, which is what we expected to see since large disturbances can occur under a Student distribution.

While Lad3 was biased compared to the other Lasso methods, this was due to a few large outliers, and it had the smallest MAD.

Table 4: Summaries

ϵ	Method	Bias	MAD	IQR	5%	95%
Normal	OLS	0.2408	0.2097	0.4042	0.6778	2.3800
	MEDIAN	0.2903	0.3803	0.6850	0.4864	2.4636
	LASSO	-0.0051	0.0223	0.0299	0.9413	1.0557
	ENET	-0.0058	0.0215	0.0286	0.9433	1.0476
	LAD1	-0.6818	0.0481	0.0650	0.2433	0.4045
	LAD2	-0.0099	0.0316	0.0417	0.8916	1.0841
	LAD3	-0.0097	0.0210	0.0290	0.9299	1.0471
Laplace	OLS	0.0855	0.1531	0.2285	0.6337	1.7842
	MEDIAN	0.0463	0.2456	0.3422	0.3578	1.8460
	LASSO	-0.0060	0.0149	0.0202	0.9532	1.0375
	ENET	-0.0062	0.0153	0.0208	0.9519	1.0350
	LAD1	-0.1292	0.0343	0.0692	0.2147	1.0493
	LAD2	-0.0098	0.0176	0.0231	0.9375	1.0366
	LAD3	-0.0105	0.0136	0.0196	0.9414	1.0325
Student	OLS	1.4773	0.3732	1.1566	0.7023	5.9364
	MEDIAN	1.4721	0.2806	0.6867	0.7708	3.7727
	LASSO	0.1388	0.0358	0.0495	0.9080	1.2092
	ENET	0.0769	0.0336	0.0469	0.9033	1.2161
	LAD1	-0.4637	0.0959	0.1430	0.3269	1.0510
	LAD2	-0.0131	0.0407	0.0544	0.6219	1.1518
	LAD3	0.1745	0.0249	0.0333	0.9271	1.1268

Table 5 reports type I and type II errors for the Lasso methods. Type I error is the percentage of unimportant lags, on average, selected by each method out of the total number of unimportant lags that could have been selected. Type II error represents the percentage of important lags, on average, that each method failed to select. We were interested in how Lasso, Elastic-net and Lad-Lasso compared in "balancing" the trade off between type I and II errors.

Both Lasso and E-net were more conservative in dropping variables from the model, while the percentages for type II errors were large but not compared to the percentages for Lad-lasso in type I. Lad3 was the only method to offer a comparable balance between type I and type II. It is also clear from this table that the first choice of a tuning parameter was

too large of a penalty since the method eliminated all of the important variables almost 100 percent of the time.

Table 5: Type I and II Errors

Type	ϵ	LASSO	ENET	LAD1	LAD2	LAD3
I	Normal	15.40	13.34	100.0	99.17	72.48
	Laplace	13.77	10.75	90.6	80.09	63.84
	Student	30.18	28.55	100.0	95.46	65.86
II	Normal	79.12	82.11	4.17	10.17	22.83
	Laplace	74.48	79.51	7.30	14.51	22.47
	Student	67.94	70.11	4.37	12.50	21.87

In both simulation studies, results varied substantially between Lad1, Lad2 and Lad3. Because Lad-lasso uses least absolute deviation, selecting the appropriate λ by cross-validation is computationally expensive. We believe that with further investigation results for Lad-lasso could be improved. Consideration should also be given to adaptive Lad-lasso, which uses a vector of tuning parameters $\boldsymbol{\lambda}$ that are weighted by previous values of the estimates for the coefficients. Wang et al. (2007) recommend $\lambda_j = \log(n)/(n|\hat{\beta}_j|)$, where $\hat{\beta}_j$ is the unpenalized Lad estimate for the j^{th} predictor, and show that Lad-lasso is $\sqrt{n}$-consistent without requiring any moment assumptions on the error terms.

7 MODELING URBAN TRAFFIC VOLUMES

Here we review some results after applying our method to data on traffic volumes in an urban network. The data was adjusted for seasonality by removing the weakly profiles. Measurements were taken at three minute time intervals.

Inference began with a Keenan's test for linearity (P-value<0.000). The next step was to identify the threshold lag. This was done my fitting SETAR models, on a small portion of the data set using the conventional OLS method up to lag 10. The threshold lag that produced a model with the smallest residual sum of squares was selected for further analyses. This procedure lead us to select lag 9 as the threshold variable.

We used the first three weeks to fit our models, and the following 4,571 measurements as testing data to evaluate each method. For comparison, we also fit an auto-regressive integrated moving average (ARIMA(p,d,q)) model. The latter assumes that current observation are determined by a linear combination of p previous observations and q previous disturbances, where d is a difference parameter, which helps to remove non-stationarity. Time-series with regime-switching dynamics behave similarly to non-stationary time-series. ARIMA estimation can perform well as a linear alternative to SETAR estimation.

Table 6 displays the chosen threshold for each of the estimated SETAR models. We also include the proportion of observations that fell in the second regime as defined by each model. In Table 7 we report the root mean squared prediction error (RMSE), mean absolute prediction error (MAE), and the median absolute prediction error (MedAPE). Lasso-methods outperformed OLS. However Lad-Lasso was the only non-linear method to outperform ARIMA. It is interesting to note that $d = 0$, indicating that differencing did not improve the fit.

Table 6: Estimated Threshold

	OLS	LASSO	ENET	LAD
$\hat{\gamma}$	1.21	11.53	11.53	3.62
$\hat{P}$	.56	.25	.25	.42

Table 7: Out of Sample Statistics

	OLS	LASSO	ENET	LAD	ARIMA
RMSE	12.9716	11.9623	11.9621	11.8430	11.9012
MAE	8.6134	7.5823	7.5858	7.3718	7.5333
MedAPE	5.8272	4.3315	4.3309	4.0517	4.3462

Table 8: Tests for Differences in Prediction Accuracy

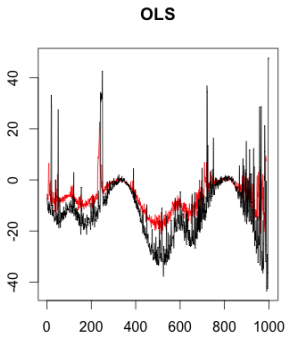
	LASSO	ENET	LAD	ARIMA
OLS	0.000000	0.000000	0.000000	0.000000
LASSO		0.710471	0.011894	0.001405
ENET			0.012028	0.001123
LAD				0.273875

We were also interested in testing whether there was a difference in the prediction accuracy of our methods. For this analyses, we used the well-known Diebold-Mariano (DM) test. Table 8 displays the matrix of (two-sided) comparison tests. While the difference between OLS and all other methods was highly significant, there were no significant differences between Lasso and E-net, nor between Lad-lasso and ARIMA. A possible explanation is that ARIMA and Lad-lasso have more in common, in the sense that they were the two most parsimonious models, while Lasso and Elastic-net retained more lags in each regime.

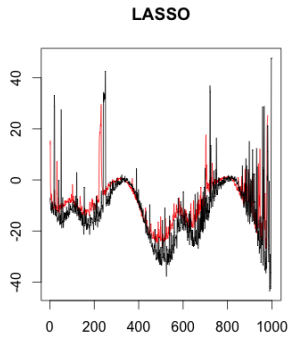
The plots in Figure 3 show a section of the out-of-sample time-series (black), and the one-step-ahead predictions (red) for each method.

Figure 3: Predicted Values

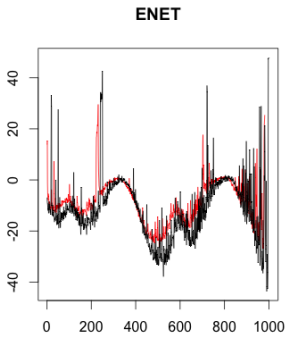
(a) OLS



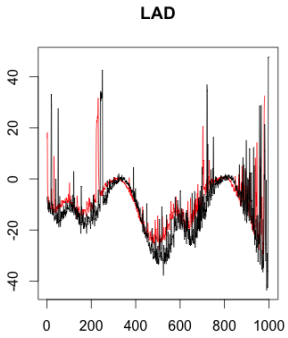
(b) LASSO



(c) ENET



(d) LAD



(e) ARIMA

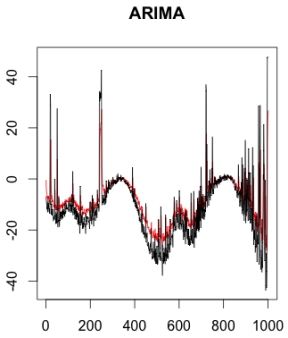
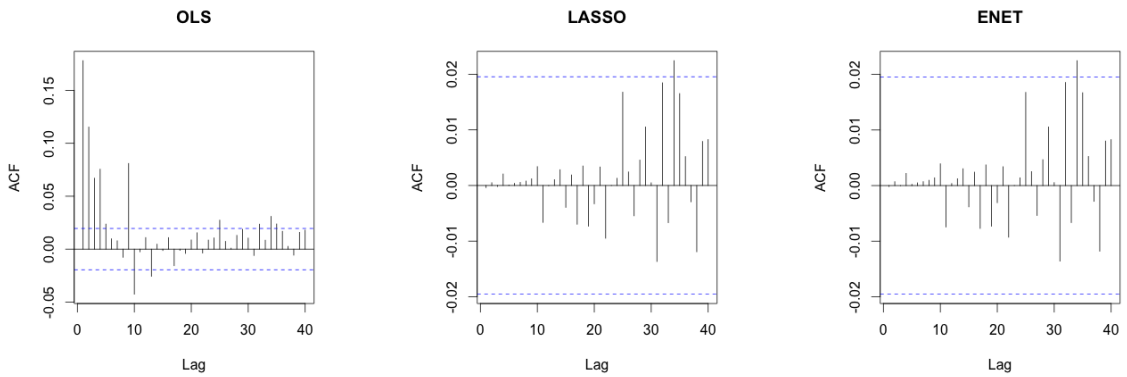


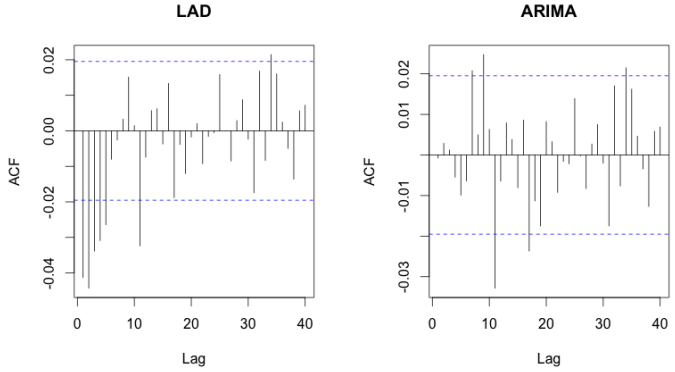
Figure 4 shows the autocorrelation functions (ACF) of the standardized residuals. If the correct model has been specified we would expect to see estimated correlations close to zero. Lasso and E-net provide the most satisfactory results, followed by ARIMA. The plots for the OLS and Lad-lasso models indicate left over serial correlation among the error terms.The reason for the discrepancy between the residual analyses and the out-of-sample performance of Lad-lasso, may lie with the choice of λ. A smaller tuning parameter could provide a better balance, by retaining important lags while avoiding overfitting.

Figure 4: ACF for Standardized Residuals

(a) OLS (b) LASSO (c) ENET



(d) LAD (e) ARIMA



Below we provide the estimated models for each method. The Lasso and E-net models kept the same lags in the model. All of the Lasso methods estimated lag orders up to 20 in the first regime. In the second regime, Lasso and E-net estimated change in slope coefficients up to lag 18, and did not include a change in intercept coefficient. Lad3 on the other hand included a change in intercept, while it only estimated a change in slope up to lag 5. The estimated ARIMA model combined the first two lags and the two previous error disturbances, without differencing.

$$
\begin{aligned}
\text{OLS:}\quad y_t = &-0.06 + 0.23y_{t-1} + 0.07y_{t-2} + 0.07y_{t-3} - 0.05y_{t-4} + 0.06y_{t-5} + 0.01y_{t-6} \\
&+ 0.08y_{t-7} + 0.06y_{t-8} - 0.22y_{t-9} + 0.26y_{t-10} - 0.06y_{t-11} - 0.03y_{t-12} + 0.08y_{t-13} \\
&+ (0.62 + 0.10y_{t-1} + 0.03y_{t-2} - 0.03y_{t-3} + 0.08y_{t-4} - 0.05y_{t-5} + 0.04y_{t-6} \\
&- 0.06y_{t-7} - 0.02y_{t-8} + 0.23y_{t-9} - 0.29y_{t-10} + 0.07y_{t-11} + 0.07y_{t-12} - 0.07y_{t-13})I(y_{t-9} > 1.21) + \epsilon_t
\end{aligned}
$$

$$
\begin{aligned}
\text{LASSO:}\quad y_t = &0.43 + 0.44y_{t-1} + 0.12y_{t-2} + 0.03y_{t-3} + 0.01y_{t-4} + 0.01y_{t-5} + 0.01y_{t-6} + 0.04y_{t-7} \\
&+ 0.01y_{t-8} + 0.04y_{t-9} + 0.02y_{t-10} - 0.003y_{t-11} + 0.02y_{t-13} + 0.004y_{t-14} + 0.01y_{t-16} \\
&- 0.01y_{t-17} + .001y_{t-18} + 0.02y_{t-20} \\
&(0.013y_{t-1} - 0.01y_{t-2} + 0.02y_{t-3} + 0.03y_{t-4} + 0.02y_{t-5} - 0.02y_{t-7} \\
&- 0.04y_{t-10} - 0.05y_{t-11}0.04y_{t-12} - 0.01y_{t-16} - 0.01y_{t-17} - 0.01y_{t-18})I(y_{t-9} > 11.52) + \epsilon_t
\end{aligned}
$$

$$
\begin{aligned}
\text{ENET:}\quad y_t = &0.42 + 0.44y_{t-1} + 0.12y_{t-2} + 0.04y_{t-3} + 0.01y_{t-4} + 0.01y_{t-5} + 0.01y_{t-6} + 0.04y_{t-7} \\
&+ 0.01y_{t-8} + 0.04y_{t-9} + 0.02y_{t-10} - 0.003y_{t-11} + 0.02y_{t-13} + 0.004y_{t-14} \\
&+ 0.01y_{t-16} - 0.01y_{t-17} + 0.0002y_{t-18} + 0.02y_{t-20} \\
&+ (0.01y_{t-1} - 0.01y_{t-2} + 0.02y_{t-3} + 0.03y_{t-4} + 0.02y_{t-5} \\
&- 0.02y_{t-7} - 0.03y_{t-10} - 0.05y_{t-11} + 0.04y_{t-12} - 0.01y_{t-16} - 0.01y_{t-17} - 0.01y_{t-18})I(y_{t-9} > 11.52) + \epsilon_t
\end{aligned}
$$

$$
\begin{aligned}
\text{LAD:}\quad y_t = &-0.004 + 0.49y_{t-1} + 0.14y_{t-2} + 0.05y_{t-3} + 0.03y_{t-4} + 0.0005y_{t-5} + 0.005y_{t-6} + 0.03y_{t-7} \\
&+ 0.01y_{t-10} + 0.01y_{t-12} + 0.01y_{t-13} + 0.02y_{t-20} \\
&+ (0.55 + 0.01y_{t-1} + 0.04y_{t-5})I(y_{t-9} > 3.62) + \epsilon_t
\end{aligned}
$$

$$
\text{ARIMA:}\quad y_t = 1.22 + 1.52y_{t-1} - 0.54y_{t-2} - 1.07e_{t-1} + 0.17e_{t-2} + \epsilon_t
$$

While tests for non-linearity were highly significant, the estimated ARIMA model provides a better fit, based on out-of-sample performance and residual analyes, than the conventional, OLS method for estimating a SETAR model. One explanation for why the linear model outperformed the SETAR model when we used the conventional method is that OLS cannot perform variable selection. This reasoning is re-enforced by the improvements we observe when Lasso-type penalties are introduced into SETAR estimation.

8 CONCLUSION

The goal of this paper was to investigate how Lasso-type methods would perform compared to OLS in threshold estimation. We considered two different frameworks: correlated variables and SETAR. Because Lasso-type methods can perform variable selection, we expected to see a reduction in the variance of $\hat{\gamma}$ when there were extra variables included in the estimation. We also were interested in how our methods would perform when the random errors had heavy-tailed distributions.

Our findings from the first simulation study suggest that Lasso-methods, in particular Lad-lasso, can improve the bias of the threshold estimate. However, in measures of spread, large sample sizes are necessary for our methods to compare favorably to the least squares method. This result is surprising since we expected to see an improvement in the variance of $\hat{\gamma}$, possibly at the expense of some bias. We also see the need for a careful choice of λ in Lad-lasso. For the second simulation study, our results were much more definitive. Introducing the Lasso penalty improved threshold estimation dramatically.

In time-series analyses, there may be evidence in favor of a regime-switching model. However, in practice the conventional method for estimating a SETAR model can be unsatisfactory compared to fitting a more parsimonious, ARIMA model. Our application to modeling traffic volumes suggest that estimating a SETAR model using Lasso-type methods not only improve upon OLS, but offer comparable results to an ARIMA model.

In the future we will study the impact of Lasso methods on tests for non-linearity. By eliminating unnecessary variables or lags, Lasso methods could improve the power of these tests. Or, it is possible that by shrinking the δ coefficients to 0 when the true underlying model is linear, these tests can be avoided altogether.

References

[1] Gao, X. and J. Huang (2010). Asymptotic Analysis of High-Dimensional Lad Regression with Lasso. *Statistica Sinica, 20*, 1485-1506.

[2] Hansen, B. (1997). Inference in TAR Models. *Studies in Nonlinear Dynamics and Econometrics, 2*(1), 1-14.

[3] Hansen, B. (2011). Threshold autoregression in economics. *Statistics and Its Interface,4*, 123-127.

[4] Hansen, B. (2011). Estimation and Inference in Threshold Type Regime Switching Models. *Handbook of Research Methods and Applications in Empirical Macroeconomics*, 189-204.

[5] Lee, S. Seo, M. H., and Y, Shin (2014). The Lasso for High-Dimensional Regression with a Possible Change-Point. arXiv:1209.4875v4.

[6] Nardi, Y. and A. Rinaldo (2010). Autoregressive Process Modeling via the Lasso Procedure. *Journal of Multivariate Analysis, 102*(3), 528-549.

[7] Pitarakis, J.Y. (2006). Model Selection Uncertainty and Detection of Threshold Effects. *Studies in Nonlinear Dynamics and Econometrics,10*1-25.

[8] Gonzalo, J. and J.Y. Pitarakis (2013). Estimation and Inference in Threshold Type Regime Switching Models. *Handbook of Research Methods and Applications in Empirical Macroeconomics*, 189-204.

[9] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological),58*(1),267-288.

[10] Wang, H. Li, G. and G. Jiang (2007). Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso. *Journal of Business & Economic Statistics, 25*(3), 347-355.

[11] Wang, L. (2013). The L_1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis, 120*, 135-151.

[12] Zou, H. and T. Hastie (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology),67*(2), 301-320.