AARHUS UNIVERSITY
BUSINESS AND SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS AND BUSINESS

CREATES
Center for Research in Econometric Analysis of Time Series

On the Oracle Property of the Adaptive LASSO in Stationary and Nonstationary Autoregressions

Anders Bredahl Kock

CREATES Research Paper 2012-05

Department of Economics and Business
Aarhus University
Bartholins Allé 10
DK-8000 Aarhus C
Denmark

Email: oekonomi@au.dk
Tel: +45 8716 5515

ON THE ORACLE PROPERTY OF THE ADAPTIVE LASSO IN STATIONARY AND NONSTATIONARY AUTOREGRESSIONS

ANDERS BREDAHL KOCK
AARHUS UNIVERSITY AND CREATES

Abstract. We show that the Adaptive LASSO is oracle efficient in stationary and non-stationary autoregressions. This means that it estimates parameters consistently, selects the correct sparsity pattern, and estimates the coefficients belonging to the relevant variables at the same asymptotic efficiency as if only these had been included in the model from the outset. In particular this implies that it is able to discriminate between stationary and non-stationary autoregressions and it thereby constitutes an addition to the set of unit root tests.

However, it is also shown that the Adaptive LASSO has no power against shrinking alternatives of the form c/T where c is a constant and T the sample size if it is tuned to perform consistent model selection. We show that if the Adaptive LASSO is tuned to performed conservative model selection it has power even against shrinking alternatives of this form.

Monte Carlo experiments reveal that the Adaptive LASSO performs particularly well in the presence of a unit root while being at par with its competitors in the stationary setting.

Keywords: Adaptive LASSO, Oracle efficiency, Consistent model selection, Conservative model selection, autoregression, shrinkage.

AMS 2000 classification: 62F7, 62F10, 62F12, 62J07

JEL classification: C13, C22

1. Introduction

Variable selection in high-dimensional systems has received a lot of attention in the statistics literature in the recent 10-15 years or so and it is also becoming increasingly popular in econometrics. As traditional computational methods are computationally infeasible if the number of covariates is large, focus has been on penalized or shrinkage type of estimators of which the most famous is probably the LASSO of Tibshirani (1996). This paper sparked a flurry of research in the theoretical properties of LASSO-type estimators, the first of which were Knight and Fu (2000). Subsequently, many other shrinkage estimators have been analyzed: the SCAD of Fan and Li (2001), the Bridge and Marginal Bridge Estimator in Huang et al. (2008), the Dantzig selector of Candes and Tao (2007) and the Sure Independence Screening of Fan and Lv (2008) to mention just a few. For a recent review with particular emphasis on the LASSO see Bühlmann and Van De Geer (2011). The focus of these papers is to establish the so-called oracle property for the proposed estimators. This entails showing that the estimators are consistent, perform correct variable selection and establishing that the limiting distribution of the non-zero coefficients is the same as if only the

Date: February 2, 2012.

Part of this research was carried out while I was visiting the Australian National University. I wish to thank Tue Gørgens and the Research School of Economics for inviting me and creating a pleasant environmnent. I am also indebted to Svend Erik Graversen and Jørgen Hoffmann Jørgensen for help. I wish to thank CREATES, funded by the Danish National Research Foundation, for providing financial support.

relevant variables had been included in the model. Put differently, the inference is as efficient as if an oracle had revealed the true model to us and estimation had been carried out using only the relevant variables.

Most focus in the statistics literature has been on establishing the oracle property for cross sectional data. An exception is Wang et al. (2007) who consider the LASSO for stationary autoregressions while Kock (2012) has shown that the oracle efficiency of the Bridge and Marginal Bridge estimator carry over to linear random and fixed effects panel data settings.

In this paper we show that the Adaptive LASSO of Zou (2006) possesses the oracle property in stationary as well as nonstationary autoregressions. We focus on the Adaptive LASSO since the original LASSO is only oracle efficient under rather restrictive assumptions which exclude too high dependence – an assumption which is unlikely to be satisfied in time series models.

We shall consider a model of the form

$$\Delta y_t = \rho^* y_{t-1} + \sum_{j=1}^{p} \beta_j^* \Delta y_{t-j} + \epsilon_t \tag{1}$$

which is sometimes called a Dickey-Fuller regression. ϵ_t is the error term to be discussed further in the next section. (1) is said to have a unit root if $\rho^* = 0$. When testing for a unit root, one usually first determines the number of lagged differences (p) to be included. This can be done either by information criteria, or modifications hereof, Ng and Perron (2001). Having selected the lags one tests whether $\rho^* = 0$. The oracle efficient estimators create new possibilities of carrying out such tests since testing for a unit root is basically a variable selection problem: Is y_{t-1} to be left out of the model ($\rho^* = 0$), or not? Hence, establishing the oracle property for the Adaptive LASSO means that we can choose the number of lagged differences to be included (and leaving out irrelevant intermediate lags) and test for a unit root at the same time. Knight and Fu (2011) made this point and have used it to construct a unit root test based on the Bridge Estimator in the setting we shall call conservative model selection.

We show: (i) The Adaptive LASSO possesses the oracle property in stationary and nonstationary autoregressions. (ii) Carry out out detailed finite sample and local to unity analysis in the stationary, nonstationary and local to unity setting. The local to unity setting reveals that the Adaptive LASSO is not exempt from the critique by Leeb and Pötscher (2005, 2008) of consistent model selection techniques. (iii) This problem, due to nonuniformity in the asymptotics, can be alleviated if one is willing to use tune the Adaptive LASSO to perform conservative model selection instead of consistent model selection[1]. The properties of conservative model selection are investigated in the stationary as well as the nonstationary setting.

The plan of the paper is as follows. Section 2 introduces the Adaptive LASSO and some notation. Section 3 states the oracle theorems for the Adaptive LASSO for stationary and nonstationary autoregressions while Section 4 carries out a detailed finite sample and local analysis under various settings. Section 5 considers the properties of the Adaptive LASSO when tuned to perform conservative model selection, 6 contains some Monte Carlos and Section 7 concludes. All proofs are deferred to the Appendix.

[1] We shall make mathematically precise definitions of consistent and conservative model selection in Section 2.

2. Setup and Notation

The most famous shrinkage estimator is without doubt the LASSO – the Least Absolute Shrinkage and Selection Operator. Its popularity is due to the fact that it carries out variable selection and parameter estimation in one step. However, it has been shown that the LASSO is only Oracle efficient under rather strict conditions, see Meinshausen and Bühlmann (2006), Zhao and Yu (2006) and Zou (2006) which don't allow too high correlations between covariates. This motivates using other procedures for variable selection than the LASSO. In particular, the problem of the LASSO is that it penalizes all parameters equally. Hence, Zou (2006) proposed the Adaptive LASSO which applies more intelligent data-driven penalization and proves that it is oracle efficient in a fixed regressor setting. In our context the Adaptive LASSO is defined as the minimizer of the following objective function.

$$\Psi_T(\rho,\beta) = \sum_{t=1}^{T}\left(\Delta y_t - \rho y_{t-1} - \sum_{j=1}^{p}\beta_j \Delta y_{t-j}\right)^2 + \lambda_T w_1^{\gamma_1}|\rho| + \lambda_T \sum_{j=1}^{p} w_{2j}^{\gamma_2}\left|\beta_j\right|, \quad \gamma_1,\gamma_2 > 0 \tag{2}$$

where $w_1 = 1/|\hat{\rho}_I|$ and $w_{2j} = 1/|\hat{\beta}_{I,j}|$ for $j = 1,...,p$ and $\hat{\rho}_I$ and $\hat{\beta}_{I,j}$ denote some initial estimator of the parameters in (2). We shall use the least squares estimator in this paper but other estimators can be used as well. Hence, the Adaptive LASSO minimizes the least squares objective function plus a penalty term which penalizes parameters that are different from 0. Due to this extra penalty term the minimizers of (2) are shrunk towards zero compared to the least squares estimator – hence the name shrinkage estimator. The size of the shrinkage depends on the penalty term, which in turn depends on the initial least squares estimates: the smaller the initial estimate, the larger the penalty and the more likely it is that the Adaptive LASSO shrinks the parameter exactly to zero. The size of the penalty also depends on the sequence λ_T which must be chosen in an appropriate manner in order to get the oracle efficiency. In particular, λ_T must grow fast enough to shrink the estimates of truly zero parameters to zero, but slow enough in order not to introduce asymptotic bias in the estimators of the non-zero coefficients. The details are given in Section 3.

In this paper we don't include deterministic components such as constants and trends to focus on the main idea of consistent and conservative model selection in stationary and nonstationary autoregressions. However, deterministics could be handled using standard detrending ideas, see e.g. Hamilton (1994).

2.1. **Notation.** We shall consider $T + p$ observations from a time series y_t generated by (1). $\mathcal{A} = \left\{1 \leq j \leq p : \beta_j^* \neq 0\right\}$ denotes the active set of lagged differences, i.e. those lagged differences with non-zero coefficients. Let $z_t = (\Delta y_{t-1}, ..., \Delta y_{t-p})'$ be the $(p \times 1)$ vector of lagged differences and let $x_t = (y_{t-1}, z_t')'$ denote the vector of all covariates. Let $\Sigma = E(z_t z_t')$ [2] and let $\Sigma_{\mathcal{A}}$ denote the matrix that has picked out all elements in columns and rows indexed by $\mathcal{A}$. So if $p = 5$ and $\mathcal{A} = (1, 3, 4)$, $\Sigma_{\mathcal{A}}$ equals the (3×3) matrix that has picked out rows and columns 1,3 and 4 out of the (5×5) matrix Σ. Similarly, $\mathcal{A}$ indexes vectors by picking out the elements with index in $\mathcal{A}$.

[2] Of course the actual value of this expectations depends on whether y_t is stationary or not. However, irrespective of this z_t is stationary.

Let $\Delta y = (\Delta y_T, ..., \Delta y_1)'$, $y_{-1} = (y_{T-1}, ..., y_0)'$ and $\Delta y_{-j} = (\Delta y_{T-j}, ..., \Delta y_{1-j})'$, $j = 1, ..., p$ [3]. Let $X_T = (y_{-1}, \Delta y_{-1}, ..., \Delta y_{-p})$ be the $T \times (p+1)$ matrix of covariates and $\epsilon = (\epsilon_T, ..., \epsilon_1)'$ the vector of error terms.

Let Z be a $p \times 1$ vector such that $Z \sim N_p(0, \sigma^2 \Sigma)$ where $\sigma^2 = E(\epsilon_t^2)$. Furthermore, $(W_r)_{r=0}^1$ denotes the standard Wiener process on $[0,1]$.

$S_T = diag(T, \sqrt{T}, ..., \sqrt{T})$ denotes a $(p+1 \times p+1)$ scaling matrix, $\tilde{\to}$ denotes weak convergence (convergence in law) and $\overset{p}{\to}$ convergence in probability. Let $(\hat{\rho}, \hat{\beta})$ denote the minimizer of (2). Of course $(\hat{\rho}, \hat{\beta})$ depends on T but to keep notation simple we suppress this in the sequel. Where no confusion arises this is also done for other quantities.

For any $x \in \mathbb{R}^n$, $||x||_{\ell_2} = \sqrt{\sum_{i=1}^n x_i^2}$ denotes the standard Euclidean ℓ_2 norm stemming from the inner product $< x, x >= \sum_{i=1}^n x_i^2$.

Letting $\mathcal{M}_0$ denote the true model and $\hat{\mathcal{M}}$ the estimated one, we shall say that a procedure is consistent if for all (ρ^*, β^*), $P(\hat{\mathcal{M}} = \mathcal{M}_0) \to 1$. A procedure is said to be conservative if for all (ρ^*, β^*), $P(\mathcal{M}_0 \not\subseteq \hat{\mathcal{M}}) \to 0$, i.e. the probability of excluding relevant variables tends to zero.

3. Oracle results

This section establishes and discusses the oracle property of the Adaptive LASSO for stationary as well as nonstationary autoregressions. The results open the possibility to use the Adapative LASSO to distinguish between these two and hence the Adaptive LASSO can also be seen as a new way of testing for unit roots.

We begin with the nonstationary case:

Theorem 1 (Consistent model selection under nonstationarity)**.** *Assume that* $\rho^* = 0$ *and that* ϵ_t *is i.i.d with* $E(\epsilon_1) = 0$ *and* $E(\epsilon_1^4) < \infty$. *Then, if* $\frac{\lambda_T}{T^{1-\gamma_1}} \to \infty$, $\frac{\lambda_T}{T^{1/2-\gamma_2/2}} \to \infty$ *and* $\frac{\lambda_T}{T^{1/2}} \to 0$

1. *Consistency:* $\left\| S_T \left[\left(\hat{\rho}, \hat{\beta}' \right)' - \left(0, \beta^{*\prime} \right)' \right] \right\|_{\ell_2} \in O_p(1)$
2. *Oracle (i):* $P(\hat{\rho} = 0) \to 1$ *and* $P(\hat{\beta}_{\mathcal{A}^c} = 0) \to 1$
3. *Oracle (ii):* $\sqrt{T}(\hat{\beta}_{\mathcal{A}} - \beta_{\mathcal{A}}) \tilde{\to} N\left(0, \sigma^2 [\Sigma_{\mathcal{A}}]^{-1}\right)$

$\frac{\lambda_T}{T^{1-\gamma_1}} \to \infty$ enables us to set $\hat{\rho} = 0$ with probability tending to one if $\rho^* = 0$. Likewise, $\frac{\lambda_T}{T^{1/2-\gamma_2/2}} \to \infty$ is needed to shrink the estimates of truly zero β_js to zero. Notice that both conditions require λ_T to grow sufficiently fast, i.e. the size of the penalty term must be sufficiently large to shrink the estimates of the zero parameters to zero. $\frac{\lambda_T}{T^{1/2}} \to 0$ on the other hand tells us that λ_T can not grow too fast. For if λ_T grows too fast even non-zero parameters will be shrunk to zero. In order for all three conditions to be satisfied simultaneously we need $\gamma_1 > 1/2$ and $\gamma_2 > 0$. It is of interest that the requirements on γ_1 and γ_2 are not the same. The reason for this difference is that $\hat{\rho}_I$ converges at a rate of $1/T$ while $\hat{\beta}_j$ converges at a rate of $1/\sqrt{T}$.

Theorem 1 states that $\hat{\rho}$ and $\hat{\beta}$ are estimated consistently at rates $1/T$ and $1/\sqrt{T}$, respectively. Furthermore, the estimators of the zero coefficients don't only converge to zero in probability – they are set exactly equal to zero with probability tending to one. Hence, the Adaptive LASSO performs variable selection and consistent estimation (the correct sparsity pattern is detected asymptotically). Finally, the asymptotic distribution of the nonzero β_js is the same as if the true model had been known and only the relevant variables (those with nonzero coefficients) had been included and least squares applied to that model. In other words, the Adaptive LASSO possesses the oracle property:

[3] The dependence on T is suppressed for some of the quantities where no confusion arises.

It sets all parameters that are zero exactly equal to zero and the asymptotic distribution of the estimators of the non-zero coefficients is the same as if only the relevant variables had been included in the model. So the Adaptive LASSO performs as well as if an oracle had revealed the correct sparsity pattern prior to estimation. This sounds too good to be true – and in some sense it is as we shall see in section 4.

The assumption that ϵ_t is *i.i.d* can be relaxed as long as $S_T^{-1}X_T'X_TS_T^{-1}$ and $S_T^{-1}X_T'\epsilon$ converge weakly. Of course the limits might change but since we establish that the Adaptive LASSO is asymptotically equivalent to the least squares estimator *only including the relevant variables* we would still conclude that it performs as well as if the true sparsity pattern had been known.

Next, we consider the Adaptive LASSO for stationary autoregressions.

Theorem 2 (Consistent model selection under stationarity)**.** *Assume that* $\rho^* \in (-2,0)$ *and that* ϵ_t *is i.i.d with* $E(\epsilon_1)=0$ *and* $E(\epsilon_1^4)<\infty$. *Then, if* $\frac{\lambda_T}{T^{1/2-\gamma_2/2}} \to \infty$ *and* $\frac{\lambda_T}{T^{1/2}} \to 0$

1. *Consistency:* $\left\| \sqrt{T}\left[\left(\hat{\rho},\hat{\beta}'\right)' - \left(\rho^*,\beta^{*\prime}\right)'\right]\right\|_{\ell_2} \in O_p(1)$
2. *Oracle (i):* $P(\hat{\rho}=0)\to 0$ *and* $P(\hat{\beta}_{\mathcal{A}^c}=0)\to 1$
3. *Oracle (ii):* $\begin{pmatrix} \sqrt{T}(\hat{\rho}-\rho^*) \\ \sqrt{T}(\hat{\beta}_{\mathcal{A}}-\beta^*_{\mathcal{A}}) \end{pmatrix} \tilde{\to} N\left(0,\sigma^2[Q_{(1,\mathcal{A}+1)}]^{-1}\right)$

   where $Q=E(x_tx_t')$ *of dimension* $(p+1\times p+1)$ [4]

Since the assumptions on λ_T are a subset of those made in Theorem 1, the resulting requirements on γ_1 and γ_2 are a fortiori satisfied.

The *i.i.d* assumption on ϵ_t can be relaxed as in the nonstationary setting as long as $\frac{1}{T}X_T'X_T$ converges in probability and $\frac{1}{\sqrt{T}}X_T'\epsilon$ converges weakly.

As in Theorem 1 the Adaptive LASSO gives consistent parameter estimates and as usual for stationary autoregressions the rate of convergence of $\hat{\rho}$ is slowed down to the standard $1/\sqrt{T}$ rate compared to $1/T$ in the nonstationary case. The probability of falsely classifying $\hat{\rho}=0$ tends to 0. As in Theorem 1 all irrelevant lagged differences will be classified as such with probability tending to one.

Theorem 1 and 2 show that the Adaptive LASSO can perform oracle efficient variable selection and estimation in stationary and nonstationary autoregressions. The theorems also suggest that the Adaptive LASSO can be used to discriminate between stationary and nonstationary autoregressions and so opens the possibility to use the Adaptive LASSO for unit root testing. The practical performance will be investigated in Section 6.

4. Finite sample and local analysis

In this section we analyze the finite sample behavior of the Adaptive LASSO. This is most conveniently done in the setting of an AR(1) to keep the focus on the main points and the expressions simple. The expressions we obtain for the finite sample selection probabilities also allow us to describe the local behavior of the Adaptive LASSO easily. We give a complete description for all possible limiting values of the regularization parameter λ_T.

Since $\gamma_1=1$ is in accordance with Theorems 1 and 2, i.e. we obtain the oracle property in the stationary as well as the nonstationary setting for this choice of γ_1, we shall focus on this value in

[4] In accordance with previous notation, $Q_{(1,\mathcal{A}+1)}$ is the matrix consisting of all rows and columns with indexes in the set $(1,\mathcal{A}+1)$ where the addition is to be understood elementwise.

the sequel. Similar calculations can be made for other admissible values of γ_1. Since there are no lagged differences γ_2 is redundant in this setting.

To be precise, we will consider the model

$$\Delta y_t = \rho^* y_{t-1} + \epsilon_t \tag{3}$$

where ϵ_t can be quite general. In particular it just needs to allow for a central limit theorem to apply in the stationary case ($\rho^* \in (-2, 0)$) and a functional central limit theorem to apply in the unit root as well as the local to unity setting. Appropriate assumptions can be found in Phillips (1987a) and Phillips (1987b) who allows for quite general dependence structures in the sequence $\{\epsilon_t\}$. In the following we shall assume that $\{\epsilon_t\}$ is i.i.d. but keep in mind that the results carry over to much more general assumptions on $\{\epsilon_t\}$. ρ^* is estimated by minimizing

$$L(\rho) = \sum_{t=1}^{T} (\Delta y_t - \rho y_{t-1})^2 + 2\lambda_T \frac{|\rho|}{|\hat{\rho}_I|} \tag{4}$$

which is the AR(1) equivalent to (2) except for a factor of 2 in front of λ_T whose only purpose is to make expressions simpler. Without any confusion, we let $\hat{\rho}$ denote the minimizer of (4) and $\hat{\rho}_I$ the least squares estimate of ρ^*.

The first theorem gives the exact finite sample probability of setting $\hat{\rho}$ equal to zero.

Theorem 3. *Let $\Delta y_t = \rho^* y_{t-1} + \epsilon_t$ and let $\hat{\rho}$ denote the minimizer of (4). Then*

$$P\left(\hat{\rho} = 0\right) = P\left(\left[{\rho^*}^2 + \left(\frac{\sum_{t=1}^{T} y_{t-1}\epsilon_t}{\sum_{t=1}^{T} y_{t-1}^2}\right)^2 + 2\rho^* \frac{\sum_{t=1}^{T} y_{t-1}\epsilon_t}{\sum_{t=1}^{T} y_{t-1}^2}\right] \sum_{t=1}^{T} y_{t-1}^2 \leq \lambda_T\right)$$

Theorem 3 characterizes the *exact finite sample* probability of setting $\hat{\rho}$ to zero. It is sensible that this is an increasing function in the regularization/shrinkage parameter λ_T since the larger λ_T is, the larger is the shrinkage.

The following theorems quantify the asymptotic behavior of $P(\hat{\rho} = 0)$ in case of 1) unit root, 2) stationarity, and 3) local to unity behavior in (3).

The first theorem is concerned with the nonstationary case where $\rho^* = 0$. Hence, we would like to classify $\hat{\rho} = 0$. This of course requires λ_T to be sufficiently large.

Theorem 4. *Let $\Delta y_t = \rho^* y_{t-1} + \epsilon_t$ with $\rho^* = 0$ and let $\hat{\rho}$ denote the minimizer of (4).*

1. *If $\lambda_T \to 0$ then $P(\hat{\rho} = 0) \to 0$*
2. *If $\lambda_T \to \lambda \in (0, \infty)$ then $P(\hat{\rho} = 0) \to p \in (0, 1)$*
3. *If $\lambda_T \to \infty$ then $P(\hat{\rho} = 0) \to 1$*

Theorem 4 reveals that in the presence of a unit root, $\lambda_T \to \infty$ yields consistent model selection, i.e. $P(\hat{\rho} = 0) \to 1$. If λ_T tends to a finite constant $\hat{\rho}$ has mass at 0 in the limit but the mass is not one. If $\lambda_T \to 0$, $\hat{\rho}$ is asymptotically equivalent to the least squares estimator and in fact even $T\hat{\rho}$ is asymptotically equivalent to T times the least squares estimator. Since this does not have any mass at 0 in the limit it is sensible that $P(\hat{\rho} = 0) \to 0$ when $\lambda_T \to 0$. In particular, $\hat{\rho}$ equals the least squares estimator if $\lambda_T = 0$ for every $T < \infty$ which can be seen from (4). Since $\rho^* = 0$, Theorem 4 does not impose any restrictions on λ_T in order to obtain conservative model selection since there are no relevant variables to be excluded.

The next theorem concerns the stationary case. In this case we do not want $\hat{\rho}$ to possess any mass at zero asymptotically. This naturally restricts the rate at which λ_T can increase as seen below.

Theorem 5. *Let $\Delta y_t = \rho^* y_{t-1} + \epsilon_t$ with $\rho^* \in (-2, 0)$ and let $\hat{\rho}$ denote the minimizer of (4)*

(1) *If $\lambda_T/T \to 0$ then $P(\hat{\rho} = 0) \to 0$*
(2) *If $\lambda_T/T \to \lambda$ then $P(\hat{\rho} = 0) \to \begin{cases} 0 \text{ if } \rho^{*2}E(y_{t-1}^2) > \lambda \\ 1 \text{ if } \rho^{*2}E(y_{t-1}^2) < \lambda \end{cases}$*
(3) *If $\lambda_T/T \to \infty$ then $P(\hat{\rho} = 0) \to 1$*

Part 1 of Theorem 5 shows that in order for $\hat{\rho}$ not to possess any mass at 0 asymptotically, it is sufficient that $\lambda_T \in o(T)$. If $\lambda/T \to \infty$, then $\hat{\rho}$ will be set to zero with probability tending to one even though $\rho^* \neq 0$ as can be seen from part 3 of the theorem. Part 2 of Theorem 5 is markedly different from part 2 of Theorem 4. This is due to the fact that the random variable which "decides" whether $\hat{\rho}$ is to be classified as zero or not converges to a point mass at $\rho^{*2}E(y_{t-1}^2)$ in the stationary setting while it converges to a nondegenerate distribution in the nonstationary setting. This constant is the knife edge on which $P(\hat{\rho} = 0)$ switches between 0 and 1. See the appendix for details. Also notice, that no classification is possible when $\lambda = \rho^{*2}E(y_{t-1}^2)$ in the stationary setting since $\rho^{*2}E(y_{t-1}^2)$ is a discontinuity point of the limiting distribution of the variable that "decides" whether $\hat{\rho}$ is to be classified as zero or not. We suspect that $P(\hat{\rho} = 0)$ depends not only on $\lambda = \rho^{*2}E(y_{t-1}^2)$ but on the concrete fashion in which λ_T converges to $\rho^{*2}E(y_{t-1}^2)$.

Taken together, theorems 4 and 5 show that for the Adaptive LASSO to act as a consistent model selection procedure it is sufficient that $\lambda_T \to \infty$ (by Theorem 4) and $\lambda_T/T \to \lambda$ for some $\lambda < \rho^{*2}E(y_{t-1}^2)$ (by Theorem 5). Since $\rho^* \neq 0$ in Theorem 5, $\lambda = 0$ works in particular. Hence, $\lambda_T = T^\alpha$ is admissible for all $0 < \alpha < 1$.

In order to make the Adaptive LASSO work as a conservative models selection device, Theorem 4 does not pose any restrictions on λ_T since $\rho^* = 0$ in that theorem so there are no relevant variables to be excluded. Hence, the only requirement is $\lambda_T/T \to \lambda$ for some $\lambda < \rho^{*2}E(y_{t-1}^2)$ (by Theorem 5). Note that in particular $\lambda_T \to a \geq 0$ or $\lambda_T = 0$ for all T work in this setting. $\lambda_T = 0$ amounts to no shrinkage at all and hence a zero probability of excluding relevant variables. These two (limiting) values of λ are ruled out by consistent model selection and will play a crucial role in highlighting the difference between consistent and conservative model selection in Theorem 6 below.

Next, compare the requirements on λ_T from theorems 4 and 5 for consistent model selection ($\lambda_T \to \infty$ and $\lambda_T/T \to \lambda < \rho^{*2}E(y_{t-1}^2)$) to those resulting from Theorems 1 and 2 (with $\gamma_1 = \gamma_2 = 1$). Note that $\lambda_T \to \infty$ in both groups of theorems. However, Theorems 1 and 2 are more restrictive on the growth rate of λ_T in that they require $\lambda_T/\sqrt{T} \to 0$ while Theorems 4 and 5 only require $\lambda_T/T \to 0$. This is not too surprising since Theorems 1 and 2 deliver more. They yield consistent model selection, consistent parameter estimation as well as the oracle efficient distribution. It is not hard to show that the requirements made in theorems 4 and 5 are also sufficient to yield consistent parameter estimation. However, only requiring $\lambda_T/T \to \infty$ is not enough to ensure that $\sqrt{T}\hat{\rho}$ is asymptotically equivalent to $\sqrt{T}$ times the least squares estimator when $\rho^* \neq 0$ since we penalize and hence shrink too much. The result is that $\hat{\rho}$ no longer obtains the oracle efficient distribution asymptotically. $\lambda_T/\sqrt{T}$ is needed to obtain the oracle efficient distribution. Knight and Fu (2000) made a similar observation in a deterministic cross sectional setting.

The next theorem concerns the local to unity situation. So far all results have been for pointwise asymptotics. The local to unity setting is a harder test for am estimator of ρ^* in the sense that it must perform well on a sequence of shrinking alternatives instead of only at a single point in the parameter space.

Theorem 6. *Let $\Delta y_t = \rho^* y_{t-1} + \epsilon_t$ with $\rho^* = c/T$ for some $c \neq 0$ and let $\hat{\rho}$ denote the minimizer of (4).*

1. *If $\lambda_T \to 0$ then $P(\hat{\rho} = 0) \to 0$*
2. *If $\lambda_T \to \lambda \in (0, \infty)$ then $P(\hat{\rho} = 0) \to p \in (0, 1)$*
3. *If $\lambda_T \to \infty$ then $P(\hat{\rho} = 0) \to 1$*

Consistent model selection requires that $\lambda_T \to \infty$ (by Theorem 4). By part 3 in Theorem 6 this implies that $P(\hat{\rho} = 0) \to 1$. Hence, the Adaptive LASSO tuned to perform consistent model selection has no power against deviations from 0 of the form c/T. This is a negative result since $c/T \neq 0$ for all T. This poor local performance is the flip side of shrinkage estimators (tuned to perform consistent model selection) which is reminiscent of Hodge's estimator, see e.g Lehmann and Casella (1998). This phenomenon has already been observed by Leeb and Pötscher (2005, 2008); Pötscher and Leeb (2009) in a different context.

Recall however, that $\lambda_T \to \infty$ is required in Theorems 1, 2 and 4 in order to achieve enough shrinkage to obtain consistent model selection. The price paid for a shrinkage of this size is that even parameters that are local to zero at a rate of O(1/T) will be shrunk to zero with probability tending to one. The Adaptive LASSO tuned to perform consistent model selection has no power against such alternatives.

On the other hand, the Adaptive LASSO tuned to perform conservative model selection *does* have power against deviations from 0 of this form. This becomes clear from parts 1 and 2 of Theorem 6 since $\lambda_T \to 0$ and $\lambda_T \to \lambda \in (0, \infty)$ are both in line with conservative model selection (see the discussion between theorems 5 and 6).

If $\lambda_T \to \lambda \in [0, \infty)$ then the probability of setting $\hat{\rho}$ exactly equal to zero no longer tends to one for $\rho^* = c/T$. However, by Theorem 4, the same is the case when $\rho^* = 0$. If this tradeoff is preferred to consistent model selection, the next section gives the properties in the of the Adaptive LASSO in the AR(p) model (1) when tuned to perform conservative model selection.

5. Conservative model selection

We continue to consider the case $\gamma_1 = \gamma_2 = 1$ since this keeps expressions simple. When tuned to perform conservative model selection the properties of the Adaptive LASSO are as follows.

Theorem 7 (Conservative model selection under nonstationarity)**.** *Assume that $\rho^* = 0$ and that ϵ_t is i.i.d with $E(\epsilon_1) = 0$ and $E(\epsilon_1^4) < \infty$. Let $\gamma_1 = \gamma_2 = 1$[5]. Then, if $\lambda_T \to \lambda \in [0, \infty)$*

$$S_T\left[\left(\hat{\rho}, \hat{\beta}'\right)' - \left(0, \beta^{*\prime}\right)'\right] \overset{\sim}{\to} \arg\min \Psi(u)$$

which implies

$$\left\| S_T\left[\left(\hat{\rho}, \hat{\beta}'\right)' - \left(0, \beta^{*\prime}\right)'\right] \right\|_{\ell_2} \in O_p(1)$$

where

[5] This assumption is not essential at all. It is only made to ensure $\frac{\lambda_T}{T^{1-\gamma_1}} = \frac{\lambda_T}{T^{1/2-\gamma_2/2}} = \lambda_T \to \lambda$ such that we don't have to deal with different cases for the size of $\frac{\lambda_T}{T^{1-\gamma_1}}$ and $\frac{\lambda_T}{T^{1-\gamma_2/2}}$.

$\Psi(u) = u'Au - 2Bu + \lambda\frac{|u_1|}{C_1} + \lambda\sum_{j=1}^{p}\frac{|u_{2j}|}{C_{2j}}\mathbf{1}_{\left\{\beta_j^*=0\right\}}$
with

$$A = \begin{pmatrix} \frac{\sigma^2}{(1-\sum_{j=1}^{p}\beta_j^*)^2}\int_0^1 W_r^2 dr & 0 \\ 0 & \Sigma \end{pmatrix}, \; B = \begin{pmatrix} \frac{\sigma^2}{(1-\sum_{j=1}^{p}\beta_j^*)}\int_0^1 W_r dW_r \\ Z \end{pmatrix}$$

$$C_1 \sim \frac{(1-\sum_{j=1}^{p}\beta_j^*)\int_0^1 W_s dW_s}{\int_0^1 W_s^2 ds} \text{ and } C_{2j} \sim N\left(0, \sigma^2(\Sigma^{-1})_{(j,j)}\right)$$

Theorem 7 reveals that $\left(\hat{\rho}, \hat{\beta}'\right)'$ converges at the same rate as the leas squares estimator – but to the minimizer of $\Psi(u)$. Note that no shrinkage is applied to u_{2j} for $j \in \mathcal{A}$ which is a desirable property.

For $\lambda = 0$, Theorem 7 reveals that the asymptotic distribution of the Adaptive LASSO estimator is identical to the minimizer of $u'Au - 2Bu$. This in turn reveals that in this case the limit law of the Adaptive LASSO estimator is identical to the one of the least squares estimator in the model including all variables. This result is of course not surprising since $\lambda = 0$ implies that asymptotically there is no penalty on nonzero parameters and hence no shrinkage which implies that the objective function of the Adaptive LASSO, (2), approaches the least squares objective function. The absence of shrinkage also implies that no parameters will be set exactly equal to 0 (or, more precisely, the probability of a parameter being set to 0 is 0).

If $\lambda \in (0, \infty)$, the penalty terms do no longer vanish asymptotically (except for the nonzero β_j^*). Hence, with positive probability[6] the minimizer of $\Psi(u)$ has entries with value zero.

Next, consider conservative model selection in the stationary case

Theorem 8 (Conservative model selection under stationarity)**.** *Assume that $\rho^* \in (-2, 0)$ and that ϵ_t is i.i.d with $E(\epsilon_1) = 0$ and $E(\epsilon_1^4) < \infty$. Let $\gamma_1 = \gamma_2 = 1$[7]. Then, if $\lambda_T \to \lambda \in [0, \infty)$*

$$\sqrt{T}\left[\left(\hat{\rho}, \hat{\beta}'\right)' - \left(\rho^*, \beta^{*\prime}\right)\right] \tilde{\to} \arg\min \tilde{\Psi}(u)$$

which implies

$$\left\|\sqrt{T}\left[\left(\hat{\rho}, \hat{\beta}'\right)' - \left(\rho^*, \beta^{*\prime}\right)'\right]\right\|_{\ell_2} \in O_p(1)$$

where

$\tilde{\Psi}(u) = u'Qu - 2\tilde{B}u + \lambda\sum_{j=1}^{p}\frac{|u_{2j}|}{\tilde{C}_{2j}}\mathbf{1}_{\left\{\beta_j^*=0\right\}}$
with

$$\tilde{B} \sim N_{p+1}(0, \sigma^2 Q) \text{ and } \tilde{C}_{2j} \sim N_{p+1}\left(0, \sigma^2(Q^{-1})_{(1+j,1+j)}\right)$$

[6] Actually calculating this probability seems to be non-trivial.

[7] As in Theorem 7 this assumption is not essential at all. It is only made to ensure $\frac{\lambda_T}{T^{1-\gamma_1}} = \frac{\lambda_T}{T^{1/2-\gamma_2/2}} = \lambda_T \to \lambda$ such that we don't have to deal with different cases for the size of $\frac{\lambda_T}{T^{1-\gamma_1}}$ and $\frac{\lambda_T}{T^{1-\gamma_2/2}}$.

As in the nonstationary case $\left(\hat{\rho}, \hat{\beta}'\right)'$ converges at the same rate as the least squares estimator – but to the minimizer of $\tilde{\Psi}(u)$. Note that no shrinkage is applied to u_{2j} for $j \in \mathcal{A}$. More importantly, no shrinkage is applied to u_1 since now $\rho^* \neq 0$.

Similar to the nonstationary case $\left(\hat{\rho}, \hat{\beta}'\right)'$ converges to the same limit as the least squares estimator if $\lambda = 0$. A particular instance of this is of course $\lambda_T =$ for all T in which case the Adaptive LASSO estimate is equal to the least squares estimate so their limiting laws are a fortiori identical.

6. Monte Carlo

This section illustrates the above results by means of Monte Carlo experiments. The Adaptive LASSO is implemented by means of the algorithm proposed in Zou (2006). Its performance is compared to the LASSO implemented by the LARS algorithm of Efron et al. (2004) using the publicly available package at `cran.r-project.org`. Furthermore, a comparison is made to the BIC only selecting over the lagged differences, i.e. y_{t-1} is always included in the model. Using the model chosen by BIC an Augmented Dickey-Fuller test is carried out for the presence of a unit root at a 5% significance level. The results for this procedure are denoted BICDF. Finally, all these procedures are compared to the "OLS Oracle" (OLSO) which carries out least squares only including the relevant variables.

The Adaptive LASSO is implemented with $\gamma = \gamma_1 = \gamma_2 = 0.51, 1, 10$. $\gamma = 0.51$ is included since it is in the lowest end of the values of γ which are in accordance with theorems 1 and 2. We also experimented with values of γ larger than 10 but the performance of the Adaptive LASSO was not improved by these. Finally, the Adaptive LASSO was also implemented by selecting γ by BIC from the above values.

The above procedures are compared along the following dimensions.

(1) Sparsity pattern: How often does the procedure detect the correct sparsity pattern, i.e. how often does it include all relevant variables while not including any irrelevant ones?
(2) Unit root: How often does the procedure make the correct decision on inclusion/exclusion of y_{t-1}? Or put differently, how well do the procedures classify whether $\rho^* = 0$ or not.
(3) Relevant included: How often does the procedure include all relevant variables in the model? Even though the correct sparsity pattern is not detected it is of interest to know whether the procedure at least does not exclude any relevant variables from the model. Or in the jargon of the previous sections does the procedure at least perform conservative model selection.
(4) Loss: How accurate does the estimated model predict on a hold out sample? Here we generate data from the same data generating process as used for the specification and estimation and use the estimated parameters to make predictions on this hold out sample.

To gauge the performance along the above dimensions we carry out the following experiments with sample sizes of $T = 100$ and 1000. The number of Monte Carlo replications is 1000 in all cases.

- Experiment A: $\rho^* = 0$ and $\beta' = (0.4, 0.3, 0.2, 0, 0, 0, 0, 0, 0, 0)$. A unit root setting with three relevant lagged differences.
- Experiment B: $\rho^* = -0.05$ and $\beta' = (0.4, 0.3, 0.2, 0, 0, 0, 0, 0, 0, 0)$. A stationary, but close to unit root setting with three relevant lagged differences. This should be a challenging setting since the data generating process is stationary but still close to the unit root setting which makes it harder to classify $\rho^* \neq 0$.
- Experiment C: $\rho^* = 0$ and $\beta' = (0.4, 0.3, 0.2, 0, 0, 0, -0.2, 0, 0, 0.2, 0, 0)$. This experiment is carried out to investigate how well the methods fare when there is a gap in the lag structure.

6.1. **Results.** The best performer in terms of choosing the correct sparsity pattern in Experiment A when $T = 100$ is the Adaptive LASSO with $\gamma = 10$ (see Table 1). It is also superior when it comes to correctly classifying $\rho^* = 0$ or $\rho \neq 0$ – it always makes the correct classification. This is better than the BICDF which classifies ρ^* correctly in 90% of the instances. The parameter estimates by the Adaptive LASSO using $\gamma = 10$ also yield the best out of sample predictive accuracy (lowest Loss), outperforming all other procedures except for the infeasible OLS Oracle. It is seen that classifying ρ^* correctly plays an important role in obtaining a low Loss since procedures with low success in classifying ρ^* incur the biggest losses (BIC and LASSO) and vice versa. For the Adaptive LASSO no choice of γ retains all the relevant variables more than 50% of the time and interestingly $\gamma = 10$ is the worst choice along this dimension. Among all procedures, the LASSO does by far the best job at retaining relevant variables when $T = 100$.

For $T = 1000$ the Adaptive LASSO performs as well as the OLS Oracle along all dimensions underscoring its oracle property. It chooses the correct sparsity pattern almost every time (except when $\gamma = 0.51$) always classifies ρ^* correctly.

Experiment A also underscores that the LASSO is not a consistent variable selection procedure – as the sample size increases the fraction of correctly selected sparsity patterns remains constant.

	Experiment A				Adaptive LASSO				
		BIC	BICDF	LASSO	$\gamma = 0.51$	$\gamma = 1$	$\gamma = 10$	BIC	OLSO
T=100	Sparsity Pattern	0.000	0.108	0.020	0.083	0.224	0.149	0.147	1.000
	Unit Root	0.000	0.904	0.081	0.266	0.858	1.000	0.919	1.000
	Relevant Retained	0.158	0.158	0.798	0.439	0.353	0.214	0.222	1.000
	Loss	1.203	1.142	1.249	1.202	1.103	1.045	1.096	1.018
T=1000	Sparsity Pattern	0.000	0.896	0.026	0.430	0.947	0.986	0.964	1.000
	Unit Root	0.000	0.947	0.062	0.449	0.990	1.000	0.991	1.000
	Relevant Retained	1.000	1.000	1.000	1.000	1.000	0.989	1.000	1.000
	Loss	1.007	1.005	1.010	1.007	1.003	1.002	1.002	1.002

TABLE 1. $\rho^* = 0$ and $\beta^{*\prime} = (0.4, 0.3, 0.2, 0, 0, 0, 0, 0, 0, 0)$

The first thing one notices in Experiment B (Table 2) is that now $\gamma = .51$ is the preferred value of γ for the Adaptive LASSO. This is in opposition to the nonstationary setting in Experiment A where $\gamma = 10$ yielded the best performance. It is also interesting that the LASSO is actually the strongest performer when T=100. It selects the correct sparsity pattern most often, classifies ρ^* correctly, and retains all relevant variable in about 80% of the cases. However, its lack of variable selection consistency is underscored by the fact that the fraction of times it selects the correct sparsity pattern does not tend to one even as the sample size increases[8]. In this stationary setting it seems less important to classify ρ^* correctly in order to achieve a low Loss on the hold out sample since all procedures incur roughly the same Loss. This is in opposition to Experiment A.

For the Adaptive LASSO as well as the BIC all quantities approach the ones of the OLS Oracle as the sample size increases illustrating the oracle efficiency of these estimators. The two procedures perform about equally well in this stationary setting. The poor performance of the Adaptive LASSO for $\gamma = 10$ may seem to be in opposition to Theorem 2 but for $T = 10.000$ (results not reported

[8] The estimations were also carried out for $T = 10.000$ (not reported here) and even then the LASSO only detected the correct sparsity pattern in 47% of the instances.

here) the Adaptive LASSO detects the correct sparsity pattern in 92% of the instances even with this value of γ.

	Experiment B				Adaptive LASSO				
		BIC	BICDF	LASSO	$\gamma = 0.51$	$\gamma = 1$	$\gamma = 10$	BIC	OLSO
T=100	Sparsity Pattern	0.214	0.214	0.478	0.340	0.269	0.001	0.281	1.000
	Unit Root	1.000	0.956	0.994	0.980	0.923	0.009	0.798	1.000
	Relevant Retained	0.265	0.257	0.814	0.518	0.436	0.008	0.382	1.000
	Loss	1.047	1.048	1.048	1.050	1.054	1.105	1.056	1.024
T=1000	Sparsity Pattern	0.946	0.946	0.754	0.914	0.922	0.000	0.924	1.000
	Unit Root	1.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000
	Relevant Retained	1.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000
	Loss	1.002	1.002	1.004	1.003	1.003	1.083	1.003	1.002

Table 2. $\rho^* = -0.05$ and $\beta^{*\prime} = (0.4, 0.3, 0.2, 0, 0, 0, 0, 0, 0, 0)$

As in Experiment A the data generating process possesses a unit root in Experiment C. Considering the results for T=100 in Table 3 the findings from Experiment A are roughly confirmed. Choosing $\gamma = 10$ for the Adaptive LASSO seems to be a wise choice in the unit root setting even with gaps in the lag structure. The Adaptive LASSO always classifies ρ^* correctly with this value of γ and outperforms its closest competitor (the adaptive LASSO using BIC to choose γ by almost 10%). By considering the Loss on the hold out sample it is also seen that a correct classification of ρ^* results in a big gain in predictive power in the presence of a unit root confirming the finding in Experiment A.

As the sample size increases the the Adaptive LASSO and the BIC perform better while the performance of the LASSO does not get better underscoring the oracle property of the first two procedures and the lack of the same for the LASSO. Note that the BICDF can not be expected to classify ρ^* correctly more often than in 95% of the cases in the presence of a unit root since testing is carried out at a 5% significance level. On the other hand, the correct classification probability of the Adaptive LASSO tends to one. Furthermore, the BIC is considerably slower to implement since the number of regressions to be run increases exponentially in the number of potential explanatory variables.

7. Conclusion

This paper has shown that the Adaptive LASSO can be tuned to perform consistent model selection in stationary and nonstationary autoregressions. The estimator of the parameters converges at the oracle efficient rate, i.e. as fast as if an oracle had revealed the true model prior to estimation and only the relevant variables had bee included in a least squares estimation. This enables us to use the Adaptive LASSO to distinguish between stationary and nonstationary autoregressions.

However, the Adaptive LASSO has no power against alternatives in a shrinking neighborhood around 0 when tuned to perform consistent variable selection. This problem can be alleviated by tuning the Adaptive LASSO to perform conservative model selection. The price paid compared to consistent model selection is that truly zero parameters are no longer set to zero with probability tending to one (but still with positive probability).

	Experiment C				Adaptive LASSO				
		BIC	BICDF	LASSO	$\gamma = 0.51$	$\gamma = 1$	$\gamma = 10$	BIC	OLSO
T=100	Sparsity Pattern	0.000	0.039	0.000	0.007	0.033	0.057	0.052	1.000
	Unit Root	0.000	0.904	0.069	0.246	0.851	1.000	0.932	1.000
	Relevant Retained	0.050	0.050	0.104	0.095	0.090	0.088	0.087	1.000
	Loss	1.242	1.158	1.298	1.233	1.133	1.083	1.132	1.033
T=1000	Sparsity Pattern	0.000	0.903	0.007	0.189	0.853	0.968	0.952	1.000
	Unit Root	0.000	0.955	0.026	0.233	0.974	1.000	0.994	1.000
	Relevant Retained	0.998	0.998	0.997	1.000	1.000	0.971	0.998	1.000
	Loss	1.008	1.006	1.012	1.010	1.005	1.003	1.003	1.003

Table 3. $\rho^* = 0$ and $\beta^{*\prime} = (0.4, 0.3, 0.2, 0, 0, 0, -0.2, 0, 0, 0.2, 0, 0)$

Monte Carlo experiments confirm that the Adaptive LASSO performs well compared to standard competitors. This is the case in particular for nonstationary data.

8. Appendix

Proof of Theorem 1. For the proof of this theorem we will need the following results which can be found in e.g. Hamilton (1994).

$$S_T^{-1}X_T'X_TS_T^{-1}\tilde{\to}\begin{pmatrix} \frac{\sigma^2}{(1-\sum_{j=1}^p \beta_j^*)^2}\int_0^1 W_r^2 dr & 0 \\ 0 & \Sigma \end{pmatrix} := A, \tag{5}$$

$$S_T^{-1}X_T'\epsilon\tilde{\to}\begin{pmatrix} \frac{\sigma^2}{(1-\sum_{j=1}^p \beta_j)}\int_0^1 W_r dW_r \\ Z \end{pmatrix} := B. \tag{6}$$

We shall also make use of the fact that the least squares estimator, $(\hat{\rho}_I, \hat{\beta}_I')$, of $(\rho^*, \beta^{*\prime})$ in (1) satisfies that $\left\|S_T\left[\left(\hat{\rho}_I, \hat{\beta}_I'\right)' - \left(\rho^*, \beta^{*\prime}\right)'\right]\right\|_{\ell_2} \in O_p(1)$

First, let $u = (u_1, u_2')'$ where u_1 is a scalar and u_2 a $p \times 1$ vector. Set $\rho = u_1/T$ and $\beta_j = \beta_j^* + u_{2j}/\sqrt{T}$ which implies that (2) as a function of u can be written as

$$\begin{aligned}\Psi_T(u) = &\left\|\Delta y - \frac{u_1}{T}y_{-1} - \sum_{j=1}^p \left(\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right)\Delta y_{-j}\right\|_{\ell_2}^2 \\ &+ \lambda_T w_1^{\gamma_1}\left|\frac{u_1}{T}\right| + \lambda_T \sum_{j=1}^p w_{2j}^{\gamma_2}\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right|.\end{aligned}$$

Let $\hat{u} = (\hat{u}_1, \hat{u}_2')' = \arg\min \Psi_T(u)$ and notice that $\hat{u}_1 = T\hat{\rho}$ and $\hat{u}_{2j} = \sqrt{T}(\hat{\beta}_j - \beta_j^*)$ for $j = 1, ..., p$. Define

$$\begin{aligned}V_T(u) &= \Psi_T(u) - \Psi_T(0)\\ &= u'S_T^{-1}X_T'X_TS_T^{-1}u - 2u'S_T^{-1}X_T'\epsilon + \lambda_T w_1^{\gamma_1}\left|\frac{u_1}{T}\right| + \lambda_T\sum_{j=1}^{p} w_{2j}^{\gamma_2}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right).\end{aligned}$$

Consider the first two terms in the above display. It follows from (5) and (6) that

$$u'S_T^{-1}X_T'X_TS_T^{-1}u - 2u'S_T^{-1}X_T'\epsilon \tilde{\rightarrow} u'Au - 2u'B. \tag{7}$$

Furthermore,

$$\lambda_T w_1^{\gamma_1}\left|\frac{u_1}{T}\right| = \lambda_T\frac{1}{|\hat{\rho}_I|^{\gamma_1}}\left|\frac{u_1}{T}\right| = |u_1|\frac{\lambda_T}{T^{1-\gamma_1}}\frac{1}{|T\hat{\rho}_I|^{\gamma_1}} \to \begin{cases}\infty \text{ in probability if } u_1 \neq 0\\ 0 \text{ in probability if } u_1 = 0\end{cases} \tag{8}$$

since $T\hat{\rho}_I$ is tight. Also, if $\beta_j^* \neq 0$

$$\begin{aligned}\lambda_T w_{2j}^{\gamma_2}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) &= \lambda_T\left|\frac{1}{\hat{\beta}_{I,j}}\right|^{\gamma_2}\frac{u_{2j}}{\sqrt{T}}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right)/\left(\frac{u_{2j}}{\sqrt{T}}\right)\\ &= \frac{\lambda_T}{T^{1/2}}\left|\frac{1}{\hat{\beta}_{I,j}}\right|^{\gamma_2}u_{2j}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right)/\left(\frac{u_{2j}}{\sqrt{T}}\right)\\ &\to 0 \text{ in probability}\end{aligned} \tag{9}$$

since (i): $\lambda_T/T^{1/2} \to 0$, (ii): $\left|1/\hat{\beta}_{I,j}\right|^{\gamma_2} \to \left|1/\beta_j^*\right|^{\gamma_2} < \infty$ in probability and (iii): $u_{2j}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right)/\left(\frac{u_{2j}}{\sqrt{T}}\right) \to u_{2j}\text{sign}(\beta_j^*)$.

Finally, if $\beta_j^* = 0$,

$$\begin{aligned}\lambda_T w_{2j}^{\gamma_2}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) &= \frac{\lambda_T}{T^{1/2}}\left|\frac{1}{\hat{\beta}_{I,j}}\right|^{\gamma_2}|u_{2j}| = \frac{\lambda_T}{T^{1/2-\gamma_2/2}}\left|\frac{1}{\sqrt{T}\hat{\beta}_{I,j}}\right|^{\gamma_2}|u_{2j}|\\ &\to \begin{cases}\infty \text{ in probability if } u_{2j} \neq 0\\ 0 \text{ in probability if } u_{2j} = 0\end{cases}\end{aligned} \tag{10}$$

since (i): $\frac{\lambda_T}{T^{1/2-\gamma_2/2}} \to \infty$ and (ii): $\sqrt{T}\hat{\beta}_{I,j}$ is tight.

Putting together (7)-(10) one concludes:

$$V_T(u) \tilde{\rightarrow} \Psi(u) = \begin{cases}u'Au - 2u'B \text{ if } u_1 = 0 \text{ and } u_{2j} = 0 \text{ for all } j \in \mathcal{A}^c\\ \infty \quad \text{if } u_1 \neq 0 \text{ or } u_{2j} \neq 0 \text{ for some } j \in \mathcal{A}^c\end{cases}$$

Since $V_T(u)$ is convex and $\Psi(u)$ has a unique minimum it follows from Knight (1999) that $\arg\min V_T(u) \tilde{\rightarrow} \arg\min \Psi(u)$. Hence,

$$\hat{u}_1 \overset{\sim}{\to} \delta_0 \tag{11}$$

$$\hat{u}_{2\mathcal{A}^c} \overset{\sim}{\to} \delta_0^{|\mathcal{A}^c|} \tag{12}$$

$$\hat{u}_{2\mathcal{A}} \overset{\sim}{\to} N(0, \sigma^2[\Sigma_{\mathcal{A}}]^{-1}) \tag{13}$$

where δ_0 is the Dirac measure at 0 and $|\mathcal{A}^c|$ is the cardinality of $\mathcal{A}^c$ (hence, $\delta_0^{|\mathcal{A}^c|}$ is the $|\mathcal{A}^c|$-dimensional Dirac measure at 0). Notice that (11) and (12) imply that $\hat{u}_1 \to 0$ in probability and $\hat{u}_{2\mathcal{A}^c} \to 0$ in probability. An equivalent formulation of (11)-(13) is

$$T\hat{\rho} \overset{\sim}{\to} \delta_0 \tag{14}$$

$$\sqrt{T}(\hat{\beta}_{\mathcal{A}^c} - \beta^*_{\mathcal{A}^c}) \overset{\sim}{\to} \delta_0^{|\mathcal{A}^c|} \tag{15}$$

$$\sqrt{T}(\hat{\beta}_{\mathcal{A}} - \beta^*_{\mathcal{A}}) \overset{\sim}{\to} N(0, \sigma^2[\Sigma_{\mathcal{A}}]^{-1}) \tag{16}$$

(14)-(16) yield the consistency part of the theorem at the rate of T for $\hat{\rho}$ and $\sqrt{T}$ for $\hat{\beta}$. Notice that this also implies that no $\hat{\beta}_j,\ j \in \mathcal{A}$ will be set equal to 0 since for all $j \in \mathcal{A},\ \hat{\beta}_j$ converges in probability to $\beta_j^* \neq 0$. (16) also yields the oracle efficient asymptotic distribution, i.e. part (3) of the theorem. It remains to show part (2) of the theorem; $P(\hat{\rho}_T = 0) \to 1$ and $P(\hat{\beta}_{T,\mathcal{A}^c} = 0) \to 1$. Both proofs are by contradiction.

First, assume $\hat{\rho} \neq 0$. Then the first order conditions for a minimum read:

$$2y'_{-1}\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right) + \lambda_T w_1^{\gamma_1}\text{sign}(\hat{\rho}) = 0$$

which is equivalent to

$$\frac{2y'_{-1}\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T} + \frac{\lambda_T w_1^{\gamma_1}\text{sign}(\hat{\rho})}{T} = 0$$

Consider first the second term:

$$\left|\frac{\lambda_T w_1^{\gamma_1}\text{sign}(\hat{\rho})}{T}\right| = \frac{\lambda_T}{T^{1-\gamma_1}}\frac{1}{|T\hat{\rho}_I|^{\gamma_1}} \to \infty \text{ in probability}$$

since $T\hat{\rho}_I$ is tight. For the first term one has:

$$\frac{2y'_{-1}\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T} = \frac{2y'_{-1}\left(\epsilon - X_T S_T^{-1} S_T[\hat{\rho}, \hat{\beta}' - \beta^{*\prime}]'\right)}{T}$$
$$= \frac{2y'_{-1}\epsilon}{T} - \frac{2y'_{-1}X_T S_T^{-1} S_T[\hat{\rho}, \hat{\beta}' - \beta^{*\prime}]'}{T}$$

By (6), $\frac{y'_{-1}\epsilon}{T} \overset{\sim}{\to} \frac{\sigma^2}{1-\sum_{j=1}^p \beta_j^*}\int_0^1 W_r dW_r$. Furthermore, $\frac{y'_{-1}X_T S_T^{-1}}{T} \overset{\sim}{\to} \left(\left(\frac{\sigma}{1-\sum_{j=1}^p \beta_j^*}\right)^2 \int_0^1 W_r^2 dr, 0, ..., 0\right)$ by (5). Hence, $\frac{y'_{-1}\epsilon}{T}$ and $\frac{y'_{-1}X_T S_T^{-1}}{T}$ are tight. We also know that $S_T[\hat{\rho}_T, \hat{\beta}'_T - \beta^{*\prime}]'$ converges weakly by (14)-(16) which implies it is tight as well. Taken together, $\frac{2y'_{-1}\left(\Delta y - X_T(\hat{\rho}_T, \hat{\beta}'_T)'\right)}{T}$ is tight and so

$$P(\hat{\rho}_T \neq 0) \leq P\left(\frac{2y_{-1}'\left(\Delta y - X_T(\hat{\rho}_T, \hat{\beta}_T')'\right)}{T} + \frac{\lambda_T w_1^{\gamma_1}\text{sign}(\hat{\rho}_T)}{T} = 0\right) \to 0$$

Next, assume $\hat{\beta}_j \neq 0$ for $j \in \mathcal{A}^c$. From the first order conditions

$$\Delta y_{-j}'(\Delta y - X_T(\hat{\rho}, \hat{\beta}')') + \lambda_T w_{2j}^{\gamma_2}\text{sign}(\hat{\beta}_j) = 0$$

or equivalently,

$$\frac{2\Delta y_{-j}'\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T^{1/2}} + \frac{\lambda_T w_{2j}^{\gamma_2}\text{sign}(\hat{\beta}_j)}{T^{1/2}} = 0$$

First, consider the second term

$$\left|\frac{\lambda_T w_{2j}^{\gamma_2}\text{sign}(\hat{\beta}_j)}{T^{1/2}}\right| = \frac{\lambda_T w_{2j}^{\gamma_2}}{T^{1/2}} = \frac{\lambda_T}{T^{1/2-\gamma_2/2}\left|T^{1/2}\hat{\beta}_{I,j}\right|^{\gamma_2}} \to \infty$$

since $\sqrt{T}\hat{\beta}_{I,j}$ is tight. Regarding the first term,

$$\frac{2\Delta y_{-j}'\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T^{1/2}} = \frac{2\Delta y_{-j}'\left(\epsilon - X_T S_T^{-1} S_T[\hat{\rho}_T, \hat{\beta}' - \beta^{*\prime}]'\right)}{T^{1/2}}$$
$$= \frac{2\Delta y_{-j}'\epsilon}{T^{1/2}} - \frac{2\Delta y_{-j}' X_T S_T^{-1} S_T[\hat{\rho}, \hat{\beta}' - \beta^{*\prime}]'}{T^{1/2}}$$

By (6) $\frac{\Delta y_{-j}'\epsilon}{T^{1/2}} \tilde{\to} N(0, \sigma^2\Sigma_j)$ where in accordance with previous notation Σ_j is the jth diagonal element of Σ. $\frac{\Delta y_{-j}' X_T S_T^{-1}}{T^{1/2}} \tilde{\to} (0, \Sigma_{(j,1)}, ..., \Sigma_{(j,p)})$ by (5). Hence, $\frac{\Delta y_{-j}'\epsilon}{T^{1/2}}$ and $\frac{\Delta y_{-j}' X_T S_T^{-1}}{T^{1/2}}$ are tight. The same is the case for $S_T[\hat{\rho}, \hat{\beta}' - \beta^{*\prime}]$ since it converges weakly by (14)-(16). Taken together, $\frac{2\Delta y_{-j}'\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T^{1/2}}$ is tight and so

$$P(\hat{\beta}_j \neq 0) \leq P\left(\frac{2\Delta y_{-j}'\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T^{1/2}} + \frac{\lambda_T w_{2j}^{\gamma_2}\text{sign}(\hat{\rho}_T)}{T^{1/2}} = 0\right) \to 0$$

□

Proof of Theorem 2. The proof runs along the same lines as the proof of Theorem 1. For the proof we will need (17) and (18) below which can be found in e.g. Hamilton (1994), Chapter 8. Notice that by definition of $x_t = (y_{t-1}, z_t')'$ the lower right hand $(p \times p)$ block of Q is Σ.

We shall make use of the following limit results:

$$\frac{1}{T}X_T'X_T \xrightarrow{p} Q \tag{17}$$

$$\frac{1}{\sqrt{T}}X_T'\epsilon \tilde{\to} N_{p+1}(0,\sigma^2 Q) := \tilde{B} \tag{18}$$

where the definition of $\tilde{B}$ means that $\tilde{B}$ is a random vector distributed as $N_{p+1}(0,\sigma^2 Q)$ We shall also make use of the fact that the least squares estimator is $\sqrt{T}$ consistent under stationarity, i.e. $\left\|\sqrt{T}\left[\left(\hat{\rho}_I,\hat{\beta}_I'\right)' - \left(\rho^*,\beta^{*\prime}\right)'\right]\right\|_{\ell_2} \in O_p(1)$

First, let $u = (u_1, u_2')'$ where u_1 is a scalar and u_2 a $p\times 1$ vector. Set $\rho = \rho^* + u_1/\sqrt{T}$ and $\beta_j = \beta_j^* + u_{2j}/\sqrt{T}$ and

$$\Psi_T(u) = \left\|\Delta y - \left(\rho^* + \frac{u_1}{\sqrt{T}}\right)y_{-1} - \sum_{j=1}^{p}\left(\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right)\Delta y_{-j}\right\|_{\ell_2}^2 + \lambda_T w_1^{\gamma_1}\left|\rho^* + \frac{u_1}{\sqrt{T}}\right| + \lambda_T\sum_{j=1}^{p} w_{2j}^{\gamma_2}\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right|$$

Let $\hat{u} = (\hat{u}_1, \hat{u}_2')' = \arg\min \Psi_T(u)$ and notice that $\hat{u}_1 = \sqrt{T}(\hat{\rho} - \rho^*)$ and $\hat{u}_{2j} = \sqrt{T}(\hat{\beta}_j - \beta_j^*)$ for $j = 1, ..., p$. Define

$$\begin{aligned}\tilde{V}_T(u) &= \Psi_T(u) - \Psi_T(0)\\ &= \frac{1}{T}u'X_T'X_Tu - 2\frac{1}{\sqrt{T}}u'X_T'\epsilon + \lambda_T w_1^{\gamma_1}\left(\left|\rho^* + \frac{u_1}{\sqrt{T}}\right| - |\rho^*|\right) + \lambda_T\sum_{j=1}^{p} w_{2j}^{\gamma_2}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right)\end{aligned}$$

Consider the first two terms in the above display. It follows from (17) and (18) that

$$\frac{1}{T}u'X_T'X_Tu - 2\frac{1}{\sqrt{T}}u'X_T'\epsilon \tilde{\to} u'Qu - 2u'\tilde{B} \tag{19}$$

Furthermore, since $\rho^* \neq 0$

$$\begin{aligned}\lambda_T w_1^{\gamma_1}\left(\left|\rho^* + \frac{u_1}{\sqrt{T}}\right| - |\rho^*|\right) &= \lambda_T\left|\frac{1}{\hat{\rho}_I}\right|^{\gamma_1}\frac{u_1}{\sqrt{T}}\left(\left|\rho^* + \frac{u_1}{\sqrt{T}}\right| - |\rho^*|\right)/\left(\frac{u_1}{\sqrt{T}}\right)\\ &= \frac{\lambda_T}{T^{1/2}}\left|\frac{1}{\hat{\rho}_I}\right|^{\gamma_1}u_1\left(\left|\rho^* + \frac{u_1}{\sqrt{T}}\right| - |\rho^*|\right)/\left(\frac{u_1}{\sqrt{T}}\right)\\ &\to 0 \text{ in probability} \qquad (20)\end{aligned}$$

since (i): $\lambda_T/T^{1/2} \to 0$, (ii): $\left|1/\hat{\rho}_I\right|^{\gamma_1} \to \left|1/\rho^*\right|^{\gamma_1} < \infty$ in probability and
(iii): $u_1\left(\left|\rho^* + \frac{u_1}{\sqrt{T}}\right| - |\rho^*|\right)/\left(\frac{u_1}{\sqrt{T}}\right) \to u_1\text{sign}(\rho^*)$.

Similarly, if $\beta_j^* \neq 0$

$$\begin{aligned}\lambda_T w_{2j}^{\gamma_2}\left(\left|\beta_j^*+\frac{u_{2j}}{\sqrt{T}}\right|-\left|\beta_j^*\right|\right) &= \lambda_T\left|\frac{1}{\hat{\beta}_{I,j}}\right|^{\gamma_2}\frac{u_{2j}}{\sqrt{T}}\left(\left|\beta_j^*+\frac{u_{2j}}{\sqrt{T}}\right|-\left|\beta_j^*\right|\right)/\left(\frac{u_{2j}}{\sqrt{T}}\right)\\ &= \frac{\lambda_T}{T^{1/2}}\left|\frac{1}{\hat{\beta}_{I,j}}\right|^{\gamma_2}u_{2j}\left(\left|\beta_j^*+\frac{u_{2j}}{\sqrt{T}}\right|-\left|\beta_j^*\right|\right)/\left(\frac{u_{2j}}{\sqrt{T}}\right)\\ &\to 0 \text{ in probability} \end{aligned}\tag{21}$$

since (i): $\lambda_T/T^{1/2} \to 0$, (ii): $\left|1/\hat{\beta}_{I,j}\right|^{\gamma_2} \to \left|1/\beta_j^*\right|^{\gamma_2} < \infty$ in probability and (iii): $u_{2j}\left(\left|\beta_j^*+\frac{u_{2j}}{\sqrt{T}}\right|-\left|\beta_j^*\right|\right)/\left(\frac{u_{2j}}{\sqrt{T}}\right) \to u_{2j}\text{sign}(\beta_j^*)$.

Finally, if $\beta_j^* = 0$,

$$\begin{aligned}\lambda_T w_{2j}^{\gamma_2}\left(\left|\beta_j^*+\frac{u_{2j}}{\sqrt{T}}\right|-\left|\beta_j^*\right|\right) &= \frac{\lambda_T}{T^{1/2}}\left|\frac{1}{\hat{\beta}_{I,j}}\right|^{\gamma_2}\left|u_{2j}\right| = \frac{\lambda_T}{T^{1/2-\gamma_2/2}}\left|\frac{1}{\sqrt{T}\hat{\beta}_{I,j}}\right|^{\gamma_2}\left|u_{2j}\right|\\ &\to \begin{cases}\infty \text{ in probability if } u_{2j} \neq 0\\ 0 \text{ in probability if } u_{2j} = 0\end{cases}\end{aligned}\tag{22}$$

since (i): $\frac{\lambda_T}{T^{1/2-\gamma_2/2}} \to \infty$ and (ii) $\sqrt{T}\hat{\beta}_{I,j}$ is tight.

Putting (19)-(22) together one concludes:

$$\tilde{V}_T(u)\tilde{\rightarrow}\Psi(u) = \begin{cases}u'Qu - 2u'\tilde{B} \text{ if } u_{2j} = 0 \text{ for all } j \in \mathcal{A}^c\\ \infty \text{ if } u_{2j} \neq 0 \text{ for some } j \in \mathcal{A}^c\end{cases}$$

Since $\tilde{V}_T(u)$ is convex and $\Psi(u)$ has a unique minimum it follows from Knight (1999) that $\arg\min \tilde{V}_T(u)\tilde{\rightarrow}\arg\min\Psi(u)$. Hence,

$$(\hat{u}_1, \hat{u}_{2\mathcal{A}}')'\tilde{\rightarrow}N\left(0, \sigma^2[Q_{(1,\mathcal{A}+1)}]^{-1}\right)\tag{23}$$

$$\hat{u}_{2\mathcal{A}^c}\tilde{\rightarrow}\delta_0^{|\mathcal{A}^c|}\tag{24}$$

where δ_0 is the Dirac measure at 0 and $|\mathcal{A}^c|$ is the cardinality of $\mathcal{A}^c$ (hence, $\delta_0^{|\mathcal{A}^c|}$ is the $|\mathcal{A}^c|$-dimensional Dirac measure at 0). Notice that (24) implies that $\hat{u}_{2\mathcal{A}^c} \to 0$ in probability. An equivalent formulation of (23) and (24) is

$$\begin{pmatrix}\sqrt{T}(\hat{\rho}-\rho^*)\\ \sqrt{T}(\hat{\beta}_{\mathcal{A}}-\beta_{\mathcal{A}}^*)\end{pmatrix}\tilde{\rightarrow}N\left(0, \sigma^2[Q_{(1,\mathcal{A}+1)}]^{-1}\right)\tag{25}$$

$$\sqrt{T}(\hat{\beta}_{\mathcal{A}^c}-\beta_{\mathcal{A}^c}^*)\tilde{\rightarrow}\delta_0^{|\mathcal{A}^c|}\tag{26}$$

(25) and (26) establish the consistency part of the theorem at the oracle rate of $\sqrt{T}$. Note that this also implies that for no $j \in \mathcal{A}$ will $\hat{\beta}_j$ be set equal to 0 since for each $j \in \mathcal{A}$, $\hat{\beta}_j$ converges in probability to $\beta_j^* \neq 0$. The same is true for $\hat{\rho}$. (25) also yields the oracle efficient asymptotic distribution, i.e. part (3) of the theorem. It remains to show part (2) of the theorem; $P(\hat{\beta}_{\mathcal{A}^c} = 0) \to 1$. The proof is by contradiction.

Assume $\hat{\beta}_j \neq 0$ for $j \in \mathcal{A}^c$. From the first order conditions

$$2\Delta y'_{-j}(\Delta y - X_T(\hat{\rho}, \hat{\beta}')') + \lambda_T w_{2j}^{\gamma_2} \text{sign}(\hat{\beta}_j) = 0$$

or equivalently,

$$\frac{2\Delta y'_{-j}\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T^{1/2}} + \frac{\lambda_T w_{2j}^{\gamma_2} \text{sign}(\hat{\beta}_j)}{T^{1/2}} = 0$$

First, consider the second term

$$\left|\frac{\lambda_T w_{2j}^{\gamma_2} \text{sign}(\hat{\beta}_j)}{T^{1/2}}\right| = \frac{\lambda_T w_{2j}^{\gamma_2}}{T^{1/2}} = \frac{\lambda_T}{T^{1/2-\gamma_2/2}\left|T^{1/2}\hat{\beta}_{I,j}\right|^{\gamma_2}} \to \infty$$

since $\sqrt{T}\hat{\beta}_{I,j}$is tight. Regarding the first term,

$$\frac{2\Delta y'_{-j}\left(\Delta y - X_T(\hat{\rho}, \hat{\beta}')'\right)}{T^{1/2}} = \frac{2\Delta y'_{-j}\left(\epsilon - X_T[\hat{\rho} - \rho^*, \hat{\beta}' - \beta^{*\prime}]'\right)}{T^{1/2}}$$
$$= \frac{2\Delta y'_{-j}\epsilon}{T^{1/2}} - \frac{2\Delta y'_{-j}X_T\sqrt{T}[\hat{\rho} - \rho^*, \hat{\beta}' - \beta^{*\prime}]'}{T}$$

By (18), $\frac{\Delta y'_{-j}\epsilon}{T^{1/2}} \overset{\sim}{\to} N(0, \sigma^2 Q_{(j+1)})$ where in accordance with previous notation $Q_{(j+1)}$ is the $(j+1)$th diagonal element of Q. $\frac{\Delta y'_{-j}X_T}{T} \xrightarrow{p} (Q_{(j+1,1)}, ..., Q_{(j+1,p+1)})$ by (17). Hence, $\frac{\Delta y'_{-j}\epsilon}{T^{1/2}}$ and $\frac{\Delta y'_{-j}X_T}{T}$ are tight. The same is the case for $\sqrt{T}[\hat{\rho}-\rho^*, \hat{\beta}'-\beta^{*\prime}]$ since it converges weakly by (25)-(26). Hence,

$$P(\hat{\beta}_j \neq 0) \leq P\left(\frac{2\Delta y'_{-j}\left(\Delta y - X_T(\hat{\rho}_T, \hat{\beta}'_T)'\right)}{T^{1/2}} + \frac{\lambda_T w_{2j}^{\gamma_2} \text{sign}(\hat{\rho}_T)}{T^{1/2}} = 0\right) \to 0$$

□

Before proving Theorem 3 we prove the following lemma. Let $(x)_+ = \max(x, 0)$.

Lemma 1. *Let* $g : \mathbb{R} \to \mathbb{R}$ *be given by* $g(u) = u^2 - 2au + 2\lambda|u|$, $\lambda \geq 0$, $a \neq 0$. *Then* $\arg\min g = 0$ *if and only if* $\lambda \geq |a|$. *More precisely,* $\arg\min g = \text{sign}(a)\left(|a| - \lambda\right)_+$.

Proof. Assume $a > 0$. Since $g'(u) = 2u - 2a + 2\lambda\text{sign}(u)$ is strictly negative for $u < 0$ $\arg\min g \in [0, \infty)$. $\tilde{u} > 0$ is a local minimum (and hence a global minimum since g is strictly convex) if and only if it is a stationary point, i.e. $g'(\tilde{u}) = 2\tilde{u} - 2a + 2\lambda = 0$ which is equivalent to $0 < \tilde{u} = a - \lambda = |a| - \lambda$. This contradicts $\lambda \geq |a|$ and so $\arg\min g = 0$ when $\lambda \geq |a|$. In total, the above shows that $\arg\min g = (a - \lambda)_+ = \text{sign}(a)\left(|a| - \lambda\right)_+$ for $a > 0$. Similar arguments establish the result for $a < 0$.

□

Proof of Theorem 3. $\hat{\rho}$ minimizes

$$L(\rho)=\sum_{t=1}^{T}(\Delta y_t-\rho y_{t-1})^2+2\lambda_T\frac{|\rho|}{\left|\hat{\rho}_I\right|}=\sum_{t=1}^{T}\Delta y_t^2+\rho^2\sum_{t=1}^{T}y_{t-1}^2-2\rho\sum_{t=1}^{T}\Delta y_t y_{t-1}+2\lambda_T\frac{|\rho|}{\left|\hat{\rho}_I\right|}$$

which is equvalent to minimizing

$$\rho^2-2\rho\frac{\sum_{t=1}^{T}\Delta y_t y_{t-1}}{\sum_{t=1}^{T}y_{t-1}^2}+2\lambda_T\frac{|\rho|}{\left|\hat{\rho}_I\right|\sum_{t=1}^{T}y_{t-1}^2}=\rho^2-2\rho\hat{\rho}_I+2\lambda_T\frac{|\rho|}{\left|\hat{\rho}_I\right|\sum_{t=1}^{T}y_{t-1}^2}$$

It follows from Lemma 1 that $\hat{\rho}=0$ if and only if

$$\left|\hat{\rho}_I\right|\le\frac{\lambda_T}{\left|\hat{\rho}_I\right|\sum_{t=1}^{T}y_{t-1}^2}\Leftrightarrow\hat{\rho}_I{}^2\sum_{t=1}^{T}y_{t-1}^2\le\lambda_T$$

Hence, recalling that $\hat{\rho}_I=\sum_{t=1}^{T}\Delta y_t y_{t-1}/\sum_{t=1}^{T}y_{t-1}^2$ (the least squares estimator)

$$\begin{aligned}P(\hat{\rho}=0)&=P\left(\hat{\rho}_I^2\sum_{t=1}^{T}y_{t-1}^2\le\lambda_T\right)=P\left(\left[\frac{\sum_{t=1}^{T}\Delta y_t y_{t-1}}{\sum_{t=1}^{T}y_{t-1}^2}\right]^2\sum_{t=1}^{T}y_{t-1}^2\le\lambda_T\right)\\&=P\left(\left[\rho^*+\frac{\sum_{t=1}^{T}y_{t-1}\epsilon_t}{\sum_{t=1}^{T}y_{t-1}^2}\right]^2\sum_{t=1}^{T}y_{t-1}^2\le\lambda_T\right)\\&=P\left(\left[\rho^{*2}+\left[\frac{\sum_{t=1}^{T}y_{t-1}\epsilon_t}{\sum_{t=1}^{T}y_{t-1}^2}\right]^2+2\rho^*\frac{\sum_{t=1}^{T}y_{t-1}\epsilon_t}{\sum_{t=1}^{T}y_{t-1}^2}\right]\sum_{t=1}^{T}y_{t-1}^2\le\lambda_T\right)\end{aligned}$$

□

Proof of Theorem 4. From Phillips (1987a) one has

$$\left(\frac{1}{T}\sum_{t=1}^{T}y_{t-1}\epsilon_t,\frac{1}{T^2}\sum_{t=1}^{T}y_{t-1}^2\right)\tilde{\to}\left(\sigma^2\int_0^1W_sdW_s,\sigma^2\int_0^1W_s^2ds\right)\tag{27}$$

Using Theorem 3 with $\rho^*=0$ yields

$$P(\hat{\rho}=0)=P\left(\left[\frac{\frac{1}{T}\sum_{t=1}^{T}y_{t-1}\epsilon_t}{\frac{1}{T^2}\sum_{t=1}^{T}y_{t-1}^2}\right]^2\frac{1}{T^2}\sum_{t=1}^{T}y_{t-1}^2\le\lambda_T\right)$$

From (27) and the continuous mapping theorem it follows that

$$G_T:=\left[\frac{\frac{1}{T}\sum_{t=1}^{T}y_{t-1}\epsilon_t}{\frac{1}{T^2}\sum_{t=1}^{T}y_{t-1}^2}\right]^2\frac{1}{T^2}\sum_{t=1}^{T}y_{t-1}^2\tilde{\to}\left[\frac{\int_0^1W_sdW_s}{\int_0^1W_s^2ds}\right]^2\sigma^2\int_0^1W_s^2ds:=G\tag{28}$$

where the last definition means that G is a random variable distributed as $\left[\frac{\int_0^1W_sdW_s}{\int_0^1W_s^2ds}\right]^2\sigma^2\int_0^1W_s^2ds$.

Case 1: $\lambda_T \to 0$. Since the right hand side in (28) is absolutely continuous with respect to the Lebesgue measure it has no mass points and so

$$P(\hat{\rho}=0) = P\left(G_T \le \lambda_T\right) = F_{G_T}(\lambda_T) \to F_G(0) = 0$$

if $\lambda_T \to 0$.
Case 2:$\lambda_T \to \lambda \in (0,\infty)$. By the same reasoning as in case 1 it follows that

$$P(\hat{\rho}=0) = P\left(G_T \le \lambda_T\right) = F_{G_T}(\lambda_T) \to F_G(\lambda) = p \in (0,1)$$

since G is supported on all of $\mathbb{R}_+$.
Case 3: $\lambda_T \to \infty$. Since G_T converges weakly it is tight and the result follows. □

Proof of Theorem 5. By standard results (see e.g. Hamilton (1994))

$$\frac{1}{T^{1/2}}\sum_{t=1}^{T} y_{t-1}\epsilon_t \tilde{\rightarrow} N(0,\sigma^2 E[y_{t-1}^2])$$
$$\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2 \xrightarrow{p} E(y_{t-1}^2)$$

Using Theorem 3 with $\rho^* \in (-2,0)$ yields

$$P(\hat{\rho}=0) = P\left(\left[T\rho^{*2} + \left[\frac{\frac{1}{T^{1/2}}\sum_{t=1}^{T} y_{t-1}\epsilon_t}{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2}\right]^2 + 2T^{1/2}\rho\frac{\frac{1}{T^{1/2}}\sum_{t=1}^{T} y_{t-1}\epsilon_t}{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2}\right]\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2 \le \lambda_T\right)$$
$$= P\left(\left[\rho^{*2} + \frac{1}{T}\left[\frac{\frac{1}{T^{1/2}}\sum_{t=1}^{T} y_{t-1}\epsilon_t}{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2}\right]^2 + 2\frac{1}{T^{1/2}}\rho\frac{\frac{1}{T^{1/2}}\sum_{t=1}^{T} y_{t-1}\epsilon_t}{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2}\right]\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2 \le \frac{\lambda_T}{T}\right)$$

Since $\frac{\frac{1}{T^{1/2}}\sum_{t=1}^{T} y_{t-1}\epsilon_t}{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2}$ is tight it follows that

$$\frac{1}{T}\left[\frac{\frac{1}{T^{1/2}}\sum_{t=1}^{T} y_{t-1}\epsilon_t}{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2}\right]^2 + 2\frac{1}{T^{1/2}}\rho\frac{\frac{1}{T^{1/2}}\sum_{t=1}^{T} y_{t-1}\epsilon_t}{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2} \xrightarrow{p} 0$$

which implies that

$$H_T := \left[\rho^{*2} + \frac{1}{T}\left[\frac{\frac{1}{T^{1/2}}\sum_{t=1}^{T} y_{t-1}\epsilon_t}{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2}\right]^2 + 2\frac{1}{T^{1/2}}\rho\frac{\frac{1}{T^{1/2}}\sum_{t=1}^{T} y_{t-1}\epsilon_t}{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2}\right]\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2 \xrightarrow{p} \rho^{*2}E(y_{t-1}^2) := L$$

Case 1: $\lambda_T/T \to 0$. Since H_T converges in probability to L it follows that

$$P(\hat{\rho}=0) = P(H_T \le \lambda_T/T) \le P(H_T \le L - L/2) \to 0$$

where the estimate holds for T sufficiently large since $\lambda_T/T \to 0$.

Case 2: $\lambda_T/T \to \lambda \in (0, \infty)$.

$$P(\hat{\rho} = 0) = P\left(H_T \leq \lambda_T\right) \begin{cases} \leq P(H_T \leq \lambda + (L-\lambda)/2) \leq P(H_T \leq L - (L-\lambda)/4) \to 0, \ \lambda < L \\ \geq P(H_T \leq \lambda - (\lambda - L)/2) \geq P(H_T \leq L + (\lambda - L)/4) \to 1, \ \lambda > L \end{cases}$$

where the first estimate in each of the cases holds from a certain step and onwards.

Case 3: $\lambda_T/T \to \infty$. Since H_T converges in probability it is tight and the result follows.

□

Proof of Theorem 6. By Phillips (1987b)

$$\left(\frac{1}{T}\sum_{t=1}^{T} y_{t-1}\epsilon_t, \frac{1}{T^2}\sum_{t=1}^{T} y_{t-1}^2\right) \tilde{\to} \left(\sigma^2 \int_0^1 J_c(r)dW(r), \sigma^2 \int_0^1 J_c^2(r)dr\right) \tag{29}$$

where J_c is the Ornstein-Uhlenbeck process with parameter c. Notice how the only difference to (27) is that the integrand process now is $J_c(r)$ instead of $W(r)$. For $c = 0$ they are identical as the Ornstein-Uhlenbeck process collapses to the Wiener process.

From Theorem 3 with $\rho^* = c/T$ it follows

$$\begin{aligned} P(\hat{\rho} = 0) &= P\left(\left[(c/T)^2 + \left[\frac{\sum_{t=1}^T y_{t-1}\epsilon_t}{\sum_{t=1}^T y_{t-1}^2}\right]^2 + 2c/T\frac{\sum_{t=1}^T y_{t-1}\epsilon_t}{\sum_{t=1}^T y_{t-1}^2}\right]\sum_{t=1}^{T} y_{t-1}^2 \leq \lambda_T\right) \\ &= P\left(\left[c^2 + \left[\frac{\frac{1}{T}\sum_{t=1}^T y_{t-1}\epsilon_t}{\frac{1}{T^2}\sum_{t=1}^T y_{t-1}^2}\right]^2 + 2c\frac{\frac{1}{T}\sum_{t=1}^T y_{t-1}\epsilon_t}{\frac{1}{T^2}\sum_{t=1}^T y_{t-1}^2}\right]\frac{1}{T^2}\sum_{t=1}^{T} y_{t-1}^2 \leq \lambda_T\right) \end{aligned}$$

By the continuous mapping theorem

$$\begin{aligned} K_T &:= \left[c^2 + \left[\frac{\frac{1}{T}\sum_{t=1}^T y_{t-1}\epsilon_t}{\frac{1}{T^2}\sum_{t=1}^T y_{t-1}^2}\right]^2 + 2c\frac{\frac{1}{T}\sum_{t=1}^T y_{t-1}\epsilon_t}{\frac{1}{T^2}\sum_{t=1}^T y_{t-1}^2}\right]\frac{1}{T^2}\sum_{t=1}^{T} y_{t-1}^2 \\ &\tilde{\to} \left[c^2 + \left[\frac{\int_0^1 J_c(r)dW(r)}{\int_0^1 J_c^2(r)dr}\right]^2 + 2c\frac{\int_0^1 J_c(r)dW(r)}{\int_0^1 J_c^2(r)dr}\right]\sigma^2\int_0^1 J_c^2(r)dr := K \end{aligned}$$

where the last definition means that K is a random variable distributed as the weak limit of K_T.

Case 1: $\lambda_T \to 0$. Since K is absolutely continuous with respect to the Lebesgue measure it has no mass points and so

$$P(\hat{\rho} = 0) = P(K_T \leq \lambda_T) = F_{K_T}(\lambda_T) \to F_K(0) = 0$$

Case 2: $\lambda_T \to \lambda \in (0, \infty)$. By the same reasoning as in Case 1 it follows that

$$P(\hat{\rho} = 0) = P(K_T \leq \lambda_T) = F_{K_T}(\lambda_T) \to F_K(\lambda) \in (0, 1)$$

since K is supported on all of $\mathbb{R}_+$.

Case 3: $\lambda_T \to \infty$. Since K_T converges weakly it is tight and the result follows.

□

Proof of Theorem 7. The setting is the same as in the proof of Theorem 1. Follow the proof of that theorem, with identical notation, until (7) with $\gamma_1 = \gamma_2 = 1$. Next, notice that

$$\lambda_T w_1 \left|\frac{u_1}{T}\right| = \lambda_T \frac{1}{|\hat{\rho}_I|}\left|\frac{u_1}{T}\right| = |u_1|\,\lambda_T \frac{1}{|T\hat{\rho}_I|} \tilde{\to} \lambda \frac{|u_1|}{|C_1|} \tag{30}$$

by (5) and (6) since C_1 has no mass at 0.

Furthermore, if $\beta_j^* \neq 0$

$$\begin{aligned} \lambda_T w_{2j}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) &= \lambda_T \left|\frac{1}{\hat{\beta}_{I,j}}\right| \frac{u_{2j}}{\sqrt{T}} \left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) / \left(\frac{u_{2j}}{\sqrt{T}}\right) \\ &= \frac{\lambda_T}{T^{1/2}}\left|\frac{1}{\hat{\beta}_{I,j}}\right| u_{2j} \left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) / \left(\frac{u_{2j}}{\sqrt{T}}\right) \\ &\to 0 \text{ in probability} \end{aligned} \tag{31}$$

since (i): $\lambda_T / T^{1/2} \to 0$, (ii): $\left|1/\hat{\beta}_{I,j}\right| \to \left|1/\beta_j^*\right| < \infty$ in probability and (iii): $u_{2j}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) / \left(\frac{u_{2j}}{\sqrt{T}}\right) \to u_{2j}\text{sign}(\beta_j^*)$.

Finally, if $\beta_j^* = 0$,

$$\lambda_T w_{2j}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) = \frac{\lambda_T}{T^{1/2}}\left|\frac{1}{\hat{\beta}_{I,j}}\right| \left|u_{2j}\right| = \lambda_T \left|\frac{1}{\sqrt{T}\hat{\beta}_{I,j}}\right| \left|u_{2j}\right| \tilde{\to} \lambda \frac{|u_{2j}|}{|C_{2j}|} \tag{32}$$

by (5) and (6) since (i): $\lambda_T \to \lambda$ and (ii): C_{2j} is 0 with probability 0 such that $x \mapsto \left|1/x\right|$ is continuous almost everywhere with respect to the limiting measure.

Putting together (7) and(30)-(32) one concludes

$$V_T(u) \tilde{\to} u'Au - 2u'B + \lambda \frac{|u_1|}{|C_1|} + \lambda \sum_{j=1}^{p} \frac{|u_{2j}|}{|C_{2j}|} \mathbf{1}_{\left\{\beta_j^* = 0\right\}} := \Psi(u)$$

Hence, since $V_T(u)$ is convex and $\Psi(u)$ has a unique minimum it follows from Knight (1999) that $\arg\min V_T(u) \tilde{\to} \arg\min \Psi(u)$

□

Proof of Theorem 8. The setting is the same as in the proof of Theorem 2. Follow the proof of that theorem, with identical notation, until (21) with $\gamma_1 = \gamma_2 = 1$. For the case of $\beta_j^* = 0$ one has

$$\lambda_T w_{2j}\left(\left|\beta_j^* + \frac{u_{2j}}{\sqrt{T}}\right| - \left|\beta_j^*\right|\right) = \frac{\lambda_T}{T^{1/2}}\left|\frac{1}{\hat{\beta}_{I,j}}\right| \left|u_{2j}\right| = \lambda_T \left|\frac{1}{\sqrt{T}\hat{\beta}_{I,j}}\right| \left|u_{2j}\right| \tilde{\to} \lambda \frac{|u_{2j}|}{|\tilde{C}_{2j}|} \tag{33}$$

by (17) and (18) since (i): $\lambda_T \to \lambda$, (ii): $\tilde{C}_{2j}$ is 0 with probability 0 such that $x \mapsto \left|1/x\right|$ is continuous almost everywhere with respect to the limiting measure.

Putting together (19)-(21) and (33) one concludes

$$\tilde{V}_T(u) \tilde{\to} u'Qu - 2u'\tilde{B} + \lambda \sum_{j=1}^{p} \frac{|u_{2j}|}{|\tilde{C}_{2j}|} \mathbf{1}_{\left\{\beta_j^* = 0\right\}} := \tilde{\Psi}(u)$$

Hence, since $\tilde{V}_T(u)$ is convex and $\tilde{\Psi}(u)$ has a unique minimum it follows from Knight (1999) that $\arg\min \tilde{V}_T(u) \tilde{\rightarrow} \arg\min \tilde{\Psi}(u)$

□

References

Bühlmann, P. and S. Van De Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer-Verlag, New York.

Candes, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*, 2313–2351.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of statistics 32*, 407–499.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*, 849–911.

Hamilton, J. D. (1994). *Time Series Analysis*. Cambridge University Press, Cambridge.

Huang, J., J. L. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics 36*, 587–613.

Knight, K. (1999). Epi-convergence in distribution and stochastic equi-semicontinuity. *Unpublished manuscript*.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, 1356–1378.

Knight, K. and W. Fu (2011). An alternative to unit root tests: Bridge estimators differentiate between nonstationary versus stationary models and select optimal lag. *Working paper*.

Kock, A. B. (2012). Oracle efficient variable selection in random and fixed effects panel data models. *Econometric Theory (forthcoming)*.

Leeb, H. and B. Pötscher (2008). Sparse estimators and the oracle property, or the return of hodges' estimator. *Journal of Econometrics 142*, 201–211.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*, 21–59.

Lehmann, E. L. and G. Casella (1998). *Theory of point estimation*, Volume 31. Springer Verlag.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*, 1436–1462.

Ng, S. and P. Perron (2001). Lag length selection and the construction of unit root tests with good size and power. *Econometrica 69*, 1519–1554.

Phillips, P. C. B. (1987a). Time series regression with a unit root. *Econometrica*, 277–301.

Phillips, P. C. B. (1987b). Towards a unified asymptotic theory for autoregression. *Biometrika 74*, 535–547.

Pötscher, B. and H. Leeb (2009). On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. *Journal of Multivariate Analysis 100*, 2065–2082.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Wang, H., G. Li, and C. L. Tsai (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*, 63–78.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research 7*, 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

Research Papers 2012

CREATES
Center for Research in Econometric Analysis of Time Series

2011-43: Peter Christoffersen, Ruslan Goyenko, Kris Jacobs, Mehdi Karoui: Illiquidity Premia in the Equity Options Market

2011-44: Diego Amaya, Peter Christoffersen, Kris Jacobs and Aurelio Vasquez: Do Realized Skewness and Kurtosis Predict the Cross-Section of Equity Returns?

2011-45: Peter Christoffersen and Hugues Langlois: The Joint Dynamics of Equity Market Factors

2011-46: Peter Christoffersen, Kris Jacobs and Bo Young Chang: Forecasting with Option Implied Information

2011-47: Kim Christensen and Mark Podolskij: Asymptotic theory of range-based multipower variation

2011-48: Christian M. Dahl, Daniel le Maire and Jakob R. Munch: Wage Dispersion and Decentralization of Wage Bargaining

2011-49: Torben G. Andersen, Oleg Bondarenko and Maria T. Gonzalez-Perez: Coherent Model-Free Implied Volatility: A Corridor Fix for High-Frequency VIX

2011-50: Torben G. Andersen and Oleg Bondarenko: VPIN and the Flash Crash

2011-51: Tim Bollerslev, Daniela Osterrieder, Natalia Sizova and George Tauchen: Risk and Return: Long-Run Relationships, Fractional Cointegration, and Return Predictability

2011-52: Lars Stentoft: What we can learn from pricing 139,879 Individual Stock Options

2011-53: Kim Christensen, Mark Podolskij and Mathias Vetter: On covariation estimation for multivariate continuous Itô semimartingales with noise in non-synchronous observation schemes

2012-01: Matei Demetrescu and Robinson Kruse: The Power of Unit Root Tests Against Nonlinear Local Alternatives

2012-02: Matias D. Cattaneo, Michael Jansson and Whitney K. Newey: Alternative Asymptotics and the Partially Linear Model with Many Regressors

2012-03: Matt P. Dziubinski: Conditionally-Uniform Feasible Grid Search Algorithm

2012-04: Jeroen V.K. Rombouts, Lars Stentoft and Francesco Violante: The Value of Multivariate Model Sophistication: An Application to pricing Dow Jones Industrial Average options

2012-05: Anders Bredahl Kock: On the Oracle Property of the Adaptive LASSO in Stationary and Nonstationary Autoregressions