

Consistent and Conservative Model Selection with 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 nonstationary 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 nonstationary autoregressions and it thereby constitutes an addition to the set of unit root tests. Next, and important in practice, we show that choosing the tuning parameter by BIC results in consistent model selection.

However, it is also shown that the adaptive Lasso has no power against shrinking alternatives of the form c/T if it is tuned to perform consistent model selection. We show that if the adaptive Lasso is tuned to perform conservative model selection it has power even against shrinking alternatives of this form and compare it to the plain Lasso.

Keywords: adaptive Lasso, Autoregressions, Conservative model selection, Consistent model selection, Oracle efficiency, Time Series.

JEL classification: C13, C18, 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. As traditional methods are computationally infeasible if the number of covariates is large, focus has been on penalized or shrinkage

I am indebted to Svend Erik Graversen, Niels Haldrup and Jørgen Hoffmann-Jørgensen, and Timo Teräsvirta for help and suggestions. The paper has also benefitted greatly from many constructive comments by three anonymous referees as well as the editors Victor Chernozhukov and Peter Phillips. Part of this research was carried out while I was visiting the Australian National University in 2011. I wish to thank Tue Gørgens and the Research School of Economics for inviting me and creating a pleasant environment. I also wish to thank CREATES, funded by the Danish National Research Foundation, for providing financial support.

type 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, of which Knight and Fu (2000) were among the first. 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. The range of applications of these estimators is widening rapidly, for example Caner (2011) has used the bridge estimator to select factors in approximate factor models. For recent reviews see Bühlmann and van de Geer (2011) and Belloni and Chernozhukov (2011). A lot of focus has been on establishing the oracle property. This entails showing that the estimators are consistent, perform correct variable selection and that the limiting distribution of the non-zero coefficients is the same as if only the relevant variables had been included in the model. Put differently, the inference is as efficient as if an oracle had revealed the true model and estimation had been carried out using only the relevant variables.

Most focus in the literature has been on establishing the oracle property for either deterministic covariates or in an *i.i.d* setting. Exceptions are Wang et al. (2007) who consider the Lasso for stationary autoregressions while Ren and Zhang (2010) have considered a variant of the adaptive Lasso similar to ours for stationary vector autoregressions in the case we shall refer to as consistent model selection. Medeiros and Mendes (2012) have investigated the properties of the adaptive Lasso in linear stationary models while Liao and Phillips (2013) have used shrinkage estimation in a clever way in the context of the cointegrated VAR model. In particular, Liao and Phillips (2013) show how shrinkage estimation can be used for simultaneous selection of the cointegration rank and the autoregressive order. Thus, the work in Section 6 of their paper actually contains our Theorem 1a) below as a special case. Oracle inequalities for high-dimensional stationary VAR models have been established in Kock and Callot (2012).

In this paper we show that the adaptive Lasso of Zou (2006) possesses the oracle property in stationary as well as nonstationary autoregressions which are workhorse models in empirical macroeconomics. For example, they are frequently applied as benchmark models in forecast competitions. We shall consider a model of the form

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

Equation (1) is sometimes called a Dickey-Fuller regression¹ where ϵ_t is the error term to be discussed further in the next section and (1) is said to have a unit root, and hence a fortiori to be nonstationary, if $\rho^* = 0$. When testing for a unit root, one usually first determines the number of lagged differences to be included. This can be done either by information criteria, or modifications thereof, see 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

¹The Dickey-Fuller form of the autoregression is simply a reparameterization of the usual $y_t = \sum_{j=1}^{p+1} \phi_j y_{t-j} + \epsilon_t$ which facilitates the comparison between stationary and nonstationary autoregressions. See Hamilton (1994) for the exact connection between the two ways of writing the model

adaptive Lasso means that we can choose the number of lagged differences to be included (and leave out irrelevant intermediate lags) and test for a unit root at the same time. Caner and Knight (2013) 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. Liao and Phillips (2013) took this point further and used it in the context of the cointegrated VAR model.

We show: (i) The adaptive Lasso possesses the oracle property in stationary and nonstationary autoregressions. Hence, it can distinguish between stationary and nonstationary autoregressions which is extremely important when choosing the right model for forecasting. (ii) Show that choosing the tuning parameter by BIC results in consistent model selection. (iii) Analyze the asymptotic behavior of the probability of classifying ρ^* as 0 in the stationary, nonstationary and local to unity setting ($\rho^* = c/T$). The local to unity setting reveals that in a time series setting the adaptive Lasso is not exempt from the critique by Pötscher and Schneider (2009). (iv) This problem, due to nonuniformity in the asymptotics, can be alleviated if one is willing to tune the adaptive Lasso to perform conservative model selection instead of consistent model selection². (v) The properties of conservative model selection are investigated in the stationary as well as the nonstationary setting and compared to that of the plain Lasso.

2 Setup and Notation

We employ the following variant of the adaptive Lasso of Zou (2006) which is defined as the minimizer of

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

where $w_1 = 1/|\hat{\rho}_I|$ and $w_{2j} = 1/|\hat{\beta}_{I,j}|$ for $j = 1, \dots, p$ and $\hat{\rho}_I$ and $\hat{\beta}_{I,j}$ denote some initial estimators of the parameters in (2). We shall use the least squares estimator in this paper but other estimators can be used as well. Notice that the objective function (2) is modified compared to the usual adaptive Lasso since it penalizes ρ , the coefficient on the potentially nonstationary variable y_t , different from the coefficients on the stationary variables.

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

2.1 Notation

We consider $T + p + 1$ observations from a time series y_t generated by (1). The $p + 1$ initial values are all assumed to be zero as in Hamilton (1994). For other choices of initial values we refer to Phillips (1987a). Let $\eta^* = (\rho^*, \beta^{*'})'$. $\mathcal{B} = \{1 \leq j \leq p + 1 : \eta_j^* \neq 0\}$ denotes the active set of variables. Similarly, denote by $\mathcal{A} = \{1 \leq j \leq p : \beta_j^* \neq 0\}$ the active set of lagged differences. Let $z_t = (\Delta y_{t-1}, \dots, \Delta y_{t-p})'$ be the $(p \times 1)$ vector of lagged differences and let

²We shall make precise definitions of consistent and conservative model selection in Section 2.

$x_t = (y_{t-1}, z_t')'$. Let $\Sigma = E(z_t z_t')$ ³. For any square matrix A and sets \mathcal{R} and \mathcal{S} , let $A_{\mathcal{R}, \mathcal{S}}$ denote the matrix which consists of all rows of A indexed \mathcal{R} and columns indexed by \mathcal{S} . If $\mathcal{R} = \mathcal{S}$ we write $A_{\mathcal{S}}$ for short. Similarly, \mathcal{S} indexes vectors by picking out the elements with index in \mathcal{S} . $|\mathcal{S}|$ denotes the cardinality of \mathcal{S} . For any $x \in \mathbb{R}^n$, $\|x\|_{\ell_2} = \sqrt{\sum_{i=1}^n x_i^2}$.

Let $\Delta y = (\Delta y_T, \dots, \Delta y_1)'$, $y_{-1} = (y_{T-1}, \dots, y_0)'$ and $\Delta y_{-j} = (\Delta y_{T-j}, \dots, \Delta y_{1-j})'$, $j = 1, \dots, p$ ⁴. Let $X_T = (y_{-1}, \Delta y_{-1}, \dots, \Delta y_{-p})$ be the $T \times (p+1)$ matrix of covariates and $\epsilon = (\epsilon_T, \dots, \epsilon_1)'$ the vector of error terms.

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

$S_T = \text{diag}(T, \sqrt{T}, \dots, \sqrt{T})$ denotes a $(p+1) \times (p+1)$ scaling matrix, $\tilde{\rightarrow}$ denotes weak convergence (convergence in law) and \xrightarrow{P} convergence in probability. Let $\hat{\eta} = (\hat{\rho}, \hat{\beta}')'$ denote the minimizer of (2). Of course this depends on λ_T but to keep notation simple we suppress this except for in Section 3.1 dealing with tuning parameter selection.

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^*) , $P(\hat{\mathcal{M}} = \mathcal{M}_0) \rightarrow 1$. A procedure is said to be conservative if for all (ρ^*, β^*) , $P(\mathcal{M}_0 \not\subseteq \hat{\mathcal{M}}) \rightarrow 0$, i.e. the probability of excluding relevant variables tends to zero. This is clearly a weaker condition than consistent model selection but still a relevant measure of performance since it is of importance to at least keep the relevant variables in the model if one can not select exactly the right ones.

3 Oracle property

This section establishes and discusses the oracle property of the adaptive Lasso for stationary and nonstationary autoregressions. The results open the possibility to use the adaptive Lasso to distinguish between these two and hence it can also be used as a new way of testing for unit roots.

Theorem 1 (Consistent model selection). *Assume that ϵ_t is i.i.d with $E(\epsilon_1) = 0$ and $E(\epsilon_1^4) < \infty$.*

a) (Nonstationary case): *Then, if $\rho^* = 0$, $\frac{\lambda_T}{T^{1-\gamma_1}} \rightarrow \infty$, $\frac{\lambda_T}{T^{1/2-\gamma_2/2}} \rightarrow \infty$, and $\frac{\lambda_T}{T^{1/2}} \rightarrow 0$*

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

2. Oracle (i): $P(\hat{\rho} = 0) \rightarrow 1$ and $P(\hat{\beta}_{\mathcal{A}^c} = 0) \rightarrow 1$

3. Oracle (ii): $\sqrt{T}(\hat{\beta}_{\mathcal{A}} - \beta_{\mathcal{A}}) \tilde{\rightarrow} N(0, \sigma^2[\Sigma_{\mathcal{A}}]^{-1})$

b) (Stationary case): *Then, if y_t is stationary⁵ such that $\rho^* \neq 0$, $\frac{\lambda_T}{T^{1/2-\gamma_2/2}} \rightarrow \infty$, and $\frac{\lambda_T}{T^{1/2}} \rightarrow 0$*

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

³Of course the actual value of this expectation depends on whether y_t is stationary or not. However, irrespective of this z_t is stationary.

⁴The dependence on T is suppressed for some of the quantities where no confusion arises.

⁵In our notation this corresponds to all roots of $(1-z) - \rho^* z - \sum_{j=1}^p (1-z)z^j \beta_j^*$ satisfying $|z| > 1$. This rules out $\rho^* = 0$.

2. Oracle (i): $P(\hat{\rho} = 0) \rightarrow 0$ and $P(\hat{\beta}_{\mathcal{A}^c} = 0) \rightarrow 1$

3. Oracle (ii): $\begin{pmatrix} \sqrt{T}(\hat{\rho} - \rho^*) \\ \sqrt{T}(\hat{\beta}_{\mathcal{A}} - \beta_{\mathcal{A}}^*) \end{pmatrix} \rightsquigarrow N(0, \sigma^2[Q_{\mathcal{B}}]^{-1})$

where $Q = E(x_t x_t')$ of dimension $(p+1 \times p+1)$

For a proof of Theorem 1 see the supplementary material. The assumption that ϵ_t is *i.i.d* can be relaxed as long as $S_T^{-1} X_T' X_T S_T^{-1}$ and $S_T^{-1} X_T' \epsilon$ converge weakly. Requiring $\frac{\lambda_T}{T^{1-\gamma_1}} \rightarrow \infty$ in part a) enables us to set $\hat{\rho} = 0$ with probability tending to one if $\rho^* = 0$. Likewise, $\frac{\lambda_T}{T^{1/2-\gamma_2/2}} \rightarrow \infty$ is needed to shrink the estimates of truly zero β_j^* s to zero. $\frac{\lambda_T}{T^{1/2}} \rightarrow 0$ on the other hand requires 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 of part a) to be satisfied simultaneously we need $\gamma_1 > 1/2$ and $\gamma_2 > 0$ while part b) does not put any restrictions on γ_1 . The reason for the different requirements on γ_1 and γ_2 is that $\hat{\rho}_I$ converges at the rate $1/T$ while $\hat{\beta}_{I,j}$ converges at the rate $1/\sqrt{T}$.

The consistency part of Theorem 1a) reveals that $\hat{\rho}$ and $\hat{\beta}$ are estimated consistently at rates $1/T$ and $1/\sqrt{T}$, respectively. In case of stationarity the rate of consistency is $1/\sqrt{T}$ for all parameters. Furthermore, the estimates of the zero coefficients do not 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 simultaneously in the stationary as well as the non-stationary setting. Finally, in both settings the asymptotic distribution of the estimators of the nonzero coefficients is the same as if the true model had been known and only the relevant variables had been included and least squares applied to that model. Thus, the adaptive Lasso is oracle efficient and can be used to discriminate between stationary and nonstationary autoregressions and this opens the possibility to use it for unit root testing. This sounds almost too good to be true – and in some sense it is as we shall see in Section 4.

As already mentioned in the Introduction the results of Theorem 1b) are as expected in the light of the work of Ren and Zhang (2010) but we have chosen to include them for the sake of completeness and comparison with the nonstationary case. More precisely, the assumptions of Theorem 1 correspond to the assumptions Ren and Zhang (2010) make in their Theorem 3.1.

Liao and Phillips (2013) have shown that consistent model selection is possible by means of shrinkage even in the context of a VAR model. In particular, the work in Section 6 of their paper on adaptive tuning parameter selection extends Theorem 1a) to the multivariate case and thus contains our result as a special case.

3.1 Tuning parameter selection

Theorem 1 established the oracle property of the adaptive Lasso when the tuning parameter λ_T grows at the right rate. However, this rate provides only limited practical guidance towards the choice of λ_T . We suggest using BIC. To be precise, let $\hat{\epsilon}_\lambda = \Delta y_t - \hat{\rho}_\lambda y_{t-1} - \sum_{j=1}^p \hat{\beta}_{\lambda,j} \Delta y_{t-j}$ be the error term resulting from the adaptive Lasso with tuning parameter λ and corresponding parameter estimate $(\hat{\rho}_\lambda, \hat{\beta}_\lambda)'$. $\hat{\mathcal{B}}_\lambda = \{1 \leq j \leq p+1 : \hat{\eta}_{\lambda,j} \neq 0\}$ denotes the indices of the

coefficients estimated to be non-zero. The BIC minimizes

$$BIC_\lambda = \log \left(\frac{\hat{\epsilon}'_\lambda \hat{\epsilon}_\lambda}{T} \right) + |\hat{\mathcal{B}}_\lambda| \frac{\log(T)}{T}$$

with respect to λ . Denote this minimizer by $\hat{\lambda}_{BIC}$.

Theorem 2. *Let the assumptions of Theorem 1a) or 1b) be satisfied. Then, as $T \rightarrow \infty$,*

$$P \left(\hat{\mathcal{B}}_{\hat{\lambda}_{BIC}} = \mathcal{B} \right) \rightarrow 1.$$

Theorem 2 justifies choosing the tuning parameter by BIC, as this choice leads to consistent variable selection irrespective of whether y_t is stationary or not. In the language of unit root testing this means that we can distinguish between $\rho^* = 0$ and $\rho^* \neq 0$. Thus, the simulation results reported in the supplementary material use BIC to choose the tuning parameter. The proof of Theorem 2 is deferred to the supplementary material.

4 Local to zero analysis

In this section we analyse the behaviour of $P(\hat{\rho} = 0)$ for $\rho^* = 0$, $\rho^* \in (-2, 0)$ and $\rho^* = c/T$ for some $c \neq 0$. To be precise, we give a complete description of the limits of $P(\hat{\rho} = 0)$ for all possible limiting values of the regularization parameter λ_T and illustrate a shortcoming of the adaptive Lasso tuned to perform consistent variable selection. This is most conveniently done in the AR(1) model to keep the focus on the main points and the expressions simple.

Since $\gamma_1 = 1$ is in accordance with Theorem 1, we focus on this value in the sequel. Similar calculations can be made for other admissible values of γ_1 . To be specific, we 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). In the following we 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 of (2) except for a factor of 2 in front of λ_T whose only purpose is to make expressions simpler. Let $\hat{\rho}$ denote the minimizer of (4) and $\hat{\rho}_I$ the least squares estimator of ρ^* .

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

Theorem 3. Let $\Delta y_t = \rho^* y_{t-1} + \epsilon_t$ with $\rho^* = 0$ and let $\hat{\rho}$ denote the minimizer of (4).

1. If $\lambda_T \rightarrow 0$ then $P(\hat{\rho} = 0) \rightarrow 0$
2. If $\lambda_T \rightarrow \lambda \in (0, \infty)$ then $P(\hat{\rho} = 0) \rightarrow F_G(\lambda) \in (0, 1)$ where $F_G(\lambda)$ is the distribution function of the random variable G defined in the proof of this result.
3. If $\lambda_T \rightarrow \infty$ then $P(\hat{\rho} = 0) \rightarrow 1$

Theorem 3 reveals that if $\lambda_T \rightarrow 0$, $\hat{\rho}$ is never classified as zero asymptotically. In particular, $\hat{\rho}$ equals the least squares estimator if $\lambda_T = 0$. Part 2 shows that if λ_T tends to a finite constant, $\hat{\rho}$ has mass at 0 in the limit but the mass is not one. Finally, $\hat{\rho}$ will correctly be classified as zero if $\lambda_T \rightarrow \infty$. The next theorem concerns the stationary case.

Theorem 4. 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 $\lambda_T/T \rightarrow \lambda \in [0, \rho^{*2} E(y_{t-1}^2))$ then $P(\hat{\rho} = 0) \rightarrow 0$
2. If $\lambda_T/T \rightarrow \rho^{*2} E(y_{t-1}^2)$ and $\sqrt{T} \left(-\rho^* - \sqrt{\frac{\lambda_T/T}{\frac{1}{T} \sum_{t=1}^T y_{t-1}^2}} \right) \xrightarrow{p} r$, then $P(\hat{\rho} = 0) \rightarrow F \left(\frac{-r \sqrt{E(y_{t-1}^2)}}{\sigma} \right)$ where F is the cdf of the standard normal distribution.
3. If $\lambda_T/T \rightarrow \lambda \in (\rho^{*2} E(y_{t-1}^2), \infty]$, then $P(\hat{\rho} = 0) \rightarrow 1$

Part 1 of Theorem 4 shows that in order for $\hat{\rho}$ not to possess any mass at 0 asymptotically, it certainly suffices that $\lambda_T \in o(T)$. If, on the other hand, $\lambda_T/T \rightarrow \infty$, then $\hat{\rho}$ will be set to zero with probability tending to one even though $\rho^* \neq 0$. Part 2 deals with the behavior of $P(\hat{\rho} = 0)$ when λ_T/T converges to $\rho^{*2} E(y_{t-1}^2)$, i.e. the knife edge just between Part 1 and 3. In this case the probability of setting $\hat{\rho}$ equal to zero can converge to anything between 0 and 1, depending on the size of r .

Taken together, theorems 3 and 4 show that for the adaptive Lasso to act as a *consistent* model selection procedure it is sufficient that $\lambda_T \rightarrow \infty$ (by Theorem 3 $P(\hat{\rho} = 0) \rightarrow 1$ for $\rho^* = 0$ if $\lambda_T \rightarrow \infty$) and $\lambda_T/T \rightarrow \lambda$ for some $\lambda < \rho^{*2} E(y_{t-1}^2)$ (by Theorem 4 $P(\hat{\rho} = 0) \rightarrow 0$ for $\rho^* \neq 0$ if $\lambda_T/T \rightarrow \lambda < \rho^{*2} E(y_{t-1}^2)$). Note that these requirements are less strict than the ones taken directly from Theorem 1 (with $\gamma_1 = \gamma_2 = 1$). To be precise, the requirement $\lambda_T \rightarrow \infty$ is the same while Theorem 4 only requires $\lambda_T/T \rightarrow 0$ as opposed to $\lambda_T/\sqrt{T} \rightarrow 0$ in Theorem 1. The reason for the stricter requirements in Theorem 1 is that this also delivers stronger results as it is not only concerned with model selection. To be precise, only requiring $\lambda_T/T \rightarrow 0$ would allow λ_T to be too big to get the oracle efficient limiting distribution in Theorem 1.

In order to make the adaptive Lasso work as a *conservative* model selection device, Theorem 3 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 $\lambda_T/T \rightarrow \lambda$ for some $\lambda < \rho^{*2} E(y_{t-1}^2)$ (by Theorem 4). The next theorem is concerned with the local to unity situation.

Theorem 5. 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 $\lambda_T \rightarrow 0$ then $P(\hat{\rho} = 0) \rightarrow 0$
2. If $\lambda_T \rightarrow \lambda \in (0, \infty)$ then $P(\hat{\rho} = 0) \rightarrow F_K(\lambda) \in (0, 1)$ where $F_K(\lambda)$ is the distribution function of the random variable K defined in the proof of this result.
3. If $\lambda_T \rightarrow \infty$ then $P(\hat{\rho} = 0) \rightarrow 1$

Consistent model selection requires that $\lambda_T \rightarrow \infty$ (by Theorem 3). By part 3 in Theorem 5 this implies that $P(\hat{\rho} = 0) \rightarrow 1$. Hence, the adaptive Lasso tuned to perform consistent model selection has no power against deviations from 0 of the form c/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 is similar to the one discussed in Pötscher and Schneider (2009).

On the other hand, the adaptive Lasso tuned to perform conservative model selection *does* have power against deviations from 0 of the form c/T . This becomes clear from parts 1 and 2 of Theorem 5 since $\lambda_T \rightarrow 0$ and $\lambda_T \rightarrow \lambda \in (0, \infty)$ are both in line with conservative model selection (see the discussion between theorems 4 and 5).

However, if the goal is not model selection per se, not much harm might be done by classifying ρ^* as zero even though $\rho^* = c/T$. For example, sparse models are known to often forecast better than dense ones and in fact Ploberger and Phillips (2003)⁶ have shown that this is even more so the case in a non-stationary environment. To be precise, the minimal attainable distance of the forecast error of *any* empirical procedure to the one based on knowing the true parameters increases much faster in the number of non-stationary variables than in the number of stationary variables. Thus, sparsity is particularly valuable in a non-stationary environment. These results build on the work of Rissanen (1986, 1987).

5 Conservative model selection

Motivated by the properties of conservative model selection in the local to unity setting we next investigate the properties of the adaptive Lasso in the AR(p) model (1) when tuned to perform conservative model selection.

Theorem 6 (Conservative model selection). *Assume 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 and $\lambda_T \rightarrow \lambda \in [0, \infty)$.*

a) (Nonstationary case): If $\rho^ = 0$*

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

which implies

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

⁶We would like to thank an anonymous referee for pointing us to this reference.

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

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

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

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

b) (Stationary case): Next, if y_t is stationary,

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

which implies

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

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

$$\tilde{B} \sim N_{p+1}(0, \sigma^2 Q), \quad \tilde{C}_{2j} \sim N_{p+1}(0, \sigma^2(Q^{-1})_{(1+j, 1+j)}) \text{ and } Q = E(x_t x_t')$$

Part a) of Theorem 6 reveals that $(\hat{\rho}, \hat{\beta}')'$ converges at the same rate as the least squares estimator. Note that no shrinkage is applied to u_{2j} for $j \in \mathcal{A}$ which is a desirable property. Similar observations apply in the stationary setting in part b) of Theorem 6.

For $\lambda = 0$, Theorem 6 reveals that the asymptotic distribution of the adaptive Lasso estimator is identical to that of the minimizer of $u' Au - 2Bu$. This, in turn, shows 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. If $\lambda \in (0, \infty)$, the penalty terms no longer vanish asymptotically (except for on the nonzero β_j^* s). Hence, with a positive probability⁸ the minimizer of $\Psi(u)$ has entries equal to zero.

Caner and Knight (2013) have shown results similar to the ones in Theorem 6 for the Bridge estimator. In particular, they show that if $\rho^* = 0$ then $\hat{\rho}$ has positive mass at zero in the limit corresponding to the result in Theorem 6a). If, on the other hand, $\rho^* \neq 0$ they show that the Bridge estimator applies no asymptotic shrinkage to $\hat{\rho}$ corresponding to the result in Theorem 6b).

Note that Caner and Knight (2013) focus exclusively on the setting which we call conservative model selection. However, they do not consider consistent model selection. At a more general level the Bridge estimator differs from the adaptive Lasso by being a non-convex minimization problem which makes the estimation more difficult. However, and as opposed to the adaptive Lasso, it does not need an initial estimator⁹. In conclusion, none of the two estimators can be strictly recommended over the other on these grounds, though it does seem to be the case that the adaptive Lasso and other estimators resulting from convex minimization problems have been more popular in recent years.

⁸Actually calculating this probability seems to be non-trivial.

⁹We would like to thank an anonymous referee for stressing this difference.

5.1 Comparison to Lasso

Finally, we compare the above results to the ones obtainable by using the plain Lasso which minimizes

$$\Phi_T(\rho, \beta) = \sum_{t=1}^T \left(\Delta y_t - \rho y_{t-1} - \sum_{j=1}^p \beta_j \Delta y_{t-j} \right)^2 + \lambda_T |\rho| + \mu_T \sum_{j=1}^p |\beta_j|$$

with respect to ρ and β . Denote the minimizer by $(\hat{\rho}_L, \hat{\beta}'_L)'$. Note that the weights w_1 and $w_{2,j}$, $j = 1, \dots, p$ in (2) are no longer present. However, we need to allow for separate tuning parameters λ_T and μ_T for y_{t-1} and the lagged difference in order to get non-trivial limits in Theorem 7 to follow. Let the notation be as in Theorem 6 above.

Theorem 7. *a) (Non-stationary case): If $\rho^* = 0$, $\lambda_T/T \rightarrow \lambda \in [0, \infty)$, and $\mu_T/\sqrt{T} \rightarrow \mu \in [0, \infty)$, then*

$$S_T \left[(\hat{\rho}_L, \hat{\beta}'_L)' - (0, \beta^{*'})' \right] \rightharpoonup \arg \min \Phi(u)$$

where

$$\Phi(u) = u' A u - 2 B u + \lambda |u_1| + \mu \sum_{j=1}^p \left(|u_{2j}| \mathbf{1}_{\{\beta_j^* = 0\}} + u_{2j} \text{sign}(\beta_j^*) \mathbf{1}_{\{\beta_j^* \neq 0\}} \right)$$

b) (Stationary case): If y_t is stationary, $\lambda_T/\sqrt{T} \rightarrow \lambda \in [0, \infty)$, and $\mu_T/\sqrt{T} \rightarrow \mu \in [0, \infty)$, then

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

$$\text{where } \tilde{\Phi}(u) = u' Q u - 2 \tilde{B} u - \lambda u_1 + \mu \sum_{j=1}^p \left(|u_{2j}| \mathbf{1}_{\{\beta_j^* = 0\}} + u_{2j} \text{sign}(\beta_j^*) \mathbf{1}_{\{\beta_j^* \neq 0\}} \right)$$

Assume that $\lambda, \mu > 0$ to rule out an unpenalized limit (otherwise, one might as well apply least squares). Then, Theorem 7 reveals that no matter whether the data is stationary or not, the Lasso applies shrinkage to all parameters in the limit – including the non-zero ones. This is in contrast to the adaptive Lasso tuned to perform conservative model selection which only penalizes the truly zero parameters in the limit, c.f. Theorem 6. As a concrete example consider the case of stationary data and assume for concreteness that $\beta_j^* \neq 0$ for all $j = 1, \dots, p$. Then Theorem 6b) yields

$$\sqrt{T} \left[(\hat{\rho}, \hat{\beta}')' - (\rho^*, \beta^{*'})' \right] \rightharpoonup N(0, \sigma^2 Q^{-1})$$

while Theorem 7b) yields

$$\sqrt{T} \left[(\hat{\rho}_L, \hat{\beta}'_L)' - (\rho^*, \beta^{*'})' \right] \rightharpoonup N(Q^{-1}(\lambda, -\mu \cdot \text{sign}(\beta^*))', \sigma^2 Q^{-1})$$

To conclude, the adaptive Lasso tuned to perform conservative model selection obtains the same limiting distribution as the unpenalized least squares estimator while the plain Lasso is asymptotically biased (its variance does, however, correspond to the one of the efficient least squares estimator). The reason is, as already indicated, that non-zero parameters are not penalized by the adaptive Lasso, while they are penalized by the plain Lasso.

6 Monte Carlo

The supplementary material contains a simulation exercise. The main findings are that the adaptive Lasso and the Bridge estimator perform equally well when it comes to classifying ρ^* correctly but the former is better at finding the correct sparsity pattern in general. The plain Lasso is too conservative to be useful as a unit root test, as it severely overclassifies ρ^* as non-zero in the presence of a unit root.

7 Conclusion

In this paper we showed that the adaptive Lasso can be tuned to perform consistent model selection in stationary as well as non-stationary autoregressions. The results for the non-stationary case are a special case of the ones in Liao and Phillips (2013). Based on the consistent model selection results we explained how the adaptive Lasso could be used to distinguish between stationary and non-stationary autoregressions and how BIC can be used to select a tuning parameter resulting in consistent model selection.

In the case where the adaptive Lasso is tuned to perform conservative model selection we compared it to the plain Lasso. We showed that an asymptotic bias is present in the weak limit of the latter while being absent in the weak limit of former.

We believe that fruitful avenues for future research include developing procedures for inference in high-dimensional time series models (where the number of parameters is more numerous than the number of observations) a la the desparsified Lasso of van de Geer et al. (2014) or the desparsified conservative Lasso of Caner and Kock (2014). Alternatively, one could use post selection procedures like in, e.g., Belloni et al. (2014). We are currently investigating these possibilities.

8 Appendix

The proofs of Theorems 1, 2, 6 and 7 can be found in the supplementary material. Before proving Theorems 3-5 we prove the following lemma. Let $(x)_+ = \max(x, 0)$.

Lemma 1. *Let $g : \mathbb{R} \rightarrow \mathbb{R}$ 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, $\arg \min g = \text{sign}(a) (|a| - \lambda)_+$.*

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 is equivalent to $\arg \min g = 0$ if and only if $\lambda \geq |a|$. In total, the above shows that $\arg \min g = (a - \lambda)_+ = \text{sign}(a) (|a| - \lambda)_+$ for $a > 0$. Similar arguments establish the result for $a < 0$. □

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

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

Proof of Lemma 2. $\hat{\rho}$ minimizes

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

which is equivalent to minimizing

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

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

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

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)

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

□

Proof of Theorem 3. From Phillips (1987a) one has

$$\left(\frac{1}{T} \sum_{t=1}^T y_{t-1}\epsilon_t, \frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2\right) \rightsquigarrow \left(\sigma^2 \int_0^1 W_s dW_s, \sigma^2 \int_0^1 W_s^2 ds\right) \quad (5)$$

Using Lemma 2 with $\rho^* = 0$ yields

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

From (5) and the continuous mapping theorem it follows that

$$G_T := \left[\frac{\frac{1}{T} \sum_{t=1}^T y_{t-1} \epsilon_t}{\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2} \right]^2 \frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2 \rightsquigarrow \left[\frac{\int_0^1 W_s dW_s}{\int_0^1 W_s^2 ds} \right]^2 \sigma^2 \int_0^1 W_s^2 ds =: G \quad (6)$$

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

Case 1: $\lambda_T \rightarrow 0$. Since the right hand side in (6) is absolutely continuous with respect to the Lebesgue measure it has no mass points and so

$$P(\hat{\rho} = 0) = P(G_T \leq \lambda_T) = F_{G_T}(\lambda_T) \rightarrow F_G(0) = 0$$

if $\lambda_T \rightarrow 0$.

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

$$P(\hat{\rho} = 0) = P(G_T \leq \lambda_T) = F_{G_T}(\lambda_T) \rightarrow F_G(\lambda) \in (0, 1)$$

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

Case 3: $\lambda_T \rightarrow \infty$. Since G_T converges weakly it is tight and the result follows. \square

Proof of Theorem 4. By standard results (see e.g. Hall and Heyde (1980) or Phillips and Solo (1992))

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

and so

$$Z_T := \frac{\frac{1}{T^{1/2}} \sum_{t=1}^T y_{t-1} \epsilon_t}{\frac{1}{T} \sum_{t=1}^T y_{t-1}^2} \rightsquigarrow N\left(0, \frac{\sigma^2}{E(y_{t-1}^2)}\right)$$

Now note that by Lemma 2 and the definition of Z_T

$$\begin{aligned} P(\hat{\rho} = 0) &= P\left(\left[T\rho^{*2} + \left[\frac{\frac{1}{T^{1/2}} \sum_{t=1}^T y_{t-1} \epsilon_t}{\frac{1}{T} \sum_{t=1}^T y_{t-1}^2}\right]^2 + 2T^{1/2}\rho^* \frac{\frac{1}{T^{1/2}} \sum_{t=1}^T y_{t-1} \epsilon_t}{\frac{1}{T} \sum_{t=1}^T y_{t-1}^2}\right] \frac{1}{T} \sum_{t=1}^T y_{t-1}^2 \leq \lambda_T\right) \\ &= P\left(\left[\sqrt{T}\rho^* + Z_T\right]^2 \leq \frac{\lambda_T}{\frac{1}{T} \sum_{t=1}^T y_{t-1}^2}\right) \\ &= P\left(\left|-\sqrt{T}\rho^* - Z_T\right| \leq \sqrt{\frac{\lambda_T}{\frac{1}{T} \sum_{t=1}^T y_{t-1}^2}}\right) \\ &= P\left(-\sqrt{\frac{\lambda_T}{\frac{1}{T} \sum_{t=1}^T y_{t-1}^2}} \leq -\sqrt{T}\rho^* - Z_T \leq \sqrt{\frac{\lambda_T}{\frac{1}{T} \sum_{t=1}^T y_{t-1}^2}}\right) \\ &= P\left(-\sqrt{T}\rho^* - Z_T \leq \sqrt{\frac{\lambda_T}{\frac{1}{T} \sum_{t=1}^T y_{t-1}^2}}\right) - P\left(-\sqrt{T}\rho^* - Z_T < -\sqrt{\frac{\lambda_T}{\frac{1}{T} \sum_{t=1}^T y_{t-1}^2}}\right) \end{aligned} \quad (7)$$

Since

$$P\left(-\sqrt{T}\rho^* - Z_T < -\sqrt{\frac{\lambda_T}{\frac{1}{T}\sum_{t=1}^T y_{t-1}^2}}\right) = P\left(-Z_T + \sqrt{T}\left(-\rho^* + \sqrt{\frac{\lambda_T/T}{\frac{1}{T}\sum_{t=1}^T y_{t-1}^2}}\right) < 0\right) \rightarrow 0$$

for all configurations of λ_T/T (because Z_T is tight since it converges in distribution and $\sqrt{T}\left(-\rho^* + \sqrt{\frac{\lambda_T/T}{\frac{1}{T}\sum_{t=1}^T y_{t-1}^2}}\right)$ diverges to infinity (recall $\rho^* \in (-2, 0)$)) it suffices to consider the asymptotic behavior of first term in (7). For this, one has

$$P\left(-\sqrt{T}\rho^* - Z_T \leq \sqrt{\frac{\lambda_T}{\frac{1}{T}\sum_{t=1}^T y_{t-1}^2}}\right) = P\left(-Z_T + \sqrt{T}\left(-\rho^* - \sqrt{\frac{\lambda_T/T}{\frac{1}{T}\sum_{t=1}^T y_{t-1}^2}}\right) \leq 0\right) \quad (8)$$

and the conclusions of part 1) and 3) of the Theorem follow from the tightness of Z_T and the divergence to ∞ or $-\infty$ of $\sqrt{T}\left(-\rho^* - \sqrt{\frac{\lambda_T/T}{\frac{1}{T}\sum_{t=1}^T y_{t-1}^2}}\right)$ when $\lambda_T/T \rightarrow \lambda \in [0, \rho^{*2}E(y_{t-1}^2))$ or $\lambda \in (\rho^{*2}E(y_{t-1}^2), \infty]$, respectively. Part 2) follows from Slutsky's theorem since $Z_T \xrightarrow{d} N\left(0, \frac{\sigma^2}{E(y_{t-1}^2)}\right)$ and $\sqrt{T}\left(-\rho^* - \sqrt{\frac{\lambda_T/T}{\frac{1}{T}\sum_{t=1}^T y_{t-1}^2}}\right) \xrightarrow{p} r$. \square

Proof of Theorem 5. By Phillips (1987b)

$$\left(\frac{1}{T}\sum_{t=1}^T y_{t-1}\epsilon_t, \frac{1}{T^2}\sum_{t=1}^T y_{t-1}^2\right) \xrightarrow{d} \left(\sigma^2 \int_0^1 J_c(r)dW(r), \sigma^2 \int_0^1 J_c^2(r)dr\right) \quad (9)$$

where J_c is the Ornstein-Uhlenbeck process with parameter c . Notice how the only difference to (5) 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 Lemma 2 with $\rho^* = c/T$ it follows

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

By the continuous mapping theorem

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

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

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

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

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

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

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

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

References

- Belloni, A. and V. Chernozhukov (2011). High dimensional sparse econometric models: An introduction. *Inverse Problems and High-Dimensional Estimation*, 121–156.
- Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* 81(2), 608–650.
- 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.
- Caner, M. (2011). Selecting the correct number of factors in approximate factor models: The large panel case with bridge estimators. Technical report, Mimeo. North Carolina State University, Raleigh, NC.
- Caner, M. and K. Knight (2013). An alternative to unit root tests: Bridge estimators differentiate between nonstationary versus stationary models and select optimal lag. *Journal of Statistical Planning and Inference*.
- Caner, M. and A. B. Kock (2014). Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso. *arXiv preprint arXiv:1410.4208*.
- 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.
- Hall, P. and C. Heyde (1980). *Martingale limit theory and its application*. Academic Press.
- 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. and W. Fu (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, 1356–1378.
- Kock, A. B. and L. Callot (2012). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics (forthcoming)* 16.
- Lehmann, E. L. and G. Casella (1998). *Theory of point estimation*, Volume 31. Springer Verlag.
- Liao, Z. and P. Phillips (2013). Automated estimation of vector error correction models. *Econometric Theory (forthcoming)*.
- Medeiros, M. and E. Mendes (2012). Estimating high-dimensional time series models. Technical report.
- 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. and V. Solo (1992). Asymptotics for linear processes. *The Annals of Statistics*, 971–1001.
- 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.
- Ploberger, W. and P. C. Phillips (2003). Empirical limits for time series econometric models. *Econometrica* 71(2), 627–673.
- Pötscher, B. and U. Schneider (2009). On the distribution of the adaptive lasso estimator. *Journal of Statistical Planning and Inference* 139(8), 2775–2790.
- Ren, Y. and X. Zhang (2010). Subset selection for vector autoregressive processes via adaptive lasso. *Statistics & probability letters* 80(23), 1705–1712.
- Rissanen, J. (1986). Stochastic complexity and modeling. *The Annals of Statistics*, 1080–1100.
- Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society. Series B (Methodological)*, 223–239.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.
- van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* 42(3), 1166–1202.

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