AARHUS
UNIVERSITY
BUSINESS AND SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS AND BUSINESS

CREATES
Center for Research in Econometric Analysis of Time Series

Oracle Inequalities for High Dimensional Vector Autoregressions

Anders Bredahl Kock and Laurent A.F. Callot

CREATES Research Paper 2012-16

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

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

ORACLE INEQUALITIES FOR HIGH DIMENSIONAL VECTOR AUTOREGRESSIONS

ANDERS BREDAHL KOCK AND LAURENT A.F. CALLOT

Abstract. This paper establishes non-asymptotic oracle inequalities for the prediction error and estimation accuracy of the LASSO in stationary vector autoregressive models. These inequalities are used to establish consistency of the LASSO even when the number of parameters is of a much larger order of magnitude than the sample size. We show that the number of variables selected is of the right order of magnitude and that no relevant variables are excluded.

Next, non-asymptotic probabilities are given for the Adaptive LASSO to select the correct sparsity pattern. We then give conditions under which the Adaptive LASSO reveals the correct sparsity pattern asymptotically. We establish that the estimates of the non-zero coefficients are asymptotically equivalent to the oracle assisted least squares estimator. This is used to show that the rate of convergence of the estimates of the non-zero coefficients is identical to the one of least squares only including the relevant covariates.

Key words: Vector autoregression, VAR, LASSO, Adaptive LASSO, Oracle inequality, Variable selection, Time series, Model selection.
JEL classifications: C01, C02, C13, C32.

1. Introduction

The last 10-15 years have witnessed a surge of research in high-dimensional statistics and econometrics. This is the study of models where the the number of parameters is of a much larger order of magnitude than the sample size. However, often only a few of the parameters are non-zero, that is the model is sparse, and one wants to be able to separate these from the zero ones. In particular, a lot of attention has been devoted to penalized estimators of which the most famous is probably the LASSO of Tibshirani (1996). Other prominent examples are the SCAD of Fan and Li (2001), the Adaptive LASSO of Zou (2006), the Bridge and Marginal Bridge estimators of Huang et al. (2008), the Dantzig selector of Candes and Tao (2007), and the Sure Independence Screening of Fan and Lv (2008). These procedures have become popular since they are computationally feasible and perform variable selection and parameter estimation at the same time. For a recent review with particular focus on the LASSO see Bühlmann and Van De Geer (2011).

Much effort has been devoted to establishing the conditions under which these procedures possess the oracle property. Here the oracle property is understood as the procedure correctly detecting the sparsity pattern, i.e. setting all zero parameters *exactly* equal to zero while not doing so for any of the non-zero ones. Furthermore, the non-zero parameters are estimated at the same asymptotic efficiency as if only the relevant variables had been included in the model from the outset. In other words the non-zero parameters are estimated as efficiently as if one had been assisted by an oracle that had revealed the true sparsity pattern prior to estimation.

Even though a lot of progress has been made in this direction most focus has been devoted to very simple data types such as the linear regression model with fixed covariates or sometimes (gaussian) independently distributed covariates. Some exceptions are Wang et al. (2007) and Nardi and Rinaldo (2011) who consider the LASSO in a stationary autoregression and Kock

Date: May 29, 2012.
We would like to thank Michael Jansson, Søren Johansen, Jørgen Hoffmann-Jørgensen, Marcelo Medeiros and Timo Teräsvirta for help, comments and discussions. Financial support from the Center for Research in the Econometric Analysis of Time Series (CREATES) is gratefully acknowledged.
Both authors are affiliated with Aarhus University and CREATES, Bartholins Alle 10, 8000 Aarhus C, Denmark. Corresponding author: Anders Bredahl Kock, e-mail: akock@creates.au.dk.

(2012a) who investigates the oracle property of the Adaptive LASSO in stationary and non-stationary autoregressions. However, these papers consider autoregressions of a fixed or slowly increasing length – i.e. a low-dimensional setting.

In this paper we are concerned with the estimation of high-dimensional stationary vector autoregressions (VAR), i.e. models of the form

$$y_t = \sum_{l=1}^{p_T} \Gamma_l y_{t-l} + \epsilon_t, \ t = 1, ..., T \tag{1}$$

where $y_t = (y_{t,1}, y_{t,2}, ..., y_{t,k_T})'$ is the $k_T \times 1$ vector of variables in the model. $\Gamma_1, ..., \Gamma_{p_T}$ are $k_T \times k_T$ parameter matrices. Even though this is suppressed in the notation these may vary with T. So we are analyzing a triangular array of models where the parameters may vary across the rows, T, but remain constant within each row, $t = 1, ..., T$. ϵ_t is assumed to be a sequence of *i.i.d.* error terms with an $N_{k_T}(0, \Sigma)$ distribution. Notice that the number of variables as well as the number of lags is indexed by T indicating that both of these are allowed to increase as the sample size increases – and in particular may be a lot larger than T. Equation (1) could easily be augmented by a vector of constants but here we omit this to keep the notation simple[1].

The VAR is without doubt one of the central pillars in macroeconometrics and is widely used for e.g. forecasting, impulse response and policy analysis. However, it suffers from the fact that many macroeconomic variables are observed at a relatively low frequency such as quarterly or annually leaving few observations for estimation. On the other hand, the number of parameters, $k_T^2 p_T$, may increase very fast if many variables are included in the model which is often the case in order to ensure satisfactory modeling of the dynamics of the variables of interest. Hence, the applied researcher may find himself in a situation where the number of parameters is much larger than the number of observations. If $T < k_T p_T$ equation by equation least squares is not even feasible since the design is singular by construction. Even if the model is possible to estimate the number of regressions which have to be run in order to calculate the information criterion for every subset of variables increases exponentially in the number of parameters and hence becomes computationally infeasible. Furthermore, these subset selection criteria are known to be inherently unstable, see e.g. Breiman (1996).

In a seminal paper Stock and Watson (2002) used factors to reduce dimensionality and obtain more precise forecasts of macro variables while Bernanke et al. (2005) popularized the inclusion of factors in the VAR in order to avoid leaving out relevant information when evaluating monetary policy. For surveys on factor analysis in the context of time series see Stock and Watson (2006), Bai and Ng (2008) and Stock and Watson (2011). Our results show that it is not necessary to augment the VAR by factors in order to handle high-dimensionality. We prove oracle inequalities for the LASSO and the Adaptive LASSO even when the number of parameters is of a much larger order of magnitude than the sample size. In particular,

i) we establish *non-asymptotic* oracle inequalities for the prediction error and estimation accuracy of the LASSO. Specifically, we give lower bounds on the probability with which these quantities are bounded from above by something which is nearly as good as the upper bound on the least squares estimator using only the relevant variables.
ii) we use the finite sample upper bounds to derive asymptotic bounds on the prediction error and estimation accuracy of the LASSO. As a byproduct it is shown that even when k_T and p_T increase at a subexponential rate it is possible to estimate the parameters consistently. The fact that k_T may increase very fast is of particular importance for state of the art macroeconometric modeling of big systems. Conditions for the number of variables selected by the LASSO being of the right order of magnitude and no relevant variables being excluded are also given.
iii) we establish *non-asymptotic* lower bounds on the probability with which the Adaptive LASSO unveils the correct sign pattern and use these bounds to give conditions under which the correct sign pattern is detected with probability tending to one. This result is shown to hold even when k_T and p_T increase at a subexponential rate.

[1]Similarly, we conjecture that a trend could be included by writing the model in deviations from the trend. But to focus on the main idea of the results this has been omitted.

iv) we show that the Adaptive LASSO is asymptotically equivalent to the oracle assisted least squares estimator. This implies that the estimates of the non-zero coefficients converge at the same rate as if least squares had been applied to a model only including the relevant covariates. Furthermore, it shows that the Adaptive LASSO is asymptotically as efficient as the oracle assisted least squares estimator.
v) the appendix contains some maximal inequalities for vector autoregressions, Lemmas 4 and 6, which might be of independent interest.
vi) similar results for autoregressions follow as a special case by simply setting $k_T = 1$ in our theorems.

We believe that these results will be of much use for the applied researcher who often faces the curse of dimensionality when building VAR models since the number of parameters increases quadratically in the number of variables included. The LASSO and the Adaptive LASSO are shown to have attractive finite sample and asymptotic properties even in these situations. Furthermore, the implementation is easy since one simply includes the whole set of potential variables in the model.

Note that since the LASSO and the Adaptive LASSO can be estimated with fewer observations than parameters one may choose to simply include the most recent observations – say 10-20 years – in the model used for forecasting instead of using the whole data set. This could be useful since observations far back in time may be conjectured to be less informative about the near future than the recent past is. Finally, it should be noted that no (significance) testing is involved in the procedures but the underlying assumption is that there exists a sparse representation of the data.

The plan of the paper is as follows. Section 2 lays out the model in more detail and gives necessary background notation. Sections 3 and 4 contain the main results of the paper on the LASSO and the Adaptive LASSO. A Monte Carlo study investigating the validity of our finite sample results can be found in Section 5 while Section 6 concludes. The proofs can be found in the Appendix.

2. Model and notation

We shall suppress the dependence of k_T and p_T on T to simplify notation. As mentioned in the introduction we are concerned with stationary VARs, meaning that the roots of $|I_k - \sum_{j=1}^{p} \Gamma_j z^j|$ lie outside the unit circle. It is convenient to write the model in stacked form. To do so let $Z_t = (y'_{t-1}, ..., y'_{t-p})'$ the $kp \times 1$ vector of explanatory variables at time t in each equation and $Z = (Z_T, ..., Z_1)'$ the $T \times kp$ matrix of covariates for each equation. Let $\tilde{y}_j = (y_{T,j}, ..., y_{1,j})'$ be the $T \times 1$ vector of observations on the jth variable. Hence, $y = (\tilde{y}'_1, ..., \tilde{y}'_k)'$ is the $Tk \times 1$ vector of equationwisely stacked left hand side variables. Similarly let ϵ be the $Tk \times 1$ equationwisely stacked vector of error terms. The fact that y inherits the gaussianity from ϵ shall be particularly useful since this means that y has slim tails. Let $X = (I_k \otimes Z)$ (where I_k denotes the k dimensional identity matrix and $\otimes$ the Kronecker product). Finally, $\beta^* = vec((\Gamma_1, ..., \Gamma_p)')$ is the k^2p dimensional parameter vector of true parameters which also implicitly depends on T. Hence, we may write (1) equivalently as

$$y = X\beta^* + \epsilon \tag{2}$$

Here the parameter vector β^* can potentially be of a much larger order of magnitude than the sample size T. A practical example occurs when building macroeconomic models on relatively infrequent time series (say quarterly or annual data). Then one will often only have 50-200 observations while for $k = 50$ and $p = 5$ the number of parameters is as large as 12,500. Traditional methods such as least squares will be inadequate in such a situation.

Even though there are k^2p parameters in the model only a subset of them might be non-zero. For example, one could imagine that only a few lags – not necessarily consecutive – are necessary to satisfactorily model the dynamics in y_t. This means that β^* is a sparse vector. Also, β^* need not be constant over time which is reasonable since k and p are not constant.

2.1. Further notation.

Let $J = \{j : \beta^*_j \neq 0\} \subseteq \{1, ..., k^2p\}$ denote the set of non-zero parameters and $s = |J|$ its cardinality. Of course J and s are allowed to depend on T but this is

suppressed for ease of notation. $\beta_{\min} = \min\{|\beta_j^*| : j \in J\}$ denotes the minimum non-zero entry of β^* which also implicitly depends on T.

For any $x \in \mathbb{R}^n$, $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$, $\|x\|_{\ell_1} = \sum_{i=1}^n |x_i|$ and $\|x\|_{\ell_\infty} = \max_{1\leq i\leq n} |x_i|$ denote ℓ_2, ℓ_1 and ℓ_∞ norms, respectively (most often $n = k^2p$ or $n = s$ in the sequel). For any symmetric square matrix M, $\phi_{\min}(M)$ and $\phi_{\max}(M)$ denote the minimal and maximal eigenvalues of M.

Let $\Psi_T = \frac{1}{T}X'X$ be the $k^2p \times k^2p$ scaled Gramian of X. For $R, S \subseteq \{1, ..., k^2p\}$, X_R and X_S denote the submatrices of X which consist of the columns of X indexed by R and S, respectively. Furthermore, $\Psi_{R,S} = \frac{1}{T}X_R'X_S$. For any vector δ in $\mathbb{R}^n$ and a subset $J \subseteq \{1, ..., n\}$ we shall let δ_J denote the vector consisting only of those elements of δ indexed by J.

For any two real numbers a and b, $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$ and for any $x \in \mathbb{R}^n$ let $\text{sign}(x)$ denote the sign function applied to each component of x.

Let $\sigma_{j,y}^2$ denote the variance of $y_{t,j}$ and $\sigma_{j,\epsilon}^2$ the variance of $\epsilon_{t,j}$, $1 \leq j \leq k$. Then define $\sigma_T = \max_{1\leq j\leq k}(\sigma_{j,y} \vee \sigma_{j,\epsilon})$.

3. The Lasso

The LASSO was proposed by Tibshirani (1996). Its theoretical properties have been studied intensively since then, see e.g. Zhao and Yu (2006), Meinshausen and Bühlmann (2006), Bickel et al. (2009), and Bühlmann and Van De Geer (2011) to mention just a few. It is known that it only selects the correct model asymptotically under rather restrictive conditions on the dependence structure of the covariates. However, we shall see that it can still serve as an effective screening device in these situations. Put differently, it can remove many irrelevant covariates while still maintaining the relevant ones and estimating the coefficients of these with high precision. The LASSO estimates β^* in (2) by minimizing the following objective function

$$L(\beta) = \frac{1}{T}\|y - X\beta\|^2 + 2\lambda_T\|\beta\|_{\ell_1} \tag{3}$$

where λ_T is a sequence to be defined exactly below. (3) is basically the least squares objective function plus an extra term penalizing parameters that are different from zero. Let $\hat{\beta}$ denote the minimizer of (3) and let $J(\hat{\beta}) = \{j : \hat{\beta}_j \neq 0\}$ be the indices of the parameters that are estimated to be nonzero. $|J(\hat{\beta})|$ denotes the cardinality of $J(\hat{\beta})$.

3.1. Results without conditions on the Gram matrix.

We begin by giving *non-asymptotic* bounds on the performance of the LASSO. Notice that these bounds are valid without any conditions on the design matrix or k, p, s and T.

Theorem 1 (LASSO). *Let* $\lambda_T = \sqrt{8\ln(1+T)^5\ln(1+k)^4\ln(1+p)^2\ln(k^2p)\sigma_T^4/T}$. *Then, with probability at least* $1 - 2(k^2p)^{1-\ln(1+T)} - 2(1+T)^{-1/A}$ *the following inequalities hold for some positive constant* A.[2]

$$\frac{1}{T}\left\|X\hat{\beta} - X\beta^*\right\|^2 + \lambda_T\left\|\hat{\beta} - \beta^*\right\|_{\ell_1} \leq 2\lambda_T\left(\left\|\hat{\beta} - \beta^*\right\|_{\ell_1} + \left\|\beta^*\right\|_{\ell_1} - \left\|\hat{\beta}\right\|_{\ell_1}\right) \tag{4}$$

$$\frac{1}{T}\left\|X\hat{\beta} - X\beta^*\right\|^2 + \lambda_T\left\|\hat{\beta} - \beta^*\right\|_{\ell_1} \leq 4\lambda_T\left[\left\|\hat{\beta}_J - \beta_J^*\right\|_{\ell_1} \wedge \left\|\beta_J^*\right\|_{\ell_1}\right] \tag{5}$$

$$\left\|\hat{\beta}_{J^c} - \beta_{J^c}^*\right\|_{\ell_1} \leq 3\left\|\hat{\beta}_J - \beta_J^*\right\|_{\ell_1} \tag{6}$$

$$|J(\hat{\beta})| \leq \frac{4}{T\lambda_T^2}\phi_{\max}(\Psi_T)\left\|X(\hat{\beta} - \beta^*)\right\|^2 \tag{7}$$

The lower bound on the probability with which inequalities (4)-(7) hold can be increased by choosing a larger value of λ_T. However, we shall see in Theorem 2 below that smaller values of λ_T yield faster rates of convergence[3].

[2] At the cost of a slightly more involved expression on the lower bound on the probability with which the expressions hold λ_T may be reduced to $\sqrt{8\ln(1+T)^{3+\delta}\ln(1+k)^4\ln(1+p)^2\ln(k^2p)\sigma_T^2/T}$ for any $\delta > 0$. This remark is equally valid for all theorems in the sequel.

[3] In general, there is a tradeoff between λ_T being small and the lower bound on the probability with which inequalities (4)-(7) hold being large.

Notice that Theorem 1 holds *without any assumptions* on the Gram matrix. Furthermore, the lower bound on the probability with which inequalities (4)-(7) hold is *nonasymptotic* – it holds for every T – and the above inequalities hold for *any* configuration of k, p, M and T. Note that the lower bound on the probability with which the estimates hold tends to one as $T \to \infty$. In the course of the proof of Theorem 1 we derive a maximal inequality, Lemma 4 in the appendix, for vector autoregressions which might be of independent interest.

Inequalities (4) and (5) give immediate upper bounds on the prediction error, $\frac{1}{T}\left\|X\hat{\beta} - X\beta^*\right\|^2$, as well as the estimation accuracy, $\|\hat{\beta} - \beta^*\|_{\ell_1}$ of the LASSO. In particular, we shall use (5) to derive oracle inequalities for these two quantities in Theorem 2 below. Equation (6) is also of interest in its own right since it shows that an upper bound on the estimation error of the non-zero parameters will result in an upper bound on the estimation error of the zero parameters. This is remarkable since there may be many more zero parameters than non-zero ones in a sparsity scenario and since the bound does not depend on the relative size of the two groups of parameters. The bound (7) will be useful for deriving upper bounds on the number of variables selected by the LASSO.

Theorem 1 will play an important role in establishing the oracle inequalities in Theorem 2.

3.2. **Restricted eigenvalue condition.** Theorem 1 did not pose any conditions on the (scaled) Gram matrix Ψ_T. If $kp > T$ the Gram matrix Ψ_T is singular, or equivalently,

$$\min_{\delta \in \mathbb{R}^{k^2p} \setminus \{0\}} \frac{\delta' \Psi_T \delta}{\|\delta\|^2} = \min_{\delta \in \mathbb{R}^{k^2p} \setminus \{0\}} \frac{\|X\delta\|^2}{T\|\delta\|^2} = 0 \tag{8}$$

In that case ordinary least squares is infeasible. However, for the LASSO Bickel et al. (2009) observed that the minimum in (8) can be replaced by a minimum over a much smaller set. The same is the case for the LASSO in the VAR since we have written the VAR as a regression model. In particular we shall make use of the

Restricted Eigenvalue Condition: RE(r).

$$\kappa(r) = \min \left\{ \frac{\|X\delta\|}{\sqrt{T}\|\delta_R\|} : |R| \leq r,\ \delta \in \mathbb{R}^{k^2p} \setminus \{0\},\ \|\delta_{R^c}\|_{\ell_1} \leq 3\lambda_T \|\delta_R\|_{\ell_1} \right\} > 0 \tag{9}$$

where $R \subseteq \left\{1, ..., k^2p\right\}$ and $|R|$ is its cardinality. Instead of minimizing over all of $\mathbb{R}^{k^2p}$ the minimum is restricted to those vectors which satisfy $\|\delta_{R^c}\|_{\ell_1} \leq 3\lambda_T \|\delta_R\|_{\ell_1}$ and where R has cardinality at most r. This implies that $\kappa(r)$ in (9) can be larger than the Rayleigh-Ritz ratio in (8) even when the latter is zero. Of course $\kappa(r)$ implicitly depends on T but this shall be suppressed in the sequel.

Notice that the restricted eigenvalue condition is trivially satisfied if Ψ_T has full rank since $\delta_R' \delta_R \leq \delta'\delta$ for every $\delta \in \mathbb{R}^{k^2p}$ and so,

$$\frac{\|X\delta\|^2}{T\|\delta_R\|^2} \geq \frac{\|X\delta\|^2}{T\|\delta\|^2} \geq \min_{\delta \in \mathbb{R}^{k^2p}} \frac{\|X\delta\|^2}{T\|\delta\|^2} > 0$$

This means that in the traditional setting of fewer variables per equation than observations the restricted eigenvalue condition is satisfied if $Z'Z$, or equivalently $X'X$, is nonsingular. Hence, the results are applicable in this setting but also in many others. Bickel et al. (2009) give further conditions under which the restricted eigenvalue condition is satisfied. For example, it suffices that the ratio of the smallest and largest eigenvalue of certain submatrices of Ψ_T is not too small. Or that the correlations between the variable are not too large compared to the smallest eigenvalue of submatrices of size $2r$ of Ψ_T. We shall be using the restricted eigenvalue condition with $r = s$.

If the restricted eigenvalue RE(s) condition is satisfied the LASSO satisfies the following oracle inequalities with $\kappa = \kappa(s)$.

Theorem 2. *Let the restricted eigenvalue condition RE(s) be satisfied and let λ_T be as in Theorem 1. Then with probability at least $1 - 2(k^2p)^{1-\ln(1+T)} - 2(1+T)^{-1/A}$ the following*

inequalities hold[4].

$$\frac{1}{T}\left\|X\hat{\beta} - X\beta^*\right\|^2 \leq \frac{16}{\kappa^2} s\lambda_T^2 \tag{10}$$

$$\left\|\hat{\beta} - \beta^*\right\|_{\ell_1} \leq \frac{16}{\kappa^2} s\lambda_T \tag{11}$$

$$|J(\hat{\beta})| \leq \frac{64\phi_{\max}(\Psi_T)}{\kappa^2} s \tag{12}$$

Furthermore, no relevant variables will be excluded from the model if $\beta_{\min} > \|\hat{\beta} - \beta^*\|_{\ell_1}$.

Notice that as in Theorem 1 the bounds are non-asymptotic and hold on a set for which lower bounds on its probability are given. Inequality (10) gives an upper bound on the prediction error compared to the hypothetical situation with knowledge of the true parameter vector. The more the restricted eigenvalue κ is bounded away from zero, the smaller the upper bound on the prediction error. On the other hand, the prediction error is increasing in the number of non-zero parameters s. This is sensible, since it is to be expected that the larger the dimension of the true model the harder it will be to obtain a fit as good as if the true parameter vector had been known. Finally, the prediction error is increasing in λ_T but recall that $\lambda_T = \sqrt{8\ln(1+T)^5\ln(1+k)^4\ln(1+p)^2\ln(k^2p)\sigma_T^4/T}$ which implies λ_T will be small for σ_T, k and p small and T large. A more detailed discussion of the role of σ_T, k and p can be found in the discussion following Corollary 3 below.

Inequality (11) gives an upper bound on the estimation error of the LASSO. To illustrate this result Lemma 1 below gives a corresponding result for the least squares estimator *only including the relevant variables* – i.e. least squares after the true sparsity pattern has been revealed by an oracle. To this end let β_{OLS} denote the least squares estimator only including the relevant variables.

Lemma 1. *Let* $\tilde{\lambda}_T = \sqrt{8\ln(1+T)^5\ln(1+s)^2\ln(s)\sigma_T^4/T}$. *If the true sparsity pattern is known and only the relevant variables are included in the model with their coefficients estimated by least squares,*

$$\left\|\hat{\beta}_{OLS} - \beta_J^*\right\|_{\ell_1} \leq \frac{\tilde{\lambda}_T}{\phi_{\min}(\Psi_{J,J})} s \tag{13}$$

with probability at least $1 - 2s^{1-\ln(1+T)} - 2(1+T)^{-1/A}$.

Comparing (11) to (13) one notices that the upper bounds are very similar. Both expressions consist of s multiplied by some term. Clearly this term is smaller for oracle assisted least squares, $\frac{\tilde{\lambda}_T}{\phi_{\min}(\Psi_{J,J})}$, than for the LASSO, $\frac{16\lambda_T}{\kappa^2}$, since $\tilde{\lambda}_T \leq \lambda_T$ and $\kappa^2 \leq \phi_{\min}(\Psi_{J,J})$ under RE(s). However, λ_T need not be much larger than $\tilde{\lambda}_T$ even if k^2p is a lot larger than s since the logarithmic function increases very slowly. Furthermore, if there exists a $c > 0$ such that $\phi_{\min}(\Psi_{J,J}) \geq \kappa^2 > c$ then the distance due to differences in $\kappa(s)^2$ and $\phi_{\min}(\Psi_{J,J})$ is bounded from above. In conclusion, it is reasonable to call (11) an oracle inequality since it shows, in a non-asymptotic manner, that the LASSO performs almost as well as if one had known the true sparsity pattern and estimated the non-zero parameters by least squares. The price paid for not doing so is the ratio $\frac{16\lambda_T}{\kappa^2} / \frac{\tilde{\lambda}_T}{\phi_{\min}(\Psi_{J,J})}$.

Also notice that the upper bounds on the ℓ_1 estimation error in (11) trivially yield upper bounds on the ℓ_p estimation error for any $p \geq 1$ since $\|\cdot\|_{\ell_p} \leq \|\cdot\|_{\ell_1}$ for any $1 \leq p \leq \infty$. This observation is equally valid for all ℓ_1 bounds in the sequel.

Inequality (12) gives, with high probability, an upper bound on the number of variables selected by the LASSO. In particular, one notices that the number of variables chosen is the true number of variables times $\frac{64\phi_{\max}(\Psi_T)}{\kappa^2}$. We shall see that this inequality can be used to deduce the asymptotic properties of the LASSO in Corollary 3 below since the the bound indicates that the number of variables selected is of the right order if $\frac{64\phi_{\max}(\Psi_T)}{\kappa^2}$ is bounded. It remains to be shown that the variables chosen are the relevant ones.

[4]The constants 16 and 64 can be improved and are used only since the proof of Theorem 1 becomes more transparent.

The last statement of Theorem 2 says that under the "beta-min" condition $\beta_{\min} > \|\hat{\beta} - \beta^*\|_{\ell_1}$ no relevant variables will be left out of the model. It is sensible that the beta-min condition is needed in order to be able to distinguish zero from non-zero parameters since the condition basically requires the two groups to be sufficiently separated – the non-zero coefficients can't be too close to zero. In particular they must be bounded away from zero by a little more than the upper bound on the ℓ_1 estimation error of the LASSO estimator.

Remark 1: Close inspection of the proof of Theorem 2 below reveals that the restricted eigenvalue condition can be replaced by the

Restricted Eigenvalue Condition: RE*(s).

$$\kappa^*(s) = \frac{\left\|X(\hat{\beta} - \beta^*)\right\|}{\sqrt{T}\left\|\hat{\beta}_J - \beta^*_J\right\|} > 0$$

So even the restricted eigenvalue condition can be weakened.

Remark 2: Even though k^2p can be a lot larger than T the parameter β^* in Theorem 2 is still uniquely defined since RE(s) is assumed valid. This follows from an observation similar to observation 2 page 1721 in Bickel et al. (2009).

Remark 3: The above bounds also yield corresponding results for univariate autoregressions, i.e. for $k = 1$. These follow trivially by setting $k = 1$ in all the above bounds.

Remark 4: It is easy to show that the bounds in Theorem 2 also hold equationwisely with unaltered probability and with s replaced by $s_i,\ i = 1, ..., k$ and $s = \sum_{i=1}^{k} s_i$ by using the LASSO on each equation separately but the value of λ_T unaltered. The proof is identical to the present one but applied to each equation separately. Of course, summing up the the individual bounds recovers (10)-(12).

3.3. **Asymptotic properties of the Lasso.** All preceding results have been for finite samples. In this section we shall utilize these results to describe the asymptotic properties of the LASSO as $T \to \infty$.

Theorem 3 (Asymptotic properties of the LASSO)**.** *Assume that there exists a constant $c > 0$ such that $\kappa \geq c$ almost surely and that $s\lambda_T \to 0$. Then, as $T \to \infty$*

i) $\frac{1}{T}\left\|X\hat{\beta} - X\beta^*\right\|^2 \to 0$ *in probability*
ii) $\|\hat{\beta} - \beta^*\|_{\ell_1} \to 0$ *in probability*
iii) $|J(\hat{\beta})| \in O_p(s)$ *if* $\phi_{\max}(\Psi_T) \in O_p(1)$.
iv) *With probability tending to one no relevant variables will be excluded from the model if* $\beta_{\min} > \frac{16}{c^2} s\lambda_T$.

To get a feeling for the size of the models that are compatible with $s\lambda_T \to 0$ in Theorem 3 assume that p and k are both of order $O(e^{T^a})$ and $s \in O(T^b)$ for some $a, b \geq 0$ and that $\sup_T \sigma_T < \infty$. Then Lemma 5 in the Appendix shows that $s\lambda_T \to 0$ if $7a + 2b < 1$. If one is only interested in the average prediction error tending to zero in probability, it suffices that $7a + b < 1$. This example shows that one can have both p and k increasing very fast – in fact sub-exponentially – and at the same time letting the number of relevant variables arrive at a polynomial rate. The setting where the total number of parameters increases sub-exponentially in the sample size is sometimes referred to as ultra-high or non-polynomial dimensionality. By choosing a sufficiently close to 0, it is clear that any $b < 1/2$ can be accommodated while still having the estimation error tending to zero in probability. A similar remark was made for the Marginal Bridge estimator in a very different context by Huang et al. (2008) and Kock (2012b). In order for the prediction error to tend to zero it suffices that $b < 1$ for a sufficiently small. Hence, the number of relevant variables can increase almost as fast as the sample size. In the perhaps more realistic setting where only k increases at a sub-exponential rate while p stays fixed arguments along the same lines as in Lemma 5 show that it suffices that $5a + 2b < 1$ in order for $s\lambda_T \to 0$.

iii) of Theorem 3 says that under the stated conditions the number of variables selected by the LASSO is of the right order. This shows that the LASSO can be used as a strong tool for an initial screening and dimension reduction to yield a model of a more manageable size.

Continuing the above example where $k, p \in O(e^{T^a})$, $s \in O(T^b)$ and $\sup_T \sigma_T < \infty$ iii) yields that even when the total number of parameters is subexponential the size of the problem can be reduced to a polynomial one.

iv) gives a sufficient condition for avoiding excluding relevant variables asymptotically. Combining this with iii) shows that the LASSO can be used as a strong screening device in vector autoregressions since iii) and iv) together say that all relevant and not too many irrelevant variables will be selected asymptotically. As argued in the discussion after Theorem 2 such a "beta-min" condition is natural since one can't expect to be able to distinguish between zero and non-zero parameters if the distance between these is less than the precision of the estimator.

At this stage it is worth mentioning that the conditions in Theorem 3 are merely sufficient. For example, it would also suffice for i) and ii) that $\frac{s\lambda_T}{\kappa^2} \to 0$ in probability which is of course implied by $\kappa \geq c > 0$ and $s\lambda_T \to 0$. In other words, κ bounded away from zero is not necessary to establish asymptotic prediction error optimality and consistency of $\hat{\beta}$. Also notice that since $s\lambda_T \to 0$ the beta-min condition in iv) is satisfied in particular if there exists a constant $\hat{c} > 0$ such that $\beta_{\min} \geq \hat{c}$. Hence, the beta-min condition is satisfied if there exists a constant $\hat{c}$ that separates the zero and non-zero coefficients. Of course this constant can be arbitrarily small.

Again the case $k = 1$ gives results corresponding to univariate autoregressions.

4. The Adaptive Lasso

The LASSO penalizes all parameters equally much. If it were possible to penalize the truly zero parameters more than the non-zero ones one would expect a better performance. Zou (2006) used this idea to propose the Adaptive LASSO in the standard linear regression model with a fixed number of non-random regressors. He established that the Adaptive LASSO is asymptotically oracle efficient in this setting – with probability tending to one it selects the correct sparsity pattern. We now apply the Adaptive LASSO to our vector autoregressive model. However, we shall give lower bounds on the finite sample probabilities of selecting the correct model. Then these bounds are used to establish that with probability tending to one the correct sparsity pattern (and a little bit more) is unveiled.

The Adaptive LASSO estimates β^* by minimizing the following objective function

$$\tilde{L}(\beta) = \frac{1}{T}\left\|y - X_{J(\hat{\beta})}\beta_{J(\hat{\beta})}\right\|^2 + 2\lambda_T \sum_{j \in J(\hat{\beta})} \frac{|\beta_j|}{|\hat{\beta}_j|} \tag{14}$$

where $\hat{\beta}_j$ denotes the LASSO estimator of β_j^* from the previous section. Let $\tilde{\beta}$ denote the minimizer of (14). Note that if $\hat{\beta}_j = 0$ the j'th variable is excluded from the model. So if the first stage LASSO estimator classifies a parameter as zero it is not even included in the second step resulting in a problem of a much smaller size. If $\beta_j^* = 0$ then $\hat{\beta}_j$ is likely to be small by (11) and consistency of the LASSO. Hence, $1/\hat{\beta}_j$ is large and the penalty on β_j is large. If $\beta_j^* \neq 0$, $\hat{\beta}_j$ is not too close to zero and the penalty is small. In short, the Adaptive LASSO is a two step estimator with which greater penalties are applied to the truly zero parameters. These more intelligent weights allow us to show that it is sign consistent with high probability, i.e. $P\left(\text{sign}(\tilde{\beta}) = \text{sign}(\beta^*)\right)$ is close to one. This in particular implies that the correct sparsity pattern is chosen.

Even though we use the LASSO as our initial estimator, this is not necessary. All we shall make use of is the upper bound on its ℓ_1 estimation error. Hence, the results in Theorem 4 below can be improved if an estimator with tighter bounds is used.

The first Theorem gives lower bounds on the *finite sample probability* of the Adaptive LASSO being sign-consistent.

Theorem 4. *Let λ_T be as above and assume that*[5] $\beta_{\min} \geq 2\left\|\hat{\beta} - \beta^*\right\|_{\ell_1}$ *and*

$$\frac{sK_T}{\phi_{\min}(\Psi_{J,J})}\left(\frac{1}{2} + \frac{2}{\beta_{\min}}\right)\left\|\hat{\beta} - \beta^*\right\|_{\ell_1} + \frac{\left\|\hat{\beta} - \beta^*\right\|_{\ell_1}}{2} \leq 1 \tag{15}$$

$$\frac{\sqrt{s}}{\phi_{\min}(\Psi_{J,J})}\left(\frac{\lambda_T}{2} + \frac{2\lambda_T}{\beta_{\min}}\right) \leq \beta_{\min} \tag{16}$$

where $K_T = \ln(1+k)^2\ln(1+p)^2\ln(T)$. Then, with probability at least $1 - 2(k^2p)^{1-\ln(1+T)} - 2(1+T)^{-1/A} - 2T^{-1/A}$ it holds that $\operatorname{sign}(\tilde{\beta}) = \operatorname{sign}(\beta^*)$.

Here we have chosen to keep the expressions at a high level instead of inserting the upper bound on $\left\|\hat{\beta} - \beta^*\right\|_{\ell_1}$ from Theorem 2 since this facilitates the interpretation. Clearly, the more precise the initial estimator, the smaller the left hand side in (15). On the other hand a small $\beta_{\min}$ makes the inequality harder to satisfy. This is sensible since the correct sign pattern is harder to detect if the non-zero parameters are close to zero. K_T is increasing in the dimension of the model and so large k and p make it harder to detect the correct sign pattern. Note (Bickel et al. (2009), page 1710) that $\phi_{\min}(\Psi_{J,J}) > 0$ when the RE(s) is satisfied since all submatrices of size $2s$ are nonsingular. It is reasonable that the closer $\phi_{\min}(\Psi_{J,J})$ is to being singular the harder it is to unveil the correct sign pattern. The interpretation of (16) is similar since λ_T is increasing in the dimension of the model. Notice again that the assumption $\beta_{\min} \geq 2\left\|\hat{\beta} - \beta^*\right\|_{\ell_1}$ is a reasonable one: one can't expect to detect the correct sign pattern if the precision of the initial estimator is smaller than the distance the smallest non-zero coefficient is bounded away from zero since otherwise the initial LASSO estimator may falsely classify non-zero parameters as zero.

Also notice that by the last assertion of Theorem 2, $\beta_{\min} \geq 2\left\|\hat{\beta} - \beta^*\right\|_{\ell_1}$ ensures that the initial LASSO estimator will not exclude any relevant variables. This is of course a necessary condition for the second stage Adaptive LASSO to select the correct sign pattern.

4.1. **Asymptotic properties of the Adaptive Lasso.** The results in Theorem 4 are non-asymptotic but can be used to obtain the following sufficient conditions for asymptotic sign consistency of the Adaptive LASSO.

Theorem 5 (Asymptotic sign consistency of the Adaptive LASSO)**.** $P(\operatorname{sign}(\tilde{\beta}) = \operatorname{sign}(\beta^*)) \to 1$ *if either of the following conditions is satisfied.*

i) *There exists a sequence $a_T \to \infty$ such that $\beta_{\min} \geq a_T \frac{s^2 K_T \lambda_T}{\phi_{\min}(\Psi_{J,J})\kappa^2} \to 0$ and $\beta_{\min} \geq a_T \frac{s^{1/4}\lambda_T^{1/2}}{\sqrt{\phi_{\min}(\Psi_{J,J})}} \to 0$ where both estimates and convergence assertions hold with probability tending to one.*

ii) *There exists a $\tilde{c} > 0$ such that $\phi_{\min}(\Psi_{J,J})$, $\kappa \geq \tilde{c}$ almost surely and a $\tilde{d} < \infty$ such that $\sup_T \sigma_T < \tilde{d}$. Furthermore, $k, p \in O(e^{T^a})$ as well as $s \in O(T^b)$ for some $a, b \geq 0$ satisfying $15a + 4b < 1$ and $\beta_{\min} \in \Omega(\ln(T)[b_T \vee c_T])$ for $b_T = T^{2b}T^{4a+(7/2)a-1/2}\ln(T)^{1+5/2}$ and $c_T = T^{b/4}T^{(7/4)a-1/4}\ln(T)^{5/4}$.*[6]

In i) it is not essential that the right hand sides of the inequalities tend to zero. However, the assumption simplifies the proof and is also in line with thinking of $\beta_{\min}$ approaching zero as the sample size increases. That $\beta_{\min}$ has to be bounded from below is in line with previous explanations. The sequence a_T can tend to infinity as slowly as desired which is utilized in the proof of part ii) of this theorem.

ii) quantifies how fast k, p and s can increase if one wishes to detect the correct sign pattern asymptotically by the Adaptive LASSO. The number of relevant variables must be $o(T^{1/4})$ while k and p can still increase at a subexponential rate. Hence, models with many more predictors than observations can be handled by the Adaptive LASSO.

How small can $\beta_{\min}$ be? Consider a model with fixed k and p corresponding to $a = b = 0$. In this case $\beta_{\min} \geq c\ln(T)[b_T \vee c_T]$ for some constant $c > 0$. This means that $\beta_{\min} \in$

[5]It suffices that $\beta_{\min} > \left\|\hat{\beta} - \beta^*\right\|_{\ell_1}$ such that $\beta_{\min} \geq q\left\|\hat{\beta} - \beta^*\right\|_{\ell_1}$ for some $q > 1$.

[6]Here $f(T) \in \Omega(g(T))$ means that there exists a constant c such that $f(T) \geq cg(T)$ from a certain T_0 and onwards. So there exists a constant c such that $\beta_{\min} \geq c\ln(T)[b_T \vee c_T]$ from a T_0 and onwards.

$\Omega(\ln(T)^{9/4}T^{-1/4})$. Of course this is the case in particular if there exists a $d > 0$ such that $\beta_{\min} \geq d$.

The above conditions are merely sufficient. For example it is possible to relax $\sup_T \sigma_T < \tilde{d}$ in ii) at the price of slower growth rates for s, k and p.

Finally, we show that the estimates of the non-zero parameters of the Adaptive LASSO are asymptotically equivalent to the least squares ones only including the relevant variables. Hence, the limiting distribution of the non-zero coefficients is identical to the oracle assisted least squares estimator.

Theorem 6. *Let the assumptions of part ii) of Theorem 5 be satisfied. Then, for any $s \times 1$ vector α with unit norm*

$$\left|\sqrt{T}\alpha'(\tilde{\beta}_J - \beta_J^*) - \sqrt{T}\alpha'(\hat{\beta}_{OLS} - \beta_J^*)\right| \in o_p(1)$$

where $o_p(1)$ is a term that converges to zero in probability uniformly in α.

Theorem 6 reveals that $\sqrt{T}\alpha'(\tilde{\beta}_J - \beta_J^*)$ is asymptotically equivalent to $\sqrt{T}\alpha'(\hat{\beta}_{OLS} - \beta_J^*)$. So inference is asymptotically as efficient as oracle assisted least squares. As seen from the discussion following Theorem 5 this is the case in even very high-dimensional models.

By combining Theorem 6 and Lemma 1 one obtains the following upper bound on the rate of convergence of $\tilde{\beta}$ to β^*.

Corollary 1. *Let the assumptions of part ii) of Theorem 5 be satisfied. Then,*

$$\|\tilde{\beta}_J - \beta_J^*\|_{\ell_1} \in O_p\left(\tilde{\lambda}_T s\right)$$

where as in Lemma 1 $\tilde{\lambda}_T = \sqrt{8\ln(1+T)^5 \ln(1+s)^2 \ln(s)/T}$.

Notice that the rate of convergence is as fast as the one for the oracle assisted least squares estimator obtained by Lemma 1. Hence, the Adaptive LASSO improves further on the LASSO by selecting the correct sparsity pattern and estimating the non-zero coefficients at same rate as the least squares oracle. It is not difficult to show that in the case of fixed covariates the least squares estimator satisfies $\|\tilde{\beta}_{OLS} - \beta_J^*\| \in O_p(s/\sqrt{T})$. Hence, we conjecture that it may be possible to decrease $\tilde{\lambda}_T$ in Theorem 1 to $1/\sqrt{T}$ but in any case the current additional factors are merely logarithmic.

5. Monte Carlo

This section explores the finite sample properties of the LASSO and the Adaptive LASSO. We compare the performance of these procedures to oracle assisted least squares which is least squares including only the relevant variables. This estimator is of course infeasible in practice but is nevertheless a useful benchmark. Whenever the sample size permits it we also implement least squares including all variables, i.e. without any variable selection whatsoever. This is at the other extreme of Oracle OLS. The LASSO and the Adaptive LASSO are implemented using the publicly available R package `glmnet`. λ_T is chosen by BIC. We also experimented with cross validation but this did not improve the results while being considerably slower. All procedures are implemented equation by equation and their performance is measured along the following dimensions which are reported for the whole system.

(1) Correct sparsity pattern: How often does a procedure select the correct sparsity pattern, i.e. how often does it include all the relevant variables while discarding all irrelevant variables.
(2) True model included: How often does a procedure retain all the relevant variables. This is a relevant measure in practice since even if a procedure does not detect the correct sparsity pattern it may still be able to retain all relevant variables while hopefully leaving many irrelevant variables out and hence reducing the dimension of the model.
(3) Fraction of relevant variables included. If a procedure wrongly discards a relevant variable, how big is the fraction of relevant variables retained?
(4) Number of variables included: How many variables does each procedure include on average. This measures how well a procedure reduces the dimension of the problem.

(5) RMSE: The root mean square error of the parameter estimates calculated as $\sqrt{\frac{1}{MC}\sum_{i=1}^{MC}\left\|\hat{\beta}^{(i)}-\beta^*\right\|^2}$ where MC denotes the number of Monte Carlo replication and $\hat{\beta}^{(i)}$ is the estimated parameter vector in the ith Monte Carlo replication by any of the above mentioned procedures.

(6) 1-step ahead RMSFE: For every Monte Carlo replication the estimated parameters are used to make a one step ahead forecast of the whole vector $y_{T+1}^{(i)}$ denoted $\hat{y}_{T+1,T}^{(i)}$. The root mean square forecast error (RMSFE) is calculated as $\sqrt{\frac{1}{k}\frac{1}{MC}\sum_{i=1}^{MC}\left\|\hat{y}_{T+1,T}^{(i)}-y_{T+1}^{(i)}\right\|^2}$.

The following three Experiments are considered where the covariance matrix of the error terms is diagonal with .01 on the diagonal in all settings. The sample sizes are $T = 50, 100$ and 500.

- Experiment A: The data is generated from a VAR(1) model with $\Gamma_1 = diag(0.5, ..., 0.5)$ and with $k = 10, 20, 50$ and 100. This is a truly sparse model where the behavior of each variable only depends on its own past. The case $k = 100$ illustrates a high dimensional setting where each equation has 99 redundant variables.
- Experiment B: The data is generated from a VAR(4) model where Γ_1 and Γ_4 have a block diagonal structure. In particular, the blocks are 5×5 matrices with all entries of the blocks of Γ_1 equal to .15 and all elements of the blocks of Γ_4 equal to $-.1$. $\Gamma_2 = \Gamma_3 = 0$. The largest root of the companion matrix of the system is .98 indicating a very persistent behavior of the system. This structure could be motivated by a model build on quarterly data as is often the case in macroeconometrics. $k = 10, 20$ and 50.
- Experiment C: The data is generated from a VAR(5) model where $\Gamma_1 = diag(.95, ..., .95)$ and $\Gamma_j = (-.95)^{(j-1)}\Gamma_1,\ j = 2, .., 5$. This results in a system with a companion matrix that has a maximal eigenvalue of .92. There are no gaps but the absolute value of the coefficients on each lag get smaller with the lag length reflecting the conventional wisdom that more recent lags are more important than more distant ones. $k = 10, 20$ and 50.

Table 1 contains the results for Experiment A. Blank entries indicate settings where least squares including all variables was not feasible.

Neither the LASSO nor the Adaptive LASSO unveil the correct sparsity pattern very often. However, in accordance with Theorem 5, the Adaptive LASSO shows a clear improvement along this dimension as the sample size increases when $k = 10$. As mentioned previously detecting exactly the correct model might be asking for too much. This is illustrated by the fact that the LASSO as well as the Adaptive LASSO very often include all relevant variables. Table 1 also shows that even in the cases where the true model is not included in the set chosen by the LASSO or the Adaptive LASSO, the share of relevant variables included is still relatively high. The worst performance may be found for $k = 100$ and $T = 50$ where the share of relevant variables included by the Adaptive LASSO is 38 percent. Also notice that since the LASSO is used as the initial estimator for the Adaptive LASSO the latter can perform no better along this dimension than the former (variables excluded in the first step are also excluded in the second step). In this light it is encouraging that the Adaptive LASSO actually performs almost as good as the LASSO – it rarely discards any relevant variables in the second step. But how many variables are included in total, or put differently, how well do the procedures reduce the dimension of the model? For this measure the results are quite encouraging. Even when $k = 100$ only 134 variables out $10,000$ possible are included by the Adaptive LASSO when $T = 500$. Since the relevant variables are always included this means that only 34 redundant variables, an average of .34 per equation, are included.

This dimension reduction can result in a large reduction in RMSE compared to the least squares estimator including all variables. The LASSO and the Adaptive LASSO are always more precise than this alternative. The Adaptive LASSO tends to be more precise than the LASSO due to its more intelligent weights in the second step. However, it is still a little less precise than the oracle estimator – a result which stems from the occasional inclusion of irrelevant variables[7]. The two shrinkage procedures forecast as precisely as the oracle estimator except for

[7]We also experimented with using least squares including all variables as initial estimator. However, it did not uniformly dominate the LASSO while being infeasible in settings with few observations relative to the number of variables.

	LASSO			Adaptive LASSO			Oracle OLS			Full OLS		
T	50	100	500	50	100	500	50	100	500	50	100	500
k	Correct sparsity pattern											
10	0.00	0.00	0.09	0.00	0.03	0.34	1	1	1	0	0	0
20	0.00	0.00	0.00	0.00	0.00	0.03	1	1	1	0	0	0
50	0.00	0.00	0.00	0.00	0.00	0.00	1	1	1		0	0
100	0.00	0.00	0.00	0.00	0.00	0.00	1	1	1			0
	True model included											
10	0.06	0.78	1.00	0.05	0.78	1.00	1	1	1	1	1	1
20	0.00	0.44	1.00	0.00	0.44	1.00	1	1	1	1	1	1
50	0.00	0.03	1.00	0.00	0.03	1.00	1	1	1		1	1
100	0.00	0.00	1.00	0.00	0.00	1.00	1	1	1			1
	Fraction of relevant variables included											
10	0.75	0.97	1.00	0.74	0.97	1.00	1	1	1	1	1	1
20	0.67	0.96	1.00	0.65	0.96	1.00	1	1	1	1	1	1
50	0.69	0.93	1.00	0.64	0.93	1.00	1	1	1		1	1
100	0.49	0.90	1.00	0.38	0.90	1.00	1	1	1			1
	Number of variables included											
10	16	16	13	13	13	11	10	10	10	100	100	100
20	37	34	26	32	31	24	20	20	20	400	400	400
50	923	93	67	832	89	65	50	50	50		2500	2500
100	4769	215	135	3825	208	134	100	100	100			10000
	RMSE											
10	1.13	0.72	0.28	1.12	0.56	0.17	0.40	0.28	0.12	1.61	1.00	0.40
20	1.78	1.15	0.45	1.91	0.98	0.28	0.56	0.40	0.17	3.96	2.24	0.82
50	8.36	2.08	0.81	10.09	1.90	0.52	0.89	0.62	0.27		7.68	2.21
100	12.07	3.27	1.26	12.91	3.16	0.81	1.26	0.88	0.39			4.95
	1-step ahead RMSFE											
10	0.106	0.102	0.101	0.106	0.101	0.100	0.100	0.100	0.100	0.113	0.105	0.101
20	0.109	0.104	0.100	0.110	0.103	0.100	0.101	0.101	0.100	0.135	0.113	0.102
50	0.154	0.105	0.100	0.174	0.104	0.100	0.101	0.100	0.100		0.147	0.105
100	0.151	0.106	0.101	0.158	0.105	0.100	0.101	0.100	0.100			0.113

TABLE 1. The results for Experiment A measured along the dimensions discussed in the main text.

the most difficult settings. As a consequence, they are more precise than least squares including all variables.

Figure 1 contains the densities of the estimates over the 1000 Monte Carlo replications of the first parameter in the first equation. The true value of this parameter is .5. The upper two plots are for $k = 10$ and reveal that all procedures except for the LASSO are centered at the right place. The LASSO is centered too far to the left due to its shrinkage. The Adaptive LASSO does not suffer from this since its weights are chosen more intelligently.

The bottom two plots are concerned with a high dimensional setting where $k = 50$. Results for $k = 100$ are not reported since least squares including all variables is only applicable for $T = 500$ here. Two things are observed for $T = 100$ when $k = 50$. First, the least squares estimator including all variables has a very big variance and is not even centered the correct place. The Adaptive LASSO does not suffer from this problem and is only slightly downwards biased compared to the least squares oracle. However, (and secondly) the LASSO and the Adaptive LASSO have bimodal densities due to the occasional wrong exclusion of the non-zero parameter. Increasing the sample size to 500 eliminates this problem and now the density of the Adaptive LASSO sits almost on top of the one of the least squares oracle while the LASSO and full least squares procedures are still biased to the left.

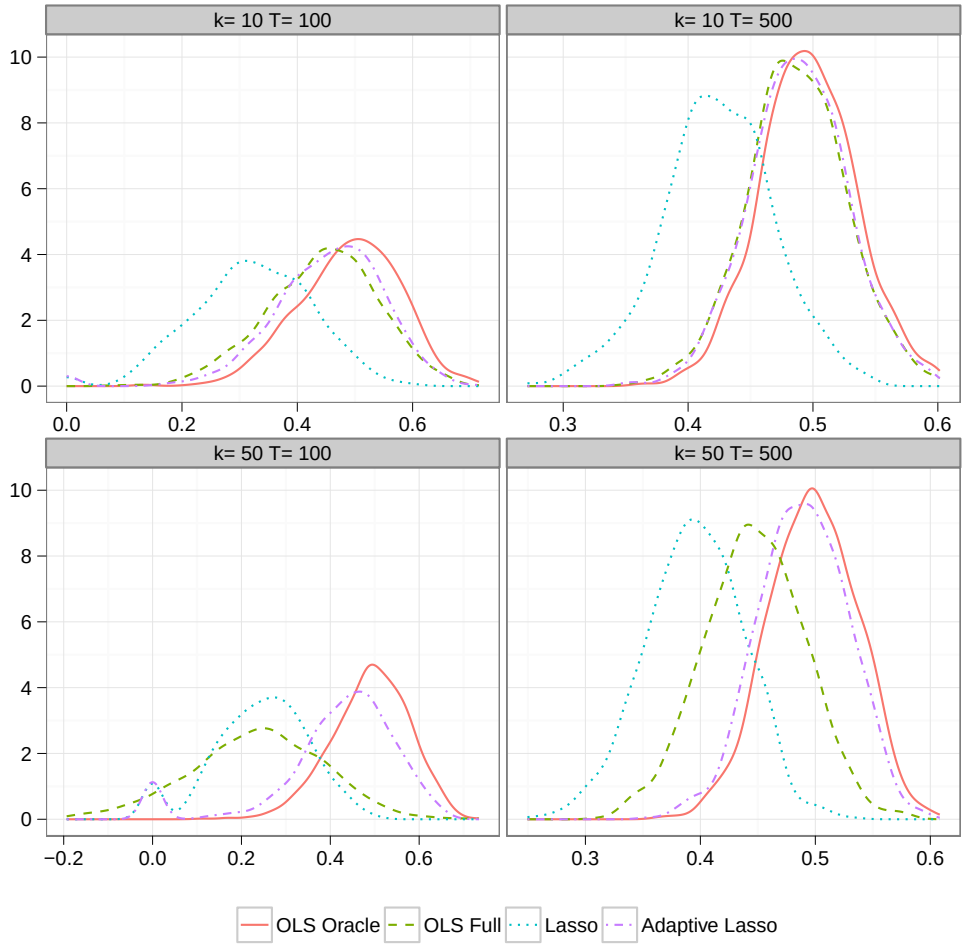

FIGURE 1. Density of the estimates of the first parameter in the first equation. The true value of the parameter is .5.

Table 2 contains the results for Experiment B. This setting is more difficult than the one in in Experiment A since the model is less sparse and the system possesses a root close to the unit circle.

Notice that neither the LASSO nor the Adaptive LASSO ever find exactly the true model. Both procedures leave out relevant variables even for $T = 500$. However, the fraction of relevant variables included tends to be increasing in the sample size. In opposition to Experiment A the Adaptive LASSO does discard relevant variables in the second step that were included by the LASSO in the first step. This can be deduced from the fact that the share of relevant variables included is lower than the one for the LASSO. This results in an interesting situation ($T = 500$) where the number of variables included by the LASSO tends to be slightly larger than the ideal one while the opposite is the case for the Adaptive LASSO.

As in Experiment A the LASSO as well as the Adaptive LASSO have much lower RMSE than OLS including all covariates. However, the LASSO is now slightly more precise than the Adaptive LASSO. This finding is due to the fact that the LASSO tends to discard slightly fewer relevant variables than the Adaptive LASSO. The LASSO is actually almost as precise as Oracle OLS for $T = 500$. This results in forecasts that are as precise as the one produced by the least squares oracle.

Table 3 contains the results for Experiment C. As was the case in Experiment B, neither the LASSO nor the Adaptive LASSO unveil the true model. However, they tend to at least retain the relevant variables as the sample size increases and the share of relevant variables is also

	LASSO			Adaptive LASSO			Oracle OLS			Full OLS		
T	50	100	500	50	100	500	50	100	500	50	100	500
k	Correct sparsity pattern											
10	0.00	0.00	0.00	0.00	0.00	0.00	1	1	1	0	0	0
20	0.00	0.00	0.00	0.00	0.00	0.00	1	1	1		0	0
50	0.00	0.00	0.00	0.00	0.00	0.00	1	1	1			0
	True model included											
10	0.00	0.00	0.22	0.00	0.00	0.00	1	1	1	1	1	1
20	0.00	0.00	0.03	0.00	0.00	0.00	1	1	1		1	1
50	0.00	0.00	0.00	0.00	0.00	0.00	1	1	1			1
	Fraction of relevant variables included											
10	0.45	0.61	0.98	0.36	0.43	0.88	1	1	1	1	1	1
20	0.59	0.52	0.98	0.52	0.37	0.87	1	1	1		1	1
50	0.33	0.56	0.98	0.24	0.46	0.87	1	1	1			1
	Number of variables included											
10	128	80	110	103	52	90	100	100	100	400	400	400
20	841	175	232	728	110	184	200	200	200		1600	1600
50	2309	3007	622	1542	2525	477	500	500	500			10000
	RMSE											
10	4.37	1.02	0.46	4.31	1.34	0.60	1.55	0.99	0.41	8.89	2.71	0.94
20	5.23	1.56	0.67	5.94	2.02	0.86	2.19	1.40	0.58		10.37	1.97
50	6.21	6.02	1.12	6.68	6.97	1.39	3.47	2.21	0.92			5.86
	1-step ahead RMSFE											
10	0.177	0.111	0.103	0.174	0.112	0.103	0.116	0.105	0.102	0.311	0.132	0.105
20	0.162	0.118	0.103	0.175	0.117	0.103	0.113	0.107	0.101		0.253	0.110
50	0.140	0.140	0.104	0.145	0.148	0.103	0.114	0.106	0.101			0.131

TABLE 2. The results for Experiment B measured along the dimensions discussed in the main text.

always above 90 percent when $T = 500$. As in Experiment A, the Adaptive LASSO does not discard many relevant variables in the second estimation step. In fact, turning to the number of variables selected, this second step is very useful since it often greatly reduces the number of irrelevant variables included by the LASSO in the first step. Put differently, the LASSO carries out the rough initial screening in the first step while the Adaptive LASSO fine tunes this in the second step.

The Adaptive LASSO always estimates the parameters more precisely than full OLS (and is also more precise than the LASSO for $T = 500$). As in the previous experiments this results in forecasts that are as precise as the OLS oracle for $T = 500$.

6. Conclusions

This paper is concerned with estimation of high-dimensional stationary vector autoregressions. In particular, the focus is on the LASSO and the Adaptive LASSO. We establish upper bounds for the prediction and estimation error of the LASSO. The novelty in these upper bounds is that they are non-asymptotic. Under further conditions it is shown that the LASSO will not select too many redundant variables and that all relevant variables are retained with high probability. A comparison to oracle assisted least squares is made and it is seen that the LASSO does not perform much worse than this infeasible procedure. The finite sample results are then used to establish equivalent asymptotic results. It is seen that the LASSO is consistent even when the number of parameters grows sub-exponentially with the sample size.

Next, lower bounds on the probability with which the Adaptive LASSO unveils the correct sign pattern are given. Again these results are non-asymptotic but they can be used to establish asymptotic sign consistency of the Adaptive LASSO. As for the LASSO the number of parameters is allowed to grow sub-exponentially fast with the sample size. Finally, we show that the

	LASSO			Adaptive LASSO			Oracle OLS			Full OLS		
T	50	100	500	50	100	500	50	100	500	50	100	500
k	Correct sparsity pattern											
10	0.00	0.00	0.00	0.00	0.00	0.01	1	1	1		0	0
20	0.00	0.00	0.00	0.00	0.00	0.00	1	1	1			0
50	0.00	0.00	0.00	0.00	0.00	0.00	1	1	1			0
	True model included											
10	0.00	0.00	1.00	0.00	0.00	1.00	1	1	1		1	1
20	0.00	0.00	0.94	0.00	0.00	0.94	1	1	1			1
50	0.00	0.00	0.00	0.00	0.00	0.00	1	1	1			1
	Fraction of relevant variables included											
10	0.69	0.62	1.00	0.61	0.59	1.00	1	1	1		1	1
20	0.43	0.57	1.00	0.36	0.52	1.00	1	1	1			1
50	0.19	0.46	0.93	0.15	0.38	0.93	1	1	1			1
	Number of variables included											
10	307	144	200	263	94	56	50	50	50	500	500	500
20	742	664	560	624	527	122	100	100	100	2000	2000	2000
50	1958	3824	1766	1532	3065	406	250	250	250	12500	12500	12500
	RMSE											
10	6.26	4.81	1.66	6.85	4.46	0.48	1.12	0.73	0.30		4.59	1.15
20	8.85	7.80	3.26	9.18	7.99	0.87	1.58	1.01	0.41			2.76
50	13.75	13.59	7.83	13.98	14.10	4.44	2.50	1.61	0.66			9.99
	1-step ahead RMSFE											
10	0.184	0.141	0.106	0.206	0.140	0.101	0.107	0.102	0.100		0.155	0.106
20	0.156	0.155	0.110	0.164	0.164	0.101	0.107	0.103	0.101			0.114
50	0.138	0.143	0.121	0.142	0.151	0.109	0.107	0.102	0.101			0.153

TABLE 3. The results for Experiment C measured along the dimensions discussed in the main text.

estimates of the non-zero coefficients are asymptotically equivalent to those obtained by least squares applied to the model only including the relevant covariates.

We believe that these results may be useful for the applied researcher who is often faced with the curse of dimensionality when building VAR models since the number of parameters increases quadratically with the number of variables included. However, the LASSO and the Adaptive LASSO are applicable even in these situations.

In many applications it may be conjectured that all parameters belonging to a particular lag length are zero. In (1) this amounts to $\Gamma_l = 0$ for some $1 \leq l \leq p$. Imposing the correct restriction of a whole group of parameters being zero may thus be relevant in practice. We are currently working on this extension.

Furthermore, this paper has been concerned with stationary vector autoregressions and it is of interest to investigate if similar oracle inequalities may hold for non-stationary VARs.

7. Appendix

We start by stating a couple of preparatory lemmas. The first lemma bounds the probability of the maximum of all possible cross terms between explanatory variables and error terms becoming large. This bound will be used in the proof of Lemma 4 below.

Lemma 2. *For any $L_T > 0$,*

$$P\left(\max_{1\leq t\leq T}\max_{1\leq i\leq k}\max_{1\leq l\leq p}\max_{1\leq j\leq k}|y_{t-l,i}\epsilon_{t,j}| \geq L_T\right) \leq 2\exp\left(\frac{-L_T}{A\ln(1+T)\ln(1+k)^2\ln(1+p)\sigma_T^2}\right)$$

for some positive constant A.

In order to prove Lemma 2 Orlicz norms turn out to be useful since random variables with bounded Orlicz norms obey useful maximal inequalities. Let ψ be a non-decreasing convex function with $\psi(0) = 0$. Then, the Orlicz norm of a random variable X is given by

$$\|X\|_\psi = \inf\left\{C > 0 : E\psi\left(|X|/C\right) \leq 1\right\}$$

where, as usual, $\inf \emptyset = \infty$. By choosing $\psi(x) = x^p$ the Orlicz norm reduces to the usual L^p-norm since for $X \in L^p$, C equals $E(|X|^p)^{1/p}$. However, for our purpose $\psi(x) = e^x - 1$. One has the following maximal inequality:

Lemma 3 (Lemma 2.2.2 from Van Der Vaart and Wellner (1996))**.** *Let $\psi(x)$ be a convex, non-decreasing, non-zero function with $\psi(0) = 0$ and $\limsup_{x,y\to\infty} \psi(x)\psi(y)/\psi(cxy) < \infty$ for some constant c. Then for any random variables, $X_1, ..., X_m$,*

$$\left\|\max_{1\leq i\leq m} X_i\right\|_\psi \leq K\psi^{-1}(m) \max_{1\leq i\leq m} \|X_i\|_\psi$$

for a constant K depending only on ψ.

Notice that this result is particularly useful if $\psi^{-1}(x)$ only increases slowly which is the case when $\psi(x)$ increases very fast as in our case.

Proof of Lemma 2. Let $\psi(x) = e^x - 1$. First we show that $\left\|\max_{1\leq t\leq T} \max_{1\leq i\leq k} \max_{1\leq l\leq p} \max_{1\leq j\leq k} y_{t-l,i}\epsilon_{t,j}\right\|_\psi < \infty$. Repeated application of Lemma 3 yields

$$\begin{aligned}&\left\|\max_{1\leq t\leq T} \max_{1\leq i\leq k} \max_{1\leq l\leq p} \max_{1\leq j\leq k} y_{t-l,i}\epsilon_{t,j}\right\|_\psi \\ &\leq K^4 \ln(1+T)\ln(1+k)^2\ln(1+p) \max_{1\leq t\leq T} \max_{1\leq i\leq k} \max_{1\leq l\leq p} \max_{1\leq j\leq k} \left\|y_{t-l,i}\epsilon_{t,j}\right\|_\psi \end{aligned} \tag{17}$$

Next, we turn to bounding $\left\|y_{t-l,i}\epsilon_{t,j}\right\|_\psi$ uniformly in $1 \leq i, j \leq k,\ 1 \leq l \leq p$ and $1 \leq t \leq T$. Since $y_{t-l,i}$ and $\epsilon_{t,j}$ are both gaussian with mean 0 and variances $\sigma^2_{i,y}$ and $\sigma^2_{j,\epsilon}$ respectively it follows by a standard estimate on gaussian tails (see e.g. Billingsley (1999) page 263) that for any $x > 0$

$$\begin{aligned}P\left(|y_{t-l,i}\epsilon_{t,j}| > x\right) &\leq P\left(|y_{t-l,i}| > \sqrt{x}\right) + P\left(|\epsilon_{t,i}| > \sqrt{x}\right) \leq 2e^{-x/2\sigma^2_{i,y}} + 2e^{-x/2\sigma^2_{j,\epsilon}} \\ &\leq 4e^{\frac{-x}{2\sigma^2_T}}\end{aligned}$$

Hence, $\left\{y_{t-l,i}\epsilon_{t,j}\right\}$ has subexponential tails[8] and it follows from Lemma 2.2.1 in Van Der Vaart and Wellner (1996) that

$$\left\|y_{t-l,i}\epsilon_{t,j}\right\|_\psi \leq 10\sigma^2_T$$

Using this in (17) yields

$$\begin{aligned}\left\|\max_{1\leq t\leq T} \max_{1\leq i\leq k} \max_{1\leq l\leq p} \max_{1\leq j\leq k} y_{t-l,i}\epsilon_{t,j}\right\|_\psi &\leq K^4 \ln(1+T)\ln(1+k)^2\ln(1+p)10\sigma^2_T \\ &= A\ln(1+T)\ln(1+k)^2\ln(1+p)\sigma^2_T := f(T)\end{aligned}$$

where $A := 10K^4$. Finally, by Markov's inequality, the definition of the Orlicz norm, and the fact that $1 \wedge \psi(x)^{-1} = 1 \wedge (e^x - 1)^{-1} \leq 2e^{-x}$,

[8] A random variable X is said to have subexponential tails if there exists constants K and C such that for every $x > 0$, $P(|X| > x) \leq Ke^{-Cx}$.

$$
\begin{aligned}
&P\left(\max_{1\le t\le T}\max_{1\le i\le k}\max_{1\le l\le p}\max_{1\le j\le k}|y_{t-l,i}\epsilon_{t,j}|\ge L_T\right)\\
&=P\left(\psi\left(\max_{1\le t\le T}\max_{1\le i\le k}\max_{1\le l\le p}\max_{1\le j\le k}|y_{t-l,i}\epsilon_{t,j}|/f(T)\right)\ge\psi\left(L_T/f(T)\right)\right)\\
&\le 1\wedge\frac{E\psi\left(\max_{1\le t\le T}\max_{1\le i\le k}\max_{1\le l\le p}\max_{1\le j\le k}|y_{t-l,i}\epsilon_{t,j}|/f(T)\right)}{\psi\left[L_T/f(T)\right]}\\
&\le 1\wedge\frac{1}{\psi\left[L_T/f(T)\right]}\\
&\le 2\exp\left(-L_T/f(T)\right)\\
&=2\exp(-L_T/[A\ln(1+T)\ln(1+k)^2\ln(1+p)\sigma_T^2])
\end{aligned}
$$

□

Lemma 4. *Let* $\lambda_T=\sqrt{8\ln(1+T)^5\ln(1+k)^4\ln(1+p)^2\ln(k^2p)\sigma_T^4/T}$. *Then,*

$$
P\left(\max_{1\le i\le k}\max_{1\le l\le p}\max_{1\le j\le k}\left|\frac{1}{T}\sum_{t=1}^{T}y_{t-l,i}\epsilon_{t,j}\right|\ge\frac{\lambda_T}{2}\right)\le 2(k^2p)^{1-\ln(1+T)}+2(1+T)^{-A}
$$

Proof. Observe that for $\mathcal{F}_t=\sigma\left(\{\epsilon_s,\ s=1,...,t;\ y_s,\ s=1,...,t\}\right)$ being the natural filtration $\left\{y_{t-l,i}\epsilon_{t,j},\mathcal{F}_t\right\}_{t=1}^{\infty}$ defines a martingale difference sequence for every $1\le i,j\le k$ and $1\le l\le p$. Hence, by subadditivity of the probability measure it follows that for any $L_T>0$,

$$
\begin{aligned}
&P\left(\max_{1\le i\le k}\max_{1\le l\le p}\max_{1\le j\le k}\left|\frac{1}{T}\sum_{t=1}^{T}y_{t-l,i}\epsilon_{t,j}\right|\ge\frac{\lambda_T}{2}\right)\\
&=P\left(\bigcup_{i=1}^{k}\bigcup_{l=1}^{p}\bigcup_{j=1}^{k}\left\{\left|\frac{1}{T}\sum_{t=1}^{T}y_{t-l,i}\epsilon_{t,j}\right|\ge\frac{\lambda_T}{2}\right\}\right)\\
&\le P\left(\bigcup_{i=1}^{k}\bigcup_{l=1}^{p}\bigcup_{j=1}^{k}\left\{\left|\frac{1}{T}\sum_{t=1}^{T}y_{t-l,i}\epsilon_{t,j}\right|\ge\frac{\lambda_T}{2}\right\}\cap\bigcap_{t=1}^{T}\bigcap_{i=1}^{k}\bigcap_{l=1}^{p}\bigcap_{j=1}^{k}\left\{|y_{t-l,i}\epsilon_{t,j}|<L_T\right\}\right)\\
&+P\left(\left\{\bigcap_{t=1}^{T}\bigcap_{i=1}^{k}\bigcap_{l=1}^{p}\bigcap_{j=1}^{k}\left\{|y_{t-l,i}\epsilon_{t,j}|<L_T\right\}\right\}^c\right)\\
&\le\sum_{i=1}^{k}\sum_{l=1}^{p}\sum_{j=1}^{k}P\left(\left|\frac{1}{T}\sum_{t=1}^{T}y_{t-l,i}\epsilon_{t,j}\right|\ge\frac{\lambda_T}{2},\bigcap_{t=1}^{T}\left\{|y_{t-l,i}\epsilon_{t,j}|<L_T\right\}\right)\\
&+P\left(\max_{1\le t\le T}\max_{1\le i\le k}\max_{1\le l\le p}\max_{1\le j\le k}|y_{t-l,i}\epsilon_{t,j}|\ge L_T\right)
\end{aligned}
$$

Next, using the Azuma-Hoeffding inequality[9] on the first term and Lemma 2 on the second term with $L_T=\ln(1+T)^2\ln(1+k)^2\ln(1+p)\sigma_T^2$ yields,

[9]The Azuma-Hoeffding inequality is now applicable since we apply it on the set where the summands are bounded by L_T.

$$
\begin{aligned}
&P\left(\max_{1\leq i\leq k}\max_{1\leq l\leq p}\max_{1\leq j\leq k}\left|\frac{1}{T}\sum_{t=1}^{T}y_{t-l,i}\epsilon_{t,j}\right|\geq\frac{\lambda_T}{2}\right)\\
&\leq k^2p\cdot 2\exp\left(-\frac{T\lambda_T^2}{8L_T^2}\right)+2\exp\left(-\frac{\ln(1+T)}{A}\right)\\
&=2k^2p\cdot\exp\Big(-\ln(1+T)\ln(k^2p)\Big)+2(1+T)^{-1/A}\\
&=2(k^2p)^{1-\ln(1+T)}+2(1+T)^{-1/A}
\end{aligned}
$$

□

Proof of Theorem 1. By the minimizing property of $\hat{\beta}$ it follows that

$$\frac{1}{T}\|y-X\hat{\beta}\|^2+2\lambda_T\|\hat{\beta}\|_{\ell_1}\leq\frac{1}{T}\|y-X\beta^*\|^2+2\lambda_T\|\beta^*\|_{\ell_1}$$

which using that $y=X\beta^*+\epsilon$ yields

$$\frac{1}{T}\|\epsilon\|^2+\frac{1}{T}\|X(\hat{\beta}-\beta^*)\|^2-\frac{2}{T}\epsilon'X(\hat{\beta}-\beta^*)+2\lambda_T\|\hat{\beta}\|_{\ell_1}\leq\frac{1}{T}\|\epsilon\|^2+2\lambda_T\|\beta^*\|_{\ell_1}$$

Or, equivalently

$$\frac{1}{T}\|X(\hat{\beta}-\beta^*)\|^2\leq\frac{2}{T}\epsilon'X(\hat{\beta}-\beta^*)+2\lambda_T\left(\|\beta^*\|_{\ell_1}-\|\hat{\beta}\|_{\ell_1}\right)\tag{18}$$

So to bound $\frac{1}{T}\|X(\hat{\beta}-\beta^*)\|^2$ one must bound on $\frac{2}{T}\epsilon'X(\hat{\beta}-\beta^*)$. Note that on the set

$$\mathcal{B}_T=\left\{\max_{1\leq i\leq k}\max_{1\leq l\leq p}\max_{1\leq j\leq k}\left|\frac{1}{T}\sum_{t=1}^{T}y_{t-l,i}\epsilon_{t,j}\right|<\frac{\lambda_T}{2}\right\}\tag{19}$$

where λ_T is as in Lemma 4, one has

$$\frac{2}{T}\epsilon'X(\hat{\beta}-\beta^*)\leq 2\left\|\frac{1}{T}\epsilon'X\right\|_{\ell_\infty}\|\hat{\beta}-\beta^*\|_{\ell_1}\leq\lambda_T\|\hat{\beta}-\beta^*\|_{\ell_1}$$

Putting things together, on $\mathcal{B}_T$,

$$\frac{1}{T}\|X(\hat{\beta}-\beta^*)\|^2\leq\lambda_T\|\hat{\beta}-\beta^*\|_{\ell_1}+2\lambda_T\left(\|\beta^*\|_{\ell_1}-\|\hat{\beta}\|_{\ell_1}\right)$$

Adding $\lambda_T\|\hat{\beta}-\beta^*\|_{\ell_1}$ yields

$$\frac{1}{T}\|X(\hat{\beta}-\beta^*)\|^2+\lambda_T\|\hat{\beta}-\beta^*\|_{\ell_1}\leq 2\lambda_T\left(\|\hat{\beta}-\beta^*\|_{\ell_1}+\|\beta^*\|_{\ell_1}-\|\hat{\beta}\|_{\ell_1}\right)\tag{20}$$

which is inequality (4). To obtain inequality (5) notice that

$$\|\hat{\beta}-\beta^*\|_{\ell_1}+\|\beta^*\|_{\ell_1}-\|\hat{\beta}\|_{\ell_1}=\|\hat{\beta}_J-\beta^*_J\|_{\ell_1}+\|\beta^*_J\|_{\ell_1}-\|\hat{\beta}_J\|_{\ell_1}$$

In addition,

$$\|\hat{\beta}_J-\beta^*_J\|_{\ell_1}+\|\beta^*_J\|_{\ell_1}-\|\hat{\beta}_J\|_{\ell_1}\leq 2\|\hat{\beta}_J-\beta^*_J\|_{\ell_1}$$

by continuity of the norm. Furthermore,

$$\|\hat{\beta}_J-\beta^*_J\|_{\ell_1}+\|\beta^*_J\|_{\ell_1}-\|\hat{\beta}_J\|_{\ell_1}\leq 2\|\beta^*_J\|_{\ell_1}$$

by subadditivity of the norm. Using the above two estimates in (20) yields inequality (5). Next notice that (5) gives

$$\lambda_T\|\hat{\beta}-\beta^*\|_{\ell_1}\leq 4\lambda_T\|\hat{\beta}_J-\beta^*_J\|_{\ell_1}$$

which is equivalent to

$$\|\hat{\beta}_{J^c}-\beta^*_{J^c}\|_{\ell_1}\leq 3\|\hat{\beta}_J-\beta^*_J\|_{\ell_1}$$

and establishes (6). To get inequality (7) notice that $\frac{1}{T}\left|X_j'(y - X\hat{\beta})\right| = \lambda_T$ for $j \in J(\hat{\beta})$ by the first order conditions for a minimum in (3). Hence, on $\mathcal{B}_T$,

$$\frac{1}{T}\left|X_j'X(\beta^* - \hat{\beta})\right| = \frac{1}{T}\left|X_j'(y - X\hat{\beta}_j) - X_j'\epsilon\right| \geq \frac{1}{T}\left|X_j'(y - X\hat{\beta})\right| - \frac{1}{T}\left|X_j'\epsilon\right| \geq \frac{1}{2}\lambda_T$$

This implies $\frac{2}{T\lambda_T}\left|X_j'X(\hat{\beta} - \beta^*)\right| \geq 1$ and so on $\mathcal{B}_T$

$$\begin{aligned}
|J(\hat{\beta})| &= \sum_{j \in J(\hat{\beta})} 1 \\
&\leq \frac{4}{T^2\lambda_T^2} \sum_{j \in J(\hat{\beta})} \left|X_j'X(\hat{\beta} - \beta^*)\right|^2 \\
&\leq \frac{4}{T^2\lambda_T^2} \left\|X'X(\hat{\beta} - \beta^*)\right\|^2 \\
&\leq \frac{4}{T\lambda_T^2} \phi_{\max}\left(\frac{XX'}{T}\right) \left\|X(\hat{\beta} - \beta^*)\right\|^2 \\
&= \frac{4}{T\lambda_T^2} \phi_{\max}(\Psi_T) \left\|X(\hat{\beta} - \beta^*)\right\|^2
\end{aligned}$$

where the last equality follows from $\phi_{\max}\left(\frac{XX'}{T}\right) = \phi_{\max}\left(\frac{X'X}{T}\right) = \phi_{\max}(\Psi_T)$. The lower bound on the probability with which (4)-(7) hold follows from the fact that $P(\mathcal{B}_T) \geq 1 - 2(k^2p)^{1-\ln(1+T)} - 2(1+T)^{-1/A}$ by Lemma 4. □

Proof of Theorem 2. By (5), Jensen's inequality and the restricted eigenvalue condition (which is applicable due to (6))

$$\frac{1}{T}\left\|X(\hat{\beta} - \beta^*)\right\|^2 \leq 4\lambda_T \left\|\hat{\beta}_J - \beta_J^*\right\|_{\ell_1} \leq 4\lambda_T\sqrt{s}\left\|\hat{\beta}_J - \beta_J^*\right\| \leq 4\lambda_T\sqrt{s}\frac{\left\|X(\hat{\beta} - \beta^*)\right\|}{\kappa\sqrt{T}}$$

Rearranging, yields (10). To establish (11) use (6), Jensen's inequality and (10):

$$\left\|\hat{\beta} - \beta^*\right\|_{\ell_1} \leq 4\left\|\hat{\beta}_J - \beta_J^*\right\|_{\ell_1} \leq 4\sqrt{s}\left\|\hat{\beta}_J - \beta_J^*\right\| \leq 4\sqrt{s}\frac{\left\|X(\hat{\beta} - \beta^*)\right\|}{\kappa\sqrt{T}} \leq \frac{16}{\kappa^2}s\lambda_T$$

Inequality (12) follows by inserting (10) into (7).

Regarding the last statement, let $\beta_{\min} > \|\hat{\beta} - \beta^*\|_{\ell_1}$. If there exists a $j \in J$ such that $\hat{\beta}_j = 0$, then $\|\hat{\beta} - \beta^*\|_{\ell_1} \geq \beta_{\min}$ which contradicts $\beta_{\min} > \|\hat{\beta} - \beta^*\|_{\ell_1}$. Hence, $\hat{\beta}_j \neq 0$ for any $j \in J$ □

Proof of Lemma 1. Let X_J denote the matrix consisting of the columns of X indexed by J. Then for $1 \leq i_h, j_h \leq k$ and $1 \leq l_h \leq p$ on

$$\tilde{\mathcal{B}}_T = \left\{ \max_{1 \leq h \leq s} \left| \frac{1}{T}\sum_{t=1}^{T} y_{t-l_h,i_h}\epsilon_{t,j_h} \right| < \frac{\tilde{\lambda}_T}{2} \right\}$$

one has, regarding $\left(\frac{1}{T}X_J'X_J\right)^{-1}$ as a bounded linear operator from $\ell_2(\mathbb{R}^s)$ to $\ell_2(\mathbb{R}^s)$ with induced operator norm given by $\phi_{\max}\left((\frac{1}{T}X_J'X_J)^{-1}\right) = 1/\phi_{\min}(\frac{1}{T}X_J'X_J)$,

$$\begin{aligned}
\left\|\hat{\beta}_{OLS} - \beta_J^*\right\|_{\ell_1} &\leq \sqrt{s}\left\|\left(\frac{1}{T}X_J'X_J\right)^{-1}\frac{1}{T}X'\epsilon\right\|_{\ell_2} \\
&\leq \sqrt{s}\left\|\left(\frac{1}{T}X_J'X_J\right)^{-1}\right\|_{\ell_2}\left\|\frac{1}{T}X_J'\epsilon\right\|_{\ell_2} \\
&\leq s\left\|\left(\frac{1}{T}X_J'X_J\right)^{-1}\right\|_{\ell_2}\left\|\frac{1}{T}X_J'\epsilon\right\|_{\ell_\infty} \\
&\leq \frac{s}{\phi_{\min}(\Psi_{J,J})}\tilde{\lambda}_T
\end{aligned}$$

That the probability of $\tilde{\mathcal{B}}_T$ must be at least $1-2s^{1-\ln(1+T)}-2(1+T)^{-1/A}$ follows from a slight modification of Lemmas 2 and 4 using that one only has to bound terms associated with relevant variables. □

Proof of Theorem 3. Observe that $s\lambda_T \to 0$ implies $\lambda_T < 1$ from a certain step and onwards and so $s\lambda_T^2 \to 0$. Hence, i) and ii) follow from (10) and (11) of Theorem 2 noting that the probability with which (10) and (11) hold tends to one. iii) follows trivially from (12) of the same theorem.

Regarding iv) let $\beta_{\min} > \frac{16}{c^2} s\lambda_T$. Then,

$$P\left(\bigcup_{j\in J}\left\{\hat{\beta}_j = 0\right\}\right) \leq P\left(\left\|\hat{\beta}-\beta^*\right\|_{\ell_1} > \frac{16}{c^2}s\lambda_T\right)$$
$$= 1 - P\left(\left\|\hat{\beta}-\beta^*\right\|_{\ell_1} \leq \frac{16}{c^2}s\lambda_T\right) \to 0$$

where the convergence to zero is a consequence of (11) in Theorem 2 and the fact that probability with which (11) holds tends to one. This establishes iv). □

Lemma 5. *If $k, p \in O(e^{T^a})$ and $s \in O(T^b)$ for some $a, b \geq 0$ satisfying $2b + 7a < 1$, then $s\lambda_T \to 0$.*

Proof of Lemma 5. $k, p \in O(e^{T^a})$ for some $a \geq 0$ implies that $1 + k, 1 + p \in O(e^{T^a})$ since e^{T^a} is bounded away from 0 (it tends to ∞ for $a > 0$). Hence[10],

$$s^2\lambda_T^2 \in O\left(T^{2b}\frac{\ln(1+T)^5 T^{4a} T^{2a}(2T^a + T^a)}{T}\right) = O\left(\ln(1+T)^5 T^{7a+2b-1}\right) \tag{21}$$

So it suffices to show that

$$\ln(1+T)^5 T^{7a+2b-1} \to 0$$

which is true under the hypothesis of the lemma. □

Lemma 6. *Let $K_T = \ln(1+k)^2 \ln(1+p)^2 \ln(T)\sigma_T^2$. Then,*

$$P\left(\max_{1\leq i,j\leq k}\max_{1\leq l,\tilde{l}\leq p}\left|\frac{1}{T}\sum_{t=1}^{T} y_{t-l,i}y_{t-\tilde{l},j}\right| \geq K_T\right) \leq 2T^{-1/A}$$

for some constant $A > 0$.

Proof. The proof is based on the same idea as in Lemma 2 in its use of Orlicz norms. First bound

$$\left\|\max_{1\leq i,j\leq k}\max_{1\leq l,\tilde{l}\leq p}\left|\frac{1}{T}\sum_{t=1}^{T} y_{t-l,i}y_{t-\tilde{l},j}\right|\right\|_{\psi}$$

where $\|\cdot\|_{\psi}$ denotes the same Orlicz norm as in Lemma 2. To this end, notice that by the gaussianity of the $y_{t-l,i}$ for any $x > 0$

$$P\left(|y_{t-l,i}y_{t-\tilde{l},j}| \geq x\right) \leq P\left(|y_{t-l,i}| \geq \sqrt{x}\right) + P\left(|y_{t-\tilde{l},j}| \geq \sqrt{x}\right)$$
$$\leq 2\exp(-x/\sigma_{i,y}^2) + 2\exp(-x/\sigma_{j,y}^2) \leq 4\exp(-x/\sigma_T^2)$$

Hence, $\{y_{t-l,i}y_{t-\tilde{l},j},\ t = 1, ..., T\}$ has subexponential tails and it follows from Lemma 2.2.1 in Van Der Vaart and Wellner (1996) that $\left\|y_{t-l,i}y_{t-\tilde{l},j}\right\|_{\psi} \leq 10\sigma_T^2$. This implies that

$$\left\|\frac{1}{T}\sum_{t=1}^{T} y_{t-l,i}y_{t-\tilde{l},j}\right\|_{\psi} \leq \frac{1}{T}\sum_{t=1}^{T}\left\|y_{t-l,i}y_{t-\tilde{l},j}\right\|_{\psi} \leq 10\sigma_T^2$$

[10]By the definition of $\lambda_T = \sqrt{8\ln(1+T)^5\ln(1+k)^4\ln(1+p)^2\ln(k^2p)\sigma_T^4/T}$ one has that $\lambda_T \in O\left(\sqrt{\frac{\ln(1+T)^5 T^{4a} T^{2a}(2T^a+T^a)}{T}}\right) = O\left(\sqrt{\frac{\ln(1+T)^5 T^{7a}}{T}}\right)$.

By Lemma 3 this implies that

$$\left\| \max_{1\le i,j\le k} \max_{1\le l,\tilde{l}\le p} \frac{1}{T}\sum_{t=1}^{T} y_{t-l,i}y_{t-\tilde{l},j} \right\|_{\psi} \le K^4 \ln(1+k)^2 \ln(1+p)^2 10\sigma_T^2$$
$$= A\ln(1+k)^2\ln(1+p)^2\sigma_T^2$$

where $A = 10K^4$. By the same trick as in Lemma 2

$$P\left(\max_{1\le i,j\le k} \max_{1\le l,\tilde{l}\le p} \left| \frac{1}{T}\sum_{t=1}^{T} y_{t-l,i}y_{t-\tilde{l},j} \right| \ge K_T \right) \le 2\exp(-\ln(T)/A) = 2T^{-1/A}$$

□

Proof of Theorem 4. Set $w = (1/|\hat{\beta}_1|, ..., 1/|\hat{\beta}_{k^2p}|)$ and $b = (\text{sign}(\beta_j^*)w_j)_{j\in J}$. From Zhou et al. (2009) $\text{sign}(\hat{\beta}) = \text{sign}(\beta^*)$ if and only if

$$\forall j \in J^c : \left| \Psi_{j,J}(\Psi_{J,J})^{-1}\left(\frac{X_J'\epsilon}{T} - \lambda_T b \right) - \frac{X_j'\epsilon}{T} \right| \le \lambda_T w_j \tag{22}$$

and

$$\text{sign}\left(\beta_J^* + (\Psi_{J,J})^{-1}\left[\frac{X_J'\epsilon}{T} - \lambda_T b \right] \right) = \text{sign}(\beta_J^*) \tag{23}$$

Let

$$\mathcal{C}_T = \left\{ \max_{1\le i,j\le k} \max_{1\le l,\tilde{l}\le p} \left| \frac{1}{T}\sum_{t=1}^{T} y_{t-l,i}y_{t-\tilde{l},j} \right| < K_T \right\}$$

where K_T is as in Lemma 6. We shall be working on $\mathcal{C}_T \cap \mathcal{B}_T$ where $\mathcal{B}_T$ is as defined in (19). Consider (22) for a given $j \in J^c$. By the triangle inequality it suffices to show that

$$\left| \Psi_{j,J}(\Psi_{J,J})^{-1}\left(\frac{X_J'\epsilon}{T} - \lambda_T b \right) \right| + \left| \frac{X_j'\epsilon}{T} \right| \le \lambda_T w_j \tag{24}$$

Bound the first term on the left hand side as follows:

$$\left| \Psi_{j,J}(\Psi_{J,J})^{-1}\left(\frac{X_J'\epsilon}{T} - \lambda_T b \right) \right| \le \left\| \Psi_{j,J}(\Psi_{J,J})^{-1} \right\|_{\ell_1} \left\| \frac{X_J'\epsilon}{T} - \lambda_T b \right\|_{\ell_\infty}$$
$$\le \sqrt{s}\left\| \Psi_{j,J}(\Psi_{J,J})^{-1} \right\|_{\ell_2} \left(\left\| \frac{X_J'\epsilon}{T} \right\|_{\ell_\infty} + \|\lambda_T b\|_{\ell_\infty} \right)$$

Considering $(\Psi_{J,J})^{-1}$ as a bounded linear operator $\ell_2(\mathbb{R}^s) \to \ell_2(\mathbb{R}^s)$, the induced operator norm is given by $\phi_{\max}((\Psi_{J,J})^{-1}) = 1/\phi_{\min}(\Psi_{J,J})$ and so

$$\left\| \Psi_{j,J}(\Psi_{J,J})^{-1} \right\|_{\ell_2} \le \frac{\left\| \Psi_{j,J} \right\|_{\ell_2}}{\phi_{\min}(\Psi_{J,J})} \le \frac{\sqrt{s}\left\| \Psi_{j,J} \right\|_{\ell_\infty}}{\phi_{\min}(\Psi_{J,J})} \le \frac{\sqrt{s}K_T}{\phi_{\min}(\Psi_{J,J})}$$

where the last estimate holds on $\mathcal{C}_T$. By Lemma 4 it follows that on $\mathcal{B}_T$

$$\left\| \frac{X_J'\epsilon}{T} \right\|_{\ell_\infty} \le \frac{\lambda_T}{2} \tag{25}$$

Next, since $\beta_{\min} \ge 2\left\| \hat{\beta} - \beta^* \right\|_{\ell_1}$ one has for all $j \in J$,

$$|\hat{\beta}_j| \ge |\beta_j^*| - |\hat{\beta}_j - \beta_j^*| \ge \beta_{\min} - \left\| \hat{\beta}_j - \beta_j^* \right\|_{\ell_1} \ge \beta_{\min}/2$$

one gets

$$\|\lambda_T b\|_{\ell_\infty} = \|\lambda_T w_J\|_{\ell_\infty} = \lambda_T \max_{j\in J}\left| \frac{1}{\hat{\beta}_j} \right| \le \frac{2\lambda_T}{\beta_{\min}}. \tag{26}$$

Lastly, on $\mathcal{B}_T$,

$$\left|\frac{X_j'\epsilon}{T}\right| \leq \frac{\lambda_T}{2}$$

for every $j \in J^c$. Hence, uniformly in $j \in J^c$,

$$\left|\Psi_{j,J}(\Psi_{J,J})^{-1}\left(\frac{X_J'\epsilon}{T} - \lambda_T b\right)\right| + \left|\frac{X_j'\epsilon}{T}\right| \leq \frac{sK_T}{\phi_{\min}(\Psi_{J,J})}\left(\frac{\lambda_T}{2} + \frac{2\lambda_T}{\beta_{\min}}\right) + \frac{\lambda_T}{2}$$

Now bound the right hand side in (24) from below. For every $j \in J^c$

$$|\lambda_T w_j| = \lambda_T \frac{1}{|\hat{\beta}_j|} \geq \lambda_T \frac{1}{\|\hat{\beta} - \beta^*\|_{\ell_1}}$$

This implies that (24), and hence (22), is satisfied if

$$\frac{sK_T}{\phi_{\min}(\Psi_{J,J})}\left(\frac{\lambda_T}{2} + \frac{2\lambda_T}{\beta_{\min}}\right) + \frac{\lambda_T}{2} \leq \lambda_T \frac{1}{\|\hat{\beta} - \beta^*\|_{\ell_1}}$$

or equivalently

$$\frac{sK_T}{\phi_{\min}(\Psi_{J,J})}\left(\frac{1}{2} + \frac{2}{\beta_{\min}}\right)\|\hat{\beta} - \beta^*\|_{\ell_1} + \frac{\|\hat{\beta} - \beta^*\|_{\ell_1}}{2} \leq 1$$

which is (15). To verify (16) it suffices to show that

$$\left\|(\Psi_{J,J})^{-1}\left(\frac{X_J'\epsilon}{T} - \lambda_T b\right)\right\|_{\ell_\infty} \leq \beta_{\min} \tag{27}$$

Considering $(\Psi_{J,J})^{-1}$ as a bounded linear operator $\ell_\infty(\mathbb{R}^s) \to \ell_\infty(\mathbb{R}^s)$ it follows that:

$$\begin{aligned}
\left\|(\Psi_{J,J})^{-1}\left(\frac{X_J'\epsilon}{T} - \lambda_T b\right)\right\|_{\ell_\infty} &\leq \left\|(\Psi_{J,J})^{-1}\right\|_{\ell_\infty}\left\|\frac{X_J'\epsilon}{T} - \lambda_T b\right\|_{\ell_\infty} \\
&\leq \sqrt{s}\left\|(\Psi_{J,J})^{-1}\right\|_{\ell_2}\left(\left\|\frac{X_J'\epsilon}{T}\right\|_{\ell_\infty} + \|\lambda_T b\|_{\ell_\infty}\right) \\
&\leq \frac{\sqrt{s}}{\phi_{\min}(\Psi_{J,J})}\left(\frac{\lambda_T}{2} + \frac{2\lambda_T}{\beta_{\min}}\right)
\end{aligned}$$

where the first estimate uses that $\left\|(\Psi_{J,J})^{-1}\right\|_{\ell_\infty} \leq \sqrt{s}\left\|(\Psi_{J,J})^{-1}\right\|_{\ell_2}$, c.f. Horn and Johnson (1990) page 314, and the last estimate follows from (25) and (26). Inserting into (27) completes the proof □

Proof of Theorem 5. First we prove i). To do so we show that $\beta_{\min} \geq a_T \frac{s^2 K_T \lambda_T}{\phi_{\min}(\Psi_{J,J})\kappa^2} \to 0$ and $\beta_{\min} \geq a_T \frac{s^{1/4}\lambda_T^{1/2}}{\sqrt{\phi_{\min}(\Psi_{J,J})}} \to 0$ in probability imply that (15) and (16) are valid asymptotically. We shall work on the set $\mathcal{B}_T \cap \mathcal{C}_T$ from Theorem 4 which has measure one asymptotically. Since it is assumed that $\beta_{\min}$ is bounded from below by sequences that converges to 0, $\beta_{\min}$ may itself be assumed to tend to zero[11]. Furthermore, on $\mathcal{C}_T$, $\phi_{\min}(\Psi_{J,J})$ is bounded from above by the smallest diagonal element of $\Psi_{J,J}$ which by Lemma 6 is less than K_T. Hence, the leading term in (15) is

$$\frac{2sK_T}{\phi_{\min}(\Psi_{J,J})}\frac{\|\hat{\beta} - \beta^*\|_{\ell_1}}{\beta_{\min}} \leq \frac{32s^2 K_T \lambda_T}{\phi_{\min}(\Psi_{J,J})\kappa^2 \beta_{\min}}$$

which converges to 0 in probability by assumption. Using again that $\phi_{\min}(\Psi_{J,J})$ is bounded from above by K_T with probability tending to one, the above display also yields that $\frac{\|\hat{\beta} - \beta^*\|_{\ell_1}}{\beta_{\min}}$ tends to zero with probability tending to one.

[11]This is without loss of generality since if $\beta_{\min}$ does not tend to zero the left hand sides in (15) and (16) will be smaller hence making these inequalities easier to satisfy since a smaller term will then be the leading one. Furthermore, this is the setting we are interested in since we want to be able to distinguish even small non-zero coefficients from the zero ones.

Since $\beta_{\min}$ can be assumed tending to zero, the leading term on the left hand side in (16) is

$$\frac{\sqrt{s}}{\phi_{\min}(\Psi_{J,J})}\frac{2\lambda_T}{\beta_{\min}}$$

which is less than $\beta_{\min}$ asymptotically since $\beta_{\min} \geq a_T \frac{s^{1/4}\lambda_T^{1/2}}{\sqrt{\phi_{\min}(\Psi_{J,J})}} \to 0$ with probability tending to one. This establishes i). To establish ii) it suffices to show that under the stated conditions the requirements of part i) are met. Hence, one must show that there exists a sequence $a_T \to \infty$ such that with probability tending to one

$$\beta_{\min} \geq a_T \frac{s^2 K_T \lambda_T}{\phi_{\min}(\Psi_{J,J})\kappa^2} \to 0$$
$$\beta_{\min} \geq a_T \frac{s^{1/4}\lambda_T^{1/2}}{\sqrt{\phi_{\min}(\Psi_{J,J})}} \to 0$$

Notice that by the definitions of λ_T and K_T and the assumption that $k, p \in O(e^{T^a})$ and $s \in O(T^b)$ for some $a, b \geq 0$ it follows in the same way as in footnote 10 that $K_T \in O(T^{4a}\ln(T))$ and $\lambda_T \in O\left(\sqrt{\ln(T)^5 T^{7a}/T}\right)$. Hence,

$$\frac{s^2 K_T \lambda_T}{\phi_{\min}(\Psi_{J,J})\kappa^2} \in O_p\left(s^2 K_T \lambda_T\right) \subseteq O_p\left(T^{2b}T^{4a+(7/2)a-1/2}\ln(T)^{1+5/2}\right) = O_p(b_T)$$

and

$$\frac{s^{1/4}\lambda_T^{1/2}}{\sqrt{\phi_{\min}(\Psi_{J,J})}} \in O_p\left(s^{1/4}\lambda_T^{1/2}\right) \subseteq O_p\left(T^{b/4}T^{(7/4)a-1/4}\ln(T)^{5/4}\right) = O_p(c_T)$$

which implies that

$$\ln(T)^{1/2}\frac{s^2 K_T \lambda_T}{\phi_{\min}(\Psi_{J,J})\kappa^2} \in O_p(\ln(T)^{1/2} b_T) \subseteq o_p(\ln(T) b_T) \tag{28}$$

and

$$\ln(T)^{1/2}\frac{s^{1/4}\lambda_T^{1/2}}{\sqrt{\phi_{\min}(\Psi_{J,J})}} \in O_p(\ln(T)^{1/2} c_T) \subseteq o_p(\ln(T) c_T) \tag{29}$$

Next notice that $\ln(T)^{1/2} b_T \to 0$ since

$$(\ln(T)^{1/2} b_T)^2 = T^{4b+15a-1}\ln(T)^8 \to 0$$

since $4b + 15a < 1$. Similarly, $\ln(T)^{1/2} c_T \to 0$ since

$$(\ln(T)^{1/2} c_T)^4 = T^{b+7a-1}\ln(T)^7 \to 0$$

since $b + 7a \leq 4b + 15a < 1$. These observations imply that

$$\ln(T)^{1/2}\frac{s^2 K_T \lambda_T}{\phi_{\min}(\Psi_{J,J})\kappa^2} \in o_p(1) \text{ and } \ln(T)^{1/2}\frac{s^{1/4}\lambda_T^{1/2}}{\sqrt{\phi_{\min}(\Psi_{J,J})}} \in o_p(1)$$

Hence, setting $a_T = \ln(T)^{1/2}$, on a set with arbitrarily large probability for some constant c,

$$\begin{aligned}\beta_{\min} &\geq c\ln(T)[b_T \vee c_T] \\ &\geq c\ln(T) b_T \\ &\geq \ln(T)^{1/2}\frac{s^2 K_T \lambda_T}{\phi_{\min}(\Psi_{J,J})\kappa^2} \to 0\end{aligned}$$

where the third estimate follows from (28). In the same way

$$\begin{aligned}\beta_{\min} &\geq c\ln(T)[b_T \vee c_T] \\ &\geq c\ln(T) c_T \\ &\geq \ln(T)^{1/2}\frac{s^{1/4}\lambda_T^{1/2}}{\sqrt{\phi_{\min}(\Psi_{J,J})}} \to 0\end{aligned}$$

which verifies the hypotheses from part i) of this theorem. □

Proof of Theorem 6. The notation is as in the statement of Theorem 5 part ii). Under the assumptions of that theorem, the Adaptive LASSO is sign consistent. Hence, with probability tending to one, the estimates of the non-zero coefficients satisfy the first order condition

$$\frac{\partial L(\beta)}{\partial \beta_J} = -2X_J'(Y - X\tilde{\beta}) + 2\lambda_T b = 0$$

where $b = (\text{sign}(\beta_j^*)/|\hat{\beta}_j|)_{j\in J}$. Using that $Y = X_J\beta_J^* + \epsilon$ this is equivalent to

$$-2X_J'\left(\epsilon - X_J(\tilde{\beta}_J - \beta_J^*) - X_{J^c}\tilde{\beta}_{J^c}\right) + 2\lambda_T b = 0$$

and so, with probability tending to one, for any $s \times 1$ vector α with norm 1 one has

$$\begin{aligned}&\sqrt{T}\alpha'(\tilde{\beta}_J - \beta_J^*)\\ &= \frac{1}{\sqrt{T}}\alpha'(\Psi_{J,J})^{-1}X_J'\epsilon - \sqrt{T}\alpha'(\Psi_{J,J})^{-1}\Psi_{J,J^c}\tilde{\beta}_{J^c} - \frac{\lambda_T}{\sqrt{T}}\alpha'(\Psi_{J,J})^{-1}b \qquad (30)\end{aligned}$$

The first term on the right hand side in (30) is recognized as $\sqrt{T}\alpha'(\hat{\beta}_{OLS} - \beta_J^*)$. Hence, to establish the theorem, it suffices to show that the second and the third term on the right hand side tend to zero in probability. Since $P(\tilde{\beta}_{J^c} = 0) \to 1$ the second term vanishes in probability. Regarding the third term, notice that

$$\left|\frac{\lambda_T}{\sqrt{T}}\alpha'(\Psi_{J,J})^{-1}b\right| \leq \frac{\lambda_T}{\sqrt{T}}\left|\alpha'(\Psi_{J,J})^{-1}b\right| \leq \frac{\lambda_T}{\sqrt{T}}\sqrt{\alpha'(\Psi_{J,J})^{-2}\alpha b'b}$$

Now, almost surely,

$$\alpha'\Psi_{J,J}^{-2}\alpha \leq \alpha'\alpha\phi_{\max}((\Psi_{J,J})^{-2}) = \alpha'\alpha 1/\phi_{\min}((\Psi_{J,J})^2) \leq 1/\tilde{c}^2$$

since α has norm one. Note that for all $j \in J$

$$|\hat{\beta}_j| \geq |\beta_k^*| - |\hat{\beta}_j - \beta_j^*| \geq |\beta_{\min}| - \|\hat{\beta} - \beta^*\|_{\ell_1}$$

and so by subadditivity of $x \mapsto \sqrt{x}$

$$\frac{\lambda_T}{\sqrt{T}}\sqrt{b'b} = \frac{\lambda_T}{\sqrt{T}}\sqrt{\sum_{j\in J}\frac{1}{\hat{\beta}_j^2}} \leq \frac{\lambda_T}{\sqrt{T}}\frac{s}{|\beta_{\min}| - \|\hat{\beta} - \beta^*\|_{\ell_1}} = \frac{s\lambda_T}{\sqrt{T}|\beta_{\min}|}\frac{1}{1 - \|\hat{\beta} - \beta^*\|_{\ell_1}/|\beta_{\min}|}$$

tends to zero in probability. Since $\kappa \geq \tilde{c}$ it follows from (11) that $\|\hat{\beta} - \beta^*\|_{\ell_1} \in O_p(s\lambda_T)$. But $s\lambda_T \in O(\ln(1+T)^{5/2}T^{(7/2)a+b-1/2})$ by (21). Hence, $\|\hat{\beta} - \beta^*\|_{\ell_1} \in O_p(\ln(1+T)^{5/2}T^{(7/2)a+b-1/2})$ and so, since $\beta_{\min} \in \Omega(\ln(T)[b_T \vee c_T]) \subseteq \Omega(\ln(T)b_T) = \Omega(T^{2b}T^{4a+(7/2)a-1/2}\ln(T)^{2+5/2})$,

$$\frac{\|\hat{\beta} - \beta^*\|_{\ell_1}}{|\beta_{\min}|} \in O_p\left(\frac{\ln(1+T)^{5/2}T^{(7/2)a+b-1/2}}{T^{2b}T^{4a+(7/2)a-1/2}\ln(T)^{2+5/2}}\right) = O_p\left(\frac{\ln(1+T)^{5/2}}{\ln(T)^{2+5/2}}T^{-4a-b}\right)$$

Since $\frac{\ln(1+T)^{5/2}}{\ln(T)^{2+5/2}}T^{-4a-b} \to 0$ it follows that $\frac{\|\hat{\beta}-\beta^*\|_{\ell_1}}{\beta_{\min}} \in o_p(1)$. Also, $15a + 4b < 1$ is more than sufficient for Lemma 5 to yield that $s\lambda_T \to 0$ and $\beta_{\min} \in \Omega\left(\ln(T)([b_T \vee c_T]\right) \subseteq \Omega\left(\ln(T)T^{-1/4}\right)$ implies that $\sqrt{T}\beta_{\min} \to \infty$ and the theorem follows. □

Proof of Corollary 1. Let $\epsilon > 0$ be given. Then,

$$\begin{aligned}\left\{\sup_{\alpha:\|\alpha\|\leq 1}\sqrt{T}\left|\alpha'(\tilde{\beta}_J - \beta_{OLS})\right| < \epsilon\right\} &\subseteq \bigcap_{j\in J}\left\{\sqrt{T}\left|\tilde{\beta}_j - \beta_{OLS,j}\right| < \epsilon\right\}\\ &\subseteq \left\{\sqrt{T}\left\|(\tilde{\beta}_J - \beta_{OLS})\right\|_{\ell_1} < s\epsilon\right\}\\ &= \left\{\frac{\sqrt{T}}{s}\left\|(\tilde{\beta}_J - \beta_{OLS})\right\|_{\ell_1} < \epsilon\right\}\end{aligned}$$

And so

$$P\left(\frac{\sqrt{T}}{s}\left\|(\tilde{\beta}_J - \beta_{OLS})\right\|_{\ell_1} < \epsilon\right) \geq P\left(\sup_{\alpha:\|\alpha\|\leq 1} \sqrt{T}|\alpha'(\tilde{\beta}_J - \beta_{OLS})| < \epsilon\right) \to 1$$

by Theorem 6. Since $\epsilon > 0$ was arbitrary, this proves that $\left\|(\tilde{\beta}_J - \beta_{OLS})\right\|_{\ell_1} \in o_p\left(\frac{s}{\sqrt{T}}\right)$. Hence, by the triangle inequality and Lemma 1

$$\|\tilde{\beta}_J - \beta_J^*\|_{\ell_1} \leq \|\tilde{\beta}_J - \beta_{OLS}\|_{\ell_1} + \|\tilde{\beta}_{OLS} - \beta_J^*\|_{\ell_1} \in O_p(\tilde{\lambda}_T s)$$

since $\tilde{\lambda}_T s > \frac{s}{\sqrt{T}}$. □

References

Bai, J. and S. Ng (2008). Large dimensional factor analysis. *Foundations and Trends in Econometrics 3*, 89–163.

Bernanke, B., J. Boivin, and P. Eliasz (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach. *The Quarterly Journal of Economics 120*(1), 387–422.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Billingsley, P. (1999). *Convergence of Probability Measures* (second ed.). John Wiley & Sons.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics 24*(6), 2350–2383.

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.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 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.

Horn, R. and C. Johnson (1990). *Matrix Analysis*. Cambridge University Press.

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.

Kock, A. B. (2012a). Consistent and conservative model selection in stationary and non-stationary autoregressions. *Submitted*.

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

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

Nardi, Y. and A. Rinaldo (2011). Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis 102*(3), 528–549.

Stock, J. and M. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*(460), 1167–1179.

Stock, J. and M. Watson (2006). Forecasting with many predictors. In G. Elliott, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 1, pp. 515–554. Elsevier, Amsterdam.

Stock, J. and M. Watson (2011). Dynamic factor models. In M. Clements and D. Hendry (Eds.), *Oxford Handbook of Economic Forecasting*, Volume 1, pp. 35–59. Oxford University Press, Oxford.

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

Van Der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes*. Springer Verlag.

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.

Zhou, S., S. Van De Geer, and P. Bühlmann (2009). Adaptive lasso for high dimensional regression and gaussian graphical modeling. *Arxiv preprint ArXiv:0903.2515*.

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

2011-53: 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

2012-06: Christian Bach and Matt P. Dziubinski: Commodity derivatives pricing with inventory effects

2012-07: Cristina Amado and Timo Teräsvirta: Modelling Changes in the Unconditional Variance of Long Stock Return Series

2012-08: Anne Opschoor, Michel van der Wel, Dick van Dijk and Nick Taylor: On the Effects of Private Information on Volatility

2012-09: Annastiina Silvennoinen and Timo Teräsvirta: Modelling conditional correlations of asset returns: A smooth transition approach

2012-10: Peter Exterkate: Model Selection in Kernel Ridge Regression

2012-11: Torben G. Andersen, Nicola Fusari and Viktor Todorov: Parametric Inference and Dynamic State Recovery from Option Panels

2012-12: Mark Podolskij and Katrin Wasmuth: Goodness-of-fit testing for fractional diffusions

2012-13: Almut E. D. Veraart and Luitgard A. M. Veraart: Modelling electricity day-ahead prices by multivariate Lévy

2012-14: Niels Haldrup, Robinson Kruse, Timo Teräsvirta and Rasmus T. Varneskov: Unit roots, nonlinearities and structural breaks

2012-15: Matt P. Dziubinski and Stefano Grassi: Heterogeneous Computing in Economics: A Simplified Approach

2012-16: Anders Bredahl Kock and Laurent A.F. Callot: Oracle Inequalities for High Dimensional Vector Autoregressions