TEXTO PARA DISCUSSÃO

No. 636

ℓ_1- Regularization of high-dimensional time-series models with flexible innovations

Marcelo C. Medeiros
Eduardo Mendes

DEPARTAMENTO DE ECONOMIA
www.econ.puc-rio.br

ℓ_1-REGULARIZATION OF HIGH-DIMENSIONAL TIME-SERIES MODELS WITH FLEXIBLE INNOVATIONS

Marcelo C. Medeiros
Department of Economics
Pontifical Catholic University of Rio de Janeiro
Rua Marquês de São Vicente 225, Gávea
Rio de Janeiro, 22451-900, BRAZIL
E-mail: mcm@econ.puc-rio.br

Eduardo F. Mendes
Department of Economics
Australian School of Business
University of New South Wales
E-mail: eduardo.mendes@unsw.edu.au

JEL: C22.

Keywords: sparse models, shrinkage, LASSO, adaLASSO, time series, forecasting, GARCH.

Acknowledgements: The authors would like to thank Anders Kock, Laurent Callot, Marcelo Fernandes, Marcelo J. Moreira, Emmanuel Guerre, Wenxin Jiang, Martin Tanner, Eric Hillebrand, Asger Lunde, Francesco Audrino, Simon Knaus, and Thiago Ferreira for insightful discussions. M. C. Medeiros acknowledges support from CREATES funded by the Danish National Research Foundation and partial support from CNPq/Brazil. E. F. Mendes acknowledges partial support from ARC grant DP0988579. Part of this work was carried out while the first author was visiting CREATES at the University of Aarhus. Its kind hospitality if greatly acknowledged. We are indebt to Gabriel Vasconcelos for superb research assistance.

Abstract: We study the asymptotic properties of the Adaptive LASSO (adaLASSO) in sparse, high-dimensional, linear time-series models. We assume that both the number of covariates in the model and the number of candidate variables can increase with the sample size (polynomially or geometrically). In other words, we let the number of candidate variables to be larger than the number of observations. We show the adaLASSO consistently chooses the relevant variables as the number of observations increases (model selection consistency) and has the oracle property, even when the errors are non-Gaussian and conditionally heteroskedastic. This allows the adaLASSO to be applied to a myriad of applications in empirical finance and macroeconomics. A simulation study shows that the method performs well in very general settings with t-distributed and heteroskedastic errors as well with highly correlated regressors. Finally, we consider an application to forecast monthly US inflation with many predictors. The model estimated by the adaLASSO delivers superior forecasts than traditional benchmark competitors such as autoregressive and factor models.

1. Introduction

We consider the problem of estimating single-equation linear dynamic time-series models with non-Gaussian and conditionally heteroskedastic errors when the number of regressors is larger than the sample size (high-dimensionality), but only some of the explanatory variables are relevant (sparsity). We focus on the ℓ_1-penalized least squares estimator and derive conditions under which the method is model selection consistent and has the oracle property. By model selection consistency we mean that the correct set of regressors are selected asymptotically. The oracle property means that the penalized estimator has the same asymptotic distribution as the ordinary least squares (OLS) estimator under the knowledge of the relevant subset of regressors (Fan and Li 2001). Since our results are asymptotic, the high-dimension is understood as a polynomial increase in the number of candidate variables. Finally, we also study the case where the number of candidates variables increases exponentially with the sample size. In the latter case, stricter conditions on the error term as well as on the regressors should be imposed. However, in most economic applications the polynomial rate of growth does not seem to be restrictive. For example, when the candidate variables are lags of a fixed set of covariates, the increase is linear with respect to the sample size. Furthermore, even when other explanatory variables apart from lags are included, the number of regressors does not grow exponentially fast (Stock and Watson 2002b, Bernanke et al. 2005).

Traditionally, one chooses the set of explanatory variables using an information criterion or some sequential testing procedure. Although these approaches work well in small dimensions, the total number of models to evaluate gets exponentially large as the number of candidate variables increases. Moreover, if the number of covariates is larger than the number of observations, sequential testing fails to recover the true model structure.

A successful approach to estimate models in large dimensions is to use *shrinkage* methods. The idea is to *shrink to zero* the irrelevant parameters. Therefore, under some conditions, it is possible to handle more variables than observations. Among shrinkage methods, the Least Absolute Shrinkage and Selection Operator (LASSO), introduced by Tibshirani (1996), and the adaptive LASSO (adaLASSO), proposed by Zou (2006), have received particular attention. It has been shown that the LASSO can handle more variables than observations and the most parsimonious subset of relevant variables can be selected (Efron et al. 2004, Zhao and Yu 2006, Meinshausen and Yu 2009). As noted in Zhao and Yu (2006) and Zou (2006), for attaining model selection

consistency, the LASSO requires a rather strong condition denoted "Irrepresentable Condition" and does not have the oracle property. Zou (2006) proposes the adaLASSO to amend these deficiencies. The adaLASSO is a two-step methodology which, broadly speaking, uses a first-step estimator to weight the relative importance of the regressors. In the original framework, the number of candidate variables was smaller than the sample size, the number of relevant covariates was fixed, and the results were derived for a fixed design regression with independent and identically distributed (IID) errors. Huang et al. (2008) extended these results to a high-dimensional framework with IID errors. Recently, Fan et al. (2012) proposed a robust version of shrinkage estimators in order to deal with heavy tailed data. They also show that the adaLASSO is a one-step implementation of folded concave penalized least-squares.

We demonstrate that the adaLASSO can be applied to time-series models in a framework much more general than the one currently available in the literature. First, we allow the errors to be non-Gaussian and conditionally heteroskedastic, which is of great importance when financial or macroeconomic data are considered. Second, the number of variables (candidate and relevant ones) increase with the number of observations, which means that the dimension of the model may increase as we gather information, e.g., the number of lags in an autoregressive process (Nardi and Rinaldo 2011). Finally, the number of candidate covariates can grow at a polynomial rate with the sample size. Under flexible conditions, we show that the adaLASSO is model selection consistent (asymptotically chooses the most parsimonious model) and enjoys the oracle property. These findings allow the adaLASSO to be applied in a general time-series setup. Geometric increase in the number of candidate variables is also achieved under stronger conditions on the errors and data generating process of the covariates. Although very complicated at first sight, our conditions for model selection consistency and oracle results can be simplified dramatically as long as the error structure becomes more restrictive. Finally, in a recent paper Audrino and Camponovo (2013) proved very useful asymptotic results for the estimates of relevant and non-relevant variables in a very general time-series setting, which allows for much more general hypothesis tests. However, they consider only the case where there are less variables than observations and the number of included variables is fixed. Therefore, our results are more general and nests previous findings in the literature.

Our theoretical results are illustrated in a simulation experiment as well as in an economic application. In the simulation experiment we consider a model with fat-tailed GARCH errors and highly correlated candidate regressors. The outcome of the simulations is quite promising, pointing that the adaLASSO with properly chosen initial weights (first step) works reasonably well even in very adverse situations which are common in macroeconomics and finance. We also consider quarterly US inflation forecasting using many predictors. The models estimated by the adaLASSO procedure delivered forecasts significantly superior than traditional benchmarks.

Our results render a number of possible applications. Forecasting macroeconomic variables with many predictors as in Stock and Watson (2002a,b, 2012) and Bai and Ng (2008) is one of them. The construction of predictive regressions for financial returns can be also considered (Rapach et al. 2010). In this case, handling non-Gaussian conditional heteroskedastic errors is of great importance. Other applications include the selection of factors in approximate factor models, as in Bai and Ng (2002), Cheng and Hansen (2012), and Cheng et al. (2013); variable selection in non-linear models (Rech et al. 2001); forecast combination of many forecasters (Issler and Lima 2009, Samuels and Sekkel 2013); time-series network models (Barigozzi and Brownlees 2013, Lam and Souza 2014a,b); and forecasting large covariance matrices as in Callot et al. (2014). Finally, instrumental variable estimation in a data rich environment with dependent data is also a potential application; see Belloni et al. (2012).

Most advances in the shrinkage methods literature are valid only in the classical IID framework, often with fixed design. Recently, a large effort has been given to adapt LASSO-based methods to the time-series case; see, for example, Wang et al. (2007a) and Hsu et al. (2008). These authors consider only the case where the number of candidate variables is smaller than the sample size. Nardi and Rinaldo (2011) considered the estimation of autoregressive (AR) models when the number of regressors increases with the sample size. However, their work differs from ours in many directions. The most significant one being that their focus is only on AR models with restrictive assumptions on the error term. Audrino and Knaus (2012) adapted the results of Nardi and Rinaldo (2011) to the case of realized volatility forecasting with the heterogenous AR (HAR) model proposed by Corsi (2009). Our results are useful in this setting as realized volatility data are conditionally heteroskedastic and non-Gaussian. Furthermore, our results allow for the inclusion of external variables as potential predictors. Wang et al. (2007b) considered regression models with autoregressive

errors. Notwithstanding, in their case the number of regressors was kept fixed. Song and Bickel (2011) and Kock and Callot (2012) studied the estimation of vector AR (VAR) models. The former used LASSO and group-LASSO for estimating VARs where the number of candidate variables were a function of the sample size. However, the number of relevant variables was fixed. Kock and Callot (2012) relaxed this assumption but assumed the errors to be independent and normally distributed. As a direct consequence of the VAR dynamics, in Kock and Callot (2012) all the covariates were Gaussian. Barigozzi and Brownlees (2013) also assumed normality and homoskedasticity of the errors. Although, our model is nested in the VAR specification, we show the oracle property under a more general setting as the above authors do not consider the inclusion of exogenous regressors. On the other hand, Kock and Callot (2012) derive non-asymptotic oracle inequalities which are not discussed here. All our results are asymptotic. Kock (2012) considered adaLASSO estimation in stationary and non-stationary AR models with a fixed number of variables.

It is important to make the following remarks. First, the adaLASSO is a two-step procedure and there is no agreement in the literature how to choose the first-step estimator. In this paper we use the LASSO as a possible solution (Zou and Hastie 2005)[1]. We show that, under regularity conditions, the LASSO can be used as an initial estimate, at a cost of possibly reducing the pool of candidate variables. Our simulation results indicate that the LASSO works quite well. Second, all the hyper-parameters in the estimation procedure (such as the penalty term) are selected via the Bayesian Information Criterion (BIC) which delivers superior results, both in terms of accuracy and computing time, than cross-validation methods. Finally, similar to other papers in the literature, all our asymptotic results are derived under pointwise convergence as shrinkage estimators suffer from lack of uniformity; see, for example, Leeb and Pötscher(2008, 2009).

The paper is organized as follows. In Section 2 we introduce the notation and assumptions. In Section 3 we present the main results. The case where the number of candidate variables grows exponentially with the sample size is discussed n Section 4. In Section 5 we discuss the selection of the weights for the adaLASSO procedure and in Section 7 we describe how our set of assumptions can be satisfied in some special cases. In Section 8 we present simulation results, followed by the real data application in Section 9. Finally, Section 10 concludes. All the proofs are postponed to the appendix. In the Appendix we also discuss how to satisfy the main assumptions of the paper.

[1]Other pre-estimators have been considered (Elastic-net, Ridge, OLS) but the LASSO delivered robust results.

2. Definition, Notation and Assumptions

Consider the following linear model

$$y_t = \alpha_0 + \boldsymbol{\theta}' \boldsymbol{x}_t + u_t, \tag{1}$$

where $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{n_T t})'$ is a n_T-vector of covariates, possibly containing lags of y_t, and u_t is a martingale difference process. We are interested in estimating the parameter vector $\boldsymbol{\theta}$ when n_T is large, possibly larger than the sample size T, but only a small number of elements of $\boldsymbol{\theta}$ is non-zero ($\boldsymbol{\theta}$ is sparse). We assume, without loss of generality, that α_0 is zero. Model (1) encompasses many linear specifications, such as sparse AR and AR distributed lag (ARDL) models, or simple predictive regressions. Equation (1) may also be a reduced-form for first-stage estimation in a two-stage least squares environment where $\boldsymbol{x}_t$ includes a set of instruments and y_t is an endogenous variable. Another possibility is to consider $\boldsymbol{x}_t$ as a set of individual forecasts, in which equation (1) represents a forecast combination problem.

The number of candidate covariates is $n \equiv n_T$, the number of non-zero parameters is $s \equiv s_T$ and the number of irrelevant variables is $n - s$. The omission of the dependence on T is just aesthetic. For any t, $\boldsymbol{x}_t = [\boldsymbol{x}_t(1)', \boldsymbol{x}_t(2)']'$ and $\boldsymbol{X} = [\boldsymbol{X}(1), \boldsymbol{X}(2)]$, where $\boldsymbol{X}(1)$ is the $(T \times s)$ matrix with the relevant variables and $\boldsymbol{X}(2)$ is the $[T \times (n-s)]$ matrix with the irrelevant ones. Write $\boldsymbol{\theta} = [\boldsymbol{\theta}(1)', \boldsymbol{\theta}(2)']'$ where $\boldsymbol{\theta}(1) \in \mathbb{R}^s$ and $\boldsymbol{\theta}(2) \in \mathbb{R}^{n-s}$. $\boldsymbol{\theta}_0$ is the *true* parameter vector, where $\boldsymbol{\theta}_0 = [\boldsymbol{\theta}_0(1)', \mathbf{0}']'$, with $\boldsymbol{\theta}_0(1) \neq \mathbf{0}$. The parameters are assumed ordered to simplify the exposition.

We make the following assumption about the processes $\{\boldsymbol{x}_t\}$, $\{y_t\}$, and $\{u_t\}$:

Assumption (DGP)**.** *Write* $\boldsymbol{z}_t = (y_t, \boldsymbol{x}_t', u_t)'$.

(1) $\{\boldsymbol{z}_t\}$ *is a zero-mean weakly stationary process.*

(2) $\mathbb{E}(u_t|\mathcal{F}_t) = 0$, $t = 1, 2, \ldots,$ *where* $\mathcal{F}_t = \sigma\{\boldsymbol{x}_t, \boldsymbol{z}_{t-1}, \boldsymbol{z}_{t-2}, \ldots\}$.

(3) With probability converging to one,

$$\max_{1 \leq i \leq n} T^{-1} \sum_{t=1}^{T} \left[x_{it}^2 - \mathbb{E}(x_{it}^2)\right] \to 0, \quad T \to \infty.$$

(4) For some finite, positive constant c_m *and some* $m \geq 1$,

$$\mathbb{E}\left|x_{it} u_t\right|^m \leq c_m, \, \forall \, i \in \{1, \ldots, n\} \text{ and } \forall \, t.$$

Assumptions DGP(1) and DGP(2) are standard in time series regressions. Note that DGP(2) does not rule out conditional heteroskedasticity, such as GARCH effects. Furthermore, $\boldsymbol{x}_t$ may also contain lagged values of y_t. Assumption DGP(3) defines a tail condition on the marginal distributions of $x_{1t}, \ldots, x_{nt}$, and DGP(4) is a moment condition on the process $\{u_t x_{jt}\}$. The possible number of candidate variables n depends both on DGP(3) and DGP(4), i.e., on the tail assumptions on $\boldsymbol{x}_t$ and u_t.

Remark 1. *DGP(3) is a condition on the concentration properties of the variances of the covariates (recall that $\mathbb{E}(x_{it}) = 0$), and has a substantial effect on the number of candidate variables. It is well understood in the literature that tail conditions on the error process $\{u_t\}$ are determinant of the number of candidate variables. What is less understood is how the total number of candidate variables depends on tail conditions and "memory" properties of the regressors. In Appendix A.1 we show that DGP(3) is satisfied under different sets of assumptions and how it can influence the possible number of candidate variables.*

Remark 2. *Sufficient conditions for DGP(4) to be satisfied are easily derived. Let $c_{m,1}$ and $c_{m,2}$ be two positive and finite constants. Assume that $\mathbb{E}\left(u_t^{2m}|\mathcal{F}_t\right) \leq c_{m,1}$, $\forall\, t$ and $\mathbb{E}\left(x_{it}^{2m}\right) \leq c_{m,2}$, $\forall\, t$ and $\forall\, i \in \{1, \ldots, n\}$. Therefore, by the law of iterated expectations DGP(4) is satisfied with $c_m = c_{m,1}c_{m,2}$. Alternatively, assume there exist $\rho > 0$ and $\gamma > 1/\rho$ such that $\mathbb{E}\left[u_t^{2m(1+\rho)}\right] \leq c_{m,2}$, $\forall\, t$ and $\mathbb{E}\left[x_{it}^{2m(1+\gamma)}\right] \leq c_{m,1}$ $\forall\, t$ and $\forall\, i \in \{1, \ldots, n\}$, then DGP(4) follows after simple application of the Hölder's inequality with $c_m = c_{m,1}^{1/(1+\rho)} c_{m,2}^{1/(1+\gamma)}$.*

Assumption (DESIGN)**.** *The following conditions hold jointly.*

(1) The true parameter vector $\boldsymbol{\theta}_0$ is an element of an open subset $\boldsymbol{\Theta}_n \in \mathbb{R}^n$ that contains the element $\mathbf{0}$.

(2) There exists $\theta_{\min} > 0$ such that $\min_{i=1,\ldots,s} |\theta_{0,i}| > \theta_{\min}$.

(3) a. Write $\boldsymbol{\Omega}_{11} = \mathbb{E}[\boldsymbol{x}_t(1)\boldsymbol{x}_t(1)']$. There exist constants $0 < \phi_{\min} < 1$ such that

$$\inf_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1} \boldsymbol{\alpha}'\boldsymbol{\Omega}_{11}\boldsymbol{\alpha} > 2\phi_{\min}.$$

b. Let $\widehat{\boldsymbol{\Omega}}_{11} = \boldsymbol{X}(1)'\boldsymbol{X}(1)/T$ denote the scaled Gram matrix of the relevant variables,

$$\max_{1\leq i,j\leq s} \left[\left|\widehat{\boldsymbol{\Omega}}_{11} - \boldsymbol{\Omega}_{11}\right|\right]_{i,j} \leq \frac{\phi_{\min}}{s},$$

with probability converging to one as $T \to \infty$

Assumption DESIGN(1) is standard. DESIGN(2) controls the lower bound of the non-zero parameters and is traditionally referred as *beta-min condition*; see, for example, Bülhmann and van der Geer (2011). We define lower bounds on $\theta_{\min}$ on Theorem 1. This lower bound can decrease with T and lower bounds on $\phi_{\min}$. DESIGN(3) imposes a lower bound, $\phi_{\min}$, on the minimal eigenvalue of the covariance matrix of the relevant variables, that may depend on T. In practice, quantifying the rate in which $\phi_{\min}$ decreases is difficult and problem specific and it is frequently assumed constant, e.g., Theorems 3 and 4 in Kock and Callot (2012) assume $\phi_{\min} > c > 0$ in a VAR(p) model with Gaussian innovations.

Condition DESIGN(3) explicitly defines how the *compatibility constant* depends on the number of variables in the true active set. In general, this dependence is implicit and appears in the oracle bounds. DESIGN(3) part (a) is related to the *restricted eigenvalue condition* (Bickel et al. 2009). If the restricted eigenvalue condition is satisfied for any constant $L > 0$ and $S = \{1, \ldots, s\}$ with compatibility constant $\phi(L, S, s) \geq 2\phi_{\min}$, then Lemma 6.25 in Bülhmann and van der Geer (2011) implies that DESIGN(3) part (a) is also satisfied. Condition DESIGN(3) part (b) can be satisfied by imposing conditions on the dependence and tail structures of the variables in the active set. In Appendix A.2 we show sufficient conditions for satisfying DESIGN(3) part (b).

The adaLASSO estimator of the $(n \times 1)$ parameter vector $\boldsymbol{\theta}$ is given by

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|Y - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda \sum_{i=1}^{n} w_i |\theta_i|, \tag{2}$$

where $\boldsymbol{Y} = (y_1, \ldots, y_T)'$, $\boldsymbol{X}$ is the $(T \times n)$ data matrix, $w_i = |\theta_{I,i}|^{-\tau}$, $\tau > 0$, and $\theta_{I,i}$ is an initial parameter estimate. When $w_i = 1$ $(i = 1, \ldots, n)$, (2) becomes the usual LASSO.

The minimization problem in (2) is equivalent to a constrained concave minimization problem and necessary and (almost) sufficient conditions for existence of a solution can be derived from the Karush-Kuhn-Tucker conditions (Zhao and Yu 2006, Zou 2006). The necessary condition for the model selection consistency for the LASSO ($w_i = 1$, $i = 1, \ldots, n$) is denoted the "Irrepresentable Condition" which is known to be easily violated in the presence of highly correlated covariates (Zhao and Yu 2006, Meinshausen and Yu 2009). The adaLASSO overcomes the "Irrepresentable Condition", by using weighted ℓ_1-penalty where the weights diverge for the zero parameters and do not diverge for the non-zero parameters. Zou (2006) suggest using the inverse of the OLS estimator

of the parameters as the weight. Nonetheless, such estimator is not available when the number of candidate variables is larger than the number of observations. Huang et al. (2008) introduce the notion of *zero-consistent* estimator, i.e., there exists an estimator that is arbitrarily small for the zero parameters as T increases, and converge to a non-zero constant for the non-zero parameters. We use a similar assumption here.

Assumption (WEIGHTS)**.** *The weights $w_1, \ldots, w_n$ satisfy:*

(1) There exist $0 < \xi < 1$, and a sufficiently large, positive constant $c_{w(2)}$, such that

$$\min_{i=s+1,\ldots,n} T^{-\xi/2} w_i > c_{w(2)} \sqrt{\frac{s}{\phi}},$$

with probability converging to one as $T \to \infty$.

(2) There exists $w_{\max} < T^{\xi/2}$ such that

$$\sum_{i=1}^{s} w_i^2 < s w_{\max}^2,$$

with probability converging to one as $T \to \infty$.

Assumption WEIGHTS(1) requires that the weights associated with the non-relevant variables $\{x_{jt} : j = s+1, \ldots, n\}$ to diverge at some rate, while WEIGHTS(2) restricts the weights associated with the relevant variables to be bounded by above by a non-decreasing sequence $w_{\max}$. This requirement is the most difficult to be satisfied in practice. In the case when the number of candidate variables n is smaller than the number of observations T, we can estimate the weights using OLS.

Huang et al. (2008) show that if the variables with zero and non-zero coefficients are only weakly correlated (partial orthogonality condition), the marginal regressions of y_t on x_{it}, $i = 1, \ldots, n$, give reasonable weights. This condition, however, is not realistig in a time series setting, in which lags of the dependent and the independent variables are in the pool of candidate variables. If the correlation matrix of regressors is Toeplitz, than the "Irrepresentable Condition" is valid and LASSO may perform reasonably well (Nardi and Rinaldo 2011, Audrino and Knaus 2012)[2].

[2]Under week regularity conditions, the "Irrepresentable Condition" yield oracle bounds (Van De Geer and Bühlmann 2009, Section 6).

Assumption REG imposes constraints on the rate of increase of number of candidate variables in terms of λ. These bounds involve m, $\phi_{\min}$, ξ, and $w_{\max}$, defined in Assumptions DGP(4), DESIGN(3), WEIGHTS(1) and WEIGHTS(2), respectively.

Assumption (REG)**.** *The regularization parameter λ and the number of candidate variables n satisfy:*

$$\frac{n^{1/m}T^{(1-\xi)/2}}{\lambda} \to 0 \quad \textit{and} \quad \frac{s^{1/2}w_{\max}}{\phi_{\min}}\frac{\lambda}{\sqrt{T}} \to 0,$$

as $T \to \infty$.

This assumption is satisfied if we take $\lambda \propto \sqrt{T} \times n^{1/m}T^{-\xi(1/2-1/m)}$, assume that $s^{1/2}w_{\max}/\phi_{\min} = O(n^{b/m})$ for some $b > 0$, and impose $n = o\left[T^{\xi(m-2)/2(b+1)}\right]$. If we further assume an oracle bound of the form of Proposition 1 in Section 5, we may take $\xi = \alpha m(b+1)/(m+2b)$, for any $0 < \alpha < 1 - 2\log_T(s/\phi_{\min})$. Combining the bounds, $n = o\left[T^{\alpha m(m-2)/(2m+4b)}\right]$. Improving these rates is possible, but have no impact on the main results of the paper.

The imposition on the number of candidate variables to be polynomial on T is a consequence of $|u_t x_{it}|$ having polynomially decreasing tails. When stronger bounds are imposed on x_{it} and u_t, it is possible to allow the number of candidate variables to grow at a faster rate. This condition only imposes an upper-bound on the rate of increase of candidate variables, which is further retracted by DGP(3) and DESIGN(3).

3. Main Results

In this section we present the main results of the paper: model selection consistency and oracle property. We follow the standard practice in the literature and show *sign consistency*, which implies model selection consistency.

Definition (Sign Consistency)**.** *We say that $\widehat{\boldsymbol{\theta}}$ is sign consistent to $\boldsymbol{\theta}$ if*

$$\Pr\left[\text{sign}(\widehat{\boldsymbol{\theta}}) = \text{sign}(\boldsymbol{\theta})\right] \to 1, \text{ element-wise as } T \to \infty,$$

where $\text{sign}(x) = I(x > 0) - I(x < 0)$*, and the identity is taken element-wise.*

Next theorem is the main result in the paper and shows that, under the previous assumptions, the adaLASSO consistently selects the correct subset of variables.

Theorem 1. *Under Assumptions DGP, DESIGN, WEIGHTS and REG, and If*

$$\theta_{\min} > \frac{\lambda}{T^{1-\xi/2}} \frac{s^{1/2}}{\phi_{\min}},$$

then

$$\Pr\left[\text{sign}(\widehat{\boldsymbol{\theta}}) = \text{sign}(\boldsymbol{\theta}_0)\right] \to 1, \text{ as } T \to \infty.$$

In Theorem 2 we show that the adaLASSO estimator for time-series has the oracle property, in the sense that it converges to the same distribution as the OLS estimator as $T \to \infty$. The relevance of this result is that one can carry out inference about the parameters as if one had used OLS in the model with only the relevant variables included.

Theorem 2 (Oracle Property)**.** *Let $\widehat{\boldsymbol{\theta}}_{ols}(1)$ denote the OLS estimator of $\boldsymbol{\theta}_0(1)$. Under Assumptions DGP, WEIGHTS, DESIGN, and REG, if $\theta_{\min} > (\lambda/T^{1-\xi/2})(s^{1/2}/\phi_{\min})$ we have*

$$\sqrt{T}\boldsymbol{\alpha}'\left[\widehat{\boldsymbol{\theta}}(1) - \boldsymbol{\theta}_0(1)\right] = \sqrt{T}\boldsymbol{\alpha}'\left[\widehat{\boldsymbol{\theta}}_{ols}(1) - \boldsymbol{\theta}_0(1)\right] + o_p(1).$$

for any s-dimensional vector $\boldsymbol{\alpha}$ with Euclidean norm 1.

4. Exponentially large number of covariates

Conditions in the previous section imply that the number of candidate variables n may increase at a polynomial rate. Under stronger assumptions, n may increase sub-exponentially fast with T. Note that the actual rate of increase also depends on the distribution of the candidate variables themselves. In this section we introduce new assumptions and restate the main results.

Assumption (DGP(5))**.** *The processes $\{x_{it}\}$, $i = 1, \ldots, n$, and $\{u_t\}$ are such that*

$$\Pr(|x_{it}| > c) \leq b_{1i}\exp(-b_{2i}c) \quad \textit{and} \quad \Pr(|u_t| > c) \leq b_3\exp(-b_4 c),$$

for all $i = 1, \ldots, n$ and every t, and for positive constants c, b_{1i}, b_{2i}, b_3, and b_4.

Assumption DGP(5) requires that the marginal distribution of the candidate variables and error term have exponential tails, which is more general then the IID Gaussian innovations. It is satisfied when the dynamics of $\boldsymbol{x}_t$ is driven by stationary vector autoregressions (VAR) with Gaussian

innovations as in Kock and Callot (2012). Alternatively, if $\boldsymbol{x}_t$ admits an infinite-order vector moving average, VMA(∞), decomposition with bounded conditional variances, Lemma 9 in the appendix shows conditions under which it has sub-exponential tails. Same arguments hold for u_t.

Assumption REG incorporates the new rate of increase in the number of irrelevant covariates. The biggest change is that it allows n to increase sub-exponentially with T, instead of polynomially.

Assumption (REG'). *The regularization parameter λ and the number of candidate variables n satisfy:*

$$\frac{(\log n + \alpha \log T)^{5/2} T^{(1-\xi)/2}}{\lambda} \to 0 \quad \textit{and} \quad \frac{s^{1/2} w_{\max}}{\phi_{\min}} \frac{\lambda}{\sqrt{T}} \to 0,$$

as $T \to \infty$, for some $\alpha > 0$.

The term $\alpha \log T$ simplifies the calculation of finite sample bounds and can be dropped if $\log n > \alpha \log T$, which is often the case for T sufficiently large. The assumption is satisfied if we take $\lambda = \log T (\log n + \xi \log T)^{5/2} T^{(1-\xi)/2}$, $s^{1/2} w_{\max}/\phi_{\min} = O\left[(\log n + \xi \log T)^{b/2}\right]$ for some $b > 0$, and impose $\log n = o\left[(T/\log T)^{2\xi/(b+5)}\right]$. If we further assume that an oracle bound of the form of Proposition 1 holds than, if $(s/\phi_{\min}) = O\left\{(\log n)^{[b+10(1-\xi)]/6\xi}\right\}$, the rate in which n increases remain unchanged. As in the previous case, improving the bounds has no impact on the main results of the paper.

This rate of increase in the total number of candidate variables is only an upper bound. The total number of variables is further constrained by DGP(3) and DESIGN(3), that depends on the distribution of the covariates.

Theorem 3. *Under Assumptions DGP(1-3), DGP(5), WEIGHTS, DESIGN and REG', if $\theta_{\min} > (\lambda/T^{1-\xi/2})(s^{1/2}/\phi_{\min})$,*

$$P\left[\text{sign}(\widehat{\boldsymbol{\theta}}) = \text{sign}(\boldsymbol{\theta}_0)\right] \to 1, \ \textit{as } T \to \infty.$$

Furthermore,

$$\sqrt{T}\boldsymbol{\alpha}' \left[\widehat{\boldsymbol{\theta}}(1) - \boldsymbol{\theta}_0(1)\right] = \sqrt{T}\boldsymbol{\alpha}' \left[\widehat{\boldsymbol{\theta}}_{ols}(1) - \boldsymbol{\theta}_0(1)\right] + o_p(1).$$

for some s-dimensional vector $\boldsymbol{\alpha}$ with Euclidean norm 1.

5. Initial weights

The choice of initial weights is critical and, often, the hardest condition to be satisfied. In this section we show that under a stronger set of conditions, one can use the LASSO as the initial

estimator to construct the weights. Furthermore, sufficient conditions for the consistency of the LASSO estimator also imply DESIGN(3). In this section we relate oracle bounds on the ℓ_1 norm of the LASSO estimates to condition WEIGHTS.

Oracle inequalities for the LASSO estimator have been derived under different assumptions on the design matrix. Van De Geer and Bühlmann (2009) study how these different assumptions relate to each other, in particular, they show that the *restricted eigenvalue condition* of Bickel et al. (2009) imply the *compatibility condition*, used for deriving oracle bounds. If the scaled Gramm matrix $\widehat{\boldsymbol{\Omega}} = \boldsymbol{X}'\boldsymbol{X}/T$, is sufficiently close to its expectation, then ℓ_1 oracle bounds follow after conditions on the smallest eigenvalue of $\boldsymbol{\Omega} = \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')$.

For any vector $\boldsymbol{v} = (v_1, \ldots, v_n)' \in \mathbb{R}^n$ and $S \subseteq \{1, \ldots, n\}$, $\boldsymbol{v}_S = (v_i,\, i \in S)'$, $\boldsymbol{v}_{S^c} = (v_i,\, i \notin S)'$, and $\|\boldsymbol{v}_S\|_1 = \sum_{i\in S} |v_i|$. We say the *restricted eigenvalue condition* is satisfied for some $1 \leq s \leq n$ if

$$\phi_T(s) = \min_{S\subseteq\{1,\ldots,n\},|S|\leq s} \; \min_{\boldsymbol{v}\in\mathbb{R}^n\setminus\{\boldsymbol{0}\},\|\boldsymbol{v}_{S^c}\|_1\leq 3\|\boldsymbol{v}_S\|_1} \frac{\boldsymbol{v}'\widehat{\boldsymbol{\Omega}}\boldsymbol{v}}{\boldsymbol{v}_S{}'\boldsymbol{v}_S} > 0.$$

First, verify that if $\widehat{\boldsymbol{\Omega}}$ is positive definite, than the restricted eigenvalue condition is satisfied. Alternatively, it suffices to impose conditions on the population covariance matrix $\boldsymbol{\Omega}$ and approximation rate between $\widehat{\boldsymbol{\Omega}}$ and $\boldsymbol{\Omega}$.

Lemma 1. *Assume that*

$$\phi_0 = \frac{1}{2} \min_{S\subseteq\{1,\ldots,n\},|S|\leq s} \; \min_{\boldsymbol{v}\in\mathbb{R}^n\setminus\{\boldsymbol{0}\},\|\boldsymbol{v}_{S^c}\|_1\leq 3\|\boldsymbol{v}_S\|_1} \frac{\boldsymbol{v}'\boldsymbol{\Omega}\boldsymbol{v}}{\boldsymbol{v}_S{}'\boldsymbol{v}_S} > 0,$$

and that,

$$\max_{ij} |[\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}]_{ij}| < \frac{\phi_0}{16s},$$

with probability converging to one as $T \to \infty$. Then, the restricted eigenvalue condition is satisfied with $\phi_T(s) = \phi_0$, and DESIGN(3) is satisfied with $\phi_{\min} = \phi_0/16$.

Next result relates the restricted eigenvalue condition to the ℓ_1 bounds on the estimated parameters using the LASSO.

Lemma 2. *Denote $\mathcal{E}_T(\lambda_0) = \left\{2\max_{i=1,\ldots,n} T^{-1/2}\left|\sum_{t=1}^T x_{it}u_t\right| < \lambda_0\right\}$, and assume that the restricted eigenvalue condition holds with probability converging to one. Then, inside $\mathcal{E}_T(\lambda_0)$,*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 \leq 4\frac{\lambda}{T}\frac{s}{\phi_T(s)},$$

for any $\lambda > 2\sqrt{T}\lambda_0$*, with probability converging to one.*

Furthermore, for $\alpha > 0$*, either assume:*

(a) DGP(1), DGP(2), DGP(4), and let $\lambda \geq n^{1/m}T^{(1-\xi)/2+\alpha/m}$*, or*

(b) DGP(1), DGP(2), DGP(5), and let $\lambda \geq c'(\log n_\alpha \log T)^{5/2}T^{1-\xi)/2}$ *for* $c' > 0$ *sufficiently large,*

then, $\Pr\{\mathcal{E}_T[\lambda T^{(1-\xi)/2}]\} \geq 1 - c_1 T^{-\alpha}$*, for some* $c_1 > 0$*.*

Assumption WEIGHTS is intimately related to finite sample oracle inequalities for the LASSO. Kock and Callot (2012) consider the LASSO as the initial estimator and derive finite sample, oracle inequalities, bounding the ℓ_1 distance between the true and estimated parameters, which is directly applicable to our problem if we impose more restrictive assumptions. Proposition 1 shows the relationship between ℓ_1 oracle inequalities and assumption WEIGHTS.

Proposition 1. *Let* $\widehat{\boldsymbol{\theta}}_I = (\hat{\theta}_{I,1}, \ldots, \hat{\theta}_{I,n})'$ *denote an initial estimate of* $\boldsymbol{\theta}_0$ *and let the weights* $w_i = |\hat{\theta}_{I,i}|^{-\tau}$*. Assume that* $w_{\max}^{1/\tau} > 2/\theta_{\min}$ *and* $\theta_{\min} > 2c_1(\lambda/T)(s/\phi_{\min})$*. Then, if*

$$\sum_{i=1}^{n} |\hat{\theta}_{I,i} - \theta_{0,i}| \leq c_1 \frac{\lambda}{T}\frac{s}{\phi_{\min}},$$

for some $c_1 > 0$*, with probability converging to one, Assumption WEIGHTS hold whenever*

$$\lambda \leq c_2 T^{1-\xi/2\tau}\left(\frac{\phi_{\min}}{s}\right)^{1+1/2\tau},$$

for some small $c_2 \leq \min(.25, c_{w(2)}/c_1)$ *and all* T *sufficiently large.*

Note that there is no contradiction is assuming that $(2/\theta_{\min})^\tau < w_{\max} < T^{\xi/2}$ and $\theta_{\min} > 2c_1(\lambda/T)(s/\phi_{\min})$, as far as $0.5T^{-\xi/2\tau} > 2c_1(\lambda/T)(s/\phi_{\min})$, which is satisfied by the assumption on λ. Conditions on the rate of increase in λ, in REG, are not violated.

6. Selection of hyper-parameters

The selection of the regularization parameter λ and the weighting parameter τ is critical. Traditionally, one employs cross-validation and selects (λ, τ) within a grid that maximizes some predictive measure. In a time-dependent framework cross-validation is more complicated. An alternative approach that has received more attention in recent years is to choose the (λ, τ) using information criteria, such as the BIC. Zou et al. (2007), Wang et al. (2007a) and Zhang et al. (2010) study such

method. Zou et al. (2007) show that the number of effective parameters is a consistent estimator of the degrees of freedom of the model. Wang et al. (2007a) show that this method works in the AR-LASSO framework. Finally, Zhang et al. (2010) study a more general criterion (Generalized Information Criterion) and show that the BIC is consistent in selecting the regularization parameter, but not asymptotically loss-efficient. We adopt the BIC to select all the hyper-parameters of the adaLASSO procedure. Although we do not derive theoretical results for consistency of such methods, we conjecture that the same properties derived in Zhang et al. (2010) should hold in our framework. Furthermore, the method performs well in Monte Carlo simulations presented in the next section.

7. Examples

7.1. **Regression with exogenous variables and GARCH errors.** In this section we check under which conditions the requires assumptions hold for a linear regression model with weakly exogenous regressors and GARCH errors defined as

$$\begin{aligned} y_t &= \boldsymbol{\theta}' \boldsymbol{x}_t + u_t \\ u_t &= h_t^{1/2} \epsilon_t \\ h_t &= \pi_0 + \pi_1 h_{t-1} + \pi_2 u_{t-1}^2. \end{aligned} \tag{3}$$

where $\mathbb{E}(u_t|\boldsymbol{x}_t) = 0$ and $\{\epsilon_t\} \sim \mathsf{iid}(0,1)$, with $\mathbb{E}(\varepsilon_t^{2m}) < \infty$.

Furthermore, consider the following set of assumptions.

Assumption (EXAMPLE 1)**.** *The GARCH process is such that:*

(1) The parameters of the GARCH model satisfy the restrictions: $\pi_0 > 0$, $\pi_1 \geq 0$, *and* $\pi_2 \geq 0$*; and* $\mathbb{E}\left[(\pi_1 + \pi_2 \epsilon_{t-1}^2)^m\right] < \infty$.

(2) $\boldsymbol{x}_t \in \mathbb{R}^n$ *is a stable and invertible, finite-order, vector ARMA (VARMA) process*

$$\boldsymbol{A}(L)\boldsymbol{x}_t = \boldsymbol{M}(L)\boldsymbol{v}_t,$$

such that:

(a) The process $\{\boldsymbol{v}_t\}$ *is a martingale difference sequence, where* $\mathbb{E}(\boldsymbol{v}_t \boldsymbol{v}_t'|\mathcal{F}_{v,t-1}) = \boldsymbol{\Sigma}$ *and* $\mathcal{F}_{v,t} = \sigma\{\boldsymbol{v}_{t-1}, \boldsymbol{v}_{t-2}, \dots\}$, *and* $\mathbb{E}(v_{jt}^{2m}) < \infty$.

(b) The matrix operators $\boldsymbol{M}(z)$ *and* $\boldsymbol{A}(z)$ *are left co-prime. Moreover,* $\det \boldsymbol{M}(z) \neq \boldsymbol{0}$ *and* $\det \boldsymbol{A}(z) \neq \boldsymbol{0}$ *for* $z \in \mathbb{C}$, $|z| \leq 1$.

(c) There exists a constant $\rho > 0$ *such that* $\rho^{-1} \leq \rho_{\min}(\boldsymbol{\Sigma}) < \rho_{\max}(\boldsymbol{\Sigma}) \leq \rho$.

Under the specification above, $\boldsymbol{x}_t$ admits a canonical VMA(∞) representation as in Appendix A (Lütkepohl 2007, Chapter 11). The coefficients of this representation converge to zero exponentially fast [3], i.e., $\log \zeta_{i,r} \propto -r^{\zeta}$, and also $\boldsymbol{x}_t$ has $2m$ moments. Finally, let $\rho_{\max}(\boldsymbol{B})$ and $\rho_{\min}(\boldsymbol{B})$ denote the minimum and maximum eigenvalues of the square matrix $\boldsymbol{B}$.

This condition implies that the eigenvalues of $\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')$ are bounded. It implies that $0 < \rho_{\min}\left(\boldsymbol{\Psi}_0\boldsymbol{\Psi}_0'\right)/\rho \leq \rho_{\min}[\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')] \leq \rho_{\max}[\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')] \leq \rho \sum_{j=0}^{\infty} \rho_{\max}(\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j') < \infty$. The last inequality follows because the operator norm of $\boldsymbol{\Psi}_j$ decreases geometrically and the first one follows because one can always construct the VMA decomposition with $\boldsymbol{\Psi}_0 = \boldsymbol{I}$. The remaining inequalities follow trivially.

Note that, if $\pi_1 + \pi_2 < 1$ and under Assumption EXAMPLE 1, Assumption DGP(1) holds. In addition, Assumption DGP(2) is trivially satisfied. Under EXAMPLE 1(1), $\mathbb{E}\left(u_t^{2m}\right) < \infty$ by the results in He and Teräsvirta (1999) and Ling and McAleer (2002). Therefore, Assumption DGP(4) is valid under Example 1. Assumption DESIGN(1) is satisfied by hypothesis as well as Assumption DESIGN(2).

If conditions of Lemma 1 are satisfied, then DGP(3) and DESIGN(3) are also satisfied. It follows from EXAMPLE 1, and the results of Appendix A.3 that setting $p = 1$, we can take $s = o(T^{\delta/2})$ and $n = o[T^{(1-\delta)(m-1)/2}]$ for some $0 < \delta < 1$. These conditions are sufficient to satisfy WEIGHTS, DGP(3), and DESIGN(3). Moreover, the LASSO can be used as the initial estimator.

7.2. **Autoregressive distributed lag models with GARCH errors.** In Medeiros and Mendes (2015) the authors consider the $ARDL(p,q) - GARCH(1,1)$

$$y_t = \sum_{i=1}^{p} \phi_{0i} y_{t-i} + \sum_{i=0}^{q} \boldsymbol{\beta}_{0i}' \boldsymbol{z}_{t-i} + \varepsilon_t = \boldsymbol{\theta}_0' \boldsymbol{x}_t + u_t, \tag{4}$$

where

$$u_t = \sqrt{h_t}\varepsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} \mathsf{N}(0,1), \quad h_t = \alpha_0 + \alpha_1 u_{t-i}^2 + \beta_1 h_{t-i}. \tag{5}$$

[3] In fact, the operator norm of $\boldsymbol{\Psi}_j$, the j^{th} MA coefficient, decreases exponentially fast.

Under the set of assumptions below, Assumptions DGP, DESIGN and WEIGHTS are satisfied, using the LASSO an initial estimator. Assuming that $\tau = 1$ and $p = O(T^{1/8})$ and $s = O(1)$, then the number of candidate variables can be $n = o[T^{(m-1)/8}]$ and $\xi < .25(m-1)/(m-2)$.

Assumption (EXAMPLE 2)**.** *The DGP is such that*

(1) The roots of the polynomial $1 - \sum_{i=1}^{p} \phi_{0i} L^i$ *are outside the unity circle.*

(2) The vector of exogenous covariates admits a VARMA decomposition $\boldsymbol{A}(L)\boldsymbol{z}_t = \boldsymbol{M}(L)\boldsymbol{v}_t$, $\boldsymbol{v}_t \in \mathbb{R}^q$*, satisfying EXAMPLE 1.*

(3) The coefficients of the GARCH model satisfy EXAMPLE 1(1).

(4) Moreover, there exists $c > 0$ *independent of* T *such that* $c^{-1} < \min_{i \in S}(\theta_{0i}) \leq \max_{i \in S} \theta_{0i} < c$*, where* $S = \{j : \theta_{0j} \neq 0\} \cap \{1, \ldots, n\}$.

(5) Let $\|\boldsymbol{B}\|$ *denote the operator norm of* $\boldsymbol{B}$ *and*

$$\mathbb{E}(\boldsymbol{x}_1 \boldsymbol{x}_1') = \begin{pmatrix} \boldsymbol{\Sigma_{YY}} & \boldsymbol{\Sigma_{YZ}} \\ \boldsymbol{\Sigma_{YZ}}' & \boldsymbol{\Sigma_{ZZ}} \end{pmatrix}.$$

a. For some $\rho > 0$, $\rho_{\min}(\boldsymbol{\Sigma_{YY}}) > \rho^{-1}$.

b. For some $0 < \nu < 1$, $\|\boldsymbol{\Sigma_{YZ}}\|^2 \leq \nu^2 \rho_{\min}(\boldsymbol{\Sigma_{YY}}) \rho_{\min}(\boldsymbol{\Sigma_{ZZ}})$.

The discussion in the first example implies that EXAMPLE 2(1–4) satisfy Assumptions (A1)–(A4) in Medeiros and Mendes (2015). We show that under EXAMPLE 2, $\boldsymbol{\Sigma_{XX}} = \mathbb{E}(\boldsymbol{x}_1 \boldsymbol{x}_1')$ is positive definite. Under (A1)–(A3) in Medeiros and Mendes (2015), $\rho_{\min}(\boldsymbol{\Sigma_{ZZ}}) > \rho^{-1}$, and, by assumption, $\boldsymbol{\Sigma_{YY}} \geq \rho^{-1}$. It follows from Kierzkowski and Smoktunowicz (2011, Corollary 2.5) that the smallest eigenvalue of $\boldsymbol{\Sigma_{XX}}$ is bounded by below by $(1-\nu)/\rho > 0$, proving the claim. EXAMPLE 2(5) can be improved using Kierzkowski and Smoktunowicz (2011, Theorem 2.9). A simpler proof is as follows. The matrix $\boldsymbol{\Sigma_{XX}}$ is positive definite if and only if $\boldsymbol{\Sigma_{YY}} - \boldsymbol{\Sigma_{YZ}} \boldsymbol{\Sigma_{ZZ}}^{-1} \boldsymbol{\Sigma_{ZY}} > 0$. The chain of inequalities follows

$$\begin{aligned} \rho_{\min}(\boldsymbol{\Sigma_{YY}} - \boldsymbol{\Sigma_{YZ}} \boldsymbol{\Sigma_{ZZ}}^{-1} \boldsymbol{\Sigma_{ZY}}) &\geq \rho_{\min}(\boldsymbol{\Sigma_{YY}}) - \rho_{\max}\left(\boldsymbol{\Sigma_{YZ}} \boldsymbol{\Sigma_{ZZ}}^{-1} \boldsymbol{\Sigma_{ZY}}\right) \\ &\geq \rho_{\min}(\boldsymbol{\Sigma_{YY}}) - \rho_{\min}(\boldsymbol{\Sigma_{ZZ}})^{-1} \|\boldsymbol{\Sigma_{YZ}}\|^2 \\ &\geq \rho_{\min}(\boldsymbol{\Sigma_{YY}}) - \rho_{\min}(\boldsymbol{\Sigma_{ZZ}})^{-1} \nu^2 \rho_{\min}(\boldsymbol{\Sigma_{ZZ}}) \rho_{\min}\left(\boldsymbol{\Sigma_{YY}}\right) \\ &\geq (1 - \nu^2)/\rho > 0. \end{aligned}$$

8. Simulation

Consider the following data generating process (DGP):

$$y_t = \phi y_{t-1} + \boldsymbol{\beta}' \boldsymbol{x}_{t-1}(1) + u_t, \tag{6}$$

$$u_t = h_t^{1/2} \varepsilon_t,\ \varepsilon_t \overset{\text{iid}}{\sim} \mathsf{t}^*(5) \tag{7}$$

$$h_t = 5 \times 10^{-4} + 0.9 h_{t-1} + 0.05 u_{t-1}^2 \tag{8}$$

$$\boldsymbol{x}_t = \begin{bmatrix} \boldsymbol{x}_t(1) \\ \boldsymbol{x}_t(2) \end{bmatrix} = \boldsymbol{A}_1 \begin{bmatrix} \boldsymbol{x}_{t-1}(1) \\ \boldsymbol{x}_{t-1}(2) \end{bmatrix} + \boldsymbol{A}_4 \begin{bmatrix} \boldsymbol{x}_{t-4}(1) \\ \boldsymbol{x}_{t-4}(2) \end{bmatrix} + \boldsymbol{v}_t,\ \boldsymbol{v}_t \overset{\text{iid}}{\sim} \mathsf{t}^*(5), \tag{9}$$

where $\phi = 0.6$ and the typical element of $\boldsymbol{\beta}$ is given by $\beta_i = \frac{1}{\sqrt{s}}(-1)^i$. $\boldsymbol{x}_t(1)$ is a $(s-1) \times 1$ vector of included (relevant) variables. The vector $\boldsymbol{x}_t = [\boldsymbol{x}_t(1)', \boldsymbol{x}_t(2)']' \in \mathbb{R}^{(n-1)}$, has $n - s$ irrelevant variables and follows a fourth-order VAR model with t-distributed errors. Apart from the error distribution, the DGP for the vector $\boldsymbol{x}_t$ is similar to the one considered in Kock and Callot (2012). The matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are block diagonal with each block of dimension 5×5 and typical element 0.15 and -0.1, respectively. All the errors in the model are t-distributed with 5 degrees of freedom. $\mathsf{t}^*(5)$ denotes an standardized t-distribution with 5 degrees of freedom, such that all the errors have zero mean and unit variance. The vector of candidate variables is $\boldsymbol{w}_t = (y_{t-1}, \boldsymbol{x}_{t-1}')'$. Furthermore, ε_t and $\boldsymbol{v}_t$ are mutually not correlated. Note that this is a very adverse setting as the errors are not normal, fat-tailed, conditionally heteroskedastic and moments of order higher than five do not exist.

We simulate $T = 50, 100, 300, 1000$ observations of DGP (6)–(9) for different combinations of candidate (n) and relevant (s) variables. We consider $n = 100, 300, 1000$ and $s = 5, 10, 15, 20$. The models are estimated by the adaLASSO method with τ and λ selected by the BIC. The initial weights are estimated using the LASSO procedure.

We start by analyzing the properties of the estimators for the parameter ϕ in (6)) Figures 1–4 illustrate the distribution of the oracle and adaLASSO estimators for different sample sizes. Several facts emerge from the plots. Firstly, both bias and variance are very low. For $T = 50$ and $s = 5$, the distribution of the adaLASSO estimator is very close to the distribution of the oracle. For the other values of s, the adaLASSO distribution presents fat-tails and multi-modality. For $T = 100$, the adaLASSO distribution is closer to the oracle one when $s = 5$ or $s = 10$. However, there still

outliers. When $T = 300$ the number of outliers reduces and the adaLASSO distribution gets closer to the oracle, specially for $s = 5$ or $s = 10$. For $T = 1000$ the distributions are almost identical.

Table 1 shows the average absolute bias and the average mean squared error (MSE) for the adaLASSO estimator over the Monte Carlo simulations and the candidate variables, i.e.,

$$\text{Bias} = \frac{1}{1000n}\sum_{j=1}^{1000}\left[\widehat{\phi} - 0.6 + \sum_{i=1}^{n-1}\left(\widehat{\beta}_i - \beta_i\right)\right] \text{ and}$$

$$\text{MSE} = \frac{1}{1000n}\sum_{j=1}^{1000}\left[\left(\widehat{\phi} - 0.6\right)^2 + \sum_{i=1}^{n-1}\left(\widehat{\beta}_i - \beta_i\right)^2\right].$$

It is clear that both variance and bias are very low. This is explained, as expected, by the large number of zero estimates. Finally, the bias and MSE decrease with the sample size. The MSE of the estimators increase with the number of candidate variables as well as with the number of relevant variables. Finally, it is quite clear that the estimates are very precise in large samples.

Table 2 presents model selection results. Panel (a) presents the fraction of replications where the correct model has been selected, i.e., all the relevant variables included and all the irrelevant regressors excluded from the final model (correct sparsity pattern). It is clear the performance of the adaLASSO improves with the sample size and gets worse as the number of relevant variables increases. Furthermore, there is a slightly deterioration as the number of candidate regressors increases. Panel (b) shows the fraction of replications where the relevant variables are all included. For $T = 300$ and $T = 1000$, the true model is included almost every time. For smaller sample sizes the performance decreases as s increases. Panel (c) presents the fraction of relevant variables included and Panel (d) shows the fraction of irrelevant variables excluded. It is clear that the fraction of included relevant variables is extremely high, as well as the fraction of excluded irrelevant regressors. Panel (e) presents the average number of included variables. Finally, Panel (f) shows the average number of included irrelevant regressors. As sample size increases, the performance of the adaLASSO improves. Overall, the results in Table 2 show that the adaLASSO is a viable alternative to model selection in high-dimensional time series models with non-Gaussian and conditionally heteroskedastic errors.

Table 3 shows the MSE for one-step-ahead out-of-sample forecasts for both the adaLASSO and oracle models. We consider a total of 100 out-of-sample observations. As expected, for low values of s, the adaLASSO has a similar performance than the oracle. For higher values of s the results

are reasonable only for $T = 300$ or $T = 1000$. The performance of the adaLASSO also improves as the sample size increases.

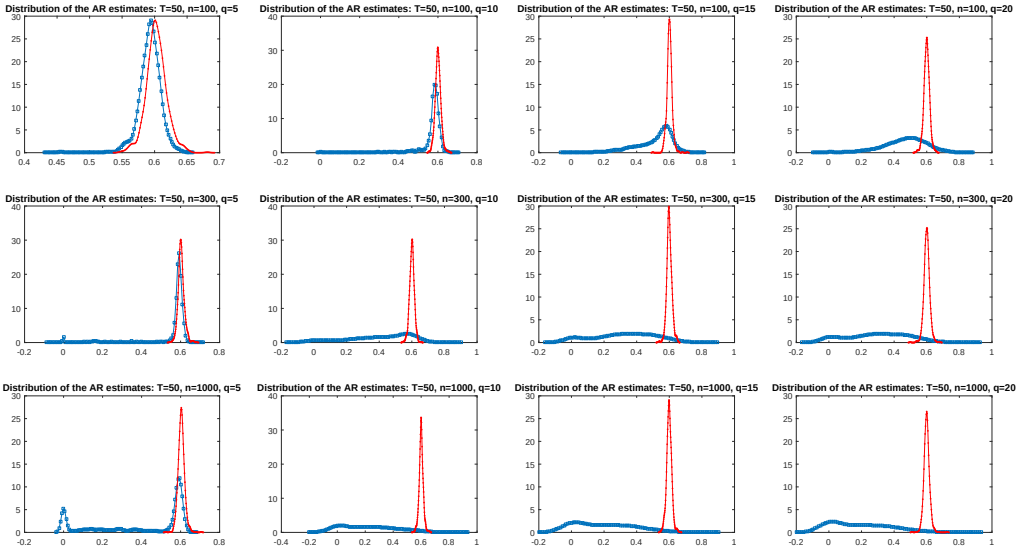

FIGURE 1. Distribution of the adaLASSO and Oracle estimators for the parameter ϕ over 1000 Monte Carlo replications. Different combinations of candidate and relevant variables. The sample size equals 50 observations.

TABLE 1. PARAMETER ESTIMATES: DESCRIPTIVE STATISTICS.

The table reports for each different sample size, the average absolute bias, Panel (a), and the average mean squared error (MSE), Panel (b), over all parameter estimates and Monte Carlo simulations. n is the number of candidate variables whereas s is the number of relevant regressors.

	$T = 50$			$T = 100$			$T = 300$			$T = 500$		
$s \backslash n$	100	300	1000	100	300	1000	100	300	1000	100	300	1000
						Panel (a): Bias $\times 10^{-3}$						
5	-0.1102	-0.2046	-0.1727	-0.0282	-0.0232	-0.0091	-0.0226	-0.0062	-0.0066	-0.0188	-0.0063	-0.0020
10	-0.3165	-0.4970	-0.2177	-0.0593	-0.0375	-0.0248	-0.0142	-0.0048	-0.0075	-0.0138	-0.0034	-0.0013
15	-1.0180	-1.0111	-0.4128	-0.0956	-0.0547	-0.1005	-0.0212	-0.0086	-0.0102	-0.0091	-0.0026	-0.0010
20	-1.3864	-0.6249	-0.2784	-0.1002	-0.1201	-0.0916	-0.0279	-0.0074	-0.0072	-0.0084	-0.0027	-0.0007
						Panel (b): MSE $\times 10^{-3}$						
5	0.0428	0.3666	0.4179	0.0068	0.0049	0.0029	0.0020	0.0007	0.0010	0.0010	0.0003	0.0001
10	0.8712	2.2258	1.1597	0.0279	0.0439	0.0620	0.0042	0.0015	0.0083	0.0018	0.0006	0.0002
15	3.8501	3.4686	1.3496	0.0477	0.2037	0.4073	0.0063	0.0024	0.0247	0.0024	0.0008	0.0002
20	6.8882	3.9529	1.4388	0.0801	0.7102	0.7621	0.0088	0.0032	0.0515	0.0029	0.0010	0.0003

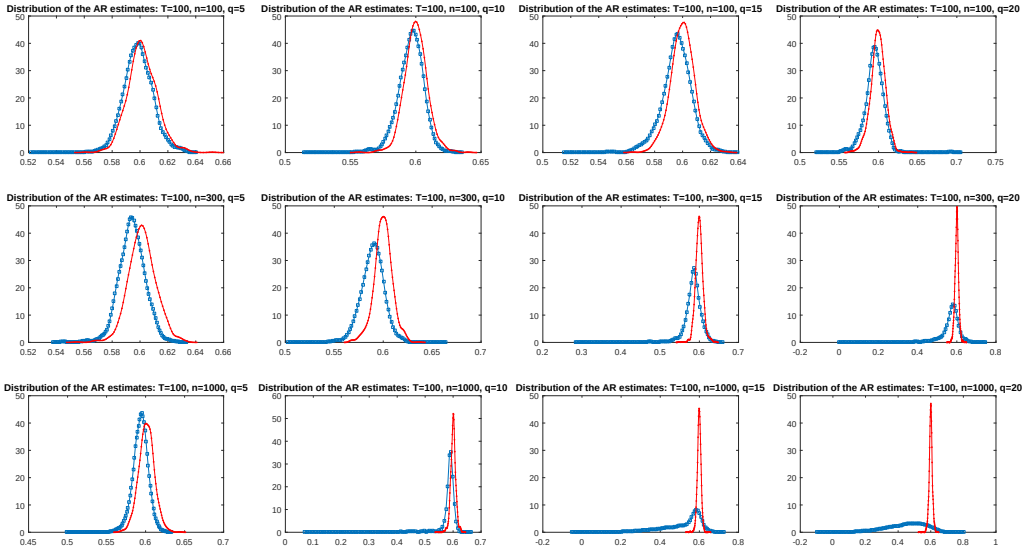

FIGURE 2. Distribution of the adaLASSO and Oracle estimators for the parameter ϕ over 1000 Monte Carlo replications. Different combinations of candidate and relevant variables. The sample size equals 100 observations.

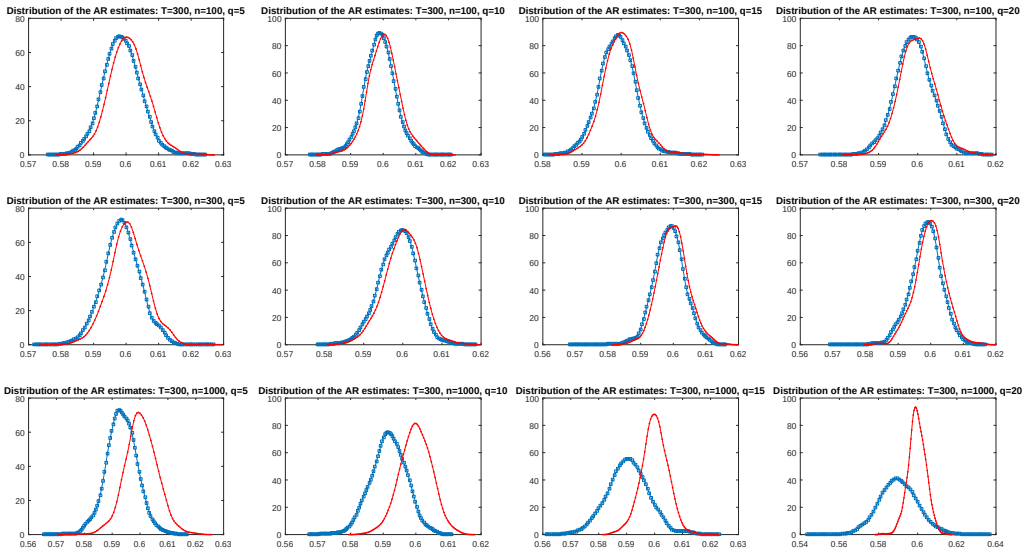

FIGURE 3. Distribution of the adaLASSO and Oracle estimators for the parameter ϕ over 1000 Monte Carlo replications. Different combinations of candidate and relevant variables. The sample size equals 300 observations.

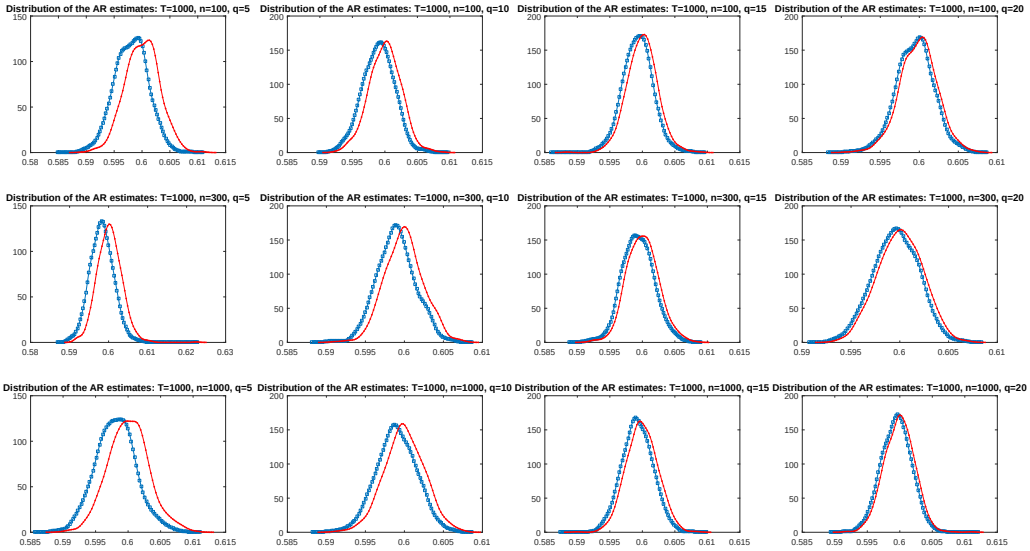

FIGURE 4. Distribution of the adaLASSO and Oracle estimators for the parameter ϕ over 1000 Monte Carlo replications. Different combinations of candidate and relevant variables. The sample size equals 1000 observations.

9. Inflation Forecasting

We consider monthly inflation forecasting with many predictors. The data consists of 131 macroeconomic variables and has been obtained from Sydney Ludvigson's webpage[4]. The dataset is the same used in Jurado et al. (2013) and is an update version of the one considered in Ludvigson and Ng (2009). The observations start in January 1960 and end in December 2011, a total of 624 time periods. The predictive regression is written as

$$\pi_{t+1} = \beta_0 + \boldsymbol{\beta}\boldsymbol{x}_t + u_{t+1},$$

where π_t is the monthly inflation at time t (percentage changes of the Consumer Price Index, CPI, for all items) and $\boldsymbol{x}_t$ is the vector of predictors (four lags of inflation plus four lags of 131 predictors. We also include four lagged factors computed as the first four principal components of the 131 predictors. Apart from the price index data which have been differenced only once, all the remaining variables were transformed according to Ludvigson and Ng (2009). We consider one step ahead forecasts computed in a rolling window scheme with 474 observations. The forecasting period starts in January 2000.

[4]http://www.econ.nyu.edu/user/ludvigsons/

Table 2. Model Selection: Descriptive Statistics.

The table reports for each different sample size, several statistics concerning model selection. Panel (a) presents the fraction of replications where the correct model has been selected, i.e., all the relevant variables included and all the irrelevant regressors excluded from the final model. Panel (b) shows the fraction of replications where the relevant variables are all included. Panel (c) presents the fraction of relevant variables included. Panel (d) shows the fraction of irrelevant variables excluded. Panel (e) presents the average number of included variables. Finally, Panel (f) shows the average number of included irrelevant regressors.

	$T = 50$			$T = 100$			$T = 300$			$T = 1000$		
$s\backslash n$	100	300	1000	100	300	1000	100	300	1000	100	300	1000
						Panel (a): Correct Sparsity Pattern						
5	0.8000	0.6540	0.4160	0.8720	0.9760	0.9430	0.9930	0.9900	1.0000	1.0000	1.0000	1.0000
10	0.7420	0.1790	0.0130	0.4900	0.9870	0.8510	0.9120	0.9090	1.0000	1.0000	1.0000	1.0000
15	0.2670	0.0050	0	0.2370	0.9260	0.3220	0.7530	0.7210	1.0000	0.9990	0.9990	0.9980
20	0.0210	0	0	0.0640	0.5860	0.0230	0.5560	0.5070	1.0000	0.9920	0.9870	0.9860
						Panel (b): True Model Included						
5	0.9990	0.8880	0.6110	1.0000	1.0000	0.9990	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	0.8820	0.2370	0.0180	1.0000	0.9990	0.9050	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
15	0.3680	0.0050	0	1.0000	0.9460	0.3760	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
20	0.0330	0	0	1.0000	0.6440	0.0280	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
						Panel (c): Fraction of Relevant Variables Included						
5	0.9994	0.9516	0.7866	1.0000	1.0000	0.9992	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	0.9659	0.6678	0.3717	1.0000	0.9999	0.9736	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
15	0.8114	0.4340	0.2193	1.0000	0.9901	0.7737	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
20	0.6024	0.3243	0.1587	1.0000	0.9154	0.5193	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
						Panel (d): Fraction of Irrelevant Excluded						
5	0.9947	0.9893	0.9907	0.9959	0.9998	0.9998	0.9998	0.9999	1.0000	1.0000	1.0000	1.0000
10	0.9823	0.9591	0.9807	0.9697	0.9998	0.9983	0.9979	0.9991	1.0000	1.0000	1.0000	1.0000
15	0.9373	0.9444	0.9796	0.9299	0.9989	0.9892	0.9938	0.9969	1.0000	1.0000	1.0000	1.0000
20	0.8987	0.9407	0.9794	0.8852	0.9911	0.9815	0.9871	0.9948	1.0000	0.9998	0.9999	1.0000
						Panel (e): Average Number of Included Variables						
5	5.4970	7.9040	13.1400	5.3860	5.0470	5.1980	5.0200	5.0250	5.0000	5.0000	5.0000	5.0000
10	11.2510	18.5250	22.7770	12.7230	10.0460	11.4390	10.1920	10.2610	10.0000	10.0000	10.0000	10.0000
15	17.5040	22.3600	23.3820	20.9560	15.1740	22.2900	15.5290	15.8860	15.0000	15.0020	15.0010	15.0040
20	20.1560	23.0770	23.3580	29.1820	20.7920	28.5220	21.0320	21.4570	20.0000	20.0160	20.0270	20.0390
						Panel (f): Average Number of Included Irrelevant Variables						
5	0.5000	3.1460	9.2070	0.3860	0.0470	0.2020	0.0200	0.0250	0	0	0	0
10	1.5920	11.8470	19.0600	2.7230	0.0470	1.7030	0.1920	0.2610	0	0	0	0
15	5.3330	15.8500	20.0930	5.9560	0.3230	10.6840	0.5290	0.8860	0	0.0020	0.0010	0.0040
20	8.1080	16.5920	20.1850	9.1820	2.4840	18.1360	1.0320	1.4570	0	0.0160	0.0270	0.0390

The forecasting results are shown in Table 4. We consider as benchmark models a linear model will all the regressors and estimated by reduced rank regression, an autoregressive (AR) model of order four, and an AR(4) model augmented by four factors. As competitors we include a model with all the variables plus the factors estimated by the LASSO procedure, the adaLASSO with LASSO initial weights and adaLASSO with Elastic-Net initial weights. The Elastic-Net is a combination

Table 3. Forecasting: Descriptive Statistics.

The table reports for each different sample size, the one-step-ahead mean squared error (MSE) for the adaLASSO, Panel(a), and the Oracle, Panel (b), estimators. n is the number of candidate variables whereas s is the number of relevant regressors.

	$T=50$			$T=100$			$T=300$			$T=500$		
$s\backslash n$	100	300	1000	100	300	1000	100	300	1000	100	300	1000
						MSE - adaLASSO						
5	0.0143	0.1299	0.6119	0.0132	0.0129	0.0128	0.0112	0.0106	0.0114	0.0103	0.0104	0.0104
10	0.1068	0.9734	1.8081	0.0140	0.0255	0.0870	0.0110	0.0110	0.0196	0.0113	0.0104	0.0109
15	0.5893	1.7082	2.2456	0.0196	0.0780	0.5545	0.0112	0.0119	0.0360	0.0104	0.0106	0.0107
20	1.0327	2.0361	2.4658	0.0268	0.2507	1.1363	0.0119	0.0118	0.0642	0.0105	0.0105	0.0105
						MSE - Oracle						
5	0.0122	0.0123	0.0126	0.0113	0.0118	0.0117	0.0112	0.0105	0.0106	0.0102	0.0103	0.0104
10	0.0155	0.0148	0.0145	0.0122	0.0125	0.0122	0.0109	0.0109	0.0114	0.0112	0.0103	0.0108
15	0.0180	0.0179	0.0177	0.0133	0.0134	0.0132	0.0110	0.0116	0.0113	0.0103	0.0104	0.0106
20	0.0235	0.0226	0.0219	0.0147	0.0148	0.0147	0.0116	0.0114	0.0115	0.0104	0.0104	0.0104

LASSO and Ridge regression are the parameters of the model are estimated as

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left[\|Y - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \alpha\lambda \sum_{i=1}^{n} |\theta_i| + (1-\alpha)\lambda \sum_{i=1}^{n} |\theta_i|^2 \right].$$

The table displays the Median Absolute Deviation (MAD), the Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) for the full forecasting period. From the inspection of the table it is clear that the LASSO-based models outperform all the benchmarks, specially when the MAE is considered.

In order to check if the differences in forecasting performance among different models are statistically significant or not we ran pairwise Giacomini-White tests for equal predictive ability. The results are summarized in Table 5. The table shows the p-value of the tests when the column model is compared to the row model according to the absolute forecasting errors (upper panel) and squared forecasting errors (lower panel). It is evident from the results that the LASSO-based models are statistically superior than the benchmark alternatives. The only case where a benchmark specification performs similarly to a competitor is then a factor model is compared to the adaLASSO with respect to the squared errors.

Figure 5 reports the cumulative absolute and squared errors for different models. There is one large error during the forecasting period (December 2008) and all models display large errors. However, the LASSO-based models continue to deliver the lowest forecasting errors. Figure 6 shows the number of variables selected by the LASSO and the adaLASSO. As expected the adaLASSO delivers more parsimonious models.

TABLE 4. Forecasting Results: Summary Statistics.

The table reports for each model the Median Absolute Deviation (MAD), the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE) and the minimum and the maximum of the out-of-sample errors.

	MAD	MAE	RMSE	Min	Max
Benchmark Models:					
All Regressors	0.0044	0.0055	0.0077	-0.0140	0.0467
AR(4)	0.0027	0.0027	0.0031	-0.0095	0.0082
AR(4) + 4 Factors	0.0027	0.0028	0.0031	-0.0081	0.0072
LASSO and adaLASSO:					
LASSO	0.0016	0.0021	0.0029	-0.0110	0.0100
adaLASSO	0.0014	0.0021	0.0031	-0.0146	0.0101
flexible adaLASSO	0.0015	0.0021	0.0028	-0.0088	0.0102

TABLE 5. p-values for Giacomini-White Test of Equal Predictive Ability.

The table reports the p-value of the Giacomini-White test for equal predictive ability. The null hypothesis is that the column model has the the same forecasting performance.

Absolute Errors

	All Regressors	AR(4)	AR(4) + 4 Factors	LASSO	adaLASSO	flex. adaLASSO
All Regressors	–	0.0000	0.0000	0.0000	0.0000	0.0000
AR(4)		–	0.3807	0.0113	0.0047	0.0401
AR(4) + 4 Factors			–	0.0088	0.0266	0.0091
LASSO				–	0.3713	0.3584
adaLASSO					–	0.3248
flex. adaLASSO						–

Squared Errors

	All Regressors	AR(4)	AR(4) + 4 Factors	LASSO	adaLASSO	flex. adaLASSO
All Regressors	–	0.0003	0.0003	0.0003	0.0003	0.0003
AR(4)		–	0.4181	0.0408	0.0121	0.0554
AR(4) + 4 Factors			–	0.0108	0.1624	0.0054
LASSO				–	0.1900	0.2518
adaLASSO					–	0.1749
flex. adaLASSO						–

10. Conclusion

We studied the asymptotic properties of the adaLASSO estimator in sparse, high-dimensional, linear time series model when both the number of covariates in the model and candidate variables can increase with the sample size. Furthermore, the number of candidate predictors is possibly larger than the number of observations. The results in this paper extend the literature by providing conditions under which the adaLASSO correctly selects the relevant variables and has the oracle property in a time-series framework with a very general error term. As a technical by-product some conditions in this paper are improvements on the frequently adopted in the shrinkage literature.

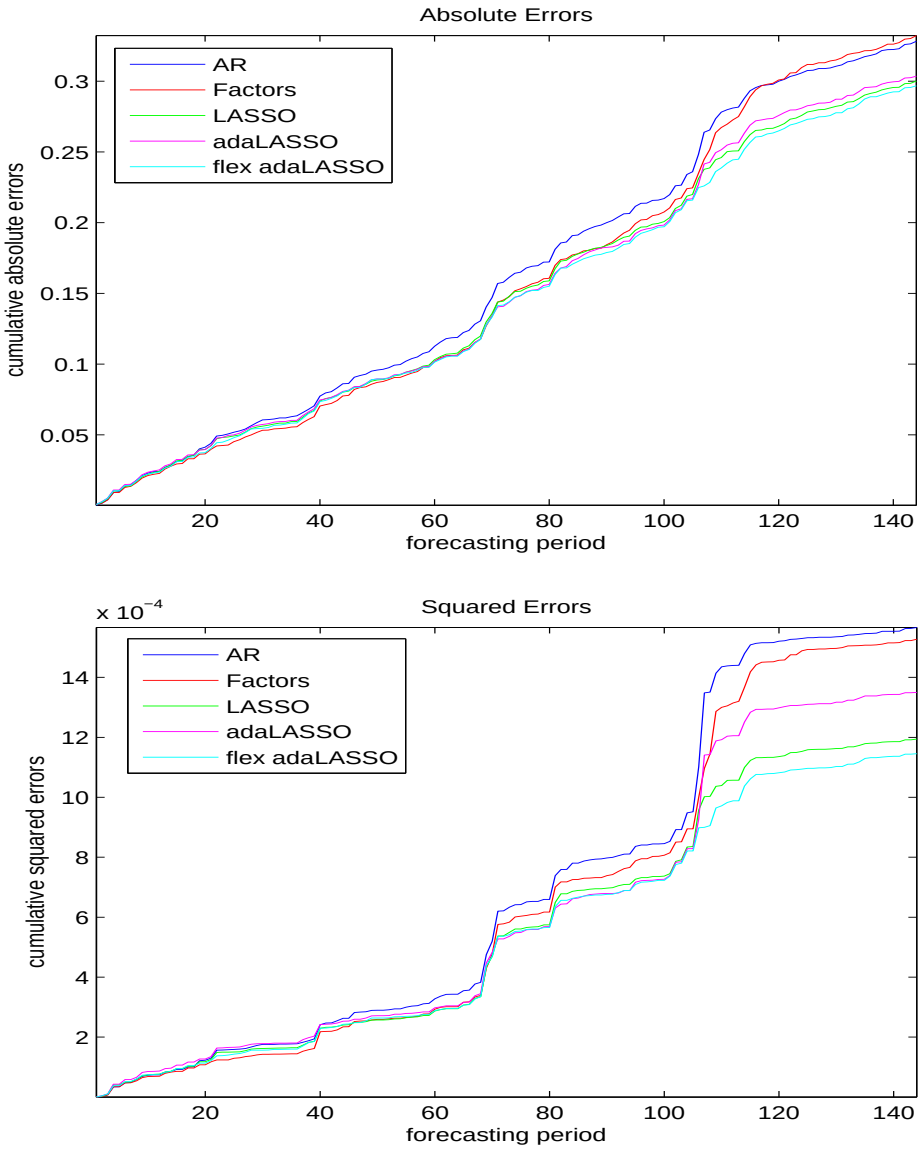

FIGURE 5. Panel (a): Cumulative absolute errors. Panel (b): Cumulative squared errors.

The main results presented in this paper are based on the assumption that only a few number of candidate variables are in fact relevant to explain the dynamics of the dependent variable (sparsity). This is a key difference from the factor models literature. The estimation of factors relies on the assumption that the loading matrix is dense, i.e., almost all variables are important for the factor determination. When the loading matrix is sparse, the usual asymptotic results for factor estimation do not hold anymore. Therefore, penalized estimation based on the adaLASSO and similar methods

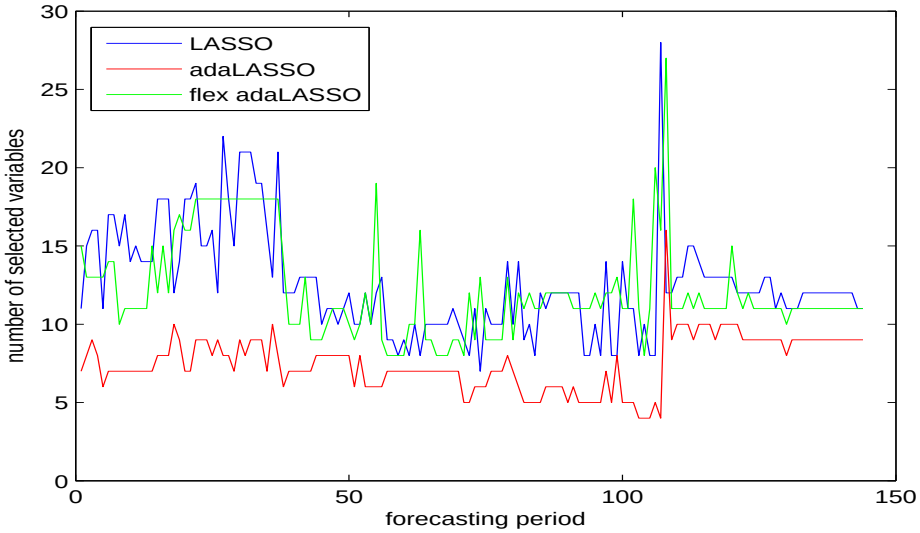

FIGURE 6. Number of parameters.

are of extreme importance. However, when the structure of the model is dense, then factor models are a better alternative.

APPENDIX A. SATISFYING ASSUMPTIONS

Let $\{\boldsymbol{z}_t\}$ denote a zero mean, weakly stationary process taking values on $\mathbb{R}^d$ ($d \in \mathbb{N}$), that admits the VMA(∞) decomposition

$$\boldsymbol{z}_t = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\epsilon}_{t-j},$$

where $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \dots, \epsilon_{dt})'$, $\mathbb{E}(\boldsymbol{\epsilon}_t | \mathcal{F}_{\epsilon,t-1}) = 0$, $\mathbb{E}(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t' | \mathcal{F}_{\epsilon,t-1}) = \boldsymbol{\Sigma}_t$, $\boldsymbol{\Psi}_j = \text{diag}(\psi_{j1}, \cdots, \psi_{jd})$ [5], and let $\mathcal{F}_{\epsilon,t} = \sigma\{\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_{t-1}, \dots\}$. In order to ensure $\mathbb{E}[\boldsymbol{z}_1 \boldsymbol{z}_1']$ is bounded independently of d, we also require that the largest eigenvalue of each Σ_t is bounded independently of d and that $\sum_{j=0}^{\infty} \psi_{ji}^2 < \infty$. It implicitly requires that the correlation matrices Σ_t are not dense or, at least, that most elements are sufficiently small, which is standard in the ℓ_1 regularization literature. As in C, we characterize the dependence of the series through $\zeta_{i,r} = \sum_{j=r}^{\infty} |\psi_{ij}|$ ($i = 1, \dots, d$), and it is assumed that they decrease either polynomially or geometrically with r.

Assume that the set of candidate variables are $\boldsymbol{x}_t = (\boldsymbol{z}_t', \dots, \boldsymbol{z}_{t-p}')'$ for some $p < T$, usually much smaller. To simplify the exposition, let $\boldsymbol{x}_t(1) = (x_{i,t},\, x \in S)$, where S denote the active set, i.e.,

[5]The assumption that $\boldsymbol{\Psi}_j$ is diagonal is only to simplify our calculations and can be relaxed, i.e., let $\boldsymbol{\Gamma}\boldsymbol{\Delta} = \boldsymbol{I}$ and $\boldsymbol{\eta}_t = \Delta \boldsymbol{\epsilon}_t$, then taking $\boldsymbol{\Psi}_j^* = \boldsymbol{\Psi}_j \boldsymbol{\Gamma}$ yield an equivalent representation.

the set of all included regressors. The number of candidate variables is $n = p \times d$ and both may increase with T. Throughout the section, we use c, c_s, $c_1, c_2, \ldots$ as positive and finite constants.

A.1. **Satisfying DGP(3).** Assumption DGP(3) is equivalent to

$$\sum_{k=0}^{p}\sum_{j=1}^{d}\Pr\left\{\left|\sum_{t=p+1}^{T}\left[z_{j,t-p}^2-\mathbb{E}(z_{j1}^2)\right]\right|>Tc\right\}\to 0,\quad T\to\infty.$$

We use the triplex inequality (19) and results in Appendix C to find conditions that satisfy DGP(3) in this setting.

Table 6 shows conditions to satisfy DGP(3). We impose conditions on $\{\epsilon_{it}\}$ $(i = 1, \ldots, d)$, the mixingale dependence term $\zeta_{i,r} = \sum_{j=r}^{\infty}|\psi_{ij}|$ $(i = 1, \ldots, d)$, and the rates of increase of n and p.

TABLE 6. Conditions on the rate of increase of n and p, the mixingale dependence terms $\{\zeta_{i,r} : i = 1, \ldots, n\}$, and the tail behaviour of ϵ_{it} $(t = 1, \ldots, T,\ i = 1, \cdots, n)$, for satisfying DGP(3). The conditions hold for any $0 < \delta < 1$, some $d \geq 1$, and some $u > 0$.

Dependence \ Tail	$\mathbb{E}\left(\lvert\epsilon_{it}\rvert^{2m}\right)<\infty$	$\mathbb{E}\left[\exp(u\lvert\epsilon_{it}\rvert)-1-u\lvert\epsilon_{it}\rvert\lvert\mathcal{F}_{\epsilon,t-1}\right]\leq f(u)\sigma_t^2$
$\zeta_{i,r}=0,\ r>r_0$	$n=o\left[T^{\delta(m-1)}\right]$	$\log n=o\left(T^{1/5}\right)$
	$p=o(T)$	$p=o(T)$
$\zeta_{i,r}\propto r^{-\zeta}$	$n=o\left[T^{\delta\left(\frac{1}{m-1}+\frac{1}{\zeta}\right)^{-1}}\right]$	$n=o\left(T^{\delta\zeta}\right)$
	$p=o\left[T^{\frac{\delta}{2\zeta}\left(\frac{1}{m-1}+\frac{1}{\zeta}\right)^{-1}}\right]$	$p=o\left(T^{\delta/2}\right)$
$\log\zeta_{i,r}\propto -r^{\zeta}$	$n=o\left[T^{\delta(m-1)}\right]$	$\log n=o\left(T^{\frac{\zeta}{5\zeta+2}}\right)$
	$p=o\left(T^{\frac{\delta}{2+\zeta}}\right)$	$p=o\left(T^{\frac{2}{2+5\zeta}}\right)$

The derivations are mechanical and the same method is applied in each of the six combinations of conditions. The first step is to adapt the triplex inequality to the problem in hand. Assume that $r \propto T^{\gamma_1/2}$ and $C_T \propto T^{\gamma_2/2}$ in (19),

$$\Pr\left[\left|\sum_{t=p+1}^{T}z_{j,t-k}^2-\mathbb{E}\left(z_{j1}^2\right)\right|>T\varepsilon\right]\leq 2c_1T^{\gamma_1/2}\exp\left(-\frac{c_2\varepsilon^2T^{1-\gamma_1-\gamma_2}}{288}\right)+D_{k,T}+E_T,$$

$D_{k,T}$ is the dependence term of (19), and E_T its tail term[6]. We use results in Appendix C.3 to bound $D_{k,T}$ and E_T. If $p = o(r)$, then $r-k > r-p > r/2$ for T sufficiently large. If the dependence

[6] $D_{k,T}=(6/\varepsilon)T^{-1}\sum_{t=1}^{T}|\mathbb{E}(z_{j,t-k}^2|\mathcal{F}_{\epsilon,t-r})-\mathbb{E}(z_{j1}^2)|$ and $E_T=(15/\varepsilon)T^{-1}\sum_{t=1}^{T}\mathbb{E}[|z_{j,t-k}^2|I(z_{j,t-k}^2>C_T)]$.

vanishes after r_0 lags, i.e., $D_T = 0$ for $r > r_0$, we can set $\gamma_1 = 0$. If the dependence decreases polynomially with r then $D_T \leq O(T^{-\zeta\gamma_1})$; if the dependence term decreases exponentially with r then $-\log D_T \leq O(T^{\zeta\gamma_1/2})$. As for the tail, $E_T \leq O\left[T^{-\gamma_2(m-1)}\right]$ in the polynomial case, and $-\log E_T \leq O(T^{\gamma_2/4})$ in the exponential case.

We optimize the convergence rate choosing the pair (γ_1, γ_2) that makes all three terms decrease at the same rate. In the case both dependence ($D_{k,T}$) and tail (E_T) terms decrease exponentially, we solve the system $1 - \gamma_1 - \gamma_2 = \gamma_2/4$ and $\gamma_2 = 2\zeta\gamma_1$. The RHS of (19) is bounded by

$$\max_{0\leq k\leq p, 1\leq j\leq d} \Pr\left\{\left|\sum_{t=p+1}^{T}\left[z_{j,t-p}^2 - \mathbb{E}(z_{j1}^2)\right]\right| > Tc\right\} \leq c_1 \exp\left(-c_2\, T^{\frac{\zeta}{5\zeta+2}}\right),$$

for positive and finite constants c_1 and c_2. Therefore,

$$\sum_{k=0}^{p}\sum_{j=1}^{d} \Pr\left\{\left|\sum_{t=p+1}^{T}\left[z_{j,t-p}^2 - \mathbb{E}(z_{j1}^2)\right]\right| > Tc\right\} \leq \sum_{k=0}^{p}\sum_{j=1}^{d} c_2 \exp\left(-c_4\, T^{\frac{\zeta}{5\zeta+2}}\right) \to 0,$$

as $T \to \infty$ if we assume $\log n = o\left[T^{\zeta/(5\zeta+2)}\right]$.

The remaining terms in Table 6 are derived similarly.

A.2. **Satisfying DESIGN(3).** Using the union bound on DESIGN(3) part (b),

$$\Pr\left[\max_{i,j\in S}\left|\sum_{t=p+1}^{T} x_{it}x_{jt} - \mathbb{E}(x_{it}x_{jt})\right| \geq \frac{T\phi_{\min}}{s}\right] \leq s^2 \max_{i,j\in S} \Pr\left[\left|\sum_{t=q+1}^{T} x_{it}x_{jt} - \mathbb{E}(x_{it}x_{jt})\right| \geq \frac{T\phi_{\min}}{s}\right].$$

Recall that $x_{it}x_{jt} = z_{l_1,t-k_1}z_{l_2,t-k_2}$ where $1 \leq l_1, l_2 \leq d$ and $0 \leq k_1, k_2 \leq p$. While the actual terms l_1, l_2, k_1, and k_2 are unimportant, S may contain terms with lags up to p, which has an influence on the dependence term as in the previous section. The RHS of the previous display is bounded by

$$s^2 \max_{0\leq k\leq p,\, 1\leq i,j\leq d} \Pr\left[\left|\sum_{t=p+1}^{T} z_{it}z_{j,t-k} - \mathbb{E}(z_{it}z_{j,t-k})\right| \geq \frac{T\phi_{\min}}{s}\right]. \tag{10}$$

We use (19) and results in Appendix C to find conditions that satisfy DESIGN(3) in this setting.

The first step is to adapt the triplex inequality to the problem in hand. Assume that $r \propto T^{\gamma_1/2}$ and $C_T \propto T^{\gamma_2/2}$ in (19),

$$(10) \leq 2c_1 s^2 T^{\gamma_1/2} \exp\left[-\frac{c_2(\phi_{\min}/s)^2 T^{1-\gamma_1-\gamma_2}}{288}\right] + c_3\frac{s^3}{\phi_{\min}}D_{k,T} + \frac{s^3}{\phi_{\min}}E_T,$$

$D_{k,T}$ is the dependence term of the triplex inequality, and E_T its tail term[7]. We use results in Appendix C.3 to bound $D_{k,T}$ and E_T. If $p = o(r)$, then $r - k > r - p > r/2$ for T sufficiently large. If the dependence vanishes after r_0 lags, i.e., $D_T = 0$ for $r > r_0$, we can set $\gamma_1 = 0$. If the dependence decreases polynomially with r then $D_T \leq O(T^{-\zeta\gamma_1})$; if the dependence term decreases exponentially with r then $-\log D_T \leq O(T^{\zeta\gamma_1/2})$. As for the tail, $E_T \leq O\left[T^{-\gamma_2(m-1)}\right]$ in the polynomial case, and $-\log E_T \leq O(T^{\gamma_2/4})$ in the exponential case.

The derivation of the bounds follow the same steps as in A.1, with the further constraint that $(s/\phi_{\min}) = o(T^{\delta/2})$ where $0 < \delta = 1-\gamma_1-\gamma_2 < 1$. If $D_{k,T} = 0$, the condition on s and p are already satisfied independently of E_T. If $D_{k,T}$ and E_T are polynomial, then we need $p = o\left(T^{\frac{1-\delta}{4\zeta}H(d,\zeta)}\right)$ and $s = O\left\{T^{\frac{\delta}{2}\wedge\frac{1}{4}[(1-\delta)H(d,\zeta)-\delta]}\right\}$, where $H(m,\zeta) = 2/[1/(m-1)+1/\zeta]$ is the harmonic average of $m-1$ and ζ. If $D_{k,T}$ and E_T decrease geometrically, then the condition is satisfied for $p = o\left[T^{1/(2+5\zeta)}\right]$. Now, let $0 < \delta < 1-\nu < 1$, for some $0 < \nu < 1$. If $D_{k,T}$ is polynomial and E_T geometric, we need $p = o\left[T^{(1-\nu-\delta)/2}\right]$ and $s = O\left(T^{\frac{\delta}{2}\wedge\zeta\frac{1-\nu-\delta}{2}}\right)$. Finally, if $D_{k,T}$ is geometric and E_T is polynomial, then $p = o(T^{\nu/2})$ and $s = O\left[T^{\frac{\delta}{2}\wedge(1-\nu-\delta)(d-1)}\right]$. The previous bounds hold with $\nu = \delta$, in which case $0 < \delta < 1/2$.

A.3. **Satisfying condition WEIGHTS.** We show that under stronger assumptions on the covariance matrix of the candidate variables and number of covariates, conditions on Lemma 1 are satisfied. These conditions also imply that DGP(3) and DESIGN(3) are satisfied. Kock and Callot (2012) show that these conditions are satisfied if the covariates are generated form a Gaussian VAR. The approach we use here is similar. Assume that the smallest eigenvalue of the population covariance matrix is bounded away from zero and that $\phi_0 > 16c > 0$ in Lemma 1. We show that

$$\Pr\left[\max_{1\leq i,j\leq n}\left|\sum_{t=p+1}^{T} x_{it}x_{jt} - \mathbb{E}(x_{it}x_{jt})\right| \geq \frac{Tc}{s}\right] \to 0,$$

as $T \to \infty$.

We use the triplex inequality again. The argument is the same as used in A.1 and A.2. The arguments are as in A.2, and we bound (10) with $\phi_{\min}$ replaced by a constant c, and s^2 replaced by n^2, because we are now dealing with the full empirical covariance matrix of the variables, instead of only the ones that enter in the model.

[7] $D_{k,T} = T^{-1}\sum_{t=1}^{T}|\mathbb{E}(z_{it}z_{j,t-k}|\mathcal{F}_{\epsilon,t-r}) - \mathbb{E}(z_{i1}z_{j1})|$ and $E_T = T^{-1}\sum_{t=1}^{T}\mathbb{E}[|z_{it}z_{j,t-k}|I(z_{it}z_{j,t-k} > C_T)]$.

The mechanics is the same as before. We assume that $s = o(T^{\delta/2})$ for some $1 < \delta < 1$, and that $p = o(r)$. If $D_{k,T} = 0$ and E_T is polynomial, p is not constrained and $n = o\left[T^{(1-\delta)(1-m)/2}\right]$. If E_T decreases geometrically, $\log n = o\left[T^{(1-\delta)/5}\right]$. If $D_{k,T}$ and E_T decrease geometrically, then $\log n = o\left[T^{\zeta(1-\delta)/(2+4\zeta)}\right]$ and $p = o\left[T^{(1-\delta)/(2+4\zeta)}\right]$. If both $D_{k,T}$ and E_T decrease polynomially, $n = o\left[T^{H(m,\zeta)(1-\delta)/4}\right]$ and $p = o\left[T^{H(m,\zeta)(1-\delta)/(4\zeta)}\right]$, where $H(m,\zeta) = 2/[1/(m-1) + 1/\zeta]$ is the harmonic average of $m-1$ and ζ. Now, let $0 < \delta < 1-\nu < 1$, for some $0 < \nu < 1$ and assume that $s = o(T^{\delta/2})$. If $D_{k,T}$ is polynomial and E_T geometric, we need $p = o\left(T^{(1-\nu-\delta)/2}\right)$ and $n = O\left[T^{\zeta(1-\nu-\delta)/2}\right]$. Finally, if $D_{k,T}$ is geometric and E_T is polynomial, then $p = o(T^{\nu/2})$ and $n = O\left[T^{(1-\nu-\delta)(m-1)/2}\right]$. The previous bounds hold with $\nu = \delta$, in which case $0 < \delta < 1/2$.

If we add bounded ε_t to the list, the tail term E_T is zero, and the number of variables would depend only on the dependence term. In any case, the number of variables depend both on the dependence structure of the covariates and their tail behaviour. For instance, if the tail does not decrease geometrically, the number of candidate variables cannot increase sub-exponentially. Similarly, when the dependence term is polynomial, n increases at most polynomially.

Hence, under the previous conditions on the increase rate of n, p, and s, and assuming the population covariance matrix of all covariates satisfy the *restricted eigenvalue condition*, the LASSO can be used as initial estimator and the condition WEIGHTS is satisfied.

Appendix B. Proofs

B.1. **Initial weights.**

Proof of Lemma 1. The first statement follows directly from Lemma 6 in Kock and Callot (2012) and the second one from comparing DESIGN(3) and the conditions in the lemma, in a set with probability one. □

Proof of Lemma 2. The proof follows after Theorem 6.1 in Bülhmann and van der Geer (2011) and the relationship between *restricted eigenvalue condition* and *compatibility condition*. The second part follows because (a) satisfies conditions of Lemma 4 and (b) satisfies conditions of 5. □

Proof of Proposition 1. The weights are given by $|\theta_{I,i}|^{-\tau}$, which means that WEIGHTS(1) is equivalent to

$$\max_{s+1\leq i\leq n} |\theta_{I,i}| \leq \sum_{i=s+1}^{n} |\theta_{I,i} - \theta_{0,i}| \leq T^{-\xi/2\tau} c_{w(2)}^{-1/\tau},$$

because $\theta_{i,0} = 0$ for all $q+1 \leq i \leq n$. Hence, WEIGHTS(1) is satisfied whenever

$$T^{-\xi/2\tau} c_{w(2)}^{-1/\tau} \geq c_1 \frac{\lambda}{T} \frac{s}{\phi_{\min}},$$

which holds under assumption on λ.

Let $x, y \in \mathbb{R}$ and $2(x-y)^2 \leq y^2$, $y^2 \leq 2\left(x^2 + (x-y)^2\right)$ which means that $x^2 \geq y^2/2 - (x-y)^2 > 0$. Moreover, $x^{2\tau} \geq \left(y^2/2 - (x-y)^2\right)^\tau$. Under the conditions on $\theta_{\min}$ and the bound on $|\theta_{I,i} - \theta_{0i}|$, $|\theta_{I,i}|^{2\tau} \geq \left(0.5|\theta_{0i}|^2 - |\theta_{I,i} - \theta_{0i}|^2\right)^\tau$. The left hand side of WEIGHTS(2) is upper bounded by

$$s^{-1} \sum_{i=1}^{s} w_i^2 \leq \max_{1 \leq i \leq s} |\theta_{I,i}|^{-2\tau} \leq \left(0.5 \min_{1 \leq i \leq s} |\theta_{0,i}|^2 - \max_{1 \leq i \leq s} |\theta_{I,i} - \theta_{0,i}|^2\right)^{-\tau}.$$

Substituting this bound on WEIGHTS(2),

$$\max_{1 \leq i \leq s} |\theta_{I,i} - \theta_{0,i}|^2 \leq 0.5\theta_{\min}^2 - w_{\max}^{-2/\tau}.$$

It follows by assumption that $(\theta_{\min}^2/2 - w_{\max}^{-2/\tau})^{1/2} \geq \frac{1}{2}\theta_{\min} \geq c_1(\lambda/T)(s/\phi_{\min})$. Therefore,

$$\max_{1 \leq i \leq s} |\theta_{I,i} - \theta_{0,i}| \leq c_1 \frac{\lambda}{T} \frac{s}{\phi_{\min}},$$

is a sufficient condition for WEIGHTS(2), which is satisfied by the ℓ_1 oracle bound. □

B.2. **Minimal Eigenvalue.** Condition DESIGN(3) imply that the smallest eigenvalue of $\widehat{\boldsymbol{\Omega}}_{11}$ is lower bounded by $\phi_{\min}$. We use this result throughout the proofs in this section.

Lemma 3. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ denote two non-negative definite, r-dimensional square matrices. If $\max_{1 \leq i,j \leq r} |A_{ij} - B_{ij}| \leq \delta$, then*

$$\inf_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1} \boldsymbol{\alpha}'\boldsymbol{B}\boldsymbol{\alpha} > \inf_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1} \boldsymbol{\alpha}'\boldsymbol{A}\boldsymbol{\alpha} - r\delta.$$

Proof. The proof is parallel to Lemma 6.17 in Bülhmann and van der Geer (2011). Let $\boldsymbol{\alpha} \in \mathbb{R}^r \setminus \{0\}$.

$$\boldsymbol{\alpha}'\boldsymbol{A}\boldsymbol{\alpha} - \boldsymbol{\alpha}'\boldsymbol{B}\boldsymbol{\alpha} \leq |\boldsymbol{\alpha}'(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{\alpha}| \leq |\boldsymbol{\alpha}|_1 |(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{\alpha}|_\infty \leq |\boldsymbol{\alpha}|_1^2 \delta \leq r\boldsymbol{\alpha}'\boldsymbol{\alpha}\delta,$$

where $|\cdot|_1$ and $|\cdot|_\infty$ are the ℓ_1 and sup norm, respectively. Rearranging the terms,

$$\frac{\boldsymbol{\alpha}'\boldsymbol{B}\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\boldsymbol{\alpha}} \geq \frac{\boldsymbol{\alpha}'\boldsymbol{A}\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\boldsymbol{\alpha}} - r\delta.$$

The result follows by minimizing over $\boldsymbol{\alpha} \in \mathbb{R}^r \setminus \{0\}$. □

Under condition DESIGN(3), in a set with probability converging to one

$$\inf_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1} \boldsymbol{\alpha}'\widehat{\boldsymbol{\Omega}}_{11}\boldsymbol{\alpha} > \inf_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1} \boldsymbol{\alpha}'\boldsymbol{\Omega}_{11}\boldsymbol{\alpha} - s\frac{\phi_{\min}}{s} > \phi_{\min}.$$

B.3. **Bounding the empirical process.** The regularization parameter λ is intrinsically connected to probability bounds on the event

$$\mathcal{E}_T(\lambda_0) = \left\{ 2 \max_{i=1,\ldots,n} T^{-1/2} \left| \sum_{i=1}^{T} x_{it} u_t \right| < \lambda_0 \right\}.$$

We derive bounds for the event $\mathcal{E}_T\left[\lambda T^{-(1-\xi)/2}\right]$ under (i) Assumptions DGP(1)-DGP(2) and DGP(4) (case i), and (ii) Asumptions DGP(1)–DGP(2) and DGP(5). More precisely, we show that

$$\Pr\left(\max_{i=1,\ldots,n} T^{-1/2} \left| \sum_{i=1}^{T} x_{it} u_t \right| > \frac{\lambda}{2\sqrt{T}} T^{\xi/2} \right) \leq c\, h_T(n,\xi),$$

for some positive constant c. The sequence $h_T(n,\xi) = \left[n^{1/m} T^{(1-\xi)/2}/\lambda\right]^m$ in case (i), and $h_T(n,\xi) = \exp\left[\log n - c_2 \lambda^{2/5} T^{-(1-\xi)/5}\right]$ in case (ii), for some constant c_2. Then, under conditions on the lower bound of λ, we find that $\Pr\left\{\mathcal{E}_T\left[\lambda T^{-(1-\xi)/2}\right]\right\} \leq 1 - cT^{-\alpha}$, for some $\alpha > 0$.

Lemma 4. *Under Assumptions DGP(1), DGP(2), DGP(3), and $\lambda \geq n^{1/m} T^{(1-\xi)/2+\alpha/m}$, for some $\alpha > 0$,*

$$\Pr\left\{\mathcal{E}_T\left[\lambda T^{-(1-\xi)/2}\right]\right\} \geq 1 - \frac{2^d c}{T^\alpha}, \tag{11}$$

for some positive constant c.

Proof. Write $\Pr[\mathcal{E}_T(\lambda_0)] = 1 - \Pr[\mathcal{E}_T^c(\lambda_0)]$. Simple application of the union bound and the Markov inequality yield

$$\begin{aligned} \Pr[\mathcal{E}_T^c(\lambda_0)] &\leq \sum_{i=1}^{n} \Pr\left(2T^{-1/2} \left| \sum_{i=1}^{T} x_{it} u_t \right| > \lambda_0 \right) \\ &\leq 2^m n^{-1} \sum_{i=1}^{n} \mathbb{E} \left| \frac{1}{\sqrt{T}} \sum_{i=1}^{T} x_{it} u_t \right|^m \left(\frac{n^{1/m}}{\lambda_0} \right)^m. \end{aligned}$$

Under Assumption DGP(2), $\{u_t x_{jt}\}$ is a martingale difference process with respect to $\mathcal{F}_t$. Application of the Burkhölder-Davis-Gundy inequality and the C_r-inequality, yield

$$\mathbb{E}\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_t x_{jt}\right|^m \leq C'_m \mathbb{E}\left|\frac{1}{T}\sum_{t=1}^{T}(u_t x_{jt})^2\right|^{m/2} \leq C_m \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}|u_t x_{jt}|^m.$$

Under DGP(3) and setting $\lambda_0 = \lambda T^{-(1-\xi)/2}$

$$\Pr[\mathcal{E}_T^c(\lambda_0)] \leq 2^m C_m c_m \left[\frac{n^{1/m} T^{(1-\xi)/2}}{\lambda}\right]^m \leq \frac{2^m C_m c_m}{T^\alpha},$$

for $\lambda \geq n^{1/m} T^{(1-\xi)/2+\alpha/m}$. □

Lemma 5. *Assume that DGP(1), DGP(2) and DGP(5) hold jointly, and that T is sufficiently large. Let $\log n \geq (\log T)^2$ and, for any $\alpha > 0$, $\lambda \geq c'(\log n + \alpha \log T)^{5/2} T^{(1-\xi)/2}$, for $c' > 0$ sufficiently large, than*

$$\Pr\left\{\mathcal{E}_T\left[\lambda T^{-(1-\xi)/2}\right]\right\} \geq 1 - \frac{c_1}{T^\alpha}, \tag{12}$$

for some positive constant c_1.

Proof. Write $\Pr[\mathcal{E}_T(\lambda_0)] = 1 - \Pr[\mathcal{E}_T^c(\lambda_0)]$. Using the union bound and Lemma 6 gives

$$\begin{aligned}\Pr[\mathcal{E}_T^c(\lambda_0)] &\leq \sum_{i=1}^{n} \Pr\left(2T^{-1/2}\left|\sum_{i=1}^{T} x_{it} u_t\right| > \lambda_0\right) \\ &\leq c_1 \exp\left(\log n - c_2 \lambda_0^{2/5}\right).\end{aligned}$$

The use of Lemma 6 is justified because whenever $\log n > (\log T)^2$, the condition on h_T is satisfied with $\log(h_T T^{-1/2})/h_T^{2/5} > 1/\log T \to 0$ as $T \to \infty$. Choose $c' > [\min(b_3, b_4, 1/27)/4]^{-5/2}$ and set $\lambda_0 = \lambda T^{-(1-\xi)/2}$ with $\lambda = c'(\log n + \alpha \log T)^{5/2} T^{(1-\xi)/2}$. Then $\Pr(\mathcal{E}_T^c) \leq c_1/T^\alpha$. The result follows. □

In the proof of Lemma 5 we used the following lemma, that defines a sub-exponential bound on the probability of the empirical process.

Lemma 6. *Let $\{\mathcal{F}_t\}_{-\infty}^{\infty}$ denote as sequence of increasing σ-fields and let $\{x_t u_t\}_{t=-\infty}^{T}$ be a difference martingale sequence with respect to $\{\mathcal{F}_t\}_{t=-\infty}^{T}$, for each T. Assume that there exist positive constants $b_1, \ldots, b_4$ such that for any c sufficiently large, $\Pr(u_t > c) \leq b_1 \exp(-b_2 c)$ and*

$\Pr(x_t > c) \leq b_3 \exp(-b_4 c)$. *Then,*

$$\Pr\left(\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t u_t\right| > h_T\right) \leq c_1 \exp\left(-c_2 h_T^{2/5}\right), \tag{13}$$

where c_1 *and* c_2 *are positive constants,* h_T *is a non-decreasing sequence, and* $\log T/h_T^{2/5} \to 0$ *as* $T \to \infty$.

Proof. The proof consists in applying the Triplex inequality (19) and optimizing the right hand side. The second term on the Triplex inequality is zero because $\mathbb{E}(x_t u_t|\mathcal{F}_t) = 0$. Take $r = 1$, $\varepsilon_T = h_T/\sqrt{T}$ and $C_T = h_T^{\gamma}$, then

$$\Pr\left(\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t u_t\right| > h_T\right) \leq 2\exp\left(-h_T^{2-2\gamma}/288\right) + \frac{15\sqrt{T}}{h_T}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[|x_t u_t| I(|x_t u_t| > h_T^{\gamma})].$$

The expectation in the RHS is bounded by

$$\mathbb{E}|x_t u_t| I(|x_t u_t| > h_T^{\gamma}) \leq \left(\mathbb{E}|x_t u_t|^2\right)^{1/2} \times \left[\Pr\left(|u_t| > h_T^{\gamma/2}\right) + \Pr\left(|x_t| > h_T^{\gamma/2}\right)\right]^{1/2}.$$

The first term on the RHS is bounded because both $\Pr(x_t > c)$ and $\Pr(u_t > c)$ decrease exponentially, as for the second term

$$\begin{aligned}\left[\Pr\left(|u_t| > h_T^{\gamma/2}\right) + \Pr\left(|x_t| > h_T^{\gamma/2}\right)\right]^{1/2} &\leq \left[b_1 \exp\left(-b_2 h_T^{\gamma/2}\right) + b_3 \exp\left(-b_4 h_t^{\gamma/2}\right)\right]^{1/2} \\ &\leq b_5 \exp\left(-b_6 h_T^{\gamma/2}/2\right),\end{aligned}$$

where $b_5^2 = 2(b_1 + b_3)$ and $b_6 = \min(b_2, b_4)$.

Therefore,

$$Pr\left(\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t u_t\right| > h_T\right) \leq 2\exp\left(-h_T^{2-2\gamma}/288\right) + b_7\frac{\sqrt{T}}{h_T}\exp\left(-b_6 h_T^{\gamma/2}/2\right),$$

where $b_7 \geq 15 b_5 T^{-1}\sum_{t=1}^{T}(\mathbb{E}|x_t u_t|^2)^{1/2}$. Choosing $\gamma = 4/5$ optimizes the convergence rate and the result follows from the assumption on h_T and by taking $c_1 = 2 + b_7$ and $c_2 = \min(1/288, b_6/4)$. □

B.4. **Proof of Theorem 1.** Write $\widehat{\boldsymbol{\Omega}} = \frac{\boldsymbol{X}'\boldsymbol{X}}{T}$, $\widehat{\boldsymbol{\Omega}}_{11} = \frac{\boldsymbol{X}(1)'\boldsymbol{X}(1)}{T}$, $\widehat{\boldsymbol{\Omega}}_{22} = \frac{\boldsymbol{X}(2)'\boldsymbol{X}(2)}{T}$ and $\widehat{\boldsymbol{\Omega}}_{21} = \widehat{\boldsymbol{\Omega}}_{12}' = \frac{\boldsymbol{X}(2)'\boldsymbol{X}(1)}{T}$. Set $\boldsymbol{\nu}_0 = \text{sign}[\boldsymbol{\theta}_0(1)]$, where $\text{sign}(x) = I(x > 0) - I(x < 0)$. Let $\boldsymbol{W}(1) = \text{diag}(w_1, \ldots, w_s)$.

The Karush-Kuhn-Tucker conditions characterize the solution of the optimization problem in equation (2). These conditions are standard in the LASSO literature and have been used in Zhao

and Yu (2006), Huang et al. (2008), among many others, to find sufficient conditions for sign consistency. Proposition 2 provides a lower bound on the probability that the signs of the estimated and true parameters are equal, and follows the same construction as Zhao and Yu (2006).

Proposition 2. *Let* $\boldsymbol{W}(1) = \text{diag}(w_1, \ldots, w_s)$ *and* $\boldsymbol{\nu}_0 = \text{sign}[\boldsymbol{\theta}_0(1)]$. *Then*

$$\Pr\left[\text{sign}(\widehat{\boldsymbol{\theta}}) = \text{sign}(\boldsymbol{\theta})\right] \geq P\left(\mathcal{A}_T \cap \mathcal{B}_T\right),$$

where

$$\mathcal{A}_T = \bigcap_{i=1}^{s} \left\{ \frac{1}{\sqrt{T}} |[\widehat{\boldsymbol{\Omega}}_{11}^{-1} \boldsymbol{X}(1)' \boldsymbol{U}]_i| < \sqrt{T}|\boldsymbol{\theta}_{0i}| - \frac{\lambda}{2\sqrt{T}} |[\widehat{\boldsymbol{\Omega}}_{11}^{-1} \boldsymbol{W}(1)\boldsymbol{\nu}_0]_i| \right\}, \tag{14a}$$

$$\mathcal{B}_T = \bigcap_{i=s+1}^{n} \left\{ 2|\frac{1}{\sqrt{T}} \boldsymbol{X}_i' \boldsymbol{M}(1)\boldsymbol{U}| < \frac{1}{\sqrt{T}} \lambda \left[w_i - |T^{-1} \boldsymbol{X}_i' \boldsymbol{X}(1) \widehat{\boldsymbol{\Omega}}_{11}^{-1} \boldsymbol{W}(1)\boldsymbol{\nu}_0| \right] \right\}, \tag{14b}$$

where $\boldsymbol{U} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}_0$, $\boldsymbol{M}(1) = \boldsymbol{I}_T - \boldsymbol{X}(1)(\boldsymbol{X}(1)'\boldsymbol{X}(1))^{-1}\boldsymbol{X}(1)'$.

Proof of Proposition 2. The proof follows as in Proposition 1 of Zhao and Yu (2006). □

The sets $\mathcal{A}_T$ and $\mathcal{B}_T$ and loosely interpreted as "keeping relevant variables inside the model" and "leave irrelevant variables outside the model". Proposition 2 provides a lower bound on the probability of selecting the correct model:

$$P\left[\text{sign}(\widehat{\boldsymbol{\theta}}) = \text{sign}(\boldsymbol{\theta}_0)\right] \geq P\left(\mathcal{A}_T \cap \mathcal{B}_T\right) \geq 1 - P\left(\mathcal{A}_T^c\right) - P\left(\mathcal{B}_T^c\right),$$

where $\mathcal{A}_T^c$ and $\mathcal{B}_T^c$ are the complements of $\mathcal{A}_T$ and $\mathcal{B}_T$ respectively. Theorem 1 follows by showing that $\mathcal{A}_t \cap \mathcal{B}_t \supseteq \mathcal{E}_T[\lambda T^{-(1-\xi)/2}]$ and claim Lemma 4 to conclude.

Lemma 7. *Assume DESIGN and WEIGHTS(2) hold jointly, and that* $\theta_{\min} > (\lambda/T^{1-\xi/2})(s^{1/2}/\phi_{\min})$. *Then* $\mathcal{A}_T \supseteq \mathcal{E}_T[\lambda T^{-(1-\xi)/2}]$.

Proof. Write the event $\mathcal{A}_T^c$

$$\mathcal{A}_T^c = \bigcup_{i=1}^{s} \left\{ \frac{1}{\sqrt{T}} |[\widehat{\boldsymbol{\Omega}}_{11}^{-1} \boldsymbol{X}(1)' \boldsymbol{U}]_i| \geq \sqrt{T}|\boldsymbol{\theta}_{0i}| - \frac{\lambda}{2\sqrt{T}} |[\widehat{\boldsymbol{\Omega}}_{11}^{-1} \boldsymbol{W}(1)\boldsymbol{\nu}_0]_i| \right\}.$$

Simple application of the Cauchy-Schwartz inequality to the left hand side of the inequality above yields

$$\begin{aligned}\left|T^{-1/2}[\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{X}(1)'\boldsymbol{U}]_i\right| &= T^{-1/2}\sup_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1}\boldsymbol{\alpha}'\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{X}(1)'\boldsymbol{U} \\ &\leq \sup_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1}(\boldsymbol{\alpha}'\widehat{\boldsymbol{\Omega}}_{11}^{-2}\boldsymbol{\alpha})^{1/2}\left[\sum_{j=1}^{s}\left(T^{-1/2}\boldsymbol{X}_j'\boldsymbol{U}\right)^2\right]^{1/2} \\ &\leq \phi_{\min}^{-1}\times\left[\sum_{j=1}^{s}\left(T^{-1/2}\boldsymbol{X}_j'\boldsymbol{U}\right)^2\right]^{1/2}. \qquad (15)\end{aligned}$$

Similarly,

$$\begin{aligned}|[\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{\nu}_0]_i| &= T^{-1/2}\sup_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1}\boldsymbol{\alpha}'\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{\nu}_0 \\ &\leq \left(\inf_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1}\boldsymbol{\alpha}'\widehat{\boldsymbol{\Omega}}_{11}\boldsymbol{\alpha}\right)^{-1}\left(\sum_{j=1}^{s}w_j^2\right)^{1/2} \qquad (16) \\ &\leq \phi_{\min}^{-1}s^{1/2}w_{\max}\leq\frac{s^{1/2}T^{\xi/2}}{\phi_{\min}} \qquad (17)\end{aligned}$$

Combining (15) and (17), and under the assumption that $\theta_{\min}>(\lambda/T^{1-\xi/2})(s^{1/2}/\phi_{\min})$,

$$\sqrt{T}|\theta_{0i}|-\frac{\lambda}{2\sqrt{T}}\left|T^{-1/2}[\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{X}(1)'\boldsymbol{U}]_i\right|\geq\frac{\lambda}{2\sqrt{T}}\frac{s^{1/2}T^{\xi/2}}{\phi_{\min}},$$

and

$$\mathcal{A}_T^c\subseteq\left\{\sum_{j=1}^{s}\left(T^{-1/2}\boldsymbol{X}_j'\boldsymbol{U}\right)^2\geq\left(\frac{\lambda}{2T^{(1-\xi)/2}}s^{1/2}\right)^2\right\}\subseteq\left\{\max_{i=1,\ldots,s}2\left|T^{-1/2}\boldsymbol{X}_j'\boldsymbol{U}\right|\geq\frac{\lambda}{T^{(1-\xi)/2}}\right\},$$

proving the claim. □

Lemma 8. *Under Assumptions DGP(3), WEIGHTS and DESIGN, $\mathcal{B}_T\supseteq\mathcal{E}_T(\lambda T^{-(1-\xi)/2})$, for all T sufficiently large.*

Proof. Write

$$\mathcal{B}_T=\bigcup_{i=s+1}^{n}\left\{2\left|\frac{1}{\sqrt{T}}\boldsymbol{X}_i'\boldsymbol{M}(1)\boldsymbol{U}\right|\geq\frac{1}{\sqrt{T}}\lambda\left[w_i-|T^{-1}\boldsymbol{X}_i'\boldsymbol{X}(1)\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{\nu}_0|\right]\right\}. \qquad (18)$$

We will bound the left and right hand side (LHS, RHS) of (18) separately and combine the bounds. Recall that $\boldsymbol{M}(1) = \boldsymbol{I}_T - \boldsymbol{X}(1)[\boldsymbol{X}(1)'\boldsymbol{X}(1)]^{-1}\boldsymbol{X}(1)'$, which means that

$$\boldsymbol{X}_i'\boldsymbol{M}(1)\boldsymbol{U} = \boldsymbol{X}_i'\boldsymbol{U} - \boldsymbol{X}_i'\boldsymbol{X}(1)[\boldsymbol{X}(1)'\boldsymbol{X}(1)]^{-1}\boldsymbol{X}(1)'\boldsymbol{U} = A_i + B_i.$$

The second term on the RHS, B_i, is bounded by

$$\begin{aligned}
|B_i| &= \left|\boldsymbol{X}_i'\boldsymbol{X}(1)[\boldsymbol{X}(1)'\boldsymbol{X}(1)]^{-1}\boldsymbol{X}(1)'\boldsymbol{U}\right| \\
&\leq \left|\sum_{t=1}^{T} x_{it}^2\right|^{1/2} \left|\boldsymbol{U}'\boldsymbol{X}(1)[\boldsymbol{X}(1)'\boldsymbol{X}(1)]^{-1}\boldsymbol{X}(1)'\boldsymbol{U}\right|^{1/2} \\
&\leq \sqrt{T}\left|\frac{1}{T}\sum_{t=1}^{T}\left[x_{it}^2 - \mathbb{E}(x_{it}^2)\right] + \max_{i=1,\ldots,n}\mathbb{E}(x_{it}^2)\right|^{1/2} \left[\frac{\sum_{i=1}^{s}\left(T^{-1/2}\boldsymbol{X}_j'\boldsymbol{U}\right)^2}{\phi_{\min}}\right]^{1/2} \\
&\leq c_1\left[\frac{\sum_{i=1}^{s}\left(T^{-1/2}\boldsymbol{X}_j'\boldsymbol{U}\right)^2}{\phi_{\min}}\right]^{1/2},
\end{aligned}$$

where $c_1 = 1 \vee \sqrt{2\max_{i=1,\ldots,n}\mathbb{E}(x_{it}^2)}$ for all T sufficiently large from DGP(3).

Therefore, the LHS of (18) is bounded by

$$2\left|\frac{1}{\sqrt{T}}\boldsymbol{X}_i'\boldsymbol{M}(1)\boldsymbol{U}\right| \leq \frac{2c_1}{\sqrt{T}}\sum_{t=1}^{T} x_{it}u_t + \frac{2c_1}{\phi_{\min}^{1/2}}\sqrt{\sum_{j=1}^{s}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_{jt}u_t\right)^2}.$$

As for the RHS, it follows from the Cauchy-Schwarz inequality, WEIGHTS(2), DESIGN(3), and DGP(3) that

$$\begin{aligned}
\left[T^{-1}\boldsymbol{X}_i'\boldsymbol{X}(1)\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{\nu}_0\right]^2 &= \left\{\boldsymbol{X}_i'\boldsymbol{X}(1)[\boldsymbol{X}(1)'\boldsymbol{X}(1)]^{-1}\boldsymbol{W}(1)\boldsymbol{\nu}_0\right\}^2 \\
&\leq \boldsymbol{\nu}_0'\boldsymbol{W}(1)(\boldsymbol{X}(1)'\boldsymbol{X}(1))^{-1}\boldsymbol{W}(1)\boldsymbol{\nu}_0 \times \boldsymbol{X}_i'\boldsymbol{X}_i \\
&\leq \frac{\sum_{j=1}^{s} w_i^2}{\inf_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1}\boldsymbol{\alpha}'\widehat{\boldsymbol{\Omega}}_{11}\boldsymbol{\alpha}} \times \frac{\sum_{t=1}^{T} x_{it}^2}{T} \\
&\leq \left(\frac{c_1 s^{1/2} T^{\xi/2}}{\phi_{\min}^{1/2}}\right)^2
\end{aligned}$$

Combining the previous bound with WEIGHTS(1) yield the following bound to the RHS of (18)

$$\frac{\lambda}{\sqrt{T}}\left(w_i - \left|T^{-1}\boldsymbol{X}_i'\boldsymbol{X}(1)\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{\nu}_0\right|\right) \geq \frac{2c_1\lambda}{T^{(1-\xi)/2}}\sqrt{\frac{s}{\phi_{\min}}}.$$

Hence,

$$\begin{aligned}\mathcal{B}_T^c &\subseteq \left\{\max_{1\le i\le s}\frac{2}{\sqrt{T}}\sum_{t=1}^{T}x_{it}u_t \ge \frac{\lambda}{T^{(1-\xi)/2}}\right\}\cup\left\{\max_{s+1\le i\le n}\frac{2}{\sqrt{T}}\sum_{t=1}^{T}x_{it}u_t \ge \frac{\lambda}{T^{(1-\xi)/2}}\sqrt{\frac{s}{\phi_{\min}}}\right\}\\ &\subseteq \left\{\max_{1\le i\le n}\frac{2}{\sqrt{T}}\sum_{t=1}^{T}x_{it}u_t \ge \frac{\lambda}{T^{(1-\xi)/2}}\right\},\end{aligned}$$

proving the claim. □

Proof of Theorem 1. Combining Proposition 2 with Lemmata 7 and 8 yield

$$\Pr\left[\operatorname{sign}(\widehat{\boldsymbol{\theta}})=\operatorname{sign}(\boldsymbol{\theta})\right]\ge P\left(\mathcal{A}_T\cap\mathcal{B}_T\right)\ge \Pr\left\{\mathcal{E}\left[\lambda T^{-(1-\xi)/2}\right]\right\}\to 1,$$

from Assumption REG and Lemma 4. □

B.5. **Proof of Theorem 2.** Write $\dot{\boldsymbol{Q}}_T(\boldsymbol{\theta})=-2\boldsymbol{X}'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\theta})+\lambda\boldsymbol{W}\boldsymbol{\nu}_\theta$, where $\boldsymbol{\nu}_\theta=[\operatorname{sign}(\theta_1),\ldots,\operatorname{sign}(\theta_n)]'$ and $\boldsymbol{W}=\operatorname{diag}(w_1,\ldots,w_n)$. Replacing $\boldsymbol{\theta}$ by the adaLASSO estimator and writing $\boldsymbol{U}=\boldsymbol{Y}-\boldsymbol{X}(1)\boldsymbol{\theta}_0(1)$, and for any $\boldsymbol{\alpha}'\boldsymbol{\alpha}=1$,

$$\sqrt{T}\boldsymbol{\alpha}'\left[\widehat{\boldsymbol{\theta}}(1)-\boldsymbol{\theta}_0(1)\right]=\frac{1}{\sqrt{T}}\boldsymbol{\alpha}'\left[\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{X}(1)'\boldsymbol{U}\right]-\sqrt{T}\boldsymbol{\alpha}'\widehat{\boldsymbol{\Omega}}_{11}^{-1}\widehat{\boldsymbol{\Omega}}_{12}\widehat{\boldsymbol{\theta}}(2)+\frac{\lambda}{2\sqrt{T}}\boldsymbol{\alpha}\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{\nu}_\theta(1).$$

The first term on the RHS equals $\sqrt{T}\boldsymbol{\alpha}'\left[\widehat{\boldsymbol{\theta}}_{ols}(1)-\boldsymbol{\theta}_0(1)\right]$. The proof consists in showing that the second and third term on the RHS converge to zero in probability. Since $\Pr[\widehat{\boldsymbol{\theta}}(2)=0]\to 1$ from Theorem 1, the second term vanishes in probability. As for the third term

$$\begin{aligned}\left(\frac{\lambda}{2\sqrt{T}}\boldsymbol{\alpha}\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{\nu}_\theta(1)\right)^2 &\le \frac{\lambda^2\sum_{i=1}^{s}w_i^2}{4T\left(\inf_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1}\boldsymbol{\alpha}'\widehat{\boldsymbol{\Omega}}_{11}\boldsymbol{\alpha}\right)^2}\\ &\le \left(\frac{\lambda}{\sqrt{T}}\frac{s^{1/2}w_{\max}}{2\phi_{\min}}\right)^2\\ &\to 0,\quad T\to\infty.\end{aligned}$$

The first line follows from the Cauchy-Schwarz inequality, the second from WEIGHTS(2) and DESIGN(3), and the last one from REG.

B.6. **Proof of Theorem 3.** The only difference between Theorem (3) and the previous results is that now the number of variables may increase faster.

The proof of (a) is identical to the proof of Theorem 1 with Lemma 4 replaced by Lemma 5. The proof of (b) is identical to the proof of Theorem 2.

Appendix C. Auxiliary Lemmata

This section we show some auxiliary results used in the previous derivation. We start with the triplex inequality (Jiang 2009, Theorem 1) and expand on how to bound its terms.

Theorem (Triplex Inequality)**.** *Let $\{\mathcal{F}_t\}_{t=-\infty}^{\infty}$ be an increasing sequence of σ-fields, and x_t be a random variable that is $\mathcal{F}_t$-measurable for each t. Then, for each ε_T, $C_T > 0$ and positive integers r and T, we have*

$$\Pr\left\{\left|\sum_{t=1}^{T}\left[x_t - \mathbb{E}(x_t)\right]\right| > T\varepsilon_T\right\} \leq 2r\exp\left[-T\varepsilon_T^2/(288r^2C_T^2)\right] + (6/\varepsilon_T)T^{-1}\sum_{t=1}^{T}\mathbb{E}\left|\mathbb{E}(x_t|\mathcal{F}_{t-r}) - \mathbb{E}(x_t)\right| + (15/\varepsilon_T)T^{-1}\sum_{t=1}^{T}\mathbb{E}\left|x_t\right| I\left(|x_t| > C_T\right), \quad (19)$$

as long as the RHS exists and is smaller than one.

The first term in the RHS is self explanatory and depends on the dependence window m, the upper bound C_T, and ε_T. The second term on the RHS is the dependence term and is described in the framework of ℓ_1-mixingale (see, e.g., Chapter 16, Davidson 1994). When $\{x_t\}$ is a martingale difference process, the dependence term vanishes. We derive bounds for the dependence term under different dependence assumptions. Finally, the third term on the RHS captures the tail behaviour of x_t and we also derive bounds for it under different tail conditions.

C.1. **Tail behaviour.** Next series of results will deal with the tail behaviour of the elements $\boldsymbol{x}_t$ under different conditions. For the sake of simplicity, and without of generality, consider $\boldsymbol{x}_t$ scalar and denote it x_t. Lemma 9 provides sufficient condition so that the tail decreases exponentially. We assume that x_t admits an MA(∞) representation where the innovations have bounded conditional variances, and impose conditions on the coefficients and innovation process. Lemma 10 follows a different direction and derives a polynomial bound assuming that $|x_t|$ has up to p moments.

Lemma 9. *Let $\{x_t\}_{t=-\infty}^{\infty}$ denote a second order, stationary process that admits an MA(∞) decomposition. Write $x_t = \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j}$, where $\{\epsilon_t\}_{t=-\infty}^{\infty}$ satisfy one of the following settings:*

(1) is an independent and identically distributed sequence (i.i.d.) of random variables, with mean zero, and $|\theta|_1 = \sum_{j=0}^{\infty} |\theta_j| < \infty$. Furthermore, ϵ_1 has a cumulant generating function $K(u) = \log \mathbb{E}\left[\exp\left(u\epsilon_1\right)\right]$ that is continuously differentiable at zero.

(2) $\{\epsilon_t, \mathcal{F}_{\epsilon,t-1}\}_{-\infty}^{\infty}$ is a martingale difference sequence, where $\mathcal{F}_{\epsilon,t-1} = \sigma\{\epsilon_{t-1}, \epsilon_{t-2}, \dots\}$, $|\boldsymbol{\theta}|_2^2 = \sum_{j=0}^{\infty} |\theta_j|^2 < \infty$, and $\mathbb{E}(\epsilon_t^2 | \mathcal{F}_{\epsilon,t-1}) = \sigma_t^2 < \infty$. Furthermore, each ϵ_t satisfies

$$\mathbb{E}\left[\exp\left(u|\epsilon_t|\right) - 1 - u|\epsilon_t| \,\middle|\mathcal{F}_{\epsilon,t-1}\right] \leq f(u)\mathbb{E}\left(\epsilon_t^2 | \mathcal{F}_{\epsilon,t-1}\right),$$

for any positive u and $f(u)$.

Then there exist positive constants b_1 and b_2 such that $\Pr(|x_t| > c) < b_1 \exp\left(-b_2 c\right)$. Moreover, the tail condition in the triplex inequality is bounded by

$$\frac{15}{\varepsilon_T} T^{-1} \sum_{t=1}^{T} \mathbb{E}\left|x_t\right| I\left(|x_t| > C_T\right) \leq \frac{15 b_1 \sqrt{\mathbb{E}|x_t|^2}}{\varepsilon_T} e^{-b_2 C_T/2}.$$

Proof. The Markov inequality yield, for some $u > 0$,

$$\Pr(|x_t| > c) = \Pr\left[\exp\left(u|x_t|\right) > \exp(uc)\right] \leq \exp\left(-uc\right) \exp\left\{\log \mathbb{E}[\exp(u|x_t|)]\right\}.$$

Under setting (1),

$$\begin{aligned}
\exp\left\{\log \mathbb{E}\left[\exp(\pm u x_t)\right]\right\} &= \exp\left[\sum_{j=0}^{\infty} \log K(\pm\theta_j u)\right] \\
&\leq \exp\left[\sum_{j-1}^{\infty} \sup_{|\nu| \leq u|\theta|_1} |K'(\nu)||u\theta_j|\right] \\
&\leq 2\exp\left[|\theta|_1 |u| \sup_{|\nu| \leq |u||\theta|_1} |K'(\nu)|\right],
\end{aligned}$$

where $K'(u) = dK(u)/du$, $b_1 = 2\exp\left[|\theta|_1 |b_2| \sup_{|\nu| \leq |b_2||\theta|_1} |K'(\nu)|\right]$ and $b_2 = u$.

Under setting (2),

$$\begin{aligned}
\mathbb{E}[\exp(u|x_t|)] &\leq \mathbb{E}\left[\exp\left(u\sum_{j=0}^{\infty}|\theta_j z_{t-j}|\right)\right] \\
&= \mathbb{E}\left(\exp\left[u\sum_{j=1}^{\infty}|\theta_j z_{t-j}|\right]\exp\left\{\log\mathbb{E}\left[\exp\left(u|\theta_0\epsilon_t|\right)|\mathcal{F}_{\epsilon,t-1}\right]\right\}\right) \\
&\leq \mathbb{E}\left(\exp\left[u\sum_{j=1}^{\infty}|\theta_j z_{t-j}|\right]\exp\left\{\mathbb{E}\left[\exp(u|\theta_0\epsilon_t|)-1-u|\theta_0\epsilon_t|\big|\mathcal{F}_{\epsilon,t-1}\right]\right\}\right) \\
&\leq \exp\left[f(u)\theta_0^2\sigma_t^2\right]\mathbb{E}\left[\exp\left(u\sum_{j=1}^{\infty}|\theta_j z_{t-j}|\right)\right] \\
&\leq \cdots \leq \exp\left(f(u)\sum_{j=0}^{\infty}\theta_j^2\sigma_{t-j}^2\right) \leq \exp\left[f(u)\sigma_{\max}^2|\theta|_2^2\right],
\end{aligned}$$

where $\sigma_{\max}^2 = \max_t \sigma_t^2$. We may choose $b_2 = u$ and $b_1 = \exp\left[f(b_2)\sigma_{\max}^2|\theta|_2^2\right]$. A tighter bound may be obtained by minimizing the right hand side $\inf_{0<u<u_0}\exp[f(u)\sigma_{\max}^2|\theta|_2^2 - C_T u]$.

The tail bound follows after the Cauchy-Schwarz inequality. □

The key assumptions are on the innovation process $\{\epsilon_t\}$. The first one requires that the cumulant generating function of the innovations, $K(u)$, is continuously differentiable at zero. This condition is satisfied, for instance, by Gaussian innovations. As for the second condition, assume each ϵ_t satisfy the Bernstein moment condition: for all $k \geq 2$,

$$\mathbb{E}\left(|\epsilon_t|^k\big|\mathcal{F}_{\epsilon,t-1}\right) \leq k!b^{k-2}\frac{\mathbb{E}\left(\epsilon_t^2|\mathcal{F}_{t-1}\right)}{2},$$

for some $b > 0$. Then, for all $0 < u < 1/b$, $\mathbb{E}\left[\exp\left(u|\epsilon_t|\right) - 1 - u|\epsilon_t|\big|\mathcal{F}_{\epsilon,t-1}\right] \leq f(u)\mathbb{E}(\epsilon_t^2|\mathcal{F}_{t-1})$ with $f(u) = u^2/2(1-ub)$ [8].

It is usually the case that the moment generating function does not exist. In such situations we require a polynomial bound on the tail term of the triplex inequality.

Lemma 10. *Assume there exist positive c_p and $p > 1$ such that $\mathbb{E}|x_t|^p < c_p$ for all t. Then, $\mathbb{E}[|x_t|I(|x_t| > C_t)] \leq c_p C_T^{-(p-1)}$. Moreover, the tail condition in the triplex inequality is satisfied*

[8] Under this condition a Bernstein-type bound may be derived, i.e., $\Pr(|x_t| > c) \leq b_1\exp\left[-c^2/2(|\theta|_2^2\sigma_{\max}^2 + cb)\right]$.

with

$$\frac{15}{\varepsilon_T} T^{-1} \sum_{t=1}^{T} \mathbb{E}\left[|x_t| \, I\left(|x_t| > C_T\right)\right] \leq \frac{15 c_p}{\varepsilon_T C_T^{p-1}}.$$

Proof. It follows after simple application of the Holder inequality:

$$\begin{aligned}
\mathbb{E}[|x_t| I(|x_t| > C_t)] &\leq \mathbb{E}(|x_t|^p)^{1/p} \Pr(|x_t| > C_t)^{(p-1)/p} \\
&\leq \mathbb{E}(|x_t|^p)^{1/p} \mathbb{E}(|x_t|^p)^{(p-1)/p} / C_t^{p(p-1)/p} \\
&= \mathbb{E}(|x_t|^p) C_t^{-(p-1)}.
\end{aligned}$$

□

The assumption that $\mathbb{E}(|x_t|^p)$ is bounded is not restrictive and, whenever x_t can be written as an MA(∞) process, we are only required that $|\theta|_1 < \infty$ and $\mathbb{E}|\epsilon_s|^p < \infty$ for all s. Under these conditions, $\mathbb{E}\left(|x_t|^p\right) \leq |\theta|_1^{p-1} \sum_{j=0}^{\infty} |\theta_j| \mathbb{E}|\epsilon_{t-j}|^p \leq |\theta|_1^p \max_t \mathbb{E}|\epsilon_t|^p$.[9] In this case, $c_p = |\theta|_1^p \max_t \mathbb{E}|\epsilon_t|^p$ in the previous lemma.

C.2. **Dependence term.** Bounds on the dependence term are derived under mixing and mixingale assumptions. We further extend the results to the processes $\{x_{it}^2\}_{t=-\infty}^{\infty}$, and $\{x_{1,t}x_{2,t}\}_{t=-\infty}^{\infty}$ where $\boldsymbol{x}_t = (x_{1,t}, x_{2,t})'$ admits a VMA(∞) decomposition.

Lemma 11. *(1) Assume the pairs $\{\boldsymbol{x}_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ form an ℓ_1-mixingale sequence with mixingale coefficients $\{\zeta_r\}_{-\infty}^{\infty}$*[10]*, then*

$$\frac{6}{\varepsilon_T T} \sum_{t=1}^{T} \mathbb{E}\Big|\mathbb{E}(\boldsymbol{x}_t|\mathcal{F}_{t-r}) - \mathbb{E}(\boldsymbol{x}_t)\Big| \leq \frac{6\, c_1}{\varepsilon_T} \zeta_r,$$

where c_1 is some positive constant.

(2) Assume $\{\boldsymbol{x}_t\}_{t=-\infty}^{\infty}$ is strong mixing with mixing coefficients $\{\alpha_r\}_0^{\infty}$, and each $\mathbb{E}|x_{it}|^p < \infty$. Then

$$\frac{6}{\varepsilon_T T} \sum_{t=1}^{T} \mathbb{E}\Big|\mathbb{E}(\boldsymbol{x}_t|\mathcal{F}_{t-r}) - \mathbb{E}(\boldsymbol{x}_t)\Big| \leq \frac{36\, c_1}{\varepsilon_T} \alpha_r^{1-1/p},$$

[9] Write $\sum_j |\theta_j \epsilon_{t-j}| = \sum_j |\theta_j|^{1-1/p} (|\theta_j|^{1/p} |\epsilon_{t-j}|)$ and use the Hölder inequality to find $\sum_j |\theta_j \epsilon_{t-j}| \leq |\theta|_1^{p-1} \sum_j |\theta_j| |\epsilon_{t-j}|^p$.

[10] $\mathbb{E}|\mathbb{E}(\boldsymbol{x}_t|\mathcal{F}_{t-r}) - \mathbb{E}(\boldsymbol{x}_t)| \leq a_t \zeta_r$, $r = \pm 1, \pm 2, \ldots$.

for some positive constant c_1*. If, instead,* $\{\boldsymbol{x}_t\}$ *is uniform mixing, the inequality holds with* $\alpha_r^{1-1/p}$ *replaced by* $\phi_r^{1-1/p}/3$*.*

Proof. Result for mixingale follows from the definition of a mixingale process. Result for strong (uniform) mixing follows after direct application of Theorems 14.2 (Theorem 14.4) in Davidson (1994). □

The terms α_r and ϕ_r are respectively the strong and the uniform mixing coefficients. Conditions on the rate of decrease of the mixing coefficients yield polynomial or exponential bounds on the second term, i.e., if $\alpha_r = O(e^{-\alpha\, r})$ then $\alpha_r^{1-1/p}/\varepsilon_T \leq c\, e^{-[\alpha(p-1)/p]\, r}/\varepsilon_T$ for some positive c. Similarly, $\alpha_r = O(r^{-\alpha})$ yield $\alpha_r^{1-1/p}/\varepsilon_T \leq c\, r^{-[\alpha(p-1)/p]}/\varepsilon_T$. If the process is strong (uniform) mixing with size $-\alpha$ $(-\phi)$, Theorem 14.1 in Davidson (1994) shows that the process $\{x_t^2\}$ is also strong (uniform) mixing with the same size. Similar results also hold for the mixingale case.

C.3. **Processes admitting a VMA(∞) decomposition.** Assume $\boldsymbol{x}_t = (x_{1,t}, x_{2,t})'$ is a second order stationary process that admits the $MA(\infty)$ decomposition

$$\boldsymbol{x}_t = \begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix} = \sum_{j=0}^{\infty} \begin{pmatrix} \theta_{1,j}\epsilon_{1,t-j} \\ \theta_{2,j}\epsilon_{2,t-j} \end{pmatrix} = \sum_{j=0}^{\infty} \boldsymbol{\Theta}_j \boldsymbol{\epsilon}_{t-j}, \tag{20}$$

where $\mathbb{E}[\boldsymbol{\epsilon}_t|\mathcal{F}_{\epsilon,t-1}] = 0$, $\mathbb{E}[\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t'|\mathcal{F}_{\epsilon,t-1}] = \Sigma_t$, with $[\Sigma_t]_{i,i} = \sigma_{it}^2$ and $[\boldsymbol{\Sigma}_t]_{i,j} = \rho_t$, and $\mathcal{F}_{z,t} = \sigma\{\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_{t-1}, \dots\}$.

We may also write

$$x_{it} = \sum_{j=0}^{\infty} \theta_{i,j}\epsilon_{i,t-j} \quad i = 1, 2,$$

which means that the marginal tail bounds on Lemmata 9 and 10 hold. It also follows that each x_{it} is a mixingale process with $\zeta_r = \sum_{j=r}^{\infty} |\theta_{i,j}|$ (see, e.g., Davidson 1994, Example 16.2) , which means that the dependence term may also be bounded as in part 1 of Lemma 11.

Lemma 12. *Let* $\{\boldsymbol{x}_t\}$ *satisfy* (20)*, with* $\sum_{j=r}^{\infty} |\theta_{i,j}| \leq \zeta_{i,r}$*. Then, for* $i, j \in \{1, 2\}$*,* $t = 1, \dots, T$*, and constants* $c_{i,j}(t) < \infty$*,*

(1) $[\mathbb{E}|\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t'|\mathcal{F}_{\epsilon,t-r}) - \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')|]_{i,j} \leq c_{i,j}(t)\zeta_{i,r}\zeta_{j,r}$*,*

(2) $\left[\mathbb{E}|\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_{t-k}'|\mathcal{F}_{\epsilon,t-r}) - \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_{t-k}')|\right]_{i,j} \leq c_{i,j}(t)\zeta_{i,r}\zeta_{i,r-k}$*,*

(3) If $\mathbb{E}\left({\epsilon_{it}}^{2p}\right) < c_{i,2p}$*, then* $\mathbb{E}\left[|x_{it}x_{jt}||I(|x_{it}x_{jt} > C_T)\right] \leq \sqrt{c_{i,2p}c_{j,2p}}/C_T^{p-1}$*.*

(4) If $\mathbb{E}\left[\exp(u|\epsilon_{it}|) - 1 - u|\epsilon_{it}||\mathcal{F}_{\epsilon,t-1}\right] \leq f(u)\sigma_{it}^2$ *then* $\mathbb{E}\left[|x_{it}x_{jt}||I(|x_{it}x_{jt} > C_T)\right] \leq b_1 \exp\left(-b_2 C_T^{1/2}\right)$, *for positive constants* b_1 *and* b_2.

Proof. (1) Write $\boldsymbol{x}_t = \boldsymbol{v}_t + \boldsymbol{w}_t$ where $\boldsymbol{v}_t = \sum_{j=0}^{r-1} \boldsymbol{\Theta}_j \boldsymbol{\epsilon}_{t-j}$ and $\boldsymbol{w}_t = \sum_{j=r}^{\infty} \boldsymbol{\Theta}_j \boldsymbol{\epsilon}_{t-j}$. Expand the product $\boldsymbol{x}_t\boldsymbol{x}_t' = \boldsymbol{v}_t\boldsymbol{v}_t' + \boldsymbol{v}_t\boldsymbol{w}_t' + \boldsymbol{w}_t\boldsymbol{v}_t' + \boldsymbol{w}_t\boldsymbol{w}_t'$. The last term is measurable with respect to $\mathcal{F}_{\epsilon,t-r}$, $\mathbb{E}[\boldsymbol{v}_t|\mathcal{F}_{\epsilon,t-r}] = 0$, and the first term is such that $\mathbb{E}(\boldsymbol{v}_t\boldsymbol{v}_t'|\mathcal{F}_{\epsilon,t-r}) = \sum_{j=0}^{r} \boldsymbol{\Theta}_j \boldsymbol{\Sigma}_{t-j} \boldsymbol{\Theta}_j' = \mathbb{E}(\boldsymbol{v}_t\boldsymbol{v}_t')$. Then $\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t'|\mathcal{F}_{\epsilon,t-r}) - \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t') = \boldsymbol{w}_t\boldsymbol{w}_t' - \mathbb{E}(\boldsymbol{w}_t\boldsymbol{w}_t')$. It follows that

$$\begin{aligned}
\mathbb{E}\left|\boldsymbol{w}_t\boldsymbol{w}_t' - \mathbb{E}[\boldsymbol{w}_t\boldsymbol{w}_t']\right| &= \mathbb{E}\left|\sum_{i=r}^{\infty}\sum_{j=r}^{\infty} \boldsymbol{\Theta}_i \left(\boldsymbol{\epsilon}_{t-i}\boldsymbol{\epsilon}_{t-j} - \mathbb{E}\boldsymbol{\epsilon}_{t-i}\boldsymbol{\epsilon}_{t-j}\right) \Theta_j\right| \\
&\leq \sum_{i=r}^{\infty}\sum_{j=r}^{\infty} |\boldsymbol{\Theta}_i| \mathbb{E}\left|\boldsymbol{\epsilon}_{t-i}\boldsymbol{\epsilon}_{t-j} - \mathbb{E}\boldsymbol{\epsilon}_{t-i}\boldsymbol{\epsilon}_{t-j}\right| |\Theta_j|.
\end{aligned}$$

Therefore,

$$[\mathbb{E}\left|\boldsymbol{w}_t\boldsymbol{w}_t' - \mathbb{E}[\boldsymbol{w}_t\boldsymbol{w}_t']\right|]_{i,j} \leq c_{i,j}(t) \sum_{k=r}^{\infty} |\theta_{i,k}| \sum_{k=r}^{\infty} |\theta_{j,k}| \leq c_{i,j}(t)\zeta_{i,m}\zeta_{j,r},$$

where $c_{i,j}(t) = \max_{r,s\geq r} \mathbb{E}\left|\epsilon_{i,t-r}\epsilon_{j,t-s} - \mathbb{E}(\epsilon_{i,t-r}\epsilon_{j,t-s})\right|$.

(2) Write $\boldsymbol{x}_{t-k} = \boldsymbol{v}_t(k) + \boldsymbol{w}_t(k)$ where $\boldsymbol{v}_t(k) = \sum_{j=k}^{r-1} \boldsymbol{\Theta}_{j-k}\boldsymbol{\epsilon}_{t-j}$ and $\boldsymbol{w}_t(k) = \sum_{j=r}^{\infty} \boldsymbol{\Theta}_{j-k}\boldsymbol{\epsilon}_{t-j}$. It follows that for all $r \geq k$, $w_t(k)$ is $\mathcal{F}_{\epsilon,t-r}$-measurable. Applying the same rational used in the first part of the proof, we find that for $r \geq k$

$$\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_{t-k}'|\mathcal{F}_{\epsilon,t-r}] - \mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_{t-k}'] = \boldsymbol{w}_t(0)\boldsymbol{w}_t(k)' - \mathbb{E}[\boldsymbol{w}_t(0)\boldsymbol{w}_t'(k)].$$

Moreover,

$$\mathbb{E}\left|\boldsymbol{w}_t(0)\boldsymbol{w}_t(k)' - \mathbb{E}[\boldsymbol{w}_t(0)\boldsymbol{w}_t'(k)]\right| \leq c_{i,j}(t)\zeta_{i,r}\zeta_{i,r-k},$$

where $c_{i,j}(t) = \max_{r,s\geq r} \mathbb{E}\left|\epsilon_{i,t-r}\epsilon_{j,t-s} - \mathbb{E}(\epsilon_{i,t-r}\epsilon_{j,t-s})\right|$.

(3) Follows after Lemma 10 and the Cauchy-Schwarz inequality.

(4) Write

$$\mathbb{E}[|xy|I(|xy| > c)] \leq (\mathbb{E}|x|^4\mathbb{E}|y|^4)^{1/4}\left[P(|x| > c^{1/2}) + P(|y| > c^{1/2})\right]^{1/2}.$$

Now, apply this inequality with $x = x_{it}$ and $y = x_{jt}$, and use the same arguments of Lemma 9 Part (2). Finally, combine the exponential bounds. □

References

Audrino, F. and Camponovo, L.: 2013, Oracle properties and finite sample inference of the adaptive lasso for time series regression models, *Discussion paper*, University of St. Gallen.

Audrino, F. and Knaus, S.: 2012, Lassoing the har model: A model selection perspective on realized volatility dynamics, *Discussion Paper 2012-24*, University of St. Gallen.

Bai, J. and Ng, S.: 2002, Determine the number of factors in approximate factor models, *Econometrica* **70**, 191–221.

Bai, J. and Ng, S.: 2008, Forecasting economic time series using targeted predictors, *Journal of Econometrics* **146**, 304–317.

Barigozzi, M. and Brownlees, C.: 2013, NETS: Network estimation for time series, *Working paper*, Pompeu Fabra University.

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C.: 2012, Sparse models and methods for instrumental regression with an application to eminent domain, *Econometrica* **80**, 2369–2430.

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

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

Bülhmann, P. and van der Geer, S.: 2011, *Statistics for High Dimensional Data*, Springer.

Callot, L., Kock, A. and Medeiros, M.: 2014, Estimation and forecasting of large realized covariance matrices and portfolio choice, *Discussion paper*, CREATES - Aarhus University.

Cheng, X. and Hansen, B.: 2012, Forecasting with factor-augmented regression: A frequentist model averaging approach, *Working Paper 12-046*, PIER.

Cheng, X., Liao, Z. and Schorfheide, F.: 2013, Shrinkage estimation of dynamic factor models with structural instabilities, *Manuscript*, Department of Economics, University of Pennsylvania.

Corsi, F.: 2009, A simple long memory model of realized volatility, *Journal of Financial Econometrics* **7**, 174–196.

Davidson, J.: 1994, *Stochastic Limit Theory*, Oxford University Press, Oxford.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R.: 2004, Least angle regression, *The Annals of Statistics* **32**(2), 407–499.

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

Fan, Y., Fan, J. and Barut, E.: 2012, Adaptive robust variable selection, *Discussion paper*, Princeton University.

He, C. and Teräsvirta, T.: 1999, Properties of moments of a family of GARCH processes, *Journal of Econometrics* **92**, 173–192.

Hsu, N., Hung, H. and Chang, Y.: 2008, Subset selection for vector autoregressive processes using lasso, *Computational Statistics & Data Analysis* **52**(7), 3645–3657.

Huang, J., Ma, S. and Zhang, C.-H.: 2008, Adaptive lasso for sparse high-dimensional regression models, *Statistica Sinica* **18**, 1603–1618.

Issler, J. and Lima, L.: 2009, A panel-data approach to economic forecasting: The bias-corrected average forecast, *Journal of Econometrics* **152**, 153–164.

Jiang, W.: 2009, On uniform deviations of general empirical risks with unboundedness, dependence and high dimensionality, *Journal of Machine Learning Research* **10**, 977–996.

Jurado, K., Ludvigson, S. and Ng, S.: 2013, Measuring uncertainty, *Working paper*, New York University.

Kierzkowski, J. and Smoktunowicz, A.: 2011, Block normal matrices and gershgorin-type discs, *Electronic Journal of Linear Algebra* **22**(1), 69.

Kock, A.: 2012, Consistent and conservative model selection in stationary and non-stationary autoregressions, *Research Paper 05*, CREATES, Aarhus University.

Kock, A. and Callot, L.: 2012, Oracle inequalities for high dimensional vector autoregressions, *Research Paper 12*, CREATES, Aarhus University.

Lam, C. and Souza, P.: 2014a, Detection and estimation of block structure in spatial weight matrix, *Econometric Reviews* . forthcoming.

Lam, C. and Souza, P.: 2014b, Regularization for high dimensional spatial models using the adaptive lasso, *Discussion paper*, London School of Economics.

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

Leeb, H. and Pötscher, B.: 2009, On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding., *Journal of Multivariate Analysis* **100**, 1065–2082.

Ling, S. and McAleer, M.: 2002, Stationarity and the existence of moments of a family of GARCH processes, *Journal of Econometrics* **106**, 109–117.

Ludvigson, S. and Ng, S.: 2009, Macro factors in bond risk premia, *The Review of Financial Studies* **22**, 5027–5067.

Lütkepohl, H.: 2007, *New Introduction to Multiple Time Series Analysis*, Springer Publishing Company, Incorporated.

Medeiros, M. C. and Mendes, E. F.: 2015, Adaptive lasso estimation for ARDL(p,q) models with GARCH innovations.

Meinshausen, N. and Yu, B.: 2009, Lasso-type recovery of sparse representations for high dimensional data, *The Annals of Statistics* **37**, 246–270.

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

Rapach, D., Strauss, J. and Zhou, G.: 2010, Out-of-sample equity premium prediction: Consistently beating the historical average, *Review of Financial Studies* **23**, 821–862.

Rech, G., Teräsvirta, T. and Tschernig, R.: 2001, A simple variable selection technique for nonlinear models, *Communications in Statistics, Theory and Methods* **30**.

Samuels, J. and Sekkel, R.: 2013, Forecasting with many models: Model confidence sets and forecast combination, *Working Paper 2013-11*, Bank of Canada.

Song, S. and Bickel, P. J.: 2011, Large vector autoregressions, *ArXiv e-prints* .

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

Stock, J. and Watson, M.: 2002b, Macroeconomic forecasting using diffusion indexes, *Journal of Business and Economic Statistics* **20**, 147–162.

Stock, J. and Watson, M.: 2012, Generalized shrinkage methods for forecasting using many predictors, *Journal of Business and Economic Statistics* **30**, 481–449.

Tibshirani, R.: 1996, Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Van De Geer, S. A. and Bühlmann, P.: 2009, On the conditions used to prove oracle results for the lasso, *Electronic Journal of Statistics* **3**, 1360–1392.

Wang, H., Li, G. and Tsai, C.: 2007a, Regression coefficient and autoregressive order shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B(Statistical Methodology)* **69**(1), 63–78.

Wang, H., Li, G. and Tsai, C.: 2007b, Regression coefficients and autoregressive order shrinkage and selection via the Lasso, *Journal of Royal Statistical Society, Series B* **69**, 63–78.

Zhang, Y., Li, R. and Tsai, C.-L.: 2010, Regularization parameter selections via generalized information criterion, *Journal of the American Statistical Association* **105**, 312–323.

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

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

Zou, H. and Hastie, T.: 2005, Regularization and variable selection via the elastic net, *Journal of Royal Statistical Society, Series B* **67**, 301–320.

Zou, H., Hastie, T. and Tibshirani, R.: 2007, On the degrees of freedom of the lasso, *Annals of Statistics* **35**, 2173–2192.

Departamento de Economia PUC-Rio
Pontifícia Universidade Católica do Rio de Janeiro
Rua Marques de Sâo Vicente 225 - Rio de Janeiro 22453-900, RJ
Tel.(21) 31141078 Fax (21) 31141084
www.econ.puc-rio.br
flavia@econ.puc-rio.br