ESTIMATING HIGH-DIMENSIONAL TIME SERIES MODELS

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 Mendes
Department of Economics
Australian School of Business
University of New South Wales
E-mail: dudafmendes@gmail.com

Creation date: 2012-09-04.

JEL: C22.

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

Acknowledgements: The authors would like to thank Anders Kock, Laurent Callot, Marcelo Fernandes, Emmanuel Guerre, Wenxin Jiang, Martin Tanner, Eric Hillebrand, Asger Lunde, and Thiago Ferreira for insightful discussions. Marcelo C. Medeiros acknowledges support from CREATES funded by the Danish National Research Foundation and partial support from CNPq/Brazil. 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.

Abstract: We study the asymptotic properties of the Adaptive LASSO (adaLASSO) in sparse, high-dimensional, linear time-series models. We assume both the number of covariates in the model and candidate variables can increase with the number of observations and the number of candidate variables is, possibly, 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. A simulation study shows the method performs well in very general settings. Finally, we consider two applications: in the first one the goal is to forecast quarterly US inflation one-step ahead, and in the second we are interested in the excess return of the S&P 500 index. The method used outperforms the usual benchmarks in the literature.

1. INTRODUCTION

We consider variable selection and parameter estimation in single-equation linear time-series models in high dimension and when the errors are possibly non-Gaussian and conditionally heteroskedastic. We focus on the case of penalized least squares estimation.

Traditionally, one chooses the set of explanatory variables using an information criterium 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 in the sense of Fan and Li (2001): the method both selects the correct subset of non-negligible variables and the estimates of non-zero parameters have the same asymptotic distribution as the ordinary least squares (OLS) estimator in a regression including only the relevant variables. Zou (2006) proposes the adaLASSO to amend these deficiencies. In their original framework, the number of candidate variables is smaller than the sample size, the number of relevant covariates is fixed, and the results are

derived for a fixed design regression with independent and identically distributed (iid) errors. Huang et al. (2008) extend these results to a high-dimensional framework with iid errors.

In this paper we demonstrate that the adaLASSO can be applied to time-series models in a framework more general than the one currently available. The main contribution is to allow the errors to be non-Gaussian, conditionally heteroskedastic, and possibly time-dependent. This is of great importance when financial or macroeconomic data are considered. We also allow the number of variables (candidate and relevant ones) to increase as a function of the sample size. Furthermore, the number of candidate covariates can be much larger than the number of observations. We show that the adaLASSO asymptotically chooses the most parsimonious model and enjoys the oracle property. These findings allow the adaLASSO to be applied in general time-series setup, which is of interest in financial and econometric modeling. Our theoretical results are illustrated in a simulation experiment as well as in two economic applications. In the first one we consider quarterly US inflation forecasting using many predictors and in the second one we apply the adaLASSO to estimate predictive regressions for the S&P500 equity premium. 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) 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); variable selection in non-linear models (Rech et al. 2001); forecast combination of many forecasters (Issler and Lima 2009). Finally, instrumental variable estimation in a data rich environment is also a potential application; see Belloni et al. (2010).

Most advances in the LASSO 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. (2007) and Hsu et al. (2008). All these authors consider only the case where the number of candidate variables are smaller than the sample size (T). Nardi and Rinaldo (2011) consider the estimation of high-dimensional autoregressive (AR) models. However, their work differs from ours in many directions. Firstly, they assume an AR model that does not include exogenous variables. Second, they require a much stronger set of assumptions that we do and some of them may be violated in a time-series context. Moreover, they assume the error term to be independent and normally distributed. Song and Bickel (2011) and Kock and Callot (2012) studied the estimation of vector AR (VAR) models. The former paper considered LASSO and group-LASSO for estimating VARs where the number of candidate variables increases with the sample size. However, the number of relevant variables is fixed. Kock and Callot (2012) relax this assumption but assume the errors to be independent and normally distributed. Although, our model is nested in their VAR specification, we show the oracle property with a more general error term. Finally, Kock (2012) considered adaLASSO estimation in stationary and non-stationary AR models with a fixed number of variables.

The paper is organized as follows. In Section 2 we introduce the notation and assumptions. In Section 3 we present the main results. In Section 4 we present simulation results. In Section 5 the real applications are presented. Finally, Section 6 concludes. The proofs are postponed to the appendix.

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 weak-stationary high-dimensional n_T-vector of covariates, possibly containing lags of y_t, and u_t is a zero-mean weak-stationary error term uncorrelated with $\boldsymbol{x}_t$. 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 handful of elements of $\boldsymbol{\theta}$ are 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. Another possibility is to consider $\boldsymbol{x}_t$ as a set of individual forecasts, in which equation (1) represents a forecast combination problem.

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

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|Y - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda \sum_{j=1}^{p} w_j |\theta_j|, \tag{2}$$

where $\boldsymbol{Y} = (y_1, \ldots, y_T)'$, $\boldsymbol{X}$ is the $(T \times n_T)$ data matrix, $w_j = |\widehat{\theta}_j^*|^{-\tau}$, $\tau > 0$, and $\widehat{\theta}_j^*$ is an initial parameter estimate. When $w_j = 1, \forall\, j$, (2) becomes the usual LASSO.

The number of candidate covariates is $n \equiv n_T$, the number of non-zero parameters is $q \equiv q_T$ and the number of zeroes is $m \equiv m_T$. 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 $(T \times q)$ partition with the relevant variables and $\boldsymbol{X}(2)$ is the $(T \times m)$ partition with the irrelevant ones. Write $\boldsymbol{\theta} = [\boldsymbol{\theta}(1)', \boldsymbol{\theta}(2)']'$ where $\boldsymbol{\theta}(1) \in \mathbb{R}^q$ and $\boldsymbol{\theta}(2) \in \mathbb{R}^m$. $\boldsymbol{\theta}_0$ is the *true* parameter, where $\boldsymbol{\theta}_0 = [\boldsymbol{\theta}_0(1)', \boldsymbol{0}']'$, with $\boldsymbol{\theta}_0(1) \neq \boldsymbol{0}$.

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 consistency when each $w_j = 1$ 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 parameter. Zou (2006) suggest using the inverse of the ordinary least squares estimator of the parameters as the weight. However, such estimator is not available when the number of candidate variables is larger than the number of observations. Ridge regression can be used as a initial estimator in this case. 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. This assumption is weaker than the existence of the OLS estimator, but still too strong in a time series framework. In this paper we use a weaker condition, denoted *Weighted Irrepresentable Condition* (WIC) (van der Geer and Bühlmann 2011).

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 weak-stationary process.*

(2) $\mathbb{E}[u_t|\boldsymbol{x}_t] = 0$.

(3) For some finite, positive constant c_d *and some* $d \geq 1$,

$$\max_{j=1,\ldots,n} \mathbb{E}\left|\tfrac{1}{\sqrt{T}}\textstyle\sum_{t=1}^{T} x_{jt}u_t\right|^{2d} \leq c_d.$$

Assumptions DGP(1) and DGP(2) are classical conditions in the the time series regression framework. Assumption DGP(3) is satisfied by a large number of distinct data generating processes. For instance, define $v_{jt} = u_t x_{jt}$ for $j = 1, \ldots, n$ and verify that $\mathbb{E}[v_{jt}] = 0$. If each $\{v_{jt}\}$ is a martingale difference sequence, one may apply the Burkholder-Davis-Gundy inequality (see, e.g., Davidson 1994, thm. 15.18) to derive the upper bound

$$\max_{j=1,\ldots,n} \mathbb{E}\left|T^{-1/2}\sum_{t=1}^{T} v_{jt}\right|^{2d} \leq c \max_{j=1,\ldots,n} \mathbb{E}\left|T^{-1}\sum_{t=1}^{T} v_{jt}^2\right|^{d},$$

for some constant c. A similar upper bound can be derived if every process $\{v_{jt}\}$, for $j = 1, \ldots, n$, satisfy the conditions of a Marcinkiewicz-Zygmund type inequality for dependent processes (see, e.g., Dedecker et al. 2007, sec. 4.3.1). Finally, Assumption DGP(3) allows for conditional heteroskedasticity such, as for example, GARCH models.

Next assumption controls the lower bound of the non-zero parameters.

Assumption (PARAM)**.** *The next conditions hold jointly.*

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

(2) There exists a constant θ_* such that $\min_{1\leq j\leq q} |\theta_{0j}| \geq \theta_*/q$.

We assume that the smallest value of a non-zero parameter is proportional to q, such that it can be as close to 0 as $q \to \infty$. This requirement is milder than the beta-min condition in the literature in which, for all $j = 1, \ldots, q$, $\theta_{0j} \geq \theta_* > 0$ for a fixed θ_* independent of T.

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 $\boldsymbol{s}_0 = \operatorname{sgn}(\boldsymbol{\theta}_0(1))$, where $s_{0j} = 1$ if $\theta_{0j} > 0$, $s_{0j} = 0$ if $\theta_{0j} = 0$, and $s_{0j} = -1$ if $\theta_{0j} < 0$. Let $\boldsymbol{W}(1) = \operatorname{diag}(w_1, \ldots, w_q)$.

Assumption (WIC)**.** *For every $j = q+1, \ldots, n$, and some $0 < \xi < (1 \wedge \tau)$, there exists a sufficiently small $\eta > 0$ satisfying,*

$$P\left(\left[T^{-\xi/2}|\widehat{\boldsymbol{\Omega}}_{21}\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{s}_0|\right]_j \leq T^{-\xi/2}w_j - \eta\right) \to 1 \quad \textit{as} \quad t \to \infty, \tag{3}$$

where $[\cdot]_j$ refers to the j^{th} element of the vector inside brackets.

In most settings it is not straightforward to show whether the WIC is satisfied. However, a simpler set of sufficient conditions can be easily derived:

Proposition 1. *Denote $\hat{\delta}_*$ the smallest eigenvalue of $\widehat{\mathbf{\Omega}}_{11}$, $\widehat{\sigma}_j$ the sample standard deviation of x_{jt}, for $j = 1, \ldots, n$, and $\hat{\rho}_{ij}$ the sample correlation between x_{it} and x_{jt}, for $i = 1, \ldots, q$ and $j = q+1, \ldots, n$. If*

$S1$. $P(\hat{\delta}_* < c_\delta\, q^{-1}) \to 0$ *as* $T \to \infty$*;*

$S2$. $P(\max_{i=1,\ldots,q} w_i > c_w\, q^{\tau}) \to 0$ *as* $T \to \infty$*; and*

$S3$. $P(\max_{j=q+1,\ldots,n} \sum_{i=1}^{q} |\hat{\rho}_{ij}|\widehat{\sigma}_i > c_\rho\, q^{\gamma}) \to 0$ *as* $T \to \infty$*, for* $0 \leq \gamma \leq 1$*, then*

$$P\left(\left[T^{-\xi/2}|\widehat{\mathbf{\Omega}}_{21}\widehat{\mathbf{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{s}_0|\right]_j \leq T^{-\xi/2}w_j - \eta\right) \geq P\left(\min_{j=q+1,\ldots,n} \frac{T^{-\xi/2}w_j}{\widehat{\sigma}_j} \geq c\,T^{-\xi/2}q^{1+\gamma+\tau} + \eta^*\right), \quad (4)$$

for some constant $c \geq c_w c_\rho / c_\delta$, and $\eta^* > \max_{j=q+1,\ldots,n} \eta/\widehat{\sigma}_j$.

Proposition 1 guarantees that under $S1$, $S2$ and $S3$, the condition

$$P\left(\min_{j=q+1,\ldots,n} \frac{T^{-\xi/2}w_j}{\widehat{\sigma}_j} \geq c\,T^{-\xi/2}q^{1+\gamma+\tau} + \eta^*\right) \to 1, \quad (5)$$

implies the WIC. It can be inferred from (5) that we do not need a zero-consistent estimator, but estimates for the weights that satisfy the previous conditions. Biased estimators of the redundant parameters can satisfy this condition if the bias is small enough. When q increases with T, the weights of the redundant variables may increase accordingly.

Following Zhao and Yu (2006), model selection consistency is equivalent to *sign consistency*.

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

$$P\left(\operatorname{sgn}(\widehat{\boldsymbol{\theta}}) = \operatorname{sgn}(\boldsymbol{\theta})\right) \to 1, \text{ element-wise as } T \to \infty.$$

Next proposition (equivalent to Proposition 1 in Huang et al. (2008)) provides a lower bound on the probability of the adaLASSO choosing the correct model.

Proposition 2. *Let* $\boldsymbol{W}(1) = \text{diag}(w_1, \ldots, w_q)$, $\boldsymbol{W}(2) = \text{diag}(w_{q+1}, \ldots, w_n)$, *and* $\boldsymbol{s}_0 = \text{sgn}(\boldsymbol{\theta}_0(1))$. *Then*

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

where

$$\mathcal{A}_T = \left\{ \frac{1}{\sqrt{T}} |\widehat{\boldsymbol{\Omega}}_{11} \boldsymbol{X}(1)' U| < \sqrt{T} |\boldsymbol{\theta}_0| - \frac{1}{2} \lambda \frac{1}{\sqrt{T}} |\widehat{\boldsymbol{\Omega}}_{11}^{-1} \boldsymbol{W}(1) \boldsymbol{s}_0| \right\}, \tag{6a}$$

$$\mathcal{B}_T = \left\{ 2 |\frac{1}{\sqrt{T}} \boldsymbol{X}(2)' \boldsymbol{M}(1) U| < \frac{1}{\sqrt{T}} \lambda \left(\boldsymbol{W}(2) \mathbf{1}_m - |\widehat{\boldsymbol{\Omega}}_{21} \widehat{\boldsymbol{\Omega}}_{11}^{-1} \boldsymbol{W}(1) \boldsymbol{s}_0| \right) \right\}, \tag{6b}$$

where $\boldsymbol{U} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}_0$, $\boldsymbol{M}(1) = \boldsymbol{I}_T - \boldsymbol{X}(1)(\boldsymbol{X}(1)'\boldsymbol{X}(1))^{-1}\boldsymbol{X}(1)'$, *and the previous inequalities hold element-wise.*

Events $\mathcal{A}_T$ and $\mathcal{B}_T$ follow from the Karush-Kuhn-Tucker conditions. We can understand the event $\mathcal{A}_T$ as "including relevant variables in the model" and $\mathcal{B}_T$ as "keeping irrelevant variables outside the model". It is straightforward to see that the WIC plays a role in equation (6b), meaning that even when this condition is violated, the adaLASSO can still capture the relevant features of the model. It is also easy to see why the WIC is a necessary condition: if WIC does not hold, then $P(\mathcal{B}_T) \to 0$.

3. Main Results

In this section we present the main results of the paper: model selection consistency and oracle property. We first present a set of technical assumptions controlling the order of q and m, the size of the weights $w_1, \ldots, w_q$, and the regularization parameter λ.

Assumption (REG)**.** *Let* λ, m, q, *and* $T \to \infty$ *such that*

$R1.$ $\left[T^{(1-\xi)/2} \left(m^{1/d} \vee q\right)\right] / \lambda \to 0$ *and* $\lambda / \sqrt{T} \to 0$

R2. Denote δ_* the smallest eigenvalue of $\widehat{\mathbf{\Omega}}_{11}$. There exists a positive, non-increasing, sequence δ_q, indexed by q, such that $P(\delta_* < \delta_q) \to 0$ as $T \to \infty$.

R3. There exists a positive, non-decreasing, sequence l_q, indexed by q, such that

$$P\left(\max_{j=1,\ldots,q} w_j > l_q\right) \to 0 \text{ as } T \to \infty.$$

R4. $\left(q^{1+1/2d} \vee \lambda l_q\right) q^{1/2}/(\sqrt{T}\,\delta_q) \to 0$ as $T \to \infty$.

Assumption R1 controls the number of candidate variables and is similar to the one employed in Huang et al. (2008) and Huang et al. (2009). Assumption R2 controls the size of the smallest eigenvalue of the estimated sample covariance matrix $\widehat{\mathbf{\Omega}}_{11}$. We allow the size of the eigenvalues to decrease as the number of the relevant variables increases, which is weaker than the *fixed* lower bound adopted in the literature. Assumption R3 defines an upper bound on the weights $w_1, \ldots, w_q$. Assumption R4 controls the relationship among l_q, δ_q, q and T. By combining the previous restrictions, one can see that the number of relevant variables can increase polynomially with T, depending on d in DGP3.

For instance, take $\lambda = T^{1/2-\xi/4}$ and, for now, assume that (i) $m^{1/d} > q$. Assumption R1 is satisfied with $m = o(T^{d\,\xi/2})$. As in proposition 1, we satisfy R2 and R3 by taking $\delta_q = q^{-1}$ and $l_q = q^{\tau}$. Condition R4 is satisfied if (ii) $q^{1+1/2d-\tau} < \lambda$ and (iii) $T^{-1/2}\lambda q^{\tau+3/2} \to 0$. Choose $\tau \geq (d+1)/2d$ and note that (iii) is satisfied if $q/(T^{1/(2\tau+3)\,\xi/2}) \to 0$. Then, substituting τ by $(d+1)/2d$, we have for $\xi = 1/2$ that the choice $m = O((T/\log T)^{d/4})$ and $q = O((T/\log T)^{d/4(4d+1)})$ satisfy REG. It follows trivially that (i) and (ii) hold for any $d \geq 1$.

Theorem 1. *Under assumptions DGP, PARAM, WIC and REG*

$$P\left(\operatorname{sgn}(\widehat{\boldsymbol{\theta}}) = \operatorname{sgn}(\boldsymbol{\theta}_0)\right) \to 1, \text{ as } T \to \infty.$$

In Theorem 2 we show that the adaLASSO estimator for time-series possess the oracle property, in a sense that converges to the same distribution as the OLS estimator as $T \to \infty$. The major 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 $\boldsymbol{\theta}_0(1)$. Then, under assumptions DGP, PARAM, WIC and REG, and for some q-dimensional vector $\boldsymbol{\alpha}$ with Euclidean norm 1, we have*

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

It follows from Proposition 2 that if one takes $l_q = c_w q^{\tau}$, $\delta_q = c_\delta q^{-1}$, and replace the WIC by (5), the previous results hold. The following corollary states this result.

Corollary 1. *Under* (5) *and Assumptions DGP, PARAM, $S1$-$S3$, $R1$, and $R4$, the results of Theorems 1 and 2 hold.*

3.1. **Selection of λ and τ.** The selection of the regularization parameter λ and the weighting parameter τ is critical. Traditionally, one employs cross-validation and selects the pair (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 pair (λ, τ) using information criteria, such as the Bayesian Information Criterion (BIC). Zou et al. (2007), Wang et al. (2007) 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. (2007) 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. 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 remarkably well in Monte Carlo simulations presented in the next section.

4. SIMULATION

Consider the following data generating process (DGP):

$$y_t = \phi y_{t-1} + \boldsymbol{\beta}' \boldsymbol{x}_{t-1}(1) + u_t, \tag{7}$$

$$u_t = h_t^{1/2} \varepsilon_t,\ \varepsilon_t \overset{\text{iid}}{\sim} \mathsf{t}^*(5) \tag{8}$$

$$h_t = 5 \times 10^{-4} + 0.9 h_{t-1} + 0.05 u_{t-1}^2 \tag{9}$$

$$\boldsymbol{x}_t = \begin{bmatrix} \boldsymbol{x}_{t-1}(1) \\ \boldsymbol{x}_{t-1}(2) \end{bmatrix} = f_t + e_t,\ e_t \overset{\text{iid}}{\sim} \mathsf{t}^*(5), \text{and} \tag{10}$$

$$f_t = 0.8 f_{t-1} + v_t,\ v_t \overset{\text{iid}}{\sim} \mathsf{t}^*(5), \tag{11}$$

where $\phi = 0.7$ and $\boldsymbol{\beta}$ is a vector of ones. The dependent y_t follows an autoregressive distributed lag (ARDL) model with non-Gaussian GARCH errors. $\boldsymbol{x}_t(1)$ is a $(q-1) \times 1$ vector of included (relevant) variables. The vector $\boldsymbol{x}_t = [\boldsymbol{x}_t(1)', \boldsymbol{x}_t(2)']' \in \mathbb{R}^{(n-1)}$, has $n - q$ irrelevant variables and follows a factor model with a single factor. The factor itself follows a first-order AR process. All the errors are serially uncorrelated and t-distributed with 5 degrees of freedom. Furthermore, ε_t, e_t, v_t are mutually not correlated. $\mathsf{t}^*(5)$ 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 $\boldsymbol{w}_t = (y_{t-1}, \boldsymbol{x}_{t-1}')'$. Note that this is a very adverse setting as the errors are not normal and are conditionally heteroskedastic. Furthermore, the candidate variables are all highly correlated, Corr$(x_{it}, x_{j,t}) = 0.83, \forall i \neq j$.

We simulate $T = 50, 100, 300, 500$ observations of DGP (7)–(11) for different combinations of candidate (n) and relevant (q) variables. We consider $n = 100, 300, 1000$ and $q = 5, 10, 15, 25$. The models are estimated by the adaLASSO method and the values of λ and τ are selected by the BIC.

We start by analyzing the properties of the estimators for the parameter α in (7). Figures 1–4 illustrates the distribution of the bias for 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 $q = 5$, the distribution of the adaLASSO estimator is very close to the distribution of the oracle. For the other values of q, the adaLASSO distribution presents fat-tails cause mainly by some outliers in the estimation. For $T = 100$, the adaLASSO distribution is closer to the oracle one when $q = 5$ or $q = 10$. However, there still outliers. When $T = 300$ the number of outliers reduces and the adaLASSO distribution gets closer to the oracle, specially for $q = 5$ or $q = 10$. For $q = 15$ or $q = 20$, the bias is or order $O(10^{-3})$. The same pattern is observed when $T = 500$.

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.,

$$\mathsf{Bias} = \frac{1}{1000n}\sum_{j=1}^{1000}\left(\left|\widehat{\phi} - 0.7\right| + \sum_{i=1}^{n-1}\left|\widehat{\beta}_i - 1\right|\right) \text{ and}$$

$$\mathsf{MSE} = \frac{1}{1000n}\sum_{j=1}^{1000}\left[\left(\widehat{\phi} - 0.7\right)^2 + \sum_{i=1}^{n-1}\left(\widehat{\beta}_i - 1\right)^2\right].$$

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.

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 $T = 500$, the true model is included almost every time. For smaller sample sizes the performance decreases dramatically as q 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.

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 q, the adaLASSO has a similar performance than the oracle. For $q = 10$ or $q = 15$, the results are reasonable only for $T = 300$ or $T = 500$. The performance of the adaLASSO also improves as the sample size increases.

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 q is the number of relevant regressors.

	$T = 50$			$T = 100$			$T = 300$			$T = 500$		
$q \backslash n$	100	300	1000	100	300	1000	100	300	1000	100	300	1000
						Panel (a): Bias						
5	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	0.020	0.025	0.025	0.001	0.001	0.001	0.001	0.000	0.000	0.001	0.000	0.000
15	0.121	0.065	0.065	0.008	0.012	0.012	0.002	0.001	0.001	0.002	0.001	0.001
20	0.213	0.097	0.097	0.049	0.051	0.051	0.005	0.003	0.003	0.004	0.002	0.002
						Panel (b): MSE						
5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	0.012	0.017	0.017	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
15	0.098	0.060	0.060	0.002	0.006	0.006	0.000	0.000	0.000	0.000	0.000	0.000
20	0.218	0.111	0.111	0.025	0.035	0.035	0.000	0.000	0.000	0.000	0.000	0.000

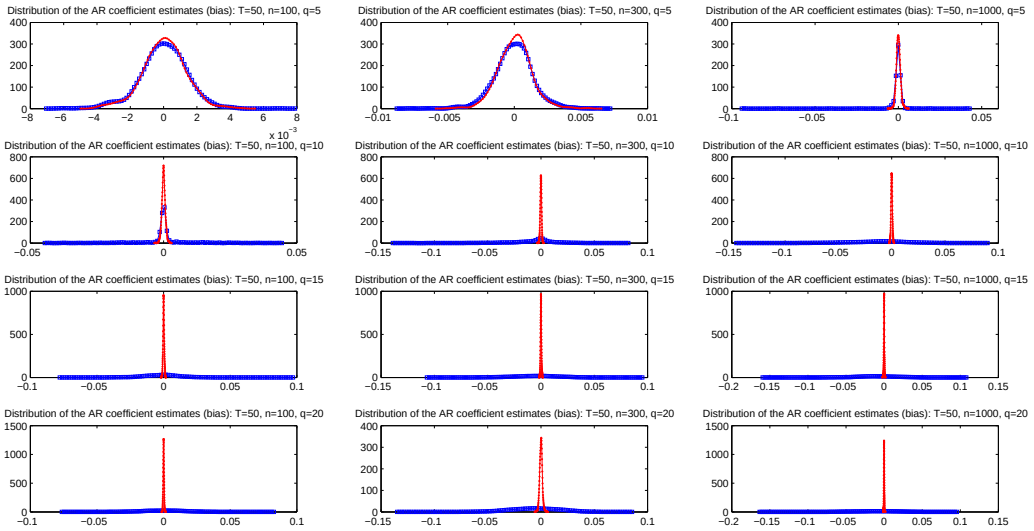


FIGURE 1. Distribution of the bias of the adaLASSO and Oracle estimators for the parameter ϕ over 1000 Monte Carlo replications. Different combinations of candidate and relevant variables. The sample size equals 50 observations.

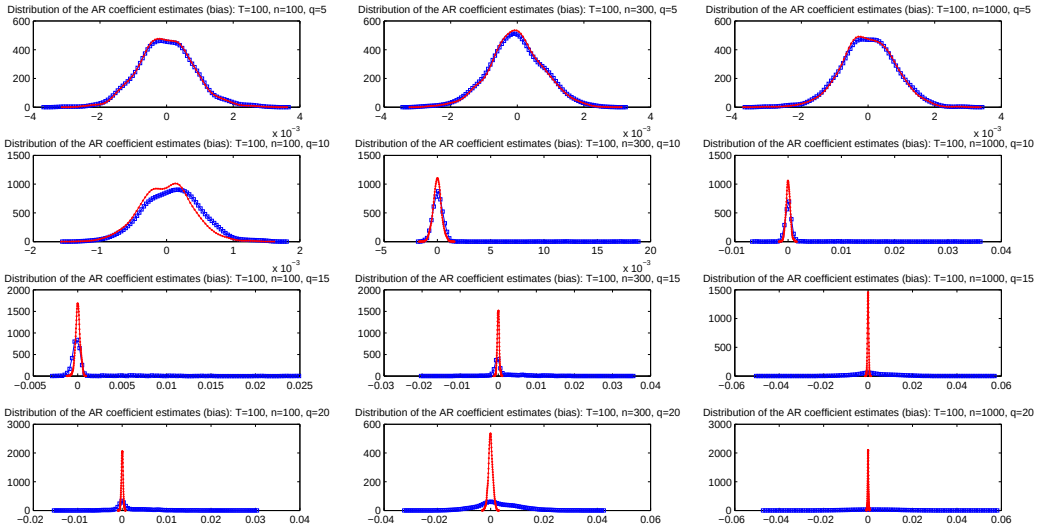


FIGURE 2. Distribution of the bias of the adaLASSO and Oracle estimators for the parameter ϕ over 1000 Monte Carlo replications. Different combinations of candidate and relevant variables. The sample size equals 100 observations.

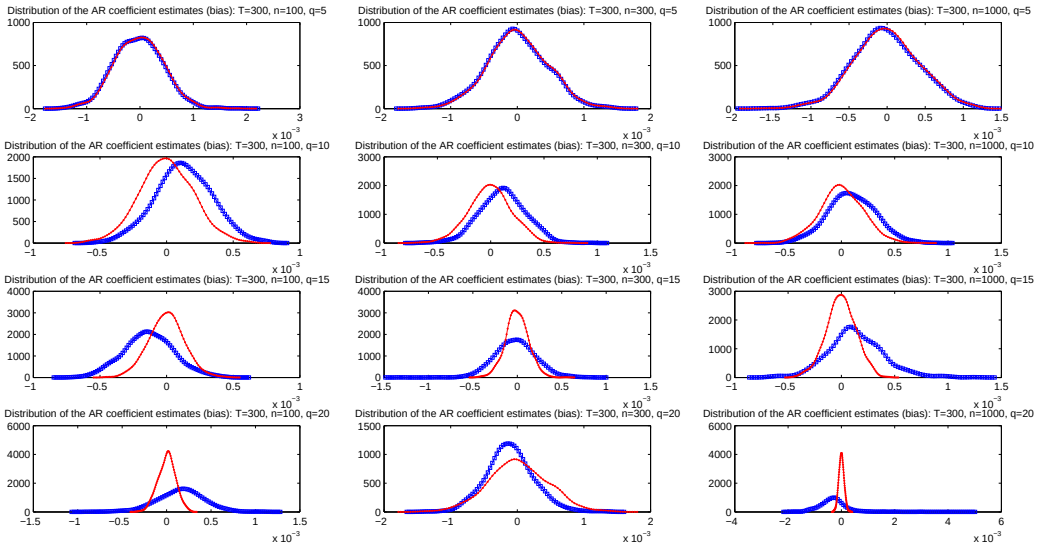


FIGURE 3. Distribution of the bias of the adaLASSO and Oracle estimators for the parameter ϕ over 1000 Monte Carlo replications. Different combinations of candidate and relevant variables. The sample size equals 300 observations.

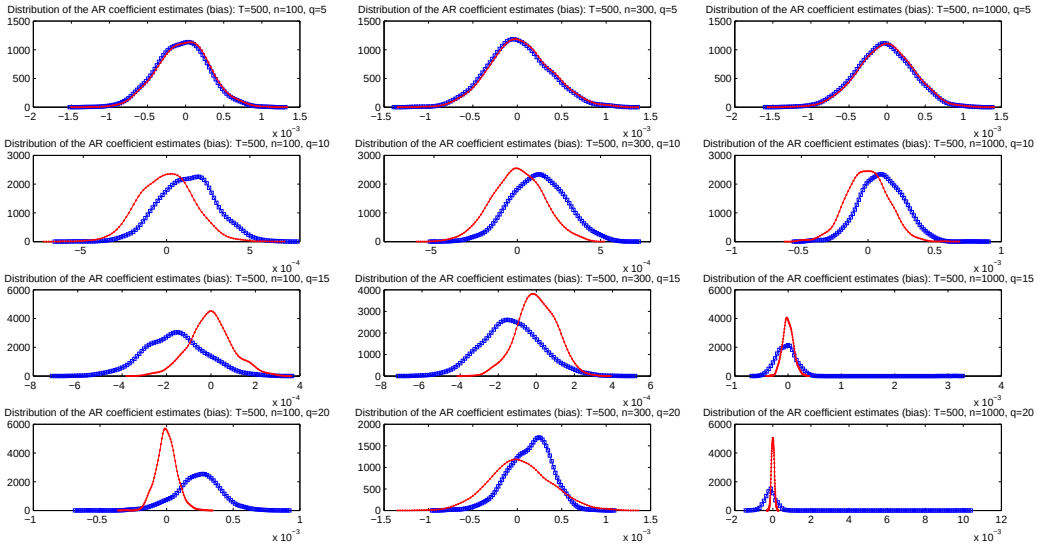


FIGURE 4. Distribution of the bias of the adaLASSO and Oracle estimators for the parameter ϕ over 1000 Monte Carlo replications. Different combinations of candidate and relevant variables. The sample size equals 500 observations.

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.

	$T=50$			$T=100$			$T=300$			$T=500$		
$q\backslash n$	100	300	1000	100	300	1000	100	300	1000	100	300	1000
						Panel (a): Correct Sparsity Pattern						
5	0.860	0.750	0.759	1	1	1	1	1	1	1	1	1
10	0.015	0.000	0.002	0.231	0.060	0.060	0.962	0.877	0.877	0.990	0.966	0.966
15	0	0	0	0.009	0	0	0.623	0.239	0.239	0.914	0.784	0.784
20	0	0	0	0	0	0	0.172	0.011	0.011	0.640	0.254	0.254
						Panel (b): True Model Included						
5	1	1	0.985	1	1	1	1	1	1	1	1	1
10	0.754	0.246	0.017	1	0.995	0.968	1	1	1	1	1	1
15	0.046	0	0	0.927	0.648	0.122	1	1	1	1	1	1
20		0	0	0.447	0.048	0	1	1	1	1	1	1
						Panel (c): Fraction of Relevant Variables Included						
5	1	1	1	1	1	1	1	1	1	1	1	1
10	0.944	0.744	0.744	1	0.999	0.999	1	1	1	1	1	1
15	0.691	0.431	0.431	0.992	0.942	0.942	1	1	1	1	1	1
20	0.524	0.300	0.300	0.929	0.730	0.730	1	1	1	1	1	1
						Panel (d): Fraction of Irrelevant Excluded						
5	0.998	0.998	0.998	1	1	1	1	1	1	1	1	1
10	0.928	0.956	0.956	0.980	0.985	0.985	0.999	0.999	0.999	0.999	0.999	0.999
15	0.884	0.948	0.948	0.924	0.948	0.948	0.994	0.994	0.994	0.999	0.999	0.999
20	0.873	0.945	0.944	0.871	0.926	0.926	0.975	0.978	0.978	0.994	0.995	0.995
						Panel (e): Number of Included Variables						
5	5.192	5.476	5.476	5.002	5.008	5.008	5	5	5	5	5	5
10	15.920	20.086	20.086	11.811	14.276	14.276	10.041	10.129	10.129	10.010	10.034	10.034
15	20.237	21.271	21.271	21.359	29.087	29.087	15.510	16.630	16.630	15.086	15.236	15.236
20	20.619	21.426	21.426	28.884	35.417	35.417	22.031	26.146	26.146	20.468	21.493	21.493
						Panel (f): Fraction of Included Irrelevant Variables						
5	0.161	0.454	0.454	0.002	0.008	0.008	0	0	0	0	0	0
10	5.774	12.181	12.181	1.598	4.150	4.150	0.037	0.121	0.121	0.009	0.033	0.033
15	9.307	14.532	14.532	6.103	14.708	14.708	0.481	1.601	1.601	0.079	0.231	0.231
20	10.143	15.429	15.429	10.298	20.815	20.815	2.031	6.146	6.146	0.468	1.493	1.493

5. APPLICATIONS

5.1. **Inflation Forecasting.** We consider quarterly inflation forecasting by many predictors. The dataset was obtained from the Federal Reserve Bank of Philadelphia and is part of the

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 q is the number of relevant regressors.

	$T=50$			$T=100$			$T=300$			$T=500$		
$q\backslash n$	100	300	1000	100	300	1000	100	300	1000	100	300	1000
						MSE - adaLASSO						
5	0.011	0.011	0.068	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010
10	1.333	6.233	12.510	0.012	0.028	0.163	0.010	0.011	0.011	0.010	0.010	0.011
15	11.516	22.391	31.890	0.237	1.988	9.843	0.013	0.015	0.024	0.013	0.013	0.129
20	26.400	43.252	55.786	2.799	11.864	26.792	0.028	0.049	0.159	0.023	0.027	1.227
						MSE - Oracle						
5	0.011	0.011	0.011	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010
10	0.012	0.012	0.012	0.011	0.011	0.011	0.010	0.010	0.010	0.010	0.010	0.010
15	0.014	0.014	0.014	0.011	0.012	0.011	0.010	0.011	0.010	0.010	0.010	0.010
20	0.017	0.017	0.017	0.012	0.012	0.012	0.011	0.010	0.010	0.010	0.010	0.010

database called "Real-Time Data Set for Macroeconomists", which consists of vintages of major macroeconomic variables. For the present work, we used only the vintage available at the third quarter of 2011, which contains data from the first quarter of 1959 and ends in the second quarter of 2011, totalling 210 observations. The dependent variable corresponds to the GDP price index and can be expressed as a ratio of nominal output and real output. There are a total of 69 variables plus one lag of inflation. The predictive regression is then written as

$$\pi_{t+1} = \phi_0 + \phi_1 \pi_t + \boldsymbol{\beta} \boldsymbol{x}_t + u_{t+1},$$

where π_t is the quarterly inflation at time t and $\boldsymbol{x}_t$ is the vector of predictors.

All variables have been pretested for unit-roots and first-differenced whenever necessary. We consider three forecasting periods starting, respectively in 1970, 1985 and 2000. An expanding window scheme is used to estimate the models recursively and to compute the one-step-ahead forecasts. We compare the adaLASSO with three different benchmark alternatives: a model with all the regressors included, a simple first-order AR model, and a factor model based on the first two principal components of the predictors. The results are shown in Table 4. It is

TABLE 4. INFLATION FORECASTING RESULTS: OUT-OF-SAMPLE R^2 ($\times 100$).

The table reports the out-of-sample R^2 multiplied by 100. Three different forecasting periods are considered. The first one starts in January 1970, the second one in January 1985, and the last one starts in January 2000.

	1970	1985	2000
Benchmark Models:			
All Regressors	25.29	19.47	29.29
AR(1)	79.54	78.09	78.60
AR(1) + PCA (two components)	76.62	69.36	73.75
LASSO and adaLASSO:			
LASSO (BIC)	86.92	88.01	90.42
adaLASSO (BIC)	85.87	88.01	90.42

clear from the table that both the LASSO and the adaLASSO models are far superior than the benchmark for all the three periods considered. Furthermore, the LASSO and adaLASSO results are almost identical.

5.2. **Equity Premium Forecasting.** Excess returns prediction has attracted academics and practitioners for many decades. In a recent paper, Goyal and Welch (2008) argued that none of the conventional predictor variables proposed in the literature seems capable of systematically predicting stock returns out-of-sample. Their empirical evidence suggests that most models were unstable or spurious, and most models are no longer significant even in-sample. However, Campbell and Thompson (2008), on the other hand, showed that many predictive regressions outperform the historical average once weak restrictions are imposed on the signs of coefficients and return forecasts. The out-of-sample explanatory advantage over the historical mean is small and usually statistically not significant, but nonetheless economically meaningful for mean-variance investors. Three recent papers corroborate the results in Campbell and Thompson (2008). Rapach et al. (2010) consider combining individual forecasts in order to attenuate the effects of model uncertainty and instability. They show, consistently over time, that simple model combination delivers statistically and economically significant out-of-sample gains

relative to the historical average. In a similar direction, Lee et al. (2008) proposed bagging estimators to reduce model instability and showed significant improvements over the historical mean. Finally, Ferreira and Santa-Clara (2011) proposed forecasting separately the three components of stock market returns: the dividend-price ratio, earnings growth, and price-earnings ratio growth.

Using the same dataset as in Goyal and Welch (2008) and Rapach et al. (2010) we apply the adaLASSO to the following monthly predictive regression:

$$r^*_{t+1} = r_{t+1} - r_{f,t+1} = \phi_0 + \phi_1 r^*_t + \boldsymbol{\theta}' \boldsymbol{x}_t + u_{t+1}, \tag{12}$$

where r^*_t represents the market returns in excess to the risk-free rate ($r_{f,t}$), $\boldsymbol{x}_{t-1}$ is a set of lagged predictors, and u_t is the error term. Stock returns are measured as continuously compounded returns on the S&P 500 index, including dividends, and the Treasury bill rate is used to compute the equity premium. With respect to the economic variables used to predict the equity premium, we consider, in addition to r^*_{t-1}, the 14 variables from Goyal and Welch (2008): Dividend-price ratio (log); dividend yield (log); earnings-price ratio (log); dividend-payout ratio (log); stock variance; book-to-market ratio; net equity expansion; treasury bill rate; long-term yield; long-term return; term spread; default yield spread; default return spread; and inflation.

We consider three different out-of-sample forecast evaluation periods: (i) a "long" out-of-sample period covering January 1965 to December 2008; (ii) a period covering the last thirty years of the full sample, January 1976 to December 2008; and (iii) a "short" forecasting period starting in January 2000. The dataset starts in January 1947. The out-of-sample R^2s for one-step-ahead forecasts are shown in Table 5. The results are impressive. The LASSO and the adaLASSO estimators are far superior that all the competitors in sample periods considered.

TABLE 5. EQUITY PREMIUM FORECASTING RESULTS: OUT-OF-SAMPLE R^2 ($\times 100$).

The table reports the out-of-sample R^2 multiplied by 100. Three different forecasting periods are considered. The first one starts in January 1965, the second one in January 1976, and the last one starts in January 2000.

	1965	1976	2000
Unrestricted Individual Predictors			
AR(1)	-0.16	-0.10	0.85
All Regressors	-0.07	-0.81	-1.45
Dividend Price Ratio	0.31	-0.82	3.80
Dividend Yield	0.40	-0.81	4.14
Earning Price Ratio	0.53	0.39	4.33
Dividend Payout Ratio	-0.51	-1.09	-0.53
Stock Variance	-0.19	0.39	4.76
Book to Market	-0.82	-0.71	1.30
Net Equity Expansion	-0.82	-0.66	-3.20
T-Bill Rate	-0.73	-2.86	-3.03
Long Term Yield	-0.76	-2.05	-0.87
Long Term Spread	0.23	-0.81	-0.93
Term Spread	-0.96	-2.59	-3.38
Default Yield Spread	-0.66	-0.66	-0.96
Default Return Spread	-0.20	-0.03	1.02
Inflation	0.72	-0.09	-3.97
Forecast Combination			
Mean	1.31	0.54	1.07
Median	0.92	0.11	0.24
Trimmed Mean	1.31	0.54	1.07
LASSO and AdaLASSO			
LASSO	7.36	6.95	13.11
adaLASSO	7.36	6.95	13.11

6. 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 features and has the oracle property in a time-series framework with a very general error term. A key ingredient is the WIC, which is necessary for sign consistency of the adaLASSO. As a technical by-product some conditions in this paper are improvements on the frequently adopted in the shrinkage literature.

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 a key difference from the factor models literature. The estimation of factors relies on the key 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 are of extreme importance. However, when the structure of the model is dense, than factor models would probably be a better alternative.

APPENDIX A. PROOFS

Proof of Proposition 1. The proof consists in showing that $\left[|\widehat{\boldsymbol{\Omega}}_{21}\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{s}_0|\right]_j \leq \sigma_j\, c_w\, q^{1+\gamma+\tau}$. Write $\left[|\widehat{\boldsymbol{\Omega}}_{21}\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{s}_0|\right]_j = |T^{-1}\boldsymbol{X}_j'\boldsymbol{X}(1)\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{s}_0|$ and $\widehat{\boldsymbol{\Omega}}_{11} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}'$, where $\boldsymbol{E}$ is a matrix of eigenvectors and $\boldsymbol{D}$ a diagonal matrix of eigenvalues. By S1, we have

$$|T^{-1}\boldsymbol{X}_j'\boldsymbol{X}(1)\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{s}_0| \leq \frac{q\sigma_j}{c_\delta}\sum_{i=1}^{q}|\hat{\rho}_{ij}|\sigma_i w_i.$$

Combining the previous equation with S2 and S3, we have

$$\frac{q\sigma_j}{c_\delta}\sum_{i=1}^{q}|\hat{\rho}_{ij}|\sigma_i\, w_i \leq \sigma_j\frac{c_w\, c_\rho}{c_\delta}q^{1+\gamma+\tau}.$$

The result follows by taking $c \geq c_w c_\rho / c_\delta$. □

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

Proof of Theorem 1. Proposition 2 provides a lower bound on the probability of selecting the correct model:

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

where $\mathcal{A}_T^c$ and $\mathcal{B}_T^c$ are the complements of $\mathcal{A}_T$ and $\mathcal{B}_T$ respectively. Therefore, to show sign consistency we have to show that $P(\mathcal{A}_T^c) \to 0$ and $P(\mathcal{B}_T^c) \to 0$ as $T \to \infty$.

Note that, under WIC,

$$\mathcal{B}_T^c \subseteq \left\{ \max_{j=q+1,\ldots,n} \left| T^{-1/2} \boldsymbol{X}_j' \boldsymbol{M}(1) U \right| > \frac{1}{2} \frac{\lambda\, \eta}{T^{(1-\xi)/2}} \right\}.$$

Denote $\tilde{\boldsymbol{\theta}}(1) = \left[\boldsymbol{X}(1)'\boldsymbol{X}(1)\right]^{-1} \boldsymbol{X}(1)'\boldsymbol{Y}$ the ordinary least squares estimator of $\boldsymbol{\theta}_0(1)$. We can bound the element on the right hand side of the inequality between brackets by

$$\begin{aligned}
|T^{-1/2} \boldsymbol{X}_j' \boldsymbol{M}(1) \boldsymbol{U}| &\leq \left| T^{-1/2} \boldsymbol{X}_j' \boldsymbol{U} \right| + \left| T^{-1/2} \boldsymbol{X}_j' \boldsymbol{P}(1) \boldsymbol{U} \right| \\
&= \left| T^{-1/2} \boldsymbol{X}_j' \boldsymbol{U} \right| + \left| T^{1/2} \boldsymbol{X}_j' \boldsymbol{X}(1) \left[\tilde{\boldsymbol{\theta}}(1) - \boldsymbol{\theta}_0(1) \right] \right| \\
&\leq \left| T^{-1/2} \boldsymbol{X}_j' \boldsymbol{U} \right| + \left| \left\{ T^{-1} \boldsymbol{X}_j' \boldsymbol{X}(1) - \mathbb{E}\left[T^{-1} \boldsymbol{X}_j' \boldsymbol{X}(1) \right] \right\} T^{1/2} \left[\tilde{\boldsymbol{\theta}}(1) - \boldsymbol{\theta}_0(1) \right] \right| \\
&\quad + \left| \mathbb{E}\left[T^{-1} \boldsymbol{X}_j' \boldsymbol{X}(1) \right] T^{1/2} \left[\tilde{\boldsymbol{\theta}}(1) - \boldsymbol{\theta}_0(1) \right] \right| \\
&= B_1 + B_2 + B_3.
\end{aligned}$$

Set $\mu_{ij} = \mathbb{E}(x_{it}x_{jt})$. Note that,

$$\begin{aligned} B_2 &= \left| \left\{ T^{-1}\boldsymbol{X}_j'\boldsymbol{X}(1) - \mathbb{E}\left[T^{-1}\boldsymbol{X}_j'\boldsymbol{X}(1)\right] \right\} T^{1/2} \left[\tilde{\boldsymbol{\theta}}(1) - \boldsymbol{\theta}_0(1) \right] \right| \\ &= \left| \sum_{i=1}^{q} \left(T^{-1} \sum_{t=1}^{T} x_{jt}x_{it} - \mu_{ij} \right) T^{1/2}(\tilde{\theta}_i - \theta_{0i}) \right| \\ &\leq o_p(1) \sum_{i=1}^{q} \left| T^{1/2} \left[\tilde{\theta}_i - \theta_{0i} \right] \right|, \end{aligned}$$

where the last line follows from the law of large numbers, as $\mathbb{E}(x_{it}x_{jt}) < \infty$, for every pair (i, j) (DGP2). Choose $\mu_2 \geq \max_{i,j} |\mu_{ij}|$, then

$$\begin{aligned} B_3 &= \left| \mathbb{E}\left[T^{-1}\boldsymbol{X}_j'\boldsymbol{X}(1)\right] T^{1/2}(\tilde{\theta}_i - \theta_{0i}) \right| \\ &= \left| \sum_{i=1}^{q} \mu_{ij} T^{1/2}(\tilde{\theta}_i - \theta_{0i}) \right| \\ &\leq \mu_2 \sum_{i=1}^{q} \left| T^{1/2}(\tilde{\theta}_i - \theta_{0i}) \right|. \end{aligned}$$

Therefore, by combining these bounds,

$$\begin{aligned} B_2 + B_3 &\leq [\mu_2 + o_p(1)] \sum_{i=1}^{q} \left| T^{1/2}(\tilde{\theta}_i - \theta_{0i}) \right| \\ &\leq [\mu_2 + o_p(1)] \sup_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=q} T^{1/2} \left| \boldsymbol{\alpha}' \left[\tilde{\boldsymbol{\theta}}(1) - \boldsymbol{\theta}_0(1) \right] \right|, \end{aligned}$$

which does not depend on $j = q+1, \ldots, n$. Thus,

$$\max_j \left| T^{-1/2}\boldsymbol{X}_j'\boldsymbol{M}(1)\boldsymbol{U} \right| \leq \max_j B_1 + [\mu_2 + o_p(1)] \sup_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=q} T^{1/2} \left| \boldsymbol{\alpha}' \left[\tilde{\boldsymbol{\theta}}(1) - \boldsymbol{\theta}_0(1) \right] \right|.$$

Define the set

$$\mathcal{C} = \left\{ (\mu_2 + o_p(1)) \sup_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=q} T^{1/2} \left| \boldsymbol{\alpha}' \left[\tilde{\boldsymbol{\theta}}(1) - \boldsymbol{\theta}_0(1) \right] \right| > \lambda\eta T^{-(1-\xi)/2}/4 \right\}.$$

Then,

$$\mathcal{B}^c \cap \mathcal{C}^c \Rightarrow \left\{ \max_{j=q+1,\ldots,n} |T^{-1/2} \boldsymbol{X}_j' \boldsymbol{U}| > \lambda\eta T^{-(1-\xi)/2}/4 \right\}.$$

Now, $\Pr(\mathcal{B}^c) \leq \Pr(\mathcal{B}^c \cap \mathcal{C}^c) + \Pr(\mathcal{C})$. We shall bound both terms on the right hand side using Markov's inequality (M), Cauchy-Schwarz inequality (CS), and the union bound (U).

$$\begin{aligned}
\Pr(\mathcal{C}) &= \Pr\left(\sup_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=q} T^{1/2} \left| \boldsymbol{\alpha}' \left[\tilde{\boldsymbol{\theta}}(1) - \boldsymbol{\theta}_0(1) \right] \right| > \frac{\lambda\eta}{T^{(1-\xi)/2} 4 \left[\mu_2 + o_p(1) \right]} \right) \\
&\overset{(M)}{\leq} 16 \left[\mu_2 + o(1) \right]^2 \frac{\mathbb{E}\left\{ \sup_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=q} \left| T^{1/2} \boldsymbol{\alpha}' \left[\tilde{\boldsymbol{\theta}}(1) - \boldsymbol{\theta}_0(1) \right] \right|^2 \right\}}{\lambda^2 \eta^2 T^{-(1-\xi)}} \\
&\overset{(CS)}{\leq} \frac{16 \left[\mu_2 + o(1) \right]^2}{\lambda^2 \eta^2 T^{-(1-\xi)}} \sup_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=q} \boldsymbol{\alpha}'\boldsymbol{\alpha} \times \sum_{i=1}^{q} \operatorname{var}[T^{1/2}(\tilde{\theta}_i - \theta_{0i})] \\
&\leq \frac{16 \left[\mu_2 + o(1) \right]^2}{\lambda^2 \eta^2 T^{-(1-\xi)}} q^2 \max_{1 \leq i \leq q} \operatorname{var} \left[T^{1/2}(\tilde{\theta}_i - \theta_{0i}) \right] \\
&= \frac{T^{1-\xi} q^2}{\lambda^2} \frac{16 \left[\mu_2 + o(1) \right]^2}{\eta^2} \max_{1 \leq i \leq q} \operatorname{var} \left[T^{1/2}(\tilde{\theta}_i - \theta_{0i}) \right] \\
&\to 0,
\end{aligned}$$

as q and $T \to \infty$ if $q\lambda / T^{(1-\xi)/2} \to 0$ (R1). For the first term on the right hand side we have

$$\begin{aligned}
\Pr(\mathcal{B}^c \cap \mathcal{C}^c) &\leq \Pr\left(\max_j |T^{-1/2} \boldsymbol{X}_j' \boldsymbol{U}| > \lambda\eta T^{-(1-\xi)/2}/4 \right) \\
&\overset{(U)}{\leq} \sum_{j=q+1}^{n} \Pr\left(|T^{-1/2} \boldsymbol{X}_j' \boldsymbol{U}| > \lambda\eta T^{-(1-\xi)/2}/4 \right) \\
&\overset{(M)}{\leq} 4^d \sum_{j=q+1}^{n} \frac{\mathbb{E}\left| T^{-1/2} \boldsymbol{X}_j' \boldsymbol{U} \right|^d}{\lambda^d \eta^d T^{-d(1-\xi)/2}} \\
&\overset{(DGP3)}{\leq} \frac{4^d c_d}{\eta^d} \frac{m T^{d(1-\xi)/2}}{\lambda^d} \\
&\to 0,
\end{aligned}$$

as m and $T \to \infty$ if $m^{1/d}T^{(1-\xi)/2}/\lambda \to 0$ (R1).

Combining these limits we have $\Pr(\mathcal{B}^c) \to 0$.

Denote $\boldsymbol{EDE}'$ the eigen-decomposition of $\widehat{\boldsymbol{\Omega}}_{11}$. Denote $\boldsymbol{\alpha}$ a $q \times 1$ non-negative vector. Under Conditions R2 and R3 we have

$$\begin{aligned}
\max_{j=1,\ldots,q}\left[\left|\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{s}_0\right|\right]_j^2 &\leq \sup_{\boldsymbol{\alpha}'\boldsymbol{\alpha}\leq 1}(\boldsymbol{\alpha}'\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{s}_0)^2 \\
&\overset{(CS)}{\leq} \sup_{\boldsymbol{\alpha}'\boldsymbol{\alpha}\leq 1}\boldsymbol{\alpha}'\widehat{\boldsymbol{\Omega}}_{11}^{-2}\boldsymbol{\alpha}\times \boldsymbol{s}_0'\boldsymbol{W}(1)^2\boldsymbol{s}_0 \\
&\overset{(R3)}{\leq} \sup_{\boldsymbol{\alpha}'\boldsymbol{\alpha}\leq 1}\boldsymbol{\alpha}'\boldsymbol{ED}^{-2}\boldsymbol{E}'\boldsymbol{\alpha}\times q\, l_q^2 \\
&\overset{(R2)}{\leq} \frac{q\, l_q^2}{\delta_q^2}
\end{aligned}$$

Therefore, $\max_{j=1,\ldots,q}\left[\left|\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{s}_0\right|\right]_j \leq \frac{q^{1/2}\, l_q}{\delta_q}$.

Applying the same reasoning and by using the Jensen's inequality (J) we have

$$\begin{aligned}
\mathbb{E}\left(\max_{j=1,\ldots,q}\left[T^{-1/2}\left|\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{X}(1)'\boldsymbol{U}\right|\right]_j\right)^2 &\leq \mathbb{E}\sup_{\boldsymbol{\alpha}'\boldsymbol{\alpha}\leq 1}T^{-1}\left(\boldsymbol{\alpha}'\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{X}(1)'\boldsymbol{U}\right)^2 \\
&\overset{(CS)}{\leq} \mathbb{E}\left[\sup_{\boldsymbol{\alpha}'\boldsymbol{\alpha}\leq 1}\boldsymbol{\alpha}'\widehat{\boldsymbol{\Omega}}_{11}^{-2}\boldsymbol{\alpha}\times q\max_{j=1,\ldots,q}(T^{-1/2}\boldsymbol{X}_j'\boldsymbol{U})^2\right] \\
&\overset{(R2)}{\leq} \delta_q^{-2}q\,\mathbb{E}\left[\max_{j=1,\ldots,q}(T^{-1/2}\boldsymbol{X}_j'\boldsymbol{U})^2\right] \\
&\overset{(J)}{\leq} \delta_q^{-2}q^{1+1/d}\max_{j=1,\ldots,q}\left(\mathbb{E}\left|T^{-1/2}\boldsymbol{X}_j'\boldsymbol{U}\right|^{2d}\right)^{1/d} \\
&\overset{(DGP3)}{\leq} \frac{q^{1+1/d}c_d^{1/d}}{\delta_q^2}.
\end{aligned}$$

Note that under DGP, PARAM, R2, R3, and R4, we have

$$\mathcal{A}_T \subseteq \left\{ \max_{j=1,\ldots,q} \left[T^{-1/2} \left| \widehat{\boldsymbol{\Omega}}_{11}^{-1} \boldsymbol{X}(1)' \boldsymbol{U} \right| \right] > \frac{\sqrt{T}\theta_*}{q} - \frac{\sqrt{q}\lambda\, l_q}{\sqrt{T}\delta_q} \right\}$$
$$\subset \left\{ \max_{j=1,\ldots,q} \left[T^{-1/2} \left| \widehat{\boldsymbol{\Omega}}_{11}^{-1} \boldsymbol{X}(1)' \boldsymbol{U} \right| \right] > \frac{\sqrt{T}\theta_*}{2q} \right\}.$$

It then follows from the Markov's inequality and R4 that

$$P(\mathcal{A}_T) \leq P\left(\max_{j=1,\ldots,q} \left[T^{-1/2} \left| \widehat{\boldsymbol{\Omega}}_{11}^{-1} \boldsymbol{X}(1)' \boldsymbol{U} \right| \right] > \frac{\sqrt{T}\theta_*}{2q} \right)$$
$$\overset{(M)}{\leq} \mathbb{E}\left(\max_{j=1,\ldots,q} \left[T^{-1/2} \left| \widehat{\boldsymbol{\Omega}}_{11}^{-1} \boldsymbol{X}(1)' \boldsymbol{U} \right| \right]_j \right)^2 \left(\frac{\sqrt{T}\theta_*}{q} \right)^{-2}$$
$$\leq \frac{4 c_d^{1/d}}{\theta_*^2} \frac{q^{3+1/q}}{T\delta_d^2} \to 0, \quad \text{as} \quad T \to \infty.$$

□

Proof of Theorem 2. Write $\dot{\boldsymbol{Q}}_T(\boldsymbol{\theta}) = -2\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}) + \lambda \boldsymbol{W} \boldsymbol{s}_\theta$, where

$$\boldsymbol{s}_\theta = (\operatorname{sgn}(\theta_1), \ldots, \operatorname{sgn}(\theta_n))'$$

and $\boldsymbol{W} = \operatorname{diag}(w_1, \ldots, w_n)$. By replacing $\boldsymbol{\theta}$ by the adaLASSO estimator and writing $\boldsymbol{U} = (\boldsymbol{Y} - \boldsymbol{X}(1)\boldsymbol{\theta}(1))$ we have that

$$\sqrt{T}(\widehat{\boldsymbol{\theta}}(1) - \boldsymbol{\theta}_0(1)) = \frac{1}{\sqrt{T}} \widehat{\boldsymbol{\Omega}}_{11}^{-1} \boldsymbol{X}(1)' \boldsymbol{U} + \frac{1}{\sqrt{T}} \widehat{\boldsymbol{\Omega}}_{11}^{-1} \widehat{\boldsymbol{\Omega}}_{12} \widehat{\boldsymbol{\theta}}(2) + \frac{\lambda}{2\sqrt{T}} \widehat{\boldsymbol{\Omega}}_{11}^{-1} \boldsymbol{W}(1)\, \boldsymbol{s}_0$$

The first term on the right hand size is, by definition, $\sqrt{T}(\hat{\boldsymbol{\theta}}_{ols}(1) - \boldsymbol{\theta}_0(1))$. It follows from the inequality $\sup_{\boldsymbol{\alpha}'\boldsymbol{\alpha} \leq 1} (\boldsymbol{\alpha}' \widehat{\boldsymbol{\Omega}}_{11}^{-1} \boldsymbol{W}(1) \boldsymbol{s}_0)^2 \leq q\, l_q^2 / \delta_q^2$ derived in the proof of Theorem 1 that the third term on the right hand side can be bounded by $\lambda\, q^{1/2}\, l_q / 2\sqrt{T}\delta_q$, which converges to 0 by

R4. Under R4 and the results in Theorem 1, an application of the Cauchy-Schwarz inequality shows that the second term on the right hand side is $o_p(1)$, i.e., $(\sqrt{T}\boldsymbol{\alpha}'\widehat{\boldsymbol{\Omega}}_{11}^{-1}\widehat{\boldsymbol{\Omega}}_{12}\widehat{\boldsymbol{\theta}}(2))^2 \leq T^{-1}(\boldsymbol{\alpha}'\widehat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{\alpha})\,(\widehat{\boldsymbol{\theta}}(2)'\widehat{\boldsymbol{\Omega}}_{22}\widehat{\boldsymbol{\theta}}(2)) = o_p(1)$. □

REFERENCES

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.

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C.: 2010, Sparse models and methods for instrumental regression, with an application to eminent domain, *Working Paper - MIT* .

Campbell, J. and Thompson, S.: 2008, Predicting the equity premium out of sample: Can anything beat the historical average?, *Review of Financial Studies* . forthcoming.

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

Dedecker, J., Doukhan, P., Lang, G., Léon, J., Louhichi, S. and Prieur, C.: 2007, *Weak dependence with examples and applications*, Springer.

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.

Ferreira, M. and Santa-Clara, P.: 2011, Forecasting stock market returns: The sum of the parts is more than the whole, *Journal of Financial Economics* **100**, 514–537.

Goyal, A. and Welch, I.: 2008, A comprehensive look at the empirical performance of equity premium prediction, *Review of Financial Studies* . forthcoming.

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., Horowitz, J. and Ma, S.: 2009, Asymptotic properties of bridge estimators in sparse high-dimensional regression models, *Annals of Statistics* **36**(2), 587–613.

Huang, J., Ma, S. and Shang, 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.

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.

Lee, T.-H., Hillebrand, E. and Medeiros, M.: 2008, Let´s do it again: Bagging equity premium predictors, *Discussion paper*, Pontifical Catholic University of Rio de Janeiro.

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**.

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.

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

van der Geer, S. and Bühlmann, P.: 2011, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Spring Series in Statistics, Springer.

Wang, H., Li, G. and Tsai, C.: 2007, 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.

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