5-2008

Autoregressive Process Modeling via the Lasso Procedure

Yuval Nardi
Carnegie Mellon University

Alessandro Rinaldo
Carnegie Mellon University, arinaldo@stat.cmu.edu

Autoregressive Process Modeling via the Lasso Procedure

Yuval Nardi*
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890 USA

Alessandro Rinaldo†
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890 USA

Abstract

The Lasso is a popular model selection and estimation procedure for linear models that enjoys nice theoretical properties. In this paper, we study the Lasso estimator for fitting autoregressive time series models. We adopt a double asymptotic framework where the maximal lag may increase with the sample size. We derive theoretical results establishing various types of consistency. In particular, we derive conditions under which the Lasso estimator for the autoregressive coefficients is model selection consistent, estimation consistent and prediction consistent. Simulation study results are reported.

1 Introduction

Classical stationary time series modeling assumes that data are a realization of a mix of autoregressive processes and moving average processes, or an ARMA model (see, e.g. Davis and Brockwell, 1991). Typically, both estimation and model fitting rely on the assumption of fixed and low dimensional parameters and include (*i*) the estimation of the appropriate coefficients under the somewhat unrealistic assumption that the orders of the AR and of the MA processes are known in advance, or (*ii*) some model selection procedures that sequentially fit models of increasing dimensions. In practice, however, it is very difficult to verify the assumption that the realized series does come from an ARMA process. Instead, it is usually assumed that the given data are a realization of a *linear* time series, which may be represented by an infinite-order autoregressive process. Some study has been done on the accuracy of an AR approximation for these processes: see Shibata (1980), Goldenshluger and Zeevi (2001) and Ing and Wei (2005). In particular, Goldenshluger and Zeevi (2001) propose a nonparametric minimax approach and assess the accuracy of a finite order AR process in terms of both estimation and prediction.

This paper is concerned with fitting autoregressive time series models with the Lasso. The Lasso procedure, proposed originally by Tibshirani (1996), is one of the most popular approach for model selection in linear and generalized linear models, and has been studied in much of the recent literature; see, e.g., Fan and Peng (2004), Zhao and Bin (2006), Zou (2006), Wainwright (2006), Lafferty et al (2007), and Nardi and Rinaldo (2008), to mention just a few. The Lasso procedure has the advantage of simultaneously performing model selection and estimation, and has been shown to be effective even in high dimensional settings where the dimension of the parameter space grows with the sample size n. In the context of an autoregressive modeling, the Lasso features become especially advantageous, as both the AR order, and the corresponding AR coefficients can be estimated simultaneously. Wang et al. (2007) study linear regression with autoregressive errors. They adapt the Lasso procedure to shrink both the regression coefficients and the autoregressive coefficients, under the assumption that the autoregressive order is fixed.

For the autoregressive models we consider in this work, the number of parameters, or equivalently, the maximal possible lag, grows with the sample size. We refer to this scheme as a double asymptotic framework. The double asymptotic framework enables us to treat the autoregressive order as virtually infinite. The

*Email: yuval@stat.cmu.edu
†Email: arinaldo@stat.cmu.edu

autoregressive time series with an increasing number of parameters lies between a fixed order AR time series and an infinite-order AR time series. This limiting process belongs to a family which is known to contain many ARMA processes (see Goldenshluger and Zeevi, 2001). In this paper we show that the Lasso procedure is particularly adequate for this double asymptotic scheme.

The rest of the paper is organized as follows. The next section formulates the autoregressive modeling scheme and defines the Lasso estimator associated with it. Asymptotic properties of the Lasso estimator are presented in Section 3. These include model selection consistency (Theorem 3.1), estimation consistency (Theorem 3.2), and prediction consistency (Corollary 3.4). Proofs are deferred to Section 6. A simulation study, given in Section 4, accompany the theoretical results. Discussion and concluding remarks appear in Section 5.

2 Penalized autoregressive modeling

In this section we describe our settings and set up the notation.

We assume that $X_1, \ldots, X_n$ are n observations from an AR(p) process:

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + Z_t \quad , \quad t = 1, \ldots, n \; , \tag{1}$$

where $\{Z_t\}$ is a random sequence of independent Gaussian variables with $\mathbb{E}Z_t = 0$, $\mathbb{E}|Z_t|^2 = \sigma^2$ and $\mathrm{cov}(Z_t, X_s) = 0$ for all $s < t$. The last requirement is standard, and rely on a reasoning under which the process $\{X_t\}$ does not depend on future values of the driving Gaussian noise. The assumption about Gaussianity of $\{Z_t\}$ is by no means necessary, and can be relaxed. It does, however, facilitate our theoretical investigation and the presentation of various results, and therefore, it is in effect throughout the article. In Section 5 we comment on how to modify our assumptions and proofs to allow for non-Gaussian innovations $\{Z_t\}$.

We further assume that $\{X_t\}$ is *causal*, meaning that there exists a sequence of constants $\{\psi_j\}$, $j = 0, 1, \ldots$, with absolutely convergent series, $\sum_{j=0}^{\infty} |\psi_j| < \infty$, such that $\{X_t\}$ has a MA(∞) representation:

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \; , \tag{2}$$

the series being absolutely convergent with probability one. Equivalently, we could stipulate that $\{X_t\}$ is purely non-deterministic, and then obtain representation (2), with $\psi_0 = 1$ and $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, directly from the Wold decomposition (see, e.g. Davis and Brockwell, 1991). A necessary and sufficient condition for causality is that $1 - \phi_1 z - \ldots - \phi_p z^p \neq 0$ for all complex z within the unit disc, $|z| \leq 1$. Notice that causality of $\{X_t\}$, and Gaussianity of $\{Z_t\}$, together imply Gaussianity of $\{X_t\}$. This follows from the fact that mean square limits of Gaussian random variables are again Gaussian. The mean and variance of X_t are given, respectively, by $\mathbb{E}X_t = 0$, $\mathbb{E}|X_t|^2 = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2$. We assume, for simplicity, and without any loss of generality, that $\mathbb{E}|X_t|^2 = 1$, so that $\sum_{j=0}^{\infty} \psi_j^2 = \sigma^{-2}$. Let $\gamma(\cdot)$ be the autocovariance function given by $\gamma(k) = \mathbb{E}X_t X_{t+k}$, and let $\Gamma_p = \big(\gamma(i-j)\big)_{i,j=1,\ldots,p}$, the $p \times p$ autocovariance matrix, of lags smaller or equal to $p-1$.

We now describe the penalized ℓ_1 least squares estimator of the AR coefficients. Let $y = (X_1, \ldots, X_n)'$, $\phi = (\phi_1, \ldots, \phi_p)'$, and $Z = (Z_1, \ldots, Z_n)'$, where apostrophe denotes transpose. Define the $n \times p$ matrix X with entry X_{t-j} in the tth row and jth column, for $t = 1, \ldots, n$ and $j = 1, \ldots, p$. The Lasso-type estimator $\hat{\phi}_n \equiv \hat{\phi}_n(\Lambda_n)$ is defined to be the minimizer of:

$$\frac{1}{2n} \|y - X\phi\|^2 + \lambda_n \sum_{j=1}^{p} \lambda_{n,j} |\phi_j| \; , \tag{3}$$

where $\Lambda_n = \{\lambda_n, \{\lambda_{n,j} \, , \, j = 1, \ldots, p\}\}$ are tuning parameters, and $\|\cdot\|$ denotes the l_2-norm. Here, λ_n is a grand tuning parameter, while the $\{\lambda_{n,j}, j = 1, \ldots, p\}$ are specific tuning parameters associated with

predictors X_{t-j}. The Lasso solution (3) will be sparse, as some of the autoregressive coefficients will be to set to (exactly) zero, depending on the choice tuning parameters Λ_n. Naturally, one may want to further impose that $\lambda_{n,j} < \lambda_{n,k}$ for lags values satisfying $j < k$, to encourage even sparser solutions, although this is not assumed throughout. The idea of using ℓ_1 regularization to penalize differently the model parameters, as we do in (3), was originally proposed by Zou (2006) under the name of adaptive Lasso. As shown in Zou (2006), from an algorithmic point of view, the solution to our adaptive Lasso (3) can be obtained by a slightly modified version of the LARS algorithm of Efron et al. (2004). A possible choice for $\lambda_{n,j}$ would be to use the inverse least squares estimates, as in Zou (2006), but this is not pursued here.

As mentioned before, we consider a double asymptotic framework, in which the number of parameters $p \equiv p_n$ grows with n at a certain rate. Clearly, the "large p small n" ($p \gg n$) scenario, which is an important subject of many of nowadays articles, is not adequate here. Indeed, one might be suspicious about the statistical properties of the proposed estimator even when p is comparable with n ($p < n$, but is close to n). Accounting for the mechanism of the autoregressive progress, one is led to think that p should grow with n at a much slower rate. This article shows that the choice of $p = \mathrm{O}(\log n)$ will lead to nice asymptotic properties of the proposed procedure (3). Such a choice of the AR order arises also in Goldenshluger and Zeevi (2001), who prove minimax optimality for a different regularized least squares estimator. Moreover, as pointed out in Goldenshluger and Zeevi (2001), the same order of p arises also in spectral density estimation (see Efromovich (1998)). Finally, similar rate appears also, in a different context, in Rothman et al. (2007).

In classical linear time series modeling, one usually attempts to fit sequentially an AR(p) with increasing orders of the maximal lag p (or by fixing p and then estimating the coefficients). The Lasso-type estimator of scheme (3) will shrink down to zero irrelevant predictors. Thus, not only that model selection and estimation will occur simultaneously, but the fitted (selected) model will be chosen among all relevant AR(p) processes, with $p = \mathrm{O}(\log n)$.

3 Asymptotic Properties of the Lasso

In this section we derive the asymptotic properties of the Lasso estimator $\hat{\phi}_n$. These include model selection consistency, estimation consistency and prediction consistency. We briefly describe each type of consistency, develop the needed notation, and present the results, with proofs relegated to Section 6.

3.1 Model Selection Consistency

We assume that the AR(p) process (1) is generated according to a true, unknown parameter $\phi^* = (\phi_1^*, \ldots, \phi_p^*)$. When p is large, it is not unreasonable to believe that this vector is sparse, meaning that only a subset of potential predictors are relevant. Model selection consistency is about recovering the sparsity structure of the true, underlying parameter ϕ^*.

For any vector $\phi \in \mathbb{R}^p$, let $\mathrm{sgn}(\phi) = (\mathrm{sgn}(\phi_1), \ldots, \mathrm{sgn}(\phi_p))$, where $\mathrm{sgn}(\phi_j)$ is the sign function taking values $-1, 0$ or 1, according to as $\phi_j < 0, \phi_j = 0$ or $\phi_j > 0$, respectively. A given estimator $\hat{\phi}_n$ is said to be *sign consistent* if $\mathrm{sgn}(\hat{\phi}_n) = \mathrm{sgn}(\phi^*)$, with probability tending to one, as n tends to infinity, i.e.,

$$\mathbb{P}(\mathrm{sgn}(\hat{\phi}_n) = \mathrm{sgn}(\phi^*)) \longrightarrow 1 \qquad , \qquad n \to \infty \, . \tag{4}$$

Let $S = \{j \, : \, \phi_j^* \neq 0\} = \mathrm{supp}(\phi^*) \subset \{1, 2, \ldots, p\}$. A weaker form of model selection consistency, implied by the sign consistency, only requires that, with probability tending to 1, ϕ^* and $\hat{\phi}_n$ have the same support.

We shall need a few more definitions. Let $s = |S|$ denote the cardinality of the set of true nonzero coefficients, and let $\nu = p - s = |S^c|$, with $S^c = \{1, \ldots, p\} \setminus S$. For a set of indexes I, we will write $x_I = \{x_i, i \in I\}$ to denote the subvector of x whose elements are indexed by the coordinates in I. Similarly, $x_I y_I$ is a vector with elements $x_i y_i$. For a $n \times p$ design matrix X, we let X_I, for any subset I of $\{1, 2, \ldots, p\}$, denote the sub-matrix of X with columns as indicated by I. Sub-matrices of the autocovariance matrix Γ_p (and of any other matrix), are denoted similarly. For example, Γ_{II^c} is $(\gamma(i-j))_{i \in I, j \notin I}$. Finally, let

$\alpha_n = \min_{j\in S} |\phi_j^*|$ denote the magnitude of the smallest nonzero coefficient. Finally, although virtually all quantities related to (3) depend on n, we do not always make this dependence explicit in our notation.

We are now ready to present our first result:

Theorem 3.1. *Consider the settings of the AR(p) process describe above. Assume that*

(i) *there exists a finite, positive constant* $C_{\max}$ *such that* $\|\Gamma_{SS}^{-1}\| \leq C_{max}$*;*

(ii) *there exists an* $\epsilon \in (0, 1]$ *such that* $\|\Gamma_{S^cS}\Gamma_{SS}^{-1}\|_\infty \leq 1 - \epsilon$.

Further, assume that the following conditions hold:

$$\limsup_{n\to\infty} \frac{\max_{i\in S} \lambda_{n,i}}{\min_{j\in S^c} \lambda_{n,j}} \leq 1 \,, \tag{5}$$

$$\frac{1}{\alpha_n}\left[\sqrt{s/n} + \lambda_n \|\lambda_{n,S}\|_\infty\right] \longrightarrow 0 \quad , \quad as \quad n \to \infty \,, \tag{6}$$

$$\frac{n\lambda_n^2(\min_{i\in S^c} \lambda_{n,i})^2}{\max\{s, \nu\}} \longrightarrow \infty \quad , \quad as \quad n \to \infty \,. \tag{7}$$

Let $p = \mathrm{O}(\log n)$*. Then, the Lasso estimator* $\hat{\phi}_n$ *is sign consistent (cf. (4)).*

Condition **(ii)** in Theorem 3.1 is assumed in various guises elsewhere in the Lasso literature (see, e.g., Wainwright (2006), Zhao and Bin (2006) and Zou (2006)). It is an incoherence condition, which controls the amount of correlation between relevant variables and irrelevant variables. Condition (5) is intuitively clear and it appears under similar form in Nardi and Rinaldo (2008). It captures the rationale , recalling that one may have $\lambda_j < \lambda_k$ for $j < k$, that (even) the largest penalty coefficient of the relevant lags should be kept asymptotically smaller than the smallest penalty coefficient of the irrelevant lags. Conditions (6) and (7) are similar to conditions appearing in Wainwright (2006), Nardi and Rinaldo (2008), and Lafferty et al (2007), to name but a few. The fraction $\sqrt{s/n}$ in (6) is in line with similar works, mentioned above. For example, under the *linear sparsity scheme*, i.e., $s = \alpha p$, with $\alpha \in (0, 1)$ (see Wainwright (2006)), and with p comparable to n, the Gaussian ensemble leads to a fraction of order $\mathrm{O}(\log n/n)$, which is similar to the fraction under the current setting, for which we have $p = \mathrm{O}(\log n)$.

3.2 Estimation and Prediction Consistency

Our next result is about *estimation consistency*. An estimator $\hat{\phi}_n$ is said to be estimation consistent, or l_2-consistent if $\|\hat{\phi}_n - \phi^*\|$ converges to zero, as n tends to infinity. We have the following:

Theorem 3.2. *Recall the settings of the AR(p) process set forth below (1). Let* $p = \mathrm{O}(\log n)$*, and* $\alpha_n = p^{1/2}(n^{-1/2} + \lambda_n\|\lambda_{n,S}\|)$*. Assume that* $\lambda_n\|\lambda_{n,S}\| = \mathrm{O}\left(n^{-1/2}\right)$*. Then, the Lasso estimator* $\hat{\phi}_n$ *is estimation consistent with a rate of order* $\mathrm{O}(\alpha_n)$.

Prediction consistency is about a similar convergence statement, but for the prediction of future values using the fitted model. Formally, prediction consistency holds if $\|X\hat{\phi}_n - X\phi^*\|$ converges to zero, as n tends to infinity. We show below a similar result when the sample autocovariance matrix $X'X$ is replaced by the (theoretical) autocovariance matrix Γ_p. The autoregressive settings assumed here are, in some sense, much more challenging than in linear (parametric or non-parametric) regression models, for two reasons. Firstly, the design matrix is not fixed as is usually assumed, and secondly, the entries of the X are not independent across rows, as is usually assumed for random designs.

The family of AR processes considered here are, in fact, a subset of a larger family of time series. In order to establish the prediction consistency result, we make an explicit use of the structure of this larger family, to which we now describe.

Following Goldenshluger and Zeevi (2001), we denote by $\mathcal{H}_\rho(l, L)$, for some $\rho > 1$, $0 < l < 1$, and $L > 1$, a family consisting of all stationary Gaussian time series with $\mathbb{E}X_t = 0$, $\mathbb{E}|X_t|^2 = 1$, and with

$$0 < l \leq |\psi(z)| \leq L \, ,$$

for every complex z with $|z| \leq \rho$, where $\psi(z)$ is the MA(∞) transfer function related to the AR polynomial by $\psi(z) = 1/\phi(z)$.

We shall need the notion of a strong mixing (or α-mixing) condition. Let $\{X_t\}$ be a time series defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For any two (sub) σ-fields $\mathcal{A}$ and $\mathcal{B}$, define

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup_{A \in \mathcal{A}, B \in \mathcal{B}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| \, .$$

Denote by $\mathcal{F}_s^t$, the σ-field generated by $(X_s, \ldots, X_t)$, for $-\infty \leq s \leq t \leq \infty$. Then, $\{X_t\}$ is said to be strongly mixing if $\alpha_X(m) \to 0$, as $m \to \infty$, where

$$\alpha_X(m) = \sup_{j \in \{0, \pm 1, \pm 2, \ldots\}} \alpha(\mathcal{F}_{-\infty}^j, \mathcal{F}_{j+m}^\infty) \, .$$

Attractiveness of $\mathcal{H}_\rho(l, L)$ comes from the fact that processes in $\mathcal{H}_\rho(l, L)$ are strong mixing with an exponential decay, i.e.

$$\alpha_X(m) \leq 2 \left(\frac{L\rho}{l(\rho - 1)} \right)^2 \rho^{-m} \, . \tag{8}$$

This follows since processes in $\mathcal{H}_\rho(l, L)$ have exponentially decaying AR coefficients as well as exponentially decaying autocovariances (see (Goldenshluger and Zeevi, 2001, Lemma 1, and in particular, expression (39))).

For every p-dimensional vector a and $p \times p$ symmetric matrix A, we denote with $\|a\|_A^2 = a'Aa$, the (squared) l_2-norm associated with A. Let C_1, C_2 be two universal constants (their explicit values are given within the proof of the following theorem). Define

$$\beta_1 = 1 + \frac{1}{\log \rho} \quad , \quad \beta_2 = 1 + \frac{L\rho}{l(\rho - 1)} \, , \qquad \text{and} \quad D = (C_1^3 C_2 \beta_1^2 \beta_2^3)^{1/5} \, . \tag{9}$$

Let $\lambda_{\min} = \min_{j=1,\ldots,p} \lambda_{n,j}$, and $\lambda_{\max} = \max_{j=1,\ldots,p} \lambda_{n,j}$. We have:

Theorem 3.3. *Recall the settings of the AR(p) process set forth below (1). Let $p = \mathrm{O}(\log n)$. Assume:*

(i) *There exists a finite, positive constant M such that $\lambda_{n,j} \leq M$, for every $j = 1, \ldots, p$.*

(ii) *For every $p \geq 2$, there exists a positive constant κ_p, such that*

$$\Gamma_p - \kappa_p \, diag(\Gamma_p)$$

is a positive semi-definite matrix.

If $\lambda_n (s/p)^{1/2} \leq D n^{-2/5}$, then there exist a constant C (depending only on M), and constants F_1 and F_2 (depending only on $C_1, C_2, \beta_1, \beta_2$), such that for all $0 < c < \infty$, and all $y > \sigma^2(n + Dn^{3/5})$,

$$\|\hat{\phi}_n - \phi^*\|_{\Gamma_p}^2 \leq C \lambda_n^2 \frac{s}{\kappa_p}$$

holds true with probability at least $1 - \pi_n$, where

$$\pi_n \leq 6p \exp\left\{ -F_1 \min\left\{ (\sigma^{-2} y - n)^{1/3}, c^2 \sigma^{-2}, \frac{n^2 \lambda_n^2 \lambda_{\min}^2}{y + cn\lambda_n \lambda_{\max}/2} \right\} \right\} + p^2 \exp\left\{ -F_2 n \lambda_n^2 (s/p^2) \right\} \, . \tag{10}$$

Condition **(ii)** has been used in the context of aggregation procedures for nonparametric regression with fixed design (Bunea *et al.* (2007a)), and also for nonparametric regression with random design (Bunea *et al.* (2007b)).

Theorem 3.3 may be utilized to show that the Lasso estimator $\hat{\phi}_n$ is prediction consistent. One only needs to make sure that the decay of the bound (10) on π_n. The theorem actually gives a whole range of possible rates of decay, by picking c and y. One *possible* choice is given below.

Corollary 3.4. *Let $\lambda_n = n^{-\alpha}$, with $\alpha \in (2/5, 1/2)$. Let $c = D_1 y/(n\lambda_n \lambda_{\max})$, and $y = D_2 n$, for positive constants D_1, D_2. If $(s/p)^{1/2} \leq D n^{\alpha - 2/5}$, then there exists an appropriate constant F, such that the bound (10) on π_n is smaller than*

$$p^2 \exp\left\{ -F \min\left\{ n^{1/3}, n^{2\alpha}/\lambda_{\max}^2, n^{1-2\alpha}\lambda_{\min}^2, n^{1-2\alpha} s/p^2 \right\} \right\},$$

which tends to zero as n goes to infinity.

4 Illustrative Simulations

We consider a sparse autoregressive time series of length 1000 obeying the model

$$X_t = 0.2X_{t-1} + 0.1X_{t-3} + 0.2X_{t-5} + 0.3X_{t-10} + 0.1X_{t-15} + Z_t, \tag{11}$$

with nonzero coefficients at lags 1, 3, 5, 10 and 15, where the innovations Z_t are i.i.d. Gaussians with mean zero and standard deviation 0.1. The coefficients were chosen to satisfy the characteristic equation for a stationary AR process.

Figure 1 shows one time series simulated according to the model (11), along with its autocorrelation and partial autocorrelation plots. For this time series, Figure 2 shows the solution paths computed using the `R` algorithm `lars` and for a value of $p = 50$. Notice that we only use one penalty parameter, i.e. we penalize equally all the autoregressive coefficients. The vertical line marks the optimal ℓ_1 threshold found by cross validation. In our simulations, we declared significant the variables whose coefficients have nonzero solution paths meeting the vertical line corresponding to the cross validation value.

Notice that, in the exemplary instance displayed in Figure 2, all the nonzero autoregressive coefficients are correctly included in the model. Furthermore, a more careful inspection of the solution paths reveals that the order at which the significant variables enter the set of active solutions match very closely the magnitude of the coefficients used in our model, with ϕ_{10} and ϕ_5, the more significant coefficients, entering almost immediately, and ϕ_3 and ϕ_{15} entering last. In contrast, Figure 3 displays the fitted values for the first 30 autoregressive coefficients computed using the Yule-Walker method implemented using `R` by the routine `ar` (note that the Yule-Walker estimator has the same asymptotic distribution as the MLE's). Notice that the solution is non-sparse. The dashed vertical gray lies indicate the true nonzero coefficients. The autoregressive order of the model was correctly estimated to be 15 using the AIC criterion.

We simulated 1000 time series from the model (11) and we selected the significant variables according to the cross-validation rule described above. Figure 4 **a)** displays the histogram of the number of selected variables. The mean and standard deviations of these numbers are 6.42 and 2.44, respectively, while the minimum, median and maximum numbers are 3, 6 and 22, respectively. In comparison, Figure 4 **b)** shows the histogram of the autoregressive orders determined by AIC in `ar`. Table 1 displays some summary statistics of our simulations. In particular, the second row shows the number of times, out 1000 simulated time series, that each of the nonzero autoregressive coefficients was correctly selected. The second row indicates the number of times the variable corresponding to each nonzero coefficient in (11) was among the first five selected variables. Notice that ϕ_{10} and ϕ_5 are always included among the selected variables, while ϕ_3 and ϕ_{15} have a significantly smaller, but nonetheless quite high, chance of being selected.

We also investigated the order at which the autoregressive coefficients entered the solution paths, the rationale being that more significant nonzero variables enter sooner, in accordance with the way the `lars`

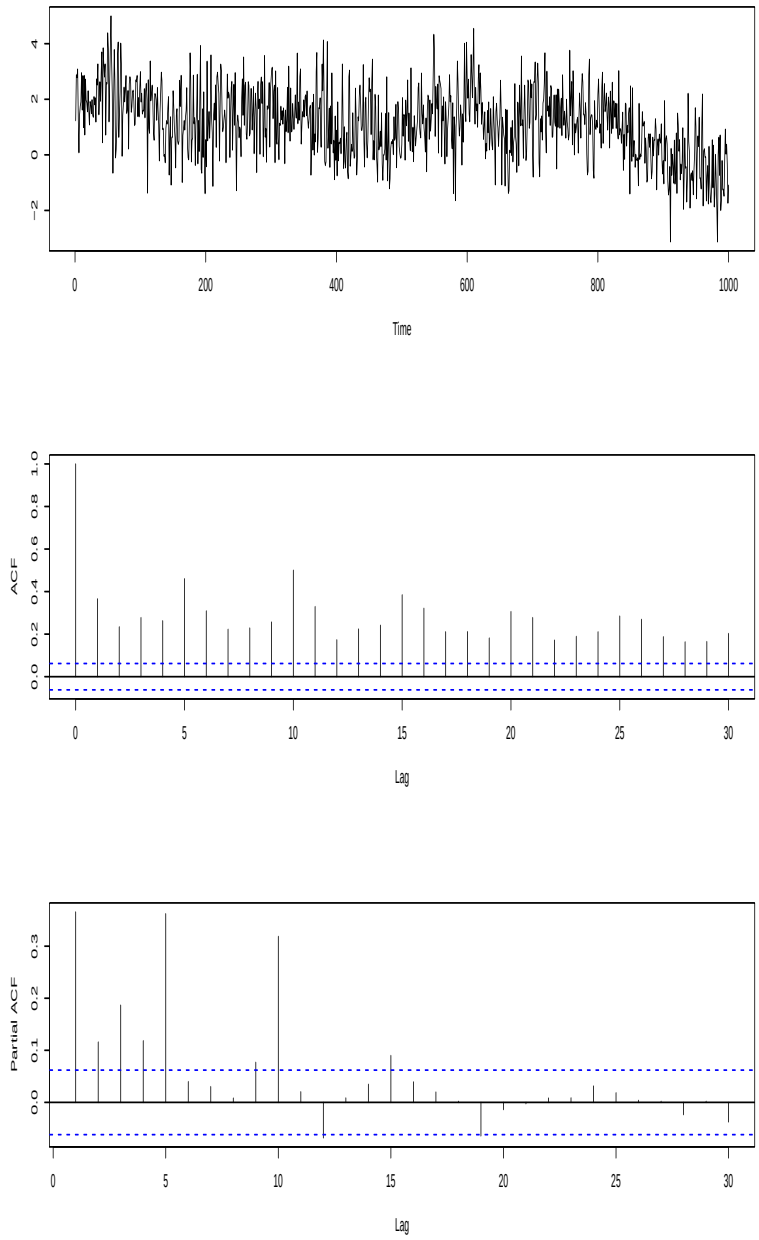


Figure 1: A time series simulated from the sparse autoregressive model (11) along with its autocorrelation and partial autocorrelation coefficients.

algorithm works (see Efron et al. (2004)). Figure 5 summarizes our findings. In each of the barplots, the x-axis indexes the steps at which the variable corresponding to the autoregressive coefficient enters the solution path, while the y-axis displays the frequency. Interestingly enough, in most cases, ϕ_{10} and ϕ_5 are selected as the first and second nonzero variables, while ϕ_{15} and, in particular, ϕ_3 enter the set of active variables later and are not even among the first five variables selected in 1.9% the and 20.2% of cases, respectively.

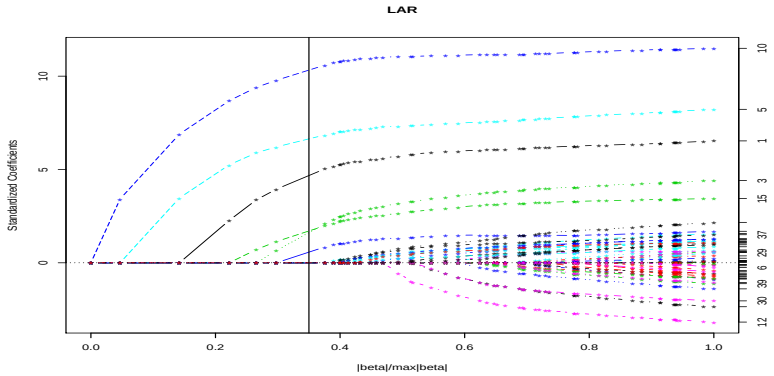


Figure 2: Solution paths of the `lars` algorithm when applied to the time series displayed in Figure 1. The vertical bar represents the optimal ℓ_1 penalty for this time series selected using cross validation.

Table 1: Number of times the nonzero autoregressive coefficients are correctly identified and number of times they are correctly selected among the first 5 variables entering the solution paths.

ϕ	ϕ_1	ϕ_3	ϕ_5	ϕ_{10}	ϕ_{15}
Value	0.2	0.1	0.2	0.3	0.1
Number of times correctly selected	992	754	1000	1000	913
Number times selected among first 5	992	602	1000	1000	895

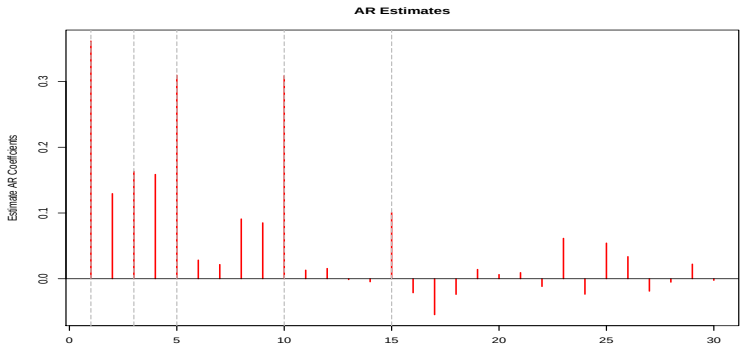
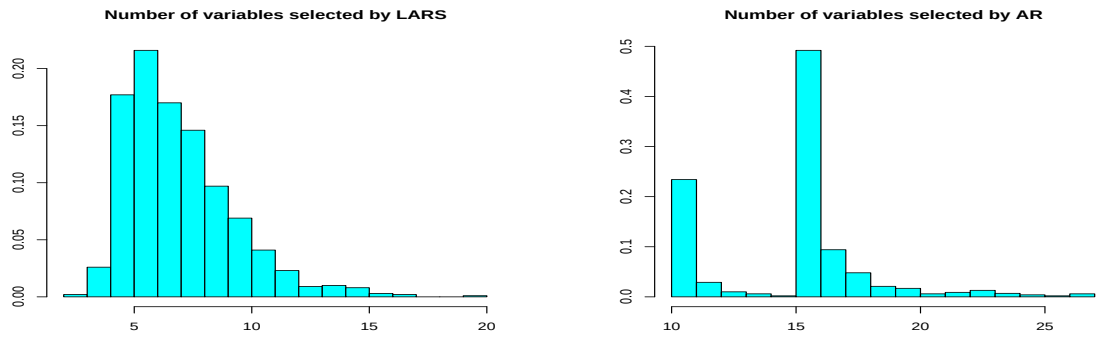


Figure 3: Autoregressive coefficients for the time series of Figure 1 obtained using the routine `ar`. The dashed vertical line mark the lags for true nonzero coefficients.

5 Discussion

We defined the Lasso procedure for fitting an autoregressive model, where the maximal lag may increase with the sample size. Under this double asymptotic framework, the Lasso estimator was shown to possess several consistency properties. In particular, when $p = \mathrm{O}(\log n)$, the Lasso estimator is model selection consistent, estimation consistent, and prediction consistency. The advantage of using the Lasso procedure in conjunction with a growing p is that the fitted model will be chosen among all possible AR models whose maximal lag is between 1 and $\mathrm{O}(\log n)$. Letting n go to infinity, we may virtually obtain a good approximation for a general linear time series.

As mentioned in Section 2, the assumption about Gaussianity of the underlying noise $\{Z_t\}$ is not necessary. The proof of the model selection consistency result (Theorem 3.1) avoids making use of Gaussianity by



a) **b)**

Figure 4: Distribution for the number of variables selected by **a)** the `lars` algorithm using cross validation and **b)** the `ar` algorithm using AIC over 1000 simulations of the time series described in (11).

using Burkholder's inequality in conjunction with a maximal moment inequality. The proof of the estimation consistency result (Theorem 3.2) requires Lemma 6.2, which does make use of the assumed Gaussianity. However, this is not crucial. In fact, we can relax the Gaussianity assumption and require only the Z_t are $IID(0, \sigma^2)$, with bounded fourth moment (see (Davis and Brockwell, 1991, p. 226-227)). In this case, instead of using Wick's formula we may apply the moving average representation $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$, along with the absolute summability of the ψ_j's. Finally, the prediction consistency result (Theorem 3.3 and Corollary 3.4) may also be obtained by relaxing the Gaussianity assumption. One only needs to impose appropriate moment conditions of the driving noise.

The autoregressive modeling via the Lasso procedure stimulates other interesting future directions. In many cases, non-linearity is evident from the data. In order to capture deviation from linearity, one may try to fit a non-linear (autoregressive) time series model to the data in the form

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \sum_{\nu=2}^{p} \{\phi^{i_1,\ldots,i_\nu} \prod_{j=1}^{\nu} X_{t-i_j}\} + Z_t \ ,$$

where we used the Einstein notation for the term in the curly brackets, to indicate summation over all $i_1 < i_2 < \ldots < i_\nu$. Notice that for even mild values of p, the number of possible interaction terms may be very large. This is a very challenging problem as one needs to obtain a solid understanding of the properties of the non-linear autoregressive process before applying the Lasso (or any other) procedure.

6 Proofs

Here we prove Theorems 3.1, Theorem 3.2, and Theorem 3.3. Recall scheme (3). This is a convex minimization problem. Denote by $M_{\Lambda_n}(\cdot)$, for $\Lambda_n = \{\lambda_n, \{\lambda_{n,j}\ , \ j = 1, \ldots p\}\}$, the objective function, i.e.,

$$M_{\Lambda_n}(\phi) = \frac{1}{2n}\|y - X\phi\|^2 + \lambda_n \sum_{j=1}^{p} \lambda_{n,j}|\phi_j| \ . \tag{12}$$

The Lasso estimator is an optimal solution to the problem $\min\{M_{\Lambda_n}(\phi)\ , \ \phi \in \mathbb{R}^p\}$. Gradient and Hessian of the least-squares part in $M_{\Lambda_n}(\cdot)$ are given, respectively, by $n^{-1}\mathfrak{X}\phi - n^{-1}\sum_{t=1}^{n} X_t \mathbf{X_t}$, and $n^{-1}\mathfrak{X}$, where $\mathfrak{X}$ (the gram matrix associated with the design matrix X), and $\mathbf{X_t}$ is a notation that we use throughout this section:

$$\mathfrak{X} = X'X \quad , \quad \mathbf{X_t} = (X_{t-1}, \ldots, X_{t-p})' \ .$$

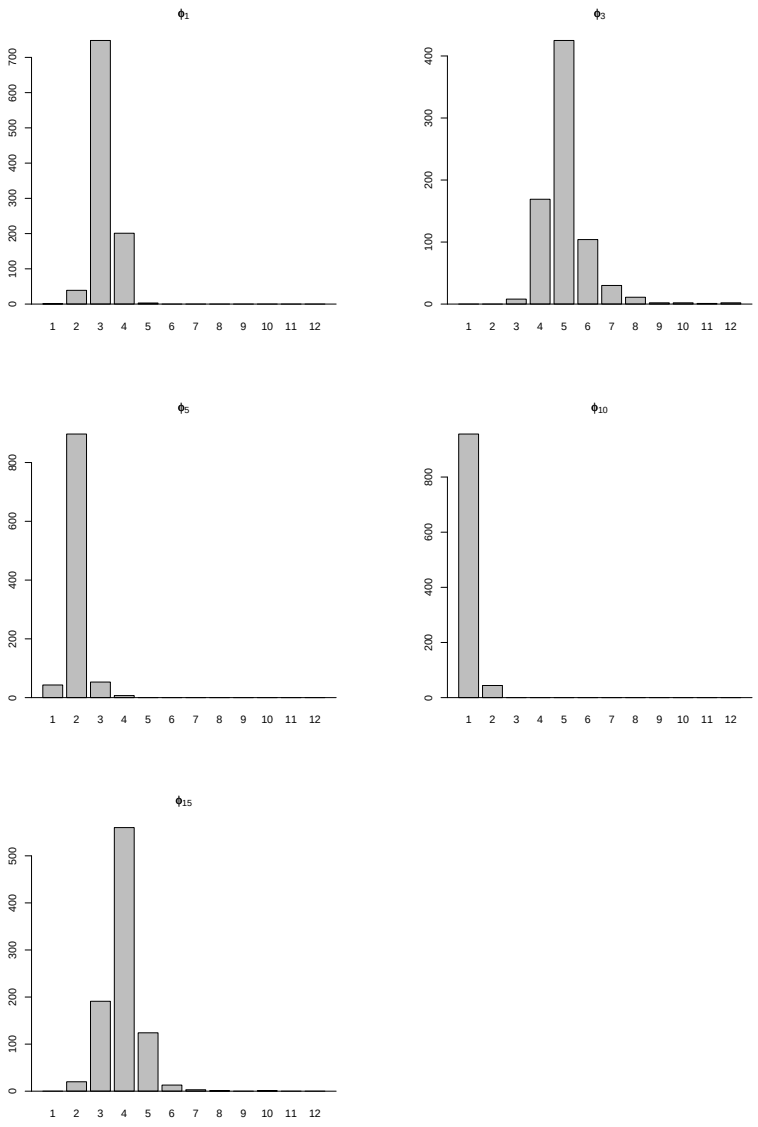


Figure 5: Frequencies of the order at which the 5 autoregressive coefficients entered the solutions paths for the `lars` algorithm over 1000 simulations of the time series described in (11).

6.1 Model Selection Consistency

Proof of Theorem 3.1. We adapt a Gaussian ensemble argument, given in Wainwright (2006), to the present setting. Standard optimality conditions for convex optimization problems imply that $\hat{\phi}_n \in \mathbb{R}^p$ is an optimal solution to the problem $\min\{M_{\Lambda_n}(\phi)\,,\ \phi \in \mathbb{R}^p\}$, if, and only if,

$$\frac{1}{n}\mathfrak{X}\hat{\phi}_n - \frac{1}{n}\sum_{t=1}^{n} X_t \mathbf{X_t} + \lambda_n \hat{\xi}_n = 0\,, \tag{13}$$

where $\hat{\xi}_n \in \mathbb{R}^p$ is a sub-gradient vector with elements $\hat{\xi}_{n,j} = \text{sgn}(\hat{\phi}_{n,j})\lambda_{n,j}$ if $\hat{\phi}_{n,j} \neq 0$, and $|\hat{\xi}_{n,j}| \leq \lambda_{n,j}$ otherwise. Plugging the model structure, $y = X\phi^* + Z$, into (13), one can see that the optimality conditions

become

$$\frac{1}{n}\mathfrak{X}(\hat{\phi}_n - \phi^*) - \frac{1}{n}\sum_{t=1}^{n} Z_t \mathbf{X_t} + \lambda_n \hat{\xi}_n = 0\,. \tag{14}$$

Recall the sparsity set, $S = \{j : \phi_j^* \neq 0\} = \mathrm{supp}(\phi^*)$, the sparsity cardinality $s = |S|$, and $\nu = p - s = |S^c|$. Decomposing the design matrix X to relevant and non-relevant variables, $X = (X_S, X_{S^c})$, we may write the gram matrix $\mathfrak{X}$ as a block matrix of the form

$$\mathfrak{X} = \begin{pmatrix} \mathfrak{X}_{SS} & \mathfrak{X}_{SS^c} \\ \mathfrak{X}_{S^cS} & \mathfrak{X}_{S^cS^c} \end{pmatrix} = \begin{pmatrix} X_S'X_S & X_S'X_{S^c} \\ X_{S^c}'X_S & X_{S^c}'X_{S^c} \end{pmatrix}\,.$$

Notice, for example, that $\mathfrak{X}_{SS} = (\sum_{t=1}^{n} X_{t-i}X_{t-j})_{i,j\in S}$. Incorporating this into the optimality conditions (14) we obtain the following two relations,

$$\begin{aligned} \frac{1}{n}\mathfrak{X}_{SS}[\hat{\phi}_{n,S} - \phi_S^*] - \frac{1}{n}\sum_{t=1}^{n} Z_t \mathbf{X_t^S} &= -\lambda_n \lambda_{n,S}\,\mathrm{sgn}(\phi_S^*)\,, \\ \frac{1}{n}\mathfrak{X}_{S^cS}[\hat{\phi}_{n,S} - \phi_S^*] - \frac{1}{n}\sum_{t=1}^{n} Z_t \mathbf{X_t^{S^c}} &= -\lambda_n \hat{\xi}_{n,S^c}\,, \end{aligned}$$

where $\mathbf{X_t^S}$, and $\mathbf{X_t^{S^c}}$ are vectors with elements $\{X_{t-i}\,,\, i \in S\}$, and $\{X_{t-i}\,,\, i \in S^c\}$, respectively. If $n - s \geq s$ then $\mathfrak{X}_{SS}$ is non-singular with probability one, and we can solve for $\hat{\phi}_{n,S}$ and $\hat{\xi}_{n,S}$,

$$\begin{aligned} \hat{\phi}_{n,S} &= \phi_S^* + \Big(\frac{1}{n}\mathfrak{X}_{SS}\Big)^{-1}\left[\frac{1}{n}\sum_{t=1}^{n} Z_t \mathbf{X_t^S} - \lambda_n \lambda_{n,S}\,\mathrm{sgn}(\phi_S^*)\right] \\ \lambda_n \hat{\xi}_{n,S^c} &= \mathfrak{X}_{S^cS}\mathfrak{X}_{SS}^{-1}\left[\frac{1}{n}\sum_{t=1}^{n} Z_t \mathbf{X_t^S} - \lambda_n \lambda_{n,S}\,\mathrm{sgn}(\phi_S^*)\right] - \frac{1}{n}\sum_{t=1}^{n} Z_t \mathbf{X_t^{S^c}}\,. \end{aligned}$$

Now, *sign consistency* is equivalent (see Lemma 1 in Wainwright (2006)) to showing that

$$\left|\phi_S^* + \big(\frac{1}{n}\mathfrak{X}_{SS}\big)^{-1}\Big[\frac{1}{n}\sum_{t=1}^{n} Z_t \mathbf{X_t^S} - \lambda_n \lambda_{n,S}\,\mathrm{sgn}(\phi_S^*)\Big]\right| > 0 \tag{15}$$

$$\left|\mathfrak{X}_{S^cS}\mathfrak{X}_{SS}^{-1}\Big[\frac{1}{n}\sum_{t=1}^{n} Z_t \mathbf{X_t^S} - \lambda_n \lambda_{n,S}\,\mathrm{sgn}(\phi_S^*)\Big] - \frac{1}{n}\sum_{t=1}^{n} Z_t \mathbf{X_t^{S^c}}\right| \leq \lambda_n \lambda_{n,S^c}\Big] \tag{16}$$

hold, elementwise, with probability tending to 1. Denote the events in (15), and in (16) by $\mathcal{A}$ and $\mathcal{B}$, respectively. The rest of the proof is devoted to showing that $\mathbb{P}(\mathcal{A}) \to 1$, and $\mathbb{P}(\mathcal{B}) \to 1$, as $n \to \infty$.

We commence with $\mathcal{A}$. Let $\alpha_n = \min_{j\in S}|\phi_j^*|$. Recall the notation $\|x_I\|_\infty$ for the l_∞ norm on a set of indices I, i.e., $\max_{i\in I}|x_i|$ (and similarly for matrices). It is enough to show that $\mathbb{P}(\|A_S\|_\infty > \alpha_n) \to 0$, as n tends to infinity, where

$$A_S = \Big(\frac{1}{n}\mathfrak{X}_{SS}\Big)^{-1}\Big[\frac{1}{n}\sum_{t=1}^{n} Z_t \mathbf{X_t^S} - \lambda_n \lambda_{n,S}\,\mathrm{sgn}(\phi_S^*)\Big]\,. \tag{17}$$

Confine attention to the matrix $\mathfrak{X}_{SS}$. The entry at row $i \in S$ and column $j \in S$ is given by $\sum_{t=1}^{n} X_{t-i}X_{t-j}$. Notice that, equivalently, we can write this as $\sum_{t=1-i}^{n-i} X_t X_{t+i-j}$. Following Davis and Brockwell (1991), one can show that $n^{-1}\mathfrak{X}_{SS} \to \Gamma_{SS}$ in probability, as $n \to \infty$, where $\Gamma_{SS} = \big(\gamma(i-j)\big)_{i\in S, j\in S}$, and $\gamma(\cdot)$ is the autocovariance function, $\gamma(h) = \mathbb{E}X_tX_{t+h}$. Therefore, by assumption **(i)** in Theorem 3.1, there exists a finite constant $C_{\max}$, such that $\|(n^{-1}\mathfrak{X}_{SS})^{-1}\|_\infty \leq o_P(1) + C_{\max}$. We continue by investigating the probability associated with the term inside the square brackets in (17).

Notice that $\|\sum_{t=1}^n Z_t \mathbf{X_t^S}\|_\infty$ is given by $\max_{i\in S}|\sum_{t=1}^n Z_t X_{t-i}|$, where Z_t and X_{t-i} are independent random variables for each $t = 1, \ldots, n$, and $i \in S$. Fix an $i \in S$, and define

$$T_n \equiv T_{n,i} = \sum_{t=1}^n Z_t X_{t-i} \ . \tag{18}$$

Let $\mathcal{F}_n = \sigma(\ldots, Z_{n-1}, Z_n)$ be the sigma-field generated by $\{\ldots, Z_{n-1}, Z_n\}$. Simple calculation shows that $\{T_n, \mathcal{F}_n\}_n$ is a martingale. Finally, Let $Y_n = T_n - T_{n-1}$ denote the martingale difference sequence associated with T_n. We quote below a result concerning martingales moment inequalities, which we shall make use of.

Theorem 6.1 (Burkholder's Inequality)**.** *Let $\{X_n, \mathcal{F}_n\}_{n=1}^\infty$ be a martingale, and $\tilde{X}_n = X_n - X_{n-1}$ be the associated martingale difference sequence. Let $q > 1$. For any finite and positive constants $c = c(q)$, and $C = C(q)$ (depending only on q), we have*

$$c\Big[\mathbb{E}\big(\sum_{i=1}^n \tilde{X}_i^2\big)^{q/2}\Big]^{1/q} \geq \big[\mathbb{E}|X_n|^q\big]^{1/q} \leq C\Big[\mathbb{E}\big(\sum_{i=1}^n \tilde{X}_i^2\big)^{q/2}\Big]^{1/q} \ . \tag{19}$$

Applying Cauchy-Schwartz inequality followed by Burkholder's inequality, we obtain

$$\mathbb{E}|T_n| \leq \Big[\mathbb{E}\Big|\sum_{t=1}^n Z_t X_{t-i}\Big|^2\Big]^{1/2} \leq C\Big[\sum_{t=1}^n \mathbb{E}|Z_t^2 X_{t-i}^2|\Big]^{1/2} \leq C\sigma\sqrt{n} \ , \tag{20}$$

where C is a finite and positive constant (from Burkholder's inequality). The last inequality follows by the independence between Z_t and X_{t-i}, and since $\mathbb{E}|X_{t-i}|^2 = 1$. Fix an arbitrary, positive $\xi < \infty$. By a trivial bound we get

$$\begin{aligned}
\mathbb{E}\max_{i\in S}|T_{n,i}| &\leq \xi + \sum_{i\in S}\int_\xi^\infty \mathbb{P}[|T_{n,i}| > y]\,dy \\
&\leq \xi + \frac{1}{\xi}\sum_{i\in S}\mathbb{E}|T_{n,i}|^2 \\
&\leq \xi + C^2\sigma^2\frac{1}{\xi}sn \ ,
\end{aligned}$$

recalling (20). Now, picking $\xi = \sqrt{sn}$, which is optimal, in the sense of obtaining an (asymptotically) smallest fraction, we have,

$$\frac{1}{n}\mathbb{E}\max_{i\in S}|T_{n,i}| \leq \sqrt{s/n} + C^2\sigma^2\sqrt{s/n} = \mathrm{O}\left(\sqrt{s/n}\right) \ . \tag{21}$$

This, in turn, implies, utilizing (17) and Markov's inequality, that $\mathbb{P}(\mathcal{A}) \to 1$, by imposing the condition:

$$\frac{1}{\alpha_n}\Big[\sqrt{s/n} + \lambda_n\|\lambda_{n,S}\|_\infty\Big] \longrightarrow 0 \quad , \quad \text{as } n \to \infty \ ,$$

which is condition (6).

We turn to the event $\mathcal{B}$. Repeating the argument below (17), it is enough to show similar assertion about the event $\mathcal{B}$, with the modification of replacing $\mathfrak{X}_{S^cS}\mathfrak{X}_{SS}^{-1}$, by $\Gamma_{S^cS}\Gamma_{SS}^{-1}$. A sufficient condition for this to hold is that $\{\|B_{S^c}\|_\infty \leq \lambda_n \min_{i\in S^c}\lambda_{n,i}\}$ happens with probability tending to one, where

$$B_{S^c} = \Gamma_{S^cS}\Gamma_{SS}^{-1}\Big[\frac{1}{n}\sum_{t=1}^n Z_t\mathbf{X_t^S} - \lambda_n\lambda_{n,S}\,\mathrm{sgn}(\phi_S^*)\Big] - \frac{1}{n}\sum_{t=1}^n Z_t\mathbf{X_t^{S^c}} \ . \tag{22}$$

Under the *incoherence condition* (condition **(ii)** in the statement of the theorem), we have the following upper bound:

$$\|B_{S^c}\|_\infty \leq (1-\epsilon)\frac{1}{n}\|\sum_{t=1}^n Z_t\mathbf{X_t^S}\|_\infty + (1-\epsilon)\lambda_n\|\lambda_{n,S}\|_\infty + \frac{1}{n}\|\sum_{t=1}^n Z_t\mathbf{X_t^{S^c}}\|_\infty \ ,$$

which leads to: $\mathbb{P}(\|B_{S^c}\|_\infty > \lambda_n \min_{i\in S^c}\lambda_{n,i}) \leq$

$$\mathbb{P}\Big(\frac{2(1-\epsilon)}{n\lambda_n \min_{i\in S^c}\lambda_{n,i}}\|\sum_{t=1}^n Z_t \mathbf{X_t^S}\|_\infty > b\Big) + \mathbb{P}\Big(\frac{2}{n\lambda_n \min_{i\in S^c}\lambda_{n,i}}\|\sum_{t=1}^n Z_t \mathbf{X_t^{S^c}}\|_\infty > b\Big)\,, \tag{23}$$

with $b = 1-(1-\epsilon)\|\lambda_{n,S}\|_\infty/\min_{i\in S^c}\lambda_{n,i}$. Note that inequality (23) follows by the inclusion $\{U+V>z\}\subset\{U>z/2\}\cup\{V>z/2\}$. Under condition (5), it would be enough to consider the right hand side of (23), replacing (the two instances of) b by ϵ. For the first term in (23) we have

$$\mathbb{P}\Big(\frac{2(1-\epsilon)}{n\lambda_n \min_{i\in S^c}\lambda_{n,i}}\|\sum_{t=1}^n Z_t \mathbf{X_t^S}\|_\infty > \epsilon\Big) \leq \frac{1-\epsilon}{\epsilon}\frac{2}{\lambda_n \min_{i\in S^c}\lambda_{n,i}}\frac{1}{n}\mathbb{E}\max_{i\in S^c}|T_{n,i}|\,, \tag{24}$$

which tends by (21) to zero once

$$\frac{n\lambda_n^2(\min_{i\in S^c}\lambda_{n,i})^2}{s} \longrightarrow \infty \quad , \quad \text{as} \quad n\to\infty\,. \tag{25}$$

The same argument may be adapted for $\max_{i\in S^c}|T_{n,i}|$. We only need to replace S by S^c. In this case we find that the condition

$$\frac{n\lambda_n^2(\min_{i\in S^c}\lambda_{n,i})^2}{\nu} \longrightarrow \infty \quad , \quad \text{as} \quad n\to\infty\,, \tag{26}$$

is sufficient for showing that the second term in (23) converges to zero. Condition (7) in the statement of the theorem guarantees both (25) and (26). The proof is now complete.

□

6.2 Estimation and Prediction Consistency

Proof of Theorem 3.2. We follow Fan and Peng (2004). In particular, denoting $\alpha_n = p^{1/2}(n^{-1/2}+\lambda_n\|\lambda_{n,S}\|)$, we will show that for every $\epsilon>0$ there exists a constant C, large enough, such that

$$\mathbb{P}\Big[\inf_{\|u\|=C} M_{\Lambda_n}(\phi^*+\alpha_n u) > M_{\Lambda_n}(\phi^*)\Big] > 1-\epsilon\,,$$

where $M_{\Lambda_n}(\cdot)$ is the objective function and is given in (12). This implies that $\|\hat{\phi}_n - \phi^*\| = O_P(\alpha_n)$.

Multiplying both sides by n clearly does not change the probability. We will show that $-n(M_{\Lambda_n}(\phi^*+\alpha_n u) - M_{\lambda_n}(\phi^*)) < 0$ holds uniformly over $\|u\| = C$. Write

$$M_{\Lambda_n}(\phi) = h(\phi) + \lambda_n\sum_{j=1}^p \lambda_{n,j}|\phi_j|\,,$$

for $h(\phi) = \|y - X\phi\|^2/2n$. We have

$$-n(M_{\Lambda_n}(\phi^*+\alpha_n u) - M_{\Lambda_n}(\phi^*)) \leq -n[h(\phi^*+\alpha_n u) - h(\phi^*)] - n\lambda_n\sum_{j\in S}\lambda_{n,j}[|\phi_j^*+\alpha_n u_j| - |\phi_j^*|]\,.$$

Consider separately the least squares term, and the term associated with the l_1-penalty. We have, exploiting the fact that $\sum_{t=1}^n X_t\mathbf{X_t} = \mathfrak{X}\phi^* + \sum_{t=1}^n Z_t\mathbf{X_t}$,

$$-n[h(\phi^*+\alpha_n u) - h(\phi^*)] = \alpha_n u'\sum_{t=1}^n Z_t\mathbf{X_t} - \alpha_n^2 u'\mathfrak{X}u/2 \equiv I_1 - I_2\,.$$

Recalling the definition of $T_{n,i} = \sum_{t=1}^n Z_t X_{t-i}$ (see (18)), and utilizing the result in (20) we obtain

$$|I_1| \leq \alpha_n\|u\|\|\sum_{t=1}^n Z_t\mathbf{X_t}\| = \|u\|O_P(\alpha_n\sqrt{pn})\,.$$

Moving on to I_2, we write

$$I_2 = \alpha_n^2 u' \mathfrak{X} u/2 = n\alpha_n^2 u'(n^{-1}\mathfrak{X} - \Gamma_p)u/2 + n\alpha_n^2 u'\Gamma_p u/2 \; . \tag{27}$$

We know that $n^{-1}\mathfrak{X}_{ij}$ tends in probability to $\gamma(i-j)$, where $\mathfrak{X}_{ij} = \sum_{t=1}^{n} X_{t-i}X_{t-j}$, the (i,j) entry of $\mathfrak{X}$. This clearly implies $\|n^{-1}\mathfrak{X} - \Gamma_p\| = o_P(1)$, in the fixed p scenario. Lemma 6.2 below shows that this may also hold true in the growing p scenario which we consider here.

Lemma 6.2. *Assume $\sum_{j=0}^{\infty} |\psi_j| < \infty$, as before. Then,*

$$\|n^{-1}\mathfrak{X} - \Gamma_p\| = o_P(1) \; . \tag{28}$$

Proof. We adopt arguments given in (Davis and Brockwell, 1991, p. 226-227). Let $\epsilon > 0$ be given. Using the fact that $\|A\| \leq \|A\|_F$, where $\|\cdot\|_F$ is the Frobenius matrix norm, $\sum_{i,j} |A_{ij}|^2$, we have

$$\mathbb{P}(\|n^{-1}\mathfrak{X} - \Gamma_p\| > \epsilon) \leq \frac{1}{\epsilon^2} \sum_{i,j=1}^{p} d_{ij} \; , \tag{29}$$

where $d_{ij} = \mathbb{E}(n^{-1}\mathfrak{X}_{ij} - \gamma(i-j))^2$. We shall make use of Wick's formula. This formula gives the expectation of a product of several centered (joint) Gaussian variables $G_1, \ldots, G_N$, in terms of the elements of their covariance matrix $C = (c_{ij})$:

$$\mathbb{E}\prod_{i=1}^{k} G_i = \sum c_{i_1 i_2} \cdots c_{i_{k-1} i_k} \; ,$$

for $k = 2m$, and zero otherwise. The sum extends over all different partitions of $\{G_1, \ldots, G_{2m}\}$ into m pairs. Applying the formula, we obtain:

$$\begin{aligned} \mathbb{E}\mathfrak{X}_{ij}^2 &= \sum_{s,t=1-i}^{n-i} \mathbb{E} X_t X_{t+i-j} X_s X_{s+i-j} \\ &= \sum_{s,t=1-i}^{n-i} \Big(\gamma^2(i-j) + \gamma^2(s-t) + \gamma(s-t+i-j)\gamma(-(s-t)+i-j)\Big) \; , \end{aligned}$$

where we have used the equivalent representation $\mathfrak{X}_{ij} = \sum_{t=1-i}^{n-i} X_t X_{t+i-j}$.

A change of variables $k = s - t$ shows that

$$\sum_{s,t=1-i}^{n-i} \Big(\gamma^2(s-t) + \gamma(s-t+i-j)\gamma(-(s-t)+i-j)\Big) =$$

$$n[\gamma^2(0) + \gamma^2(i-j)] + 2\sum_{k=1}^{n-1}(n-k)[\gamma^2(k) + \gamma(k+i-j)\gamma(-k+i-j)] \; .$$

Therefore,

$$\begin{aligned} d_{ij} = \frac{p^2}{n^2}\gamma^2(i-j) + \frac{1}{n}[\gamma^2(0) + \gamma^2(i-j)] \\ + \quad \frac{2}{n^2}\sum_{k=1}^{n-1}(n-k)[\gamma^2(k) + \gamma(k+i-j)\gamma(-k+i-j)] \; . \end{aligned} \tag{30}$$

Notice that $\sum_{k=1}^{\infty} |\gamma^2(k) + \gamma(k+i-j)\gamma(-k+i-j)| < \infty$. This may be seen by using the expression for the autocovariance function, $\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}$, and by utilizing the summability of the ψ_j's, $\sum_{j=0}^{\infty} |\psi_j| < \infty$. The expression (30) is therefore bounded by an $O(1/n)$ order term. This, in turn, shows that $d_{ij} = O(1/n)$, uniformly for every i, j. The proof is completed by recalling the RHS of (29), which is of the order of magnitude of $O(p^2/n)$. □

Using Lemma 6.2 we obtain

$$|n\alpha_n^2 u'(n^{-1}\mathfrak{X} - \Gamma_p)u/2| \leq o_P(1)n\alpha_n^2\|u\|^2 \,. \tag{31}$$

We complete the argument with a bound on the term associated with the penalties, $-n\lambda_n \sum_{j\in S}\lambda_{n,j}[|\phi_j^* + \alpha_n u_j| - |\phi_j^*|]$. Applying the Cauchy-Schwarz inequality, along with the fact that $\|a\|_1 \leq \sqrt{p}\|a\|_2$ for every $a \in \mathbb{R}^p$, it is clear that the above term is absolutely bounded by $\lambda_n\|\lambda_{n,S}\|_\infty\sqrt{s}n\alpha_n\|u\|$. Now, since the second term in I_2 (see (27)) dominates the other terms, the proof of the theorem is completed.

□

Proof of Theorem 3.3. We begin as in Bunea *et al.* (2007b). Recall that $\|a\|_A^2$ stands for $a'Aa$, for every p-dimensional vector a, and $p \times p$ symmetric matrix A. We proceed by stating and proving two lemmas.

Lemma 6.3. *Let assumptions* **(i)**, *and* **(ii)** *of Theorem 3.3 be in effect. Then,*

$$\|\hat{\phi}_n - \phi^*\|_{\mathfrak{X}/n}^2 \leq 4\lambda_n M(s\kappa_p^{-1})^{1/2}\|\hat{\phi}_n - \phi^*\|_{\Gamma_p} \tag{32}$$

holds true on

$$\mathfrak{I}_1 = \Big\{|\frac{2}{n}\sum_{t=1}^{n} X_{t-j}Z_t| \leq \lambda_n\lambda_{n,j} \quad , \quad \textit{for all} \quad j = 1,\ldots,p\Big\} \,. \tag{33}$$

Proof. By definition, the Lasso estimator $\hat{\phi}_n$ satisfies (see (12)),

$$n^{-1}\|y - X\hat{\phi}_n\|^2 + 2\lambda_n\sum_{j=1}^{p}\lambda_{n,j}|\hat{\phi}_{n,j}| \leq n^{-1}\|y - X\phi^*\|^2 + 2\lambda_n\sum_{j=1}^{p}\lambda_{n,j}|\phi_j^*| \,.$$

Recalling the model $y = X\phi^* + Z$, we obtain, by re-arrangeing the above terms,

$$\|\hat{\phi}_n - \phi^*\|_{\mathfrak{X}/n}^2 + 2\lambda_n\sum_{j=1}^{p}\lambda_{n,j}|\hat{\phi}_{n,j}| \leq 2(\hat{\phi}_n - \phi^*)'\frac{1}{n}X'Z + 2\lambda_n\sum_{j=1}^{p}\lambda_{n,j}|\phi_j^*| \,.$$

Now, since $(\hat{\phi}_n - \phi^*)'\frac{1}{n}X'Z = \sum_{j=1}^{p}(\hat{\phi}_{n,j} - \phi_j^*)\frac{1}{n}\sum_{t=1}^{n}X_{t-j}Z_t$, we have, on $\mathfrak{I}_1$,

$$\begin{aligned}\|\hat{\phi}_n - \phi^*\|_{\mathfrak{X}/n}^2 &\leq \lambda_n\sum_{j=1}^{p}\lambda_{n,j}|\hat{\phi}_{n,j} - \phi_j^*| + 2\lambda_n\sum_{j=1}^{p}\lambda_{n,j}(|\phi_j^*| - |\hat{\phi}_{n,j}|) \\ &\leq 4\lambda_n\sum_{j\in S}\lambda_{n,j}|\hat{\phi}_{n,j} - \phi_j^*| \,,\end{aligned} \tag{34}$$

where the second inequality is obtained by decomposing the summation $\sum_{j=1}^{p}$ into $\sum_{j\in S} + \sum_{j\notin S}$, and using Cauchy-Schwarz inequality.

By assumption **(ii)**, and the fact that $\gamma(0) = \mathbb{E}|X_t|^2 = 1$, we have

$$\begin{aligned}\sum_{j\in S}|\hat{\phi}_{n,j} - \phi_j^*|^2 &\leq \sum_{j=1}^{p}(\hat{\phi}_{n,j} - \phi_j^*)^2 = \|\hat{\phi}_n - \phi^*\|_{\mathrm{diag}(\Gamma_p)}^2 \\ &\leq \frac{1}{\kappa_p}\|\hat{\phi}_n - \phi^*\|_{\Gamma_p}^2 \,.\end{aligned} \tag{35}$$

The proof is completed by applying the Cauchy-Schwarz inequality on (34), and by using assumption **(i)**.

□

We turn to the second lemma.

Lemma 6.4. *Let assumptions* **(i)**, **(ii)** *of Theorem 3.3 be in effect. Let C be a constant (given explicitly in the proof) depending on M only. Put $\epsilon = \lambda_n (sp^{-1})^{1/2}$. Then,*

$$\|\hat{\phi}_n - \phi^*\|^2_{\Gamma_p} \leq C\lambda_n^2 s\kappa_p^{-1} \ , \tag{36}$$

holds true on $\mathcal{I}_1 \cap \mathcal{I}_2$, where $\mathcal{I}_1$ is given by (33), and

$$\mathcal{I}_2 = \{M_p \leq \epsilon\} \ , \tag{37}$$

with

$$M_p = \max_{1\leq i,j\leq p} \left| \frac{\mathfrak{X}_{ij}}{n} - \gamma(i-j) \right| \ . \tag{38}$$

Proof. Note that

$$\left| \|\hat{\phi}_n - \phi^*\|^2_{\mathfrak{X}/n} - \|\hat{\phi}_n - \phi^*\|^2_{\Gamma_p} \right| \leq M_p \|\hat{\phi}_n - \phi^*\|_1^2 \ .$$

Therefore,

$$\begin{aligned} \|\hat{\phi}_n - \phi^*\|^2_{\mathfrak{X}/n} &\geq \|\hat{\phi}_n - \phi^*\|^2_{\Gamma_p} - M_p p^{1/2} \|\hat{\phi}_n - \phi^*\| \\ &\geq \|\hat{\phi}_n - \phi^*\|^2_{\Gamma_p} - M_p (p\kappa_p^{-1})^{1/2} \|\hat{\phi}_n - \phi^*\|_{\Gamma_p} \ . \end{aligned}$$

The first inequality follows since $\|a\|_1 \leq n\|a\|^2$, and the second inequality is satisfied under assumption **(ii)** (see (35)). Referring back to (32), we obtain, on $\mathcal{I}_1 \cap \mathcal{I}_2$,

$$\|\hat{\phi}_n - \phi^*\|^2_{\Gamma_p} \leq 2(1/2 + 2M)\lambda_n (s\kappa_p^{-1})^{1/2} \|\hat{\phi}_n - \phi^*\|_{\Gamma_p} \ .$$

Applying the inequality $2xy \leq 2x^2 + y^2/2$ on the right-hand side of the expression above (with $x = (1/2 + 2M)\lambda_n (s\kappa_p^{-1})^{1/2}$, and $y = \|\hat{\phi}_n - \phi^*\|_{\Gamma_p}$), we establish the statement of the Lemma, with $C = 4(1/2 + 2M)^2$. □

The rest of the proof of Theorem 3.3 is devoted to showing that indeed $\|\hat{\phi}_n - \phi^*\|^2_{\Gamma_p} \leq C\lambda_n^2 s k_p^{-1}$ holds on a negligible event, i.e., that the probability of the complement of $\mathcal{I}_1 \cap \mathcal{I}_2$ is negligible. We shall commence with $\mathcal{I}_2$.

We recall here the family of time series $\{X_t\}$, denoted by $\mathcal{H}_\rho(l, L)$, for some $\rho > 1$, $0 < l < 1$, and $L > 1$ (Section 3.2). The family consists of all stationary Gaussian time series with $\mathbb{E}X_t = 0$, $\mathbb{E}|X_t|^2 = 1$, and enjoys an exponential decay of the strongly mixing coefficients (see (8)).

Lemma 6.5. *Assume that $\epsilon = \lambda_n (s/p)^{1/2} \leq Dn^{-2/5}$, where $D = (C_1^3 C_2 \beta_1^2 \beta_2^3)^{1/5}$, with C_1 and C_2 two constants explicitly specified in the proof. Then,*

$$\mathbb{P}(\mathcal{I}_2^c) \leq p^2 \exp\left\{ -n\lambda_n^2 (s/p^2)/(4C_1\beta_1\beta_2) \right\} \ .$$

Proof. We begin with

$$\mathbb{P}(|\sum_{t=1-i}^{n-i} Y_t| > \epsilon) \ ,$$

where

$$Y_t \equiv Y_{t,i,j} = \frac{1}{n}(X_t X_{t+i-j} - \gamma(i-j)) \ . \tag{39}$$

The proof is based on an application of the pair of lemmas 6.6 and 6.7, after noticing that

$$\mathbb{P}(\mathcal{I}_2^c) = \mathbb{P}(M_p > \epsilon) \leq \sum_{i,j=1}^{p} \mathbb{P}\left(|\sum_{t=1-i}^{n-i} Y_t| > \epsilon \right) \ .$$

Define $k = i - j$. It is enough to consider only $k \geq 0$ ($i \geq j$), since $\mathfrak{X}_{ij}$ and $\gamma(i-j)$ are symmetric. By the same argument below expression (39) in Goldenshluger and Zeevi (2001), one may notice that $\{Y_t\}$ is strongly mixing with the rate $\alpha_Y(m) \leq \alpha_X(m-k)$ for $m > k$, and $\alpha_Y(m) \leq 1/4$ (see Bradley (2005)), but for our purposes in would be enough to bound $\alpha_Y(m)$, for $m > k$, by simply 1.

We shall make use of the following two lemmas, adapted from Goldenshluger and Zeevi (2001).

Lemma 6.6. *Suppose $\{X_t\}$ is a strongly mixing time series, $S_n = \sum_{t=1}^n$, and $cum_r(S_n)$ is the rth order cumulant of S_n. For $\nu > 0$ define the function*

$$\Lambda_n(\alpha_X, \nu) = \max\left\{1\,,\ \sum_{m=1}^{n} (\alpha_X(m))^{1/\nu}\right\}\,.$$

If, for some $\mu \geq 0$, $H > 0$

$$\mathbb{E}|X_t|^r \leq (r!)^{\mu+1} H^r \quad , \quad t = 1, \ldots, n,\ r = 2, 3, \ldots\ ,$$

then $|cum_r(S_n)| \leq 2^{r(1+\mu)+1} 12^{r-1} (r!)^{2+\mu} H^r [\Lambda_n(\alpha_X, 2(r-1))]^{r-1} n$.

Lemma 6.7. *Let Y be a random variable with $\mathbb{E}Y = 0$. If there exist $\mu_1 \geq 0$, $H_1 > 0$ and $\Delta > 0$ such that*

$$|cum_r(Y)| \leq \left(\frac{r!}{2}\right)^{1+\mu_1} \frac{H_1}{\Delta^{r-2}}\,, \quad r = 2, 3, \ldots\ ,$$

then

$$\mathbb{P}(|Y| > y) \leq \begin{cases} \exp\{-y^2/(4H_1)\} & 0 \leq y \leq (H_1^{1+\mu_1}\Delta)^{1/(2\mu_1+1)} \\ \exp\{-(y\Delta)^{1/(1+\mu_1)}/4\} & y \geq (H_1^{1+\mu_1}\Delta)^{1/(2\mu_1+1)}\,. \end{cases}$$

Back to the proof of Lemma 6.5. Absolute moment of Y_t are bounded as follows:

$$\begin{aligned} \mathbb{E}|Y_t|^r &\leq n^{-r} 2^{r-1} \left[\mathbb{E}|X_t X_{t+k}|^r + |\gamma(k)|^r\right] \\ &\leq n^{-r} 2^{r-1} \left[\left(\mathbb{E}|X_t|^{2r} \mathbb{E}|X_{t+k}|^{2r}\right)^{1/2} + \gamma(0)\right] \\ &\leq r!(4/n)^r\,. \end{aligned}$$

The second inequality follows by the Cauchy-Schwarz inequality together with the inequality $(a+b)^j \leq 2^{j-1}(a^j + b^j)$, and the last inequality follows by the assumed Gaussianity of X_t, and the inequality $\binom{2r}{r} \leq 2^{2r}$. We have

$$\begin{aligned} \sum_{m=1}^{n} (\alpha_X(m))^{1/2(r-1)} &\leq k + \left(\frac{2Ll}{l(\rho-1)}\right)^{1/(r-1)} \sum_{m=1}^{n-k} \rho^{-m/2(r-1)} \\ &\leq k + \left(\frac{2Ll}{l(\rho-1)}\right)^{1/(r-1)} \left(1 + \frac{2(r-1)}{\log\rho}\right)\,, \end{aligned}$$

The first inequality utilizes the relationship between $\alpha_Y(m)$ and $\alpha_X(m)$, and inequality (8). The second inequality uses geometric series expression together with the inequality $\rho^x - 1 \geq x \log\rho$, for all $x \geq 0$.

Therefore, defining $\hat{k} = k$ if $k > 0$, and $\hat{k} = 1$, if $k = 0$, we obtain, after some manipulations, similar to those in Goldenshluger and Zeevi (2001),

$$[\Lambda_n(\alpha_X, 2(r-1))]^{r-1} \leq 12^{r-1} r! (\hat{k}\beta_1)^{r-1} \beta_2\,,$$

for two constants β_1 and β_2, given, respectively, by $1 + 1/\log\rho$ and $1 + L\rho/l(\rho-1)$ (see (9)). The bound results from the inequalities $(a+b)^j \leq 2^{j-1}(a^j + b^j)$, $n^n \leq n! e^n$, and other trivial inequalities.

Applying Lemma 6.6 (with $\mu = 0$ and $H = 4/n$) we have $|cum_r(\sum_{t=1-i}^{n-i})Y_t| \leq$ RHS, where RHS can be put in the form $(r!/2)^3 H\Delta^{2-r}$, with $H_1 = C_1\beta_2(\hat{k}\beta_1/n)$, $\Delta = C_2(\hat{k}\beta_1/n)^{-1}$, and $C_1 = 2^{10}12^2$, $C_2 = 2^{-3}12^{-2}$. Now, applying Lemma 6.7 (with $\mu_1 = 2$, and H_1 and Δ as above) we obtain:

$$\mathbb{P}\left(|\sum_{t=1-i}^{n-i} Y_t| > y\right) \leq \begin{cases} \exp\left\{-y^2 n/(4C_1\hat{k}\beta_1\beta_2)\right\} & 0 \leq y \leq D\hat{k}^{2/5}n^{-2/5} \\ \exp\left\{-\frac{1}{4}\left(\frac{C_2}{\hat{k}\beta_1}\right)^{1/4}(yn)^{1/3}\right\} & y \geq D\hat{k}^{2/5}n^{-2/5}\ , \end{cases} \tag{40}$$

where $D = (C_1^3 C_2 \beta_1^2 \beta_2^3)^{1/5}$. The proof is completed by applying the moderate deviation part in (40) with $y = \epsilon$, and by noticing that $1 \leq \hat{k} \leq p$.

□

We turn to evaluate the probability of the complement of the event $\mathcal{I}_1$.

Lemma 6.8. *For all $0 < c < \infty$ and $y > \sigma^2(n + Dn^{3/5})$ (where D is given by (9)),*

$$\mathbb{P}(\mathcal{I}_1^c) \leq 6p\exp\left\{-F_1\min\left\{(\sigma^{-2}y - n)^{1/3}, c^2\sigma^{-2}, \frac{n^2\lambda_n^2\lambda_{\min}^2}{y + cn\lambda_n\lambda_{\max}/2}\right\}\right\}\ ,$$

where $F_1 = \min\left\{(C_2/\beta_1)^{1/4}/4, 2^{-9}, 8^{-1}\right\}$.

Proof. Let $V_n^2 = \sigma^2\sum_{t=1}^n X_{t-i}^2 = \sigma^2\sum_{t=1-i}^{n-i} X_t^2$. Fix a $y > \sigma^2(n + Dn^{3/5})$ and a $0 < c < \infty$. Denote by $\tilde{\mathcal{I}}_1$ the event $\mathcal{I}_1$ (see (33)) with the absolute value removed. We begin by writing:

$$\begin{aligned} \mathbb{P}(\tilde{\mathcal{I}}_1^c) &\leq \sum_{j=1}^p \mathbb{P}\big(\frac{2}{n}\sum_{t=1}^n X_{t-j}Z_t > \lambda_n\lambda_{n,j}\big) \\ &\leq \sum_{j=1}^p \mathbb{P}\Big(\bigcup_{n=1}^\infty \big\{\frac{2}{n}\sum_{t=1}^n X_{t-j}Z_t > \lambda_n\lambda_{n,j}\ ,\ V_n^2 \leq y\big\}\Big) + p\,\mathbb{P}(V_n^2 > y) \\ &=: I_1 + I_2\ . \end{aligned}$$

Clearly, I_1 satisfies $I_1 \leq I_{11} + I_{12}$, with

$$\begin{aligned} I_{11} &= \sum_{j=1}^p \mathbb{P}\Big(\bigcup_{n=1}^\infty \big\{\frac{2}{n}\sum_{t=1}^n X_{t-j}Z_t > \lambda_n\lambda_{n,j}\ ,\ V_n^2 \leq y\big\}\ ,\ \bigcap_{r=3}^\infty \big\{|X_{t-j}|^{r-2}\mathbb{E}|Z_t|^r \leq \frac{r!}{2}\sigma^2 c^{r-2}\big\}\Big)\ , \\ I_{12} &= \sum_{j=1}^p \mathbb{P}\Big(\bigcup_{r=3}^\infty \big\{|X_{t-j}|^{r-2}\mathbb{E}|Z_t|^r > \frac{r!}{2}\sigma^2 c^{r-2}\big\}\Big)\ . \end{aligned}$$

We analyze $\mathbb{P}(\tilde{\mathcal{I}}_1^c)$ by investigating I_{11}, I_{12} and I_2 separately.

For I_2, we recall that $Y_t \equiv Y_{t,i,i} = (X_t^2 - \gamma(0))/n$ (see (39) and the remark below) is strongly mixing with exponential decay rate. Therefore, by the large deviation part in (40) (with $\hat{k} = 1$),

$$\begin{aligned} \mathbb{P}(V_n^2 > y) &\leq \mathbb{P}(|V_n^2 - n\sigma^2| > y - n\sigma^2) \\ &= \mathbb{P}(|\sum_{t=1-i}^{n-i} Y_t| > \sigma^{-2}n^{-1}y - 1) \\ &\leq \exp\left\{-\frac{1}{4}\left(\frac{C_2}{\beta_1}\right)^{1/4}(\sigma^{-2}y - n)^{1/3}\right\}\ . \end{aligned}$$

For I_{12}, we use the bound $\mathbb{E}|Z_t|^{2r} \leq \sigma^{2r}r!2^{2r}$ (and the Cauchy-Schwarz inequality) to obtain

$$\big\{|X_{t-j}|^{r-2}\mathbb{E}|Z_t|^r > \frac{r!}{2}\sigma^2 c^{r-2}\big\} \subset \big\{|X_{t-j}| > 2^{-(1+r)/(r-2)}\sigma^{-1}c\big\}\ .$$

Therefore, noticing that $\{2^{-(1+r)/(r-2)}\}_{r=3}^{\infty}$ is an increasing sequence, we have

$$I_{12} \leq \sum_{j=1}^{p} \mathbb{P}\big(|X_{t-j}| > 2^{-4}\sigma^{-1}c\big) \leq (2/\pi)^{1/2} p \exp\{-2^{-8}c^2/2\sigma^2\} \, .$$

For I_{11}, we use the following theorem which is a Bernstein's type of an inequality for martingales.

Theorem 6.9 (De La Peña (1999)). *Let $\{M_n, \mathcal{F}_n\}$ be a martingale, with difference $\Delta_n = M_n - M_{n-1}$. Define $V_n^2 = \sum_{i=1}^n \sigma_i^2 = \sum_{i=1}^n \mathbb{E}(\Delta_i^2 \,|\, \mathcal{F}_{i-1})$. Assume that $\mathbb{E}(|\Delta_i|^r \,|\, \mathcal{F}_{i-1}) \leq (r!/2)\sigma_i^2 c^{r-2}$ a.e. for $r \geq 3$, $0 < c < \infty$. Then, for all $x, y > 0$,*

$$\mathbb{P}\Big(\bigcup_{n=1}^{\infty} \{M_n > x \, , \, V_n^2 \leq y\} \Big) \leq \exp\left\{ -\frac{x^2}{2(y + cx)} \right\} \, . \tag{41}$$

Recall that $\sum_{t=1}^n X_{t-j} Z_t$ is a martingale (see (18)). Then, simple application of the above theorem, with $x = n\lambda_n \lambda_{n,j}/2$, leads to

$$I_{11} \leq p \exp\left\{ -\frac{n^2 \lambda_n^2 \lambda_{\min}^2}{8(y + cn\lambda_n \lambda_{\max}/2)} \right\} \, .$$

Lemma 6.8 now follows by collecting the bounds of I_{11}, I_{12}, and I_2, and by symmetry.

□

The proof of theorem 3.3 is now complete by virtue of Lemma 6.3, Lemma 6.4, Lemma 6.5, and Lemma 6.8.

□

References

Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability surveys*, Vol. 2, 107-144.

Bunea, F., Tsybakov, A. and Wegkamp M. (2007a). Aggregation for gaussian regression. *The Annals of Statistics*, 35(4), 1674-1697.

Bunea, F., Tsybakov, A. and Wegkamp M. (2007b). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, Vol. 1, 169-194.

Davis, R. A. and Brockwell, P. J. (1991). *Time series: theory and methods*, Springer-Verlag, New York.

De La Peña, V. A. (1999). A general class of exponential inequalities for martingales and ratios. *The Annals of Statistics*, 27(1), 537-564.

Efromovich, S. (1998). Data-driven efficient estimation of the spectral density. *Journal of the American Statistical Association*, 93(442), 762-769.

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

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3), 928-961.

Goldenshluger, A. and Zeevi, A. (2001). Nonasymptotic bounds for autoregressive time series modeling. *The Annals of Statistics*, 29(2), 417-444.

Hall, P. and Heyde, C. C. (1980). *Martingale limit theory and its application*, Academic Press Inc., New York.

Ing, C and Wei, C. (2005). Order selection for same-realization predictions in autoregressive processes. *The Annals of Statistics*, 33(5), 2423-2474.

Lafferty, J., Liu, H., Ravikumar, P. and Wasserman, L. (2007). Sparse additive models. Submitted.

Nardi, Y. and Rinaldo, A. (2008). The log-linear group lasso estimator and its asymptotic properties. Submitted.

Rothman A. J., Bickel, P. J., Levina, E. and Zhu, J. (2007). Sparse permutation invariant covariance estimation. Technical Report # 467, Department of Statistics, University oF Michigan.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics*, 8, 147-164.

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

Wainwright, M. J. (2006). Sharp thresholds for high-dimensional and noisy recovery of sparsity, Technical Report 708, Department of Statistics, UC Berkeley.

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.*, 69(1), 63-78.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B.*, 68 (1) , 4967.

Zhao, P. and Bin, Y. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7, 2541-2563.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418-1429.