A study of the asymptotic properties of Lasso for correlated data

F Bunea and S Gupta
Department of Statistics
Florida State University

Abstract

In this paper we investigate post-model selection properties of $\ell_1$ penalized weighted least squares estimators in regression models with a large number of variables $M$ and correlated errors. We focus on correct subset selection and on the asymptotic distribution of the penalized estimators. We first prove the result under the general setup when the errors have a weak dependency structure (Doukhan 1996, see [5]), and then discuss this result for a special case i.e. AR(1). For the problem of model selection, the number of regression variables $M$ is allowed to exceed the sample size $n$. We further investigate the asymptotic distribution of our estimates, when $M < n$, and show that under appropriate choices of the tuning parameters the limiting distribution is multivariate normal. This generalizes the result of Knight and Fu (2000), see [8], obtained for regression models with independent errors, to the case of correlated errors. We also establish that if the covariance structure is unknown but an estimator, which is componentwise consistent, is used instead, then we get result same as what we would have got if the covariance structure was known.

1 Introduction

Let $(X_1, Y_1), ..., (X_n, Y_n)$ be a sample of random pairs distributed as $(X, Y) \in (\mathcal{X}, \mathbb{R})$, where $\mathcal{X}$ is a Borel subset of $\mathbb{R}^M$; we denote the probability measure of $X$ by $\mu$. The response variable $Y$ can be represented as

$$Y = \mathbb{E}(Y|X) + Y - \mathbb{E}(Y|X).$$

Henceforth $Y - \mathbb{E}(Y|X)$ shall be denoted as $\epsilon$ and shall be referred to as the random error. We assume that

$$\mathbb{E}(Y|X) = X' \beta.$$ 

This assumption enables us to represent the $i^{th}$ response variable $Y_i$ as

$$Y_i = X_i' \beta + \epsilon, \; \forall \; i = 1, ..., n.$$ 

where, $X_i$ is a $M$ dimensional vector $(X_{i1}, ..., X_{iM})$ and, as already mentioned above, it belongs to $\mathcal{X}$ and $\beta \in \mathbb{R}^M$ are the regression coefficients. Thus

$$\tilde{Y} = X \beta + \tilde{\epsilon}$$

where

$$\tilde{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} X_1' \\ \vdots \\ X_n' \end{pmatrix} \quad \text{and} \quad \tilde{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$
In this paper we shall be investigating the statistical problems of model selection consistency of the estimated \( \beta \) and finding its asymptotic distribution when the underlying errors are weakly dependent. We estimate \( \beta \) by minimizing the penalized least squares objective function

\[
Z_n(\beta) = \frac{1}{n}(\tilde{Y} - X\beta)'(\tilde{Y} - X\beta) + \lambda_n \sum_{j=1}^{M} |\beta_j|, \tag{2}
\]

that is,

\[
\hat{\beta}_n = \arg\min_{\beta \in \mathbb{R}^M} Z_n(\beta) \tag{3}
\]

where, \( \lambda_n \) is the penalty parameter. The value \( \hat{\beta}_n \), at which \( Z_n(\beta) \) is minimized, is called the Lasso (Least Absolute Shrinkage and Selection Operator) estimate of \( \beta \). Lasso was first proposed by Tibshirani in 1996 see [14]. Let

\[
\Sigma_n = \text{Cov}(\epsilon_1, \ldots, \epsilon_n)
\]

be the covariance matrix of the random errors \( \epsilon_i \)'s. From (1) we can write,

\[
\Sigma_n^{-1/2} \tilde{Y} = \Sigma_n^{-1/2} X\beta + \Sigma_n^{-1/2} \tilde{\epsilon}.
\]

Let

\[
\tilde{Y}^* = \Sigma_n^{-1/2} \tilde{Y},
\]

\[
X^* = \Sigma_n^{-1/2} X,
\]

and

\[
\tilde{\epsilon}^* = \Sigma_n^{-1/2} \tilde{\epsilon}.
\]

Now

\[
\text{Cov}(\tilde{\epsilon}^*) = I_n
\]

with \( I_n \) being the \( n \) dimensional identity matrix. Thus we see that the components of the error vector \( \tilde{\epsilon}^* \) are uncorrelated. If we were to use the same criteria for estimating the regression coefficients \( \beta \) as we had used in (2) we would have to minimize

\[
\frac{1}{n}(\tilde{Y}^* - X^* \beta)'(\tilde{Y}^* - X^* \beta) + \lambda||\beta||_{\ell_1}
\]

with respect to \( \beta \). Observe that

\[
(\tilde{Y}^* - X^* \beta)'(\tilde{Y}^* - X^* \beta) = (\tilde{Y} - X\beta)'\Sigma_n^{-1}(\tilde{Y} - X\beta).
\]

This acts as a motivation for us to alter the minimization criterion for estimating the regression coefficients, thus when the errors are non i.i.d our estimate of the regression coefficients \( \beta \) is

\[
\hat{\beta}_n = \arg\min_{\beta \in \mathbb{R}^M} \frac{1}{n}(\tilde{Y} - X\beta)'\Sigma_n^{-1}(\tilde{Y} - X\beta) + \lambda_n||\beta||_{\ell_1} \tag{4}
\]

Observe that (3) can be obtained from (4), by taking \( \Sigma_n = \sigma^2 I_n \). Many statistical properties of the Lasso estimate, like the model selection consistency of the set of nonzero components, \( \ell_2 \) consistency, the asymptotic distribution, and the behaviour of the prediction loss of the Lasso estimate have been under investigation but
the literature concerning theoretical results on asymptotic properties of the Lasso estimator with dependent errors is very limited. When the observations $Y_i$s are independent, the problem of the asymptotic distribution of Lasso-type estimators has been studied by Knight and Fu, see [8], in the setting when $M \leq n$ and $M$ is fixed. Zou (2006), see [23], considers the same problem but the Lasso penalty is weighted to ensure model selection consistency. Huang et al., see [7], extends these results to the situation when $M > n$ and proves asymptotic normality of marginal Lasso estimator, which is the minimizer of

$$U_n(\beta) = \sum_{i=1}^{M_n} \left( y_i - \hat{x}_{ij} \beta_j \right)^2 + \lambda_n \sum_{j=1}^{M_n} |\beta_j|^\gamma.$$

In Section 2 we establish the asymptotic distribution of the estimator, and we show that

$$\sqrt{n}(\hat{\beta} - \beta)$$

has an asymptotic ($n \to \infty$) normal distribution for general weakly dependent errors if $\lambda_n = O\left(\frac{1}{\sqrt{n \ln n}}\right)$. This extends to the dependent error setting the results of Knight and Fu (2000), see [8]. A related result has been obtained by Wang Li and Tsai (2007), see [20], for a different estimator, the adaptive Lasso, for a particular family of weak dependence namely $AR(p)$. All these results including ours are obtained for $M < n$, where $M$ is the number of parameters and $n$ is the number of observations. The strength of our result lies in the fact that our problem is solved under the assumption that the errors are weakly dependent, which generalises the work of Wang Li and Tsai as they have solved the problem in the setting where the errors are $AR(p)$ i.e. autoregressive model of order $p$. In Section 2.1 we derive our results assuming that $\Sigma^{-1}_n$ is known, but for all practical purposes, one needs an estimator of $\Sigma^{-1}_n$ to compute the estimator of $\beta$ and most importantly we have shown even if the covariance matrix is unknown we can do just as well by using a suitable estimator. In Section 2.2 we show that if $\Sigma^{-1}_n$ is replaced by $\hat{\Sigma}^{-1}_n$ in definition (2) of the estimator of $\beta$, where $\hat{\Sigma}^{-1}_n$ is a consistent estimator of $\Sigma^{-1}_n$, then the results of Section 2.1 continue to hold. In Section 2.3 we provide an illustrative example.

Althought the Lasso is a potent technique for regularization and model selection this method is not free from drawbacks. One of them is that this method is not always model selection consistent. It has been shown by Leng, Lin and Wahba (2006), see [9], that in a linear regression setup when the random errors are $N(0, \sigma^2)$ and i.i.d and $M \leq n$, then if prediction accuracy is used to choose the tuning/penalty parameter $\lambda_n$, this procedure is not consistent in terms of model selection. Zou (2006) working in a linear regression setup where the errors are i.i.d $N(0, \sigma^2)$ and $M \leq n$ establishes a necessary condition which ensures model selection consistency of the Lasso estimator.

The method of using weighted Lasso to ensure model selection consistency was also proposed by Zou(2006), see [23], and Bunea et al.(2007) in [2]. Zou worked in a linear model setting and assumed that $M < n$ and Least Squares Estimate exists but Bunea et al did not make any such assumption and worked in a nonparametric setting of which linear models turns out to be a special case. Meinshausen and Buhlmann (2006), see [10], showed that for neighborhood selection in the Gaussian graphical models, under a neighborhood stability condition on the design matrix and certain additional regularity conditions, the Lasso is consistent even when the number of variables tends to infinity at a rate faster than $n$. Zhao and Yu (2006), see [22], formalized the neighborhood stability condition in the context of linear regression models as a strong irrepresentable condition in the linear regression setup where the errors are i.i.d $(0, \sigma^2)$. They proposed two conditions, namely the strong irrepresentable condition and the weak irrepresentable condition, depending mainly on the covariance of the predictor variables. Here they also introduce for the first time strong sign consistency and general sign consistency, concepts which are more general than model selection consistency.
They established consistency of selection when \( M > n \) for fixed design linear regression models and a target set \( J^* \) that corresponds to coefficients \( \lambda^*_j \) that are assumed to be lower bounded by a sequence of order \( O(n^{-\delta/2}) \) for \( 0 < \delta < 1 \). For a related notion of consistency in the fixed design case with Gaussian errors see Wainwright (2007), see [19]. Here Wainwright provides a necessary and sufficient condition to ensure an asymptotic perfect recovery of the sparsity pattern. Wainwright (2006), see [18], has proved \( (M > n) \) and errors i.i.d \( N(0, \sigma^2) \) setting that for a broad class of random Gaussian ensembles based on covariance matrices satisfying mutual incoherence conditions, the LASSO estimate is sign consistent. This work is very similar to that by Zhao and Yu (2006), see [22].

Other notions of consistency have been introduced and discussed by many others including Zhang and Huang (2007), see [21], and Meinshausen and Yu (2007), see [11]. Meinshausen and Yu (2007), see [11], have proved the \( \ell_2 \) consistency of the Lasso estimator and have also proposed a two step procedure to estimate the unknown parameter. They established the sign consistency of the resulting estimator. Bunea et al. (2007), see [2], has proved model selection consistency and oracle inequalities in a nonparametric setting when \( M > n \). Bunea (2008), see [1], proved the model selection consistency under a very general condition which was termed as the Condition Stabil. These are by far the most general conditions (and also quite easy to verify) available in literature. We have tried to prove our results assuming these conditions. For the purpose of proof in this essay we actually need Condition Identif which is a special case of Condition Stabil. The fact that Condition Identif implies Condition Stabil is shown in Lemma 2.1 of [1]. Some other papers aimed at understanding the Lasso estimator from the point of view of model selection are Donoho et al., see [4], Tropp, see [16] and [15]. Wang, Li and Tsai (2007), see [20], has proved models election consistency and asymptotic normality for Lasso estimates when \( M < n \) and the errors are \( AR(q) \). Wang et al. considered observations being generated from the model

\[
y_t = x_t^\prime \beta + e_t,
\]

where,

\[
x_t = (x_{t1}, ..., x_{tM})
\]

is the \( M \)-dimensional regression covariate and

\[
\beta = (\beta_1, ..., \beta_M)
\]

is the \( M \)-dimensional regression coefficient and

\[
e_t = \sum_{j=1}^{q} \phi_j e_{t-j} + \varepsilon_t.
\]

Note that \((\phi_1, ..., \phi_q)\) is the autoregression coefficient and \( \varepsilon_t \) are i.i.d random variables with mean 0 and variance \( \sigma^2 \). But what differentiates our work from Wang et al.’s results are: (1) We assumed that errors are weakly dependent, and \( AR(q) \) is a special case of it. (2) We also proved both model selection consistency and asymptotic normality in the setting where \( M > n \) whereas their results were only confined to the setting where \( M < n \). (3) They have proved that the vector of zero coefficients in the true model, that is

\[
P(\hat{\beta}_{S^c} = 0) \rightarrow 1,
\]

where, \( S = \{j : \beta_j \neq 0\} \), tends to zero as \( n \) goes to \( \infty \) which is slightly weaker that model selection consistency, that is the result which we have proved. Tropp, see[15] and [16], shows that the Lasso estimates are near minimax optimal. The studies by Bunea et al., see [2], Greenshtein and Ritov, see [6], van de Geer,
see [17], have focussed mainly on the behavior of the prediction loss. Bunea et al, see [2], obtained results
for random design and sharp bounds for the $\ell_1$ distance between the vector $\beta$ and its Lasso estimate $\hat{\beta}$.
Prediction loss for high dimensional regression under an $\ell_1$ penalty has been studied for a quadratic loss
function in Greenshtein and Ritov, see [6], and for general Lipschitz loss functions in van de Geer, see [17].
In Section 3 we study the consistency of selection that is,
$$
P(\hat{I} \neq I^*) \to 0 \text{ as } n \to \infty,
$$
where,
$$
\hat{I} = \{ j : \hat{\beta}_j \neq 0 \}.
$$
(5)
and
$$
I^* = \{ j : \beta_j \neq 0 \}.
$$
(6)
that is $k^*$ is the cardinality of $I^*$. This is essentially an extension to the non i.i.d setup of an existing work
by Bunea, see [1], for the general class of weakly dependent data and general $M$. In particular $M > n$ is
allowed. We show that the estimator is consistent under a more general conditions on the design matrix
than that considered in the previous section, and relaxes the need for a positive definite matrix needed for
the study of asymptotic distribution. We give the results for the general weakly dependent error structure.
The main novelty here is the usage of a concentration inequality for weakly dependent errors. As in Section
2 we also discuss the quality of $\hat{I}$ when $\Sigma_n^{-1}$ is replaced by $\hat{\Sigma}_n^{-1}$. In Section 4 we present a brief summary
along with some concluding remarks and in the Appendices we present the proofs of some of the results used
in Sections 2 and 3. It should be noted that, henceforth, whenever we refer to Lasso estimate we shall be refering to (4).

1.1 Weak Dependence
As we had mentioned before that the errors $\epsilon_i$s are assumed to be weakly dependent. This is a very general
notion of independence see Doukhan and Louhichi, see [5]. It can be shown that, see Nze, Buhlmann and
Doukhan, see [12], that $AR(p), ARMA(p,q)$ satisfy the condition of weak dependence. For the sake of
completeness and for the convenience of the readers we formally state the definition of weak dependence.
Let $E$ be some Euclidean Space $\mathbb{R}^d$ endowed with its Euclidean norm $||.||$. We consider a sequence of $E$
valued random variables $(\epsilon_n)_{n \in \mathbb{N}}$. We define $L^\infty$ as the set of measurable and bounded numerical functions
on some space $\mathbb{R}^u$ and its norm classically written as $||.||_{\infty}$. Moreover, let $u \in \mathbb{N}^*$ be a positive integer we
endow the set $F = E^u$ with the norm
$$
||(x_1, ..., x_u)||_F = ||x_1|| + ... + ||x_u||.
$$
Let now $h : F = E^u \to \mathbb{R}$ be a numerical function on $F$, we set
$$
\text{Lip}(h) = \sup_{x \neq y} \frac{|h(x) - h(y)|}{||x - y||_F},
$$
the Lipschitz modulus of $h$. Define
$$
\mathcal{L} = \bigcup_{u=1}^{\infty} \{ h \in L^\infty(\mathbb{R}^u, \mathbb{R}) : ||h||_{\infty} \leq 1, \text{ Lip}(h) < \infty \}
$$
(8)
The sequence \((\epsilon_n)_{n \in \mathbb{N}}\) is \(s\)-weakly (resp. \(a\)-weakly) dependent, if for some sequence \(\theta = (\theta_r)_{r \in \mathbb{N}}\) decreasing to zero at infinity and any \((u+2)\)-tuple \(i_1, \ldots, i_u, j_1, j_2\) with \(i_1 \leq \ldots \leq i_u < i_u + r \leq j_1 \leq j_2\), and \(h \in L^\infty\) satisfies \(||h||_\infty \leq 1\) and \(k \in \mathcal{L}\),

\[
|\text{Cov}(h(\epsilon_{i_1}, \ldots, \epsilon_{i_u}), k(\epsilon_{j_1}, \epsilon_{j_2}))| \leq \text{Lip}(k)\theta_r
\]

and, respectively, for \(h, k \in \mathcal{L}\)

\[
|\text{Cov}(h(\epsilon_{i_1}, \ldots, \epsilon_{i_u}), k(\epsilon_{j_1}, \epsilon_{j_2}))| \leq \text{Lip}(h)\text{Lip}(k)\theta_r.
\]

Weak dependence conditions are shown to hold in, either causal or noncausal frames in Doukhan and Louhichi, see [5]. For this, consider also \(v\)-tuple \((j_1, \ldots, j_v)\) with \(i_1 \leq \ldots \leq i_u < i_u + r \leq j_1, \ldots, \leq j_v\), the such weak dependence conditions are defined for functions \(h\) and \(k\) defined on \(E^u\) and \(E^v\), respectively, through inequalities

\[
|\text{Cov}(h(\epsilon_{i_1}, \ldots, \epsilon_{i_u}), k(\epsilon_{j_1}, \epsilon_{j_2}))| \leq v\text{Lip}(k)\theta_r
\]

if \(||h||_\infty \leq 1\) and \(k \in \mathcal{L}\), and,

\[
|\text{Cov}(h(\epsilon_{i_1}, \ldots, \epsilon_{i_u}), k(\epsilon_{j_1}, \epsilon_{j_2}))| \leq \min(u, v)\text{Lip}(h)\text{Lip}(k)\theta_r.
\]

if \(h, k \in \mathcal{L}\).

2 Results concerning asymptotic distribution of the Lasso estimate

In this section we present the results concerning the asymptotic distribution of \(\sqrt{n}(\hat{\beta}_n - \beta)\), where \(\hat{\beta}_n\) is the Lasso estimate, as defined in (4). Althought this problem has been already been much explored, as we had discussed in Section 1, what sets our work apart from the rest is that this is the first time this problem is being investigated under the assumption that the errors in the linear equation, (1), are weakly dependent. We have already presentend a formal definition of weak dependence as proposed by Doukhan (1996), see Section 1.1. In Section 2.1 we establish the result under the assumption that the covariance structure \(\Sigma_n\) is known and then in Section 2.2 we extend this result to the setting where \(\Sigma_n\) is unknown and a consistent estimate of this matrix, \(\hat{\Sigma}_n\) is used instead.

2.1 Study of asymptotic distribution when \(\Sigma_n\) is known

Before stating the result let us first state the assumptions under which the results hold good:

**Assumption 2.1**

\[
\frac{1}{n} \sum_{i=1}^{n} X_i \Sigma_n^{-1} X_i^T \rightarrow C
\]

as \(n \rightarrow \infty\), where \(X_i\) is the \(i^{th}\) row of the design matrix \(X\), \(C\) is a positive definite matrix and as mentioned before \(\text{Cov}(\epsilon_1, \ldots, \epsilon_n) = \Sigma_n\). This condition can be viewed as something more general than what has been used by Knight and Fu, see . Knight and Fu were working under the assumption that the errors are independent, which means \(\Sigma_n = I_n\).

6
Assumption 2.2 The random errors $\epsilon_i$ are weakly dependent. See Section 1.1 for the definition of weak dependence.

Assumption 2.3 The assumptions of Theorem A.2 in Appendix A are satisfied, with

$$g_{k,n}(x) = \frac{1}{\sqrt{n}} u^* v_k$$

where $v_k$ is the $k^{th}$ column of the matrix $X^\prime \Sigma_n^{-1}$ and $u$ is a fixed vector belonging to $\mathbb{R}^M$.

**Theorem 2.1:** Suppose $\sqrt{n}\lambda_n \rightarrow \lambda_0$ as $n \rightarrow \infty$ and let $C$ be non singular, and Assumptions 2.1, 2.2 and 2.3 are satisfied. Then

$$\sqrt{n}(\hat{\beta}_n - \beta) \overset{d}{\rightarrow} \text{argmin}(V)$$

where,

$$V(u) = -2u^\prime S + u^\prime Cu + \lambda_0 \sum_{j=1}^M [u_j \text{sgn}(\beta_j)I(\beta_j \neq 0) + |u_j|I(\beta_j = 0)]$$

and $S$ has a $M$ dimensional $N(0, \sigma^2 C)$ distribution. ■

**Note:** We can choose $\lambda_n = O(\frac{1}{\ln \sqrt{n}})$. Observe that we have put an $\ln n$ term in the denominator but could have put a polynomial of $n$ like $n^\delta$ where $\delta > 0$ or an exponential term.

**Proof:** Let

$$V_n(u) = \sum_{i=1}^n [(\epsilon_i^* - u^\prime X_i^*/\sqrt{n})^2 - \epsilon_i^{*2}] + n\lambda_n \sum_{j=1}^M [\beta_j + \frac{u_j}{\sqrt{n}} - |\beta_j|].$$

We claim that $V_n(u)$ is minimized at $\sqrt{n}(\hat{\beta}_n - \beta)$. Observe that

$$V_n(u) = \sum_{i=1}^n (\epsilon_i^* - u^\prime X_i^*/\sqrt{n})^2 + n\lambda_n \sum_{j=1}^M |\beta_j + \frac{u_j}{\sqrt{n}}| - \left\{ \sum_{i=1}^n \epsilon_i^{*2} + n\lambda_n \sum_{j=1}^M |\beta_j| \right\}$$

$$= \sum_{i=1}^n (Y_i^* - X_i^* \beta - u^\prime X_i^*/\sqrt{n})^2 + n\lambda_n \sum_{j=1}^M |\beta_j + \frac{u_j}{\sqrt{n}}| - \left\{ \sum_{i=1}^n (Y_i^* - X_i^* \beta)^2 + n\lambda_n \sum_{j=1}^M |\beta_j| \right\}$$

$$= U_n(u) - U$$

where,

$$U_n(u) = \sum_{i=1}^n (Y_i^* - X_i^* \beta - u^\prime X_i^*/\sqrt{n})^2 + n\lambda_n \sum_{j=1}^M |\beta_j + \frac{u_j}{\sqrt{n}}|$$

7
and

\[ U = \sum_{i=1}^{n} (Y_i^* - X_i^* \beta)^2 + n\lambda_n \sum_{j=1}^{M} |\beta_j|. \]

Since \( U \) is free of \( u \), minimizing \( V_n(u) \) with respect to \( u \) is equivalent to minimizing \( U_n(u) \) with respect to \( u \).

Thus in order to show that \( \sqrt{n}(\hat{\beta}_n - \beta) \) is the minimizer of \( V_n(u) \) it suffices to show that it is the minimizer of \( U_n(u) \).

\[ U_n(\sqrt{n}(\hat{\beta} - \beta)) = \sum_{i=1}^{n} (Y_i^* - (\beta + \hat{\beta} - \beta)^T X_i^*)^2 + n\lambda_n \sum_{j=1}^{M} |\beta_j + \hat{\beta}_j - \beta_j| \]

\[ = \sum_{i=1}^{n} (Y_i^* - \hat{\beta} X_i^*)^2 + n\lambda_n \sum_{j=1}^{M} |\hat{\beta}_j| \]

\[ \leq \sum_{i=1}^{n} (Y_i^* - (\beta + u \frac{1}{\sqrt{n}})^T X_i^*)^2 + n\lambda_n \sum_{j=1}^{M} |\beta_j + \frac{u_j}{\sqrt{n}}| \]

\[ = U_n(u), \text{ for all } u. \]

The inequality in the second last line in the above derivation follows from the definition of \( \hat{\beta} \). Thus we see that

\[ \arg\min_{u \in \mathbb{R}^M} V_n(u) = \sqrt{n}(\hat{\beta}_n - \beta). \]

Now we can claim, by the Argmin Theorem, that

\[ \arg\min_{u \in \mathbb{R}^M} V_n(u) \overset{d}{\to} \arg\min_{u \in \mathbb{R}^M} V(u), \]

which would give us

\[ \sqrt{n}(\hat{\beta} - \beta) \overset{d}{\to} \arg\min_{u \in \mathbb{R}^M} V(u), \]

which would prove the theorem. In what follows we show that

\[ V_n(u) \overset{d}{\to} V(u) \]

for all \( u \). Note that

\[ V_n(u) = A_n(u) + B_n(u), \]

where,

\[ A_n(u) = \sum_{i=1}^{n} [(\epsilon_i^* - u^T X_i^*)^2 / \sqrt{n}]^2 - \epsilon_i^2 \]

and

\[ B_n(u) = n\lambda_n \sum_{j=1}^{M} [\beta_j + \frac{u_j}{\sqrt{n}}] - |\beta_j|]. \]
Note that

\[ A_n(u) = \sum_{i=1}^{n} (\epsilon_i^* - u'X_i^*/\sqrt{n})^2 - \epsilon_i^2 \]

\[ = u' (\frac{1}{n} \sum_{i=1}^{n} X_i^*X_i^*)u - \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i^* u'X_i^* \].

As \( n \to \infty \), from Assumption 2.1 we have

\[ u' (\frac{1}{n} \sum_{i=1}^{n} X_i\Sigma_n^{-1}X_i^*)u \to u'Cu. \]

In order to do so we have to first show that

\[ V_n(u) \overset{d}{\to} V(u). \]

In order to show this we first need to prove that

\[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i^* u'X_i^* \]

converges in law to the normal distribution. Observe that

\[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i^* u'X_i^* = \frac{1}{\sqrt{n}} u' (X_1^*, ..., X_n^*) \left( \begin{array}{c} \epsilon_1^* \\ \vdots \\ \epsilon_n^* \end{array} \right) \]

\[ = \frac{1}{\sqrt{n}} u'X^*\epsilon^* \]

\[ = \frac{1}{\sqrt{n}} u'X^*\Sigma_n^{-1}\epsilon \]

\[ = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \]

where the \( U_i \)'s are a function of a sequence of weakly dependent random variables. This is true because

\[ U_i = (u'u_i)\epsilon_i \]

where, \( u_i \) is the \( i^{th} \) column of the matrix \( X^*\Sigma_n^{-1} \). Since \( \epsilon_i \)'s are a sequence of weakly dependent random variables, from the definition of weak dependence in Section 1.1 and Theorem A.2 in Appendix A we can conclude that

\[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \overset{d}{\to} N(0, u'Cu). \quad (13) \]

Result (13) follows from Assumption 2.2. Thus we can say that

\[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i^* u'X_i^* \overset{d}{\to} u'S, \]
where, $S$ is a $M$ dimensional multivariate normal distribution $N(0, \sigma^2 C)$. Observe that

$$u' \left( \frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right) u = u' \left( \frac{1}{n} \sum_{i=1}^{n} X_i \Sigma_n^{-1} X_i' \right) u$$

But from Assumption 2.1 we also conclude that

$$u' \left( \frac{1}{n} \sum_{i=1}^{n} X_i \Sigma_n^{-1} X_i' \right) u \rightarrow u' C u$$

as $n \rightarrow \infty$. Now we show that

$$n \lambda_n \sum_{j=1}^{M} [\beta_j + \frac{u_j}{\sqrt{n}}] - |\beta_j| \rightarrow \lambda_0 \sum_{j=1}^{M} u_j \text{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)$$

as $n \rightarrow \infty$. Consider the case when $\beta_j = 0$, here it would be enough to show

$$n \lambda_n \sum_{j=1}^{M} \{|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j|\} \rightarrow \lambda_0 \sum_{j=1}^{M} |u_j|.$$

When $\beta_j = 0$ we have

$$|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j| = |\frac{u_j}{\sqrt{n}}|,$$

thus

$$n \lambda_n \sum_{j=1}^{M} \{|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j|\} = n \lambda_n \sum_{j=1}^{M} |u_j|.$$

As $n \rightarrow \infty$, from the assumption on $\lambda_n$, we have

$$\frac{n \lambda_n}{\sqrt{n}} \sum_{j=1}^{M} |u_j| \rightarrow \lambda_0 \sum_{j=1}^{M} |u_j|.$$

Thus we have, when $\beta_j = 0$,

$$\lambda_n \sum_{j=1}^{M} \{|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j|\} \rightarrow \lambda_0 \sum_{j=1}^{M} |u_j| I(\beta_j = 0).$$

When $\beta_j \neq 0$, we shall show that

$$n \lambda_n \sum_{j=1}^{M} \{|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j|\} \rightarrow \lambda_0 \sum_{j=1}^{M} u_j \text{sgn}(\beta_j) I(\beta_j \neq 0).$$

Observe that

$$|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j| = \frac{1}{\sqrt{n}} \{ |\sqrt{n} \beta_j + u_j| - |\sqrt{n} \beta_j| \}.$$
Now when $n$ is sufficiently large, we have

$$|\sqrt{n}\beta_j + u_j| - |\sqrt{n}\beta_j| = \sqrt{n}\text{sgn}(\beta_j)\beta_j + \text{sgn}(\beta_j)u_j - |\sqrt{n}\beta_j|,$$

and since $\text{sgn}(\beta_j)\beta_j = |\beta_j|$ we can write

$$\sqrt{n}\text{sgn}(\beta_j)\beta_j + \text{sgn}(\beta_j)u_j - |\sqrt{n}\beta_j| = \text{sgn}(\beta_j)u_j.$$

So for $n$ sufficiently large we have

$$n\lambda_n\{|\beta_j + u_j\sqrt{n}| - |\beta_j|\} = n\lambda_n\text{sgn}(\beta_j)u_j,$$

which tends to $\lambda_0\text{sgn}(\beta_j)u_j$ when $n \to \infty$. We can now say that

$$\lambda_n \sum_{j=1}^p \{ |\beta_j + u_j\sqrt{n}| - |\beta_j| \} \to \lambda_0 \sum_{j=1}^p u_j\text{sgn}(\beta_j)I(\beta_j \neq 0).$$

Using Slutsky’s theorem and combining the two results we have

$$\mathcal{A}_n(u) + \mathcal{B}_n(u) \xrightarrow{d} u'C u - 2u'W + \lambda_0 \sum_{j=1}^p [u_j\text{sgn}(\beta_j)I(\beta_j \neq 0) + |u_j|I(\beta_j = 0)].$$

Thus we arrive at

$$V_n(u) \xrightarrow{d} V(u).$$

Now in order to show

$$\text{argmin}_{u \in \mathbb{R}^M} V_n(u) \xrightarrow{d} \text{argmin}_{u \in \mathbb{R}^M} V(u)$$

we invoke the Argmin Theorem. According to this theorem we need to show that for all finite subset of \{s_1, ..., s_k\} of $\mathbb{R}^M$

$$(V_n(s_1), ..., V_n(s_k)) \xrightarrow{d} (V(s_1), ..., V(s_k)). \quad (14)$$

By the Cramer Wold device (see Shorack [13] page 351 Theorem 3.2) we can say that in order to show (14) it would be necessary and sufficient to show

$$(V_n(s_1), ..., V_n(s_k))l \xrightarrow{d} (V(s_1), ..., V(s_k))l \quad (15)$$

for any $l \in \mathbb{R}^k$. Now

$$(V_n(s_1), ..., V_n(s_k))l = \sum_{j=1}^k l_j V_n(s_j) = \sum_{j=1}^k l_j \mathcal{A}_n(s_j) + \sum_{j=1}^k l_j \mathcal{B}_n(s_j).$$

Now

$$\sum_{j=1}^k l_j \mathcal{A}_n(s_j) = \sum_{j=1}^k l_j \{s_j'\left(\frac{1}{n} \sum_{i=1}^n X_i X_i'\right)s_j - \frac{2}{\sqrt{n}} \sum_{i=1}^n s_j'X_i\},$$
so, by (14) we can say that
\[ \sum_{j=1}^{k} l_j A_n(s_j) + \sum_{j=1}^{k} l_j B_n(s_j) \xrightarrow{d} \sum_{j=1}^{k} l_j V(s_j). \]

Since \( V_n(u) \) is convex and \( V(u) \) has a unique minimum it follows from the \textit{Argmin theorem} that
\[ \argmin_{u \in \mathbb{R}^M} (V_n(u)) = \sqrt{n} (\hat{\beta} - \beta) \xrightarrow{d} \argmin_{u \in \mathbb{R}^M} (V(u)). \]

In particular if \( \lambda_0 = 0 \), then we have
\[ \argmin_{u \in \mathbb{R}^M} (V(u)) = C^{-1} W \sim N(0, \sigma^2 C^{-1}) \]

2.2 Study of asymptotic distribution when \( \Sigma_n \) is replaced by an estimate

In this section we show that if \( \Sigma_n^{-1} \) in (2) is replaced by \( \hat{\Sigma}_n^{-1} \), which is a componentwise consistent estimator of \( \Sigma_n^{-1} \) that is
\[ \left( (\hat{\Sigma}_n^{-1}) \right)_{i,j} \xrightarrow{P} \left( (\Sigma_n^{-1}) \right)_{i,j} \]

where, \( ((\hat{\Sigma}_n^{-1}))_{i,j} \) is the \((i,j)\)th entry of the matrix \( \hat{\Sigma}_n^{-1} \) and \( ((\Sigma_n^{-1}))_{i,j} \) is the \((i,j)\)th entry of the matrix \( \Sigma_n^{-1} \). Then under some additional assumptions on the design matrix \( X \) we get the same theoretical results as in the previous section. The additional assumption being
\[ \frac{1}{n} \max_{1 \leq i \leq n} X_i' X_i \rightarrow 0, \]

where, \( X_i \) is the \(i^{th}\) row of the design matrix \( X \). The implication of this assumption is
\[ \frac{1}{n} \max_{1 \leq i \leq j \leq n} X_i' X_j \rightarrow 0. \]

This follows from Cauchy Schwartz Inequality
\[ \frac{1}{n} X_i' X_j \leq \sqrt{\frac{1}{n} X_i' X_i} \sqrt{\frac{1}{n} X_j' X_j} \]
\[ \frac{1}{n} \max_{1 \leq i \leq j \leq n} X_i' X_j \leq \sqrt{\frac{1}{n} \max_{1 \leq i \leq n} X_i' X_i} \sqrt{\frac{1}{n} \max_{1 \leq j \leq n} X_j' X_j}. \]

\textbf{Theorem 2.2:} Suppose \( \sqrt{n} \lambda_n \rightarrow \lambda_0 \) as \( n \rightarrow \infty \) and let \( C \) be non singular, and Assumptions 2.1, 2.2, 2.3 and (17) are satisfied. Then
\[ \sqrt{n} (\hat{\beta}_n - \beta) \xrightarrow{d} \argmin(V) \]

where
\[ V(u) = -2u' S + u' Cu + \lambda_0 \sum_{j=1}^{M} [u_j \text{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)] \]
and $S$ has a $M$ dimensional $N(0, \sigma^2C)$ distribution.

**Proof**: Recall that the crux of the proof of the theorem in *Section 2.1* lies in showing

$$V_n(u) \overset{d}{\to} V(u)$$

where $V_n(u)$ and $V(u)$ is as described in the previous section. Now that we have the new minimization criteria, by replacing $\Sigma$ by $\hat{\Sigma}$ in $V_n(u)$, we have to show that

$$M_n(u) \overset{d}{\to} V(u) \quad (18)$$

where,

$$M_n(u) = \frac{1}{n} u' \Sigma n^{-1} X u - \frac{2}{\sqrt{n}} X' \hat{\Sigma} n^{-1} \epsilon + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{M} |\beta_j + \frac{u_j}{\sqrt{n}} - |\beta_j||.$$

Now in order to establish (18) it we have proved, see *Proposition 2.1*

$$V_n(u) - M_n(u) \overset{p}{\to} 0,$$

we then combine, using Slutsky’s theorem, the fact that

$$V_n(u) \overset{d}{\to} V(u)$$

which we had proved in *Proposition 2.2*, to give us

$$M_n(u) \overset{d}{\to} V(u).$$

Now repeating the argument as before in the previous section, that is we use the Cramer Wold device and the *Argmin Theorem* and arguing as before we can conclude that

$$\sqrt{n}(\hat{\beta} - \beta) \to \text{argmin}_u(V(u))$$

thus we prove the theorem. ■

**Proposition 2.1**: Under the *Assumption 2.1, 2.2, 2.3* and (17)

$$V_n(u) - M_n(u) \overset{p}{\to} 0$$

**Proof**: Observe that

$$V_n(u) - M_n(u) = \frac{1}{n} u' X' (\hat{\Sigma} n^{-1} - \Sigma n^{-1}) X u - \frac{2}{\sqrt{n}} X' (\hat{\Sigma} n^{-1} - \Sigma n^{-1}) \epsilon.$$

Let,

$$A_n(u) = \frac{1}{n} u' X' (\hat{\Sigma} n^{-1} - \Sigma n^{-1}) X u$$

and

$$B_n(u) = \frac{2}{\sqrt{n}} X' (\hat{\Sigma} n^{-1} - \Sigma n^{-1}) \epsilon.$$
It would be enough to show that

\[ A_n(u) \xrightarrow{P} 0 \]

and

\[ B_n(u) \xrightarrow{P} 0. \]

Slutsky’s theorem would then give us \( A_n(u) + B_n(u) \xrightarrow{P} 0 \) or \( V_n(u) - M_n(u) \xrightarrow{P} 0 \).

**Proposition 2.2** Under the Assumption 2.1, 2.2, 2.3 and (17) we have

\[ A_n(u) \xrightarrow{P} 0. \]

**Proof**

\[
A_n(u) = \frac{1}{n} u' X (\hat{\Sigma}_n^{-1} - \Sigma_n^{-1}) X u
\]

where, \( m_{ij} \) is the \((i,j)\)th entry of the matrix \( X (\hat{\Sigma}_n^{-1} - \Sigma_n^{-1}) X \). Now observe that

\[
\frac{1}{n} |m_{ij}| \leq |s_{ij}| \frac{1}{n} \max_{1 \leq i \leq j \leq n} X_i X_j
\]

where \( s_{ij} \) is the \((i,j)\)th entry of the matrix \( \hat{\Sigma}_n^{-1} - \Sigma_n^{-1} \). We further note that

\[ s_{i,j} = o_P(1) \]

as \( \hat{\Sigma}_n \) is a component wise consistent estimator of \( \Sigma_n \) and

\[
\frac{1}{n} \max_{1 \leq i \leq j \leq n} X_i X_j \rightarrow 0
\]

from the given assumption. Thus we have

\[
\frac{1}{n} |m_{ij}| = o_P(1)
\]

Therefore

\[ A_n(u) \xrightarrow{P} 0 \]

**Proposition 2.3**: Under the Assumption 2.1, 2.2, 2.3 and (17)

\[ B_n(u) \xrightarrow{P} 0. \]

**Proof**: We now show that

\[ B_n(u) \xrightarrow{P} 0 \]
where,

\[
|B_n(u)| = \frac{2}{\sqrt{n}} |u' X' (\hat{\Sigma}_n^{-1} - \Sigma_n^{-1}) \epsilon| \\
= \frac{2}{\sqrt{n}} \left| \sum_{i=1}^{n} \sum_{j=1}^{n} \epsilon_i u_j X_{j,i} \right| \\
\leq \sum_{j=1}^{n} \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i s_{i,j} |u_j X_{j}| \\
\leq \sum_{j=1}^{n} \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i s_{i,j} |L| \\
\leq \left\{ \frac{2}{\sqrt{n}} \sum_{i=1}^{n} |\epsilon_i| \right\} \max_{i,j} |s_{i,j}| L \\
\leq O_P(1) o_P(1) \\
\leq o_P(1)
\]

To arrive from the third to the fourth step we had assumed that the covariates \(x_j\) are bounded in each component. Thus we have \(B_n(u) \xrightarrow{P} 0\). □

2.3 Illustrative examples

In this section we provide two examples of weak dependence structure \(\Sigma_n\) and the corresponding estimates. The two examples we consider are (1) \(AR(1)\) process (2) \(ARMA(p,q)\).

2.3.1 \(AR(1)\) processes

We know that the errors \(\epsilon_i\) are first order auto regressive i.e. \(AR(1)\)

\[
\epsilon_i = a\epsilon_{i-1} + \eta_i, \quad \text{where} \quad i = 1, ..., n
\]

(19)

and

\[
\epsilon_1 = \eta_1
\]

where, \(0 < a < 1\) and \(\eta_1, ..., \eta_n\) is a sequence of independent random variables with mean 0 and variance \(\sigma^2\).

The covariance matrix \(\Sigma_n\) of \(\epsilon_1, ..., \epsilon_n\) is

\[
\Sigma_n = \frac{\sigma^2}{1 - a^2} \begin{bmatrix}
1 & a & \cdots & a^{n-1} \\
a & 1 & \cdots & a^{n-2} \\
\vdots & \ddots & \ddots & \vdots \\
a^{n-1} & \cdots & a & 1
\end{bmatrix}
\]

and

\[
\Sigma_n^{-1} = \frac{1}{\sigma^2} \begin{bmatrix}
1 & -a & \cdots & 0 \\
-a & 1 + a^2 & \cdots & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
0 & \cdots & -a & 1
\end{bmatrix}
\]
In the $AR(1)$ setting one way to estimate $a$ would be to use the Yule Walker estimate for $a$, which is

$$\hat{a} = \frac{n^{-1} \sum_{t=1}^{n-1} (U_{t+1} - \bar{U}_n)(U_t - \bar{U}_n)}{n^{-1} \sum_{t=1}^{n-1} (U_t - \bar{U}_n)(U_t - \bar{U}_n)},$$

where,

$$U = Y - X\hat{\beta}.$$

We also know that

$$\sqrt{n}(\hat{a} - a) \rightarrow N(0, \theta^2)$$

implying

$$\sqrt{n}(\hat{a} - a) = O_P(1)$$

implying

$$(\hat{a} - a) = O_P(n^{-1/2}) = o_P(1) \quad (20)$$

which implies,

$$\hat{a} \frac{P}{\sim} a.$$

The continuous mapping theorem implies that the $(i, j)^{th}$ component of $\hat{\Sigma}_n^{-1}$ converges weakly to the $(i, j)^{th}$ component of $\Sigma_n^{-1}$. Thus we can construct an estimator of the covariance matrix and using Theorem 2.1 we can establish the asymptotic distribution of the Lasso estimate of parameters of a linear model where the errors are generated from an $AR(1)$ process.

### 2.3.2 $ARMA(p,q)$

A sequence of random variables $\epsilon_i$ where $1 \leq i \leq n$ is said to be generated from an $ARMA(p,q)$ process if $\epsilon_i$ are stationary and for every $i$ we have

$$\epsilon_i - \sum_{j=1}^{p} \phi_j \epsilon_{i-j} = Z_i + \sum_{i=1}^{q} \theta_i Z_{i-q},$$

where, $Z_i \sim N(0, 1)$. Here it is not possible to get a closed form expression of the covariance matrix $\Sigma_n$ or the concentration matrix $\Sigma_n^{-1}$. Recall that since $ARMA(p,q)$ is a stationary process therefore the $(i, j)^{th}$ entry of the covariance matrix $\Sigma_n$ is $\gamma(|i-j|)$, where $\gamma(.)$ is the auto-covariance function. In order to estimate $\Sigma_n^{-1}$ we first estimate $\Sigma_n$ and take its inverse. $\Sigma_n$ in turn is estimated by estimating $\gamma(.)$ by $\hat{\gamma}(.)$. Observe that this is possible because $\hat{\gamma}(.)$ is a positive definite function which makes $\hat{\Sigma}_n$ a positive definite matrix, thus invertible and makes it possible for us to estimate $\Sigma_n^{-1}$. One way of estimating $\gamma(h)$ would be

$$\hat{\gamma}(h) = n^{-1} \sum_{k=1}^{n-h} (U_k - \bar{U}_n)(U_{k+h} - \bar{U}_n)$$

where $0 \leq h \leq n - 1$ and $U = Y - X\hat{\beta}$. Observe that here $\hat{\beta}$ is obtained iteratively by initial value of $\Sigma_n^{-1}$ as $\Sigma_0^{-1}$. Now using the continuous mapping theorem we can find a component wise consistent estimator of $\Sigma_n^{-1}$ and using Theorem we can establish the asymptotic distribution of the Lasso estimate of parameters of a linear model where the errors are generated from an $ARMA(p,q)$ process.
3 Results concerning model selection consistency of the Lasso estimate

In this section we investigate the problem of model selection consistency of Lasso estimate when the errors $\epsilon_i$s are correlated. We adopt the same setting as mentioned in the introduction. Here we assume that there is a true regression coefficient $\beta^*$ from which the data is generated. Thus,

$$\tilde{Y} = X\beta^* + \bar{\epsilon}$$

where

$$\tilde{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} X'_1 \\ \vdots \\ X'_n \end{pmatrix} \quad \text{and} \quad \bar{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

and $\beta^* \in \mathbb{R}^M$. Let $J(\beta^*)$ or $I^*$ be the index set corresponding to the nonzero components of the true regression coefficient, whereas the remaining coefficients are assumed to be exactly 0. Let $k(\beta^*)$ henceforth written as $k^*$ denote the cardinality of the set $J(\beta^*)$. Without loss of generality we will assume that the first $k(\beta^*)$ components of $\beta^*$ are nonzero and the rest 0 i.e. $J(\beta^*) = \{1, 2, \ldots, k(\beta^*)\}$ henceforth written as $I^*$. We denote the $j^{th}$ column of $X$ as $X_{.,j}$, and

$$\text{Cov}(\epsilon_1, \ldots, \epsilon_n) = \Sigma_n$$

Observe that the i.i.d errors case is covered when $\Sigma_n = \sigma^2 I_n$.

3.1 Study of model selection consistency when $\Sigma_n$ is known

Here we present the model selection consistency result for lasso estimate when the errors are weakly dependent, for the formal of definition of weak dependence see Section 1.1. First we present the results under the assumption that the covariance structure is assumed to be known. In this section we establish the model selection consistency of a lasso estimate, $\hat{\beta}$, when the observations are dependent and the error covariance structure is known. The key result in this section is stated in the form of the theorem below. We now list the assumptions which are required for the theorem to hold good.

**Assumption 3.1:** The errors $\epsilon_i$s are weakly dependent.

**Assumption 3.2:**

$$e_1 \geq \frac{1}{\theta} \quad \text{or} \quad (e_1^{-1} \leq \theta)$$

for some $\theta > 0$, where, $e_1$ is the smallest eigenvalue of the matrix $\Sigma_n^{-1}$ (so accordingly $e_1^{-1}$ is the largest eigenvalue of $\Sigma_n$)

**Assumption 3.3:** For $k \in I^*$

$$|\beta^*_k| \geq 2\theta \lambda_n.$$

**Assumption 3.4:** Let

$$\rho_{kj} = \frac{1}{n} X'_{.,j} \Sigma_n^{-1} X_{.,k}$$
We assume that there exists a constant $0 < d < 1$ such that
\[ P(\max_{j \in I^*, k \neq j} |\rho_{kj}| \leq \frac{d}{k^*}) = 1 \]

**Theorem 3.1:** If Assumptions 3.1 to 3.4 above holds then for any $\delta > 0$ and
\[ \lambda_n \geq \max\{C_1 \sqrt{\frac{\log M}{n}}, B_1(\frac{\log M}{n})^{\mu+\nu+2}\} \]
for suitably chosen constants $C_1$ and $B_1$, we have
\[ P(\hat{I} = I^*) \geq 1 - 2\delta. \quad (21) \]

**Proof:** We know,
\[ P(\hat{I} = I^*) \geq 1 - P(I^* \not\subseteq \hat{I}) - P(\hat{I} \not\subseteq I^*) \]
From Proposition 3.1 and Proposition 3.2 below we see that each of the probabilities on the right hand side of the above inequality goes to 0 as $n$ goes to $\infty$. Thus our proof is complete. \(\square\)

**Remark:** The proofs of Proposition 3.1 and Proposition 3.2 are largely motivated by the works of Bunea(2008), see [1]. Bunea(2008), see [1], addressed this problem in a setting where the errors are i.i.d. However our present work differs from Bunea’s work along the lines of (i) here we need version of Concentration Inequality for this particular dependent structure, which we have stated in Appendix A, see Theorem A.2 and (ii) a modification of the proof of Bunea to take in account the non i.i.d structure of the errors.

**Proposition 3.1:** If the assumptions of Theorem 3.1 holds then for any $\delta > 0$ and
\[ \lambda_n \geq \max\{C_1 \sqrt{\frac{\log M}{n}}, B_1(\frac{\log M}{n})^{\mu+\nu+2}\} \]
for suitably chosen constants $C_1$ and $B_1$, we have
\[ P(I^* \not\subseteq \hat{I}) \leq \delta. \]

**Proof:** We follow the same reasoning as Proposition 3.3 of Bunea (2008), see [1]. From the definitions of $k^*$, $I^*$, $\hat{I}$ and the union bound we have
\[ P(I^* \not\subseteq \hat{I}) \leq P(\hat{I} \not\subseteq I^* for some k \not\in I^*) \]
\[ \leq P(\hat{\beta}_k = 0 and \beta_k^* \neq 0, for some k \in I^*) \]
\[ \leq \max_{k \in I^*} P(\hat{\beta}_k = 0 and \beta_k^* \neq 0). \]
It follows from Lemma 4.1 in Appendix B of Bunea (2008), see [1], that if $\hat{\beta}_k = 0$ is a component of the solution $\hat{\beta}$ then
\[ \frac{2}{n}(Y - X\hat{\beta})^\prime \Sigma_n^{-1}X_{\cdot k} \leq 2\lambda_n. \]
where, \( X_{.,k} \) is the \( k^{th} \) column of the design matrix \( X \). Therefore,

\[
P(I^* \notin \hat I) \leq k^* \max_{k \in I^*} P(\beta_{k} = 0 \cap \beta_k^* \neq 0)
\]

\[
\leq k^* \max_{k \in I^*} P(\frac{1}{n}(Y - X \hat \beta)' \Sigma^{-1}_{n} X_{.,k} \leq 2 \lambda_n \cap \beta_k^* \neq 0)
\]

\[
= k^* \max_{k \in I^*} P(\frac{1}{n}(X \beta^* + \epsilon - X \hat \beta)' \Sigma^{-1}_{n} X_{.,k} \leq \lambda_n \cap \beta_k^* \neq 0)
\]

\[
= k^* \max_{k \in I^*} P(\frac{1}{n}(X \beta^* + \epsilon - X \hat \beta)' \Sigma^{-1}_{n} X_{.,k} \leq \lambda_n \cap \beta_k^* \neq 0)
\]

\[
= k^* \max_{k \in I^*} P(\frac{1}{n} X \beta^* - \hat \beta)' \Sigma^{-1}_{n} X_{.,k} \leq \lambda_n \cap \beta_k^* \neq 0)
\]

\[
= k^* \max_{k \in I^*} P(\frac{1}{n} X \beta^* - \hat \beta)' \Sigma^{-1}_{n} X_{.,k} \leq \lambda_n \cap \beta_k^* \neq 0)
\]

\[
+ \sum_{j \neq k} (\beta_j^* - \hat \beta_j) \frac{1}{n} X_{.,j} \Sigma^{-1}_{n} X_{.,k} \leq \lambda_n \cap \beta_k^* \neq 0)
\]

\[
\leq k^* \max_{k \in I^*} P(\beta_k^* \mid \frac{1}{n} X_{.,j} \Sigma^{-1}_{n} X_{.,k} \leq \lambda_n)
\]

\[
= k^* \max_{k \in I^*} P(\frac{1}{n} X_{.,k}^* \Sigma^{-1}_{n} X_{.,k} \leq \lambda_n)
\]

\[
= \sum_{j \neq k} (\beta_j^* - \hat \beta_j) \frac{1}{n} X_{.,j} \Sigma^{-1}_{n} X_{.,k} \geq \beta_k^* \mid \frac{1}{n} X_{.,k} \Sigma^{-1}_{n} X_{.,k} \leq \lambda_n)
\]

Now

\[
2 \lambda_n \leq |\beta_k^*| \frac{1}{\theta} \leq |\beta_k^*| e_1 \leq |\beta_k^*| \frac{1}{n} X_{.,k} \Sigma^{-1}_{n} X_{.,k}
\]

and this follows from the assumptions that

\[
e_1 \geq \frac{1}{\theta} \text{ or } (e_1^{-1} \leq \theta)
\]

where, \( e_1 \) is the smallest eigenvalue of the matrix \( \Sigma^{-1}_{n} \) (so accordingly \( e_1^{-1} \) is the largest eigenvalue of \( \Sigma_{n} \)) and

\[
|\beta_k^*| \geq 2 \theta \lambda_n.
\]

The last statement here is true from the assumption on the signal level, that is, the non-zero signals are uniformly bounded from below. In order to complete this proof we need a Concentration inequality for the weak dependence structure, which we have stated, without proof, in Appendix A, for the convenience of the readers. Now observe that, from the above derived relationship, for any \( k \) in \( I^* \)

\[
P(\frac{1}{n} \epsilon \Sigma_{n}^{-1} X_{.,k} | + | \sum_{j \neq k} (\beta_j^* - \hat \beta_j) \frac{1}{n} X_{.,j} \Sigma^{-1}_{n} X_{.,k} \geq \beta_k^* \mid \frac{1}{n} X_{.,k} \Sigma^{-1}_{n} X_{.,k} \leq \lambda_n)
\]

can be bounded from above by

\[
P(\frac{1}{n} \epsilon \Sigma_{n}^{-1} X_{.,k} | + | \sum_{j \neq k} (\beta_j^* - \hat \beta_j) \frac{1}{n} X_{.,j} \Sigma^{-1}_{n} X_{.,k} \geq \lambda_n)
\]
which in turn can be bounded from above by

\[ P\left(\left|\frac{1}{n}I^\prime \Sigma_n^{-1} X_{.,k}\right| \geq \lambda_n/2\right) + P\left(\sum_{j \neq k} (\beta^*_j - \hat{\beta}_j) \frac{1}{n} X_{.,j} \Sigma_n^{-1} X_{.,k} \geq \lambda_n/2\right) \]

From the Concentration Inequality for a weak dependent sequence in Appendix A, and

\[ \lambda_n \geq \max\{C_1 \sqrt{\frac{\log \frac{8M}{\delta}}{n}}, B_1(\frac{\log \frac{8M}{\delta}}{n})^{\mu + \nu + 2}\} \]

we have

\[ P\left(\left|\frac{1}{n}I^\prime \Sigma_n^{-1} X_{.,k}\right| \geq \lambda_n/2\right) \leq \frac{\delta}{2M^2}. \]

Now, from Assumption 3.4,

\[ \left|\sum_{j \neq k} (\beta^*_j - \hat{\beta}_j) \frac{1}{n} X_{.,j} \Sigma_n^{-1} X_{.,k}\right| \leq \sum_{j \neq k} |\beta^*_j - \hat{\beta}_j| \frac{d}{k^*}. \]

Thus

\[ P\left(\left|\sum_{j \neq k} (\beta^*_j - \hat{\beta}_j) \frac{1}{n} X_{.,j} \Sigma_n^{-1} X_{.,k}\right| \geq \lambda_n/2\right) \leq P\left(\left|\hat{\beta} - \beta^*\right| \leq \frac{\lambda_n k^*}{2d}\right) \]

Now taking \( b = 10d \) it follows from Theorem B.2 of Appendix B that

\[ P\left(\left|\hat{\beta} - \beta^*\right| \leq \frac{\lambda_n k^*}{2d}\right) \leq \frac{\delta}{2M} \]

Thus we have

\[ P(I^* \not\subseteq \hat{I}) \leq k^* \frac{\delta}{2M} + k^* \frac{\delta}{2M^2} \leq \delta \]

\[ \blacksquare \]

Note: Observe that the smaller the lower limit of \( e_1 \) larger the upper limit of \( e_1 \), and also note larger the value of \( \theta \) larger is the upper limit limit of \( |\beta^*_1| \). Now also note that there is a relationship between the eigenvalues of the covariance matrix and the variances of the random errors, so as \( e_1 \) increases or as the upper bound of \( e_1 \) that is \( \theta \) increases we find that the threshold of the signal level also increases. This can be essentially interpreted as, as the noise level in the data increases more difficult does it become for model selection to be possible.

**Proposition 3.2:** If the assumptions of Theorem 3.1 hold, then for any \( \delta > 0 \) and

\[ \lambda_n \geq \max\{C_1 \sqrt{\frac{\log \frac{8M}{\delta}}{n}}, B_1(\frac{\log \frac{8M}{\delta}}{n})^{\mu + \nu + 2}\} \]

for suitably chosen constants \( C_1 \) and \( B_1 \), we have

\[ P(\hat{I} \not\subseteq I^*) \leq \delta. \]
Proof: Let

\[ h(\beta) = \frac{1}{n} \eta' \Sigma_n^{-1} \eta + 2 \lambda_n | (I : O) \beta |_{\ell_1} \]

\[ \eta = Y - X (I : O) \beta \]

where \( I \) is a \( k^* \times k^* \) identity matrix and \( O \) is a \( k^* \times M - k^* \) matrix whose every entry is 0. Define

\[ \tilde{\beta} = \arg\min_{\beta \in \mathbb{R}^{k^*}} h(\beta) \]

Let

\[ B = \bigcap_{k \notin J(\beta^*)} \left\{ \left| \frac{1}{n} X_k' \Sigma_n^{-1} (Y - X (I : O) \tilde{\beta}) \right| < 2 \lambda_n \right\} \]

\[ \mathbb{P}(\hat{I} \not\subseteq I^*) \leq \mathbb{P}(B^c) \]

\[ = \mathbb{P}( \bigcup_{k \notin J(\beta^*)} \left\{ \left| \frac{1}{n} X_k' \Sigma_n^{-1} (Y - X \left( \begin{array}{c} I \\ O \end{array} \right) \tilde{\beta}) \right| \geq \lambda_n \right\} ) \]

\[ \leq \sum_{k \notin J(\beta^*)} \mathbb{P}(\left\{ \left| \frac{1}{n} X_k' \Sigma_n^{-1} (Y - X \left( \begin{array}{c} I \\ O \end{array} \right) \tilde{\beta}) \right| \geq \lambda_n \right\} ) \]

\[ \leq \sum_{k \notin J(\beta^*)} \mathbb{P}(\left\{ \left| \frac{1}{n} X_k' \Sigma_n^{-1} (Y - X \left( \begin{array}{c} I \\ O \end{array} \right) \tilde{\beta}) \right| \geq \lambda_n \right\} ) \]

\[ = \sum_{k \notin J(\beta^*)} \mathbb{P}(\left\{ \left| \frac{1}{n} X_k' \Sigma_n^{-1} (X \beta^* + \epsilon - X \left( \begin{array}{c} I \\ O \end{array} \right) \tilde{\beta}) \right| \geq \lambda_n \right\} ) \]

\[ \leq \sum_{k \notin J(\beta^*)} \mathbb{P}(\left\{ \left| \frac{1}{n} X_k' \Sigma_n^{-1} X (\beta^* - \left( \begin{array}{c} I \\ O \end{array} \right) \tilde{\beta} \right| \geq \frac{\lambda_n}{2} \right\} ) \]

\[ + \mathbb{P}(\left\{ \left| \frac{1}{n} X_k' \Sigma_n^{-1} \epsilon \right| \geq \frac{\lambda_n}{2} \right\} ) \]

The second probability

\[ \sum_{k \notin J(\beta^*)} \mathbb{P}(\left\{ \left| X_k' \Sigma_n^{-1} \epsilon \right| \geq \frac{\lambda_n}{2} \right\} ) \leq \frac{\delta}{2M} < \frac{\delta}{2} \]

follows from Theorem A.2 of Appendix A. The first probability

\[ \sum_{k \notin J(\beta^*)} \mathbb{P}(\left\{ \left| X_k' \Sigma_n^{-1} (\beta^* - \left( \begin{array}{c} I \\ O \end{array} \right) \tilde{\beta}) \right| \geq \frac{\lambda_n}{2} \right\} ) \leq \sum_{k \notin J(\beta^*)} \mathbb{P}(\left\{ |\beta^* - \tilde{\beta}|_{\ell_1} \geq \frac{k^* \lambda_n}{d} \right\} ) \]

\[ \leq \sum_{k \notin J(\beta^*)} \frac{\delta}{2M} \]

\[ \leq \frac{\delta}{2} \]

21
This follows from Theorem B.2 in Appendix B. Thus we have
\[ P(\hat{I} \not\subset I) \leq \delta \]

3.2 An illustrative example: Consistent variable selection with AR(1) errors

In the previous section we discussed the model selection consistency for a Lasso estimator of the regression coefficients of a linear model when the errors are weakly dependent. AR(1) is a weakly dependent sequence, therefore the model selection consistency in this setup when the covariance matrix \( \Sigma_n \) is known follows from the above theorem. A formal definition of an AR(1) process is as follows. The errors \( \epsilon_i \) are first order auto regressive i.e. AR(1) if
\[ \epsilon_i = a\epsilon_{i-1} + \eta_i, \quad \text{where} \quad i = 1, ..., n \] (22)
and
\[ \epsilon_1 = \eta_1 \]
where, \( 0 < a < 1 \) and \( \eta_1, ..., \eta_n \) is a sequence of independent random variables with mean 0 and variance \( \sigma^2 \).

From the the way we have defined \( \epsilon_n \) we can show that
\[ \epsilon_n = \sum_{i=1}^{n} a^{n-i} \eta_i \]

For AR(1) setting observe that the choice of the \( \lambda_n \) is of the form, see Theorem B.3 of Appendix B,
\[ \lambda_n \geq \max\{4c \sqrt{\frac{\log 8M^2}{n}}, 8d \log \frac{8M^2}{\delta n}\} \]
where \( c = ka \) and \( d = kb \) and
\[ k = \frac{(1 + a^2)}{1 - a} \]
Now observe that as \( a \) goes to zero the errors slowly turns into a sequence of iid random variables and the penalty parameter \( \lambda_n \) is free from \( a \) which is the parameter controlling the degree of dependence in the error. However when \( a \) goes to 1 \( \lambda_n \) gets very large. Then our result holds for a very restrictive class of models as is evident from Assumption 3.3. One way to explain why this happens would be since we are in an AR(1) setup, when \( a \) goes to 1 the variances of the errors increases drastically, thus our signal tends to get drowned in the noise and hence our method is applicable on a very small class of models.

In most practical cases this is not the situation, there we need to estimate \( \Sigma_n \). However if we have the information that the errors are from an AR(1) process then our task reduces to estimating only one parameter. In the next section we present a result concerning the model selection consistency when the inverse covariance matrix \( \Sigma_n^{-1} \) is estimated by a componentwise consistent estimator.
3.2.1 Consistent variable selection for AR(1) errors with estimated $\Sigma_n$

As mentioned, here we estimate the covariance matrix. We know that the errors $\epsilon_i$ are first order auto regressive i.e. AR(1)

$$\epsilon_i = a\epsilon_{i-1} + \eta_i, \text{ where } i = 1, ..., n \quad (23)$$

and

$$\epsilon_1 = \eta_1$$

where, $0 < a < 1$ and $\eta_1, ..., \eta_n$ is a sequence of independent random variables with mean 0 and variance $\sigma^2$.

The covariance matrix $\Sigma_n$ of $\epsilon_1, ..., \epsilon_n$ is

$$\Sigma_n = \frac{\sigma^2}{1 - a^2}((a^{i-j}))$$

In the AR(1) setting one way to estimate $a$ would be to use the Yule Walker estimate for $a$, which is

$$\hat{a} = \frac{n^{-1}\sum_{t=1}^{n-1}(U_{t+1} - U_n)(U_t - U_n)}{n^{-1}\sum_{t=1}^{n-1}(U_t - U_n)(U_t - U_n)}$$

where

$$U = Y - X\hat{\beta}.$$ 

We also know that

$$\sqrt{n}(\hat{a} - a) \rightarrow N(0, \theta^2)$$

implying

$$\sqrt{n}(\hat{a} - a) = O_P(1)$$

implying

$$(\hat{a} - a) = O_P(n^{-1/2}) = o_P(1) \quad (24)$$

which implies,

$$\hat{a} \stackrel{P}{\rightarrow} a$$

Now consider the event for some $\epsilon > 0$ and fixed.

$$\mathcal{A} = \{\omega : |\hat{a}(\omega) - a| \leq \epsilon\}$$

and

$$\mathcal{A}^c = \{\omega : |\hat{a}(\omega) - a| > \epsilon\}.$$ 

Now from statement (24) it is quite clear that

$$\mathbb{P}(\mathcal{A}^c) \rightarrow 0$$

as $n \rightarrow \infty$. Now observe that from triangle inequality we have

$$|a| + |\hat{a} - a| \geq |\hat{a}|$$
So $\mathcal{A}$ along with the above inequality implies

$$|a| + \epsilon \geq |\hat{a}|$$

Now the question is with $a$ replaced by $\hat{a}$ (or equivalently replacing $\Sigma_{n}^{-1}$ by $\hat{\Sigma}_{n}^{-1}$) in the minimization criteria (4) can we still have model selection consistency or does

$$\mathbb{P}(\hat{\mathcal{I}} \neq \mathcal{I}^*) \rightarrow 0$$

as $n \to \infty$? We shall try to answer this question in this section. Observe in Section 3.1, the model selection consistency was proved by the help of Proposition 3.1 and Proposition 3.2. Now in order to ensure that the model selection consistency result holds good in our current setting with an estimated concentration matrix we need to ensure that the two proposition holds good. Combining the two propositions we had

$$\mathbb{P}(\mathcal{I}^* \neq \hat{\mathcal{I}}) \leq \delta.$$

We want to see if we can prove this in our current setting. Observe that

$$\mathbb{P}(\mathcal{I}^* \neq \hat{\mathcal{I}}) = \mathbb{P}(\{\mathcal{I}^* \neq \hat{\mathcal{I}} \cap \mathcal{A}\} + \mathbb{P}(\{\mathcal{I}^* \neq \hat{\mathcal{I}} \cap \mathcal{A}^c\))$$

since

$$\mathbb{P}(\{\mathcal{I}^* \neq \hat{\mathcal{I}} \cap \mathcal{A}^c\) \leq \mathbb{P}(\mathcal{A}^c\) \leq \delta/2.$$

as $n$ is sufficiently large. Also observe that given the event $\mathcal{A}$ we can say that

$$|((\Sigma_{n}^{-1})_{i,j}| \geq |((\hat{\Sigma}_{n}^{-1})_{i,j}|$$

where $((\Sigma_{n}^{-1})_{i,j}$ is the $(i, j)^{th}$ entry of the matrix $\Sigma_{n}^{-1}$ and $((\hat{\Sigma}_{n}^{-1})_{i,j}$ is the $(i, j)^{th}$ entry of the matrix $\hat{\Sigma}_{n}^{-1}$ and $\Sigma_{n}^{-1}$ retains the same structure as the matrix $\hat{\Sigma}_{n}^{-1}$ with $\hat{a}$ replaced by $|a| + \epsilon$. Thus instead of working with (4) we shall be effectively working with

$$\hat{\beta}_n = \arg\min_{\beta \in R^M} \left\{ \frac{1}{n} (\tilde{Y} - \tilde{X} \beta)' \Sigma_{n}^{-1} (\tilde{Y} - \tilde{X} \beta) + 2\lambda_n \sum_{j=1}^{M} |\beta_j| \right\}$$

(25)

Now recall that in order to prove consistency in the setting (4) an important assumption which was needed was of the form:

If

$$\rho_{kj} = \frac{1}{n} X_{.,j}' \Sigma_{n}^{-1} X_{.,k}$$

We assume that there exists a constant $0 < d < 1$ such that

$$\mathbb{P}(\max_{j \in \mathcal{I}^*, k \neq j} |\rho_{kj}| \leq \frac{d}{k^d}) = 1.$$

Now in order to establish model selection consistency we need to replace the above assumption by

$$\mathbb{P}(\max_{j \in \mathcal{I}^*, k \neq j} |\rho_{kj}| \leq \frac{d}{k^d}) = 1,$$

here,

$$\rho_{kj} = \frac{1}{n} X_{.,j}' \Sigma_{n}^{-1} X_{.,k}.$$
4 Conclusion

In Section 2 and Section 3 we stated and proved the asymptotic normality and model selection consistency respectively of the lasso estimates. The results which we presented for the asymptotic normality and model selection consistency of an $\ell_1$ penalized estimator included the setting of $M > n$, that is, the number of parameters exceeds the number of observation. We had discussed in Section 1 that similar kind of work has already been done before but what differentiated our work from the rest is that we were working under the setting where the data are dependent or clustered. In order to address this problem we had proposed a new minimization criteria and proved some theoretical results for the new estimator. It was also shown that the criterion which is used in iid cases is a special case of the one which we have used. Here we briefly highlight the significance of having a $\Sigma^{-1}n$ term inserted in the minimization function as in (4). To illustrate the importance of $\Sigma^{-1}$ let us confine ourselves to the setting of AR(1). Recall that in Section 3.2, we have proved that under certain assumptions, for any $\delta > 0$, there exists a $\lambda_n$ such that

$$\mathbb{P}(\hat{I} = I^*) \geq 1 - 2\delta.$$ 

The key result used to prove this theorem was the non i.i.d Concentration Inequality, see Theorem A.2 of Appendix B. For the sake of convenience let us rewrite the statement of the theorem

**Theorem:** Let $\epsilon_1, \ldots, \epsilon_n$ be a sequence of random variables as defined by (39), then for any $\lambda_n > 0$ we have

$$\mathbb{P}\left(\left|\frac{2}{n}X' \cdot \Sigma^{-1} \epsilon\right| \geq \lambda_n\right) \leq 2 \exp\left(-\frac{n\lambda_n^2}{2(c^2 + d\lambda_n)}\right)$$

for some positive constants $c$ and $d$, where $c = k_0a_1$ and $d = k_0b_1$ where

$$k_0 = \min\{\frac{(1 + a^2)}{1 - a}, (1 + a)^2\}.$$ 

Now if $\Sigma^{-1}$ was not present in (26) then the above theorem would become

**Theorem:** Let $\epsilon_1, \ldots, \epsilon_n$ be a sequence of random variables as defined by (39), then for any $\lambda_n > 0$ we have

$$\mathbb{P}\left(\left|\frac{2}{n}X' \cdot \epsilon\right| \geq \lambda_n\right) \leq 2 \exp\left(-\frac{n\lambda_n^2}{2(c_1^2 + d_1\lambda_n)}\right).$$

for some positive constants $c$ and $d$. Where $c_1 = k_1a_1$ and $d_1 = k_1b_1$ and

$$k_1 = \frac{1}{1 - a}$$

and $a_1, b_1$ are suitable constants.

From the structure of the errors, see (22), it is quite clear that as $a$ approaches 0 we get closer to an independent sequence and as $a$ approaches 1 the dependence between the random errors increases. Now observe that as $a$ is close to 1 (in fact if $a > 0.75$) then clearly

$$\frac{1}{1 - a} > (1 + a)^2$$

That is

$$\frac{k_1}{2(k_1^2a_1^2 + k_1b_1\lambda_n)} > \frac{k_0}{2(k_0^2a_1^2 + k_0b_1\lambda_n)}$$

$$\frac{n\lambda_n^2}{2(k_0^2a_1^2 + k_0b_1\lambda_n)} > \frac{n\lambda_n^2}{2(k_1^2a_1^2 + k_1b_1\lambda_n)}$$

25
\[-\frac{n\lambda_n^2}{2(k_1^2a_1^2 + k_1b_1\lambda_n)} > -\frac{n\lambda_n^2}{2(k_0^2a_1^2 + k_0b_1\lambda_n)}\]
\[\exp(-\frac{n\lambda_n^2}{2(k_1^2a_1^2 + k_1b_1\lambda_n)}) > \exp(-\frac{n\lambda_n^2}{2(k_0^2a_1^2 + k_0b_1\lambda_n)})\]

Thus it follows that
\[\mathbb{P}(\frac{2}{n}X_{\cdot j}^T\Sigma_n^{-1}\epsilon \geq \lambda_n)\]
will go to 0 faster than
\[\mathbb{P}(\frac{2}{n}X_{\cdot j}\epsilon \geq \lambda_n)\]

Thus we see inserting \(\Sigma_n^{-1}\) in the minimization criteria does improve the consistency result if the errors are highly dependent. If the errors are not highly dependent in fact if \(a < \sqrt{2} - 1\) then we have
\[k_0 = \frac{1 + a^2}{1 - a} \approx \frac{1}{1 - a} = k_1.\]

We now point out a difference in the choice of the tuning/penalty parameter in the two problems in Section 2 and Section 3. In the asymptotic normality problem the tuning parameter, see Theorem 2.1,
\[\lambda_n = C \frac{1}{\sqrt{n \log n}}.\] (28)
for some \(C > 0\) and in the model selection problem, see Theorem 2.1,
\[\lambda_n \geq \max\{4\epsilon \sqrt{\frac{\log 2M^2}{n}}, 8d \frac{\log 2M^2}{n}\}\] (29)

Now comparing the penalty term (28) for asymptotic normality and the penalty term (29) for model selection consistency, and on doing so we see they differ by a multiple of \((\log n)^2\). This difference in choice of \(\lambda_n\) is because these are two essentially different problems and one cannot be related to another in a straightforward manner.

**Appendix A**

**Concentration inequality for weakly dependent errors**

**Theorem A.1** Suppose that \(Z_1, \ldots, Z_n\) are real valued random variables with 0 mean, defined on a probability space \((\Omega, \mathcal{A}, \mathbb{P})\). Let \(\psi : \mathbb{N}^2 \rightarrow \mathbb{N}\) be one of the functions:
1. \(\psi(u, v) = 2v,\)
2. \(\psi(u, v) = u + v,\)
3. \(\psi(u, v) = uv,\)
4. \(\psi(u, v) = \alpha(u + v) + (1 - \alpha)uv,\) for some \(\alpha \in (0, 1).\)
We assume that there exists constants $K, M, L_1, L_2 < \infty, \mu, \nu \geq 0,$ and a non increasing sequence of real coefficients $\rho(n)$ such that for all $u-$tuples $(s_1, ..., s_u)$ and all $v-$tuples $(t_1, ..., t_v)$ with $1 \leq s_1 \leq ... \leq s_u \leq t_1 \leq ... \leq t_v \leq n$ the following inequalities are fulfilled:

$$|\text{Cov}(Z_{s_1}...Z_{s_u}, Z_{t_1}...Z_{t_v})| \leq K^2M^{u+v-2}((u + v)!)^\nu \psi(u, v)\rho(t_1 - s_u),$$

where,

$$\sum_{s=0}^{\infty} (s + 1)^k \rho(s) \leq L_1L_2^k (k!)^\mu \forall k \geq 0,$$

and

$$\mathbb{E}|Z_t|^k \leq (k!)^\nu M^k \forall k \geq 0$$

Then, for all $t \geq 0$,

$$\mathbb{P}(S_n \geq t) \leq \exp(-\frac{t^2/2}{A_n + B_n^{1/(\mu+\nu+2)}t(2\mu+2\nu+3)/(\mu+\nu+2)})$$

where $A_n$ can be chosen as any number greater than or equal to $\sigma_n^2$ and

$$B_n = 2(K\nu M)L_2((\frac{2^{4+\mu+\nu nK^2 L_1}}{A_n}))^{v1})$$

We now state a Central Limit Theorem for a triangular sequence of weakly dependent random variables. Let $(X_{n,k})_{k=1,\ldots,k_n}$ for $n = 1, 2, \ldots$ be a triangular array defined through an $E$ valued weakly dependent $(\epsilon_n)_{n \in \mathbb{N}}$ (for the definition of weakly dependent sequence see Appendix-C Definition-(??)) by numerical Lipschitz functions $g_{n,k}$ defined on $E$ for $k = 1, 2, \ldots, k_n$ and $n = 1, 2, \ldots$. We assume that the sequence of integers $k_n$ increases to infinity with $n$. We set

$$X_{n,k} = g_{n,k}(\epsilon_k)$$

and,

$$S_n = \sum_{j=1}^{k_n} X_{n,j}.$$  

We assume that $\mathbb{E}(X_{n,k}) = 0$. Now let $S_{k,n} = \sum_{j=1}^{k} X_{n,j}$ for $1 \leq k \leq k_n$. We also suppose that there exist constants $\sigma, \alpha > 0$ such that

$$\lim_{n \to \infty} \mathbb{V}(S_n) = \sigma^2 > 0$$

and $v_{k,n} = \mathbb{V}(S_{k,n}) - \mathbb{V}(S_{k-1,n}) \geq \frac{\alpha}{n}$,

for each $k \in {1, \ldots, k_n}$ and for any integer $n$. We shall set

$$\lambda_n = \sup_{1 \leq k \leq k_n} \text{Lip}(g_{n,k}), \ M_n = \sup_{1 \leq k \leq k_n} ||g_{k,n}||_{\infty},$$

$$\delta_n = \sup_{1 \leq k \leq k_n} \mathbb{E}|X_{k,n}| \text{ and } \Delta_n = \sup_{1 \leq k \neq l \leq k_n} \mathbb{E}|X_{k,n}X_{l,n}|.$$  

Here we present a Central Limit Theorem for weakly dependent random variables. For proofs please see Coulon-Prieur and Doukhan [3].
**Theorem A.2:** Assume that the $E$-valued sequence satisfies the s-weak (resp a-weak) dependence condition and the triangular array $(X_{n,k})_{1 \leq k \leq k_n}$ defined as before satisfies assumption (32). Then if, as $n \to \infty$, we have

$$\left(k_n M_n + \frac{k_n^2}{n^3}\right) M_n \delta_n \to 0,$$

and

$$k_n \sum_{p=1}^{k_n} \min(\lambda_n \theta, \Delta_n) \to 0$$

and

$$k_n \sum_{p=1}^{k_n} \min(M_n \lambda_n \theta, \Delta_n) \to 0$$

(for a-weak dependence the last two conditions are replaced by

$$k_n \sum_{p=1}^{k_n} \min(\lambda_n^2 \theta, \Delta_n) \to 0$$

and

$$k_n \sum_{p=1}^{k_n} \min(\lambda_n^2 \theta, \Delta_n) \to 0)$$

we obtain

$$S_n \to N(0, \sigma^2)$$

in distribution.

### Appendix B

**Results leading to the non iid version of Theorem 2.2 from Bunea(2008)**

In this section we present a non i.i.d version of the Theorem 2.2 from Bunea(2008), see [2]. As before we assume that there is a true regression coefficient $\beta^* \in \mathbb{R}^M$ from which the data is generated. Let $J(\beta^*)$ or $I^*$ be the index set corresponding to the nonzero components of the true regression coefficient, whereas the remaining coefficients are assumed to be exactly 0. Let $k(\beta^*)$ henceforth written as $k^*$ denote the cardinality of the set $J(\beta^*)$. Without loss of generality we will assume that the first $k(\beta^*)$ components of $\beta^*$ are nonzero and the rest 0 i.e. $J(\beta^*) = \{1, 2, ..., k(\beta^*)\}$. Thus,

$$\tilde{Y} = X\beta^* + \tilde{\epsilon}$$

where

$$\tilde{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} X_1' \\ \vdots \\ X_n' \end{pmatrix} \quad \text{and} \quad \tilde{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

where, $\epsilon_i$s are a sequence of weakly dependent errors and,

$$\text{Cov}(\epsilon_1, ..., \epsilon_n) = \Sigma_n$$

28
Thus when $\Sigma_n = \sigma^2 I_n$ the i.i.d errors case is covered. Now,

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^M} \left\{ ||\tilde{Y} - X'\beta||_n^2 + \text{pen}(\beta) \right\}$$  \hspace{1cm} (35)$$

and,

$$||\tilde{Y} - X'\beta||_n = \frac{1}{n} (\tilde{Y} - X\beta)' \Sigma_n^{-1} (\tilde{Y} - X\beta)$$

and

$$\text{pen}(\beta) = 2\lambda_n \sum_{j=1}^M |\beta_j|$$

**Proposition B.1**: On the set $E_1$ we have for $n \geq 1$,

$$||X'\beta^* - X'\hat{\beta}||_n^2 + \lambda_n \sum_{j=1}^M |\hat{\beta}_j - \beta^*_j| \leq 4 \sum_{j \in J(\beta^*)} |\hat{\beta}_j - \beta_j^*|$$

where

$$E_1 = \bigcap_{j=1}^M \{ 2|V_j| \leq \lambda_n \}$$

and $V_j = \frac{1}{n} X_{-,j} \Sigma_n^{-1} \epsilon$

where $X_{-,j}$ is the $j$th column of the design matrix, and

$$||X'\beta^* - X'\hat{\beta}||_n^2 = \frac{1}{n} (X\beta^* - X\hat{\beta})' \Sigma_n^{-1} (X\beta^* - X\hat{\beta})$$

**Remark**: 

$$V_j = \frac{1}{n} \epsilon' \Sigma_n^{-1} X_{-,j} \forall i = 1, ..., M.$$ 

$$V' = (V_1, ..., V_M) = \frac{1}{n} \epsilon' \Sigma_n^{-1} X$$

$$V = \begin{pmatrix} V_1 \\ \vdots \\ V_M \end{pmatrix} = \frac{1}{n} X' \Sigma_n^{-1} \epsilon$$

**Proof**: From the definition of $\hat{\beta}$, see (35), we know

$$||\tilde{Y} - X'\hat{\beta}||_n^2 + 2\lambda_n \sum_{j=1}^M |\hat{\beta}_j| \leq ||\tilde{Y} - X'\beta^*||_n^2 + 2\lambda_n \sum_{j=1}^M |\beta^*_j|$$

This simplifies to

$$||X\beta^* - X'\hat{\beta}||_n^2 \leq \frac{2}{n} \epsilon' \Sigma_n^{-1} (X\hat{\beta} - X\beta^*) + 2\lambda_n \sum_{j=1}^M |\beta^*_j| - 2\lambda_n \sum_{j=1}^M |\hat{\beta}_j|$$
Now,
\[
\frac{2}{n} \varepsilon' \Sigma_n^{-1}(X\hat{\beta} - X\beta^*) = \frac{2}{n} \varepsilon' \Sigma_n^{-1}X(\hat{\beta} - \beta^*) \\
= 2V'(\hat{\beta} - \beta^*) \\
= 2 \sum_{j=1}^{M} V_j (\hat{\beta}_j - \beta_j^*)
\]

On the set \(E_1\) we have
\[
\frac{2}{n} \varepsilon' \Sigma_n^{-1}(X\hat{\beta} - X\beta^*) = 2V'(\hat{\beta} - \beta^*) = 2 \sum_{j=1}^{M} V_j (\hat{\beta}_j - \beta_j^*) \leq \lambda_n \sum_{j=1}^{M} |\hat{\beta}_j - \beta_j^*|
\]

Thus,
\[
||X'\beta^* - X'\hat{\beta}||_n^2 \leq \lambda_n \sum_{j=1}^{M} |\hat{\beta}_j - \beta_j^*| + 2\lambda_n \sum_{j=1}^{M} |\beta_j^*| - 2\lambda_n \sum_{j=1}^{M} |\hat{\beta}_j|
\]

or
\[
||X'\beta^* - X'\hat{\beta}||_n^2 + \lambda_n \sum_{j=1}^{M} |\hat{\beta}_j - \beta_j^*| \leq 2\lambda_n \sum_{j=1}^{M} |\hat{\beta}_j - \beta_j^*| + 2\lambda_n \sum_{j=1}^{M} |\beta_j^*| - 2\lambda_n \sum_{j=1}^{M} |\hat{\beta}_j|
\]

Observe that
\[
\sum_{j=1}^{M} |\hat{\beta}_j - \beta_j^*| + \sum_{j=1}^{M} |\beta_j^*| - \sum_{j=1}^{M} |\hat{\beta}_j| = \sum_{j \in J(\beta^*)} |\hat{\beta}_j - \beta_j^*| + \sum_{j \notin J(\beta^*)} |\beta_j^*| - \sum_{j \in J(\beta^*)} |\hat{\beta}_j| \\
= \sum_{j \in J(\beta^*)} |\hat{\beta}_j - \beta_j^*| + \sum_{j \notin J(\beta^*)} \{ |\beta_j^*| - |\hat{\beta}_j| \} \\
\leq \sum_{j \in J(\beta^*)} |\hat{\beta}_j - \beta_j^*| + \sum_{j \in J(\beta^*)} |\hat{\beta}_j - \beta_j^*| \\
= 2 \sum_{j \in J(\beta^*)} |\hat{\beta}_j - \beta_j^*|
\]

Thus on the set \(E_1\) we get
\[
||X'\beta^* - X'\hat{\beta}||_n^2 + \lambda_n \sum_{j=1}^{M} |\hat{\beta}_j - \beta_j^*| \leq 4\lambda_n \sum_{j \in J(\beta^*)} |\hat{\beta}_j - \beta_j^*|
\]

\[\blacksquare\]
**Theorem B.2:** For any $\delta > 0$ we can get a $\lambda_n$ such that
\[
P(\|\hat{\beta} - \beta^*\|_{\ell_1} > 5b\lambda_n k^*) \leq \frac{\delta}{2}
\]

**Proof:** From the Proposition B.1 we can conclude that on the set $E_1$ we have
\[
\lambda_n \sum_{j=1}^{M} |\hat{\beta}_j - \beta^*_j| \leq 4\lambda_n \sum_{j \in J(\beta^*)} |\hat{\beta}_j - \beta^*_j|
\]
or equivalently
\[
\sum_{j \notin J(\beta^*)} |\hat{\beta}_j - \beta^*_j| \leq 3 \sum_{j \in J(\beta^*)} |\hat{\beta}_j - \beta^*_j|
\]
which implies
\[
\sum_{j \notin J(\beta^*)} |\hat{\beta}_j - \beta^*_j| \leq 3 \sum_{j \in J(\beta^*)} |\hat{\beta}_j - \beta^*_j| + a\lambda_n^2 k^*
\]
Thus on the set $E_1$ we have $\hat{\beta} - \beta^* \in V_{\alpha, \epsilon}$ with $V_{\alpha, \epsilon}$ as defined in (36), for $\epsilon = a\lambda_n^2 k^*$ and $\alpha = 3$.
\[
V_{\alpha, \epsilon} = \left\{ v \in \mathbb{R}^M : \sum_{j \notin I^*} |v_j| \leq \alpha \sum_{j \in I^*} + \epsilon \right\}
\]
(36)

Proposition B.1 states that
\[
||X'\beta^* - X'\hat{\beta}||_n^2 + \lambda_n \sum_{j=1}^{M} |\hat{\beta}_j - \beta^*_j| \leq 4\lambda_n \sum_{j \in J(\beta^*)} |\hat{\beta}_j - \beta^*_j|
\]
Using the Cauchy Schwarz inequality in the right hand side of the inequality above, followed by an inequality of the type $2uv \leq au^2 + v^2/a$, for any $a > 1$, we further obtain
\[
||X'\beta^* - X'\hat{\beta}||_n^2 + \lambda_n \sum_{j=1}^{M} |\hat{\beta}_j - \beta^*_j| \leq 4a\lambda_n^2 k^* + \frac{1}{a} \sum_{j \in I^*} (\beta^*_j - \hat{\beta}_j)^2.
\]
(37)

Since $\hat{\beta} - \beta^* \in V_{\alpha, \epsilon}$ by Lemma 2.1 Bunea 2008 [1] we can say that
\[
||X'\beta^* - X'\hat{\beta}||_n^2 \geq \frac{1}{a} \sum_{j \in I^*} (\beta^*_j - \hat{\beta}_j)^2 - a\lambda_n^2 k^*
\]
(38)

Combining inequalities (37) and (38) we get
\[
|\hat{\beta} - \beta^*|_{\ell_1} \leq \frac{5b}{b} \lambda_n k^*
\]
where $b = \frac{1}{5}$. Thus we observe that on the set $E_1$ we have $|\hat{\beta} - \beta^*|_{\ell_1} \leq \frac{5b}{5} \lambda_n k^*$. Thus we can conclude that
\[
P(|\hat{\beta} - \beta^*|_{\ell_1} > \frac{5b}{b} \lambda_n k^*) \leq \mathbb{P}(E_1^c)
\]
\[
= \sum_{j=1}^{M} \mathbb{P}(\frac{2}{n} X'_{-j} \Sigma^{-1}_{-j} \epsilon_j > \lambda_n)
\]
\[
\leq \frac{\delta}{2}
\]
The choice of $\lambda_n$, that is

$$\lambda_n \geq \max\{C_1\sqrt{\frac{\log \frac{8M}{\delta}}{n}}, B_1\left(\frac{\log \frac{8M}{\delta}}{n}\right)^{\mu+\nu+2}\}$$

for suitably chosen constants $C_1$ and $B_1$, and the last inequality follows from , Theorem A.2, Concentration Inequality for weakly dependent random variables, see, in Appendix A.

**Bernstein’s Inequality in $AR(1)$ setup**

Here we prove a Bernstein-type inequality in a dependence situation. Let

$$\epsilon_i = a_{i-1} + \eta_i, \ i = 2, \ldots, n$$

$$\epsilon_1 = \eta_1$$

(39)

where, $0 < a < 1$ and $\eta_1, \ldots, \eta_n$ is a sequence of mean 0 independent random variables such that

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}|\eta_i|^m \leq \frac{m!}{2} a^2 b^{m-2} \forall \ m \geq 2$$

(41)

for some positive constants $a$ and $b$. For convenience we now state, without proof, Bernstein’s Inequality for independent random variables.

**Theorem B.3**: Let $\epsilon_1, \ldots, \epsilon_n$ be a sequence of random variables as defined by (39) Then, for any $\lambda_n > 0$ we have

$$\mathbb{P}\left(\frac{2}{n} \mathbf{X}_{-,j} \Sigma^{-1}_n \epsilon \geq \lambda_n\right) \leq 2 \exp\left(-\frac{n\lambda_n^2}{2(c^2 + d\lambda_n)}\right).$$

(42)

for some positive constants $c$ and $d$.

**Proof**: Observe that

$$\mathbf{X}_{-,j} \Sigma^{-1}_n \epsilon = \sum_{i=1}^{n} \gamma_i \epsilon_i$$

(43)

where $\gamma_i$ is the $i^{th}$ component of the vector $\mathbf{X}_{-,j} \Sigma^{-1}_n$. We know, from (43), that

$$\sum_{i=1}^{n} \gamma_i \epsilon_i = \sum_{i=1}^{n} \gamma_i \sum_{j=1}^{i} a^{i-j} W_j$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{i} \gamma_i a^{i-j} W_j$$

$$= \sum_{j=1}^{n} \sum_{i=j}^{n} \gamma_i a^{i-j} W_j$$

$$= \sum_{j=1}^{n} c_j W_j$$

32
where \( c_j = \sum_{i=j}^{n} \gamma_i a^{i-j} \). Thus

\[
P\left(\left| \frac{2}{n} \sum_{j=1}^{n} c_j X_j' \Sigma_n^{-1} c_i \right| \geq \lambda_n \right) = P\left(\left| \frac{2}{n} \sum_{j=1}^{n} c_j W_j \right| \geq \lambda_n \right)
\]

where \( W_i \) as we mentioned earlier are a sequence of i.i.d random variables which satisfies the Bernstine’s condition (41). The random variables \( c_j W_j \) are a sequence of independent random variables but we need to see if they also satisfy Bernstein’s Condition (41). Now

\[
\frac{1}{n} \sum_{j=1}^{n} \mathbb{E}|c_j W_j|^m \leq \max_{1 \leq j \leq n} |c_j|^m \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}|W_j|^m
\]

\[
\leq |k|^m \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}|W_j|^m
\]

\[
\leq \frac{m!}{2} c^2 d^{m-2}
\]

where \( c = ka_1 \) and \( d = kb_1 \). The preceding inequality follows from the fact that

\[
|c_j| = \left| \sum_{i=j}^{n} \gamma_i a^{i-j} \right| \leq \left| \sum_{i=j}^{n} \gamma_i a^{i-j} \right|
\]

\[
\leq \sum_{i=j}^{n} |\gamma_i|
\]

\[
\leq (1 + a)^2
\]

and

\[
|c_j| = \left| \sum_{i=j}^{n} \gamma_i a^{i-j} \right| \leq \left| \sum_{i=j}^{n} \gamma_i a^{i-j} \right|
\]

\[
\leq \max_{1 \leq i \leq n} |\gamma_i| \sum_{i=j}^{n} a^{i-j}
\]

\[
\leq (1 + a^2) \sum_{i=j}^{n} a^{i-j}
\]

\[
\leq \frac{(1 + a^2)}{1 - a}
\]

Let

\[
k = \frac{(1 + a^2)}{1 - a}.
\]

Thus we see the random variables \( c_i W_i \) are not just independent but they satisfy the Bernstein’s Condition as well which ensures that we can have a Bernstein type inequality here, that is
\[
\mathbb{P}(\frac{2}{n} \mathbf{x}' J \Sigma_n^{-1} \epsilon_i \geq \lambda_n) = \mathbb{P}(\frac{1}{n} \sum_{i=1}^{n} \gamma_i \epsilon_i \geq \lambda_n)
\]
\[
= \mathbb{P}(\frac{1}{n} \sum_{j=1}^{n} c_j W_j \geq \frac{\lambda_n}{2})
\]
\[
\leq 2 \exp\left(-\frac{n(\lambda_n/2)^2}{2(c^2 + d(\lambda_n/2))}\right)
\]
\[
\leq 2 \exp\left(-\frac{n(\lambda_n/2)^2}{4c^2}\right) + \exp\left(-\frac{n(\lambda_n/2)}{4d}\right)
\]
\[
\leq \frac{\delta}{2M}
\]

The penultimate inequality follows from the fact that
\[
\exp\{-\frac{x}{\alpha + \beta}\} \leq \exp\{-\frac{x}{2\alpha}\} + \exp\{-\frac{x}{2\beta}\}
\]
for any \(x\) and \(\alpha, \beta > 0\). The last inequality is true for the choice of \(\lambda_n\) as below
\[
\lambda_n \geq \max\{4c \sqrt{\frac{\log 8M}{n}}, 8d \frac{\log 8M}{n}\}
\]
where \(c = ka_1\) and \(d = kb_1\) and as mentioned before
\[
k = \min\{(1 + a^2), (1 + a)^2\}
\]
That is
\[
k = (1 + a)^2 \text{ when } a > \sqrt{2} - 1
\]
\[
= \frac{1 + a^2}{1 - a} \text{ when } a \leq \sqrt{2} - 1.
\]
Observe that we can always conclude that
\[
k \leq (1 + a)^2
\]
irrespective of the value of \(a\).

\textbf{Remark:} If we had decided to bound the probability
\[
\mathbb{P}(\frac{2}{n} \mathbf{x}' J \Sigma_n^{-1} \epsilon \geq \lambda_n) \leq \frac{\delta}{2M^2},
\]
which subsequently would have implied that
\[
\mathbb{P}(|\hat{\beta} - \beta^*| \epsilon_1 > \frac{5}{6} \lambda_n k^*) \leq \frac{\delta}{2M},
\]
\[
34
\]
then our choice of $\lambda_n$ would have been

$$\lambda_n \geq \max\{4c\sqrt{\frac{\log \frac{8M^2}{\delta}}{n}}, 8d\frac{\log \frac{8M^2}{\delta}}{n}\}.$$ 

Here $c = ka$ and $d = kb$ and as mentioned before

$$k = \frac{(1 + a^2)}{1 - a}$$
References


