ESTIMATION OF STABLE PARAMETERS FOR MULTIPLE AUTOREGRESSIVE PROCESSES VIA CONVEX PROGRAMMING

Somnath Chakraborty, Johannes Lederer, Rainer von Sachs

LIDAM Discussion Paper ISBA 2023 / 37



UCLouvain

ISBA | LIDAM Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication.html

Estimation of stable parameters for multiple autoregressive processes via convex programming

SOMNATH CHAKRABORTY[1,a], JOHANNES LEDERER[2,b] and RAINER von SACHS*[3,c]

[1] *Fakultät für Mathematik, Ruhr-Universität Bochum, 44801 Bochum, Deutschland,*
[a] *somnath.chakraborty@ruhr-uni-bochum.de*
[2] *Fachbereich Mathematik, Universität Hamburg, 20146 Hamburg, Deutschland,*
[b] *johannes.lederer@uni-hamburg.de*
[3] *Institute of Statistics, Biostatistics and Actuarial Sciences, LIDAM, UCLouvain, 1348 Louvain-la-Neuve, Belgium,* [c] *rainer.vonsachs@uclouvain.be*

We develop a finite-sample theory for estimating the coefficients and for the prediction of multiple stable autoregressive processes that (i) share an unknown lag order but (ii) can differ in their individual sample sizes. Our technique is based on penalisation similar to hierarchical, overlapping group-Lasso but requires a new mathematical set-up to accommodate (i) and (ii). The set-up differs from existing work considerably, for example, in that we estimate the common lag order directly from the data rather than using extrinsic criteria. We prove that the estimated autoregressive processes enjoy stability, and we establish rates for both the estimation and prediction error that can outmatch the known rates in our setting. Our insights on the lag selection and the stability are also of interest for the case of individual autoregressive processes.

MSC2020 subject classifications: Primary 62M10; 62L99; secondary 49M29
Keywords: autoregressive process; effective noise; statistical guarantees; tuning parameter; dual norm; hierarchical-group norm; group LASSO; lag selection; false discoveries; regularised least square; sample complexity; stable AR process; gaussian innovation; restricted eigenvalue property

1. Introduction

Today's time-series data in areas such diverse as macroeconomics and finance, weather predictions, and brain imaging often come in the form of a panel or multivariate vectors. Such data are, therefore, often high dimensional. The high dimensionality motivates the search for underlying structures across the different dimensions of the data. A popular modeling framework in this context is (parametric) vector autoregression (VAR). VAR assumes that the serial dependence is limited by a common maximal lag-order for all components. However, as the number of component series is increased, VAR models have the known tendency to become overparametrized.

To address this issue, *regularized* LASSO-type approaches for estimating the parameters in these models have been proposed (Nardi and Rinaldo, 2011). These approaches tend to outmatch more traditional ones, which select a lag order through AIC, BIC, or similar, assuming that a universal lag order applies to all components simultaneously. However, it was not until recently that these Lasso-type approaches also incorporated the notion of lag structure and/or lag order selection. Recently, Nicholson et al. (2020) embedded the selection of the lag structure into a convex regularizer. Their key modeling tool has been a group LASSO with nested groups, which guarantees that the sparsity pattern of the lag coefficients honors the VAR's ordered structure; for more details on the literature on dimension reduction methods which address the VAR's overparametrization problem we refer to Section 2 of their

work. However, importantly, Nicholson et al. (2020) focus on a single multivariate stable autoregressive process and assume perfect information on the maximum lag. Another clear shortcoming is the fact that components of the observed multivariate time series need to be of the same length, a constraint that is inherent to VAR-modelling.

A field that regularly deals with time series of very different lengths is natural-language processing (NLP). Research in NLP has made tremendous progress on the machine-learning front in recent years; for example, ever since appearance of the pioneering work of Vaswani et al. (2017), attention-based time-series models such as Transformers have generated a glut of research papers over the last few years (see, for example, Li et al. (2019) and references therein). The inputs to these models are texts, which are treated as a time series consisting of individual (embedded) words or language tokens. Naturally, these time series can be considered to be independent of each other, and they can differ quite substantially in their lengths, but it is imaginable that at least similar types of text might have common lags. While our linear models are still very far from complex network models such as Transformers, the example indicates the interest in dealing with time series of different lengths.

This paper proposes a method to analyse multiple stable autoregressive processes of (potentially) *different lengths* in the framework of regularized LASSO, where the regularization is achieved via an overlapping hierarchical group-norm that induces sparsity at the group level. In doing this, we have adapted the use of hierarchical overlapping group-LASSO, proposed in Mairal et al. (2011) — and recently put to use to deal with the single VAR case in Nicholson et al. (2020). We show that, even in absence of any information on the maximum lag of the processes, the proposed framework estimates the true lag and the coefficients of the AR model accurately when the number of samples is sufficient. Moreover, we are able to prove that the model fitted with the AR coefficients returned by this proposed method is stable. As our results on statistical guarantees are essentially of non-asymptotic nature, which is interesting even in the context of observing a single time series, we first review the (sparse) literature on those non-asymptotic results in a time-series context.

Essentially, the very scarce literature dealing with algorithmic learning of single VAR-coefficients using LASSO-based methods observes that, algorithmically, the method by Mairal et al. (2011) (and its computationally faster implementation using a technique in Tseng (2008)) can be employed in learning vector autoregressive process via certain group-LASSO analysis, to address — in the presence of prior information on the true unknown lag order — estimation of order L vector-autoregressive models of dimension M, abbreviated $\mathrm{VAR}_M(L)$ in the sequel.

Our work can be seen as fitting a common (i.e. "diagonal vector") autoregressive model to a multivariate time series of dimension M and diagonal AR coefficient matrices, that is, a panel of M observed (univariate) "similar" time series — but of *in general not equal* lengths $n_m, 1 \leq m \leq M$. Our notion of similarity here resembles the idea behind factor models, namely, that unobserved (and actually not modeled) "latent" phenomena (such as an intrinsic memory or lag mechanism of a common hidden market indicator) proposes to use the *the same* lag order L for each of the (potentially different length) time series in the panel. Notice that, because of multiple AR processes of unequal lengths, one can not 'simply diagonalize' the settings of the references cited above to estimate the AR-coefficients.

We address the challenging question on how to choose, solely from the information available in the model for the observed data, the common appropriate lag order L that allows us to phrase and solve our problem via a penalised LASSO approach. We derive non-asymptotic bounds on the multivariate one-step ahead prediction error and estimate the collection of the autoregressive coefficients $\beta := (\beta_1^m, \ldots, \beta_L^m)_{1 \leq m \leq M}$ under the paradigm of sparseness. Assuming that there is a common true unknown lag order L_0 that generated our M time series, our estimation pipeline is based on a modification of a hierarchical group LASSO approach and uses the algorithm in Nicholson et al. (2020) — for fitting a common model — as a subroutine. We first determine an appropriate (minimal) upper bound $L \geq L_0$ depending essentially on the sample size $n_{\min}$ of the shortest observed time series component (and on

M, of course) from a thorough analysis of the theoretical complexity of our group-LASSO based approach. With this appropriate L, necessary to embed our autoregression problem into the framework of high-dimensional multivariate regression, we transfer existing technology on LASSO estimation (with overlapping hierarchically constructed groups) to our problem. We derive statistical guarantees for the estimators $\hat{\beta}$, solution of our aforementioned learning algorithm, and for the estimator $\hat{L}_0$ (essentially taken from the support of $\hat{\beta}$). More specifically we deliver non-asymptotic bounds on the multivariate one-step ahead prediction error, on the estimation error of β, on the false discoveries for the support of β, and quite innovatively on the stability of the fitted model. For the latter, we show that the fitted autoregressive model of order $\hat{L}_0$ with estimated coefficients $\hat{\beta}$ fulfils the conditions of the true model for stability (via a more explicit concept of ε-stability that we introduce to asses the difficulty of the statistical estimation problem).

In the following paragraphs, we are more explicit about the exact nature of our contributions motivated from the existing limitations of the current approaches we found in the literature.

Current limitations (L) **and our contributions** (C)

Some of the questions that are not sufficiently addressed in recent existing work on non-asymptotic time series analysis are as follows.

L1 Via an application of Mairal et al. (2011), it is now known that an approach based on LASSO with overlapping groups can determine the component-wise lag orders L_m of stable, high-dimensional vector autoregressive processes (including the "cross-over" lag order L_{ij} of the dependence of the i-th component on the j-th component of a VAR model). The relevant formulation based on (dual) convex programming is computationally attractive and intriguing more generally, but it requires a uniform upper-bound on the L_{ij}'s as a parameter both in their theoretical bounds and in practice. The current literature either ignores this issue altogether or sets those bounds based on model selection such as AIC or BIC or via Bayesian shrinkage, whose suitability is unclear here. In fact, if an external lag selection method outputs an incorrect lag order, then deriving statistical guarantees for the LASSO as formulated in equation (3.7) will be difficult or even impossible to achieve. Our approach avoids this problem by integrating lag order selection within parameter estimation via the LASSO. Moreover, we thereby avoid some potential loss of sample-efficiency that would occur if we used an external lag selection method such as AIC or BIC, e.g., on sample-splitting. Finally, our method has the clear advantage to be more time efficient than one that would carry out order selection via AIC/BIC for each of the M time series separately.

C1 We establish a suitable estimate of the lag order.
Our estimate of this lag order guarantees the following:
1. the smallness of the empirical prediction risk (see Theorem 4.4 and the discussion following the theorem),
2. a resulting tuning parameter that is not too large (see Corollary 4.6), and
3. the restricted eigenvalue property of the data matrix to hold (Theorem 4.7 and the Corollary 4.8 immediately after that).

In this sense, our estimated lag order is optimal for our setup.

L2 Real world applications often involve multiple univariate, decoupled time series of varying lengths — an example of such occurrence is discussed in the introduction section. This would translate to VAR modeling with diagonal coefficient matrices. However, purely transferring existing results from a $\mathrm{VAR}_M(L)$ modeling approach can be cumbersome in practice for situations in which the number of available samples for each individual component time series is not the same: obviously needing to chop off the samples in order to work with the minimal individual sample size could result in weaker theoretical guarantees and practical performances (see below).

C2 Applying a hierarchical group norm enables us to derive prediction guarantees that depend on the *average* number of samples per component (instead of the minimum number of samples per model, as would be the case had we translated naively to a M-dimensional VAR model). More specifically, availability of perfect information on the true lag L_0 and $n_1 = L_0 + T_1, \cdots, n_m = L_0 + T_M$ (respective) number of samples for the M individual components yields the following: the $\mathrm{VAR}_M(L)$ translation would result in the following provable upper-bound on error (see Basu and Michailidis (2015)), stating that with high probability

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 = O\left(\sqrt{\frac{\log(M^2 L_0)}{n_{\min} M}}\right), \tag{1.1}$$

whereas the error from the algorithm we describe is of the order (see equation (4.17) below)

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 = O\left(\sqrt{\frac{\log(M L_0)}{D}}\right). \tag{1.2}$$

Here $D = T_1 + \cdots + T_M$ is the total number of "postsamples" ($T_m := n_m - L_0$). Furthermore, our strategy results in weaker dependency of the error on the behaviour of the reverse characteristic polynomial on the unit disk, unlike in the relevant $\mathrm{VAR}_M(L)$ model translation (compare with Basu and Michailidis (2015)). Our bound of the effective noise, namely (see (4.6)),

$$\mathbb{P}\left[\frac{2}{D}\mathcal{N}_\star(X^\top \boldsymbol{U}) \geq \frac{3\eta}{2\sqrt{L}}\right] \;\leq\; \delta\,.$$

requires a set of 'post'-samples of cardinality

$$D \geq \frac{8\sigma_{\max}^2(1+\epsilon^{-2}+\epsilon^{-4})}{c_0\eta}\log\left(\frac{ML}{\delta}\right),$$

a saving of a factor of M that results due to an appropriate application of a dual norm bound, and it is not obvious to us if such a bound would follow naively from existing results for single VAR mentioned in the references cited above.

L3 Recall that a (univariate) autoregressive process $X_t = a_1 X_{t-1} + \cdots + a_L X_{t-L} + U_t$ is stable if the "reverse characteristics polynomial" $1 - a_1 z - \cdots - a_L z^L$ has no complex roots on the closed unit disk; it is known that stability implies stationarity (see Section 2 below). But what about stability for *fitted* autoregressive models? While this question has an affirmative answer in the special case of Yule-Walker estimation of the coefficients of a univariate autoregressive process (known however to be less efficient), the question of stability or stationarity of the process reconstructed from the parameters estimated by LASSO-based approaches does not seem to have been addressed in the literature. However, starting with a stable process to have generated the input observations of these devised algorithms, it is reasonable to expect that the reconstructed process be stable (and thus, multi-step predictions be reliable as well).

C3 We show, in Theorem 4.11, that the process reconstructed from the parameters returned by our algorithm is stable when the samples available as input are generated by stable processes. As Basu and Michailidis (2015) showed, a measure of stability for an autoregressive process is, equivalently, a boundedness criteria on the spectral density, and the boundedness in the Fourier domain translates in a sense to 'smoothness' of the process in the temporal domain. Thus, as, intuitively, the stable autoregressive processes form a 'smooth' subclass, it is desirable that any algorithm for

learning the parameters of processes from this smooth subclass should return estimators lying in this subclass; in this paper, this is ensured by Theorem 4.11. To the best of our knowledge, this important aspect of time series estimation has been hitherto neglected in the literature.

It is important to mention here that the results in this paper demonstrate that our overlapping group-LASSO approach yields a stable process when the underlying process is stable as well. The aim of the paper is not to choose the most optimal tuning parameters or some absolute constants, but rather to show existence of these proposing reasonable candidates for such parameters/constants.

Organization of the paper

The paper is organized as follows. In Section 2, we recall relevant definitions from existing literature, and we set notations. In Section 3, (1) we specify the model and formulate the learning problem as a regularized group-LASSO with overlapping group norm, where the data matrix is a block-diagonal matrix — each block of which consists of the data matrix that treats a least-squares problem corresponding to the associated component time series; (2) we present the learning algorithm based on the group-LASSO problem. Section 4 contains the bulk of the technical contents, in particular, the proof of the statements of the main results, already presented at the end of Section 3. This Section 4 is divided into four subsections: Subsection 4.1 presents an oracle inequality bounding the one-step-ahead prediction error (Theorem 4.2), as well as a high-probability bound on the effective noise of the model (Theorem 4.4). Subsection 4.2 starts with restricted eigenvalue bounds (Proposition 4.7) for the blocks of the data matrix, and goes on to integrate the blockwise results to finally arrive at an estimate (Theorem 4.10) of the error in estimating the AR-coefficients. Finally, combining the results from these subsections, stability of the estimated AR model (Theorem 4.11) has been established in Subsection 4.3. All proofs of technical and auxiliary results are deferred to a series of subsections of Section 7.

2. Preliminaries

2.1. General notations

- $M_d(\mathbb{F})$ denotes the ring of $d \times d$ matrices with entries in the field $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, and for $M \in M_d(\mathbb{F})$, we write $M^\top$ for the transpose; if $\mathbb{F} = \mathbb{C}$, then $M^\star$ denotes the conjugate transpose.
- $\mathbb{D}$ is the complex closed unit disk $\mathbb{D} := \{z \in \mathbb{C} : |z| \leq 1\}$, and its boundary is $\partial\mathbb{D} := \{z \in \mathbb{C} : |z| = 1\}$.
- For any integer $d > 0$, if $x \in \mathbb{R}^d$, then $\|x\|_2 = \sqrt{x^\top x}$, and $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$; in particular, under standard identification $\mathbb{C} = \mathbb{R}^2$, we have $\mathbb{S}^1 = \partial\mathbb{D}$.
- For integer $n > 0$, we will denote the set $\{1, 2, \cdots, n\}$ by $[n]$.
- For a vector $\hat{\boldsymbol{\beta}}_m = (\hat{\boldsymbol{\beta}}_{m,1}, \cdots, \hat{\boldsymbol{\beta}}_{m,L})$ and an integer $0 < L_0 \leq L$, we write

$$\hat{\boldsymbol{\beta}}_m(L_0) := (\hat{\boldsymbol{\beta}}_{m,1}, \cdots, \hat{\boldsymbol{\beta}}_{m,L_0}) \,. \tag{2.1}$$

2.2. Notations for autoregressive process

Conventions: We use $X, Y, Z, \ldots$ to denote random variables, and $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \ldots$ to denote random vectors. On the other hand, we use $a, b, c, \ldots$ to denote real (or complex) constants, and $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \ldots$ for vector-valued constants.

Notations: • For d-dimensional autoregressive process

$$X_t := A_1 X_{t-1} + \cdots + A_L X_{t-L} + U_t \,, \tag{2.2}$$

and $z \in \mathbb{C}$, we write

$$\mathcal{A}_z := I - A_1 z - \cdots - A_L z^L \,. \tag{2.3}$$

- L will denote an initially determined "ad-hoc" upper-bound on the true lag of the process in (2.2), and L_0 the true lag.

Definition 2.1 (Weak stationarity). A d-dimensional time series $\{X_t\}_{t\in\mathbb{Z}}$ is said to be *weakly stationary* if the following holds: *a)* $\mathbb{E}[\|X_t\|_2^2] < \infty$ for all $t \in \mathbb{Z}$, *b)* $\mathbb{E}[X_t] = \mu$ for all $t \in \mathbb{Z}$, and *c)* $\mathbb{E}[X_t X_{t-h}^\top] = \Gamma(h)$ for all $t, h \in \mathbb{Z}$.

Definition 2.2 (Strong stationarity). A d-dimensional time series $\{X_t\}_{t\in\mathbb{Z}}$ is said to be *strongly stationary* if for each integer $n > 0$, and all integers $t_1, \cdots, t_n, h$, the distributions of the vectors $(X_{t_1}, \cdots, X_{t_n})$ and $(X_{t_1+h}, \cdots, X_{t_n+h})$ are identical.

Definition 2.3 (Autoregressive time series). A d-dimensional time series $\{X_t\}_{t\in\mathbb{Z}}$ is autoregressive of lag at most $L > 0$ if there are $d \times d$ matrices $A_1, \cdots, A_L$ such that

$$X_t = A_1 X_{t-1} + \cdots + A_L X_{t-L} + U_t \,. \tag{2.4}$$

holds for all $t \in \mathbb{Z}$, for some random white noise process $\{U_t\}_{t\in\mathbb{Z}}$.

Associated to each d-dimensional lag-L autoregressive process $\{X_t\}_{t\in\mathbb{Z}}$ — as in equation (2.4) — is the associated order-1 process $\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{U}_t$, where

$$\mathbf{A} := \begin{pmatrix} A_{1\to L}, A_L \\ \mathbf{I}_{dL-d}\,, \mathbf{0} \end{pmatrix}, \qquad \mathbf{U}_t := \begin{pmatrix} U_t \\ \mathbf{0} \end{pmatrix}, \tag{2.5}$$

and $A_{1\to L}$ is the block matrix $(A_1 \cdots A_{L-1})$.

Definition 2.4 (Stability). A d-dimensional lag-L autoregressive process $\{X_t\}_{t\in\mathbb{Z}}$ — as in equation (2.4) — is said to be stable if $\det(\mathbf{I} - \mathbf{A}z) \neq 0$ for $|z| \leq 1$. Equivalently, the process is stable if $\det(\mathcal{A}_z) \neq 0$ for $|z| \leq 1$.

Definition 2.5 (Reverse characteristic polynomial). The polynomial $\det(\mathcal{A}_z)$ is called the *reverse characteristic polynomial* of the process in equation 2.4.

We note the equality $\det(\mathbf{I} - \mathbf{A}z) = \det(\mathcal{A}_z)$. In particular, the process in equation (2.4) is stable if and only if every eigenvalue of $\mathbf{A}$ is inside the open unit disk.

Definition 2.6 (ϵ-stability). A stable autoregressive process, as in equation (2.4), is said to be ϵ-stable for an $\epsilon \in (0, 1)$ if the following holds:

$$\epsilon \leq \min_{|z|=1} |\det(\mathcal{A}_\mathbf{z})| \leq \max_{|z|=1} |\det(\mathcal{A}_\mathbf{z})| \leq \epsilon^{-1} \,. \tag{2.6}$$

Remark (A). By maximum modulus principle, this is equivalent to saying that

$$\epsilon \leq \min_{|z|\leq 1} |\det(\mathcal{A}_{\mathbf{z}})| \leq \max_{|z|\leq 1} |\det(\mathcal{A}_{\mathbf{z}})| \leq \epsilon^{-1} .$$

The lower-bound here is a convenient quantification of the notion of stability, which demands that $\min_{|z|\leq 1} |\det(\mathcal{A}_{\mathbf{z}})| > 0$. Note that we also require the upper bound to derive our statistical guarantees.

A well-known fact about autoregressive processes is the following: see Lütkepohl (2005, proposition 2.1) for details.

Lemma 2.7 (Stability implies weak stationarity). *A stable autoregressive process is weakly stationary.*

3. Statistical Model and Estimator

This section introduces our statistical model and estimator, and presents an algorithm to learn the parameters of the model from observed samples.

3.1. Statistical Model

We start with the model. Suppose that we observe time-samples generated by M univariate autoregressive process, for which we know a uniform upper-bound L of the true lag-order. Then, we can aggregate these M univariate lag at most L autoregressive processes

$$\begin{aligned} X_t^1 &= \beta_1^1 X_{t-1}^1 + \cdots + \beta_L^1 X_{t-L}^1 + U_t^1 ; \\ &\vdots \\ X_t^M &= \beta_1^M X_{t-1}^M + \cdots + \beta_L^M X_{t-L}^M + U_t^M . \end{aligned} \tag{3.1}$$

In this paper, we work under the simplified assumption that the true lag of all the M component processes is identical, namely, L_0, and that $L \geq L_0$ is generic; neither L_0 nor L is known *a priori*. Additionally, we assume mean-zero, Gaussian white-noise innovations; that is, for each $m \in [M]$ the set $\{U_t^m\}_{t\in\mathbb{Z}}$ consists of independent mean-zero, univariate Gaussians with coordinate-wise standard deviation $\sigma_m \in (0, \infty)$. Additionally, we assume that for each $t \in \mathbb{Z}$, the noise variables $U_t^1, \ldots, U_t^M$ are independent. We summarize the parameters of the model in a matrix $\Theta \in \mathbb{R}^{M\times L}$ via $\Theta_{ml} := \beta_l^m$ to refer to groups of parameters more easily later. Our goal is 1. to estimate the parameters of *all* models simultaneously; and 2. to assess the lag L_0, which is assumed to be the same over all M processes.

We make the following assumption on the absolute value of the smallest β-coefficient, which is widely known as *β-min assumption* in the LASSO literature; see, for example, Bunea (2008). We note that this assumption will only be needed to achieve the bound in Theorem 3.1 after λ-thresholding; in particular, when no thresholding is employed, the analysis in this paper does not require the assumption.

Assumption 1 (*β-min assumption*). *There is an absolute constant $c_\beta > 0$ such that the true autoregressive coefficient vector $\boldsymbol{\beta}$ satisfies*

$$\boldsymbol{\beta}_j^m \neq 0 \implies |\boldsymbol{\beta}_j^m| \geq c_\beta . \tag{3.2}$$

Broadly speaking, this assumption ensures that the non-zero coefficients can be detected in the first place.

We now set out to define a regularizer. Let n_j denote the total number of samples from

$$X_t^j = \beta_1^j X_{t-1}^j + \cdots + \beta_L^j X_{t-L}^j + U_t^j \,,$$

and $T_j := n_j - L$. The main idea is as follows. Suppose that $L \geq L_0$ is some integer, and for each $t \in \{-L+1, \ldots, 1, \ldots, T_m\}$ and $m \in \{1, \ldots, M\}$, we have an observation x_t^m of X_t^m. Denote $\boldsymbol{\beta}_m := (\beta_1^m, \ldots, \beta_L^m)^\top$ for each $m \in [M]$, and let $\boldsymbol{\beta} := (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_m^\top)^\top$. We define $\mathcal{G}_1, \ldots, \mathcal{G}_L \subset S_{ML} := \{1, 2, \ldots, M\} \times \{1, 2, \ldots, L\}$ by

$$\mathcal{G}_l := \{1, 2, \ldots, M\} \times \big\{l, l+1 \ldots, L\big\} \tag{3.3}$$

for all $l \in \{1, \ldots, L\}$. The groups are nested: $\mathcal{G}_1 \supset \cdots \supset \mathcal{G}_L$. Let $\boldsymbol{\beta}_{\mathcal{G}_l} \in \mathbb{R}^{M \times (L-l+1)}$ be the submatrix of $\boldsymbol{\Theta}$, consisting of columns having index larger or equal to l. We set the group norm to be

$$\|\boldsymbol{\beta}\|_{\mathcal{G}} := \sum_{l=1}^{L} \sqrt{M(L-l+1)}\, \|\boldsymbol{\beta}_{\mathcal{G}_l}\|_{\mathbb{F}} \,, \tag{3.4}$$

$$\text{where} \qquad \|\boldsymbol{\beta}_{\mathcal{G}_l}\|_{\mathbb{F}} := \sqrt{\sum_{m=1}^{M} \sum_{j=l}^{L} |\beta_j^m|^2} \,.$$

is the Frobenius norm of $\boldsymbol{\beta}_{\mathcal{G}_l}$. We will alternatively write $\mathcal{N}(\boldsymbol{\beta})$ for the group norm $\|\boldsymbol{\beta}\|_{\mathcal{G}}$, for the sake of notational ease.

The overall post-sample size is denoted

$$D := T_1 + \cdots + T_M \,.$$

In order to estimate the coefficient vector $\boldsymbol{\beta}$, we propose solving the following constrained convex program

$$\text{minimize} \qquad \frac{1}{D} \sum_{m=1}^{M} \sum_{t=1}^{T_m} \big(x_t^m - \beta_1^m x_{t-1}^m - \cdots - \beta_L^m x_{t-L}^m\big)^2 + \lambda \|\boldsymbol{\beta}\|_{\mathcal{G}} \,, \tag{3.5}$$

with an appropriate tuning parameter $\lambda > 0$. The objective function can be put in a concise form. For this, we define the vector $\boldsymbol{y} \in \mathbb{R}^D$, the matrix $X \in \mathbb{R}^{D \times (ML)}$, and the parameter $\boldsymbol{\beta} \in \mathbb{R}^{ML}$, as follows:

$$\boldsymbol{y} := (x_1^1, \ldots, x_{T_1}^1, x_1^2, \ldots, x_{T_2}^2, \cdots, x_1^M, \ldots, x_{T_M}^M)^\top ;$$

$$X := \begin{pmatrix} x^1_{1-1},\dots,x^1_{1-L} & & & & \\ \vdots & & & & \\ x^1_{T_1-1},\dots,x^1_{T_1-L} & & & & \\ & x^2_{1-1},\dots,x^2_{1-L} & & & \\ & \vdots & & & \\ & x^2_{T_2-1},\dots,x^2_{T_2-L} & & & \\ & & \ddots & & \\ & & & x^M_{1-1},\dots,x^M_{1-L} \\ & & & \vdots \\ & & & x^M_{T_M-1},\dots,x^M_{T_M-L} \end{pmatrix}; \tag{3.6}$$

$$\boldsymbol{\beta} := (\beta^1_1,\dots,\beta^1_L,\beta^2_1,\dots,\beta^2_L,\cdots,\cdots,\beta^M_1,\dots,\beta^M_L)^\top.$$

It is immediate that

$$\sum_{m=1}^{M}\sum_{t=1}^{T_m}\left(x^m_t - \beta^m_1 x^m_{t-1} - \cdots - \beta^m_L x^m_{t-L}\right)^2 \;=\; \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 .$$

In conclusion, the above estimation program is equivalent to

$$\widehat{\boldsymbol{\beta}} \;\in\; \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{ML}}\left\{\frac{1}{D}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_{\mathcal{G}}\right\}. \tag{3.7}$$

Hence, the estimator can be cast as a modified group-lasso estimator, which means that we can use established group-lasso algorithms that allow for overlapping groups (Mairal et al., 2011). In essence, the above estimator generalizes the *elementwise* estimator $\mathrm{HLag}^{\mathrm{E}}$ in Nicholson et al. (2020) to multiple time series. Note that in strong contrast to ordinary LASSO estimators, as well as any group-LASSO estimators with non-overlapping groups, our estimator enforces sparsity by setting coefficients as follows: using as input only an upper-bound to the true-lag L_0 — *all* coefficients indexed between $L_0 + 1$ and L are set to zero whilst those with indices smaller or equal to L_0 are not. This is precisely achieved by the penalty obtained via the nested groups $\mathcal{G}_1 \supseteq \cdots \supseteq \mathcal{G}_L$.

Example (Regularizer). We consider the case of two univariate lag (at most) three autoregressive processes; that is, $M = 2$ and $L = 3$. The corresponding groups are the following:

$$\begin{aligned}
\mathcal{G}_1 &= \{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3)\}\,;\\
\mathcal{G}_2 &= \{(1,2),(1,3),(2,2),(2,3)\}\,;\\
\mathcal{G}_3 &= \{(1,3),(2,3)\}\,.
\end{aligned}$$

Thus, the group norm of $\boldsymbol{\beta} \in \mathbb{R}^{2\times 3}$ is

$$\begin{aligned}
\|\boldsymbol{\beta}\|_{\mathcal{G}} = \sqrt{6}\sqrt{(\beta^1_1)^2 + (\beta^2_1)^2 + (\beta^1_2)^2 + (\beta^2_2)^2 + (\beta^1_3)^2 + (\beta^2_3)^2} \\
+\; \sqrt{4}\sqrt{(\beta^1_2)^2 + (\beta^2_2)^2 + (\beta^1_3)^2 + (\beta^2_3)^2}
\end{aligned}$$

$$+ \sqrt{2}\sqrt{(\beta_3^1)^2 + (\beta_3^2)^2}\,.$$

Notice that, when the (regularized) LASSO sets a certain group (say the second group above) to $\mathbf{0}$, it automatically sets all the following groups to $\mathbf{0}$ as well. More specifically, when the (regularized) LASSO sets $\boldsymbol{\beta}_{\mathcal{G}_l}\mathbf{0}$, then the hierarchical structure $\mathcal{G}_l \supseteq \mathcal{G}_{l+1} \supseteq \cdots$ means that for all $r > 0$, each coordinate of $\boldsymbol{\beta}_{\mathcal{G}_{l+r}}$ comes as a coordinate of $\boldsymbol{\beta}_{\mathcal{G}_l}$, thus ensuring $\boldsymbol{\beta}_{\mathcal{G}_{l+r}} = \mathbf{0}$ for each $r \geq 0$.

In what follows, we use the following notations. For $m \in [M]$ and $l \in [L]$, and $t \leq T_m$, we write

$$\begin{aligned} \boldsymbol{U}^{(m)} &:= (U_1^m, \cdots, U_{T_m}^m)^\top; \\ X^{(m,l)} &:= (X_{1-l}^m, \ldots, X_{T_m-l}^m)^\top; \\ X_t^{(m)} &:= (X_{t-1}^m, \ldots, X_{t-L}^m); \\ X^{(m)} &:= (X^{(m,1)}, \ldots, X^{(m,L)})\,. \end{aligned} \tag{3.8}$$

Moreover, we will write $X_{<j}^{(m)}$ to denote any of the variables $X_{j'}^m$ for $j' < j$.

3.2. Estimation Pipeline

Learning autoregressive coefficients of multiple time series of potentially different lengths and identical true lag is more complex than just the usual group-LASSO problem, where the groups form a partition of the index set. From a methodological perspective, some immediate technical challenges are 1. deciding on what L should be taken in the formulation of the convex problem (3.5), 2. how to disentangle the dual of the group norm (in order to apply Hölder's inequality to derive oracle prediction guarantees as in subsection 4.1 below); from a practical perspective, the challenge lies in incorporating the varying number of samples into the convex problem.

We now give a high-level overview of our estimation pipeline (described below) that takes as input the multiple time series, and forms the appropriate convex problem in the form of (3.7), and solves this convex problem via using the stochastic proximal gradient method appeared in Nicholson et al. (2020) as a subroutine. In essence, the idea is to start looking into the data set to first find the samples corresponding to the component which has the minimum number of samples (breaking ties arbitrarily). We then use this component to find the initial input lag, L, as described in (4.22). In the next step, we solve the regularized least-squares problem in (3.5) — where the penalty function is the group norm

$\mathcal{N}$, discussed further below (see (3.4)). The rate of convergence of the procedure is quadratic in number of computational steps as discussed in the work cited above.

Algorithm 1: AR Coefficient Estimation Pipeline

Data: samples from component stable processes, parameters $A \geq 1$ and $\delta > 0$, stability parameter $\epsilon \in (0, 1)$

Result: estimated lag $\hat{L}_0$, and estimated autoregression coefficient vector $\tilde{\boldsymbol{\beta}}$

1. Let $n_{\min}$ be the minimum number of samples from the components. Solve for L (see equation (4.22)):
$$n_{\min} = L + 84Ae\zeta^{-2}L\log\left(\frac{ML}{\delta}\right).$$
2. Use the learning algorithm in Nicholson et al. (2020) as subroutine to solve the convex problem in (3.7) with L as above; call the output $\hat{\boldsymbol{\beta}}$.
3. If $\boldsymbol{\beta}_m = \boldsymbol{\beta}_{m'}$ for all $m, m' \in [M]$ (equivalently, the component time series are from identical AR process), return $\hat{L}_0 := \max\{j : |(\hat{\boldsymbol{\beta}}_1)_j| > \lambda\}$ and $\tilde{\boldsymbol{\beta}} := (\hat{\boldsymbol{\beta}}_0', \cdots, \hat{\boldsymbol{\beta}}_0')$, where (refer Equation (2.1) for notation)
$$\hat{\boldsymbol{\beta}}_0' := \frac{1}{M}\sum_{m=1}^{M}\hat{\boldsymbol{\beta}}_m(\hat{L}_0)\,;$$
else, return
$$\hat{L}_0 := \max\{j : |\hat{\boldsymbol{\beta}}_j^m| > \lambda \text{ for some } m \in [M]\}\,,$$
$$\text{and} \qquad \tilde{\boldsymbol{\beta}} := (\hat{\boldsymbol{\beta}}_1(\hat{L}_0), \cdots, \hat{\boldsymbol{\beta}}_m(\hat{L}_0))\,.$$

The following statement is the main theorem on the theoretical properties of the output of the estimation pipeline. This theorem is essentially a summary of the results contained in Section 4, and will be discussed in all detail in subsections 4.2.2 and 4.3, as indicated below.

Theorem 3.1 (Main Theorem). *Let $\hat{\boldsymbol{\beta}}$ be the output of the group-LASSO in* (3.7) *above, with*
$$\lambda = 24(84Ae)^{\frac{1}{2}}\zeta^{-1}\sigma_{\max}^2 C_\sharp^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}}\,,$$
where $\zeta = 6^{-3}\epsilon^4$, and $L \geq L_0$ satisfying
$$n_{\min} = L + 84Ae\zeta^{-2}L\log\left(\frac{ML}{\delta}\right).$$
Suppose that the β-min condition (3.2) *holds with $c_\beta = \lambda$. If the total number D of post-samples satisfies*
$$D \geq 3^9\cdot(84Ae)C_\sharp^5(1+\epsilon^{-2}+\epsilon^{-4})^2\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^4(\epsilon^3\zeta)^{-2}M^a L_0^2 L^3\log\left(\frac{ML}{\delta}\right)\log(2L)\,,$$

if

$$\alpha := \min_{m\in[M]} \frac{T_m\sigma_m^2}{D},$$

and if $T_{\min} \geq 84eA\zeta^{-2}L_0 \log L$*, then the following holds with high probability:*

1. *the estimation error is given by*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \leq \frac{81(84Ae)^{\frac{1}{2}}LL_0\sigma_{\max}^2 C_\sharp^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})}{\zeta\alpha\epsilon^2}\sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{D}};$$

2. *if* $S_\lambda := \{j : |\hat{\boldsymbol{\beta}}_j| > \lambda\}$*, then the false discovery is bounded by the following inequality:*

$$|S_\lambda \setminus \mathrm{supp}(\boldsymbol{\beta})| \leq \frac{243(84Ae)^{\frac{1}{2}}LL_0^{\frac{3}{2}}\sigma_{\max}^2 C_\sharp^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})}{\zeta\alpha\epsilon^2\lambda}\sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{D}}.$$

3. *the AR-models — fitted with coefficients* $\hat{\boldsymbol{\beta}}_0$ *returned by the Algorithm 1 ("AR Coefficient Estimation Pipeline") — are stable, with high probability.*

Proof of Theorem 3.1. Subject to the stated β-min condition (3.2), this follows immediately from Theorem (4.10) in combination with Theorem (4.11). □

In the pipeline above, it is necessary to consider two distinct stability parameters $A \geq 1$ and δ, as it is not possible to integrate them into a single parameter — due mainly to the different number of samples from the component processes in our set-up.

Also, we separately mention the two cases (of $\boldsymbol{\beta}_m$'s being identical (or not) for all m) in order to specifically emphasize that in the first case, our algorithm requires a smaller number of samples than in the later case (which allows savings of a factor of M in the sample complexity).

Descent algorithms like the proximal gradient method used in the subroutine above might produce small non-zero valued parameters as numerical artifacts. One could consider other types of algorithms instead, but the required number of computational steps could be much higher. Moreover, those artifacts, and statistical false positives more generally, can also be controlled by standard λ-thresholding as mentioned in the algorithm.

4. Statistical Guarantees

This section contains the main theoretical results of this paper. We begin with deriving bounds for the one-step ahead prediction error, first formulated by an oracle inequality (see Theorem 4.2), which depends on the tuning parameter λ of our least-squares penalisation approach. Then we control this tuning parameter by controlling the effective noise of our Lasso-optimisation problem. Both things together will finally yield more explicit rates of our one-step ahead prediction error. The second part of this section treats control of the estimated autoregressive coefficients, including control of false discovery for their support (Theorem 4.10). In the end we present our result on stability of the estimated AR-model (Theorem 4.11).

To start with, we briefly recall the notion of the dual norm of our group-LASSO norm $\mathcal{N}$ defined above.

Our underlying space is $\mathbb{R}^{ML}$. Observe that

$$\boldsymbol{\beta} \mapsto \mathcal{N}(\boldsymbol{\beta}) := \sum \sqrt{|\mathcal{G}_l|} \cdot \|\boldsymbol{\beta}_{\mathcal{G}_l}\|_2$$

is a norm for any $\mathcal{G} = \{\mathcal{G}_1, \cdots, \mathcal{G}_l\}$ that covers $[ML]$. However, this norm is singular at any point where all the coordinates in a group vanish. Moreover, in presence of a vanishing group, all the coordinates in all of the smaller-sized groups will vanish, which is important in our analysis, since we will basically care about sparse solutions. Thus, we can not appeal to differential techniques to get bounds on the norm, but rather need to appeal to the dual norm approach.

Definition 4.1 (Dual norm). For a norm $\mathcal{N}$ on $\mathbb{R}^{ML}$, the dual norm $\mathcal{N}_\star(\boldsymbol{\alpha})$ of $\boldsymbol{\alpha} \in \mathbb{R}^{ML}$ is the optimum solution of the following convex program:

$$\begin{aligned} &\text{maximize } \langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle \\ &\text{subject to } \mathcal{N}(\boldsymbol{\beta}) \leq 1 \,. \end{aligned}$$

The dual norm is used to encapsulate the effective noise of LASSO. More explicitly, the dual norm shows up in the form $\mathcal{N}_\star(X^\top \boldsymbol{U})$, called the *effective noise* of the LASSO in equation (3.7). Recall that the effective noise vector $X^\top \boldsymbol{U}$ can be thought of as the "projection" of the noise vector $\boldsymbol{U}$ on the column space of X, and thus, $\mathcal{N}_\star(X^\top \boldsymbol{U})$ is a measure of the "true" noise present in the data.

In Appendix 7.4, we obtain a generic bound on the dual norm, which will be used to obtain the statistical guarantees of this paper.

4.1. Prediction Error

Here we now give the announced theoretical results on non-asymptotic bounds for the one-step ahead prediction error, first formulated by an oracle inequality (see Theorem 4.2), then in the following subsection including control of the tuning parameter λ by control of the effective noise of our Lasso-optimisation problem. This enables us to formulate concrete rates of the prediction error. Note that this approach delivers an explicit way of how to select L (via equation (4.8), and subsequently (4.22)) the input parameter for Step 2 of Algorithm 1 ("AR Coefficient Estimation Pipeline").

4.1.1. Oracle Prediction Error

The problem (3.7) has a solution because, given any specific realization of the time series (thus, effectively, fixing $\boldsymbol{y}$ and X) and any $\lambda > 0$, the convex function

$$f_\lambda(\boldsymbol{\beta}) = \frac{1}{D}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_{\mathcal{G}} \tag{4.1}$$

is continuous, and

$$\lim_{\|\boldsymbol{\beta}\|_2 \to \infty} f_\lambda(\boldsymbol{\beta}) = \infty \,.$$

This also shows that we can consider this as a convex program on a compact domain, because we should (at least in theory) be able to restrict the domain of minimization to be an ℓ^2-ball of suitable radius, say $R_{r,X,\boldsymbol{y}} \in (0, \infty)$. Let $\hat{\boldsymbol{\beta}}$ be a solution of this convex program.

Now write the time series observations in the matrix form as $\mathcal{X}$. We are interested in an estimate of the 'risk', equivalently, the in-sample, one-step-ahead mean squared forecast error $\mathbb{E}[\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_2^2/D \mid \mathcal{X}]$.

In the derivations below, we follow a well-known approach for its control, as appeared (for example) in Lederer (2021, Chapter 6). We defer the proof to Subsection 7.1.

Theorem 4.2 (Prediction Guarantee). *Suppose that* $\lambda \geq \frac{2}{D}\mathcal{N}_\star(X^\top \boldsymbol{U})$*, where* $\mathcal{N}(\boldsymbol{\alpha}) = \sum_{l=1}^{L} \sqrt{|\mathcal{G}_l|} \cdot \|\boldsymbol{\alpha}_{(\geq l)}\|_2$*. Write* $\overline{\sigma}^2 = D^{-1}(T_1\sigma_1^2 + \cdots + T_M\sigma_M^2)$*; then*

$$\frac{1}{D}\mathbb{E}\left[\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_2^2 \mid \mathcal{X}\right] \leq \overline{\sigma}^2 + \min_{\boldsymbol{\alpha}\in\mathbb{R}^{ML}} \left(\frac{1}{D}\|X(\boldsymbol{\beta}-\boldsymbol{\alpha})\|_2^2 + 2\lambda\mathcal{N}(\boldsymbol{\alpha})\right),$$

In particular, the following inequality holds:

$$\mathbb{E}\left[\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_2^2 \mid \mathcal{X}\right] - \mathbb{E}\left[\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 \mid \mathcal{X}\right] \leq \min_{\boldsymbol{\alpha}\in\mathbb{R}^{ML}} \left(\|X(\boldsymbol{\beta}-\boldsymbol{\alpha})\|_2^2 + 2D\lambda\mathcal{N}(\boldsymbol{\alpha})\right).$$

If, moreover, $\lambda \geq \frac{4}{D}\mathcal{N}_\star(X^\top \boldsymbol{U})$*, then*

$$\frac{1}{D}\mathbb{E}\left[\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_2^2 \mid \mathcal{X}\right] \leq \overline{\sigma}^2 + \frac{\lambda}{2}\min\left\{3\mathcal{N}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}), 3\mathcal{N}(\boldsymbol{\beta}) - \mathcal{N}(\hat{\boldsymbol{\beta}})\right\}. \tag{4.2}$$

Consequently,

$$\frac{1}{D}\mathbb{E}\left[\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_2^2 \mid \mathcal{X}\right] \leq \overline{\sigma}^2 + 2\lambda \sum_{\ell=1}^{L_0} \sqrt{|\mathcal{G}_\ell|} \cdot \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_{\mathcal{G}_\ell}. \tag{4.3}$$

This result is in the form of standard oracle inequalities in high-dimensional statistics Lederer (2021, Chapter 6). It shows that the estimator minimizes the one-step-ahead mean-squared forecast risk up to a complexity term that is linear in the tuning parameter and the model complexity. Hence, the above bound yields an upper bound for the rate convergence once we can control the tuning parameter λ via an upper bound on the effective noise.

To proceed, we need the following bound on the dual of the overlapping hierarchical group-norm $\mathcal{N}$; its proof appears in Subsection 7.4.

Proposition 4.3 (L^∞ norm bounds dual norm, general case). *Suppose* $\boldsymbol{\alpha} \in \mathbb{R}^{ML}$*, and*

$$\mathcal{N}(\boldsymbol{\alpha}) = \sum_{l=1}^{L} \sqrt{M(L-l+1)}\|\boldsymbol{\alpha}_{(\geq l)}\|_2,$$

$$\textit{where} \qquad \|\boldsymbol{\alpha}_{(\geq l)}\|_2 := \sqrt{\sum_{j=l}^{L}\sum_{m=1}^{M} \alpha_{(m-1)+j}^2}.$$

Then $\mathcal{N}_\star(\boldsymbol{\alpha}) \leq L^{-\frac{1}{2}}\|\boldsymbol{\alpha}\|_\infty$.

Thanks to Theorem 4.2 and Proposition 4.3, in order to find the smallest tuning parameter λ fulfilling $\lambda \geq 4\mathcal{N}_\star(X^\top \boldsymbol{U})/D$ it suffices to derive a high-probability upper-bound on $D^{-1}L^{-\frac{1}{2}}\|X^\top \boldsymbol{U}\|_\infty$ (which is precisely the bound on the dual norm of $X^\top U$).

4.1.2. Control of the Effective Noise for bounding the tuning parameter

We can prove the following high-probability bound on the tuning parameter (equivalently, on the effective noise). Again, its proof appears in Subsection 7.2. Note that this is a major inequality, and while the arguments are well-known, we have applied those arguments to the case where the sparsity enforcing regularizer is induced by the overlapping group norm.

Theorem 4.4 (Bound on the Effective Noise). *Let $C_\sharp := T_{\max}/T_{\min}$. For any $\eta > 0$ satisfying*

$$\eta \geq \frac{8C_\sharp \sigma_{\max}^2 (1 + \epsilon^{-2} + \epsilon^{-4})}{M}, \tag{4.4}$$

and any $\delta > 0$, if $D = T_1 + \cdots + T_m$ satisfies

$$D \geq \frac{8\sigma_{\max}^2 (1 + \epsilon^{-2} + \epsilon^{-4})}{c_0 \eta} \log\left(\frac{ML}{\delta}\right), \tag{4.5}$$

then the following inequality holds:

$$\mathbb{P}\left[\frac{2}{D}\mathcal{N}_\star(X^\top \boldsymbol{U}) \geq \frac{3\eta}{2\sqrt{L}}\right] \leq \delta. \tag{4.6}$$

Here $c_0 > 0$ is the absolute constant from the Gaussian concentration inequality (see Proposition 7.4).

The interpretation of the above result is that for the LASSO oracle inequality (Theorem 4.2) to hold with high probability, it is sufficient to choose λ to be just as large as $(3\eta)/(2\sqrt{L})$, but it is not necessary to take it larger. Indeed, equation (4.6) shows that the probability that $2/D\mathcal{N}_\star(X^T U)$ is larger than $3\eta/2\sqrt{L}$ is small; thus, we may assume $(2/D)\mathcal{N}_\star(X^T U) < 3\eta/2\sqrt{L}$, to hold with probability $1 - \delta$.

The above result bounds the tails of the effective noise. For any $A \geq 1$, we will now set

$$\eta = C_0 \sqrt{A} C_\epsilon C_\sharp^{\frac{3}{2}} \sigma_{\max}^2 \sqrt{\frac{L \log\left(\frac{ML}{\delta}\right)}{DM}},$$

for some absolute constant $C_0 > 0$ and parameter $C_\epsilon > 0$ that depends only on the stability parameter ϵ. Henceforth, we take

$$\eta := 8(84Ae)^{\frac{1}{2}} \zeta^{-1} (1 + \epsilon^{-2} + \epsilon^{-4}) C_\sharp^{\frac{3}{2}} \sigma_{\max}^2 \sqrt{\frac{L \log\left(\frac{ML}{\delta}\right)}{DM}}, \tag{4.7}$$

which is obtained by plugging-in values of C_0 and C_ϵ — as in the proof of Lemma 4.5 below — into Equation (4.7); we do not attempt to optimize these constants.

Lemma 4.5 (Data-dependent selection of L). *There is an absolute constant $C \in (0, \infty)$ and a parameter $C_\epsilon > 0$ that depends only on the stability parameter $\epsilon > 0$ such that if η is as in* (4.7) *and*

$$T_{\min} \leq 84Ae\zeta^{-2} L \log\left(\frac{ML}{\delta}\right), \tag{4.8}$$

then the inequality (4.4) *holds.*

This results shows that there is a suitable η for our theories to hold. The question of finding such an η in practice will need to be discussed in more applied future work.

Proof. We set $C_0 := (12)^3\sqrt{84e}$, and

$$C_\epsilon := \epsilon^{-4}(1+\epsilon^{-2}+\epsilon^{-4})\,.$$

Note that

$$C_\sharp T_{\min} = T_{\max} \geq \frac{D}{M}\,. \tag{4.9}$$

Now, with η as in (4.7), the inequality (4.4) is ensured by the following sequence of inequalities:

$$\begin{aligned}
& (84Ae)^{\frac{1}{2}}\zeta^{-1}\sqrt{L\log\left(\tfrac{ML}{\delta}\right)} && \geq \sqrt{T_{\min}} \\
\Rightarrow \quad & (84Ae)^{\frac{1}{2}}\zeta^{-1}C_\sharp^{\frac{1}{2}}\sqrt{L\log\left(\tfrac{ML}{\delta}\right)} && \geq \sqrt{\frac{D}{M}} \\
\Rightarrow \quad & 8(84Ae)^{\frac{1}{2}}\zeta^{-1}(1+\epsilon^{-2}+\epsilon^{-4})C_\sharp^{\frac{3}{2}}\sigma_{\max}^2\sqrt{\tfrac{L\log\left(\frac{ML}{\delta}\right)}{DM}} && \geq \frac{8C_\sharp\sigma_{\max}^2(1+\epsilon^{-2}+\epsilon^{-4})}{M}\,.
\end{aligned}$$

□

The inequality (4.8) provides for the theoretical support of our choice of L in formulating the penalized least squares program in (3.7). The explicit nature of the parameters in the proof above is crucial here, in order to satisfy both (4.4) and (4.5) above, as well as remaining compatible with the requirements involving the restricted eigenvalue property (as in Corollary 4.8) below. Note that the current LASSO-based literature often ignores combining the requirements coming from standard oracle inequality type result (Theorem 4.2) and the restricted eigenvalue type results.

Remark (B). *On choosing L via equation* (4.8) The use of such an upper-bound L conforms with the (by now) well-known restricted isometry property of sub-sampled Gaussian matrices (that is, matrices whose entries are iid Gaussian) in the compressed-sensing literature; for more details, see (for example) Candes and Tao (2006, section 1.E) and the references therein.

For D as in (4.5), this gets us the following bound:

$$\mathbb{P}\left[\mathcal{N}_\star\left(\frac{2}{D}X^\top \boldsymbol{U}\right) > 12(84Ae)^{\frac{1}{2}}\zeta^{-1}\sigma_{\max}^2 C_\sharp{}^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}}\right] \leq \delta,$$

In view of Theorem 4.4, we correspondingly set

$$\lambda = 24(84Ae)^{\frac{1}{2}}\zeta^{-1}\sigma_{\max}^2 C_\sharp{}^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}} \tag{4.10}$$

to force (4.2) to be true.

Remark (C). From the above, we note that λ decreases when any of M, D increases (as the function $x^{-1}\log x \to 0$ when $x \to \infty$). This is intuitive since larger M requires that LASSO must not set too many coefficients to 0, and larger the D better the non-regularized estimator is as approximation of the true coefficients.

Together with Theorem 4.2, the above yields the following bound:

Corollary 4.6 (Concrete rates of Prediction Error). *Suppose that*

$$\eta = 8(84Ae)^{\frac{1}{2}}\zeta^{-1}(1+\epsilon^{-2}+\epsilon^{-4})C_{\sharp}^{\frac{3}{2}}\sigma_{\max}^2\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}},$$

as set in Equation (4.7) *above, and*

$$D \geq \frac{8\sigma_{\max}^2(1+\epsilon^{-2}+\epsilon^{-4})}{c_0\eta}\log\left(\frac{ML}{\delta}\right).$$

Then, the following inequality holds with probability at least $1-\delta$*:*

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{D}\|\boldsymbol{y}-X\hat{\boldsymbol{\beta}}\|_2^2 \mid \mathcal{X}\right] - \overline{\sigma}^2\\ \leq &\min_{\boldsymbol{\alpha}\in\mathbb{R}^{ML}}\left(\frac{1}{D}\|X(\boldsymbol{\beta}-\boldsymbol{\alpha})\|_2^2 + \frac{24}{\zeta}(84Ae)^{\frac{1}{2}}\sigma_{\max}^2 C_{\sharp}^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}}\mathcal{N}(\boldsymbol{\alpha})\right).\end{aligned}$$

4.2. Estimating error of parameters, false discoveries

We now deliver the treatment of assertions 1. and 2. on estimation error and support (via control of false discoveries) of our Main Theorem 3.1. For this we need to cope with the *Restricted Eigenvalue Property* as it typically arises in LASSO analysis (see Basu and Michailidis (2015)).

4.2.1. Control of the restricted eigenvalue property

Our prediction guarantees stated so far did not impose restrictions on the minimum number of samples from each component. For estimation and stability guarantees, however, we need to demand a minimum number of those samples. In particular, we will need the following proposition, on the restricted eigenvalue property of the Gram matrix $X^\top X$. The proof presented in Subsection 7.3 follows the same lines of arguments as in Basu and Michailidis (2015). Because of the non-uniform nature of the block dimensions of the data matrix in (3.6), as well as the individual weights assigned to the components of the coefficient $\boldsymbol{\beta}$, we can only hope for a block-wise result as stated below.

The following proposition delivers a lower bound on the error-expression $\|X_m\boldsymbol{v}_m\|_2^2$ with $\boldsymbol{v}_m$ the $m-$th component of $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}$.

Proposition 4.7 (Bound on the Restricted Eigenvalue). *For any parameter* $\delta > 0$ *and all vectors* $\boldsymbol{v} = (\boldsymbol{v}_1^\top, \ldots, \boldsymbol{v}_M^\top)^\top \in (\mathbb{R}^L)^M$*, the following inequality holds with* $\zeta := \frac{\epsilon^4}{216}$ *and* s_m *even positive integers:*

$$\mathbb{P}\left[\forall\, m \in [M] \inf_{\boldsymbol{v}_m \in \mathbb{R}^L} \left(\|X_m \boldsymbol{v}_m\|_2^2 - \frac{T_m \sigma_m^2 \epsilon^2}{2} \left(\|\boldsymbol{v}_m\|_2^2 - \frac{2}{s_m} \|\boldsymbol{v}_m\|_1^2 \right) \right) \geq 0 \right]$$

$$\geq 1 - 2 \sum_{m=1}^{M} \exp\left(-\frac{T_m \min\{\zeta, \zeta^2\}}{2} + s_m \min\{\log L, \log(21 e L / s_m)\} \right). \tag{4.11}$$

As mentioned above, a proof of Proposition 4.7 appears in Subsection 7.3.

Setting

$$s_m := 2 \left\lfloor \frac{T_m \zeta^2}{8 \log L} \right\rfloor, \tag{4.12}$$

we get the following immediate corollary to Proposition 4.7:

Corollary 4.8 (Bound on the Restricted Eigenvalue for specific s_m). *If*

$$T_{\min} \geq 84 e \zeta^{-2} \log L\,, \tag{4.13}$$

where $\zeta = 6^{-3}\epsilon^4$, then the following holds:

$$\mathbb{P}\left[\forall m \in [M] \inf_{\boldsymbol{v}_m \in \mathbb{R}^L} \left(\|X_m \boldsymbol{v}_m\|_2^2 - \frac{T_m \sigma_m^2 \epsilon^2}{2} \|\boldsymbol{v}_m\|_2^2 \left(1 - \frac{8L \log L}{T_m \zeta^2} \right) \right) \geq 0 \right]$$

$$\geq\ 1 - 2 \sum_{m=1}^{M} e^{-\frac{T_m \zeta^2}{4}}\,. \tag{4.14}$$

Proof. We use $\|\boldsymbol{v}_m\|_1^2 \leq L \|\boldsymbol{v}_m\|_2^2$ in Proposition 4.7. □

4.2.2. Bounds on estimation error of the autoregressive coefficients, and false discovery

This section contains the bound on the error of estimation of the autoregressive coefficients, and the false discovery.

The following is a compact notation used in the proof of the theorem below (also in the proof of Theorem 4.2).

Definition 4.9 (Notation).

$$\mathcal{N}_{\leq L_0}(\boldsymbol{v}) := \sum_{\ell=1}^{L_0} \sqrt{|\mathcal{G}_\ell|} \cdot \|\boldsymbol{v}\|_{\mathcal{G}_\ell}\,. \tag{4.15}$$

The theorem below bounds the ℓ^2-error in estimating the autoregressive coefficients $\boldsymbol{\beta}$, as well as the size of the set of false positives, using the penalized LASSO formulation as in (3.7).

Theorem 4.10 (Bounds on the Estimation Error and False Discovery). *Let $\hat{\boldsymbol{\beta}}$ be the solution to the group-regularized LASSO in* (3.7) *with λ as in Equation* (4.10) *above — namely,*

$$\lambda = 24(84Ae)^{\frac{1}{2}}\zeta^{-1}\sigma_{\max}^2 C_{\sharp}^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}}$$

where $\zeta = 6^{-3}\epsilon^4$, which correspond to η as in Equation (4.7)*; if*

$$\alpha := \min_{m\in[M]} \frac{T_m\sigma_m^2}{D}, \tag{4.16}$$

and

$$D \geq \frac{8\sigma_{\max}^2(1+\epsilon^{-2}+\epsilon^{-4})}{c_0\eta}\log\left(\frac{ML}{\delta}\right),$$

then, for any parameter $\delta > 0$, the inequality

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2 \leq \frac{81(84Ae)^{\frac{1}{2}}LL_0\sigma_{\max}^2 C_{\sharp}^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})}{\zeta\alpha\epsilon^2}\sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{D}} \tag{4.17}$$

holds with probability at least

$$(1-\delta)\left(1-2\sum_{m=1}^{M} L^{-\frac{T_m}{4\log L}}\right), \tag{4.18}$$

provided $T_{\min} \geq 84eA\zeta^{-2}L_0\log L$ with $\zeta = 6^{-3}\epsilon^4$.

Moreover, introducing the notation $S_\lambda := \{j : |\hat{\boldsymbol{\beta}}_j| > \lambda\}$, then the false discovery is bounded by the following inequality holding with probability as in (4.18) *above:*

$$|S_\lambda \setminus \mathrm{supp}(\boldsymbol{\beta})| \leq \frac{243(84Ae)^{\frac{1}{2}}LL_0^{\frac{3}{2}}\sigma_{\max}^2 C_{\sharp}^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})}{\zeta\alpha\epsilon^2\lambda}\sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{D}}. \tag{4.19}$$

Remark (D). In the above, the term inside the square root decreases asymptotically, and the terms outside may change or stay fixed (depending on how the number of samples for each component increases).

Proof of Theorem 4.10. For ease of notation, write

$$\boldsymbol{v} := \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}.$$

By Proposition 4.7 above (which can be applied since $T_{\min} \geq 84e\zeta^{-2}\log L$), one has

$$\begin{aligned}\frac{1}{D}\|X\boldsymbol{v}\|_2^2 &\geq \frac{1}{D}\sum_{m=1}^{M}\|X_m\boldsymbol{v}_m\|_2^2 \\ &\geq \frac{\epsilon^2}{2}\sum_{m=1}^{M}\frac{\sigma_m^2 T_m}{D}\|\boldsymbol{v}_m\|_2^2\left(1-\frac{8L\log L}{T_m\zeta^2}\right)\end{aligned}$$

$$> \frac{4\alpha\epsilon^2}{9}\|\boldsymbol{v}\|_2^2 .$$

To this, we now apply inequality (7.4), that is,

$$\frac{1}{D}\|X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\|_2^2 \leq 3\lambda \mathcal{N}_{\leq L_0}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}),$$

which yields

$$\begin{aligned} \|\boldsymbol{v}\|_2^2 &\leq \frac{27\lambda}{8\alpha\epsilon^2}\mathcal{N}_{\leq L_0}(\boldsymbol{v}) \\ &\leq \frac{27\lambda L_0\sqrt{ML}}{8\alpha\epsilon^2}\|\boldsymbol{v}\|_2 \\ \Rightarrow \qquad \|\boldsymbol{v}\|_2 &\leq \frac{27\lambda L_0\sqrt{ML}}{8\alpha\epsilon^2}. \end{aligned}$$

With the choice of λ as in equation (4.10), namely,

$$\lambda = 24(84Ae)^{\frac{1}{2}}\zeta^{-1}\sigma_{\max}^2 C_\sharp^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}},$$

it now follows that

$$\|\boldsymbol{v}\|_2 \leq \frac{81(84Ae)^{\frac{1}{2}}LL_0\sigma_{\max}^2 C_\sharp^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})}{\zeta\alpha\epsilon^2}\sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{D}},$$

as claimed. To get the bound on the false discovery, write $S := \mathrm{supp}(\boldsymbol{\beta})$; the bound on the false discovery can be obtained as follows:

$$\begin{aligned} |S_\lambda \setminus S| &= \sum_{j\in[ML]\setminus S} 1_{|\hat{\boldsymbol{\beta}}_j|>\lambda}(\hat{\boldsymbol{\beta}}) \\ &= \sum_{j\in[ML]\setminus S} 1_{|\hat{\boldsymbol{v}}_j|>\lambda}(\hat{\boldsymbol{v}}) \\ &\leq \frac{1}{\lambda}\sum_{j\in[ML]\setminus S}|\hat{\boldsymbol{v}}_j| \\ \text{by Remark (F) in Subsection (7.1)} \qquad &\leq \frac{3}{\lambda}\sum_{j\in S}|\hat{\boldsymbol{v}}_j| \\ &\leq \frac{3\sqrt{L_0}}{\lambda}\|\hat{\boldsymbol{v}}\|_2 \end{aligned}$$

which yields the bound in (4.19) by inequality (4.17). □

Remark (E). For the explicit choice of λ as in equation (4.10), namely,

$$\lambda = 24(84Ae)^{\frac{1}{2}}\zeta^{-1}\sigma_{\max}^2 C_\sharp^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})\sqrt{\frac{L\log\left(\frac{ML}{\delta}\right)}{DM}}$$

with $\zeta = 6^{-3}\epsilon^4$, the inequality (4.19) becomes

$$|S_\lambda \setminus \mathrm{supp}(\boldsymbol{\beta})| \leq \frac{81(ML)^{\frac{1}{2}} L_0^{\frac{3}{2}}}{8\alpha\epsilon^2} . \tag{4.20}$$

4.3. Stability of the estimated AR model

We finally prove that the coefficients estimated as per the overlapping-group-LASSO in (3.7), with λ as in (4.10), lie in the region for stability of univariate lag-L autoregressive processes, even when the number of post-samples is 'not too large'. More specifically, we have the following theorem.

Theorem 4.11 (Stability Guarantee). *Let $A \geq 1$ be a parameter and let $\hat{\boldsymbol{\beta}}$ be the output of the group-LASSO in* (3.7) *above, using D post-samples, where*

$$D \geq 3^9 \cdot (84Ae)C_\sharp^5(1+\epsilon^{-2}+\epsilon^{-4})^2 \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^4 (\epsilon^3\zeta)^{-2} M^a L_0^2 L^3 \log\left(\frac{ML}{\delta}\right)\log(2L) . \tag{4.21}$$

If $T_{\min} \geq 84eA\zeta^{-2}L_0 \log L$, and L satisfies

$$n_{\min} = L + 84Ae\zeta^{-2}L\log\left(\frac{ML}{\delta}\right), \tag{4.22}$$

with $\zeta = 6^{-3}\epsilon^4$, then the AR-models — fitted with coefficients $\hat{\boldsymbol{\beta}}_0$ returned by Algorithm 1 ("AR Coefficient Estimation Pipeline") — are stable, with probability as in (4.18) *— in the following scenarios:*

1. *when all the time series are different realizations of a unique underlying stochastic process, $a \geq 1$, and*
$$\hat{\boldsymbol{\beta}}_0 := \frac{1}{M}\sum_{m=1}^{M} \hat{\boldsymbol{\beta}}_m ;$$
2. *when all the time series are realizations of different underlying stochastic processes (equivalently, all the $\boldsymbol{\beta}_m$'s are different), $a \geq 2$, and $\hat{\boldsymbol{\beta}}_0 := \hat{\boldsymbol{\beta}}$.*

The lower bound on $T_{\min}$ aligns with the upper bound in Equation (4.8) as $L \geq L_0$. The stated value of $n_{\min}$ chooses the smallest value of L to satisfy both the upper and lower bound on $T_{\min}$.

Proof of Theorem 4.11. 1. We start with the first case. The idea of the proof is as follows. All the component $\hat{\boldsymbol{\beta}}_m$'s of $\hat{\boldsymbol{\beta}}$ are approximations of the same underlying $\boldsymbol{\beta}_0 := \boldsymbol{\beta}_m$ for all $m \in [M]$. Therefore, by the bound in Proposition 4.10 above and by convexity of the square function, their mean must be ℓ^2-close to $\boldsymbol{\beta}_0$, with $D = O(ML_0^2L^3\log(ML)\log(2L))$ many samples. Moreover, the $0.5L^{-1}\epsilon$ perturbation of the coefficients preserves stability of the ϵ-stable process.

More explicitly, we have

$$\left\| \left(\sum_{m=1}^{M} \frac{1}{M}\hat{\boldsymbol{\beta}}_m\right) - \boldsymbol{\beta}_0 \right\|_2^2 = \left\| \frac{1}{M}\sum_{m=1}^{M} (\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m) \right\|_2^2$$

$$\text{convexity} \Rightarrow \qquad \leq \frac{1}{M}\sum_{m=1}^{M} \|\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m\|_2^2$$

$$= \frac{1}{M}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2,$$

and by Theorem 4.10, this yields

$$\left\| \left(\sum_{m=1}^{M} \frac{1}{M} \hat{\boldsymbol{\beta}}_m \right) - \boldsymbol{\beta} \right\|_2 \leq \frac{1}{\sqrt{M}} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$$

$$\leq \frac{81(84Ae)^{\frac{1}{2}} L L_0 \sigma_{\max}^2 C_\sharp^{\frac{3}{2}} (1 + \epsilon^{-2} + \epsilon^{-4})}{\zeta \alpha \epsilon^2} \sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{MD}} \tag{4.23}$$

with high probability. Note that

$$\alpha = \min_{m \in [M]} \frac{T_m \sigma_m^2}{D} \geq \frac{T_{\min} \sigma_{\min}^2}{D} \geq \frac{T_{\min} \sigma_{\min}^2}{M T_{\max}} \geq \frac{\sigma_{\min}^2}{M C_\sharp} \,. \tag{4.24}$$

When

$$D \geq 3^9 \cdot (84Ae) C_\sharp^5 (1 + \epsilon^{-2} + \epsilon^{-4})^2 \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^4 (\epsilon^3 \zeta)^{-2} M L_0^2 L^3 \log\left(\frac{ML}{\delta}\right) \log(2L) \,,$$

this yields

$$\left\| \left(\sum_{m=1}^{M} \frac{1}{M} \hat{\boldsymbol{\beta}}_m \right) - \boldsymbol{\beta} \right\|_2 \leq \frac{81(84Ae)^{\frac{1}{2}} L L_0 \sigma_{\max}^2 C_\sharp^{\frac{3}{2}} (1 + \epsilon^{-2} + \epsilon^{-4})}{\zeta \alpha \epsilon^2} \sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{MD}}$$

$$\leq \frac{\epsilon \sigma_{\min}^2}{\sqrt{3} C_\sharp M \alpha} \sqrt{\frac{1}{L \log(2L)}}$$

$$(4.24) \Rightarrow \qquad < \frac{\epsilon}{\sqrt{3}} \sqrt{\frac{1}{L \log(2L)}} \,.$$

Since $\epsilon \in (0, 1)$, the stability follows by the triangle inequality. More explicitly, if $\hat{\boldsymbol{\beta}}_0 := M^{-1}(\hat{\boldsymbol{\beta}}_1 + \cdots + \hat{\boldsymbol{\beta}}_M)$, then for every $z \in \mathbb{D}$, we have :

$$\begin{aligned}
|f_{\hat{\boldsymbol{\beta}}_0}(z)| &= |1 - \hat{\boldsymbol{\beta}}_0 \cdot (z, z^2, \cdots, z^L)| \\
&= |1 - \boldsymbol{\beta}_0 \cdot (z, z^2, \cdots, z^L) - (\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0) \cdot (z, z^2, \cdots, z^L)| \\
&\geq |f_{\boldsymbol{\beta}_0}(z)| - \|\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\|_2 \cdot \|(z, z^2, \cdots, z^L)\|_2 \\
&> \epsilon - \frac{\epsilon}{\sqrt{3}} \sqrt{\frac{1}{\log(2L)}} \\
&> 0 \,.
\end{aligned}$$

Evidently, this shows that the autoregressive process with $\hat{\boldsymbol{\beta}}_0$ coefficients is stable.

2. The arguments are similar in the second case. Since

$$D \geq 3^9 \cdot (84Ae)C_\sharp^5(1+\epsilon^{-2}+\epsilon^{-4})^2\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^4(\epsilon^3\zeta)^{-2}M^2L_0^2L^3\log\left(\frac{ML}{\delta}\right)\log(2L)\,,$$

by Theorem 4.10, for any $m \in [M]$ we have

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m\|_2 &\leq \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \\ &\leq \frac{81(84Ae)^{\frac{1}{2}}LL_0\sigma_{\max}^2C_\sharp^{\frac{3}{2}}(1+\epsilon^{-2}+\epsilon^{-4})}{\zeta\alpha\epsilon^2}\sqrt{\frac{\log\left(\frac{ML}{\delta}\right)}{D}} \\ (4.24) \;\Rightarrow \qquad &< \frac{\epsilon}{\sqrt{3}}\sqrt{\frac{1}{L\log(2L)}}\,.\end{aligned}$$

As in the first case, this implies stability for each of the components.

□

5. Algorithmic Aspects

In this section we briefly introduce the main gazette of our algorithm — namely, the proximal operator — to understand the procedure of solving LASSO as was done in Nicholson et al. (2020). To sketch the outline of the standard procedure for solving regularized LASSO penalized with an overlapping group-norm, we start with the following definition (see Mairal et al. (2011), for example).

Definition 5.1 (Proximal operator). Given a norm $\mathcal{N}$ on $\mathbb{R}^{ML}$, and a tuning parameter λ, the associated proximal operator $\mathrm{Prox}_{\mathcal{N},\lambda}$ is defined for every $\boldsymbol{\alpha} \in \mathbb{R}^{ML}$ as the optimum value of the following convex problem:

$$\underset{\mathcal{N},\lambda}{\mathrm{Prox}}(\boldsymbol{\alpha}) := \underset{\boldsymbol{\beta}}{\mathrm{argmin}}\left\{\frac{1}{2}\|\boldsymbol{\beta}-\boldsymbol{\alpha}\|_2^2 + \lambda\mathcal{N}(\boldsymbol{\beta})\right\}.$$

As established in Zhao, Rocha and Yu (2009), the proximal operator $\mathrm{Prox}_{\mathcal{N},\lambda}$ — for $\mathcal{N}$ as defined in (3.4) — is the composition

$$\underset{\mathcal{G}_1,\lambda}{\mathrm{Prox}} \circ \cdots \circ \underset{\mathcal{G}_L,\lambda}{\mathrm{Prox}}$$

of the proximal operators for the individual groups and can be computed inductively — starting from $\mathrm{Prox}_{\mathcal{G}_L,\lambda}$, which is the well-known soft-thresholding — in $O(ML)$ computational steps. By Combettes and Wajs (2005, Proposition 3.1(iii)b), the solutions to the regularized LASSO problem in (3.7) are precisely the fixed points of the operator

$$\boldsymbol{\beta} \longmapsto \underset{\mathcal{N},\lambda}{\mathrm{Prox}}\left(\boldsymbol{\beta} + \frac{2\lambda}{D}\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right). \tag{5.1}$$

Here we use an appropriate λ (as given in (4.10) above). As Nicholson et al. (2020) observed, the proximal operator can be evaluated via duality. The proximal gradient method of Mairal et al. (2011) then finds the fixed point (which exists and is unique by convexity of (3.7)) of this proximal operator. For pseudo-code of this procedure, see Nicholson et al. (2020), where an accelerated version of the proximal descent method was employed for achieving quadratic convergence rate.

6. Conclusion

We have established a set-up (see Equation (3.7) and the discussion preceding this equation) in which LASSO — regularized with a hierarchical group norm — can be used to derive statistical guarantees in terms of the one-step ahead prediction error (Theorem 4.2) in the realms of multiple ϵ-stable (Definition (2.6)) univariate autoregressive processes of different lengths but identical true lag L_0. The results presented here assume no prior knowledge of the true lag (or any upper-bound of the true lag); in fact, we show that the sample size itself suggests a certain upper-bound $\hat{L}$ of the true lag to be used for the group-LASSO. Given an appropriately large sample size, this $\hat{L}$ will indeed be an upper-bound of the true lag L_0, hence producing, via thresholding, a convenient estimator $\hat{L}_0$ of the true lag. Moreover, this $\hat{L}$ will be of the order that is required for our theoretical guarantees to hold. We proved that the group-LASSO formulated with a suitable tuning parameter λ estimates the AR coefficients with an arbitrarily high degree of accuracy. We also showed the support of the estimated coefficient-set approximately matches the support of the original parameters (Theorem 4.10). Finally, we proved that the fitted models with coefficients as estimated by the group-LASSO are ϵ-stable (Theorem 4.11), a property that is known in the literature solely for Yule-Walker estimates of the parameters of univariate autoregressive processes.

Open Questions and Future Directions

From a theoretical perspective, it will be interesting to investigate adaptations of the group-LASSO method to the case of multiple decoupled AR processes with multivariate components. We expect that this will require, among others, (1) additional techniques to deal with the group-norm, and (2) integrating the stability issues and the restricted eigenvalue issues; these will be technically far more demanding in the multivariate components settings. On the practical front, the most important question is to get a better hold on the tuning parameter λ (see Equation (4.10)), which requires better constant/parameters than, for example, those appearing in the proof of Lemma 4.5. A better strategic control on these will enable a more realistic initial estimate of $\hat{L}$ — to be used by the group-LASSO as an upper-bound on the true lag L_0, and thereby producing a more efficient estimator $\hat{L}_0$.

A further question is to see if our approach can be extended to deal with multiple AR process of uneven lengths, where the corss-sectional innovations are correlated. In the current state, an immediate obstruction is via the restricted eigenvalue property of the data matrix — which would be much more difficult to analyze if the correlation has to be taken into account. Note that, for multiple AR time series of equal length, we can follow the lag selection strategy here and embed the learning problem as a diagonal VAR. Established work such as Basu and Michailidis (2015) will be applicable in deriving the statistical guarantees, and Mairal et al. (2011)'s algorithm can be applied to learn the AR-coefficients. Treating the unequal length cases would however require new ideas and techniques.

In a different track, it will be interesting to see if LASSO-based lag selection can be used to deal with other time series models, such as factor models, non-linear high-dimensional time series - with dependence across coordinates, GARCH-type time series, etc.

Funding

S.C. and J.L. acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG) under grant number 451920280. S.C. would like to express his gratitude to J.L. and Ruhr-Universität Bochum — having been invited on a visit during March - May of 2022, with a 'Research Assistantship Fellowship

for Early Postdocs' from Ruhr-Universität Bochum Research School by means of the German Academic Exchange Service (DAAD) STIBET funds. S.C. would like to express his gratitude to RvS and the ISBA/LIDAM of UCLouvain — having been invited on a research visit in November 2022.

7. Technical results

7.1. Proof of Theorem 4.2

Proof. Denote the data by $\mathcal{X}$. Define the random vector $\boldsymbol{U} \in \mathbb{R}^D$ as follows:

$$\boldsymbol{U} \coloneqq (U_1^1, \ldots, U_{T_1}^1, U_1^2, \ldots, U_{T_2}^2, \cdots, U_1^M, \ldots, U_{T_M}^M)^\top .$$

We note the following series of inequalities, all of which follow from linearity of expectation: here, we write $\boldsymbol{\beta}$ for the true coefficient vector of the model. One has

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_2^2 \mid X\right] \\
&= \mathbb{E}\left[\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 \mid X\right] + \mathbb{E}\left[\|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2 \mid X\right] \\
&\qquad\qquad +2\mathbb{E}\left[\langle \boldsymbol{y} - X\boldsymbol{\beta}, X\boldsymbol{\beta} - X\hat{\boldsymbol{\beta}}\rangle \mid X\right] \\
&= \mathbb{E}\left[\|\boldsymbol{U}\|_2^2 \mid X\right] + \mathbb{E}\left[\|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2 \mid X\right] \\
&\qquad\qquad +2\mathbb{E}\left[\langle \boldsymbol{U}, X\boldsymbol{\beta} - X\hat{\boldsymbol{\beta}}\rangle \mid X\right] \\
&= T_1\sigma_1^2 + \cdots + T_M\sigma_M^2 + \mathbb{E}\left[\|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2 \mid X\right] + 2\langle \mathbb{E}\left[\boldsymbol{U} \mid X\right], X\boldsymbol{\beta} - X\hat{\boldsymbol{\beta}}\rangle \\
&= T_1\sigma_1^2 + \cdots + T_M\sigma_M^2 + \|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2,
\end{aligned}$$

where the last equality follows from $\mathbb{E}\left[\boldsymbol{U} \mid X\right] = \boldsymbol{U}$, and $\boldsymbol{U} \perp (X\boldsymbol{\beta} - X\hat{\boldsymbol{\beta}})$. Therefore, we need to derive an estimate of the error $\|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2$. We notice that

$$\|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2 = \sum_{m=1}^{M}\sum_{t=1}^{T_m}\left(\sum_{l=1}^{L}(\hat{\beta}_l^m - \boldsymbol{\beta}_l^m)\, x_{t-l}^m\right)^2,$$

where $\{\boldsymbol{\beta}_l^m\}$ denotes the true-parameters of the models. Using the fact that $\hat{\boldsymbol{\beta}}$ is a minimizer of

$$\frac{1}{D}\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + \lambda\sum_{l=1}^{L}\mathcal{N}(\boldsymbol{\alpha}),$$

over $\boldsymbol{\alpha} \in \mathbb{R}^{ML}$, one derives

$$\begin{aligned}
&\tfrac{1}{D}\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_2^2 + \lambda\mathcal{N}(\hat{\boldsymbol{\beta}}) \le \frac{1}{D}\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + \lambda\mathcal{N}(\boldsymbol{\alpha}) \\
\Rightarrow \quad &\tfrac{1}{D}\|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_2^2 \le \frac{1}{D}\|X(\boldsymbol{\beta} - \boldsymbol{\alpha})\|_2^2 + \frac{2}{D}\langle \boldsymbol{U}, X(\hat{\boldsymbol{\beta}} - \boldsymbol{\alpha})\rangle + \lambda\mathcal{N}(\boldsymbol{\alpha}) - \lambda\mathcal{N}(\hat{\boldsymbol{\beta}}) \\
&\qquad\qquad \le \frac{1}{D}\|X(\boldsymbol{\beta} - \boldsymbol{\alpha})\|_2^2 + \frac{2}{D}\langle X^\top\boldsymbol{U}, \hat{\boldsymbol{\beta}} - \boldsymbol{\alpha}\rangle + \lambda\mathcal{N}(\boldsymbol{\alpha}) - \lambda\mathcal{N}(\hat{\boldsymbol{\beta}})
\end{aligned}$$

for all $\boldsymbol{\alpha} \in \mathbb{R}^{ML}$. By the Cauchy-Schwarz inequality, this implies

$$\begin{aligned}\frac{1}{D}\|X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\|_2^2 &\le \frac{1}{D}\|X(\boldsymbol{\beta}-\boldsymbol{\alpha})\|_2^2 + \frac{2}{D}\mathcal{N}_\star(X^\top \boldsymbol{U})\mathcal{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\alpha}) + \lambda(\mathcal{N}(\boldsymbol{\alpha}) - \mathcal{N}(\hat{\boldsymbol{\beta}})) \\ &\le \frac{1}{D}\|X(\boldsymbol{\beta}-\boldsymbol{\alpha})\|_2^2 + \lambda\mathcal{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\alpha}) + \lambda(\mathcal{N}(\boldsymbol{\alpha}) - \mathcal{N}(\hat{\boldsymbol{\beta}}))\end{aligned}$$

since $\lambda \ge \frac{2}{D}\mathcal{N}_\star(X^\top \boldsymbol{U})$. Finally, the triangle inequality $\mathcal{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\alpha}) \le \mathcal{N}(\hat{\boldsymbol{\beta}}) + \mathcal{N}(\boldsymbol{\alpha})$ establishes the first part of the theorem.

In the following, we will choose a slightly different threshold for the tuning parameter λ, as proposed in the statement of theorem. Let $\boldsymbol{\nu} := \boldsymbol{\beta} + \hat{\boldsymbol{\beta}}$, and notice that the true lag L_0 satisfies

$$L_0 = \max\{\ell \in [L] : \|\boldsymbol{\beta}\|_{\mathcal{G}_\ell} \neq 0\}\,.$$

We have

$$\begin{aligned}\mathcal{N}(\boldsymbol{\beta}) - \mathcal{N}(\hat{\boldsymbol{\beta}}) &= \mathcal{N}(\boldsymbol{\beta}) - \mathcal{N}(\boldsymbol{\beta}+\boldsymbol{\nu}) \\ &= \sum_{\ell=1}^{L_0}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{\beta}\|_{\mathcal{G}_\ell} + \sum_{\ell=L_0+1}^{L}\sqrt{|\mathcal{G}_\ell|}\cdot\|\overset{0}{\cancel{\boldsymbol{\beta}}}\|_{\mathcal{G}_\ell} \\ &\qquad - \sum_{\ell=1}^{L_0}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{\beta}+\boldsymbol{\nu}\|_{\mathcal{G}_\ell} - \sum_{\ell=L_0+1}^{L}\sqrt{|\mathcal{G}_\ell|}\cdot\|\overset{0}{\cancel{\boldsymbol{\beta}}}+\boldsymbol{\nu}\|_{\mathcal{G}_\ell} \\ &= \sum_{\ell=1}^{L_0}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{\beta}\|_{\mathcal{G}_\ell} - \sum_{\ell=1}^{L_0}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{\beta}+\boldsymbol{\nu}\|_{\mathcal{G}_\ell} - \sum_{\ell=L_0+1}^{L}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{\nu}\|_{\mathcal{G}_\ell} \\ &\le \sum_{\ell=1}^{L_0}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{\nu}\|_{\mathcal{G}_\ell} - \sum_{\ell=L_0+1}^{L}\sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{\nu}\|_{\mathcal{G}_\ell}, \qquad (7.1)\end{aligned}$$

where the last step is due to the triangle inequality: $\|\boldsymbol{\beta}+\boldsymbol{\nu}\|_{\mathcal{G}_\ell} \ge \|\boldsymbol{\beta}\|_{\mathcal{G}_\ell} - \|\boldsymbol{\nu}\|_{\mathcal{G}_\ell}$. Suppose that $\lambda \ge \frac{4}{D}\mathcal{N}_\star(X^\top \boldsymbol{U})$. Using the fact that $\hat{\boldsymbol{\beta}}$ is a minimizer of the objective $\frac{1}{D}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2 + \lambda\mathcal{N}(\boldsymbol{\alpha})$, over $\boldsymbol{\alpha} \in \mathbb{R}^{ML}$, one derives

$$\begin{aligned}&\tfrac{1}{D}\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_2^2 + \lambda\mathcal{N}(\hat{\boldsymbol{\beta}}) \le \frac{1}{D}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\mathcal{N}(\boldsymbol{\beta}) \\ \Rightarrow \qquad &\tfrac{1}{D}\|X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\|_2^2 \qquad \le \frac{2}{D}\langle X^\top \boldsymbol{U}, \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rangle + \lambda\mathcal{N}(\boldsymbol{\beta}) - \lambda\mathcal{N}(\hat{\boldsymbol{\beta}})\,.\end{aligned}$$

By the Cauchy-Schwarz inequality, this implies

$$\begin{aligned}\frac{1}{D}\|X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\|_2^2 &\le \frac{2}{D}\mathcal{N}_\star(X^\top \boldsymbol{U})\mathcal{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) + \lambda(\mathcal{N}(\boldsymbol{\beta}) - \mathcal{N}(\hat{\boldsymbol{\beta}})) \\ &\le \frac{\lambda}{2}\mathcal{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) + \lambda(\mathcal{N}(\boldsymbol{\beta}) - \mathcal{N}(\hat{\boldsymbol{\beta}})) \qquad \text{because } \lambda \ge \frac{4}{D}\mathcal{N}_\star(X^\top \boldsymbol{U}) \\ &\le \frac{\lambda}{2}\min\left\{3\mathcal{N}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}), 3\mathcal{N}(\boldsymbol{\beta}) - \mathcal{N}(\hat{\boldsymbol{\beta}})\right\}\,. \qquad (7.2)\end{aligned}$$

The inequalities

$$0 \le \frac{2}{\lambda D}\|X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\|_2^2 \le \mathcal{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) + 2(\mathcal{N}(\boldsymbol{\beta}) - \mathcal{N}(\hat{\boldsymbol{\beta}}))\,,$$

and

$$\mathcal{N}(\boldsymbol{\beta}) - \mathcal{N}(\hat{\boldsymbol{\beta}}) \le \sum_{\ell=1}^{L_0} \sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{v}\|_{\mathcal{G}_\ell} - \sum_{\ell=L_0+1}^{L} \sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{v}\|_{\mathcal{G}_\ell}$$

yield the following:

$$\begin{aligned}
0 &\le \mathcal{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) + 2(\mathcal{N}(\boldsymbol{\beta}) - \mathcal{N}(\hat{\boldsymbol{\beta}})) \\
&\le \mathcal{N}(\hat{\boldsymbol{v}}) + 2\sum_{\ell=1}^{L_0} \sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{v}\|_{\mathcal{G}_\ell} - 2\sum_{\ell=L_0+1}^{L} \sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{v}\|_{\mathcal{G}_\ell} \\
&= 3\sum_{\ell=1}^{L_0} \sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{v}\|_{\mathcal{G}_\ell} - \sum_{\ell=L_0+1}^{L} \sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{v}\|_{\mathcal{G}_\ell}\,.
\end{aligned}$$

That is, we have

$$\sum_{\ell=L_0+1}^{L} \sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_{\mathcal{G}_\ell} \le 3\sum_{\ell=1}^{L_0} \sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_{\mathcal{G}_\ell}\,. \tag{7.3}$$

By equation (7.2), we then have

$$\begin{aligned}
\frac{1}{D}\|X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\|_2^2 &\le 2\lambda\sum_{\ell=1}^{L_0} \sqrt{|\mathcal{G}_\ell|}\cdot\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_{\mathcal{G}_\ell} \qquad (7.4)\\
&= 2\lambda\mathcal{N}_{\le L_0}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\,.
\end{aligned}$$

□

Remark (F). If, instead of $\mathcal{N}$, we take standard ℓ^2-norm in $\mathbb{R}^{ML}$ (which correspond to the largest group in $\mathcal{N}$), the argument leading to (7.3) yields

$$\begin{aligned}
\sum_{\ell=L_0+1}^{L} |\hat{\boldsymbol{\beta}}_\ell^m| &= \sum_{\ell=L_0+1}^{L} |\boldsymbol{\beta}_\ell^m - \hat{\boldsymbol{\beta}}_\ell^m| \\
&\le 3\sum_{\ell=1}^{L_0} |\boldsymbol{\beta}_\ell^m - \hat{\boldsymbol{\beta}}_\ell^m|\,.
\end{aligned}$$

7.2. Proof of Theorem 4.4

Proof. Since Lemma 4.3 implies

$$\mathcal{N}_\star\left(\frac{2}{D}X^\top \boldsymbol{U}\right) \le \frac{1}{\sqrt{L}}\left\|\frac{2}{D}X^\top \boldsymbol{U}\right\|_\infty = \frac{1}{\sqrt{L}}\max_{m\in[M]}\max_{l\in[L]}\left|\frac{2}{D}\sum_{t=1}^{T_m} U_t^m X_{t-l}^m\right|,$$

it suffices, by an union bound argument, to find high-probability upper bound of the inner-product (see equation (3.8) for notations):

$$\tau_{m,l} := 2\langle \boldsymbol{U}^m, X^{m,l}\rangle = 2\left(\sum_{t=1}^{T_m} U_t^m X_{t-l}^m\right). \tag{7.5}$$

We now present standard bounding techniques, as appeared in (a previous version of) Basu and Michailidis (2015), for example. One has

$$\tau_{m,l} = \left\|\boldsymbol{U}^{(m)} + X^{(m,l)}\right\|_2^2 - \left\|\boldsymbol{U}^{(m)}\right\|_2^2 - \left\|X^{(m,l)}\right\|_2^2.$$

Hence, for any $\eta > 0$, an union bound argument, together with mutual independence of the Gaussian random variable U_j^m and the variables $X_{<j}^m$ for each $j \in [T_m]$, implies

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{\tau_{m,l}}{D}\right| > \frac{3}{2}\eta\right] \\
&\le \mathbb{P}\left[\left|(\boldsymbol{U}^{(m)})^\top(\boldsymbol{U}^{(m)}) - \mathrm{Var}(U^m)\right| > \frac{D}{2}\eta\right] + \mathbb{P}\left[\left|(X^{(m,l)})^\top(X^{(m,l)}) - \mathrm{Var}(X^m)\right| > \frac{D}{2}\eta\right] \\
&\qquad + \mathbb{P}\left[\left|(\boldsymbol{U}^{(m)} + X^{(m,l)})^\top(\boldsymbol{U}^{(m)} + X^{(m,l)}) - \mathrm{Var}(X^m + U^m)\right| > \frac{D}{2}\eta\right].
\end{aligned}$$

Here, $\mathrm{Var}(Y)$ denotes the trace of the covariance matrix of the random variable Y. We now estimate each summand separately, starting with

$$p_{m,l,\eta} := \mathbb{P}\left[\left|(\boldsymbol{U}^{(m)})^\top(\boldsymbol{U}^{(m)}) - \mathrm{Var}(U^{(m)})\right| > \frac{D}{2}\eta\right].$$

By the running assumptions, the vector $\boldsymbol{U}^{(m)}$ is T_m-dimensional mean zero Gaussian having covariance matrix $\sigma_m^2\mathbb{I}_{T_m}$; by the inequality in proposition 7.4, one has

$$p_{m,l,\eta} \le 2e^{-\frac{c_0 D\eta}{8\sigma_m^2}\min\{1, \frac{D\eta}{8\sigma_m^2 T_m}\}}. \tag{7.6}$$

Next, we consider

$$q_{m,l,\eta} := \mathbb{P}\left[\left|(\boldsymbol{X}^{(m,l)})^\top(\boldsymbol{X}^{(m,l)}) - \mathrm{Var}(X^{(m,l)})\right| > \frac{D}{2}\eta\right].$$

Recall that the random vector $X^{(m,l)}$ is mean-zero Gaussian with covariance matrix $\Gamma^{(m)}$ mentioned in equation (7.18) below. One has

$$\mathrm{Var}(X^{(m,l)}) = \mathrm{tr}(\Gamma^{(m)}) = T_m \mathbb{E}[(X_0^m)^2]\,.$$

From lemma 7.5, one has $\|\Gamma^{(m)}\|_{\mathrm{op}} \le \epsilon^{-2}\sigma_m^2$; thus, proposition 7.4 yields

$$q_{m,l,\eta} \le 2e^{-\frac{c_0 D\eta}{8\sigma_m^2\epsilon^{-2}}\min\{1,\frac{D\eta}{8\sigma_m^2 T_m\epsilon^{-2}}\}}\,. \tag{7.7}$$

Finally, we consider

$$r_{m,l,\eta} := \mathbb{P}\left[\left|(\boldsymbol{U}^{(m)} + X^{(m,l)})^\top(\boldsymbol{U}^{(m)} + X^{(m,l)}) - \mathrm{Var}(X^m + U^m)\right| > \frac{D}{2}\eta\right]\,.$$

Note that $\boldsymbol{U}^{(m)} + X^{(m,l)}$ is a mean zero Gaussian with symmetric covariance matrix $\tilde{\Gamma}^{(m)}$, whose (t,s)-entry (for $t \ge s$) is given by

$$\begin{aligned}\tilde{\Gamma}^{(m)}_{t,s} &:= \mathbb{E}[(X^m_{t-l} + U^m_t)(X^m_{s-l} + U^m_s)]\\ &= \mathbb{E}[(X^m_{t-l}X^m_{s-l}] + \mathbb{E}[X^m_{s-l}U^m_t] + \mathbb{E}[X^m_{t-l}U^m_s] + \mathbb{E}[U^m_t U^m_s]\\ &= \mathbb{E}[(X^m_{t-l}X^m_{s-l}] + \mathbb{E}[X^m_{t-l}U^m_s] + \sigma_m^2\mathbf{1}_{t=s}\,.\end{aligned}$$

Together with lemma 7.6, this implies

$$\begin{aligned}\|\tilde{\Gamma}^{(m)}\|_{\mathrm{op}} &\le \epsilon^{-2}\sigma_m^2 + \epsilon^{-4}\sigma_m^2 + \sigma_m^2\\ &= \sigma_m^2(1+\epsilon^{-2}+\epsilon^{-4})\,,\end{aligned}$$

and by proposition 7.4, we fetch

$$r_{m,l,\eta} \le 2e^{-\frac{c_0 D\eta}{8\sigma_m^2(1+\epsilon^{-2}+\epsilon^{-4})}\min\{1,\frac{D\eta}{8\sigma_m^2 T_m(1+\epsilon^{-2}+\epsilon^{-4})}\}}\,. \tag{7.8}$$

Now suppose that

$$\eta \ge \frac{8C_\sharp\sigma^2_{\max}(1+\epsilon^{-2}+\epsilon^{-4})}{M}\,;$$

then, the inequalities in (7.6), (7.7), and (7.8) are simplified as follows:

$$\begin{aligned} p_{m,l,\eta} &\le 2e^{-\frac{c_0 D\eta}{8\sigma_m^2}}\,;\\ q_{m,l,\eta} &\le 2e^{-\frac{c_0 D\eta}{8\sigma_m^2\epsilon^{-2}}}\,,\\ r_{m,l,\eta} &\le 2e^{-\frac{c_0 D\eta}{8\sigma_m^2(1+\epsilon^{-2}+\epsilon^{-4})}}\,.\end{aligned} \tag{7.9}$$

Of these, the right hand side is the largest in the bottom-most inequality (7.9). Note that, the union bound implies

$$\mathbb{P}\left[\sup_{m,l}\left|\frac{\tau_{m,l}}{D}\right| > \frac{3}{2}\eta\right] \le \sum_{m,l}\mathbb{P}\left[\left|\frac{\tau_{m,l}}{D}\right| > \frac{3}{2}\eta\right]$$

$$\leq \sum_{l\in[L]}\sum_{m\in[M]} \mathbb{P}\left[\left|\frac{\tau_{m,l}}{D}\right| > \frac{3}{2}\eta\right]$$

$$(7.9)\ \Rightarrow \qquad \leq 6\sum_{l\in[L]}\sum_{m\in[M]} e^{-\frac{c_0 D\eta}{8\sigma_m^2(1+\epsilon^{-2}+\epsilon^{-4})}}\,.$$

Thus, for any $\delta > 0$, in order to have the inequality

$$\mathbb{P}\left[\mathcal{N}_\star\left(\frac{2}{D}X^\top \boldsymbol{U}\right) > \frac{3\eta}{2\sqrt{L}}\right] \leq \delta\,,$$

it suffices to have

$$\exp\left(-\frac{c_0 D\eta}{8\sigma_m(1+\epsilon^{-2}+\epsilon^{-4})}\right) \leq \frac{\delta}{ML} \tag{7.10}$$

for each $m \in [M]$. Equivalently, it suffices to have the following for each $m \in [M]$:

$$\frac{c_0 D\eta}{8\sigma_m^2(1+\epsilon^{-2}+\epsilon^{-4})} \geq \log\left(\frac{ML}{\delta}\right)$$

$$\text{equivalently,} \qquad D \geq \frac{8\sigma_m^2(1+\epsilon^{-2}+\epsilon^{-4})}{c_0\eta}\log\left(\frac{ML}{\delta}\right).$$

This is equivalent to

$$D \geq \frac{8\sigma_{\max}^2(1+\epsilon^{-2}+\epsilon^{-4})}{c_0\eta}\log\left(\frac{ML}{\delta}\right). \tag{7.11}$$

It then follows from the argument above that

$$\mathbb{P}\left[\frac{2}{D}\mathcal{N}_\star(X^\top \boldsymbol{U}) \geq \frac{3\eta}{2\sqrt{L}}\right] \ \leq\ \delta \tag{7.12}$$

holds, provided

$$D \geq \frac{8\sigma_{\max}^2(1+\epsilon^{-2}+\epsilon^{-4})}{c_0\eta}\log\left(\frac{ML}{\delta}\right).$$

□

7.3. Proof of Proposition 4.7

We first observe what happens in the $M = 1$ case: this amounts to restricting X to the top left block

$$X_1 := \begin{pmatrix} x^1_{1-1} & \dots & x^1_{1-L} \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ x^1_{T_1-1} & \dots & x^1_{T_1-L} \end{pmatrix}.$$

Lemma 7.1 (Blockwise concentration inequality). *For any integer $s_1 > 0$, write $\mathbb{K}(s_1) := \mathbb{B}_0(s_1) \cap \mathbb{B}_1(1)$. Then*

$$\mathbb{P}\left[\sup_{\boldsymbol{v}_1 \in \mathbb{K}(s_1)} \left|\boldsymbol{v}_1^\top (X_1^\top X_1 - T_1\Gamma^{(1)})\boldsymbol{v}_1\right| \geq \frac{T_1^2 \zeta_1 \Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{2\mathfrak{m}(f)}\right]$$
$$\leq 2\exp\left(-\frac{T_1 \min\{\zeta_1, \zeta_1^2\}}{2} + s_1 \min\{\log L,\ \log(21eL/s_1)\}\right). \tag{7.13}$$

Proof of Lemma 7.1. Let $\boldsymbol{v}_1 \in \mathbb{R}^L$ be a fixed unit-normed vector. Then, $\boldsymbol{u}_1 := X_1\boldsymbol{v}_1$ is a random mean-zero Gaussian vector in $\mathbb{R}^{T_1}$, with covariance matrix $\Sigma_{\boldsymbol{v}_1}^{X_1} := \mathbb{E}\left[X_1\boldsymbol{v}_1\boldsymbol{v}_1^\top X_1^\top\right]$. One has

$$\begin{aligned}(\Sigma_{\boldsymbol{v}_1}^{X_1})_{r,s} &= \sum_{j,l=1}^{L} \mathbb{E}\left[X_{r-j}^1 \boldsymbol{v}_{1,j}\boldsymbol{v}_{1,l} X_{s-l}^1\right] \\ &= \sum_{j,l=1}^{L} \boldsymbol{v}_{1,j}\mathbb{E}\left[X_{r-j}^1 X_{s-l}^1\right]\boldsymbol{v}_{1,l} \\ &= \boldsymbol{v}_1^\top \Gamma_{r,s}\boldsymbol{v}_1\,,\end{aligned}$$

where $\Gamma_{r,s}$ is the covariance of the vectors $X_r^{(1)}$ and $X_s^{(1)}$. Note that

$$\begin{aligned}\operatorname{tr}(\Sigma_{\boldsymbol{v}_1}^{X_1}) &= \mathbb{E}\left[\operatorname{tr}(X_1\boldsymbol{v}_1\boldsymbol{v}_1^\top X_1^\top)\right] \\ &= T_1\mathbb{E}\left[(X_1\boldsymbol{v}_1)_{11}^2\right] \\ &= T_1\boldsymbol{v}_1^\top\Gamma^{(1)}\boldsymbol{v}_1\,,\end{aligned}$$

where $\Gamma^{(1)}$ is the autocovariance matrix of $\{X_t^1\}$. Moreover, one has by Lemma 7.5, the inequality

$$\|\Sigma_{\boldsymbol{v}_1}^{X_1}\|_{\mathrm{op}} \leq \frac{\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{m}(f)}\,.$$

By the inequality in Proposition 7.4, one has

$$\mathbb{P}\left[\left|\boldsymbol{u}_1^T\boldsymbol{u}_1 - \operatorname{tr}(\Sigma_{\boldsymbol{v}_1}^{X_1})\right| \geq 4T_1\zeta_1\|\Sigma_{\boldsymbol{v}_1}^{X_1}\|_{\mathrm{op}}\right] \leq 2e^{-\frac{T_1}{2}\min\{\zeta_1,\zeta_1^2\}}$$

for any $\zeta_1 > 0$. In particular, this implies that for any fixed $\boldsymbol{v}_1 \in \mathbb{R}^L$, the inequality

$$\mathbb{P}\left[\left|\boldsymbol{v}_1^\top(X_1^\top X_1 - T_1\Gamma^{(1)})\boldsymbol{v}_1\right| \geq \frac{4T_1\zeta_1\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{m}(f)}\right] \leq 2e^{-\frac{T_1}{2}\min\{\zeta_1,\zeta_1^2\}} \tag{7.14}$$

holds for any $\zeta_1 > 0$. Applying Lemma 7.7 with $G := X_1^\top X_1 - T_1\Gamma^{(1)}$, we conclude that

$$\mathbb{P}\left[\sup_{\boldsymbol{v}_1 \in \mathbb{K}(s_1)} \left|\boldsymbol{v}_1^\top(X_1^\top X_1 - T_1\Gamma^{(1)})\boldsymbol{v}_1\right| \geq \frac{4T_1\zeta_1\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{m}(f)}\right]$$

$$\leq 2\exp\left(-\frac{T_1\min\{\zeta_1,\zeta_1^2\}}{2}+s_1\min\{\log L,\ \log(21eL/s_1)\}\right) \tag{7.15}$$

holds for any integer $s_1 \geq 1$, and any $\zeta_1 > 0$. □

Proof of Proposition 4.7

Proof. When $s_1 > 0$ is even, it follows from Lemma 7.8 that

$$\mathbb{P}\left[\sup_{\boldsymbol{v}_1\in\mathbb{R}^L}\left|\boldsymbol{v}_1(X_1^\top X_1 - T_1\Gamma^{(1)})\boldsymbol{v}_1\right| \leq \frac{108T_1\zeta_1\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{m}(f)}\left(\|\boldsymbol{v}_1\|_2^2+\frac{2}{s_1}\|\boldsymbol{v}_1\|_1^2\right)\right]$$
$$\geq\ 1-2\exp\left(-\frac{T_1\min\{\zeta_1,\zeta_1^2\}}{2}+s_1\min\{\log L,\ \log(21eL/s_0)\}\right).$$

An application of Lemma 7.5 implies that for

$$\delta := 2e^{-\frac{T_1\min\{\zeta_1,\zeta_1^2\}}{2}+s_1\min\{\log L,\ \log(21eL/s_1)\}},$$

the conclusion of the following sequence of inequalities holds for all $\boldsymbol{v}_1 \in \mathbb{R}^L$ with probability at least $1-\delta$:

$$\begin{aligned}\|X_1\boldsymbol{v}_1\|_2^2 &\geq T_1\lambda_{\min}(\Gamma^{(1)})\|\boldsymbol{v}_1\|_2^2 - \frac{108T_1\zeta_1\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{m}(f)}\left(\|\boldsymbol{v}_1\|_2^2+\frac{2}{s_1}\|\boldsymbol{v}_1\|_1^2\right)\\ &\geq \frac{T_1\Lambda_{\min}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{M}(f)}\|\boldsymbol{v}_1\|_2^2 - \frac{108T_1\zeta_1\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{m}(f)}\left(\|\boldsymbol{v}_1\|_2^2+\frac{2}{s_1}\|\boldsymbol{v}_1\|_1^2\right)\\ &= T_1\left(\frac{\Lambda_{\min}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{M}(f_1)} - \frac{108\zeta_1\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{\mathfrak{m}(f_1)}\right)\|\boldsymbol{v}_1\|_2^2 - \frac{216T_1\zeta_1\Lambda_{\max}(\Sigma_{\boldsymbol{U}_1})}{s_1\mathfrak{m}(f_1)}\|\boldsymbol{v}_1\|_1^2.\end{aligned}$$

We specialize to the case of univariate ϵ-stable autoregressive time series, with $\Sigma_{\boldsymbol{U}_1} = \sigma_1^2$, and (by ϵ-stability)

$$\epsilon^2 \leq \mathfrak{m}(f_1) \leq \mathfrak{M}(f_1) \leq \epsilon^{-2}.$$

Letting $\zeta_1 := 6^{-3}\epsilon^4$, we obtain

$$\mathbb{P}\left[\inf_{\boldsymbol{v}_1\in\mathbb{R}^L}\left(\|X_1\boldsymbol{v}_1\|_2^2 - \frac{T_1}{2}\sigma_1^2\epsilon^2\left(\|\boldsymbol{v}_1\|_2^2-\frac{2}{s_1}\|\boldsymbol{v}_1\|_1^2\right)\right)\geq 0\right]$$
$$\geq 1-2e^{-\frac{T_1}{2}\zeta_1^2+s_0\min\{\log L,\ \log(21eL/s_0)\}}. \tag{7.16}$$

Finally, we consider all the blocks of $X^\top X$; the deviation inequality above yields

$$\mathbb{P}\left[\forall_{m\in[M]}\inf_{\boldsymbol{v}_m\in\mathbb{R}^L}\left(\|X_m\boldsymbol{v}_m\|_2^2 - \frac{T_m\sigma^2\epsilon^2}{2}\left(\|\boldsymbol{v}_m\|_2^2-\frac{2}{s_m}\|\boldsymbol{v}_m\|_1^2\right)\right)\geq 0\right]$$
$$\geq 1-2\sum_{m=1}^{M}e^{-\frac{T_m\zeta^2}{2}+s_m\min\{\log L,\ \log(21eL/s_m)\}}. \tag{7.17}$$

This proves the proposition. □

7.4. More on Dual Norms

Since $\mathcal{N}_\star(\mathbf{0}) = 0$, we will assume, in the following computation of the dual norm, that $\boldsymbol{\alpha} \neq \mathbf{0}$. This ensures $\mathcal{N}_\star(\boldsymbol{\alpha}) > 0$ too. Note that $\mathcal{N}_\star(\boldsymbol{\alpha}) = \mathcal{N}_\star(|\boldsymbol{\alpha}|)$, where $|\boldsymbol{\alpha}| = (|\alpha_1|, \cdots, |\alpha_L|)$; this is because $\mathcal{N}$ is invariant under arbitrary sign changes of the coordinates of its argument. Thus, for $\mathcal{N}_\star(\boldsymbol{\alpha}) = \boldsymbol{\alpha} \cdot \boldsymbol{\beta}$, we may assume (without loss of generality) that $|\boldsymbol{\alpha}| = \boldsymbol{\alpha}$ and $|\boldsymbol{\beta}| = \boldsymbol{\beta}$.

The advantage of the hierarchical group norm $\mathcal{N}(\boldsymbol{\beta})$ — in the setting of sparse recovery via regularized least square regression — is that, while it sets any group of parameters to zero — because of the hierarchy in the group structure — all variables in all groups that appear further down the order of hierarchy are set to zero automatically. Thus, these norms are well-suited for determination of the true lag order in the context of autoregressive process.

In order to use this group norm in our analysis, we need to collect some basic facts on the norm. Note that the norm resembles ℓ^1 at the group level, while within each group it is the ℓ^2 norm; from this, one might expect that the dual of the norm should "resemble" the ℓ^∞-norm at the group level, and within a group it should be ℓ^2. We will prove that this intuition goes quite well, in the sense that the actual dual norm can be upper-bounded by this mixed $\ell^{\infty,2}$ norm.

Lemma 7.2 (Dual norm is attained on the boundary).

$$\mathcal{N}_\star(\boldsymbol{\alpha}) = \max_{\mathcal{N}(\boldsymbol{\beta})=1} \langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle \,.$$

Proof of Lemma 7.2. Note that — by continuity of $\boldsymbol{\beta} \mapsto \langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle$, compactness of the subset $C_\beta := \{\boldsymbol{\beta} : \mathcal{N}(\boldsymbol{\beta}) \leq 1\}$, and since $\mathcal{N}_\star(\boldsymbol{\alpha}) \neq 0$ — there is nonzero $\boldsymbol{\beta}_0 \in C_\beta$ such that $\mathcal{N}_\star(\boldsymbol{\alpha}) = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}_0 \rangle$. Suppose, if possible, that $\boldsymbol{\beta}_0$ satisfies $\mathcal{N}(\boldsymbol{\beta}_0) < 1$. In $(0, +\infty)$, the function $c \mapsto \mathcal{N}(c\boldsymbol{\beta}_0)$ is strictly increasing (continuous) function; thus, there is $c > 1$ such that $c\mathcal{N}(\boldsymbol{\beta}_0) = \mathcal{N}(\boldsymbol{c\beta}_0) \leq 1$; one has

$$\begin{aligned} \max\{\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle : \mathcal{N}(\boldsymbol{\beta}) \leq 1\} &\geq \langle \boldsymbol{\alpha}, c\boldsymbol{\beta_0} \rangle \\ &= c \cdot \langle \boldsymbol{\alpha}, \boldsymbol{\beta}_0 \rangle \\ &> \langle \boldsymbol{\alpha}, \boldsymbol{\beta}_0 \rangle \\ &= \mathcal{N}_\star(\boldsymbol{\alpha}) \,, \end{aligned}$$

which is a contradiction. □

In order to facilitate our computations of the dual norm in $\mathbb{R}^{ML}$, we start with the special case of $M = 1$; after Proposition 7.3, we will extend this to the general case. Now, we define

$$\mathcal{N}^1(\boldsymbol{\alpha}) := \sum_{l=1}^{L} \sqrt{(L-l+1) \sum_{j=l}^{L} \alpha_j^2} \,.$$

Let us write $\boldsymbol{\alpha}_{\geq j} = (\mathbf{0}, \alpha_j, \alpha_{j+1}, \cdots, \alpha_L)$ for $j \in [L]$. From now on, we assume that

$$\prod_{j \in [L]} \alpha_j \neq 0 \,.$$

We have the following proposition.

Proposition 7.3 (L^∞ **norm bounds dual norm,** $M = 1$ **case).** *The following inequality holds for any* $\boldsymbol{\alpha} \in \mathbb{R}^L$*:*

$$\mathcal{N}_\star^1(\boldsymbol{\alpha}) \le L^{-\frac{1}{2}} \|\boldsymbol{\alpha}\|_\infty .$$

Proof of Proposition 7.3. We apply the elementary inequalities

$$\begin{aligned}
(\sqrt{L-l+1}\|\beta_{\ge l}\|_2 - \sqrt{L-l+1}|\beta_l|)^2 &\ge (L-l+1)\|\beta_{\ge l}\|_2^2 - (L-l+1)|\beta_l|^2 \\
&= (L-l+1)\sum_{j>l}\beta_j^2 \\
&= \left(\sqrt{(L-l+1)\sum_{j>l}\beta_j^2}\right)^2 \\
\text{Cauchy-Schwarz } \Rightarrow \qquad &\ge \left(\sum_{j>l}\beta_j\right)^2 ,
\end{aligned}$$

to derive

$$\begin{aligned}
\mathcal{N}^1(\boldsymbol{\beta}) &= \sum_{l=1}^{L}\sqrt{L-l+1}\|\beta_{\ge l}\|_2 \\
&\ge \sum_{l=1}^{L}\left(\sqrt{L-l+1}|\beta_l| + \sum_{j>l}|\beta_j|\right) \\
&= \sum_{l=1}^{L}\left(\sqrt{L-l+1}+l-1\right)|\beta_l| \\
&\ge \sqrt{L}|\beta_l| .
\end{aligned}$$

Since $\mathcal{N}_\star^1(\boldsymbol{\alpha}) \le \|\boldsymbol{\alpha}\|_\infty \mathcal{N}^1(\mathbf{1})$, where $\mathbf{1} \in \mathbb{R}^L$ is the *all-one* vector, it suffices to show that $\mathcal{N}_\star^1(\mathbf{1}) \le L^{-\frac{1}{2}}$. Since

$$\mathcal{N}_\star^1(\mathbf{1}) = \max_{\mathcal{N}^1(\boldsymbol{\beta})=1} \sum_{j\in[L]}\beta_j ,$$

and

$$\begin{aligned}
1 &= \mathcal{N}^1(\boldsymbol{\beta}) \\
&= \sum_{j\in[L]}\sqrt{L-l+1}\|\beta_{\ge j}\|_2 \\
&\ge \sqrt{L}\sum_{j\in[L]}|\beta_j| ,
\end{aligned}$$

the claim follows. □

With the above proposition dealing with the case $M = 1$, we can now derive the proof of Proposition 4.3 dealing with general M.

Proof of Proposition 4.3. Again, it suffices to show that $\mathcal{N}(\boldsymbol{\alpha}) \le L^{-\frac{1}{2}}$ when $\boldsymbol{\alpha}$ is the all-one vector.

Now, $\mathcal{N}_\star(\boldsymbol{\alpha})$ is the optimum value of the problem

$$\begin{aligned} \text{maximize} \quad & \sum_{j\in[ML]} \beta_j \\ \text{subject to} \quad & \mathcal{N}(\boldsymbol{\beta}) = 1\,. \end{aligned}$$

Write $\boldsymbol{\beta}_{(l)} := (\beta_l, \beta_{L+l}, \cdots, \beta_{(M-1)L+l})^\top$ and $x_l := \sqrt{M}\|\boldsymbol{\beta}_{(l)}\|_2$ for $l \in [L]$; by Cauchy-Schwarz, we have

$$\sum_{j\in[ML]} \beta_j \le \sum_{l=1}^{L} x_l\,.$$

Moreover, $\mathcal{N}(\boldsymbol{\beta}) = \mathcal{N}^1(\boldsymbol{x})$, where $\boldsymbol{x} := (x_1, \cdots, x_L)$; thus, the lemma follows by 7.3. □

7.5. Some known results used in the proofs

The following tail bound on the norm of Gaussian random vectors is well-known; see (Vershynin, 2018, Theorem 6.2.1) for details.

Proposition 7.4 (Wright-Hansen inequality). *There is an absolute constant $c_0 > 0$ such that the following statement holds. If $\boldsymbol{q} \sim N(\mathbf{0}, \Sigma)$ in $\mathbb{R}^d$, where Σ is symmetric positive definite, then the following holds for any $\tau > 0$:*

$$\mathbb{P}\left[\frac{1}{d}\left|\boldsymbol{q}^T\boldsymbol{q} - tr(\Sigma)\right| \ge 4\tau\|\Sigma\|_{op}\right] \le 2e^{-c_0 d\min\{\tau^2,\tau\}}\,.$$

We need to collect some information about the multiple autoregressive time series model. Let $T := \min_{m\in[M]} T_m$, and recall that we have written $D = T_1 + \cdots + T_M$. Let $\Gamma^{(m,l)}$ be the covariance matrix of the random vector $X^{(m,l)}$; this is a symmetric matrix, and for any $r, s \in \{1-l, \ldots, T_m - l\}$ with $r \le s$, the stationarity of the time series yields

$$\Gamma^{(m,l)}_{r,s} = \mathbb{E}[X^{(m,l)}_{r-l} X^{(m,l)}_{s-l}] = \mathbb{E}[X^m_0 X^m_{s-r}]\,. \tag{7.18}$$

Thus, we can write $\Gamma^{(m)} := \Gamma^{(m,l)}$ for the "auto-covariance" matrix.

Now, let further, for any two integers $T > L > 0$, ${}_{T,L}\Gamma$ denote the $T_1 L \times T_1 L$ matrix, whose (r,s)-th block is the covariance of the vectors $(X_{r-1}, X_{r-2}, \ldots, X_{r-L})$ and $(X_{s-1}, X_{s-2}, \ldots, X_{s-L})$.

The following lemma gives eigenvalue bounds of these block covariance matrices.

Lemma 7.5 (Eigenvalue bound for Toeplitz matrices). *Let $f(z)$ denote the reverse characteristic polynomial of a stable univariate AR-process $\{X_t\}$. Let $\mathfrak{m}(f) = \inf_{|z|\le 1}|f(z)|^2$ and $\mathfrak{M}(f) = \sup_{|z|\le 1}|f(z)|^2$. The following inequality holds:*

$$\frac{\Lambda_{\min}(\Sigma_\epsilon)}{\mathfrak{M}(f)} \le \Lambda_{\min}({}_{T,L}\Gamma) \le \Lambda_{\max}({}_{T,L}\Gamma) \le \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mathfrak{m}(f)}\,.$$

Proof of Lemma 7.5. This is immediate from (Basu and Michailidis, 2015, Proposition 2.3) if we consider the stable dim-L VAR process $Y_t := (X_{t-1}, \ldots, X_{t-L})$. □

The following bound appeared as of (Basu and Michailidis, 2015, Proposition 2.4b).

Lemma 7.6 (Basu-Michailidis). *Let $\Delta^{(m)}$ denote the T_m-dimensional matrix with $\Delta^{(m)}_{t,s} := \mathbb{E}[X^m_{t-l} U^m_s]$. The following inequality holds:*

$$\Lambda_{\max}(\Delta^{(m)}) \leq \frac{\Lambda_{\max}(\Sigma_{\epsilon_m}) \mathfrak{M}(f_m)}{\mathfrak{m}(f)} .$$

The following appeared as (Basu and Michailidis, 2015, Lemma F.2). Recall that $\mathbb{M}_L(\mathbb{R})$ denotes the algebra of all $L \times L$ real matrices, and $\mathbb{S}^{L-1}$ the unit sphere in $\mathbb{R}^L$.

Lemma 7.7 (Basu-Michailidis). *Suppose that $G \in \mathbb{M}_L(\mathbb{R})$ is a random symmetric matrix for which the following inequality holds for every $\boldsymbol{u} \in \mathbb{S}^{L-1}$, $T_1 \in \mathbb{N}$, and $\eta > 0$:*

$$\mathbb{P}\left[\left|\boldsymbol{u}^T G \boldsymbol{u}\right| \geq C\eta\right] \leq 2e^{-cT_1 \min\{\eta, \eta^2\}} ,$$

where $C, c > 0$ are parameters independent of η and $\boldsymbol{u}$. Then the following inequality holds for any integer $s_0 > 0$:

$$\mathbb{P}\left[\sup_{\|\boldsymbol{u}\|_0 \leq s_0,\ \|\boldsymbol{u}\|_2 \leq 1} \left|\boldsymbol{u}^T G \boldsymbol{u}\right| \geq C\eta\right] \leq 2e^{-cT_1 \min\{\eta, \eta^2\} + s_0 \min\{\log L, \log(21eL/s_0)\}} . \tag{7.19}$$

The following result was obtained as (Loh and Wainwright, 2012, Lemma 12). Recall that, for any integer $s_0 > 0$, we denote

$$\mathbb{K}(s_0) = \{\boldsymbol{v} \in \mathbb{R}^L : \|\boldsymbol{v}\|_2 \leq 1,\ \|\boldsymbol{v}\|_0 \leq s_0\} .$$

Lemma 7.8 (Loh-Wainwright). *Suppose that $G \in \mathbb{M}_L(\mathbb{R})$ is a symmetric matrix for which the following holds:*

$$\sup_{\|\boldsymbol{u}\|_0 \leq s_0,\ \|\boldsymbol{u}\|_2 \leq 1} \left|\boldsymbol{u}^T G \boldsymbol{u}\right| \leq \delta .$$

Then, the following inequality holds:

$$\sup_{\boldsymbol{u} \in \mathbb{R}^L} \left|\boldsymbol{u}^T G \boldsymbol{u}\right| \leq 27\delta \left(\|\boldsymbol{u}\|_2^2 + \frac{2}{s_0}\|\boldsymbol{u}\|_1^2\right) . \tag{7.20}$$

References

BASU, S. and MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43** 1535–1567. https://doi.org/10.1214/15-AOS1315 MR3357870

BUNEA, F. (2008). Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electron. J. Stat.* **2** 1153–1194. https://doi.org/10.1214/08-EJS287 MR2461898

CANDES, E. J. and TAO, T. (2006). Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory* **52** 5406–5425. https://doi.org/10.1109/TIT.2006.885507 MR2300700

COMBETTES, P. L. and WAJS, V. R. (2005). Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.* **4** 1168–1200. https://doi.org/10.1137/050626090 MR2203849

LEDERER, J. (2021). *Fundamentals of high-dimensional statistics—with exercises and R labs*. *Springer Texts in Statistics*. Springer, Cham. https://doi.org/10.1007/978-3-030-73792-4 MR4385517

LI, S., JIN, X., XUAN, Y., ZHOU, X., CHEN, W., WANG, Y.-X. and YAN, X. (2019). Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. In *Advances in Neural Information Processing Systems* (H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D' ALCHÉ-BUC, E. FOX and R. GARNETT, eds.) **32**. Curran Associates, Inc.

LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *Ann. Statist.* **40** 1637–1664. https://doi.org/10.1214/12-AOS1018 MR3015038

LÜTKEPOHL, H. (2005). *New introduction to multiple time series analysis*. Springer-Verlag, Berlin. https://doi.org/10.1007/978-3-540-27752-1 MR2172368

MAIRAL, J., JENATTON, R., OBOZINSKI, G. and BACH, F. (2011). Convex and network flow optimization for structured sparsity. *J. Mach. Learn. Res.* **12** 2681–2720. MR2845676

NARDI, Y. and RINALDO, A. (2011). Autoregressive process modeling via the Lasso procedure. *J. Multivariate Anal.* **102** 528–549. https://doi.org/10.1016/j.jmva.2010.10.012 MR2755014

NICHOLSON, W. B., WILMS, I., BIEN, J. and MATTESON, D. S. (2020). High dimensional forecasting via interpretable vector autoregression. *J. Mach. Learn. Res.* **21** Paper No. 166, 52. MR4209452

TSENG, P. (2008). On Accelerated Proximal Gradient Methods for Convex-Concave Optimization. available at https://www.mit.edu/ dimitrib/PTseng/papers/apgm.pdf.

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. U. and POLOSUKHIN, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems* (I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN and R. GARNETT, eds.) **30**. Curran Associates, Inc.

VERSHYNIN, R. (2018). *High-dimensional probability*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge University Press, Cambridge An introduction with applications in data science, With a foreword by Sara van de Geer. https://doi.org/10.1017/9781108231596 MR3837109

ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37** 3468–3497. https://doi.org/10.1214/07-AOS584 MR2549566