Estimation in High-dimensional Vector Autoregressive Models

Sumanta Basu and George Michailidis
Department of Statistics, University of Michigan

Abstract

Vector Autoregression (VAR) is a widely used method for learning complex interrelationship among the components of multiple time series. Over the years it has gained popularity in the fields of control theory, statistics, economics, finance, genetics and neuroscience. We consider the problem of estimating stable VAR models in a high-dimensional setting, where both the number of time series and the VAR order are allowed to grow with sample size. In addition to the “curse of dimensionality” introduced by a quadratically growing dimension of the parameter space, VAR estimation poses considerable challenges due to the temporal and cross-sectional dependence in the data. Under a sparsity assumption on the model transition matrices, we establish estimation and prediction consistency of \( \ell_1 \)-penalized least squares and likelihood based methods. Exploiting spectral properties of stationary VAR processes, we develop novel theoretical techniques that provide deeper insight into the effect of dependence on the convergence rates of the estimates. We study the impact of error correlations on the estimation problem and develop fast, parallelizable algorithms for penalized likelihood based VAR estimates.

1 Introduction

Vector Autoregression (VAR) represents a popular class of time series models in applied macroeconomics and finance, widely used for structural analysis and simultaneous forecasting of a number of temporally observed variables. Unlike structural models, VAR provides a broad framework for capturing complex temporal and cross-sectional interrelationship among the time series. In addition to economics, VAR models have been instrumental in linear system identification problems in control theory, while more recently, they have become standard tools in genomics for reconstruction of regulatory networks (Michailidis and d’Alché Buc, 2013) and neuroscience for understanding functional connectivity patterns between brain regions.

Formally, for a \( p \)-dimensional vector-valued stationary time series \( X^t = \{X^t_1, \ldots, X^t_p\} \), a VAR model of lag \( d \) (VAR(d)) with serially uncorrelated Gaussian errors takes the form

\[
X^t = A_1 X^{t-1} + \ldots + A_d X^{t-d} + \epsilon^t, \quad \epsilon^{t, i.i.d} \sim N(0, \Sigma_{\epsilon})
\]

(1.1)

where \( A_1, \ldots, A_d \) are \( p \times p \) matrices and \( \epsilon^t \) is a \( p \)-dimensional vector of possibly correlated innovation shocks. In the classical setting, estimation of VAR models is undertaken under the assumption that the process is stable, i.e., \( \det(I_p - \sum_{t=1}^d A_t z^t) \neq 0 \) for all \( z \in \mathbb{C}, |z| \leq 1 \). Note that stability
guarantees stationarity. For further discussion on stability we refer the reader to Lütkepohl (2005).

The main objective in VAR models is to estimate the order of the model $d$ and the transition matrices $A_1, \ldots, A_d$. The structure of the transition matrices provide insight into the complex temporal relationships amongst the $p$ time series and lead to efficient forecasting strategies. In many analyses, a second goal is to estimate the error correlations. The matrix $\Sigma$ captures the correlation between unknown idiosyncratic errors, which are attributed to additional contemporaneous dependency among the time series.

VAR models can also be represented as directed graphs with a partial order (induced by time) among the nodes. Figure 1 depicts a representation of (1.1) with $(d + 1)p$ nodes. The transition matrix of lag $j$ ($A_j$) is precisely the adjacency matrix of the bipartite subgraph corresponding to the nodes $\{X_{t-j}, X_t\}$. If the error process is Gaussian, $\Sigma^{-1}$ corresponds to an undirected graph of additional contemporaneous interaction among the $p$ time series. The directed graphical model

![Figure 1: Graphical representation of the VAR model (1.1): directed edges (solid) correspond to the entries of the transition matrices, undirected edges (dashed) correspond to the entries of $\Sigma^{-1}$](image)

induced by the adjacency matrices $A_1, \ldots, A_d$ is commonly known as a network of “Granger-causality”. It should be noted that this network, often learned from observational data, merely captures the transition patterns and true causal inference can be drawn only under stringent assumptions.

The advent of novel data collection and storage technologies in macroeconomics and finance has enabled researchers to study the relationship amongst a large number of time series. In addition to better forecasting performance, large VAR models also provide better insight into the structure of the economy. For example, Christiano et al. (1999) argues that the increase in price level shortly followed by unexpected monetary tightening, commonly known as “price puzzle”, is an artifact of not including forward looking variables in the model. Also in financial applications, it is often necessary to analyze a large number of time series simultaneously. The problem is that the dimensionality of VAR models grows quadratically with $p$. For example, estimating a VAR(10) model with $p = 10$ time series requires estimating $dp^2 = 1000$ parameters. However, it is often very difficult, to obtain a comparable number of stationary observations.

Common approaches to deal with this “curse of dimensionality” is to apply shrinkage on the coefficients of the transition matrices or impose some low-dimensional structural assumption on the
Figure 2: In the left panel, we consider a VAR(1) model with $p = 2$, $X_t = A_1 X_{t-1} + \varepsilon_t$, where $A_1 = [\alpha \ 0; \beta \ \alpha]$. The unbounded set (dotted) denotes the values of $(\alpha, \beta)$ for which the process is stable. The bounded region (solid) represents the VAR models that satisfy $\|A_1\| < 1$. In the right panel, we consider a VAR(2) model with $p = 1$, $X_t = 2\alpha X_{t-1} - \alpha^2 X_{t-2} + \varepsilon_t$. Equivalent formulation of this model as VAR(1) is: $Y_t = \tilde{A}_1 Y_{t-1} + \tilde{\varepsilon}_t$, where $Y_t = [X_t, X_{t-1}]'$, $\tilde{A}_1 = [2\alpha - \alpha^2; 1 0]$, and $\tilde{\varepsilon}_t = [\varepsilon_t, 0]'$. The model is stable whenever $|\alpha| < 1$ but $\|\tilde{A}_1\|$ is always greater than or equal to 1.

VAR model. Bayesian shrinkage is the most popular among such approaches (Bańbura et al., 2010).
due to its simple implementation and good forecasting performance. In the econometrics literature, a common structural assumption is the presence of few underlying factors, driving economic or market behavior. These assumptions can be incorporated through factor models or principal component analysis. However, none of these methods are capable of undertaking variable selection. Recent developments in high-dimensional statistics provide a framework to incorporate different lower-dimensional structural assumptions in the model using penalized estimation techniques (e.g., by invoking lasso, group lasso penalties, or low-rank etc.). The simplest of these structures is to assume sparsity on the edge set of the underlying graphical model. This is incorporated in the VAR model by optimizing an $\ell_1$-penalization procedure like lasso. Accurate, interpretable, computationally attractive - lasso has received a lot of attention over the past decade in the Statistics community.

In a recent work De Mol et al. (2008) have shown that lasso and ridge penalized estimates often provide forecasting accuracy that is comparable to principal component regressions.

Studying the theory of VAR estimation in the high-dimensional regime poses considerable challenge due to the dependence between the rows and columns of the design matrix, its cross-dependence with the vector of errors and also the dependence among the columns of the error matrix in the multiple regression. Some early attempts to establish consistency of VAR models under high-dimensional scaling include Song and Bickel (2011) and Kock and Callot (2012). Both papers rely on certain regularity assumptions but do not investigate their validity in the context of VAR. Recent works of Negahban and Wainwright (2011) and Loh and Wainwright (2012) provide a more in-depth analysis of the dependence and prove estimation consistency of VAR(1) models under the assumption $\|A_1\| < 1$, where $\|\cdot\|$ denotes the operator norm of a matrix. We show that even this assumption is restrictive and cannot accommodate all stable VAR(1) models. More importantly, as shown in Appendix D, these techniques cannot be directly adapted to VAR(d) model for $d > 1$, where the dependence structure is more involved. We illustrate this in Figure 2 with two examples. The theoretical properties of the penalized likelihood based estimates have not been addressed before. We provide a unified theory of the least-squares and log-likelihood based VAR estimates which can accommodate any stable VAR(d) model, for any $d \geq 1$. This is accomplished by introducing some novel techniques to quantify the unique dependence structure in VAR processes and proving sharp concentration bounds. Our results exploit the spectral properties of stationary VAR processes and provide additional insight into the effect of dependence on the convergence rates of the estimators.

Two issues need careful consideration in penalized estimation - (a) choice of loss function (least squares, negative log-likelihood, some other type of divergence etc.), (b) form of penalty (a structural assumption that is relevant to the context of the application and translates to a scalable optimization problem). The second issue, i.e., choice of penalty, is an active area of research in Statistics. Choice of loss function, however, has received less attention in the literature of high-dimensional linear regression, primarily because the response is univariate. In VAR settings, one deals with prediction of multiple responses and we show that the choice of loss function can play an important role in the modeling strategy.

The standard way to incorporate the sparsity assumption is to penalize the least square objective function ($\ell_1$-LS). This estimate does not account for the contemporaneous correlation ($\Sigma_e$) among the errors (see Figure 1). In a recent paper Davis et al. (2012) showed that a penalized log-likelihood based approach ($\ell_1$-LL), which takes into account the structure of $\Sigma_e$, sometimes performs better than $\ell_1$-LS. The proposed estimator is a maximum likelihood estimate of $A^h$, assuming $\Sigma_e$ is known. Maximizing the complete likelihood ($\ell_1$-ML) with respect to $A^h, \Sigma_e$ poses additional challenge as it involves estimating both the transition matrices $A^h$ and a high-dimensional inverse covariance matrix $\Sigma_e^{-1}$. In this work we provide theoretical results for the first two estimates and compare their
performances numerically.

Computation of the penalized least-square estimates ($\ell_1$-LS) of large VAR is straightforward since the estimation problem can be decomposed into $p$ separate LASSO programs and computation can be carried out in parallel. Computing the likelihood based estimators $\ell_1$-LL and $\ell_1$-ML, however, needs more work since the regressions are not separable. For example, the algorithm presented in Davis et al. (2012) does not take advantage of parallel computing when $p$ is large. To this end, we propose a blockwise coordinate descent algorithm that is parallelizable and is of the same order of complexity as $\ell_1$-LS.

In summary, the main contributions of this work are: (a) develop novel theoretical techniques to analyze penalized estimators for high-dimensional stable, Gaussian VAR model of any order $d$, (b) investigate analytically the properties of least square and likelihood based methods in the high-dimensional regime, and (c) develop parallellizable algorithm for likelihood based VAR estimates.

2 Modeling Framework

Consider a single realization of $\{X^0, X^1, \ldots, X^T\}$ generated according to a $p$-dimensional vector autoregressive model of order $d$ with serially uncorrelated Gaussian errors

$$X^t = A_1 X^{t-1} + \ldots + A_d X^{t-d} + \epsilon^t, \quad \epsilon^t \sim N(0, \Sigma_e), \quad \text{Cov}(\epsilon^t, \epsilon^s) = 0_{p \times p} \text{ if } t \neq s$$

Throughout the paper we will assume that the error covariance matrix $\Sigma_e$ is positive definite so that $\lambda_{\min}(\Sigma_e) > 0$ and $\lambda_{\max}(\Sigma_e) < \infty$. Here $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum eigenvalues of a symmetric or hermitian matrix, respectively.

We will also assume that the VAR process is stable, i.e., all the eigenvalues of the reverse characteristic polynomial $A(z) = I_p - \sum_{t=1}^d A_t z^t$ lie outside the unit circle $\{z \in \mathbb{C} : |z| = 1\}$. Formally,

$$\det(A(z)) = \det(I_p - \sum_{t=1}^d A_t z^t) \neq 0, \text{ for all } |z| \leq 1$$

The stability of the process guarantees that all the eigenvalues of the Hermitian matrix $A^*(z)A(z)$ are positive, whenever $|z| = 1$. We take the following quantity as a measure of stability of the VAR(d) process:

$$\mu_{\min}(A) = \min_{|z|=1} \lambda_{\min}(A^*(z)A(z))$$

We also define a second, related quantity that will be useful in our theoretical analyses:

$$\mu_{\max}(A) = \max_{|z|=1} \lambda_{\max}(A^*(z)A(z))$$

Note that $\mu_{\min}(A)$ and $\mu_{\max}(A)$ are well-defined because of the continuity of eigenvalues and compactness of the unit circle $\{z \in \mathbb{C} : |z| = 1\}$. Further, stability of the process guarantees that these are positive.

We will often use an equivalent representation of a VAR(d) model. This states that a $p$-dimensional VAR(d) process (2.1) can be represented as a $dp$-dimensional VAR(1) process $\tilde{X}^t = \tilde{A}_1 \tilde{X}^{t-1} + \tilde{\epsilon}^t$.
with

\[
\begin{bmatrix}
X_t \\
X^{t-1} \\
\vdots \\
X^{t-d+1}
\end{bmatrix}_{dp \times 1} 
\begin{bmatrix}
\tilde{A}_1 \\
\vdots \\
\tilde{A}_{d-1} \\
\tilde{A}_d
\end{bmatrix}_{dp \times dp} 
\begin{bmatrix}
e_t \\
\vdots \\
0
\end{bmatrix}_{dp \times 1}
\]

(2.3)

In view of (2.2), the process \( \tilde{X}^t \) with reverse characteristic polynomial \( \tilde{A}(z) := I_{dp} - \tilde{A}_1 z \) is stable if and only if the process \( X^t \) is stable. However, the stability measures \( \mu_{\min}(\tilde{A}), \mu_{\max}(\tilde{A}) \) are not necessarily the same as \( \mu_{\min}(\tilde{A}), \mu_{\max}(\tilde{A}) \).

A key ingredient of our frequency domain analysis is the spectral density matrix of the VAR process (2.1), defined as

\[
f(\theta) = \frac{1}{2\pi} \sum_{l=-\infty}^{\infty} \Gamma(l)e^{-il\theta}, \quad \theta \in [-\pi, \pi]
\]

(2.4)

where \( \Gamma(l) = \mathbb{E} [X^t (X^{t+l})'] \) denotes the autocovariance matrix of order \( l \). If the process is stable, the spectral density matrix has the following closed form expression (cf. equation (9.4.23), Priestley (1981))

\[
f(\theta) = \frac{1}{2\pi} \left( A(e^{-i\theta}) \right)^{-1} \Sigma_\epsilon \left( A^*(e^{-i\theta}) \right)^{-1}
\]

(2.5)

The quantities \( \mu_{\min}(A), \mu_{\max}(A), \lambda_{\min}(\Sigma_\epsilon) \) and \( \lambda_{\max}(\Sigma_\epsilon) \) play important roles in our theoretical analysis. Essentially, these quantities determine some upper and lower bounds on the extreme eigenvalues of the spectral density matrix and help capture the effect of temporal and cross-sectional dependence on the VAR estimates. As seen in Figure 1, dependence in a VAR process comes from two distinct sources: (a) the transition matrices \( A_1, \ldots, A_d \) and (b) the error covariance \( \Sigma_\epsilon \). In time domain analysis it is difficult to separate the two sources. In frequency domain analysis, however, this is accomplished very easily, as seen in (2.5). This leads to tighter control of the dependence and sharper concentration bounds.

In Appendix D we provide some useful bounds on the stability measures and discuss their connection with the assumption \( \|A\| < 1 \).

### 2.1 Notation.

Based on the data \( \{X^0, \ldots, X^T\} \), we construct autoregression

\[
\begin{bmatrix}
(X^T)' \\
\vdots \\
(X^d)'
\end{bmatrix}_{Y} 
= 
\begin{bmatrix}
(X^{T-1})' & \cdots & (X^{T-d})'
\end{bmatrix}_{X} 
\begin{bmatrix}
\begin{array}{c}
A_1' \\
\vdots \\
A_d'
\end{array}
\end{bmatrix}_{B^*} 
\begin{bmatrix}
\begin{array}{c}
(\epsilon^T)'
\end{array}
\end{bmatrix}_{E}
\]

\[
vec(Y) = \vec(X B) + \vec(\epsilon) = (I \otimes X) \vec(B^*) + \vec(\epsilon)
\]

(2.6)
We will assume that $\beta^*$ is a $k$-sparse vector, i.e., $\sum_{t=1}^{d} \|A_t\|_0 = k$. We denote the $(i, j)^{th}$ entry of $\Sigma$ and $\Sigma^{-1}$ by $\sigma_{i,j}$ and $\sigma_{i,j}^\epsilon$, respectively. For any matrix $M$, we use $M_j$ to denote its $j^{th}$ column. We use $\|A\|$ to denote the $\ell_2$-norm when $A$ is a vector and the operator norm when $A$ is a matrix. Any other choice of norm is mentioned explicitly. We use $e_i$ to denote the $i^{th}$ unit vector in $\mathbb{R}^N$.

2.2 Estimates.

We consider two different estimates for the transition matrices $A_1, \ldots, A_d$, or equivalently, for $\beta^*$. The first one is a $\ell_1$-penalized least squares estimate of VAR coefficients ($\ell_1$-LS). It is defined as

$$\arg\min_{\beta \in \mathbb{R}^q} \frac{1}{N} \|Y - Z\beta\|^2 + \lambda_N \|\beta\|_1$$

(2.7)

This estimate does not exploit the error covariance structure $\Sigma_\epsilon$.

The second one uses an $\ell_1$-penalized log-likelihood estimation ($\ell_1$-LL). This is defined as

$$\arg\min_{\beta \in \mathbb{R}^q} \frac{1}{N} (Y - Z\beta)' (\Sigma_\epsilon^{-1} \otimes I) (Y - Z\beta) + \lambda_N \|\beta\|_1$$

(2.8)

This is a maximum likelihood estimate of $\beta$, assuming the error covariance $\Sigma_\epsilon$ is known. Although it is not directly applicable in practice, we study the properties of this estimate to demonstrate some important aspects related to the choice of loss function. In section 3, we implement this method with a plug-in estimate of $\Sigma_\epsilon$, obtained using the residuals from an ordinary least squares or $\ell_1$-LS fit.

A third estimate, $\ell_1$-penalized maximum-likelihood estimate of VAR coefficients ($\ell_1$-ML), can be formulated as

$$\arg\min_{\beta \in \mathbb{R}^q, \Sigma_\epsilon^{-1} \in \mathbb{R}^{p \times p}} \frac{1}{N} (Y - Z\beta)' (\Sigma_\epsilon^{-1} \otimes I) (Y - Z\beta) - \log \det(\Sigma_\epsilon^{-1}) + \lambda_N \|\beta\|_1 + P_{\gamma}(\Sigma_\epsilon^{-1})$$

(2.9)

It maximizes the full likelihood with respect to $(\beta, \Sigma_\epsilon^{-1})$ under an $\ell_1$-regularization on $\beta$ and some convex penalty $P_{\gamma}$ on $\Sigma_\epsilon^{-1}$. The second penalty can be chosen appropriately to incorporate some relevant structural assumption (sparsity, block structure etc.) on $\Sigma_\epsilon^{-1}$. Theoretical analysis of this estimate is more involved than the previous ones as it involves estimating an undirected graph from dependent samples of residuals. In this paper we do not study this estimate theoretically. In section 3 we show that a simple variant of our algorithm for $\ell_1$-LL can be used to implement this.

3 Implementation.

The optimization problem $\ell_1$-LS in (2.7) can be expressed as $p$ separate penalized regression problems:

$$\arg\min_{\beta \in \mathbb{R}^q} \frac{1}{N} \|Y - Z\beta\|^2 + \lambda_N \|\beta\|_1$$

$$\equiv \arg\min_{B_1, \ldots, B_p} \frac{1}{N} \sum_{i=1}^{p} \|Y_i - A'B_i\|^2 + \lambda_N \sum_{i=1}^{p} \|B_i\|_1$$

7
This amounts to running \( p \) separate lasso programs, each with \( dp \) predictors: \( Y_i \sim \mathcal{X}, \; i = 1, \ldots, p. \) For large \( d \) and \( p \), the \( p \) programs can be solved in parallel.

In the optimization problem \( \ell_1 \)-LL, the above regressions are coupled through \( \Sigma_{\epsilon}^{-1} \). One way to solve the problem, as mentioned in Davis et al. (2012), is to reformulate it into a single penalized regression problem:

\[
\min_{\beta \in \mathbb{R}^q} \frac{1}{N} \left( Y - Z\beta \right)^T \left( \Sigma_{\epsilon}^{-1} \otimes I \right) \left( Y - Z\beta \right) + \lambda_N \|\beta\|_1
\]

This amounts to running a single lasso program with \( dp^2 \) predictors:

\[
Y_i \sim \mathcal{X} \quad \text{for} \quad i = 1, \ldots, p.
\]

This is computationally expensive for large \( d \) and \( p \). Unlike \( \ell_1 \)-LL, this algorithm is not parallelizable.

We propose an alternative algorithm based on blockwise coordinate descent to estimate the \( \ell_1 \)-LL coefficients. To this end, we first observe that the objective function in (2.8) can be simplified to

\[
\frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{p} \sigma_{ij} \left( Y_i - \mathcal{X} \hat{B}_i \right)^T \left( Y_j - \mathcal{X} \hat{B}_j \right) + \lambda \sum_{k=1}^{p} \| \hat{B}_k \|_1
\]

Minimizing the above objective function cyclically with respect to each \( B_i \) leads to the following algorithm for \( \ell_1 \)-LL and \( \ell_1 \)-ML (the last step is used only for \( \ell_1 \)-ML):

1. pre-select \( d \). Run \( \ell_1 \)-LS to get \( \hat{B}, \hat{\Sigma}_{\epsilon} \).

2. iterate till convergence:

   (a) For \( i = 1, \ldots, p \),

   \[
   \text{set } r_i := \left( \frac{1}{2} \hat{\sigma}_{\epsilon}^{ii} \right) \sum_{j \neq i} \hat{\sigma}_{\epsilon}^{ij} \left( Y_j - \mathcal{X} \hat{B}_j \right)
   \]

   \[
   \text{update } \hat{B}_i = \arg\min_{\hat{B}_i} \hat{\sigma}_{\epsilon}^{ii} \left( Y_i + r_i \right) - \mathcal{X} \hat{B}_i \|_2^2 + \lambda_N \| \hat{B}_i \|_1
   \]

   (b) [only for \( \ell_1 \)-ML]: update \( \hat{\Sigma}_{\epsilon} \) as

\[
\arg\min_{\Sigma_{\epsilon}^{-1} \in \mathbb{R}^{p^2} \times \mathbb{R}^{p^2}} \frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{p} \sigma_{ij} \left( Y_i - \mathcal{X} \hat{B}_i \right)^T \left( Y_j - \mathcal{X} \hat{B}_j \right) - \log \det(\Sigma_{\epsilon}^{-1}) + P_\gamma(\Sigma_{\epsilon}^{-1}) \quad (3.1)
\]

In this algorithm, a single iteration amounts to running \( p \) separate lasso programs, each with \( dp \) predictors: \( Y_i + r_i \sim \mathcal{X} \), \( i = 1, \ldots, p \). As in \( \ell_1 \)-LS, these \( p \) programs can be solved in parallel. Further, for solving \( \ell_1 \)-ML, one can incorporate structural information about \( \Sigma_{\epsilon}^{-1} \) (sparsity, block structure, bandedness etc.) using the penalty \( P_\gamma(\cdot) \).

## 4 Theoretical Properties

In this section we analyze the theoretical properties of the two penalized estimators \( \ell_1 \)-LS and \( \ell_1 \)-LL, proposed in Section 2. We provide non-asymptotic upper bounds on their estimation and in-sample prediction errors under different choices of norms. First, for a given realization of
\{X^0, \ldots, X^T\}, we derive a deterministic upper bound on the errors under standard restricted eigenvalue and deviation conditions. We then show that for a random realization of \{X^0, \ldots, X^T\} generated according to a stable VAR(d) process, these conditions hold with high probability. The first result is known from the literature of high-dimensional linear regressions. The other results, viz., verifying the two regularity conditions for general VAR(d) models, have not been studied before.

We analyze the two problems (2.7) and (2.8) under a general penalized M-estimation framework proposed in Loh and Wainwright (2012). To motivate this general framework, note that the VAR estimation problem with ordinary least squares is equivalent to the following optimization

$$\arg\min_{\beta \in \mathbb{R}^q} -2\beta' \hat{\gamma} + \beta' \hat{\Gamma} \beta,$$

where \(\hat{\Gamma} = \left( I \otimes X'X/N \right) \), \(\hat{\gamma} = \left( I \otimes X' \right) Y/N \) are unbiased estimates for their population analogues.

A more general choice of \((\hat{\gamma}, \hat{\Gamma})\) in the penalized version of the objective function leads to the following optimization problem

$$\arg\min_{\beta \in \mathbb{R}^q} -2\beta' \hat{\gamma} + \beta' \hat{\Gamma} \beta + \lambda_N \|\beta\|_1,$$

where \(\hat{\Gamma} = \left( W \otimes X'X/N \right) \), \(\hat{\gamma} = \left( W \otimes X' \right) Y/N \)

where \(W\) is a symmetric, positive definite matrix of weights. The optimization problems (2.7) and (2.8) are special cases of (4.2) with \(W = I\) and \(W = \Sigma^{-1}\), respectively.

Consistent estimation of VAR models in the low-dimensional regime \((p, d \text{ finite, } T \rightarrow \infty)\) relies on two assumptions (cf. Section 3.2.2, Lütkepohl (2005)): (a) \(X'X/N\) converges in probability to a non-singular matrix \(C\) (i.e., \(\Lambda_{\min}(C) > 0\)) and (b) \(\text{vec}(X'\epsilon)/N\) converges to zero in probability. In the high-dimensional regime \((N \ll q)\) the first assumption is never true since the design matrix is rank-deficient (more variables than observations). The second assumption is also very stringent since the dimensions of \(X'\epsilon\) grows with \(N\).

Interestingly, consistent estimation in the high-dimensional regime can be guaranteed under two analogous sufficient conditions. The first one comes from a class of conditions commonly referred to as \textbf{Restricted Eigenvalue (RE)} conditions. Different variants of the RE condition have been proposed in the literature (Bickel et al., 2009; van de Geer and Bühlmann, 2009). Roughly speaking, in the regression framework of (2.6), these assumptions require that \(\|Z(\hat{\beta} - \beta^*)\|\) is small only when \(\|\hat{\beta} - \beta^*\|\) is small. If \(\hat{\beta}, \beta^*\) are any arbitrary vectors in \(\mathbb{R}^q\), this assumption is never true since the columns of \(Z\) are correlated. However, if \(\beta^*\) is sparse and \(\lambda_N\) is appropriately chosen, it is now well-understood that the vectors \(\hat{\beta} - \beta^*\) only vary on a low-dimensional subset of the high-dimensional space \(\mathbb{R}^q\), with high probability. This indicates that the RE condition may not be very stringent after all, even though the columns of \(Z\) are correlated. Although, verifying that the assumption is indeed satisfied with high probability, is often far from trivial.

We work with the RE condition proposed in Loh and Wainwright (2012). This is a slightly stronger version of the original RE assumption proposed in (Bickel et al., 2009), as discussed in Raskutti et al. (2010).

\textbf{(A1) Lower-Restricted Eigenvalue Assumption (Loh and Wainwright, 2012):} \(\hat{\Gamma}\) satisfies restricted eigenvalue condition with curvature \(\alpha > 0\) and tolerance \(\tau(N, q) > 0\) if

$$\theta' \hat{\Gamma} \theta \geq \alpha \|\theta\|^2 - \tau(N, q) \|\theta\|^2_1, \forall \theta \in \mathbb{R}^q$$

(4.3)
The second condition ensures that $\hat{\gamma}$ and $\hat{\Gamma}$ are well-behaved in the sense that they concentrate nicely around their population means. As $\hat{\gamma}$ and $\hat{\Gamma}\beta^*$ have the same expectation, this assumption requires an upper bound on their difference. Note that in the low-dimensional context of (4.1), $\hat{\gamma} - \hat{\Gamma}\beta^*$ is precisely $\text{vec}(X'E)/N.$

(A2) Deviation Condition: There exists a deterministic function $Q(\beta^*, \Sigma_\epsilon)$ such that

$$\left\| \hat{\gamma} - \hat{\Gamma}\beta^* \right\|_\infty \leq Q(\beta^*, \Sigma_\epsilon) \sqrt{\frac{\log q}{N}}$$ (4.4)

The following proposition establishes non-asymptotic upper bounds on the estimation and prediction errors when the above conditions are satisfied.

Proposition 4.1 (Estimation and prediction error). Consider the penalized M-estimation problem (4.2) with $W = I$ or $W = \Sigma_e^{-1}$. Suppose $\hat{\Gamma}$ satisfies lower RE condition (4.3) with $k\tau(N,q) \leq \alpha/32$ and $(\hat{\Gamma}, \hat{\gamma})$ satisfies the deviation bound (4.4). Then, for any $\lambda_N \geq 4Q(\beta^*, \Sigma_\epsilon)\sqrt{\log q/N}$, any solution $\hat{\beta}$ of (4.2) satisfies

$$\|\hat{\beta} - \beta^*\|_1 \leq 64k\lambda_N/\alpha$$ (4.5)
$$\|\hat{\beta} - \beta^*\| \leq 16\sqrt{k}\lambda_N/\alpha$$ (4.6)
$$(\hat{\beta} - \beta^*)'\hat{\Gamma}(\hat{\beta} - \beta^*) \leq 128k\lambda_N^2/\alpha$$ (4.7)

Remarks. $\|\hat{\beta} - \beta^*\|$ is precisely $\sum_{t=1}^d \|\hat{A}_t - A_t\|_F$, the $\ell_2$-error in estimating the transition matrices. For $\ell_1$-LS, $(\hat{\beta} - \beta^*)'\hat{\Gamma}(\hat{\beta} - \beta^*)$ is a measure of in-sample prediction error under $\ell_2$-norm, defined as $\sum_{t=1}^T \sum_{h=1}^d (\hat{A}_h - A_h)X^{t-h}\|/N$. For $\ell_1$-LL, $(\hat{\beta} - \beta^*)'\hat{\Gamma}(\hat{\beta} - \beta^*)$ takes the form $\sum_{t=1}^T \sum_{h=1}^d (\hat{A}_h - A_h)X^{t-h}\|/N$, where $\|v\|_\Sigma := \sqrt{v'\Sigma^{-1}v}$. This can be viewed as a measure of in-sample prediction error under a Mahalanobis type distance on $\mathbb{R}^p$ induced by $\Sigma_\epsilon$.

The convergence rates are governed by two sets of parameters: (i) dimensionality parameters - dimension of the process ($p$), order of the process ($d$), number of edges in the network ($k$) and sample size ($N = T-d+1$); (ii) intrinsic parameters - restricted eigenvalue(or curvature) ($\tau(N,q)$ and the deviation bound $Q(\beta^*, \Sigma_\epsilon)$). The squared $\ell_2$-errors of estimation and prediction scale with the dimensionality parameters as $k(2\log p + \log d)/N$, same as the rates obtained when the observations are independent (Bickel et al., 2009). The temporal and cross-sectional dependence affect the rates only through the intrinsic parameters. Typically, the rates are better when $\alpha$ is large and $Q(\beta^*, \Sigma_\epsilon)$, $\tau(N,q)$ are small. In propositions 4.2 and 4.3 we investigate in detail how these quantities are related to the dependence structure of the process.

Although the above proposition is derived under the assumption that $d$ is the true order of the VAR process, the results hold if $d$ is an upper bound on the true order. This follows from the fact that a VAR($d$) model can also be viewed as VAR($\hat{d}$), for any $\hat{d} > d$, with transition matrices $A_1, \ldots, A_d, 0_{p \times p}, \ldots, 0_{p \times p}$.

Proposition 4.1 is deterministic, i.e., it assumes a fixed realization of $\{X^0, \ldots, X^T\}$. To show that these error bounds hold with high probability, one needs to verify that the assumptions (A1-2) are satisfied with high probability when $\{X^0, \ldots, X^T\}$ is a random realization from the VAR($d$) process. This is accomplished in the next two propositions.

Verifying RE assumption. The problem of verifying RE-type assumptions for correlated design matrices has been studied by several authors (Raskutti et al., 2010; Rudelson and Zhou, 2013). These results require independence among the rows of the design matrix and are not directly applicable for VAR models. Verifying RE-type assumptions on a general design matrix where both the rows and columns are dependent is indeed a daunting task. To the best of our knowledge, the
closest results are due to Negahban and Wainwright (2011) and Loh and Wainwright (2012). However, their results crucially depend on the assumption $\|A\| < 1$ and cannot be extended beyond a subclass of stable VAR(1) models (Figure 2 and Lemma D.2).

We have dealt with this challenge by quantifying the unique dependence structure in VAR processes through its spectral representation. The results rely on a key technical lemma B.1. Conducting part of our analysis in the frequency domain, we were able to prove sharper concentration bounds that helped verify RE for any stable VAR(d) model, for $d \geq 1$. The techniques are fairly general and can be adapted for a broader class of stationary processes and different penalties other than lasso.

**Proposition 4.2 (Verifying RE for $\hat{\Gamma}$).** Consider a random realization $\{X^0, \ldots, X^T\}$ generated according to a stable VAR(d) process (2.1). Then there exist universal positive constants $c_i$ such that for all $N \gtrsim \max\{1, \omega^{-2}\} k \log(dp)$, with probability at least $1 - c_1 \exp(-c_2 N \min\{\omega^2, 1\})$, the matrix

$$\hat{\Gamma} = I_p \otimes (\mathcal{X}' \mathcal{X}/N) \sim RE(\alpha, \tau),$$

where

$$\omega = \frac{\Lambda_{\text{min}}(\Sigma_e) / \Lambda_{\text{max}}(\Sigma_e)}{\mu_{\text{max}}(A) / \mu_{\text{min}}(A)}, \quad \alpha = \frac{\Lambda_{\text{min}}(\Sigma_e)}{2 \mu_{\text{max}}(A)}, \quad \tau(N, q) = c_3 \frac{\Lambda_{\text{min}}(\Sigma_e)}{\mu_{\text{max}}(A)} \max\{\omega^{-2}, 1\} \frac{\log(dp)}{N}.$$ 

Further, if $\Sigma_e^{-1}$ satisfies $\bar{\sigma}_e^{ij} = \sigma_e^{ij} - \sum_{j \neq i} \sigma_e^{ij} > 0$, for $i = 1, \ldots, p$, then, with the same probability as above, the matrix

$$\hat{\Gamma} = \Sigma_e^{-1} \otimes (\mathcal{X}' \mathcal{X}/N) \sim RE \left( \alpha \min_i \bar{\sigma}_e^{ii}, \tau(N, q) \max_i \bar{\sigma}_e^{ii} \right)$$

This proposition provides insight into the effect of temporal and cross-sectional dependence on the convergence rates obtained in Proposition 4.1. As mentioned earlier, the convergence rates are faster for larger $\alpha$ and smaller $\tau$. From the expressions of $\omega$, $\alpha$ and $\tau$, it is clear that the VAR estimates have lower error bounds when $\Lambda_{\text{max}}(\Sigma_e), \mu_{\text{max}}(A)$ are smaller and $\Lambda_{\text{min}}(\Sigma_e), \mu_{\text{min}}(A)$ are larger. We defer the proof to Appendix B.

**Verifying Deviation Condition.** The problem of verifying the deviation condition poses some unique challenges that are specific to VAR models. In the literature of high-dimensional linear regression with random design, one typically assumes that the errors are independent of the design matrix. This does not hold for VAR models since $X^t$ is always correlated with $\epsilon^{t-1}$. In the classical, fixed $p$ analysis of VAR, this is resolved by invoking martingale central limit theorem on the coordinates of $X'E$. In the non-asymptotic, high-dimensional framework, a natural alternative is to come up with sharp martingale concentration inequalities. The standard martingale inequalities involve some measure of predictable quadratic variation (Bercu and Touati, 2008), which, in our context, corresponds to $\sum_{t=d-1}^{T-1} (X^t)'^2$. To establish a meaningful deviation bound, one needs to derive a second concentration bound for this term and somehow combine the two. Although viable, this approach often leads to poor and cumbersome deviation bounds.

Interestingly, we found that taking advantage of the spectral representation of VAR helps establish useful concentration bounds required to verify the deviation condition (A2). The main idea is to express the errors in terms of the design matrix and take advantage of the spectral representation of the autocovariance function. The proof relies crucially on a key technical lemma C.1. To the best of our knowledge, this is a novel technique to establish deviation inequalities for high-dimensional autoregressive processes.
Proposition 4.3 (Deviation Bound). For any $q \geq 2$, $A > 0$, $N \gtrsim \log q$, with probability at least $1 - 12q^{-A}$, we have

$$\| \hat{\gamma} - \hat{\Gamma} \beta^* \|_{\infty} \leq Q(\beta^*, \Sigma_e) \sqrt{\frac{\log q}{N}},$$

where, for $\ell_1$-LS,

$$Q(\beta^*, \Sigma_e) = (18 + 6\sqrt{2(A + 1)}) \left[ \Lambda_{\text{max}}(\Sigma_e) + \frac{\Lambda_{\text{max}}(\Sigma_e)}{\mu_{\text{min}}(A)} + \frac{\Lambda_{\text{max}}(\Sigma_e)\mu_{\text{max}}(A)}{\mu_{\text{min}}(A)} \right],$$

and for $\ell_1$-LL,

$$Q(\beta^*, \Sigma_e) = (18 + 6\sqrt{2(A + 1)}) \left[ \frac{1}{\Lambda_{\text{min}}(\Sigma_e)} + \frac{\Lambda_{\text{max}}(\Sigma_e)}{\mu_{\text{min}}(A)} + \frac{\Lambda_{\text{max}}(\Sigma_e)\mu_{\text{max}}(A)}{\mu_{\text{min}}(A)} \right].$$

As before, this proposition shows that the VAR estimates have lower error bounds when $\Lambda_{\text{max}}(\Sigma_e)$, $\mu_{\text{max}}(A)$ are smaller and $\Lambda_{\text{min}}(\Sigma_e)$, $\mu_{\text{min}}(A)$ are larger.

5 Numerical Experiments

We evaluate the performances of $\ell_1$-LS and $\ell_1$-LL on simulated data and compare with the performances of ordinary least squares (OLS) and Ridge estimates. Implementing $\ell_1$-LL requires an estimate of $\Sigma_e$ in the first step. For this, use the residuals from $\ell_1$-LS to construct a plug-in estimate $\hat{\Sigma}_e$. To evaluate the effect of error correlation on the transition matrix estimates more precisely, we also implement an oracle version, $\ell_1$-LL-O, which uses the true $\Sigma_e$ in the estimation. Next, we describe the simulation settings, choice of performance metrics and discuss the results.

We design two sets of numerical experiments - (a) SMALL VAR ($p = 10, d = 1, T = 30, 50$) and (b) MEDIUM VAR ($p = 30, d = 1, T = 80, 120, 160$). In each setting, we generate an adjacency matrix $A_1$ with $5 \sim 10\%$ non-zero edges selected at random and rescale to ensure that the process is stable with $SNR = 2$. We generate three different error processes with covariance matrix $\Sigma_e$ from one of the following families:

1. Block-I: $\Sigma_e = (\sigma_{e,ij})_{1 \leq i,j \leq p}$ with $\sigma_{e,ii} = 1$, $\sigma_{e,ij} = \rho$ if $1 \leq i \neq j \leq p/2$, 0 otherwise;
2. Block-II: $\Sigma_e = (\sigma_{e,ij})_{1 \leq i,j \leq p}$ with $\sigma_{e,ii} = 1$, $\sigma_{e,ij} = \rho$ if $1 \leq i \neq j \leq p/2$ or $p/2 < i \neq j \leq p$, 0 otherwise;
3. Toeplitz: $\Sigma_e = (\sigma_{e,ij})_{1 \leq i,j \leq p}$ with $\sigma_{e,ij} = \rho^{\mid i-j \mid}$.

We let $\rho$ vary in $\{0.5, 0.7, 0.9\}$. Larger values of $\rho$ indicate that the error processes are more strongly correlated. Figure 3 illustrates the structure of a random transition matrix used in our simulation and the three different types of error covariance structure.

We compare the different methods for VAR estimation (OLS, $\ell_1$-LS, $\ell_1$-LL, $\ell_1$-LL-O, Ridge) based on the following performance metrics:

1. Model Selection: Area under receiving operator characteristic curve (AUROC)
2. Estimation error: Relative estimation accuracy measured by $\| \hat{\beta} - B \|_F / \| B \|_F$

We report the results for small VAR with $T = 30$ and medium VAR with $T = 120$ (averaged over 50 replicates) in Tables 1 and 2. The results in the other settings are qualitatively similar,
Although the overall accuracy changes with the sample size. We find that the regularized VAR estimates outperform ordinary least squares uniformly in all the cases.

In terms of model selection, the $\ell_1$-penalized estimates perform fairly well, as reflected in their AUROC. Ordinary least squares and ridge regression do not perform any model selection. Further, for all three choices of error covariance, the two variants of $\ell_1$-LL outperform $\ell_1$-LS. The difference in their performances is more prominent for larger values of $\rho$. Among the three covariance structures, the difference between least squares and log-likelihood based methods is more prominent in Block-II and Toeplitz family since the error processes are more strongly correlated. Finally, in all the cases, the accuracy of $\ell_1$-LL is somewhere in between $\ell_1$-LS and $\ell_1$-LL-O, which suggests that a more accurate estimation of $\Sigma_\epsilon$ might improve the model selection performance of regularized VAR estimates.

In terms of $\ell_2$-error of estimation, the conclusions are broadly the same. The effect of overfitting is reflected in the performance of ordinary least squares. In many settings, the estimation error of ordinary least squares is even twice as large as the signal strength. The performance of ordinary least squares deteriorates when the error processes are more strongly correlated (see, for example, $\rho = 0.9$ for block-II). Ridge regression performs better than ordinary least squares as it applies shrinkage on the coefficients. However, the $\ell_1$-penalized estimates show higher accuracy than Ridge in almost all the cases. This is somewhat expected as the data were simulated from a sparse model with strong signals, whereas Ridge regression tend to favor a non-sparse model with many small coefficients.

![Adjacency matrix $A_1$ and error covariance matrix $\Sigma_\epsilon$ of different types used in the simulation studies](image)

**Figure 3:** Adjacency matrix $A_1$ and error covariance matrix $\Sigma_\epsilon$ of different types used in the simulation studies

**Table 1: VAR(1) model with $p = 10$**

| AUROC | BLOCK-I | BLOCK-II | Toeplitz |
|-------|---------|----------|----------|
| $p = 10$ | $T = 30$ | $\rho = 0.5$ | $\rho = 0.7$ | $\rho = 0.9$ | $\rho = 0.5$ | $\rho = 0.7$ | $\rho = 0.9$ |
| $\ell_1$-LS | 0.77 | 0.74 | 0.7 | 0.79 | 0.76 | 0.74 | 0.82 | 0.79 | 0.77 |
| $\ell_1$-LL | 0.77 | 0.75 | 0.73 | 0.79 | 0.77 | 0.77 | 0.81 | 0.8 | 0.81 |
| $\ell_1$-LL-O | 0.8 | 0.79 | 0.76 | 0.82 | 0.8 | 0.81 | 0.85 | 0.84 | 0.84 |
| **Estimation Error** | | | | | | | | | |
| OLS | 1.24 | 1.39 | 1.77 | 1.29 | 1.63 | 2.36 | 1.32 | 1.56 | 2.38 |
| $\ell_1$-LS | 0.68 | 0.72 | 0.76 | 0.64 | 0.67 | 0.7 | 0.63 | 0.66 | 0.69 |
| $\ell_1$-LL | 0.66 | 0.66 | 0.66 | 0.57 | 0.59 | 0.53 | 0.59 | 0.56 | 0.49 |
| $\ell_1$-LL-O | 0.61 | 0.62 | 0.62 | 0.53 | 0.54 | 0.47 | 0.53 | 0.51 | 0.42 |
| ridge | 0.72 | 0.74 | 0.75 | 0.7 | 0.71 | 0.72 | 0.7 | 0.71 | 0.72 |
Table 2: VAR(1) model with $p = 30$

| AUROC | $T = 120$ | BLOCK-I | BLOCK-II | Toeplitz |
|-------|-----------|---------|----------|----------|
|       | $\rho = 0.5$ | $\rho = 0.7$ | $\rho = 0.9$ | $\rho = 0.5$ | $\rho = 0.7$ | $\rho = 0.9$ |
| $\ell_1$-LS | 0.89 | 0.85 | 0.77 | 0.87 | 0.81 | 0.69 | 0.91 | 0.87 | 0.76 |
| $\ell_1$-LL | 0.89 | 0.87 | 0.82 | 0.9 | 0.89 | 0.88 | 0.91 | 0.91 | 0.89 |
| $\ell_1$-LL-O | 0.92 | 0.9 | 0.84 | 0.93 | 0.92 | 0.9 | 0.94 | 0.93 | 0.92 |
| OLS | 1.73 | 2 | 2.93 | 1.95 | 2.53 | 4.28 | 1.82 | 2.28 | 3.88 |
| $\ell_1$-LS | 0.72 | 0.76 | 0.85 | 0.74 | 0.82 | 0.93 | 0.69 | 0.73 | 0.86 |
| $\ell_1$-LL | 0.71 | 0.71 | 0.72 | 0.68 | 0.68 | 0.65 | 0.67 | 0.63 | 0.6 |
| $\ell_1$-LL-O | 0.66 | 0.66 | 0.68 | 0.64 | 0.63 | 0.59 | 0.63 | 0.59 | 0.54 |
| Ridge | 0.81 | 0.83 | 0.85 | 0.82 | 0.85 | 0.88 | 0.81 | 0.82 | 0.86 |

### A Results of Consistency

**Proof of Proposition 4.1.** Since $\hat{\beta}$ is a minimizer of (4.2), for all $\beta \in \mathbb{R}^q$ we have

$$-2\hat{\beta}' \hat{\gamma} + \hat{\beta}' \hat{\Gamma} \hat{\beta} + \lambda_N \|\hat{\beta}\|_1 \leq -2\beta' \hat{\gamma} + \beta' \hat{\Gamma} \beta + \lambda_N \|\beta\|_1$$

For $\beta = \beta^*$, the above inequality reduces to

$$\hat{v}' \hat{\Gamma} \hat{v} \leq 2\hat{\beta}' (\hat{\gamma} - \hat{\Gamma} \beta^*) + \lambda_N \{ \|\beta^*\|_1 - \|\beta^* + \hat{v}\|_1 \} \quad (A.1)$$

where $\hat{v} = \hat{\beta} - \beta^*$.

The first term on the right hand side of (A.1) can be upper bounded by $2\|\hat{v}\|_1 q(\beta^*, \Sigma_\epsilon) \sqrt{\log q/N}$. The second term, by triangle inequality, is at most $\lambda_N \{ \|\hat{v}_S\|_1 - \|\hat{v}_{S^c}\|_1 \}$. Together with the choice of $\lambda_N$, this leads to the following inequality

$$0 \leq \hat{v}' \hat{\Gamma} \hat{v} \leq \frac{\lambda_N}{2} \{ \|\hat{v}_S\|_1 + \|\hat{v}_{S^c}\|_1 \} + \lambda_N \{ \|\hat{v}_S\|_1 - \|\hat{v}_{S^c}\|_1 \}
\leq \frac{3\lambda_N}{2} \|\hat{v}_S\|_1 - \frac{\lambda_N}{2} \|\hat{v}_{S^c}\|_1 \leq 2\lambda_N \|\hat{v}\|_1 \quad (A.2)$$

In particular, this implies $\|\hat{v}_{S^c}\|_1 \leq 3\|\hat{v}_S\|_1$ so that $\|\hat{v}\|_1 \leq 4\|\hat{v}_S\|_1 \leq 4\sqrt{k} \|\hat{v}\|$. This, together with the restricted eigenvalue and the upper bound on $k\tau(N, q)$ implies

$$\hat{v}' \hat{\Gamma} \hat{v} \geq \alpha \|\hat{v}\|^2 - \tau(N, q) \|\hat{v}\|^2 \geq (\alpha - 16k\tau(N, q)) \|\hat{v}\|^2 \geq \frac{\alpha}{2} \|\hat{v}\|^2$$

Together, the upper and lower bounds on $\hat{v}' \hat{\Gamma} \hat{v}$ guarantee that

$$\frac{\alpha}{4} \|\hat{v}\|^2 \leq \lambda_N \|\hat{v}\|_1 \leq 4\sqrt{k} \lambda_N \|\hat{v}\|$$

This implies

$$\|\hat{v}\| \leq 16\sqrt{k} \lambda_N / \alpha$$

$$\|\hat{v}\|_1 \leq 4\sqrt{k} \lambda_N \|\hat{v}\| \leq 64k \lambda_N / \alpha$$

$$\hat{v}' \hat{\Gamma} \hat{v} \leq 2\lambda_N \|\hat{v}\|_1 \leq 128k \lambda_N / \alpha$$

\[ \square \]
B Results for Restricted Eigenvalue

Recall that $\hat{\Gamma} = W \otimes (\mathcal{A}' \mathcal{A}/N)$ where $W = I$ for $\ell_1$-LS and $W = \Sigma^{-1}_e$ for $\ell_1$-LL. Since $W$ is deterministic, we first verify RE assumption on the random matrix $\mathcal{A}' \mathcal{A}/N$ and then investigate the assumptions on $W$ under which the same holds for $\hat{\Gamma}$.

Note that $S = \mathcal{A}' \mathcal{A}/N$ is an estimate for $\hat{\Gamma}(0) = \text{Var}(\hat{X}^t)$, constructed using $N$ dependent samples $\{\hat{X}^{d-1}, \ldots, \hat{X}^{T-1}\}$. To establish RE on $S$, one needs to first ensure that $\Lambda_{\min}(\hat{\Gamma}(0)) > 0$. This is accomplished in Lemma B.1. The next step is to ensure that $S$ concentrates nicely around $\hat{\Gamma}(0)$ in an appropriate sense. This task is accomplished in Lemma B.3. In particular, we show that for any $v = \beta - \beta^* \in \mathbb{R}^{dp}$ with $\|v\| = 1$, the deviation $|v'(S - \hat{\Gamma}(0))v|$ is small with high probability. Note that this is considerably challenging due to the dependence among the rows of $\mathcal{A}$. In particular, we need a sharp control on the temporal decay of the VAR process. This follows from the upper bound presented in Lemma B.1. Once we have these two pieces, the proof follows along the line of Loh and Wainwright (2012). The main idea is to approximate the lower-dimensional space of possible $(\beta - \beta^*)$s (restricted to the unit ball) by convex combinations of sparse vectors and taking an union bound over the deviation inequality.

**Lemma B.1** (Eigenvalues of a block Toeplitz matrix). Consider a $p$-dimensional stable vector autoregressive process $\{X^t\}$ of order $d$ with characteristic polynomial $A(z) = I - \sum_{t=1}^{d} A_t z^t$ and error covariance $\Sigma_e$. Denote the covariance matrix of the $np$-dimensional random vector $\{(X^n)', (X^{n-1})', \ldots, (X^1)\}'$ by $\Gamma_n^A = [\Gamma(r-s)_{p \times p}]_{1 \leq r, s \leq n}$, where $\Gamma(h) = \mathbb{E} \left[ X^t (X^{t+h})' \right]$ is the autocovariance matrix of order $h$ of the process $X^t$. Then

$$\frac{\Lambda_{\min}(\Sigma_e)}{\mu_{\max}(A)} \leq \Lambda_{\min}(\Gamma_n^A) \leq \Lambda_{\max}(\Gamma_n^A) \leq \frac{\Lambda_{\max}(\Sigma_e)}{\mu_{\min}(A)}$$

**Proof.** We consider the cross spectral density of the VAR(d) process

$$f(\theta) = \frac{1}{2\pi} \sum_{l=-\infty}^{\infty} \Gamma(l) e^{-il\theta}, \quad \theta \in [-\pi, \pi] \quad (B.1)$$

From standard results of spectral theory it is known that $\Gamma(l) = \int_{-\pi}^{\pi} e^{il\theta} f(\theta) \, d\theta$, for every integer $l$. For any vector $x \in \mathbb{R}^{np}$ with $\|x\| = 1$, we want to provide upper and lower bounds on $x' \Gamma_n^A x$. We start with a fixed vector $x$, formed by stacking the $p$-tuples $x^1, \ldots, x^n$, i.e., $x' = ((x^1)', (x^2)', \ldots, (x^n)')'$, where each $x^t \in \mathbb{R}^p$.

For every $\theta \in [-\pi, \pi]$, define $G(\theta) = \sum_{t=1}^{n} x^t e^{-it\theta}$ and note that

$$\int_{-\pi}^{\pi} G^*(\theta) G(\theta) \, d\theta = \sum_{t=1}^{n} \sum_{\tau=1}^{n} (x^t)'(x^\tau) \int_{-\pi}^{\pi} e^{i(t-\tau)\theta} \, d\theta = 2\pi \sum_{t=1}^{n} \sum_{\tau=1}^{n} (x^t)'(x^\tau) 2\pi 1_{t=\tau} = 2\pi \|x\|^2 = 2\pi$$

15
and

\[
x' \Gamma_n^A x = \sum_{t=1}^{n} \sum_{\tau=1}^{n} (x_t')' \Gamma (t - \tau) x^\tau = \sum_{t=1}^{n} \sum_{\tau=1}^{n} (x_t')' \left( \int_{-\pi}^{\pi} e^{i(t-\tau)\theta} f (\theta) \, d\theta \right) x^\tau
\]

\[
= \int_{-\pi}^{\pi} \left( \sum_{t=1}^{n} (x_t') e^{i\tau \theta} \right) f (\theta) \left( \sum_{\tau=1}^{n} (x^\tau)' e^{-i\tau \theta} \right) \, d\theta = \int_{-\pi}^{\pi} G^* (\theta) f (\theta) G (\theta) \, d\theta
\]

Hence if \( m(\theta) \) and \( M(\theta) \) denote the minimum and maximum eigenvalues of the Hermitian matrix \( f(\theta) \), then from the above representations,

\[
2\pi \min_{\theta \in [-\pi, \pi]} m(\theta) \leq x' \Gamma_n^A x \leq 2\pi \max_{\theta \in [-\pi, \pi]} M(\theta)
\]

for all \( x \in \mathbb{R}^{np}, \| x \| = 1 \). So it remains to provide uniform bounds on the eigenvalues of \( f(\theta), \theta \in [-\pi, \pi] \). To this end, we note that the spectral density of a stable VAR(d) process has the closed form expression (cf. equation (9.4.23), Priestley (1981))

\[
f(\theta) = \frac{1}{2\pi} \left( A \left( e^{-i\theta} \right) \right)^{-1} \Sigma_\epsilon \left( A^* \left( e^{-i\theta} \right) \right)^{-1}
\]

(B.2)

This implies the following upper and lower bounds on \( M(\theta) \) and \( m(\theta) \), respectively, for all \( \theta \in [-\pi, \pi] \):

\[
M(\theta) = \Lambda_{\max} (f(\theta)) \leq \frac{1}{2\pi} \| \Sigma_\epsilon \| \left\| A \left( e^{-i\theta} \right)^{-1} \right\|^2 \leq \frac{1}{2\pi} \Lambda_{\max} (\Sigma_\epsilon) \mu_{\min} (A)
\]

\[
m(\theta) = \Lambda_{\min} (f(\theta)) \geq \frac{1}{2\pi} \left[ \| \left( A \left( e^{-i\theta} \right) \right)^{-1} \Sigma_\epsilon \left( A^* \left( e^{-i\theta} \right) \right)^{-1} \right]^{-1} \geq \frac{1}{2\pi} \Lambda_{\min} (\Sigma_\epsilon) \mu_{\max} (A)
\]

\[
\square
\]

**Lemma B.2** (Concentration of Gaussian random vectors). If \( Y_{n \times 1} \sim N(0_{n \times 1}, Q) \), then for all \( \eta > 2/\sqrt{n} \),

\[
\mathbb{P} \left[ \frac{1}{n} \| Y \|^2 - tr(Q) > 4\eta \| Q \| \right] \leq 2 \exp \left( -\frac{n}{2} \left( \eta - \frac{2}{\sqrt{n}} \right)^2 \right) + 2 \exp (-n/2)
\]

**Proof.** See Supplementary Lemma I.2 in Negahban and Wainwright (2011). \( \square \)

**Lemma B.3** (RE condition for \( \mathcal{X}' \mathcal{X} / N \)). Under the assumptions of Proposition 4.2, there exist universal positive constants \( c_i \) such that the matrix \( S = \mathcal{X}' \mathcal{X} / N \) satisfies RE\((\alpha, \tau)\) with probability \( 1 - c_1 \exp(-c_2 N \min \{ \omega^2, 1 \}) \) where

\[
\omega = \frac{\Lambda_{\min} (\Sigma_\epsilon) / \Lambda_{\max} (\Sigma_\epsilon)}{\mu_{\max} (A) / \mu_{\min} (A)}, \quad \alpha = \frac{\Lambda_{\min} (\Sigma_\epsilon)}{2 \mu_{\max} (A)}, \quad \tau(N, q) = \frac{c_3 \Lambda_{\min} (\Sigma_\epsilon)}{\mu_{\max} (A)} \max \{ \omega^{-2}, 1 \} \frac{\log(dp/k)}{N}
\]

**Proof.** Every row of \( \mathcal{X}_{N \times dp} \) is a \( dp \)-dimensional zero mean Gaussian random vector with covariance matrix \( \Gamma_n^A = \Gamma_d^A \). We first note that, by Lemma B.1, \( \Lambda_{\min} (\Sigma_\epsilon) / \mu_{\max} (A) \) is a lower bound for \( \Lambda_{\min} (\Gamma_d^A) = \Lambda_{\min} (\Gamma_1^A) = \Lambda_{\min} (\Sigma_\epsilon) \).

16
Next, we show that for any $v \in \mathbb{R}^{dp}$ with $\|v\| = 1$, and any $\eta > 2/\sqrt{N}$,
\[
\mathbb{P} \left[ |u' \left( S - \Gamma \hat{A} \right) v | > 4\eta \Lambda_{\max} \left( \Gamma \hat{A} \right) \right] \leq 2 \exp \left( -\frac{N}{2} (\eta - 2/\sqrt{N})^2 \right) + 2 \exp (-N/2) \tag{B.3}
\]

To prove (B.3), note that $v' \left( S - \Gamma \hat{A} \right) v = (1/N)((\mathcal{X}v)'(\mathcal{X}v) - Nv'\Gamma \hat{A}v)$. Define $Y = \mathcal{X}v \in \mathbb{R}^N$. Then $Y$ is a zero mean normal random vector with covariance matrix $Q = ((Q_{ij}))$, where $Q_{ij} = \text{cov}(v'\hat{X}^{T-i}, v'\hat{X}^{T-j}) = v'\hat{\Gamma}(i-j)v$. Also $\text{tr}(Q) = Nv'\hat{\Gamma}(0)v = Nv'\Gamma \hat{A}v$. To apply Lemma (B.2) we need an upper bound on $Q_{ij}$.

The next step is to extend the deviation bound (B.3) for a single $v$ to an appropriate set of sparse vectors $\mathcal{K}(2s) := \{v \in \mathbb{R}^{dp}||v|| \leq 1, \|v\|_0 \leq 2s\}$, for any $s \geq 1$. A discretization argument along the line of Lemma 15 in Loh and Wainwright (2012) leads to the following deviation bound:

\[
\mathbb{P} \left[ \sup_{v \in \mathcal{K}(2s)} |v' \left( S - \Gamma \hat{A} \right) v | > 4\eta \Lambda_{\max} \left( \Gamma \hat{A} \right) \right] \leq 4 \exp \left( -\frac{N}{2} \min \left\{ 1, \left( \frac{2\nu}{9} - \frac{2}{\sqrt{N}} \right)^2 \right\} + 2s \log (9dp) \right)
\]

Next, we apply Lemma 12 in the above paper with $\delta = \Lambda_{\min}(\Sigma_{\epsilon}) / 54\mu_{\max}(A)$ and $\Gamma = S - \hat{\Gamma}(0)$ to show that

\[
v' S v \geq \alpha \|v\|^2 - \frac{\alpha}{s} \|v\|_1^2, \quad \text{for all } v \in \mathbb{R}^{dp}
\]

with probability at least $1 - 4 \exp \left[ -\frac{N}{2} \min \left\{ 1, \left( c_0 \omega - \frac{2}{\sqrt{N}} \right)^2 \right\} + 2s \log (9dp) \right]$, for some constant $c_0$. With the proposed choice of $N$, the proof then follows by setting $s := N \min \{1, \omega^2 / c \log (9dp)\}$, where $c$ is a constant, chosen sufficiently large to ensure $s \geq 1$.

\[\square\]

**Lemma B.4** (RE condition for $\hat{\Gamma}$). If $\mathcal{X}'\mathcal{X}/N \sim RE(\alpha, \tau)$, then so does $I_p \otimes \mathcal{X}'\mathcal{X}/N$. Further, if $\Sigma_{\epsilon}^{-1}$ is strictly diagonally dominant with $\bar{\sigma}_i^\epsilon := \sigma_i^\epsilon - \sum_{j \neq i} \sigma_j^\epsilon$, for $i = 1, \ldots, p$, then

\[
\Sigma_{\epsilon}^{-1} \otimes \mathcal{X}'\mathcal{X}/N \sim RE \left( \alpha \min \limits_i \bar{\sigma}_i^\epsilon, \tau \max \limits_i \bar{\sigma}_i^\epsilon \right)
\]

**Proof.** $S = \mathcal{X}'\mathcal{X}/N \sim RE(\alpha, \tau)$. Consider $\hat{\Gamma} = I_p \otimes S$. For any $\theta \in \mathbb{R}^{dp^2}$ with $\theta' = (\theta_1', \ldots, \theta_p')'$, each $\theta_i \in \mathbb{R}^{dp}$, we have

\[
\theta'(I_p \otimes S)\theta = \sum_{r=1}^{p} \theta'_r S \theta_r \geq \alpha \sum_{r=1}^{p} \|\theta_r\|^2 - \tau \sum_{r=1}^{p} \|\theta_r\|_1^2 \geq \alpha \|\theta\|^2 - \tau \|\theta\|_1^2
\]

proving the first part. To prove the second part, note that

\[
\theta' (\Sigma_{\epsilon}^{-1} \otimes S) \theta = \sum_{r,s=1}^{p} \sigma_r^\epsilon \sigma_s^\epsilon \theta'_r S \theta_s = \sum_{r=1}^{p} \sigma_r^\epsilon \theta'_r S \theta_r + \sum_{r \neq s} \sigma_r^\epsilon \sigma_s^\epsilon \theta'_r S \theta_s
\]
Since the matrix $S$ is non-negative definite, $\theta_0^T S \theta_s \geq -\frac{1}{2}(\theta_0^T S \theta_r + \theta_0^T S \theta_s)$ for every $r \neq s$. This implies

$$\theta_0^T (\Sigma^{-1}_\varepsilon \otimes S) \theta \geq \sum_{r=1}^{p} \sigma_{\varepsilon}^{rr} \theta_0^T S \theta_r - \sum_{r<s}^{p} \sigma_{\varepsilon}^{rs} (\theta_0^T S \theta_r + \theta_0^T S \theta_s)$$

$$= \sum_{r=1}^{p} \left( \sigma_{\varepsilon}^{rr} - \sum_{r \neq s}^{p} \sigma_{\varepsilon}^{rs} \right) \theta_0^T S \theta_r = \sum_{r=1}^{p} \sigma_{\varepsilon}^{rr} \theta_0^T S \theta_r$$

$$\geq \alpha \sum_{r=1}^{p} \sigma_{\varepsilon}^{rr} \|\theta_r\|^2 - \tau \sum_{r=1}^{p} \sigma_{\varepsilon}^{rr} \|\theta_r\|^2_1$$

$$\geq \left( \alpha \min_i \sigma_{\varepsilon}^i \right) \|\theta\|^2 - \left( \tau \max_i \sigma_{\varepsilon}^i \right) \|\theta\|^2_1$$

Proof of Proposition 4.2. Follows from Lemma B.3 and Lemma B.4.

C Results on Deviation Bound

Proof of Proposition 4.3. We want to establish an upper bound on $\|\hat{\gamma} - \hat{\Gamma} \beta^*\|_\infty$ that holds with high probability. To this end, we first note that in the context of (4.2),

$$\hat{\gamma} = (W \otimes \mathcal{X}') (I_p \otimes \mathcal{X}) \beta^*/N + (W \otimes \mathcal{X}') \text{vec}(E)/N$$

$$\hat{\Gamma} \beta^* = (W \otimes \mathcal{X}' \mathcal{X}/N) \beta^*$$

which implies $\hat{\gamma} - \hat{\Gamma} \beta^* = (W \otimes \mathcal{X}') \text{vec}(E)/N = \text{vec}(\mathcal{X}'E/W)/N$.

For $\ell_1$-LS, $EW = E$, a matrix with independent rows, each row $\sim N(0, \Sigma_\varepsilon)$. For $\ell_1$-LL, $EW = E$, a matrix with independent rows, each row $\sim N(0, \Sigma_\varepsilon^{-1})$. Note that in both cases $i^{th}$ row of $E$ or $\hat{E}$ is independent of the $i^{th}$ row of $\mathcal{X}$. First, we present the argument for $\ell_1$-LS.

For any $l \in \{1, \ldots, dp\}$, any $k \in \{1, \ldots, p\}$, we first derive an upper bound on $\mathbb{P}(\|\mathcal{X}_l' E_k/N\| > t)$. Taking union bound over all $l$, $k$ then leads to the final upper bound on the tail probability of the maximum

$$\mathbb{P}\left( \frac{1}{N} \|\mathcal{X}' E\|_{\max} > t \right) = \mathbb{P}\left( \max_{1 \leq t \leq dp, 1 \leq k \leq p} \frac{1}{N} \|\mathcal{X}_l' E_k\| > t \right)$$

Note that for any given $l$, $1 \leq l \leq dp$, there exist unique $j$, $h > 0$ such that $l = p(h - 1) + j$, $1 \leq h \leq d$, $1 \leq j \leq p$. The $l^{th}$ column of $\mathcal{X}$ and the $k^{th}$ column of $E$ are precisely

$$\mathcal{X}(j, h) := \mathcal{X}_l = \begin{bmatrix} \mathcal{X}_j^{T-h} \\ \mathcal{X}_j^{T-1-h} \\ \vdots \\ \mathcal{X}_j^{d-h} \end{bmatrix}, \quad E_k = \begin{bmatrix} \epsilon_k^T \\ \epsilon_k^{T-1} \\ \vdots \\ \epsilon_k^d \end{bmatrix}$$

We will use $\mathcal{X}(j, h)$ and $\mathcal{X}_l$ interchangeably for notational convenience.

$$\frac{1}{N} \|\mathcal{X}(j, h)' E_k\| \leq \frac{1}{2N} \left[ \|\mathcal{X}(j, h) + E_k\|^2 - \|\mathcal{X}(j, h)\|^2 - \|E_k\|^2 \right]$$
So
\[
\mathbb{P}\left(\frac{1}{N}|\mathcal{X}(j, h)'E_k| > t\right) = \mathbb{P}\left(\frac{1}{N}(\mathcal{X}(j, h)' + E_k)'(\mathcal{X}(j, h)' + E_k) - N \left(\text{Var}(X_j^{T-h}) + \text{Var}(\epsilon_k^T)\right) > \frac{2t}{3}\right) + \mathbb{P}\left(\frac{1}{N}|\mathcal{X}(j, h)'\mathcal{X}(j, h) - N \text{Var}(X_j^{T-h})| > \frac{2t}{3}\right) + \mathbb{P}\left(\frac{1}{N}|E_k'E_k - N \text{Var}(\epsilon_k^T)| > \frac{2t}{3}\right)
\]
(C.1)

So it suffices to concentrate each of the three terms on the right around their population means.

**Term III:** \(E_k \sim N(0, Q)\) with \(Q = (\epsilon_k^T \Sigma e_k)I_N\), so that \(\|Q\| \leq \Lambda_{\text{max}}(\Sigma_e)\). So, by Lemma B.2 with \(\eta_3 = t/6\Lambda_{\text{max}}(\Sigma_e)\), we have
\[
\mathbb{P}\left(\frac{1}{N}|E_k'E_k - N \text{Var}(\epsilon_k^T)| > \frac{2t}{3}\right) \leq 4 \exp\left[-\frac{N}{2} \min\left\{1, \left(\eta_3 - \frac{2}{\sqrt{N}}\right)^2\right\}\right]
\]

**Term II:** \(\mathcal{X}(j, h) \sim N(0, Q)\) with \(Q_{rs} = \epsilon_j^T \Gamma(r - s)e_j\), so that \(\text{tr}(Q) = N \epsilon_j^T \Gamma(0)e_j = N\text{Var}(X_j^{T-h})\), and
\[
\|Q\| = \sup_{\|u\|=1} u'Qu = \sum_{r=1}^N \sum_{s=1}^N u_r u_s \epsilon_j^T \Gamma(r - s)e_j = (u \otimes e_j)' \Gamma_A(N) (u \otimes e_j) \leq \Lambda_{\text{max}}(\Gamma_A) \leq \Lambda_{\text{max}}(\Sigma_e)/\mu_{\text{min}}(A)
\]
Again, by Lemma B.2 with \(\eta_2 = t\mu_{\text{min}}(A)/6\Lambda_{\text{max}}(\Sigma_e)\), we have
\[
\mathbb{P}\left(\frac{1}{N}|\mathcal{X}(j, h)'\mathcal{X}(j, h) - N \text{Var}(X_j^{T-h})| > \frac{2t}{3}\right) \leq 4 \exp\left[-\frac{N}{2} \min\left\{1, \left(\eta_2 - \frac{2}{\sqrt{N}}\right)^2\right\}\right]
\]

**Term I:** The \(N\)-dimensional random vector \(\mathcal{X}(j, h) + E_k\) is centered Gaussian with covariance matrix \(Q\), where
\[
Q_{rs} = \text{Cov}(X_j^{T-h-r+1}, X_j^{T-h-s+1}) + \epsilon_k^T(r - s)e_j + \epsilon_j^T \Delta(r, s)e_j + 1_{(r=s)} \epsilon_k^T \Sigma e_k + \epsilon_k^T \epsilon_k
\]
where \(\Delta(r, s)_{p \times p}\) is the \((r, s)^{th}\) block of the covariance matrix
\[
\Delta := \text{Cov}\left(\begin{bmatrix} X^{T-h} \\ X^{T-1-h} \\ \vdots \\ X^{d-h} \end{bmatrix}, \begin{bmatrix} \epsilon^T \\ \epsilon^{T-1} \\ \vdots \\ \epsilon^d \end{bmatrix}\right), \quad \Delta(r, s) = \text{Cov}(X^{T-h-r+1}, \epsilon^{T-s+1})
\] (C.2)
This implies that for any \( u \in \mathbb{R}^N, ||u|| = 1, \)

\[
u'Qu = \sum_{r=1}^{N} \sum_{s=1}^{N} u_r u_s Q_{rs} = (u \otimes e_j)' \Gamma_A^N (u \otimes e_j) + (u \otimes e_j)' \Delta (u \otimes e_k) + e_k' \Sigma e_k
\]

Since \( ||u \otimes e_j|| = ||u \otimes e_k|| = 1, \) it follows from the upper bounds in Lemma C.1 and Lemma B.1 that

\[
||Q|| \leq \Lambda_{\max}(\Sigma_e) [1 + (1 + \mu_{\max}(A))/\mu_{\min}(A)]
\]

Once again, using Lemma B.2 with \( \eta_1 = t [1 + (1 + \mu_{\max}(A))/\mu_{\min}(A)]^{-1}/6\Lambda_{\max}(\Sigma_e), \) we have

\[
\mathbb{P} \left( \frac{1}{N} \left| (\mathcal{X}(j, h) + E_k)'(\mathcal{X}(j, h) + E_k) - N \left( \text{Var}(X_j^{T-h}) + \text{Var}(\epsilon_k^T) \right) \right| > \frac{2t}{3} \right) \leq 4 \exp \left[ \left( -\frac{N}{2} \right) \min \left\{ 1, \left( \eta_1 - \frac{2}{\sqrt{N}} \right)^2 \right\} \right]
\]

Note that \( \eta_1 = \min\{\eta_1, \eta_2, \eta_3\}. \) Combining the three tail bounds from terms I - III and taking an union bound over all \( l, k, 1 \leq l \leq dp, 1 \leq k \leq p, \) we get

\[
\mathbb{P} \left( \max_{l,k} \frac{1}{N} |\Lambda_e' E_k| > t \right) \leq 12 \exp \left[ -\frac{N}{2} \min \left\{ 1, \left( \eta_1 - \frac{2}{\sqrt{N}} \right)^2 \right\} + \log q \right]
\]

Setting \( t = Q(\beta^*, \Sigma_e) \sqrt{\log q/N} \) with

\[
Q(\beta^*, \Sigma_e) = (18 + 6\sqrt{2(A + 1)})\Lambda_{\max}(\Sigma_e) [1 + (1 + \mu_{\max}(A))/\mu_{\min}(A)]
\]

for some \( A > 0, \) we have the required result for \( \ell_1-\text{LS}. \)

The proof for \( \ell_1-\text{LL} \) can be derived exactly along the same line. For term III, we have \( ||Q|| \leq 1/\Lambda_{\min}(\Sigma_e). \) For term II, \( ||Q|| \) remains the same. For term I,

\[
||Q|| \leq \frac{1}{\Lambda_{\min}(\Sigma_e)} + \frac{\Lambda_{\max}(\Sigma_e)}{\mu_{\min}(A)} + \frac{\Lambda_{\max}(\Sigma_e) \mu_{\max}(A)}{\Lambda_{\min}(\Sigma_e) \mu_{\min}(A)}
\]

The additional \( \Lambda_{\min}(\Sigma_e) \) in the denominator of the third expression appears because the \( (r, s)^{th} \) block of \( \Delta \) in lemma C.1 now changes to

\[
\Delta(r, s) = \text{Cov}(X^{T-h-r+1}, \Sigma_e^{-1} \epsilon^T) = \left[ (\Gamma(r-s+h) - \ldots - \Gamma(r-s+h-d) A_d') \right] \Sigma_e^{-1}
\]

As before, setting \( \eta_1, \eta_2, \eta_3 \) with the modified bounds, we get the result. \( \square \)

**Lemma C.1 (Bounding \( u'\Delta v \)).** Consider \( \Delta, as defined in (C.2). For any \( u, v \in \mathbb{R}^{Np} \) with \( ||u|| = ||v|| = 1, ||u'Qv|| \leq \Lambda_{\max}(\Sigma_e) \mu_{\max}(A)/\mu_{\min}(A). \)
This leads to the following upper bound on the quadratic form.

By Cauchy-Schwarz inequality,

\[ u' \Delta v = \sum_{r,s=1}^{N} (u^r)(v^s) \left( \int_{-\pi}^{\pi} f(\theta)e^{i(r-s+h)\theta}A^*(e^{i\theta})d\theta \right) \]

\[ \leq \int_{-\pi}^{\pi} \left[ \sum_{r=1}^{N} (u^r)^2 \int_{-\pi}^{\pi} f(\theta)e^{i\theta}e^{i\theta}A^*(e^{i\theta})d\theta \right] \sum_{s=1}^{N} (v^s)^2 e^{-i\theta}d\theta \]

\[ = \int_{-\pi}^{\pi} G^*(\theta)f(\theta)e^{i\theta}A^*(e^{i\theta})H(\theta) d\theta \]

By Cauchy-Schwarz inequality,

\[ \left| \int_{-\pi}^{\pi} G^*(\theta)I(\theta) d\theta \right| \leq \left( \int_{-\pi}^{\pi} G^*(\theta)G(\theta) d\theta \right)^{1/2} \left( \int_{-\pi}^{\pi} I^*(\theta)I(\theta) d\theta \right)^{1/2} \]

This leads to the following upper bound on the quadratic form.

\[ \left| u' \Delta v \right| \leq \left( \int_{-\pi}^{\pi} G^*(\theta)G(\theta) d\theta \right)^{1/2} \left( \int_{-\pi}^{\pi} H^*(\theta)A(e^{i\theta})f^2(\theta)A^*(e^{i\theta})H(\theta) d\theta \right)^{1/2} \]

\[ \leq 2\pi \max_{\theta \in [-\pi, \pi]} \Lambda_{\max}^{1/2} \left( A(e^{i\theta})f^2(\theta)A^*(e^{i\theta}) \right) \]

\[ \leq \frac{\Lambda_{\max}(\Sigma_d)\mu_{\max}(A)}{\mu_{\min}(A)} \]

where the last inequality follows from the expression of \( f(\theta) \) in (B.2). \( \square \)

### D Results on \( \mu_{\min}(A) \), \( \mu_{\max}(A) \) and \( \|A_1\| \)

In this section we discuss some properties of the stability measures introduced in Section 2 and their connection with the assumption \( \|A_1\| < 1 \). In particular, we show that the assumption \( \|A_1\| < 1 \) guarantees stability of the process, but not the other way. If, however, the transition matrix \( A_1 \) is symmetric, the assumption \( \|A_1\| < 1 \) is necessary for stability. We also show that the contraction principle is violated for all VAR(d) models, whenever \( d > 1 \). Finally, we provide an upper bound on \( \mu_{\max} \) for any VAR(d) and a lower bound on \( \mu_{\min} \) for VAR(1), when \( A_1 \) has distinct eigenvalues.

The first lemma shows that \( \|A_1\| \leq 1 \) whenever \( d > 1 \).

**Lemma D.1.** \( \|A_1\| \leq 1 \) whenever \( d > 1 \).
Proof.

\[
\tilde{A}_1 \tilde{A}'_1 = \begin{bmatrix}
\sum_{t=1}^{d} A_t A'_t & A_1 & \cdots & A_{d-1} \\
A'_1 & I_p & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
A'_{d-1} & 0 & \cdots & I_p
\end{bmatrix}_{dp \times dp}
\]

So for any \( v \in \mathbb{R}^{dp} \) with \( v' = (v_1', \ldots, v_d') \), each \( v_t \in \mathbb{R}^p \), we have

\[
v' \tilde{A}_1 \tilde{A}'_1 v = v'_1 \left( \sum_{t=1}^{d} A_t A'_t \right) v_1 + 2v'_1 \sum_{t=2}^{d} A_{t-1} v_t + \sum_{t=2}^{d} v_t^2 \]

This implies

\[
\Lambda_{\text{max}} \left( \tilde{A}_1 \tilde{A}'_1 \right) = \max_{\|v\|=1} v' \tilde{A}_1 \tilde{A}'_1 v \geq \max_{\|v\|=1} v' \tilde{A}_1 \tilde{A}'_1 v = \max_{\|v\|=1} \sum_{t=2}^{d} v_t^2 = 1
\]

Lemma D.2. A VAR(1) process is stable if \( \|A_1\| < 1 \). If \( A_1 \) is symmetric, then a VAR(1) process is stable only if \( \|A_1\| < 1 \).

Proof. If \( \|A_1\| < 1 \), then all the eigenvalues of \( A_1 \) lie inside the unit circle \( \{ z \in \mathbb{C} : |z| \leq 1 \} \). So the process is stable.

If the process is stable, then all the eigenvalues of \( A_1 \) lie inside the unit circle. In addition, if \( A_1 \) is symmetric, then this implies that \( \|A_1\| = \sqrt{\Lambda_{\text{max}}(A_1'A_1)} \leq 1 \).

Lemma D.3. For any VAR(d) process, \( \mu_{\text{max}}(A) \leq 1 + (v_{\text{in}} + v_{\text{out}})/2 \), where

\[
v_{\text{in}} = \sum_{h=1}^{d} \max_{1 \leq i \leq p} \sum_{j=1}^{p} |A_{h,ij}|, \quad v_{\text{out}} = \sum_{h=1}^{d} \max_{1 \leq j \leq p} \sum_{i=1}^{p} |A_{h,ij}|
\]

are the maximum absolute incoming and outgoing effects at a node, respectively.

Proof.

\[
\mu_{\text{max}}(A) = \max_{|z|=1} \left\| I - A_1 z - \cdots - A_d z^d \right\|
\]

\[
\leq 1 + \sum_{h=1}^{d} \|A_h\| \leq 1 + \sum_{h=1}^{d} \sqrt{\|A_h\|_1 \|A_h\|_\infty}
\]

\[
\leq 1 + \sum_{h=1}^{d} \left( \max_{1 \leq i \leq p} \sum_{j=1}^{p} |A_{h,ij}| + \max_{1 \leq j \leq p} \sum_{i=1}^{p} |A_{h,ij}| \right) / 2
\]

Lemma D.4. Consider a stable VAR(1) process \( X_t = A_1 X_{t-1} + \epsilon_t \), and assume that the eigenvalues of \( A_1 \) are distinct. Let \( A_1 = PDP^{-1} \) denote the Jordan form of \( A_1 \) and \( \rho(A_1) \) denote its
spectral radius, i.e., the maximum absolute eigenvalue of $A_1$. Then $\mu_{\min}(A) \geq \kappa_P^{-2}(1 - \rho(A_1))^2$, where $\kappa_P$ denotes the condition number of $P$, defined as $\|P\|\|P^{-1}\|$.

Proof.

$$\mu_{\min}(A) = \min_{|z|=1} \Lambda_{\min}((I - A_1z)^*(I - A_1z)) = \min_{|z|=1} \Lambda_{\min}((zI - A_1)^*(zI - A_1))$$

Let $\lambda_1, \ldots, \lambda_r$ denote the distinct eigenvalues of $A_1$ with multiplicities $m_1, \ldots, m_r$. The Jordan normal form of $A_1$ is $A_1 = PDP^{-1}$, where $D$ is a block diagonal matrix with diagonal blocks $D_1, \ldots, D_r$ given by:

$$D_j = \begin{bmatrix}
\lambda_j & 1 & 0 & \cdots & 0 \\
0 & \lambda_j & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & \lambda_j & 1 \\
0 & 0 & \cdots & 0 & \lambda_j
\end{bmatrix}_{m_j \times m_j}$$

This implies that for any $z \in \mathbb{C}$ with $|z| = 1$, $zI - A$ has a Jordan normal form $PD_zP^{-1}$, where the $j^{th}$ diagonal block of $D_z$ has $z - \lambda_j$ on the main diagonal and $-1$ on the superdiagonal. This leads to the following lower bound

$$\Lambda_{\min}((zI - A_1)^*(zI - A_1)) = \Lambda_{\min}((P^{-1})'D_z^*PD_zP^{-1}) \geq \kappa_P^{-2}\Lambda_{\min}(D_z^*D_z)$$

where $\kappa_P$ is the condition number of $P$. So it is enough to find a lower bound on the minimum eigenvalues of each of the diagonal blocks of $D_z$. The $j^{th}$ diagonal block of $D_z^*D_z$ can be expressed as

$$H_j^z = \begin{bmatrix}
|z - \lambda_j|^2 & -(z - \lambda_j)^* & 0 & 0 & \cdots & 0 \\
-(z - \lambda_j)^* & 1 + |z - \lambda_j|^2 & -(z - \lambda_j)^* & 0 & \cdots & 0 \\
0 & -(z - \lambda_j) & 1 + |z - \lambda_j|^2 & -(z - \lambda_j)^* & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & -(z - \lambda_j) & 1 + |z - \lambda_j|^2 & -(z - \lambda_j)^* \\
0 & 0 & \cdots & 0 & -(z - \lambda_j) & 1 + |z - \lambda_j|^2
\end{bmatrix}_{m_j \times m_j}$$

If $m_j = 1$, $\Lambda_{\min}(H_j^z) \geq (1 - |\lambda_j|)^2$, for any $z \in \mathbb{C}$ with $|z| = 1$.

If the eigenvalues of $A_1$ are distinct, then $m_j = 1$, for all $j$. The proof follows from the fact that $\rho(A_1) = \max_j |\lambda_j|$.

□

References

Bańbura, M., Giannone, D., and Reichlin, L. (2010), “Large Bayesian vector auto regressions,” *Journal of Applied Econometrics*, 25, 71–92.

Bercu, B. and Touati, A. (2008), “Exponential inequalities for self-normalized martingales with applications,” *Ann. Appl. Probab.*, 18, 1848–1869.
Bickel, P., Ritov, Y., and Tsybakov, A. (2009), “Simultaneous analysis of Lasso and Dantzig selector,” *The Annals of Statistics*, 37, 1705–1732.

Christiano, L. J., Eichenbaum, M., and Evans, C. L. (1999), “Monetary policy shocks: What have we learned and to what end?” in *Handbook of Macroeconomics*, eds. Taylor, J. B. and Woodford, M., Elsevier, vol. 1 of *Handbook of Macroeconomics*, chap. 2, pp. 65–148.

Davis, R. A., Zang, P., and Zheng, T. (2012), “Sparse Vector Autoregressive Modeling,” *ArXiv e-prints*.

De Mol, C., Giannone, D., and Reichlin, L. (2008), “Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?” *Journal of Econometrics*, 146, 318–328.

Kock, A. and Callot, L. (2012), “Oracle Inequalities for High Dimensional Vector Autoregressions,” .

Loh, P.-L. and Wainwright, M. J. (2012), “High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity.” *Ann. Stat.*, 40, 1637–1664.

Lütkepohl, H. (2005), *New introduction to multiple time series analysis*, Springer.

Michailidis, G. and d’Alché Buc, F. (2013), “Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues,” *Mathematical Biosciences*, in press.

Negahban, S. and Wainwright, M. J. (2011), “Estimation of (near) low-rank matrices with noise and high-dimensional scaling,” *Ann. Statist.*, 39, 1069–1097.

Priestley, M. B. (1981), *Spectral analysis and time series. Vol. 2*, London: Academic Press Inc. [Harcourt Brace Jovanovich Publishers], multivariate series, prediction and control, Probability and Mathematical Statistics.

Raskutti, G., Wainwright, M. J., and Yu, B. (2010), “Restricted eigenvalue properties for correlated Gaussian designs,” *J. Mach. Learn. Res.*, 11, 2241–2259.

Rudelson, M. and Zhou, S. (2013), “Reconstruction From Anisotropic Random Measurements,” *Information Theory, IEEE Transactions on*, 59, 3434–3447.

Song, S. and Bickel, P. J. (2011), “Large Vector Auto Regressions,” *Arxiv preprint arXiv:1106.3915v1*.

van de Geer, S. A. and Bühlmann, P. (2009), “On the conditions used to prove oracle results for the Lasso,” *Electron. J. Stat.*, 3, 1360–1392.