Biometrika (2021), **108**, 2, *pp.* 469–489
Printed in Great Britain
doi: 10.1093/biomet/asaa065
Advance Access publication 2 February 2021

Finite-time analysis of vector autoregressive models under linear restrictions

BY YAO ZHENG

Department of Statistics, University of Connecticut,
215 Glenbrook Road, Storrs, Connecticut 06269, U.S.A.
yao.zheng@uconn.edu

AND GUANG CHENG

Department of Statistics, Purdue University,
250 N. University St., West Lafayette, Indiana 47907, U.S.A.
chengg@purdue.edu

SUMMARY

This paper develops a unified finite-time theory for the ordinary least squares estimation of possibly unstable and even slightly explosive vector autoregressive models under linear restrictions, with the applicable region $\rho(A) \leqslant 1 + c/n$, where $\rho(A)$ is the spectral radius of the transition matrix A in the VAR(1) representation, n is the time horizon and $c > 0$ is a universal constant. The linear restriction framework encompasses various existing models such as banded/network vector autoregressive models. We show that the restrictions reduce the error bounds via not only the reduced dimensionality, but also a scale factor resembling the asymptotic covariance matrix of the estimator in the fixed-dimensional set-up: as long as the model is correctly specified, this scale factor is decreasing in the number of restrictions. It is revealed that the phase transition from slow to fast error rate regimes is determined by the smallest singular value of A, a measure of the least excitable mode of the system. The minimax lower bounds are derived across different regimes. The developed non-asymptotic theory not only bridges the theoretical gap between stable and unstable regimes, but precisely characterizes the effect of restrictions and its interplay with model parameters. Simulations support our theoretical results.

Some key words: Consistency; Empirical process theory; Least squares estimation; Non-asymptotic analysis; Stochastic regression; Unstable process; Vector autoregressive model.

1. INTRODUCTION

The vector autoregressive model (Sims, 1980) is arguably the most fundamental model for multivariate time series (Lütkepohl, 2005; Tsay, 2013). Applications of the model and its variants can be found in almost any field that involves learning the temporal dependency: economics and finance (Wu & Xia, 2016), energy forecasting (Dowell & Pinson, 2016), psychopathology (Bringmann et al., 2013), neuroscience (Gorrostieta et al., 2012) and reinforcement learning (Recht, 2018), among others.

Consider the vector autoregressive model of order one, VAR(1), in the following form:

$$X_{t+1} = AX_t + \eta_t,$$

where $X_t \in \mathbb{R}^d$ is the observed time series, $A \in \mathbb{R}^{d\times d}$ is the unknown transition matrix, and $\eta_t \in \mathbb{R}^d$ is the innovation. In modern applications, the dimension d is often relatively large. However, since the number of unknown parameters increases as d^2, leading to problems such as overparameterization, the model cannot provide reliable estimates or forecasts without further restrictions (Stock & Watson, 2001). A classical approach to dimensionality reduction for vector autoregressive models, which recently has enjoyed a resurgence of interest, advocates the incorporation of prior knowledge into modelling. For example, motivated by the fact that in spatiotemporal studies it is often sufficient to collect information from neighbours, Guo et al. (2016) proposed the banded vector autoregressive model, where the nonzero entries of A are assumed to form a narrow band along the main diagonal, after arranging the d components of X_t by geographic location. To analyse users' time series over large social networks, the network vector autoregressive model of Zhu et al. (2017) used the follower-followee adjacency matrix to determine the zero-nonzero pattern of A, together with equality restrictions to further reduce the dimensionality. In fact, the above models can both be incorporated by the general framework of linear restrictions

$$\mathcal{C}\mathrm{vec}(A^{\mathrm{T}}) = \mu, \tag{1}$$

where $\mathcal{C}$ is a prespecified restriction matrix, μ is a known constant vector and A^{T} is the transpose of A. This form of restrictions is traditionally well known by time series modellers; see books on multivariate time series analysis such as Reinsel (1993), Lütkepohl (2005) and Tsay (2013).

Meanwhile, drawing inspiration from recent developments in high-dimensional regression, another well-studied approach concerns penalized estimation, where the modeller is agnostic to the locations of nonzero coordinates in A while assuming a certain sparsity (Davis et al., 2015; Han et al., 2015a; Basu & Michailidis, 2015), or the directions of low-dimensional projections while, e.g., assuming a low-rank structure of A (Ahn & Reinsel, 1988; Negahban & Wainwright, 2011). In the former case, once the locations of nonzero coordinates are identified, the model can be formulated as an instance of (1). Although we focus on fully known restrictions, the framework of (1) allows us to study the theoretical properties of a much richer variety of restriction patterns in this paper.

On the other hand, in the literature on large vector autoregressive models there has been an almost exclusive focus on stable processes. Technically, this means that the spectral radius $\rho(A) < 1$, or often, more stringently, that the spectral norm $\|A\|_2 < 1$. However, the analysis of stable processes typically cannot be carried over to unstable processes. In this paper we provide a novel finite-time, non-asymptotic analysis of the ordinary least squares estimator for stable, unstable and even slightly explosive vector autoregressive models within the general framework of (1).

Our analysis sheds new light on the phase transition phenomenon of the ordinary least squares estimator across different stability regimes. This is made possible by adopting the non-asymptotic, nonmixing approach of Simchowitz et al. (2018). Resting upon a generalization of Mendelson's (2014) small-ball method, this approach is particularly attractive because: (i) it unifies stable and unstable cases, whereas in asymptotic theory these two cases would require substantially different techniques; and (ii) in contrast to existing non-asymptotic methods, it can capture well the fundamental trait that the estimation will be more accurate as $\rho(A) \to 1$. While relaxing the normality assumption in Simchowitz et al. (2018), we precisely characterize the impact of imposing restrictions on the estimation error. More importantly, we reveal for the first time that the phase transition from slow to fast error rates depends on the the smallest singular value of A, a measure of the least excitable mode of the system. In addition, we expand the applicable

region $\rho(A) \leqslant 1$ in the above paper to $\rho(A) \leqslant 1 + c/n$, where $c > 0$ is a universal constant, so slightly explosive processes are also included. Compared to Guo et al. (2016), which focused on the case with $\|A\|_2 < 1$, our assumption on the innovation distribution is stronger, and our error rate for the stable regime is larger by a logarithmic factor which, however, can be dropped under the normality assumption; see § 3.4 for details. Although Zhu et al. (2017) relied on even milder assumptions on the innovations than Guo et al. (2016), they assumed that the number of unknown parameters is fixed, and hence their theoretical analysis is not comparable to ours; see Example 5 in § 3.1.

Throughout, we denote by $\|\cdot\|$ the Euclidean norm and by $\mathcal{S}^{d-1} = \{\omega \in \mathbb{R}^d : \|\omega\| = 1\}$ the unit sphere in $\mathbb{R}^d$. For a real matrix $A = (a_{ij})$, we let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, or $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$, be its largest and smallest eigenvalues, or singular values, respectively; additionally, we let $\rho(A) = |\lambda_{\max}(A)|$, $\|A\|_2 = \sup_{\|\omega\|=1}\|A\omega\|$ and $\|A\|_{\mathrm{F}} = (\sum_{i,j} a_{ij}^2)^{1/2}$ be the spectral radius, spectral norm and Frobenius norm of A, respectively. For $x \in \mathbb{R}$, let $\lfloor x \rfloor = \max\{k \in \mathbb{Z} : k \leqslant x\}$ and $\lceil x \rceil = \min\{k \in \mathbb{Z} : k \geqslant x\}$, where $\mathbb{Z}$ is the set of integers. We write $A \succ 0$, or $A \succeq 0$, if A is a positive definite, or positive semidefinite, matrix. Moreover, for any real symmetric matrices A and B, we write $A \prec B$ (or $A \preceq B$) if $B - A \succ 0$ (or $B - A \succeq 0$), and write $A \not\prec B$ (or $A \not\preceq B$) if $A \prec B$, or $A \preceq B$, does not hold. For any quantities X and Y, we write $X \gtrsim Y$ if there exists a universal constant $c > 0$ independent of $(n, d, m, R, k, \sigma, \delta)$, whose meaning will become clear later, such that $X \geqslant cY$.

2. Linearly restricted stochastic regression

2.1. *Problem formulation*

Consider a sequence of time-dependent covariate-response pairs $\{(X_t, Y_t)\}_{t=1}^n$ following

$$Y_t = A_* X_t + \eta_t, \tag{2}$$

where $Y_t, \eta_t \in \mathbb{R}^q$, $X_t \in \mathbb{R}^d$, $A_* \in \mathbb{R}^{q \times d}$ and $E(\eta_t) = 0$. In particular, (2) becomes the VAR(1) model when $Y_t = X_{t+1}$. In model (2), the process $\{X_t, t = 1, 2, \ldots\}$ is adapted to the filtration

$$\mathcal{F}_t = \sigma\{\eta_1, \ldots, \eta_{t-1}, X_1, \ldots, X_t\}.$$

Let $\beta_* = \mathrm{vec}(A_*^{\mathrm{T}}) \in \mathbb{R}^N$, where $N = qd$. The parameter space of the linearly restricted model can be defined as

$$\mathcal{L} = \{\beta \in \mathbb{R}^N : \mathcal{C}\beta = \mu\},$$

where $\mathcal{C}$ is a known $(N - m) \times N$ matrix of rank $N - m$, representing $N - m$ independent restrictions, and $\mu \in \mathbb{R}^{N-m}$ is a known constant vector which may simply be set to zero in practice. Let $\mathcal{C}_+$ be an $m \times N$ complement of $\mathcal{C}$ such that $\mathcal{C}_{\mathrm{full}} = (\mathcal{C}_+^{\mathrm{T}}, \mathcal{C}^{\mathrm{T}})^{\mathrm{T}}$ is invertible, with its inverse partitioned into two blocks as $\mathcal{C}_{\mathrm{full}}^{-1} = (R, R_+)$, where R is the matrix of the first m columns of $\mathcal{C}_{\mathrm{full}}^{-1}$. Additionally, define $\gamma = R_+\mu$. Then $\mathcal{C}\gamma = \mathcal{C}R_+\mu = \mu$. If $\mathcal{C}\beta = \mu$, then $\beta = \mathcal{C}_{\mathrm{full}}^{-1}\mathcal{C}_{\mathrm{full}}\beta = R\mathcal{C}_+\beta + R_+\mathcal{C}\beta = R\theta + \gamma$, where $\theta = \mathcal{C}_+\beta$. Conversely, for any $\theta \in \mathbb{R}^m$, if $\beta = R\theta + \gamma$, then $\mathcal{C}\beta = \mathcal{C}R\theta + \mathcal{C}\gamma = \mu$. Thus, we have

$$\mathcal{L} = \{R\theta + \gamma : \theta \in \mathbb{R}^m\},$$

i.e., the linear space spanned by columns of the restriction matrix R, shifted by the constant vector γ. This immediately implies that, given (R, γ), there exists a unique unrestricted parameter $\theta_* \in \mathbb{R}^m$ such that $\beta_* = R\theta_* + \gamma$. Note that $\gamma = 0$ if and only if $\mu = 0$. Moreover, the unrestricted model corresponds to the special case where $R = I_N$ and $\gamma = 0$.

The following examples illustrate how the linear restrictions can be encoded by (R, γ) or $(\mathcal{C}, \mu)$, where, without loss of generality, we set $\mu = \gamma = 0$. Let β_{*i} denote the ith entry of β_*.

Example 1 (*Zero restriction*). The restriction $\beta_{*i} = 0$ may be encoded by setting the ith row of R to zero, or by setting a row of $\mathcal{C}$ to $(0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^N$, where the ith entry is one.

Example 2 (*Equality restriction*). Consider the restriction $\beta_{*i} - \beta_{*j} = 0$. Suppose that the value of $\beta_{*i} = \beta_{*j}$ is θ_{*k}, the kth entry of θ_*. Then this restriction may be encoded by setting both the ith and jth rows of R to $(0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^m$, where the kth entry is one. Alternatively, we may set a row of $\mathcal{C}$ to the $1 \times N$ vector $c(i, j)$ whose ℓth element is defined as $[c(i,j)]_\ell = 1(\ell = i) - 1(\ell = j)$, where $1(\cdot)$ is the indicator function.

Define $n \times q$ matrices $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ and $E = (\eta_1, \ldots, \eta_n)^{\mathrm{T}}$, and the $n \times d$ matrix $X = (X_1, \ldots, X_n)^{\mathrm{T}}$. Then (2) has the matrix form $Y = XA_*^{\mathrm{T}} + E$. Let $y = \mathrm{vec}(Y)$, $\eta = \mathrm{vec}(E)$, $Z = (I_q \otimes X)R$ and $\tilde{y} = y - (I_q \otimes X)\gamma$. By vectorization and reparameterization, we have

$$\tilde{y} = (I_q \otimes X)(\beta_* - \gamma) + \eta = Z\theta_* + \eta.$$

As a result, the ordinary least squares estimator of β_* for the linearly restricted model is

$$\hat{\beta} = R\hat{\theta} + \gamma, \qquad \hat{\theta} = \underset{\theta \in \mathbb{R}^m}{\arg\min} \|\tilde{y} - Z\theta\|^2, \tag{3}$$

where $Z \in \mathbb{R}^{qn \times m}$. To ensure the feasibility of (3), we need $qn \geqslant m$. Let $R = (R_1^{\mathrm{T}}, \ldots, R_q^{\mathrm{T}})^{\mathrm{T}}$ and $\gamma = (\gamma_1^{\mathrm{T}}, \ldots, \gamma_q^{\mathrm{T}})^{\mathrm{T}}$, where R_i are $d \times m$ matrices and γ_i are $d \times 1$ vectors. Then, $A_* = (I_q \otimes \theta_*^{\mathrm{T}})\tilde{R} + G$, where

$$\tilde{R} = (R_1, \ldots, R_q)^{\mathrm{T}} \in \mathbb{R}^{mq \times d}, \qquad G = (\gamma_1, \ldots, \gamma_q)^{\mathrm{T}} \in \mathbb{R}^{q \times d}.$$

Consequently, the ordinary least squares estimator of A is $\hat{A} = (I_q \otimes \hat{\theta}^{\mathrm{T}})\tilde{R} + G$.

2.2. *General upper bounds analysis*

To derive upper estimation error bounds for the stochastic regression model in § 2.1, we begin by introducing a key technical ingredient, namely the block martingale small-ball condition (Simchowitz et al., 2018). As a generalization of Mendelson's (2014) small-ball method to time-dependent data, this condition can be viewed as a non-asymptotic stability assumption for controlling the lower tail behaviour of the Gram matrix $X^{\mathrm{T}}X$, or $Z^{\mathrm{T}}Z$ in our context.

DEFINITION 1 (Block martingale small-ball condition). (i) *For a real-valued time series* $\{X_t, t = 1, 2, \ldots\}$ *adapted to the filtration* $\{\mathcal{F}_t\}$, *we say that* $\{X_t\}$ *satisfies the* (k, ν, α)-BMSB *condition if there exist an integer* $k \geqslant 1$, *and constants* $\nu > 0$ *and* $\alpha \in (0, 1)$ *such that, for every integer* $s \geqslant 0$, $k^{-1}\sum_{t=1}^{k} \mathrm{pr}(|X_{s+t}| \geqslant \nu \mid \mathcal{F}_s) \geqslant \alpha$ *with probability 1.* (ii) *For a time series* $\{X_t, t = 1, 2, \ldots\}$ *taking values in* $\mathbb{R}^d$, *we say that* $\{X_t\}$ *satisfies the* $(k, \Gamma_{\mathrm{sb}}, \alpha)$-BMSB *condition if there exists* $0 \prec \Gamma_{\mathrm{sb}} \in \mathbb{R}^{d \times d}$ *such that, for every* $\omega \in \mathcal{S}^{d-1}$, *the real-valued time series* $\{\omega^{\mathrm{T}}X_t, t = 1, 2, \ldots\}$ *satisfies the* $\{k, (w^{\mathrm{T}}\Gamma_{\mathrm{sb}}w)^{1/2}, \alpha\}$-BMSB *condition.*

The value of the probability α is unimportant for our purpose as long as it exists. The thresholding matrix Γ_{sb}, or ν in the univariate case, captures the average cumulative excitability over any size-k block; e.g., if $\{X_t\}$ is a mean-zero vector autoregressive process, then Γ_{sb} will scale proportionally no less than $k^{-1}E(\sum_{t=1}^{k} X_{s+t}X_{s+t}^{\mathrm{T}} \mid \mathcal{F}_s)$, which is constant for all s. Since every time-point is associated with a new shock to the process, $E(X_{s+t}X_{s+t}^{\mathrm{T}} \mid \mathcal{F}_s)$ will increase as t increases. Consequently, Γ_{sb} will be monotonic increasing in k; see Lemma 1 in § 3.2. Moreover, by aggregating all the size-k blocks, Γ_{sb} will essentially become the lower bound of $n^{-1}\sum_{t=1}^{n} X_tX_t^{\mathrm{T}}$. For ordinary least squares estimation, a larger lower bound on the Gram matrix will yield a sharper estimation error bound. Thus, a larger block size k is generally preferred; see Theorem 3 in § 3 for details.

Let Γ_{sb} and $\overline{\Gamma}$ be $d \times d$ positive definite matrices, and denote

$$\underline{\Gamma}_R = R^{\mathrm{T}}(I_q \otimes \Gamma_{\mathrm{sb}})R, \qquad \overline{\Gamma}_R = R^{\mathrm{T}}(I_q \otimes \overline{\Gamma})R. \tag{4}$$

In our theoretical analysis, properly rescaled matrices $\underline{\Gamma}_R$ and $\overline{\Gamma}_R$ will serve as lower and upper bounds of the Gram matrix $Z^{\mathrm{T}}Z$, respectively, and the covering numbers derived from them will give rise to the quantity $\log\det(\overline{\Gamma}_R\underline{\Gamma}_R^{-1})$ in Theorem 1. The regularity conditions underlying our upper bound analysis are listed as follows:

Assumption 1. The covariates process $\{X_t\}_{t=1}^{n}$ satisfies the $(k, \Gamma_{\mathrm{sb}}, \alpha)$-BMSB condition.

Assumption 2. For any $\delta \in (0,1)$, there exists $\overline{\Gamma}_R$ defined as in (4) such that $\mathrm{pr}(Z^{\mathrm{T}}Z \not\preccurlyeq n\overline{\Gamma}_R) \leqslant \delta$, where $Z = (I_q \otimes X)R$, and $\overline{\Gamma}_R$ is dependent on δ.

Assumption 3. For every integer $t \geqslant 1$, $\eta_t \mid \mathcal{F}_t$ is mean zero and σ^2-sub-Gaussian.

THEOREM 1. *Let $\{(X_t, Y_t)\}_{t=1}^{n}$ be generated by the linearly restricted stochastic regression model. Fix $\delta \in (0,1)$. Suppose that Assumptions 1–3 hold, $0 \prec \Gamma_{\mathrm{sb}} \preceq \overline{\Gamma}$, and*

$$n \geqslant \frac{9k}{\alpha^2}\left\{m\log\frac{27}{\alpha} + \frac{1}{2}\log\det(\overline{\Gamma}_R\underline{\Gamma}_R^{-1}) + \log q + \log\frac{1}{\delta}\right\}. \tag{5}$$

Then, with probability at least $1 - 3\delta$, we have

$$\|\hat{\beta} - \beta_*\| \leqslant \frac{9\sigma}{\alpha}\left[\frac{\lambda_{\max}(R\underline{\Gamma}_R^{-1}R^{\mathrm{T}})}{n}\left\{12m\log\frac{14}{\alpha} + 9\log\det(\overline{\Gamma}_R\underline{\Gamma}_R^{-1}) + 6\log\frac{1}{\delta}\right\}\right]^{1/2},$$

where $\|\hat{\beta} - \beta_*\| = \|\hat{A} - A_*\|_{\mathrm{F}}$. The result in Theorem 1 is new even for the unrestricted stochastic regression, where $R = I_N$. For vector autoregressive processes, we will specify the matrices $\underline{\Gamma}_R$ and $\overline{\Gamma}_R$ in § 3.2, where $\underline{\Gamma}_R$ will depend on the block size k through Γ_{sb}. As k increases, Γ_{sb} will become larger, and hence the factor $\lambda_{\max}(R\underline{\Gamma}_R^{-1}R^{\mathrm{T}})$ will become smaller, resulting in a sharper error bound. However, k cannot be too large due to condition (5). Specifically, this condition arises from applying the Chernoff bound technique to lower bound the Gram matrix via aggregation of all the size-k blocks, since the probability guarantee of the Chernoff bound will degrade as the number of blocks decreases; see the Supplementary Material. Therefore, to apply Theorem 1 to vector autoregressive processes, a crucial step will be to

derive a feasible region for k that guarantees condition (5); see the Supplementary Material for details.

Similarly, we can obtain an analogous upper bound for $\hat{A} - A_*$ in the spectral norm:

PROPOSITION 1. *Let $\{(X_t, Y_t)\}_{t=1}^n$ be generated by the linearly restricted stochastic regression model. Fix $\delta \in (0,1)$. Then, under the conditions of Theorem 1, with probability at least $1-3\delta$, we have*

$$\|\hat{A} - A_*\|_2 \leqslant \frac{9\sigma}{\alpha}\left[\frac{\lambda_{\max}\left(\sum_{i=1}^q R_i \underline{\Gamma}_R^{-1} R_i^{\mathrm{T}}\right)}{n}\left\{12m\log\frac{14}{\alpha} + 9\log\det(\overline{\Gamma}_R\underline{\Gamma}_R^{-1}) + 6\log\frac{1}{\delta}\right\}\right]^{1/2}.$$

3. LINEARLY RESTRICTED VECTOR AUTOREGRESSION

3.1. *Representative examples*

We begin by illustrating how the formulation in § 2 can be used to study vector autoregressive models. Four representative examples will be discussed: the VAR(p) model, the banded vector autoregressive model, the network vector autoregressive model and the pure unit root process.

Consider the VAR(1) model, i.e., model (2) with $Y_t = X_{t+1} \in \mathbb{R}^d$:

$$X_{t+1} = A_* X_t + \eta_t, \tag{6}$$

subject to $\beta_* = R\theta_* + \gamma$, where $\beta_* = \mathrm{vec}(A_*^{\mathrm{T}}) \in \mathbb{R}^{d^2}$, $R = (R_1^{\mathrm{T}}, \ldots, R_d^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{d^2 \times m}$ with R_i being $d \times m$ matrices, $\theta_* \in \mathbb{R}^m$ and $\gamma = (\gamma_1^{\mathrm{T}}, \ldots, \gamma_d^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{d^2}$ with $\gamma_i \in \mathbb{R}^d$.

Example 3 (VAR(p) model). Interestingly, vector autoregressive models of order $p < \infty$ can be viewed as linearly restricted VAR(1) models. Consider the VAR(p) model

$$Z_{t+1} = A_{*1}Z_t + A_{*2}Z_{t-1} + \cdots + A_{*p}Z_{t-p+1} + \varepsilon_t, \tag{7}$$

where $Z_t, \varepsilon_t \in \mathbb{R}^{d_0}$ and $A_{*i} \in \mathbb{R}^{d_0 \times d_0}$ for $i = 1, \ldots, p$. Denote $d = d_0 p$, $X_t = (Z_t^{\mathrm{T}}, Z_{t-1}^{\mathrm{T}}, \ldots, Z_{t-p+1}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^d$, $\eta_t = (\varepsilon_t^{\mathrm{T}}, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^d$ and

$$A_* = \begin{pmatrix} A_{*1} & \cdots & A_{*p-1} & A_{*p} \\ I_{d_0} & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & I_{d_0} & 0 \end{pmatrix} \in \mathbb{R}^{d \times d}. \tag{8}$$

As a result, (7) can be written into the VAR(1) form in (6). As shown in (8), all entries in the last $d - d_0$ rows of A_* are restricted to either zero or one. The restriction $\beta_{*i} = 1$ can be encoded in $(\mathcal{C}, \mu)$ in the same way as Example 1, but with the ith entry of μ set to one. Thus, with or without restrictions, the VAR(p) model can be studied by the same method as that for a linearly restricted VAR(1) model. The special structure of the innovation η_t that some entries of η_t are fixed at zero will not pose any extra difficulty.

In the following examples, we consider VAR(1) models with various structures for $A_* = (a_{*ij})_{d\times d}$, and set $\gamma = 0$ so that the restrictions are in the form of $R\theta = 0$:

Example 4 (Banded vector autoregression). Guo et al. (2016) proposed the vector autoregressive model with the following zero restrictions:

$$a_{*ij} = 0, \qquad |i-j| > k_0, \tag{9}$$

where the integer $1 \leqslant k_0 \leqslant \lfloor (d-1)/2 \rfloor$ is called the bandwidth parameter. Let $b_{*i} \in \mathbb{R}^d$ be the transpose of the ith row of A_*. Hence, $\beta_* = (b_{*1}^{\mathrm{T}}, \ldots, b_{*d}^{\mathrm{T}})^{\mathrm{T}}$. The restrictions are imposed on each b_{*i} separately. As a result, the b_{*i} are determined by non-overlapping subsets of entries in θ_*; that is, we can write $b_{*i} = R_{(i)}\vartheta_{*i}$, where $R_{(i)} \in \mathbb{R}^{d\times m_i}$, $\vartheta_{*i} \in \mathbb{R}^{m_i}$, $\theta_* = (\vartheta_{*1}^{\mathrm{T}}, \ldots, \vartheta_{*d}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^m$ and $m = \sum_{i=1}^d m_i$. In this case, R is a block diagonal matrix:

$$R = \begin{pmatrix} R_{(1)} & & 0 \\ & \ddots & \\ 0 & & R_{(d)} \end{pmatrix} \in \mathbb{R}^{d^2\times m},$$

and (9) can be encoded in R as follows: (i) $m_i = k_0 + i$ and $R_{(i)} = (I_{m_i}, 0)^{\mathrm{T}}$ if $1 \leqslant i \leqslant k_0 + 1$; (ii) $m_i = 2k_0 + 1$ and $R_{(i)} = (0_{m_i\times(i-k_0-1)}, I_{m_i}, 0_{m_i\times(d-i-k_0)})^{\mathrm{T}}$ if $k_0 + 1 < i < d - k_0$; and (iii) $m_i = k_0 + 1 + d - i$ and $R_{(i)} = (0, I_{m_i})^{\mathrm{T}}$ if $d - k_0 \leqslant i \leqslant d$.

Example 5 (Network vector autoregression). Consider the network model in Zhu et al. (2017). Let us drop the individual effect and the intercept to ease the notation. This model assumes that all diagonal entries of A_* are equal: $a_{*ii} = \theta_{*1}$ for $1 \leqslant i \leqslant d$. For the off-diagonal entries, the zero-nonzero pattern is known and completely determined by the social network: $a_{*ij} \neq 0$ if and only if individual i follows individual j. Moreover, all nonzero off-diagonal entries are assumed to be equal: $a_{*ij} = \theta_{*2}$ if $a_{*ij} \neq 0$, for $1 \leqslant i \neq j \leqslant d$. This model is actually very parsimonious, with only $m = 2$, while the network size d can be extremely large. To incorporate the above restrictions, we may define the $d^2 \times 2$ matrix R as follows: for $i = 1, \ldots, d^2$, the ith row of R is $(1, 0)$ if β_{*i} corresponds to a diagonal entry of A_*, $(0, 1)$ if β_{*i} corresponds to a nonzero off-diagonal entry of A_*, and $(0, 0)$ if β_{*i} corresponds to a zero off-diagonal entry of A_*.

Example 6 (Pure unit root process). Another simple but important case is $A_* = \rho I_d$ with $\rho \in \mathbb{R}$. Then, the smallest true model has $m = 1$, and the corresponding restrictions, $a_{*11} = \cdots = a_{*dd}$ and $a_{*ij} = 0$ for $1 \leqslant i \neq j \leqslant d$, can be imposed by setting $R = (e_1^{\mathrm{T}}, \ldots, e_d^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{d^2}$, where e_i is the $d \times 1$ vector with all elements zero except the ith being one. When $\rho = 1$, the underlying model becomes the pure unit root process, a classic example of unstable vector autoregressive processes (Hamilton, 1994). In particular, the problem of testing $A_* = I_d$, or unit root testing in panel data, has been extensively studied in the asymptotic literature; see Chang (2004) and Zhang et al. (2018) for studies in low and high dimensions, respectively. Zhang et al. (2018) focused on asymptotic distributions of the largest eigenvalues of the sample covariance matrix of the pure unit root process under $\lim_{n,d\to\infty} d/n = 0$. It does not involve parameter estimation, and hence cannot be directly compared to this paper.

The stochastic regression can also incorporate possibly time-dependent exogenous inputs such as individual effects (Zhu et al., 2017) and observable factors (Zhou et al., 2018), leading to the class of VARX models (see, e.g., Wilms et al., 2017). Since VARX models can be analysed similarly to vector autoregressive models, we do not pursue the details in this paper.

3.2. *Verification of Assumptions 1–3 in Theorem 1*

In light of the generalizability to VAR(p) models via the VAR(1) representation, to apply the general results in § 2.2 to linearly restricted vector autoregressive models, it suffices to restrict our attention to the VAR(1) model in (6) from now on.

Following the notation in § 2, let $Y = (X_2, \ldots, X_{n+1})^{\mathrm{T}}, Z = (I_d \otimes X)R$ and $A_* = (I_d \otimes \theta_*^{\mathrm{T}})\tilde{R} + G$, where $\tilde{R} = (R_1, \ldots, R_d)^{\mathrm{T}}$ and $G = (\gamma_1, \ldots, \gamma_d)^{\mathrm{T}}$, where $q = d$ for the VAR(1) model. In addition, $\{X_t\}$ is adapted to the filtration

$$\mathcal{F}_t = \sigma\{\eta_1, \ldots, \eta_{t-1}\}.$$

The following conditions on $\{X_t\}$ will be invoked in our analysis:

Assumption 4. (i) The process $\{X_t\}$ starts at $t = 0$ with $X_0 = 0$; (ii) the innovations $\{\eta_t\}$ are independent and identically distributed with $E(\eta_t) = 0$ and $\mathrm{var}(\eta_t) = \Sigma_\eta = \sigma^2 I_d$; (iii) there is a universal constant $C_0 > 0$ such that, for every $\nu \in \mathbb{R}^d$ with $\nu^{\mathrm{T}}\Sigma_\eta \nu \neq 0$, the density of $\nu^{\mathrm{T}}\eta_t/(\nu^{\mathrm{T}}\Sigma_\eta\nu)^{1/2}$ is bounded from above by C_0 almost everywhere; and (iv) $\{\eta_t\}$ are σ^2-sub-Gaussian.

Under Assumption 4(i), we can simply write X_t in the finite-order moving average form, $X_t = \sum_{s=0}^{t-1} A_*^s \eta_{t-s-1}$ for any $t \geqslant 1$. Then, by Assumption 4(ii), $\mathrm{var}(X_t) = \sigma^2 \Gamma_t$, where

$$\Gamma_t = \sum_{s=0}^{t-1} A_*^s (A_*^{\mathrm{T}})^s \tag{10}$$

is called the finite-time controllability Gramian (Simchowitz et al., 2018). Here $\mathrm{var}(X_t) < \infty$ for any A_*. By contrast, the typical set-up in asymptotic theory of stable processes assumes that $\{X_t\}$ starts at $t = -\infty$. In this case, $\{X_t\}_{t\in\mathbb{Z}}$ has the infinite-order moving average form $X_t = \sum_{s=0}^{\infty} A_*^s \eta_{t-s-1}$, so $\mathrm{var}(X_t) = \sigma^2 \sum_{s=0}^{\infty} A_*^s (A_*^{\mathrm{T}})^s = \sigma^2 \lim_{t\to\infty} \Gamma_t < \infty$ if and only if $\rho(A_*) < 1$. Thus, an important benefit of Assumption 4(i) is that it allows us to capture the possibly explosive behaviour of $\mathrm{var}(X_t)$ over any finite time horizon, and derive upper bounds of the Gram matrix over different stability regimes; see Lemmas 2 and 3 in this subsection.

The condition $\Sigma_\eta = \sigma^2 I_d$ in Assumption 4(ii) is imposed for simplicity. However, we can easily extend all the proofs in this paper to the general case with any symmetric matrix $\Sigma_\eta \succeq 0$: we only need to rederive all results with the role of Γ_t replaced by $\sum_{s=0}^{t-1} A_*^s \Sigma_\eta (A_*^{\mathrm{T}})^s$. Assumption 4(iii) is used to establish the block martingale small-ball condition, i.e., Assumption 1, for vector autoregressive processes, and it allows us to lower bound the small-ball probability by leveraging Theorem 1.2 in Rudelson & Vershynin (2015) on densities of sums of independent random variables; see also Remark 3. Essentially, Assumption 4(iii) only requires that the distribution of any one-dimensional projection of the innovation is well spread on the real line. Examples of such distributions include multivariate normal and multivariate t (Kotz & Nadarajah, 2004) distributions and, more generally, elliptical distributions (Fang et al., 1990) with the consistency property in Kano (1994). Lastly, it is clear that Assumptions 4(ii) and (iv) guarantee Assumption 3.

Remark 1. In asymptotic theory, stable and unstable processes require substantially different techniques, and results derived under $\rho(A_*) < 1$ typically cannot be carried over to unstable processes. For example, the convergence rate of the ordinary least squares estimator for fixed-dimensional unstable VAR(1) processes is n instead of $n^{1/2}$, and the limiting distribution is no longer normal (Hamilton, 1994).

Remark 2. The controllability Gramian Γ_t is interpretable even without Assumption 4(i). By recursion, we have $X_{s+t} = \sum_{\ell=0}^{t-1} A_*^\ell \eta_{s+t-\ell-1} + A_*^t X_s$ for any time-point s and duration $t \geqslant 1$. As a result, $\mathrm{var}(X_{s+t} \mid \mathcal{F}_s) = \sum_{\ell=0}^{t-1} A_*^\ell \Sigma_\eta (A_*^{\mathrm{T}})^\ell$, and it simply becomes $\sigma^2 \Gamma_t$ if $\Sigma_\eta = \sigma^2 I_d$; $\mathrm{var}(X_{s+t} \mid \mathcal{F}_s)$, or equivalently Γ_t, is a partial sum of a geometric sequence due to the autoregressive structure. Roughly speaking, larger A_* means more persistent impact of η_t.

In the following, we present three lemmas for the linearly restricted vector autoregressive model. Lemma 1 establishes the block martingale small-ball condition by specifying Γ_{sb}, while Lemmas 2 and 3 verify Assumption 2 by providing two possible specifications of $\overline{\Gamma}$.

Some additional notation to be used in Lemma 3 is introduced as follows: denote by Σ_X the covariance matrix of the $dn \times 1$ vector $\mathrm{vec}(X^{\mathrm{T}}) = (X_1^{\mathrm{T}}, \ldots, X_n^{\mathrm{T}})^{\mathrm{T}}$, so the (t,s)th $d \times d$ block of Σ_X is $E(X_t X_s^{\mathrm{T}})$, for $1 \leqslant t, s \leqslant n$. Then define

$$\xi = \xi(m,d,n,\delta) = 2\left\{\frac{\lambda_{\max}(\Gamma_n)\psi(m,d,\delta)\|\Sigma_X\|_2}{\sigma^2 n}\right\}^{1/2} + \frac{2\psi(m,d,\delta)\|\Sigma_X\|_2}{\sigma^2 n}, \tag{11}$$

where $\psi(m,d,\delta) = C_1\{m\log 9 + \log d + \log(2/\delta)\}$ and $C_1 > 0$ is a universal constant.

LEMMA 1. *Suppose that* $\{X_t\}_{t=1}^{n+1}$ *follows* $X_{t+1} = A_* X_t + \eta_t$ *for* $t = 0, 1, \ldots, n$. *Under Assumptions* 4(ii) *and* (iii), *for any* $1 \leqslant k \leqslant \lfloor n/2 \rfloor$, $\{X_t\}_{t=1}^n$ *satisfies the* $(2k, \Gamma_{\mathrm{sb}}, 1/10)$-BMSB *condition, where* $\Gamma_{\mathrm{sb}} = \sigma^2 \Gamma_k / (4C_0)^2$.

LEMMA 2. *Let* $\{X_t\}_{t=1}^{n+1}$ *be generated by the linearly restricted vector autoregressive model. Under Assumptions* 4(i) *and* (ii), *for any* $\delta \in (0,1)$ *we have* $\mathrm{pr}(Z^{\mathrm{T}} Z \not\preceq n\overline{\Gamma}_R) \leqslant \delta$, *where* $\overline{\Gamma}_R = R^{\mathrm{T}}(I_d \otimes \overline{\Gamma})R$, *with* $\overline{\Gamma} = \sigma^2 m \Gamma_n / \delta$.

LEMMA 3. *Let* $\{X_t\}_{t=1}^{n+1}$ *be generated by the linearly restricted vector autoregressive model. Under Assumptions* 4(i) *and* (ii), *if* $\{\eta_t\}$ *are normal, then for any* $\delta \in (0,1)$ *we have* $\mathrm{pr}(Z^{\mathrm{T}} Z \not\preceq n\overline{\Gamma}_R) \leqslant \delta$, *where* $\overline{\Gamma}_R = R^{\mathrm{T}}(I_d \otimes \overline{\Gamma})R$, *with* $\overline{\Gamma} = \sigma^2 \Gamma_n + \sigma^2 \xi I_d$ *and* $\xi = \xi(m,d,n,\delta)$ *as defined in* (11).

Remark 3. Unlike Simchowitz et al. (2018), by leveraging Rudelson & Vershynin (2015), we establish the block martingale small-ball condition without the normality assumption. If Assumption 4(ii) is relaxed to the general $\mathrm{var}(\eta_t) = \Sigma_\eta \succeq 0$, by a straightforward extension of the proof of Lemma 1, we can show that Lemma 1 holds with $\Gamma_{\mathrm{sb}} = \sum_{\ell=0}^{k-1} A_*^\ell \Sigma_\eta (A_*^{\mathrm{T}})^\ell / (4C_0)^2$.

Remark 4. Lemma 2 is a simple consequence of the Markov inequality and the property that $\lambda_{\max}(\cdot) \leqslant \mathrm{tr}(\cdot)$, so no distributional assumption on η_t is required. However, Lemma 3 relies on the Hanson–Wright inequality (Vershynin, 2018), where the normality assumption is invoked. Although adopting the $\overline{\Gamma}$ in Lemma 3 can eliminate a factor of $\log m$ in the resulting estimation error bounds, the $\overline{\Gamma}$ in Lemma 2 actually leads to sharper bounds under certain conditions on A_*; see § 3.3 and § 3.4 for details.

By Lemma 1, for any $1 \leqslant k \leqslant \lfloor n/2 \rfloor$, the matrix $\underline{\Gamma}_R$ in Theorem 1 can be specified as

$$\underline{\Gamma}_R = \sigma^2 R^{\mathrm{T}}(I_d \otimes \Gamma_k)R / (4C_0)^2. \tag{12}$$

By Lemmas 2 and 3, the matrix $\overline{\Gamma}_R$ in Theorem 1 can be chosen as $\overline{\Gamma}_R = \overline{\Gamma}_R^{(1)}$ or $\overline{\Gamma}_R^{(2)}$, where

$$\overline{\Gamma}_R^{(1)} = \sigma^2 m R^{\mathrm{T}}(I_d \otimes \Gamma_n)R/\delta, \qquad \overline{\Gamma}_R^{(2)} = \sigma^2 R^{\mathrm{T}}(I_d \otimes \Gamma_n)R + \sigma^2 \xi(m,d,n,\delta) R^{\mathrm{T}} R; \tag{13}$$

recall the definitions of $\underline{\Gamma}_R$ and $\overline{\Gamma}_R$ in (4), where $q = d$ for the VAR(1) model. Furthermore, observe that $\underline{\Gamma}_R$ in (12) and the two $\overline{\Gamma}_R$s in (13), which serve as lower and upper bounds of $Z^{\mathrm{T}}Z$, respectively, are all related to the controllability Gramian Γ_t, and $0 \prec I_d \preceq \Gamma_k \preceq \Gamma_n$.

3.3. *Feasible region of k*

With $\underline{\Gamma}_R$ and $\overline{\Gamma}_R$ chosen as in (12) and (13), the term $\log\det(\overline{\Gamma}_R\underline{\Gamma}_R^{-1})$ in condition (5) in Theorem 1 is intricately dependent on both k and n. Thus, in order to apply the theorem to model (6), we need to verify the existence of the block size k satisfying (5). This boils down to deriving an explicit upper bound of $\log\det(\overline{\Gamma}_R\underline{\Gamma}_R^{-1})$ free of k.

By (10) and (12), it is easy to show that $\det(\underline{\Gamma}_R)$ is monotonic increasing in k. Then, as $\overline{\Gamma}_R$ is free of k, $\log\det(\overline{\Gamma}_R\underline{\Gamma}_R^{-1})$ is maximized when $k = 1$. As a result, we can first upper bound $\log\det(\overline{\Gamma}_R\underline{\Gamma}_R^{-1})$ by its value at $k = 1$ to get rid of its dependence on k. That is,

$$\log\det(\overline{\Gamma}_R\underline{\Gamma}_R^{-1}) \leqslant \log\det\{\overline{\Gamma}_R(\sigma^2 R^{\mathrm{T}}R)^{-1}(4C_0)^2\}. \tag{14}$$

Now it suffices to upper bound the right-hand side of (14), where $\overline{\Gamma}_R$ can be chosen from $\overline{\Gamma}_R^{(1)}$ and $\overline{\Gamma}_R^{(2)}$ in (13). As shown in the Supplementary Material,

$$\log\det\{\overline{\Gamma}_R(\sigma^2 R^{\mathrm{T}}R)^{-1}(4C_0)^2\} \lesssim \begin{cases} m\log(m/\delta) + \kappa, & \text{if } \overline{\Gamma}_R = \overline{\Gamma}_R^{(1)}, \\ m\log\{2\max(1,\xi)\} + \kappa, & \text{if } \overline{\Gamma}_R = \overline{\Gamma}_R^{(2)}, \end{cases} \tag{15}$$

where $\xi = \xi(m, d, n, \delta)$ is defined as in (11), and

$$\kappa = \log\det\left\{R^{\mathrm{T}}(I_d \otimes \Gamma_n)R(R^{\mathrm{T}}R)^{-1}\right\}. \tag{16}$$

Obviously, without imposing normality, we can only choose $\overline{\Gamma}_R = \overline{\Gamma}_R^{(1)}$; see Lemmas 2 and 3 in § 3.2. However, if $\{\eta_t\}$ are normal, $\overline{\Gamma}_R$ can be set to whichever of $\overline{\Gamma}_R^{(1)}$ and $\overline{\Gamma}_R^{(2)}$ delivers the sharper upper bound. Both κ and ξ depend on n, and their growth rates with respect to n depend on the magnitude of A_*. This will ultimately affect the choice between $\overline{\Gamma}_R^{(1)}$ and $\overline{\Gamma}_R^{(2)}$; e.g., if κ is the dominating term in both upper bounds in (15), we will be indifferent between the two. Assumptions 5–6′ below summarize the three cases of A_* we consider:

Assumption 5. $\rho(A_*) \leqslant 1 + c/n$, where $c > 0$ is a universal constant;

Assumption 6. $\rho(A_*) \leqslant \bar{\rho} < 1$ and $\|A_*\|_2 \leqslant C$, where $\bar{\rho}, C > 0$ are universal constants;

Assumption 6′. $\rho(A_*) \leqslant \bar{\rho} < 1$, $\mu_{\min}(\mathcal{A}) = \inf_{\|z\|=1}\lambda_{\min}\{\mathcal{A}^*(z)\mathcal{A}(z)\} \geqslant \mu_1$ and $\|A_*^t\|_2 \leqslant C\varrho^t$ for any integer $1 \leqslant t \leqslant n$, where $\bar{\rho}, \mu_1, C > 0$ and $\varrho \in (0, 1)$ are universal constants, and $\mathcal{A}(z) = I_d - A_* z$ for any complex number z.

Assumption 5 is the most general case among the above three, and Assumption 6 is weaker than Assumption 6′. Assumption 6′ does not require $\|A_*\|_2 < 1$ because C may be greater than one. Guo et al. (2016) assumed $\|A_*^t\|_2 \leqslant \varrho^t$ for $\varrho \in (0, 1)$ and any positive integer t, while it is unclear if this can be relaxed to $\|A_*^t\|_2 \leqslant C\varrho^t$ as in Assumption 6′. We need $\mu_{\min}(\mathcal{A})$ to be bounded away from zero in order to derive a sharp upper bound on $\|\Sigma_X\|_2$; see Remark 6. This condition is also necessary for the estimation error rates derived in Basu & Michailidis (2015)

for a similar reason. In particular, if A_* is diagonalizable, it is shown in Proposition 2.2 therein that $\mu_{\min}(\mathcal{A}) \geqslant [\{1-\rho(A_*)\}/\operatorname{cond}(S)]^2$, where S is defined in (17).

Remark 5. From (16), κ is dependent on n through Γ_n. If $\rho(A_*) < 1$ then $\Gamma_n \preceq \Gamma_\infty = \lim_{n\to\infty}\Gamma_n < \infty$ and $\kappa \leqslant \log\det\{R^{\mathrm{T}}(I_d \otimes \Gamma_\infty)R(R^{\mathrm{T}}R)^{-1}\}$, an upper bound free of n. By contrast, if $\rho(A_*) \geqslant 1$ then Γ_∞ no longer exists, and we need to carefully control the growth rate of Γ_n; see Lemma S7 in the Supplementary Material. This is achieved via the Jordan decomposition of A_* in (17), and the mildest condition we need is Assumption 5. The upper bound of κ under Assumption 5 or 6 is given in the Supplementary Material.

Remark 6. For ξ defined in (11), $\|\Sigma_X\|_2$ depends on n, as $\Sigma_X = [E(X_tX_s^{\mathrm{T}})]_{1\leqslant t,s\leqslant n}$, where $E(X_tX_s^{\mathrm{T}}) = \sigma^2 A_*^{t-s}\Gamma_s$ for $1 \leqslant s \leqslant t \leqslant n$ under Assumptions 4(i) and (ii). Unlike κ discussed in Remark 5, even under Assumption 6, $\|\Sigma_X\|_2$ is not guaranteed to be bounded by a constant free of n; indeed, we need Assumption 6′ for this purpose, since $\|\Sigma_X\|_2$ is affected by not only the growing diagonal blocks $\sigma^2\Gamma_1, \ldots, \sigma^2\Gamma_n$, but also the growing off-diagonal blocks; see the Supplementary Material for details. The upper bound of ξ under Assumption 5 or 6′ is given in the Supplementary Material.

Let the Jordan decomposition of A_* be

$$A_* = SJS^{-1}, \tag{17}$$

where J has L blocks with sizes $1 \leqslant b_1, \ldots, b_L \leqslant d$, and both J and S are $d \times d$ complex matrices. Let $b_{\max} = \max_{1\leqslant\ell\leqslant L} b_\ell$, and denote the condition number of S by $\operatorname{cond}(S) = \{\lambda_{\max}(S^*S)/\lambda_{\min}(S^*S)\}^{1/2}$, where S^* is the conjugate transpose of S.

The following proposition, which follows from (14), (15) and upper bounds of κ and ξ under Assumptions 5, 6 or 6′, is proved in the Supplementary Material.

PROPOSITION 2. *For any $A_* \in \mathbb{R}^{d\times d}$, under Assumption 5 we have $\log\det(\overline{\Gamma}_R\underline{\Gamma}_R^{-1}) \lesssim m\,[\log\{d\operatorname{cond}(S)/\delta\} + b_{\max}\log n]$ for $\overline{\Gamma}_R = \overline{\Gamma}_R^{(1)}$ or $\overline{\Gamma}_R^{(2)}$. Moreover, if Assumption 6 holds then $\log\det(\overline{\Gamma}_R^{(1)}\underline{\Gamma}_R^{-1}) \lesssim m\log(m/\delta)$. Furthermore, if Assumption 6′ holds and $n \gtrsim m + \log(d/\delta)$, then $\log\det(\overline{\Gamma}_R^{(2)}\underline{\Gamma}_R^{-1}) \lesssim m$.*

The condition $n \gtrsim m + \log(d/\delta)$ in Proposition 2 is not stringent, because it is necessary for condition (5) in Theorem 1, where $q = d$ for the VAR(1) model. By Proposition 2, $\overline{\Gamma}_R^{(2)}$ yields a sharper upper bound of $\log\det(\overline{\Gamma}_R\underline{\Gamma}_R^{-1})$ than does $\overline{\Gamma}_R^{(1)}$ only under Assumption 6′. Thus, we shall always set $\overline{\Gamma}_R = \overline{\Gamma}_R^{(1)}$ unless Assumption 6′ holds and $\{\eta_t\}$ are normal. As a result, under Assumption 4, the feasible region of k that is sufficient for condition (5) in Theorem 1 is

$$k \lesssim \begin{cases} \dfrac{n}{m[\log\{d\operatorname{cond}(S)/\delta\} + b_{\max}\log n]}, & \text{if Assumption 5 holds,} \\ \dfrac{n}{m\log(m/\delta) + \log d}, & \text{if Assumption 6 holds,} \\ \dfrac{n}{m + \log(d/\delta)}, & \text{if Assumption 6′ holds and } \{\eta_t\} \text{ are normal.} \end{cases} \tag{18}$$

Because the upper bound of $\log\det(\overline{\Gamma}_R\underline{\Gamma}_R^{-1})$ and the feasible region of k are both dependent on the assumption on A_* and whether $\{\eta_t\}$ are normal, the resulting estimation error bounds will vary slightly under different conditions; see Theorems 2 and 3 in § 3.4.

3.4. *Analysis of upper bounds in vector autoregression*

We focus on the upper bound analysis of $\|\hat{\beta} - \beta_*\|$; nevertheless, from Proposition 1 we can readily obtain analogous results for $\|\hat{A} - A_*\|_2$, which are omitted here. For simplicity, denote

$$\Gamma_{R,k} = R\left\{R^{\mathrm{T}}(I_d \otimes \Gamma_k)R\right\}^{-1} R^{\mathrm{T}}.$$

The first theorem follows directly from Theorem 1, Lemmas 1–3 and Proposition 2.

THEOREM 2. *Let* $\{X_t\}_{t=1}^{n+1}$ *be generated by the linearly restricted vector autoregressive model. Fix* $\delta \in (0,1)$. *For any* $1 \leqslant k \leqslant \lfloor n/2 \rfloor$ *satisfying* (18), *under Assumption* 4, *we have the following results:* (i) *If Assumption* 5 *holds, with probability at least* $1 - 3\delta$,

$$\|\hat{\beta} - \beta_*\| \lesssim \left(\lambda_{\max}(\Gamma_{R,k})\frac{m\left[\log\{d\,\mathrm{cond}(S)/\delta\} + b_{\max}\log n\right]}{n}\right)^{1/2}.$$

(ii) *If Assumption* 6 *holds, with probability at least* $1 - 3\delta$,

$$\|\hat{\beta} - \beta_*\| \lesssim \left\{\lambda_{\max}(\Gamma_{R,k})\frac{m\log(m/\delta)}{n}\right\}^{1/2}.$$

(iii) *If Assumption* 6′ *holds and* $\{\eta_t\}$ *are normal, with probability at least* $1 - 3\delta$,

$$\|\hat{\beta} - \beta_*\| \lesssim \left\{\lambda_{\max}(\Gamma_{R,k})\frac{m + \log(1/\delta)}{n}\right\}^{1/2}.$$

To gain an intuitive understanding of the factor $\lambda_{\max}(\Gamma_{R,k})$ in Theorem 2, consider the asymptotic distribution of $\hat{\beta}$ under the assumptions that $\rho(A_*) < 1$ and that d, m and A_* are all fixed:

$$n^{1/2}(\hat{\beta} - \beta_*) \to N\big[0, \underbrace{R\{R^{\mathrm{T}}(I_d \otimes \Gamma_\infty)R\}^{-1}R^{\mathrm{T}}}_{\Gamma_{R,\infty}}\big] \tag{19}$$

in distribution as $n \to \infty$, where $\Gamma_\infty = \lim_{k\to\infty}\Gamma_k$; see Lütkepohl (2005). Thus, $\lambda_{\max}(\Gamma_{R,k})$ in Theorem 2 resembles the limiting covariance matrix $\Gamma_{R,\infty}$. However, by adopting a nonasymptotic approach, Theorem 2 retains the dependence of the estimation error on $\Gamma_{R,k}$ across stable, unstable and slightly explosive regimes. Similarly to (19), the error bounds in Theorem 2 are free of σ^2, as the scaling effect of σ^2 on η_t is cancelled out by that on X_t due to the autoregressive structure.

As a special case, $b_{\max} = 1$ if A_* is diagonalizable. Moreover, if $A_* = \rho I_d$, then $b_{\max} = 1$, $\mathrm{cond}(S) = 1$, $\Gamma_k = \gamma_k(\rho)I_d$, with $\gamma_k(\rho) = \sum_{s=0}^{k-1}\rho^{2s}$, and thus

$$\lambda_{\max}(\Gamma_{R,k}) = \gamma_k^{-1}(\rho)\lambda_{\max}\{R(R^{\mathrm{T}}R)^{-1}R^{\mathrm{T}}\} = \gamma_k^{-1}(\rho)\lambda_{\max}\{(R^{\mathrm{T}}R)^{-1}R^{\mathrm{T}}R\} = \gamma_k^{-1}(\rho), \tag{20}$$

where the second equality is due to the fact that, for any matrices $A \in \mathbb{R}^{N\times m}$ and $B \in \mathbb{R}^{m\times N}$, AB and BA have the same nonzero eigenvalues (Theorem 1.3.20, Horn & Johnson, 1985).

By Theorem 2, the linear restrictions affect the error bounds through both the factor $\lambda_{\max}(\Gamma_{R,k})$ and the explicit rate function of m and n. To further illustrate this, suppose that

$$\beta_* = R\theta_* + \gamma = R^{(1)}R^{(2)}\theta_* + \gamma,$$

where $R^{(1)} \in \mathbb{R}^{d^2 \times \tilde{m}}$ has rank $\tilde{m}$, and $R^{(2)} \in \mathbb{R}^{\tilde{m} \times m}$ has rank m, with $\tilde{m} \geqslant m + 1$. Then $\mathcal{L}^{(1)} = \{R^{(1)}\theta + \gamma : \theta \in \mathbb{R}^{\tilde{m}}\} \supseteq \mathcal{L} = \{R\theta + \gamma : \theta \in \mathbb{R}^m\}$. By an argument similar to that in Lütkepohl (2005, p. 199), we can show that $\Gamma_{R,k} \preceq \Gamma_{R^{(1)},k}$, so

$$\lambda_{\max}(\Gamma_{R,k}) \leqslant \lambda_{\max}(\Gamma_{R^{(1)},k}). \tag{21}$$

The parameter space $\mathcal{L}^{(1)}$ has fewer restrictions than $\mathcal{L}$. Therefore, with fewer restrictions, the effective model size will increase from m to $\tilde{m}$, and meanwhile $\lambda_{\max}(\Gamma_{R,k})$ will increase to $\lambda_{\max}(\Gamma_{R^{(1)},k})$, both leading to deterioration of the error bound.

Remark 7. The preservation of the factor $\lambda_{\max}(\Gamma_{R,k})$ in Theorem 2 is achieved by bounding $Z^{\mathrm{T}}Z$ and $Z^{\mathrm{T}}\eta$ simultaneously through the Moore–Penrose pseudoinverse $Z^{\dagger}$, where $Z^{\dagger} = (Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}$ if $Z^{\mathrm{T}}Z \succ 0$; see also Simchowitz et al. (2018). This key advantage is not enjoyed by the non-asymptotic analyses in Basu & Michailidis (2015) and Faradonbeh et al. (2018). In their analyses, $X^{\mathrm{T}}X$ and $X^{\mathrm{T}}E$, or $Z^{\mathrm{T}}Z$ and $Z^{\mathrm{T}}\eta$ in our context, were bounded separately. This would not only break down $\Gamma_{R,k}$, but also cause degradation of the error bound as $\rho(A_*) \to 1$ due to the inevitable involvement of the condition number of $X^{\mathrm{T}}X$ in the resulting error bound.

Remark 8. If $A_* = \rho I_d$, then (21) becomes an equality. However, the equality generally does not hold even for diagonal matrices A_*. For example, if $A_* = \mathrm{diag}(\rho_1, \rho_2) \in \mathbb{R}^{2\times 2}$, where $|\rho_1| > |\rho_2|$, then $\Gamma_k = \mathrm{diag}\{\gamma_k(\rho_1), \gamma_k(\rho_2)\}$. Let $R = (1, 0, 0, 0)^{\mathrm{T}} = R^{(1)}R^{(2)}$, where $R^{(1)} = (I_2, 0)^{\mathrm{T}} \in \mathbb{R}^{4\times 2}$ and $R^{(2)} = (1, 0)^{\mathrm{T}}$. Consequently, (21) is a strict inequality: $\lambda_{\max}(\Gamma_{R,k}) = \gamma_k^{-1}(\rho_1) < \gamma_k^{-1}(\rho_2) = \lambda_{\max}(\Gamma_{R^{(1)},k})$.

The next theorem sharpens the error bounds in Theorem 2 by utilizing the largest possible k, since $\lambda_{\max}(\Gamma_{R,k})$ is monotonic decreasing in k. The dependence of $\lambda_{\max}(\Gamma_{R,k})$ on A_* will be captured by $\sigma_{\min}(A_*)$, a measure of the least excitable mode of the underlying dynamics.

THEOREM 3. *Suppose that the conditions of Theorem 2 hold. Fix $\delta \in (0, 1)$, and let $c_1 > 0$ be a universal constant.*

(i) *Under Assumption 5, if*

$$\sigma_{\min}(A_*) \leqslant 1 - \frac{c_1 m\,[\log\{d\,\mathrm{cond}(S)/\delta\} + b_{\max}\log n]}{n}, \tag{22}$$

then, with probability at least $1 - 3\delta$,

$$\|\hat{\beta} - \beta_*\| \lesssim \left(\frac{\{1 - \sigma_{\min}^2(A_*)\}m\,[\log\{d\,\mathrm{cond}(S)/\delta\} + b_{\max}\log n]}{n}\right)^{1/2}, \tag{23}$$

and if inequality (22) holds in the reverse direction, then, with probability at least $1 - 3\delta$,

$$\|\hat{\beta} - \beta_*\| \lesssim \frac{m\,[\log\{d\,\mathrm{cond}(S)/\delta\} + b_{\max}\log n]}{n}. \tag{24}$$

(ii) *Under Assumption 6, if*

$$\sigma_{\min}(A_*) \leqslant 1 - \frac{c_1\{m\log(m/\delta) + \log d\}}{n}, \tag{25}$$

then, with probability at least $1-3\delta$,

$$\|\hat{\beta}-\beta_*\| \lesssim \left[\frac{\{1-\sigma_{\min}^2(A_*)\}m\log(m/\delta)}{n}\right]^{1/2}, \tag{26}$$

and if inequality (25) *holds in the reverse direction, then, with probability at least* $1-3\delta$,

$$\|\hat{\beta}-\beta_*\| \lesssim \frac{m\log(m/\delta)+\log d}{n}. \tag{27}$$

(iii) *Under Assumption* 6′, *if* $\{\eta_t\}$ *are normal and*

$$\sigma_{\min}(A_*) \leqslant 1-\frac{c_1\{m+\log(d/\delta)\}}{n}, \tag{28}$$

then, with probability at least $1-3\delta$,

$$\|\hat{\beta}-\beta_*\| \lesssim \left[\frac{\{1-\sigma_{\min}^2(A_*)\}\{m+\log(1/\delta)\}}{n}\right]^{1/2}, \tag{29}$$

and if inequality (28) *holds in the reverse direction, then, with probability at least* $1-3\delta$,

$$\|\hat{\beta}-\beta_*\| \lesssim \frac{m+\log(d/\delta)}{n}. \tag{30}$$

Theorem 3 reveals an interesting phenomenon of phase transition from slow to fast error rate regimes, i.e., from about $O\{(m/n)^{1/2}\}$ as in (23), (26) and (29) to about $O(m/n)$ as in (24), (27) and (30), up to logarithmic factors. Within the slow-rate regime, the estimation error decreases as $\sigma_{\min}(A_*)$ increases. The slow rates in (23), (26) and (29) differ from each other only by logarithmic factors, and so do the fast rates in (24), (27) and (30). Moreover, the point at which the transition occurs is dependent on $\sigma_{\min}(A_*)$ instead of $\rho(A_*)$; see conditions (22), (25) and (28). Since $\sigma_{\min}(A_*) \leqslant \rho(A_*)$, conditions (22), (26) and (28) may be mild as long as $\rho(A_*)$ is not too large. However, the fast rates require the opposite of (22), (25) and (28), which cannot be directly inferred from $\rho(A_*)$.

Remark 9. For the special case of $A_* = \rho I_d$ with $\rho \in \mathbb{R}$, Assumptions 6 and 6′ both simply reduce to $|\rho| < 1$, and Assumption 5 to $|\rho| \leqslant 1 + O(1/n)$. Also, $\mathrm{cond}(S) = 1$ and $b_{\max} = 1$. Thus, under Assumption 4, by (24), (26) and (27), the following holds with high probability:

$$\|\hat{\beta}-\beta_*\| \lesssim \begin{cases} O[\{(1-\rho^2)m\log m/n\}^{1/2}], & \text{if } |\rho| \leqslant 1-O\{(m\log m+\log d)/n\}, \\ O\{(m\log m+\log d)/n\}, & \text{if } 1-O\{(m\log m+\log d)/n\} \leqslant |\rho| < 1, \\ O\{m\log(dn)/n\}, & \text{if } 1 \leqslant |\rho| \leqslant 1+O(1/n). \end{cases}$$

Moreover, if $\{\eta_t\}$ are normal, by (29) and (30), we can eliminate all factors of $\log m$ in the above results; that is, every $m\log m$ will be replaced by m.

Remark 10. Faradonbeh et al. (2018) derived error bounds for the ordinary least squares estimator of explosive unrestricted vector autoregressive processes when (i) $|\lambda_{\min}(A_*)| > 1$ or

(ii) A_* has no unit eigenvalue. In contrast to case (i), we focus on slightly explosive processes with $\rho(A_*) \leqslant 1 + O(1/n)$. Moreover, the no unit root requirement of case (ii) may be quite restrictive; e.g., it excludes the case of $|\rho| = 1$ in Remark 9. Thus, the conditions in Theorem 3 may be more reasonable in practice.

Remark 11. Guo et al. (2016) obtained the error rate $O_p\{(m/n)^{1/2}\}$ for the banded vector autoregressive model under weaker conditions on $\{\eta_t\}$ yet stronger conditions on A_* than those in this paper; see Theorem 2 therein. In particular, they required $\|A_*\|_2 < 1$. This rate matches (29). In view of the lower bounds to be presented in § 4, we conjecture that the rate in (26) is larger than the actual rate by a factor of $\log m$.

4. Analysis of lower bounds

For any $\theta \in \mathbb{R}^m$, let $\beta = R\theta + \gamma$, and the corresponding transition matrix is denoted by $A(\theta) = (I_d \otimes \theta^{\mathrm{T}})\tilde{R} + G$, where $R, \gamma, \tilde{R}$ and G are defined as in § 3.1. As β is completely determined by θ, it is more convenient to index the probability law of the model by the unrestricted parameter θ. Thus, we denote by $\mathrm{pr}_\theta^{(n)}$ the distribution of the sample $(X_1, \ldots, X_{n+1})$ on $(\mathcal{X}^{n+1}, \mathcal{F}_{n+1})$, where $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{F}_{n+1} = \sigma\{\eta_1, \ldots, \eta_n\}$. For any fixed $\bar{\rho} > 0$, we write the subspace of θ such that the spectral radius of $A(\theta)$ is bounded above by $\bar{\rho}$ as

$$\Theta(\bar{\rho}) = \{\theta \in \mathbb{R}^m : \rho\{A(\theta)\} \leqslant \bar{\rho}\}.$$

The corresponding linearly restricted subspace of β is denoted by $\mathcal{L}(\bar{\rho}) = \{R\theta + \gamma : \theta \in \Theta(\bar{\rho})\}$.

The minimax rate of estimation over $\beta \in \mathcal{L}(\bar{\rho})$, or $\theta \in \Theta(\bar{\rho})$, is provided by the next theorem.

Theorem 4. *Suppose that $\{X_t\}_{t=1}^{n+1}$ follow the vector autoregressive model $X_{t+1} = AX_t + \eta_t$ with linear restrictions defined as in § 3. In addition, Assumptions 4*(i) *and* (ii) *hold, and $\{\eta_t\}$ are normally distributed. Fix $\delta \in (0, 1/4)$ and $\bar{\rho} > 0$. Let $\gamma_n(\bar{\rho}) = \sum_{s=0}^{n-1} \bar{\rho}^{2s}$. Then, for any $\epsilon \in (0, \bar{\rho}/4]$, we have*

$$\inf_{\hat{\beta}} \sup_{\theta \in \Theta(\bar{\rho})} \mathrm{pr}_\theta^{(n)} \left\{ \|\hat{\beta} - \beta\| \geqslant \epsilon \right\} \geqslant \delta,$$

where the infimum is taken over all estimators of β subject to $\beta \in \{R\theta + \gamma : \theta \in \mathbb{R}^m\}$, for any n such that

$$n\gamma_n(\bar{\rho}) \lesssim \frac{m + \log(1/\delta)}{\epsilon^2}.$$

As a result, we have the following minimax rates of estimation across different values of $\bar{\rho}$.

Corollary 1. *For the linearly restricted vector autoregressive model in Theorem 4, the minimax rates of estimation over $\beta \in \mathcal{L}(\bar{\rho})$ are given as follows:*

(i) $\{(1 - \bar{\rho}^2)m/n\}^{1/2}$, *if* $0 < \bar{\rho} < (1 - 1/n)^{1/2}$;

(ii) $m^{1/2}/n$, *if* $(1 - 1/n)^{1/2} \leqslant \bar{\rho} \leqslant 1 + c/n$ *for a fixed* $c > 0$; *and*

(iii) $\bar{\rho}^{-n}\{(\bar{\rho}^2 - 1)m/n\}^{1/2}$, *if* $\bar{\rho} > 1 + c/n$.

While the phase transition in Corollary 1 depends on $\rho(A_*)$ instead of $\sigma_{\min}(A_*)$, we may still compare the lower bounds to the upper bounds in Theorem 3. For case (i) in Corollary 1, since

Table 1. *Comparison of upper and lower bounds*

Range of $\lvert\rho\rvert$	Lower bound	Upper bound
$(0, 1 - O\{(m + \log d)/n\}]$	$\Omega[\{(1-\rho^2)m/n\}^{1/2}]$	$O[\{(1-\rho^2)m/n\}^{1/2}]$
$[1 - O\{(m + \log d)/n\}, (1 - 1/n)^{1/2})$	$\Omega[\{(1-\rho^2)m/n\}^{1/2}]$	$O\{(m + \log d)/n\}$
$[(1 - 1/n)^{1/2}, 1)$	$\Omega(m^{1/2}/n)$	$O\{(m + \log d)/n\}$
$[1, 1 + O(1/n)]$	$\Omega(m^{1/2}/n)$	$O\{m \log(dn)/n\}$
$(1 + O(1/n), \infty)$	$\Omega[\lvert\rho\rvert^{-n}\{(\rho^2 - 1)m/n\}^{1/2}]$	—

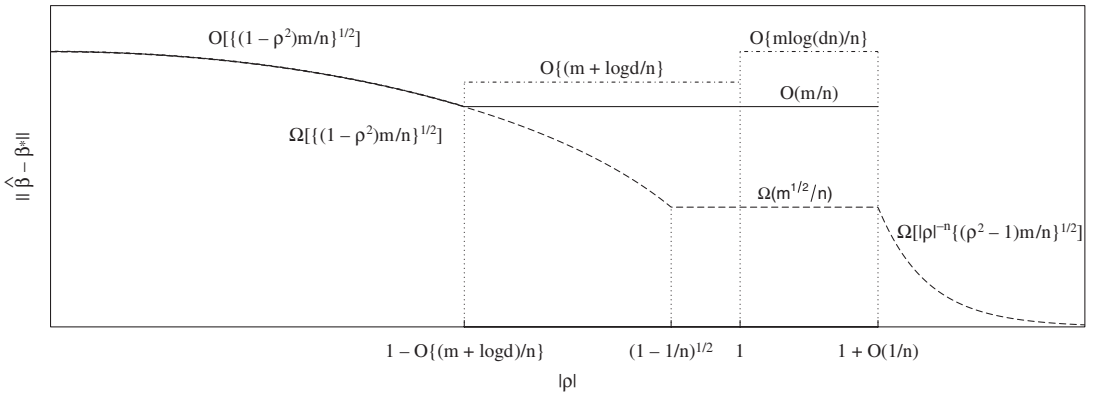

Fig. 1. Illustration of theoretical upper (dot-dashed) and lower (dashed) bounds and the actual rates (solid) suggested by simulation results in § 5 for the VAR(1) model with $A_* = \rho I_d$ and normal innovations.

$\sigma_{\min}(A_*) \leqslant \rho(A_*) < (1 - 1/n)^{1/2} < 1$, we may expect that condition (25) will hold in most cases, and hence the upper bound $O\{(m \log m/n)^{1/2}\}$ differs from the lower bound by a factor of $\log m$. However, for case (ii), the upper bound may lie in either the slow- or fast-rate regime, depending on the magnitude of $\sigma_{\min}(A_*)$, whereas the lower bound lies in the fast-rate regime. As shown by our first experiment in § 5, the transition from slow to fast error rates actually depends on $\sigma_{\min}(A_*)$ instead of $\rho(A_*)$. This also suggests that the results in Theorem 3 are sharp in the sense that they correctly capture the transition behaviour.

Remark 12. If $A_* = \rho I_d$ with $\rho \in \mathbb{R}$ and $\{\eta_t\}$ are normal, in view of Remark 9 and Corollary 1, a more straightforward comparison of the upper and lower bounds can be made; see Table 1. See Fig. 1 for an illustration of the theoretical bounds and actual rates suggested by simulation results in § 5. The actual rates and the theoretical upper and lower bounds exactly match when $0 < |\rho| \leqslant 1 - O\{(m + \log d)/n\}$. In addition, the suggested actual rate is m/n for $1 - O\{(m + \log d)/n\} < |\rho| < 1 + O(1/n)$ and even faster for $|\rho|$ beyond this range.

Remark 13. Han et al. (2015b) considered the estimation of a class of copula-based stationary vector autoregressive processes which includes the Gaussian VAR(1) process as a special case, and extended the theoretical properties to VAR(p) processes by arguments similar to those in Example 3. Under a strong-mixing condition on the process and the low-rank assumption $\text{rank}(A_*) \leqslant r$, the proposed estimator $\tilde{A}$ was proved to attain the minimax error rate $\|\tilde{A} - A_*\|_{\text{F}} = O\{(dr/n)^{1/2}\}$. While we consider different restrictions and estimation methods, the rate $\{(1 - \bar{\rho}^2)m/n\}^{1/2}$ in Corollary 1 for the stable regime resembles that in Han et al. (2015b) if we regard dr as the effective model size m of the low-rank VAR(1) model. However, similarly

to our upper bound analysis in § 3, the factor of $(1-\bar{\rho}^2)^{1/2}$ in our lower bound also reveals that the estimation error may decrease as A_* approaches the stability boundary.

5. Simulation experiments

We conduct three simulation experiments to verify the theoretical results in the previous sections, including the estimation error rates, the transition from slow- to fast-rate regimes, and the effect of the ambient dimension d on the estimation. The data are generated from VAR(1) processes with $\{\eta_t\}$ drawn independently from $N(0, I_d)$ and the following structures of A_*:

DGP1: Banded structure defined by the zero restrictions $a_{*ij} = 0$ if $|i-j| > k_0$, where $1 \leqslant k_0 \leqslant \lfloor (d-1)/2 \rfloor$ is the bandwidth parameter; see Example 4 in § 3.1. As a result, if all restrictions are imposed, the size of the model is $m = d + (2d-1)k_0 - k_0^2$.

DGP2: Group structure defined by equality restrictions as follows. Partition the index set $\mathcal{V} = \{1, \ldots, d\}$ of the coordinates of X_t into K groups of size $b = d/K$ as $\mathcal{V} = \bigcup_{k=1}^{K} \mathcal{G}_k$, where

$$\mathcal{G}_k = \{(k-1)b + 1, \ldots, kb\} \qquad (k = 1, \ldots, K).$$

In each row of A_*, the off-diagonal entries a_{*ij} with j belonging to the same group are assumed to be equal: for any $1 \leqslant k \leqslant K$ and $1 \leqslant i \leqslant d$, all elements of $\{a_{*ij}, j \in \mathcal{G}_k, j \neq i\}$ are equal. Thus, $m = (K+1)d$, as there are $(K+1)$ free parameters in each row of A_*.

DGP3: $A_* = \rho I_d$, where $\rho \in \mathbb{R}$. The smallest true model with size $m = 1$ results from imposing zero restrictions on all off-diagonal entries of A_* and equality restrictions on all diagonal entries; see Example 6 in § 3.1.

Throughout the experiments, the ℓ_2 estimation error $\|\hat{\beta} - \beta_*\|$ is calculated by averaging over 1000 replications. Except for DGP3, nonzero entries of A_* are generated independently from $U[-1, 1]$ and then rescaled such that $\rho(A_*)$ is equal to a certain value.

The first experiment aims to verify the error rates in Theorem 3 and the implication of Theorem 2 that the restrictions can reduce the estimation error through both the explicit rate $(m/n)^{1/2}$ and the decrease in the factor $\lambda_{\max}(\Gamma_{R,k})$. Fixing $d = 24$ and $\rho(A_*) = 0.2$, 0.8 or 1, we generate data from DGP1 with $k_0 = 1$, DGP2 with $K = 2$, and DGP3. For DGP1 and DGP3, we fit banded vector autoregressive models with $k_0 = 1$, 5 or 7 such that $m = 70$, 156 or 304, respectively. For DGP2, we fit the group-structured model with $K = 2$, 8 or 12 such that $m = 72$, 120 or 312, respectively. For DGP1 and DGP2, $\sigma_{\min}(A_*) \leqslant 0.1$ even when the randomly generated matrix A_* has $\rho(A_*) = 1$. However, for DGP3, $\sigma_{\min}(A_*) = \rho(A_*)$. The ℓ_2 estimation error $\|\hat{\beta} - \beta_*\|$ is plotted against $(m/n)^{1/2}$ in Fig. 2, where we consider $(m/n)^{1/2} \in \{0.15, 0.35, 0.55, 0.75, 0.95\}$. Our findings are summarized below.

(i) When $\rho(A_*) = 0.2$, for all the data-generating processes the lines for different m coincide completely with each other and scale perfectly linearly with $(m/n)^{1/2}$. This suggests that the actual error rate is $(m/n)^{1/2}$ when $\sigma_{\min}(A_*)$ lies in the slow-rate regime of Theorem 3.

(ii) When $\rho(A_*) = 0.8$ or 1, for DGP1 and DGP2, although $\|\hat{\beta} - \beta_*\|$ is still proportional to $(m/n)^{1/2}$, the three lines for the same $\rho(A_*)$, but different m do not coincide: fixing $\rho(A_*)$, the slope increases as m increases, and the variation in slope is greater as $\rho(A_*)$ is larger. For DGP1 and DGP2, $\sigma_{\min}(A_*)$ is very small. As finding (i) suggests that the actual error rate is $(m/n)^{1/2}$ for small $\sigma_{\min}(A_*)$, this extra variation in slope may be partially explained by the factor $\lambda_{\max}(\Gamma_{R,k})$ in the error bound in Theorem 2 due to the effect of R. However, $\|\hat{\beta} - \beta_*\|$ actually depends on the spectrum of $\Gamma_{R,k}$, and its largest eigenvalue merely serves as an upper bound. When $\rho(A_*)$

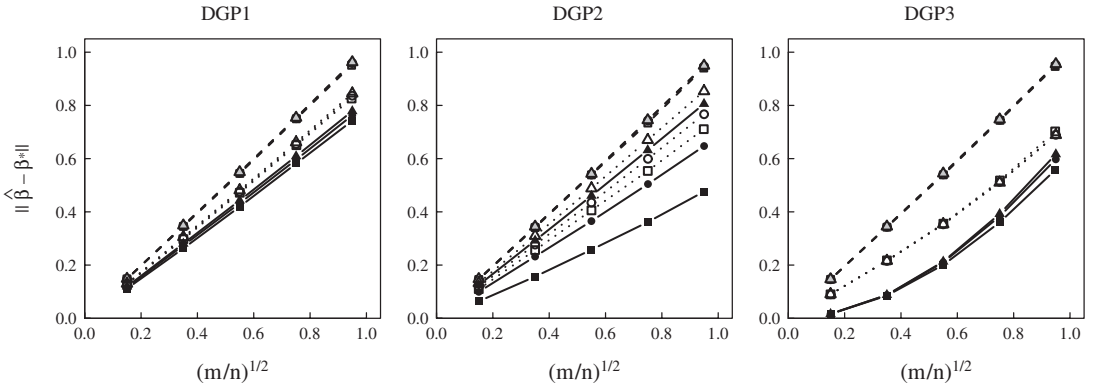

Fig. 2. Plots of $\|\hat{\beta} - \beta_*\|$ against $(m/n)^{1/2}$ for three data-generating processes with $\rho(A_*) = 0.2$ (dashed lines, grey-filled symbols), 0.8 (dotted lines, unfilled symbols) or 1 (solid lines, black-filled symbols) and different m. DGP1 and DGP3 were fitted as banded vector autoregressive models with $m = 70$ (squares), 156 (circles) or 304 (triangles), and DGP2 was fitted as grouped vector autoregressive models with $m = 72$ (squares), 120 (circles) or 312 (triangles).

is smaller, we will have more control over the spectrum of A_*, and hence that of $\Gamma_{R,k}$. This may explain why the variation in slope is smaller when $\rho(A_*)$ is smaller.

(iii) For DGP3 with $\rho(A_*) = 0.2$ or 0.8, the three lines corresponding to different m still completely coincide with each other. This can be explained by the fact that $\lambda_{\max}(\Gamma_{R,k})$ is independent of R when $A_* = \rho I_d$; see (20).

(iv) For DGP3 with $\rho(A_*) = 1$, in sharp contrast to all other cases, the error rate appears to be a quadratic function of $(m/n)^{1/2}$. This matches the implication of Theorem 3 that when $\sigma_{\min}(A_*) = 1$, the error rate falls into the fast-rate regime.

(v) Fixing both m and n, $\|\hat{\beta} - \beta_*\|$ always decreases as $\rho(A_*)$ increases. Moreover, when $\sigma_{\min}(A_*) < 1$, fixing m, the lines become less steep as $\rho(A_*)$ increases. Note that $\sigma_{\min}(A_*)$ is larger when $\rho(A_*)$ is, due to our method of generating A_*. Thus, this finding can be explained by the factor $\{1 - \sigma^2_{\min}(A_*)\}^{1/2}$ in the error bound for the slow-rate regime in Theorem 3.

In the second experiment, we focus on DGP3 to further investigate the error rates and the phase transition. We set $d = 24$ and $m = 1$, 70, 156 or 304, where $m = 1$ results from fitting the smallest true model, and $m = 70$, 156 or 304 from fitting a banded model with $k_0 = 1$, 3 or 7, respectively. Figures 3 and 4 display the results, where we have the following findings:

(i) The combination of results from the first experiment and Fig. 3(a) suggests that $\|\hat{\beta} - \beta_*\|$ scales as $O[\{(1 - \rho^2)m/n\}^{1/2}]$ when $|\rho|$ is fixed at a level well below one.

(ii) Figure 4 suggests that when $|\rho| = 1$ the actual error rate is m/n. Specifically, when m is fixed, Fig. 4(a) shows that $n\|\hat{\beta} - \beta_*\|$ becomes stable for n sufficiently large, while $\|\hat{\beta} - \beta_*\|$ multiplied by $n^{1/2}$ or $n/\log n$ appears to diminish as $n \to \infty$. On the other hand, when n is fixed, Fig. 4(b) shows that $\|\hat{\beta} - \beta_*\|/m$ becomes stable for m sufficiently large.

(iii) Figure 3(b) suggests that the regime of rate m/n is reached as early as $|\rho| = 1 - O\{(m + \log d)/n\}$ and maintained even as the process becomes slightly explosive with $|\rho| = 1 + O(1/n)$. This lends support to the boundaries of the fast-rate regime suggested by Theorem 3; see Remarks 9 and 12. By contrast, when $\rho = 0.99$ the rate appears to be $(m/n)^{1/2}$, similar to our findings in the first experiment. On the other hand, when ρ is fixed at a level slightly above 1, the rate becomes even faster than m/n. This matches the conclusion in Remark 12 that the corresponding lower bound diminishes at a rate faster than $|\rho|^n$ as n increases.

The third experiment aims to check whether the ambient dimension d directly affects the estimation error. We generate data from DGP3 with $\rho = 0.2$, 0.8 or 1, $n = 100$ or 500 and

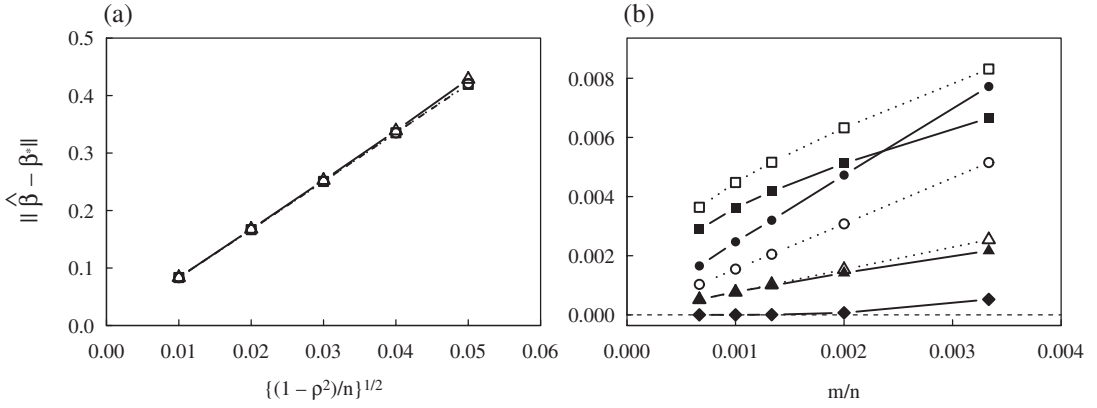

Fig. 3. Error rates for DGP3 as ρ is fixed or approaching 1 at different rates. (a) Plot of $\|\hat{\beta} - \beta_*\|$ against $\{(1-\rho^2)/n\}^{1/2}$ with $\rho = 0.2$ (dashed lines, squares), 0.4 (dotted lines, circles) or 0.6 (solid lines, triangles), and $m = 70$. (b) Plot of $\|\hat{\beta} - \beta_*\|$ against m/n with $\rho = 0.99$ (squares), $1 - (m + \log d)/n$ (circles), $1 + 1/n$ (triangles) or 1.01 (diamonds), and $m = 1$ (solid lines, filled symbols) or 70 (dotted lines, unfilled symbols). The case of $(m, \rho) = (70, 1.01)$ is omitted as the process becomes so explosive that the computation is numerically infeasible.

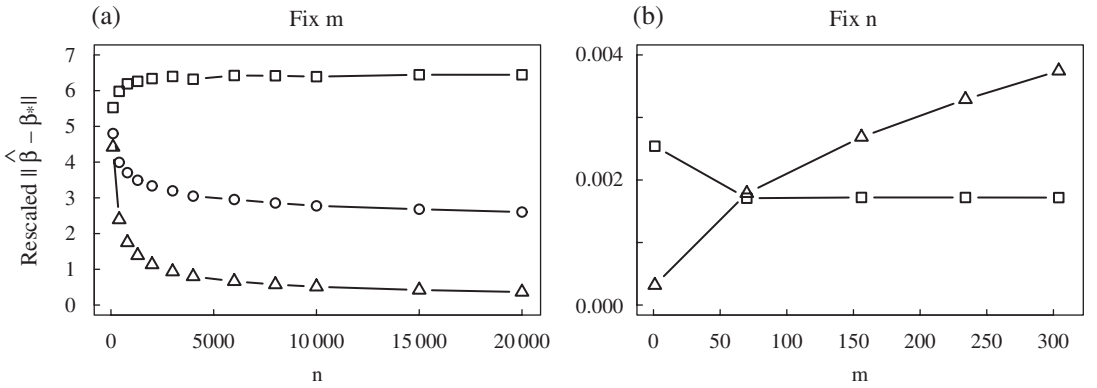

Fig. 4. Error rates for DGP3 when $\rho = 1$. (a) Plot of $\|\hat{\beta} - \beta_*\|$ multiplied by $n/8$ (squares), $n/(2\log n)$ (circles) or $n^{1/2}$ (triangles) against n, fixing $m = 70$. (b) Plot of $\|\hat{\beta} - \beta_*\|$ divided by m (squares) or $8m^{1/2}$ (triangles) against m, fixing $n = 400$.

$d \in [25, 500]$. For the estimation, we consider $m = 1$ or 20, where $m = 1$ corresponds to the smallest true model, and $m = 20$ corresponds to a model subject to (i) $a_{*11} = \cdots = a_{*dd}$, and (ii) the restriction that all but $m - 1$ of the off-diagonal entries of A_* are zero. To generate the pattern in (ii), we sample the $m - 1$ positions uniformly without replacement from all off-diagonal positions of A_*.

Figure 5 shows that the estimation error is constant in d for almost all cases. This confirms that d does not affect the estimation error when $|\rho| < 1 - O\{(m + \log d)/n\}$; see Remark 12. Moreover, the extra factor of $\log d$ in the theoretical upper bounds for the other regimes might not be necessary. For $(n, m, \rho) = (100, 20, 1)$, the estimation error seems less stable when $d \leqslant 75$; see Fig. 5(a). This might be explained by the indirect effect of R on $\lambda_{\max}(\Gamma_{R,k})$ in Theorem 2. As m is fixed at 20, different d corresponds to different R. The spectrum of $\Gamma_{R,k}$ may be more sensitive to R when d is smaller, and the resulting impact on the estimation error may be more pronounced when n is smaller. However, as d grows, the restrictions will become relatively more sparse, so eventually the spectrum of $\Gamma_{R,k}$ will be stable, and the impact of d will be negligible.

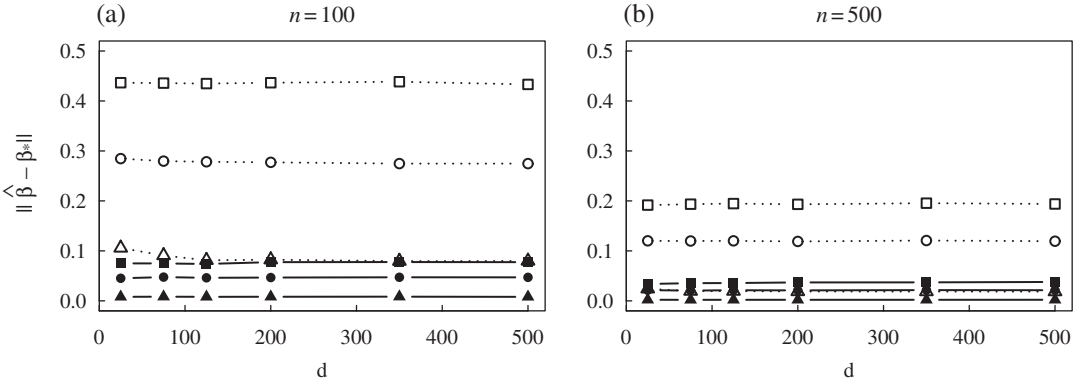

Fig. 5. Plots of $\|\hat{\beta} - \beta_*\|$ against d for DGP3 with (a) $n = 100$ and (b) $n = 500$, when $\rho = 0.2$ (squares), 0.8 (circles) or 1 (triangles), and $m = 1$ (solid lines, filled symbols) or 20 (dotted lines, unfilled symbols).

6. DISCUSSION

An interesting future direction is dimensionality reduction for vector autoregressive models with data-driven restrictions. Such a procedure involves first suggesting possible restrictions based on subject knowledge and then selecting the true restrictions by a data-driven approach. The lasso method (Basu & Michailidis, 2015; Davis et al., 2015) can be viewed as a procedure where zero restrictions are initially suggested for all entries of A, and then the true zeros are identified by penalized estimation. Adopting a more general point of view, the modeller can initially suggest the general linear restrictions (1) instead. This will enable a more flexible and data-driven integration of expert knowledge. On the other hand, if it is known that only zero and equality restrictions are true, yet the locations of the restrictions are unknown, we can select the true restrictions efficiently by the delete or merge regressors algorithm proposed by Maj-Kańska et al. (2015) based on the Bayesian information criterion. The consistency of this procedure can be easily extended to vector autoregressive models.

ACKNOWLEDGEMENT

We thank the editor, associate editor and two referees for their invaluable comments, which have led to substantial improvements of our paper. Cheng's research was partially supported by the U.S. National Science Foundation, an Adobe Data Science Faculty Award and the Office of Naval Research, and he wishes to thank the Institute for Advanced Study at Princeton for its hospitality during his visit in Fall 2019.

SUPPLEMENTARY MATERIAL

Supplementary Material available at *Biometrika* online includes all technical proofs of this paper.

REFERENCES

AHN, S. K. & REINSEL, G. C. (1988). Nested reduced-rank autogressive models for multiple time series. *J. Am. Statist. Assoc.* **83**, 849–56.

Basu, S. & Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43**, 1535–67.

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D. & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE* **8**, e60188.

Chang, Y. (2004). Bootstrap unit root tests in panels with cross-sectional dependency. *J. Econometrics* **120**, 263–93.

Davis, R. A., Zang, P. & Zheng, T. (2015). Sparse vector autoregressive modeling. *J. Comp. Graph. Statist.* **25**, 1077–96.

Dowell, J. & Pinson, P. (2016). Very-short-term probabilistic wind power forecasts by sparse vector autoregression. *IEEE Trans. Smart Grid* **7**, 763–70.

Fang, K.-T., Kotz, S. & Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*. New York: Chapman and Hall/CRC.

Faradonbeh, M. K. S., Tewari, A. & Michailidis, G. (2018). Finite time identification in unstable linear systems. *Automatica* **96**, 342–53.

Gorrostieta, C., Ombao, H., Bédard, P. & Sanes, J. N. (2012). Investigating brain connectivity using mixed effects vector autoregressive models. *NeuroImage* **59**, 3347–55.

Guo, S., Wang, Y. & Yao, Q. (2016). High-dimensional and banded vector autoregressions. *Biometrika* **103**, 889–903.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton: Princeton University Press.

Han, F., Lu, H. & Liu, H. (2015a). A direct estimation of high dimensional stationary vector autoregressions. *J. Mach. Learn. Res.* **16**, 3115–50.

Han, F., Xu, S. & Liu, H. (2015b). Rate-optimal estimation of a high-dimensional semiparametric time series model. Preprint, University of Washington. https://sites.stat.washington.edu/people/fanghan/paper-VAR.pdf.

Horn, R. A. & Johnson, C. R. (1985). *Matrix Analysis*. New York: Cambridge University Press.

Kano, Y. (1994). Consistency property of elliptical probability density functions. *J. Mult. Anal.* **51**, 139–47.

Kotz, S. & Nadarajah, S. (2004). *Multivariate t-Distributions and Their Applications*. Cambridge: Cambridge University Press.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.

Maj-Kańska, A., Pokarowski, P. & Prochenka, A. (2015). Delete or merge regressors for linear model selection. *Electron. J. Statist.* **9**, 1749–78.

Mendelson, S. (2014). Learning without concentration. *Proc. Mach. Learn. Res.* **35**, 25–39.

Negahban, S. & Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39**, 1069–97.

Recht, B. (2018). A tour of reinforcement learning: The view from continuous control. *arXiv:*1806.09460v2.

Reinsel, G. C. (1993). *Elements of Multivariate Time Series Analysis*. New York: Springer.

Rudelson, M. & Vershynin, R. (2015). Small ball probabilities for linear images of high-dimensional distributions. *Int. Math. Res. Not.* **2015**, 9594–617.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. & Recht, B. (2018). Learning without mixing: Towards a sharp analysis of linear system identification. *Proc. Mach. Learn. Res.* **75**, 439–73.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica* **48**, 1–48.

Stock, J. H. & Watson, M. W. (2001). Vector autoregressions. *J. Econ. Perspect.* **15**, 101–15.

Tsay, R. S. (2013). *Multivariate Time Series Analysis: With R and Financial Applications*. Chichester: John Wiley & Sons.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge: Cambridge University Press.

Wilms, I., Basu, S., Bien, J. & Matteson, D. S. (2017). Interpretable vector autoregressions with exogenous time series. In *Proc. Symp. Interpretable Machine Learning, 31st Conf. Neural Information Processing Systems (NIPS 2017)*.

Wu, J. C. & Xia, F. D. (2016). Measuring the macroeconomic impact of monetary policy at the zero lower bound. *J. Money, Credit Banking* **48**, 253–91.

Zhang, B., Pan, G. & Gao, J. (2018). CLT for largest eigenvalues and unit root testing for high-dimensional nonstationary time series. *Ann. Statist.* **46**, 2186–215.

Zhou, W.-X., Bose, K., Fan, J. & Liu, H. (2018). A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *Ann. Statist.* **46**, 1904–31.

Zhu, X., Pan, R., Li, G., Liu, Y. & Wang, H. (2017). Network vector autoregression. *Ann. Statist.* **45**, 1096–123.

[*Received on* 4 *April* 2019. *Editorial decision on* 18 *May* 2020]