

*The Annals of Statistics*
2015, Vol. 43, No. 4, 1535–1567
DOI: 10.1214/15-AOS1315


# REGULARIZED ESTIMATION IN SPARSE HIGH-DIMENSIONAL TIME SERIES MODELS

BY SUMANTA BASU AND GEORGE MICHAILIDIS[1]

*University of Michigan*

Many scientific and economic problems involve the analysis of high-dimensional time series datasets. However, theoretical studies in high-dimensional statistics to date rely primarily on the assumption of independent and identically distributed (i.i.d.) samples. In this work, we focus on stable Gaussian processes and investigate the theoretical properties of $\ell_1$-regularized estimates in two important statistical problems in the context of high-dimensional time series: (a) stochastic regression with serially correlated errors and (b) transition matrix estimation in vector autoregressive (VAR) models. We derive nonasymptotic upper bounds on the estimation errors of the regularized estimates and establish that consistent estimation under high-dimensional scaling is possible via $\ell_1$-regularization for a large class of stable processes under sparsity constraints. A key technical contribution of the work is to introduce a measure of stability for stationary processes using their spectral properties that provides insight into the effect of dependence on the accuracy of the regularized estimates. With this proposed stability measure, we establish some useful deviation bounds for dependent data, which can be used to study several important regularized estimates in a time series setting.

**1. Introduction.** Recent advances in information technology have made high-dimensional time series data sets increasingly common in numerous applications. Examples include structural analysis and forecasting with a large number of macroeconomic variables [De Mol, Giannone and Reichlin (2008)], reconstruction of gene regulatory networks from time course microarray data [Michailidis and d'Alché-Buc (2013)], portfolio selection and volatility matrix estimation in finance [Fan, Lv and Qi (2011)] and studying co-activation networks in human brains using task-based or resting state fMRI data [Smith (2012)]. These applications require analyzing a large number of temporally observed variables using small to moderate sample sizes (number of time points), and the techniques used for the respective learning tasks include classical regression, vector autorgressive modeling and covariance estimation. Meaningful inference in such settings

Received February 2014; revised January 2015.

[1]Supported by NSA Grant H98230-10-1-0203 and NSF Grants DMS-11-61838 and DMS-12-28164.

*MSC2010 subject classifications.* Primary 62M10, 62J99; secondary 2M15.

*Key words and phrases.* High-dimensional time series, stochastic regression, vector autoregression, covariance estimation, lasso.

is often impossible without imposing some lower-dimensional structural assumption on the data generating mechanism, the most common being that of sparsity on the model parameter space. In high-dimensional regression and VAR problems, the notion of sparsity is often incorporated into the estimation procedure by $\ell_1$-penalization procedures like lasso and its variants [Bickel, Ritov and Tsybakov (2009), van de Geer, Bühlmann and Zhou (2011)], while for covariance matrix estimation problems, sparsity is enforced via hard thresholding [Bickel and Levina (2008)].

Theoretical properties of such regularized estimates under high-dimensional scaling have been investigated in numerous studies over the last few years, under the key assumption that the samples are independent and identically distributed (i.i.d.). On the other hand, theoretical analysis of these estimates in a time series context, where the data exhibit *temporal* and *cross-sectional dependence*, is rather incomplete. A central challenge is to assess how the underlying dependence structure affects the performance of these regularized estimates.

In this paper, we focus on stationary Gaussian time series and use their *spectral properties* to propose a measure of stability. Using this measure of stability, we establish necessary concentration bounds for dependent data and study, in a nonasymptotic framework, the theoretical properties of regularized estimates in the following key statistical models: (a) $\ell_1$-penalized sparse stochastic regression with exogenous predictors and serially correlated errors and (b) $\ell_1$-penalized least squares and log likelihood based estimation of sparse VAR models. We establish nonasymptotic upper bounds on the estimation error and show that lasso can perform consistent estimation in high-dimensional settings under a mild stability assumption on the underlying processes that is common in the classical literature of low-dimensional time series. Our results also provide new insights into how the convergence rates are affected by the presence of temporal dependence in the data.

Next, we introduce the two models analyzed in this paper and highlight the main contributions of our work to the existing literature. Although the main interest of this work is to study VAR models in high dimensions, a key stepping stone to our analysis comes from stochastic regression models, which are of independent interest.

*Stochastic regression.* We start with this canonical problem in time series analysis [Hamilton (1994)], a linear regression model of the form

$$y^t = \langle \beta^*, X^t \rangle + \varepsilon^t, \qquad t = 1, \ldots, n, \tag{1.1}$$

where the $p$-dimensional predictors $\{X^t\}$ and the errors $\{\varepsilon^t\}$ are generated according to independent, centered, Gaussian stationary processes. Under a sparsity assumption on $\beta^*$, we study the properties of the lasso estimate

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \|Y - \mathcal{X}\beta\|^2 + \lambda_n \|\beta\|_1, \tag{1.2}$$

where $Y = [y^n : \ldots : y^1]'$, $\mathcal{X} = [X^n : \ldots : X^1]'$ and $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$. Theoretical properties of lasso have been studied for fixed design regression $Y = \mathcal{X}\beta^* + E$, with $E = [e^n : \ldots : e^1]'$, by several authors [Bickel, Ritov and Tsybakov (2009), Loh and Wainwright (2012), Negahban et al. (2012)]. They establish consistency of lasso estimates in a high-dimensional regime under some form of restricted eigenvalue (RE) or restricted strong convexity (RSC) assumption on $S = \mathcal{X}'\mathcal{X}/n$ and suitable deviation conditions on $\mathcal{X}'E/n$.

In general, for a given design matrix $\mathcal{X}$, verifying that $\mathcal{X}$ satisfies an RE condition [Dobriban and Fan (2013)] is an NP-hard problem. In the case where the rows of $\mathcal{X}$ are independently generated from a common Gaussian/sub-Gaussian ensemble, these assumptions are known to hold with high probability under mild conditions [Raskutti, Wainwright and Yu (2010), Rudelson and Zhou (2013)]. It is not clear, however, whether similar regularity conditions are satisfied with high probability when the observations are dependent.

Asymptotic properties of lasso for high-dimensional time series have been considered by [Loh and Wainwright (2012), Wu and Wu (2014)], and we provide detailed comparisons with those studies in Section 3. In short, these works either assume RE conditions or establish their validity within a very restricted class of VAR(1) models, as illustrated in Figure 1 and Lemma E.2 in Appendix E (supplementary material [Basu and Michailidis (2015)]).

A major contribution of the present study is to establish the validity of suitable RE and deviation conditions for a large class of stationary Gaussian processes $\{X^t\}$ and $\{\varepsilon^t\}$. As a result, this work extends existing results to a much larger class of time series models and provides deeper insights into the effect of dependence on the estimation error of lasso.

*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 [Bernanke, Boivin and Eliasz (2005), Sims (1980), Stock and Watson (2005)]. Unlike structural models, VAR provides a broad framework for capturing complex temporal and cross-sectional interrelationship among the time series [Bańbura, Giannone and Reichlin (2010)]. In addition to economics, VAR models have been instrumental in linear system identification problems in control theory [Kumar and Varaiya (1986)], while more recently, they have become standard tools in functional genomics for reconstruction of regulatory networks [Michailidis and d'Alché-Buc (2013), Shojaie and Michailidis (2010)] and in neuroscience for understanding effective connectivity patterns between brain regions [Friston (2009), Seth, Chorley and Barnett (2013), Smith (2012)].

Formally, for a $p$-dimensional vector-valued stationary time series $\{X^t\} = \{(X_1^t, \ldots, X_p^t)\}$, a VAR model of lag $d$ [VAR($d$)] with serially uncorrelated Gaussian errors takes the form

$$X^t = A_1 X^{t-1} + \cdots + A_d X^{t-d} + \varepsilon^t, \qquad \varepsilon^t \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \Sigma_\varepsilon), \tag{1.3}$$

where $A_1, \ldots, A_d$ are $p \times p$ matrices and $\varepsilon^t$ is a $p$-dimensional vector of possibly correlated innovation shocks. The main objective in VAR models is to estimate the transition matrices $A_1, \ldots, A_d$, together with the order of the model $d$, based on realizations $\{X^0, X^1, \ldots, X^T\}$. The structure of the transition matrices provides insight into the complex temporal relationships amongst the $p$ time series and leads to efficient forecasting strategies.

VAR estimation is a natural high-dimensional problem, since the dimensionality of the parameter space ($dp^2$) grows quadratically with $p$. For example, estimating a VAR(2) model with $p = 20$ time series requires estimating $dp^2 = 800$ parameters. However, a comparable number of stationary observations is rarely available in practice. In the low-dimensional setting, VAR estimation is carried out by reformulating it as a multivariate regression problem [Lütkepohl (2005)]. Under high-dimensional scaling and sparsity assumptions on the transition matrices, a natural strategy is to resort to $\ell_1$-penalized least squares or log-likelihood based methods [Davis, Zang and Zheng (2012), Song and Bickel (2011)].

Compared to stochastic regression, theoretical analysis of large VAR requires two important considerations. First, since the response variable is multivariate, the choice of the loss function (least squares, negative log-likelihood) plays an important role in estimation and prediction, especially when the multivariate error process has correlated components. Second, correlation of the error process with the process of predictors $\mathrm{Cov}(X^t, \varepsilon^t) \neq 0$ makes the theoretical analysis more involved. Existing work on high-dimensional VAR models requires stringent assumptions on the dependence structure [Song and Bickel (2011)], or on the transition matrix [Negahban and Wainwright (2011)], which are violated by many stable VAR models, as discussed in Section 4. Our results show that consistent estimation is possible with $\ell_1$-penalization for *both* least squares and log-likelihood based choices of loss functions under high-dimensional scaling for *any* stable VAR($d$) models. Interestingly, the latter choice of loss function leads to an $M$-estimation problem that does not fit into the stochastic regression framework. As in the case of stochastic regression, we establish the validity of suitable restricted eigenvalue and deviation conditions using the stability measures introduced in this work.

A central theme of our theoretical results is that the effect of dependence on the behavior of these regularized estimates can be nicely captured by the spectral properties of the underlying multivariate processes. In particular, we show that the estimation error of lasso in the time series models scales at the same rate as for i.i.d. data, modulo a "price" of dependence, which can be interpreted as a measure of "narrowness" of the underlying spectra. This agrees with a fundamental phenomenon in the signal processing literature—a flatter autocorrelation function (slower decay of temporal dependence) corresponds to a narrower spectrum and vice versa. Moreover, for linear ARMA models, our spectral approach has an added advantage of interpretability, since the spectral density of this class allows a closed form expression in terms of the model parameters.

At the core of our theoretical results are some novel deviation bounds for dependent data established in Section 2. These deviation bounds serve two important purposes. First, they help verify routinely used restricted eigenvalue and deviation conditions used in the lasso literature for a large class of time series models and help develop a theory independent of abstract regularity assumptions. Second, these deviation bounds are general enough to seamlessly integrate with the existing theory of other regularization mechanisms and hence extend the available results to time series setting. Examples include sparse covariance estimation via hard thresholding, nonconvex penalties like SCAD and MCP for sparse modeling, group lasso for structured sparsity and nuclear norm minimization for low-rank modeling, as discussed in Section 7. It is worth noting that many of these regularization mechanisms have been applied on time series data with good empirical performance [Bickel and Levina (2008), Fan, Lv and Qi (2011), Song and Bickel (2011)].

*Outline of the paper.* The remainder of the paper is organized as follows. In Section 2, we first demonstrate via simulation how lasso errors scale in low and high-dimensional regimes for time series data which motivates the proposed stability measure, discuss relevant spectral properties of stationary processes, introduce our measures of stability and present the main deviation bounds used in subsequent analyses. In Section 3 we derive nonasymptotic upper bounds on the estimation error of lasso in stochastic regression with serially correlated errors. Section 4 is devoted to the modeling, estimation and theoretical analysis of sparse VAR models. We examine both least squares and likelihood based regularized estimation of VAR models and their consistency properties. In Section 5, we discuss extensions of the current framework to other regularized estimation problems in high-dimensional time series models. Finally, Section 6 illustrates the performance of lasso estimates in stochastic regression and VAR estimation through simulation studies. We delegate many of the technical proofs to the Appendices in the supplement [Basu and Michailidis (2015)].

*Notation.* Throughout this paper, $\mathbb{Z}$, $\mathbb{R}$ and $\mathbb{C}$ denote the sets of integers, real numbers and complex numbers, respectively. We denote the cardinality of a set by $J$ by $|J|$. For a vector $v \in \mathbb{R}^p$, we denote $\ell_q$ norms by $\|v\|_q := (\sum_{j=1}^p |v_j|^q)^{1/q}$, for $q > 0$. We use $\|v\|_0$ to denote $|\operatorname{supp}(v)| = \sum_{i=1}^p \mathbf{1}[v_j \neq 0]$ and $\|v\|_\infty$ to denote $\max_j |v_j|$. Unless mentioned otherwise, we always use $\|\cdot\|$ to denote $\ell_2$-norm of a vector $v$. For a matrix $A$, $\rho(A)$, $\|A\|$ and $\|A\|_F$ will denote its spectral radius $|\Lambda_{\max}(A)|$, operator norm $\sqrt{\Lambda_{\max}(A'A)}$ and Frobenius norm $\sqrt{\operatorname{tr}(A'A)}$, respectively. We will also use $\|A\|_{\max}$, $\|A\|_1$ and $\|A\|_\infty$ to denote the coordinate-wise maximum (in absolute value), maximum absolute row sum and maximum absolute column sum of a matrix, respectively. For any $p \geq 1$, $q \geq 0$, $r > 0$, we denote the unit balls by $\mathbb{B}_q(r) := \{v \in \mathbb{R}^p : \|v\|_q \leq r\}$. For any $J \subset \{1, \ldots, p\}$ and $\kappa > 0$,

we define the cone set $\mathcal{C}(S,\kappa) = \{v \in \mathbb{R}^p : \|v_{S^c}\|_1 \leq \kappa \|v_S\|_1\}$ and the sparse set $\mathcal{K}(s) = \mathbb{B}_0(s) \cap \mathbb{B}_2(1)$, for any $s \geq 1$. For any set $V$, we denote its closure and convex hull by $\text{cl}\{V\}$ and $\text{conv}\{V\}$. For a symmetric or Hermitian matrix $A$, we denote its maximum and minimum eigenvalues by $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$. We use $e_i$ to denote the $i$th unit vector in $\mathbb{R}^p$. Throughout the paper, we write $A \gtrsim B$ if there exists an absolute constant $c$, independent of the model parameters, such that $A \geq cB$. We use $A \asymp B$ to denote $A \gtrsim B$ and $B \gtrsim A$.

## 2. Deviation bounds for multivariate Gaussian time series.

2.1. *Effect of temporal dependence on lasso errors.* Whereas in classical asymptotic analysis of time series, the quantification of temporal dependence and its impact on the limiting behavior of the model parameter estimates are typically achieved by assuming some mixing condition on the underlying stochastic process, this route is hard to follow in a high-dimensional context, even for standard ARMA processes. In recent work, Wu and Wu (2014) and Chen, Xu and Wu (2013) investigate the asymptotic properties of lasso and covariance thresholding in the time series context, assuming a specific rate of decay on the functional dependence measure [Wu (2005)] of the underlying stationary process. For VAR(1) processes $X^t = A_1 X^{t-1} + \varepsilon^t$, the mixing rates and the functional dependence measure are known to scale with the spectral radius $\rho(A)$ [Chen, Xu and Wu (2013), Liebscher (2005)]. The following two simulation experiments show that dependence in the data affect the convergence rates of lasso estimates in a more *intricate* manner, not completely captured by $\rho(A)$. Further, several authors [Han and Liu (2013), Loh and Wainwright (2012), Negahban and Wainwright (2011)] conducted nonasymptotic analysis of high-dimensional VAR(1) models, assuming $\|A\| < 1$. In Appendix E (supplementary material [Basu and Michailidis (2015)]) (see Figure 1 and Lemma E.2), we show that this assumption is restrictive and is violated by many stable VAR(1) models. More importantly, such an assumption does not generalize beyond VAR(1).

EXAMPLE 1. We generate data from the stochastic regression model (1.1) with $p = 200$ predictors and i.i.d. errors $\{\varepsilon^t\}$. The process of predictors comes from a Gaussian VAR(1) model $X^t = AX^{t-1} + \xi^t$, where $A$ is an upper triangular matrix with $\alpha = 0.2$ on the diagonal and $\gamma$ on the two upper off-diagonal bands. We generate processes with different levels of cross-correlation among the predictors by changing $\gamma$ and plot the average estimation error of lasso (over multiple iterates) against different sample sizes $n$ in Figure 1.

The spectral radius is *common* ($\alpha = 0.2$) across all models. Consistently with the classical low-dimensional asymptotics, the lasso errors for different processes seem to converge as $n$ goes to infinity. However, for small to moderate $n$, as is common in high-dimensional regimes, lasso errors are considerably different for

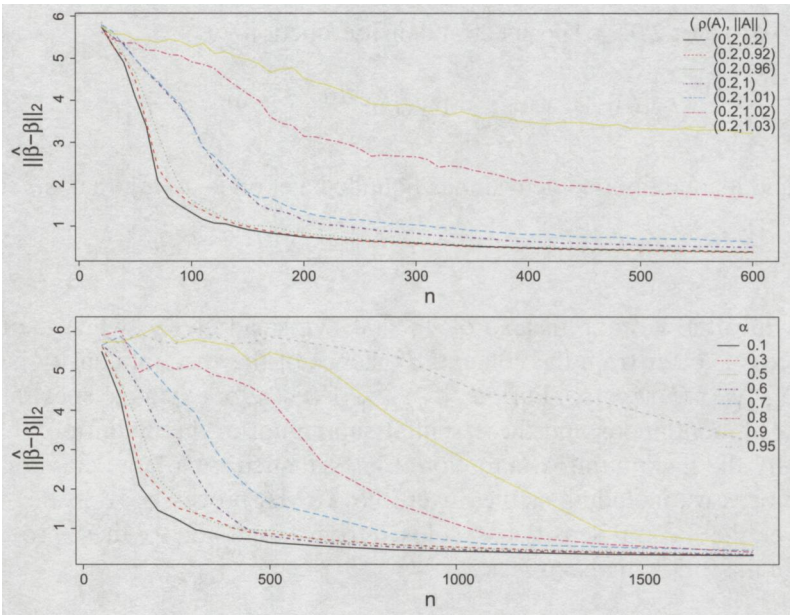


FIG. 1. *Estimation error of lasso in stochastic regression. Top panel*: *Example* 1, VAR(1) *process of predictors with cross-sectional dependence. Bottom panel*: *Example* 2, VAR(2) *process of predictors with no cross-sectional dependence.*

different processes. Capturing the effect of cross-dependence via $\|A\| < 1$ has limitations, as discussed above. We also see that the errors decay even when $\|A\|$ exceeds 1. This motivates a new approach to capture the cross-dependence among the univariate components.

EXAMPLE 2. Even in the absence of cross-dependence, lasso errors exhibit interesting behavior in different regimes, as we show in the next example. Here we generate a similar regression model with $p = 500$ predictors, each generated independently from a Gaussian VAR(2) process $X_j^t = 2\alpha X_{j-1}^t - \alpha^2 X_{j-2}^t + \xi^t$, $0 < \alpha < 1$, $\Gamma_X(0) = 1$. The assumption $\|A\| < 1$ is not applicable here. The processes with different $\alpha$ exhibit different behavior for small to moderate $n$, as predicted by their mixing rates and the functional dependence measures, although it seems the effect of this dependence is significantly reduced when the sample size is large (Figure 1).

These examples motivate us to introduce a different measure to quantify dependence that reconciles the observed behavior of the lasso errors.

2.2. *Measure of stability.* Consider a $p$-dimensional discrete time, centered, covariance-stationary process $\{X^t\}_{t\in\mathbb{Z}}$ with autocovariance function $\Gamma_X(h) = \mathrm{Cov}(X^t, X^{t+h})$, $t, h \in \mathbb{Z}$. We make the following assumption:

ASSUMPTION 2.1. The spectral density function

$$f_X(\theta) := \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \Gamma_X(\ell) e^{-i\ell\theta}, \qquad \theta \in [-\pi, \pi] \tag{2.1}$$

exists, and its maximum eigenvalue is bounded a.e. on $[-\pi, \pi]$, that is,

$$\mathcal{M}(f_X) := \operatorname*{ess\,sup}_{\theta \in [-\pi,\pi]} \Lambda_{\max}\big(f_X(\theta)\big) < \infty. \tag{2.2}$$

We will often write $f$ instead of $f_X$ and $\Gamma$ instead of $\Gamma_X$, when the underlying process is clear from the context. Existence of the spectral density is guaranteed if $\sum_{l=0}^{\infty} \|\Gamma(l)\|^2 < \infty$. Further, if $\sum_{l=0}^{\infty} \|\Gamma(l)\| < \infty$, then the spectral density is bounded, continuous and the essential supremum in the definition of $\mathcal{M}(f_X)$ is actually the maximum. Assumption 2.1 is satisfied by a large class of general linear processes, including stable, invertible ARMA processes [Priestley (1981)]. Moreover, the spectral density has a closed form expression for these processes, as shown in the following examples.

EXAMPLE. An ARMA$(d, \ell)$ process $\{X^t\}$

$$\begin{aligned} X^t = A_1 X^{t-1} + A_2 X^{t-2} + \cdots + A_d X^{t-d} \\ + \varepsilon^t - B_1 \varepsilon^{t-1} - B_2 \varepsilon^{t-2} - \cdots - B_\ell \varepsilon^{t-\ell} \end{aligned} \tag{2.3}$$

is stable and invertible if the matrix valued polynomials $\mathcal{A}(z) := I_p - \sum_{t=1}^{d} A_t z^t$ and $\mathcal{B}(z) := I_p - \sum_{t=1}^{\ell} B_t z^t$ satisfy $\det(\mathcal{A}(z)) \neq 0$ and $\det(\mathcal{B}(z)) \neq 0$ on the unit circle of the complex plane $\{z \in \mathbb{C} : |z| = 1\}$.

For a stable, invertible ARMA process, the spectral density takes the form

$$f_X(\theta) = \frac{1}{2\pi} (\mathcal{A}^{-1}(e^{-i\theta})) \mathcal{B}(e^{-i\theta}) \Sigma_\varepsilon \mathcal{B}^*(e^{-i\theta}) (\mathcal{A}^{-1}(e^{-i\theta}))^*. \tag{2.4}$$

In Appendix E (supplementary material [Basu and Michailidis (2015)]), we provide more details on general linear processes and connection with mixing conditions.

Existence of the spectral density ensures the following representation of the autocovariance matrices

$$\Gamma_X(\ell) = \int_{-\pi}^{\pi} f_X(\theta) e^{i\ell\theta} \, d\theta \qquad \text{for all } \ell \in \mathbb{Z}. \tag{2.5}$$

Since the autocovariance function characterizes a centered Gaussian process, it can be used to quantify the temporal and cross-sectional dependence for this class of models. In particular, spectral density provides insight into the stability of the process, as illustrated and explained in the caption of Figure 2. The upshot is that the peak of the spectral density can be used as a measure of stability of the process.

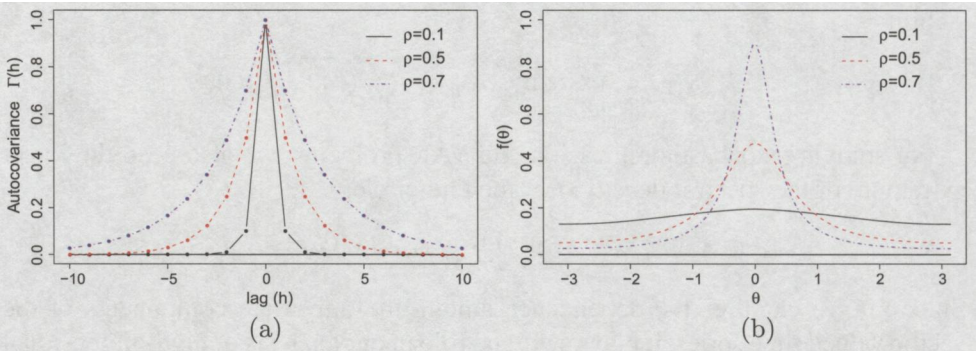


FIG. 2. *Autocovariance* $\Gamma(h)$ *and spectral density* $f(\theta)$ *of a univariate* AR(1) *process* $X^t = \rho X^{t-1} + \varepsilon^t$, $0 < \rho < 1$, $\Gamma_X(0) = 1 = \int_{-\pi}^{\pi} f(\theta)\, d\theta$. *Processes with stronger temporal dependence, that is, with larger* $\rho$, *have flatter* $\Gamma$ *and narrower* $f$. *For* $\rho = 1$, *the process is unstable, and the spectral density does not exist.* (a) *Autocovariance of* AR(1), (b) *spectral density of* AR(1).

More generally, for a $p$-dimensional time series $\{X^t\}$, a natural analogue of the "peak" is the maximum eigenvalue of the (matrix-valued) spectral density function over the unit circle, as defined in (2.2).

In our analysis of high-dimensional time series, we will use $\mathcal{M}(f_X)$ as a *measure of stability* of the process. Processes with larger $\mathcal{M}(f_X)$ will be considered less stable.

For any $k$-dimensional subset $J$ of $\{1, \ldots, p\}$, we can similarly measure the stability of the subprocess $\{X(J)\} = \{(X_j^t) : j \in J\}_{t\in\mathbb{Z}}$ as $\mathcal{M}(f_{X(J)})$. We will measure the stability of all $k$-dimensional subprocesses of $\{X^t\}$ using

$$\mathcal{M}(f_X, k) := \max_{J\subseteq\{1,\ldots,p\},|J|\le k} \mathcal{M}(f_{X(J)}).$$

Clearly, $\mathcal{M}(f_X) = \mathcal{M}(f_X, p)$. For completeness, we define $\mathcal{M}(f_X, k)$ to be $\mathcal{M}(f_X)$, for all $k \ge p$. It follows from the definitions that

$$\mathcal{M}(f_X, 1) \le \mathcal{M}(f_X, 2) \le \cdots \le \mathcal{M}(f_X, p) = \mathcal{M}(f_X).$$

If $\{X^t\}$ and $\{Y^t\}$ are independent $p$-dimensional time series satisfying Assumption 2.1 and $Z^t = X^t + Y^t$, then $f_Z = f_X + f_Y$. Consequently,

$$\mathcal{M}(f_Z) = \mathcal{M}(f_X) + \mathcal{M}(f_Y).$$

More generally, for any two $p$-dimensional processes $\{X^t\}$ and $\{Y^t\}$, the cross-spectral density is defined as

$$f_{X,Y}(\theta) = (1/2\pi) \sum_{l=-\infty}^{\infty} \Gamma_{X,Y}(l) e^{-il\theta}, \qquad \theta \in [-\pi, \pi],$$

where $\Gamma_{X,Y}(h) = \mathrm{Cov}(X^t, Y^{t+h})$, $h \in \mathbb{Z}$. If the joint process $W^t = [(X^t)', (Y^t)']'$ satisfies Assumption 2.1, we can similarly define the cross-spectral measure of

stability

$$\mathcal{M}(f_{X,Y}) = \operatorname*{ess\,sup}_{\theta\in[-\pi,\pi]} \sqrt{\Lambda_{\max}\big(f^*_{X,Y}(\theta) f_{X,Y}(\theta)\big)}.$$

For studying stochastic regression and VAR problems, we also need the lower extremum of the spectral density over the unit circle,

$$\mathrm{m}(f_X) := \operatorname*{ess\,inf}_{\theta\in[-\pi,\pi]} \Lambda_{\min}\big(f_X(\theta)\big).$$

Since $\mathrm{m}(f_X)$ captures the dependence among the univariate components of the vector-valued time series, it plays a crucial role in our analysis of high-dimensional regression in quantifying dependence among the columns of the design matrix.

For stable, invertible ARMA processes and general linear processes with stable transfer functions, the spectral density is bounded and continuous. In these cases, the essential supremum (infimum) in the above definitions of $\mathrm{m}(f_X)$ and $\mathcal{M}(f_X)$ reduce to maximum (minimum) because of the continuity of eigenvalues and the compactness of the unit circle $\{z \in \mathbb{C} : |z| = 1\}$.

Note that $\mathrm{m}(f_X)$ and $\mathcal{M}(f_X)$ may not have closed form expressions for general stationary processes. However, for a stationary ARMA process (2.3), we have the following bounds:

$$\begin{aligned}
\mathrm{m}(f_X) &\geq \frac{1}{2\pi} \frac{\Lambda_{\min}(\Sigma_\varepsilon)\mu_{\min}(\mathcal{B})}{\mu_{\max}(\mathcal{A})}, \\
\mathcal{M}(f_X) &\leq \frac{1}{2\pi} \frac{\Lambda_{\max}(\Sigma_\varepsilon)\mu_{\max}(\mathcal{B})}{\mu_{\min}(\mathcal{A})} \\
\mu_{\min}(\mathcal{A}) &:= \min_{|z|=1} \Lambda_{\min}\big(\mathcal{A}^*(z)\mathcal{A}(z)\big), \\
\mu_{\max}(\mathcal{A}) &:= \max_{|z|=1} \Lambda_{\max}\big(\mathcal{A}^*(z)\mathcal{A}(z)\big),
\end{aligned} \tag{2.6}$$

and $\mu_{\min}(\mathcal{B})$, $\mu_{\max}(\mathcal{B})$ are defined accordingly.

It is often easier to work with $\mu_{\min}(\mathcal{A})$ and $\mu_{\max}(\mathcal{A})$ instead of $\mathrm{m}(f_X)$ and $\mathcal{M}(f_X)$. In particular, we have the following bounds:

PROPOSITION 2.2. *Consider a polynomial* $\mathcal{A}(z) = I_p - \sum_{t=1}^d A_t z^t$, $z \in \mathbb{C}$, *satisfying* $\det(\mathcal{A}(z)) \neq 0$ *for all* $|z| \leq 1$:

(i) *For any* $d \geq 1$, $\mu_{\max}(\mathcal{A}) \leq [1 + (v_{\text{in}} + v_{\text{out}})/2]^2$, *where*

$$v_{\text{in}} = \sum_{h=1}^{d} \max_{1\leq i\leq p} \sum_{j=1}^{p} |A_h(i,j)|, \qquad v_{\text{out}} = \sum_{h=1}^{d} \max_{1\leq j\leq p} \sum_{i=1}^{p} |A_h(i,j)|.$$

(ii) *If* $d = 1$, *and* $A_1$ *is diagonalizable, then*

$$\mu_{\min}(\mathcal{A}) \geq \big(1 - \rho(A_1)\big)^2 \|P\|^{-2} \|P^{-1}\|^{-2},$$

*where $\rho(A_1)$ is the spectral radius (maximum absolute eigenvalue) of $A_1$, and the columns of $P$ are eigenvectors of $A_1$.*

Proposition 2.2, together with (2.6), demonstrate how $\mathfrak{m}(f_X)$ and $\mathcal{M}(f_X)$ behave for ARMA models. For instance, for a VAR(1) process, these quantities are bounded away from zero and infinity as long as the noise covariance structure and the matrix of eigenvectors of $A_1$ are well conditioned, the spectral radius of $A_1$ is bounded away from 1 and the entries of $A_1$ do not concentrate on a single row or column. The proof is delegated to Appendix E (supplementary material [Basu and Michailidis (2015)]).

2.3. *Deviation bounds.* Based on realizations of $\{X^t\}_{t=1}^n$ generated according to a stationary process satisfying Assumption 2.1, we construct the data matrix $\mathcal{X} = [X^n : \ldots : X^1]'$ and the sample Gram matrix $S = \mathcal{X}'\mathcal{X}/n$. Deriving suitable concentration bounds on $S$ is a key step for studying regression and VAR estimation problems in high dimension. In the time series context, this is particularly challenging, since both the rows and columns of the data matrix $\mathcal{X}$ are dependent on each other. When the underlying process is Gaussian, this dependence can be expressed using the covariance matrix of the random vector $\mathrm{vec}(\mathcal{X}')$. We denote this covariance matrix by $\Upsilon_n^X := \mathrm{Cov}(\mathrm{vec}(\mathcal{X}'), \mathrm{vec}(\mathcal{X}'))_{np\times np}$.

The next proposition provides bounds on the extreme eigenvalues of $\Upsilon_n^X$ and generalizes analogous results in univariate analysis presented in Xiao and Wu (2012) and Grenander and Szegö (1958). A similar result for block Toeplitz forms under slightly different conditions can be found in Parter (1961). Note that these bounds depend only on the spectral density $f_X$ and are independent of the sample size $n$.

PROPOSITION 2.3. *For any $n \geq 1$, $p \geq 1$,*

$$2\pi\,\mathfrak{m}(f_X) \leq \Lambda_{\min}\big(\Upsilon_n^X\big) \leq \Lambda_{\max}\big(\Upsilon_n^X\big) \leq 2\pi\,\mathcal{M}(f_X).$$

*In particular, for $n = 1$,*

$$2\pi\,\mathfrak{m}(f_X) \leq \Lambda_{\min}\big(\Gamma_X(0)\big) \leq \Lambda_{\max}\big(\Gamma_X(0)\big) \leq 2\pi\,\mathcal{M}(f_X).$$

Next, we establish some deviation bounds on $S = \mathcal{X}'\mathcal{X}/n$ and $\mathcal{X}'E/n$. These bounds serve as starting points for analyzing regression and covariance estimation problems. In part (a), the first deviation bound shows how $\|\mathcal{X}v\|^2/n\|v\|^2$ concentrates around its expectation, where $v \in \mathbb{R}^p$ is a fixed vector. This will be used to verify restricted eigenvalue assumptions for stochastic regression and VAR estimation problems. The second deviation bound is about the concentration of the entries of $S$ around their expectations. This will be useful for estimating sparse covariance matrices. In part (b), we establish deviation bounds on how $\mathcal{X}'\mathcal{Y}/n$

concentrates around zero ($\mathcal{Y}$ is the data matrix from another process $\{Y^t\}$). In regression and VAR problems, applying this bound with $\{Y^t\}$ as the error process enables the derivation of necessary deviation bounds on $\mathcal{X}'E/n$ under different norms.

PROPOSITION 2.4. (a) *For a stationary, centered Gaussian time series* $\{X^t\}_{t\in\mathbb{Z}}$ *satisfying Assumption* 2.1, *there exists a constant* $c > 0$ *such that for any k-sparse vectors* $u, v \in \mathbb{R}^p$ *with* $\|u\| \le 1$, $\|v\| \le 1$, $k \ge 1$, *and any* $\eta \ge 0$,

$$\mathbb{P}\big[|v'(S - \Gamma_X(0))v| > 2\pi\mathcal{M}(f_X, k)\eta\big] \le 2\exp\big[-cn\min\{\eta^2, \eta\}\big], \tag{2.7}$$

$$\mathbb{P}\big[|u'(S - \Gamma_X(0))v| > 6\pi\mathcal{M}(f_X, 2k)\eta\big] \le 6\exp\big[-cn\min\{\eta^2, \eta\}\big]. \tag{2.8}$$

*In particular, for any* $i, j \in \{1, \ldots, p\}$, *we have*

$$\mathbb{P}\big[|S_{ij} - \Gamma_{ij}(0)| > 6\pi\mathcal{M}(f_X, 2)\eta\big] \le 6\exp\big[-cn\min\{\eta^2, \eta\}\big]. \tag{2.9}$$

(b) *Consider two p-dimensional, centered, stationary Gaussian processes* $\{X^t\}_{t\in\mathbb{Z}}$ *and* $\{Y^t\}_{t\in\mathbb{Z}}$ *with* $\mathrm{Cov}(X^t, Y^t) = 0$ *for every* $t \in \mathbb{Z}$ *and the joint process* $[(X^t)', (Y^t)']'$ *satisfying Assumption* 2.1. *Let* $\mathcal{X} = [X^n : \ldots : X^1]'$ *and* $\mathcal{Y} = [Y^n : \ldots : Y^1]'$ *be the data matrices. Then there exists a constant* $c > 0$ *such that for any* $u, v \in \mathbb{R}^p$ *with* $\|u\| \le 1$, $\|v\| \le 1$, *we have*

$$\begin{aligned}&\mathbb{P}\big[|u'(\mathcal{X}'\mathcal{Y}/n)v| > 2\pi\big(\mathcal{M}(f_X) + \mathcal{M}(f_Y) + \mathcal{M}(f_{X,Y})\big)\eta\big]\\ &\qquad \le 6\exp\big[-cn\min\{\eta, \eta^2\}\big].\end{aligned} \tag{2.10}$$

*In particular, for any stable* VAR($d$) *model* (1.3) *with* $\mathcal{X} = [X^n : \ldots : X^1]'$ *and* $E = [\varepsilon^{n+h} : \ldots : \varepsilon^{1+h}]'$, $h > 0$, *we have*

$$\begin{aligned}&\mathbb{P}\left[|u'(\mathcal{X}'E/n)v| > 2\pi\left(\Lambda_{\max}(\Sigma_\varepsilon)\left(1 + \frac{1 + \mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\right)\right)\eta\right]\\ &\qquad \le 6\exp\big[-cn\min\{\eta, \eta^2\}\big].\end{aligned} \tag{2.11}$$

Next, we give the proofs of the these two key propositions that employ techniques in spectral theory of multivariate time series and nonasymptotic random matrix theory results.

PROOF OF PROPOSITION 2.3. For $1 \le r, s \le n$, the $(r, s)$th block of the $np \times np$ matrix $\Upsilon_n^X$ is a $p \times p$ matrix

$$\Gamma_X(r - s) = \mathrm{Cov}\big(X^{n-r+1}, X^{n-s+1}\big).$$

For any $x \in \mathbb{R}^{np}$, $\|x\| = 1$, write $x$ as $x = \{(x^1)', (x^2)', \ldots, (x^p)'\}'$, where each $x^i \in \mathbb{R}^p$. Define $G(\theta) = \sum_{r=1}^n x^r e^{-ir\theta}$, for $\theta \in [-\pi, \pi]$. Note that

$$\begin{aligned}\int_{-\pi}^{\pi} G^*(\theta)G(\theta)\,d\theta &= \sum_{r=1}^n\sum_{s=1}^n \int_{-\pi}^{\pi} (x^r)'(x^s)e^{i(r-s)\theta}\,d\theta\\ &= \sum_{r=1}^n \|x^r\|^2 2\pi = 2\pi.\end{aligned} \tag{2.12}$$

Also,

$$\begin{aligned} x'\Upsilon_n^X x &= \sum_{r=1}^{n}\sum_{s=1}^{n}(x^r)'\Gamma_X(r-s)(x^s) \\ &= \sum_{r=1}^{n}\sum_{s=1}^{n}\int_{-\pi}^{\pi}(x^r)' f_X(\theta)e^{i(r-s)\theta}(x^s)\,d\theta \qquad \text{using (2.5)} \\ &= \int_{-\pi}^{\pi} G^*(\theta) f_X(\theta) G(\theta)\,d\theta. \end{aligned}$$

Since $f_X(\theta)$ is Hermitian, $G^*(\theta) f_X(\theta) G(\theta)$ is real, for all $\theta \in [-\pi, \pi]$, and

$$\mathrm{m}(f_X) G^*(\theta)G(\theta) \le G^*(\theta) f_X(\theta) G(\theta) \le \mathcal{M}(f_X) G^*(\theta)G(\theta).$$

This, together with (2.12), implies

$$2\pi\,\mathrm{m}(f_X) \le x'\Upsilon_n^X x \le 2\pi\,\mathcal{M}(f_X)$$

for all $x \in \mathbb{R}^{np}$, $\|x\| = 1$. □

PROOF OF PROPOSITION 2.4. (a) First, note that it is enough to prove (2.7) for $\|v\| = 1$. For any $v \in \mathbb{R}^p$, $\|v\| = 1$, let $J$ denote its support $\mathrm{supp}(v)$ so that $|J| = k$. define $Y = \mathcal{X}v = \mathcal{X}_J v_J$. Then $Y \sim N(0_{n\times 1}, Q_{n\times n})$ with

$$Q_{rs} = v_J' \operatorname{Cov}(X_J^{n-r+1}, X_J^{n-s+1}) v_J = v_J' \Gamma_{X(J)}(r-s) v_J \qquad \text{for all } 1 \le r, s \le n.$$

Note that $v'Sv = (1/n)Y'Y = (1/n)Z'QZ$ where $Z \sim N(0, I_n)$. Also, $v'\Gamma_X(0)v = v_J'\Gamma_{X(J)}(0)v_J = \mathbb{E}[Z'QZ/n]$.

So, by the Hanson–Wright inequality of Rudelson and Vershynin (2013), with $\|Z_i\|_{\psi_2} \le 1$ since $Z_i \sim N(0,1)$, we get

$$\begin{aligned} \mathbb{P}\big[|v'(S - \Gamma_X(0))v| > \zeta\big] &= \mathbb{P}\big[|Z'QZ - \mathbb{E}[Z'QZ]| > n\zeta\big] \\ &\le 2\exp\left[-cn\min\left\{\frac{n^2\zeta^2}{\|Q\|_F^2}, \frac{n\zeta}{\|Q\|}\right\}\right]. \end{aligned} \tag{2.13}$$

Since $\|Q\|_F^2/n \le \|Q\|^2$, setting $\zeta = \|Q\|\eta$, we obtain

$$\mathbb{P}\big[|v'(S - \Gamma_X(0))v| > \eta\|Q\|\big] \le 2\exp\big[-cn\min\{\eta, \eta^2\}\big].$$

Also, for any $w \in \mathbb{R}^n$, $\|w\| = 1$, we have

$$\begin{aligned} w'Qw &= \sum_{r=1}^{n}\sum_{s=1}^{n} w_r w_s Q_{rs} = \sum_{r=1}^{n}\sum_{s=1}^{n} w_r w_s v_J' \Gamma_{X(J)}(r-s) v_J \\ &= (w\otimes v)'\Upsilon_n^{X(J)}(w\otimes v) \\ &\le \Lambda_{\max}(\Upsilon_n^{X(J)}) \qquad \text{since } \|w\otimes v\| = 1 \\ &\le 2\pi\,\mathcal{M}(f_{X(J)}) \qquad \text{by Proposition 2.3} \\ &\le 2\pi\,\mathcal{M}(f_X, k). \end{aligned}$$

This establishes an upper bound on the operator norm $\|Q\| \leq 2\pi \mathcal{M}(f_X, k)$.

To prove (2.8), note that

$$2|u'(S - \Gamma_X(0))v| \leq |u'(S - \Gamma_X(0))u| + |v'(S - \Gamma_X(0))v| + |(u+v)'(S - \Gamma_X(0))(u+v)|$$

and $u + v$ is $2k$-sparse with $\|u + v\| \leq 2$. The result follows by applying (2.7) separately on each of the three terms on the right.

The element-wise deviation bound (2.9) is obtained by choosing $u = e_i$, $v = e_j$.

(b) Note that $u'(\mathcal{X}'\mathcal{Y}/n)v$ can be viewed as $(1/n)\sum_{t=1}^n w^t z^t$, where $w^t = \langle u, X^t \rangle$, $z^t = \langle v, Y^t \rangle$ are two univariate stationary processes with spectral densities $f_w(\theta) = u' f_X(\theta) u$ and $f_z(\theta) = v' f_Y(\theta) v$. Since $\mathrm{Cov}(w^t, z^t) = 0$, we have the following decomposition:

$$\frac{2}{n}\sum_{t=1}^n w^t z^t = \left[\frac{1}{n}\sum_{t=1}^n (w^t + z^t)^2 - \mathrm{Var}(w^1 + z^1)\right] - \left[\frac{1}{n}\sum_{t=1}^n (w^t)^2 - \mathrm{Var}(w^1)\right] - \left[\frac{1}{n}\sum_{t=1}^n (z^t)^2 - \mathrm{Var}(z^1)\right],$$

and it suffices to concentrate the three terms separately. Applying (2.7) on the process $w^t = \langle u, X^t \rangle$ and noting that $\mathcal{M}(f_w) \leq \mathcal{M}(f_X)$, we have

$$\mathbb{P}\left[\left|(1/n)\sum_{t=1}^n (w^t)^2 - \mathrm{Var}(w^1)\right| > 2\pi \mathcal{M}(f_X)\eta\right] > 2\exp[-cn\min\{\eta, \eta^2\}].$$

A similar argument for $\{z^t\}$ leads to

$$\mathbb{P}\left[\left|(1/n)\sum_{t=1}^n (z^t)^2 - \mathrm{Var}(z^1)\right| > 2\pi \mathcal{M}(f_Y)\eta\right] > 2\exp[-cn\min\{\eta, \eta^2\}].$$

To concentrate the first term, note that the process $\{w^t + z^t\}$ has a spectral density given by

$$f_{w+z}(\theta) = \begin{bmatrix} u' & v' \end{bmatrix} \begin{bmatrix} f_X(\theta) & f_{X,Y}(\theta) \\ f^*_{X,Y}(\theta) & f_Y(\theta) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = u' f_X(\theta) u + v' f_Y(\theta) v + u' f_{X,Y}(\theta) v + v' f^*_{X,Y}(\theta) u.$$

Since $\|u\| \leq 1$, $\|v\| \leq 1$, $\mathcal{M}(f_{w+z}) \leq \mathcal{M}(f_X) + \mathcal{M}(f_Y) + 2\mathcal{M}(f_{X,Y})$, where the last term is obtained by applying the Cauchy–Schwarz inequality on each of the cross-product terms. Applying (2.7) separately on $\{w^t\}$, $\{z^t\}$ and $\{w^t + z^t\}$ with the above bounds on the respective stability measures leads to the final result.

In the special case of a VAR($d$) process, set $\tilde{\varepsilon}^t := \varepsilon^{t+h}$ so that $\mathrm{Cov}(X^t, \tilde{\varepsilon}^t) = 0$. Then it suffices to establish upper bounds on $\mathcal{M}(f_X)$, $\mathcal{M}(f_{\tilde{\varepsilon}})$ and $\mathcal{M}(f_{X,\tilde{\varepsilon}})$. From (2.6), $2\pi\mathcal{M}(f_X)$ is upper bounded by $\Lambda_{\max}(\Sigma_\varepsilon)/\mu_{\min}(\mathcal{A})$. The process $\{\tilde{\varepsilon}^t\}$

is serially uncorrelated, so $\mathcal{M}(f_{\tilde{\varepsilon}})$ is the same as $\Lambda_{\max}(\Sigma_\varepsilon)$. To derive an upper bound on the cross-spectral measure of stability, note that

$$\begin{aligned}\operatorname{Cov}(X^t, \varepsilon^{t+h+l}) &= \operatorname{Cov}(X^t, X^{t+h+l} - A_1 X^{t+h+l-1} - \cdots - A_d X^{t+h+l-d}) \\ &= \Gamma_X(h+l) - \Gamma_X(h+l-1)A_1' - \cdots - \Gamma_X(h+l-d)A_d'.\end{aligned}$$

Hence, the cross-spectrum of $\{X^t\}$ and $\{\tilde{\varepsilon}^t\}$ can be expressed as

$$\begin{aligned}&f_{X,\tilde{\varepsilon}}(\theta) \\ &\quad= \frac{1}{2\pi}\sum_{l=-\infty}^{\infty}[\Gamma_X(h+l) - \Gamma_X(h+l-1)A_1' - \cdots - \Gamma_X(h+l-d)A_d']e^{-il\theta} \\ &\quad= f_X(\theta)e^{ih\theta}[I - A_1'e^{-i\theta} - \cdots - A_d'e^{-id\theta}] \\ &\quad= e^{ih\theta} f_X(\theta)\mathcal{A}^*(e^{i\theta}).\end{aligned}$$

Hence $\mathcal{M}(f_{X,\tilde{\varepsilon}})$ is bounded above by $\mathcal{M}(f_X)\mu_{\max}(\mathcal{A})$. Combining the three upper bounds on the stability measures and replacing $\mathcal{M}(f_X)$ with its upper bound in (2.6), establishes the final result. □

*Role of the two tails in* (2.13) *and sharpness of the bounds.* The convergence rates of lasso and other regularized estimates in high-dimensional settings depend on how $S$ concentrates around $\Gamma_X(0)$ and $\mathcal{X}'E/n$ around 0, as is evident in subsequent proofs. In the bounds established above, the effect of dependence is captured by $\mathcal{M}(f_X)$. In the special case of no temporal and cross-sectional dependence, our results recover the bounds of lasso for i.i.d. data, as we remark in Section 3. For processes with strong dependence, however, we believe this bound can be further sharpened, although a closed form solution of the exact rate was not established. Next, we provide an asymptotic argument for a fixed $p$ case and demonstrate that in a low-dimensional setting with very large sample sizes, the effect of dependence can be captured by the integrated spectrum, which provides a tighter bound.

The sub-Gaussian and sub-exponential tails in the main concentration inequality (2.13) suggest an interesting phenomenon, that temporal dependence in the data may affect the concentration property and in turn the convergence rates of the regularized estimates in two different ways, depending on which term in the tail bound is dominant.

In the special case of no temporal dependence, that is, $X^t \overset{\text{i.i.d.}}{\sim} N(0, \Sigma)$, the matrix $Q$ is diagonal and $\|Q\|_F/\sqrt{n} = \|Q\|$. So, setting $\zeta = \eta\|Q\|_F/\sqrt{n}$ or $\zeta = \eta\|Q\|$ leads to the same bound, and we recover the Bernstein-type tail bounds for subexponential random variables [Vershynin (2010)].

In the presence of temporal dependence, the two norms $\|Q\|_F$ and $\|Q\|$ behave differently, and this affects the rates. To illustrate this further, we need additional notation. First note that $\mathcal{M}(f_X)$ can be viewed as $\sup_{\|v\|=1}\|f_y\|_\infty$ where $y^t = \langle v, X^t\rangle$ and $\|\cdot\|_\infty$ denotes the $L_\infty$ or sup norm of a function. A related quantity

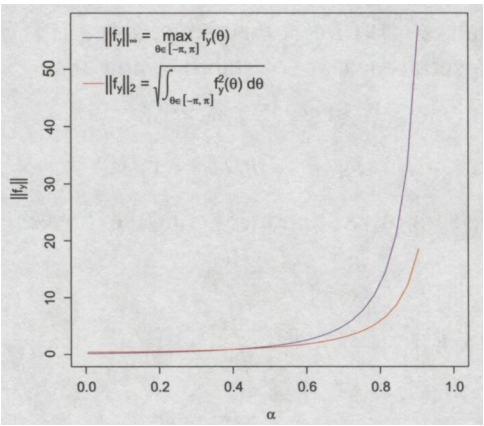


FIG. 3. $\|f_y\|_2$ *and* $\|f_y\|_\infty$ *for a univariate Gaussian* AR(2) *process* $y^t = 2\alpha y^{t-1} - \alpha^2 y^{t-2} + \xi^t$, $\Gamma_y(0) = 1, 0 < \alpha < 1$.

that will be useful for studying the tails is the Euclidean or $L_2$ norm $\|f_y\|_2 = (\int_{-\pi}^{\pi} f_y^2(\theta)\, d\theta)^{1/2}$. For any univariate Gaussian process $\{y^t\}$, it is easy to see that $\|f_y\|_2 \leq \sqrt{2\pi}\|f_y\|_\infty$, and they coincide when the process is serially uncorrelated, that is, the spectrum is flat a.e. With stronger temporal dependence, the spectrum becomes more spiky and $\|f_y\|_\infty$ changes more sharply than $\|f_y\|_2$. In Figure 3, we demonstrate this on a family of AR(2) processes $y^t = 2\alpha y^{t-1} - \alpha^2 y^{t-2} + \xi^t$, $\Gamma_y(0) = 1, 0 < \alpha < 1$.

Coming back to the behavior of the two tails, note that

$$\mathbb{P}[|v'(S - \Gamma(0))v| > \zeta] \leq 2\exp\left[-c\min\left\{\frac{n\zeta^2}{\|Q\|_F^2/n}, \frac{n\zeta}{\|Q\|}\right\}\right].$$

We consider a low-dimensional, fixed $p$ regime. It is known that [cf. Chapter 5, Grenander and Szegö (1958)] for large $n$, $\|Q\|_F^2/n$ approaches $2\pi\|f_y\|_2^2$ and $\|Q\|$ approaches $2\pi\|f_y\|_\infty$. With a choice of $\zeta \asymp \sqrt{\log p/n}$, the tail probability on the right-hand side can be approximated by

$$2\exp\left[-c\min\left\{\frac{\log p}{c_1\|f_y\|_2^2}, \frac{\sqrt{n\log p}}{\|f_y\|_\infty}\right\}\right].$$

This indicates that for very large $n$, the first term will be smaller, and the tail probability will scale with $\|f_y\|_2$. So processes with various levels of dependence should behave similarly in terms of estimation errors. For strongly dependent processes, where $\|f_y\|_2 \ll \|f_y\|_\infty$, it would take more samples $n$ for the first term to offset the second term. With a smaller sample size, the tail behavior will be driven by $\|f_y\|_\infty$, and the effect of dependence will be more prominent in the estimation error of the regularized estimates. Interestingly, this is the same pattern reflected in Figure 1.

**3. Stochastic regression.** In the presence of serially correlated errors, and under a sparsity assumption on $\beta^*$, we use the deviation bounds of Section 2 to derive an upper bound on the estimation error of lasso. Our results show that consistent estimation of $\beta^*$ is possible, as long as the predictor and noise processes are stable. We consider the lasso estimate (1.2) for the stochastic regression model (1.1). Further, we assume that both $f_X$ and $f_\varepsilon$ satisfy Assumption 2.1, and $\beta^*$ is $k$-sparse, with support $J$, that is, $|J| = k$.

Note that in the low-dimensional regime, consistent estimation relies on the following assumptions:

(a) $\mathcal{X}'\mathcal{X}/n$ converges to a nonsingular matrix ($\lim_{n\to\infty} \Lambda_{\min}(\frac{\mathcal{X}'\mathcal{X}}{N}) > 0$).
(b) $\mathcal{X}'E/n$ converges to zero.

In the high-dimensional regime ($n \ll p$), the first assumption is never true since the design matrix is rank-deficient (i.e., more variables than observations). The second assumption is also very stringent, since the dimension of $\mathcal{X}'E$ grows with $n$ and $p$. Interestingly, consistent estimation in the high-dimensional regime can be ensured under two analogous sufficient conditions. The first one comes from a class of conditions commonly referred to as *restricted eigenvalue* (RE) conditions [Bickel, Ritov and Tsybakov (2009), van de Geer and Bühlmann (2009)]. Roughly speaking, these assumptions require that $\|\mathcal{X}(\hat\beta - \beta^*)\|$ is small only when $\|\hat\beta - \beta^*\|$ is small. For sparse $\beta^*$ and $\lambda_n$ appropriately chosen, it is now well understood that the vectors $v = \hat\beta - \beta^*$ only vary on a small subset of the high-dimensional space $\mathbb{R}^p$ [Negahban et al. (2012)]. As shown in the proof of Proposition 3.3, the error vectors $v$ in stochastic regression lie in a cone

$$\mathcal{C}(J,3) = \{v \in \mathbb{R}^p : \|v_{J^c}\|_1 \le 3\|v_J\|_1\},$$

whenever $\lambda_n \ge 4\|\mathcal{X}'E/n\|_\infty$. This indicates that the RE condition may not be very stringent after all, even though $\mathcal{X}$ is singular. Note that verifying that the assumption indeed holds with high probability is a nontrivial task.

The next proposition shows that a restricted eigenvalue (RE) condition holds with high probability when the sample size is sufficiently large and the process of predictors $\{X^t\}$ is stable, with a full-rank spectral density.

PROPOSITION 3.1 (Restricted eigenvalue). *If* $\mathrm{m}(f_X) > 0$, *then there exist constants* $c_i > 0$ *such that for* $n \gtrsim \max\{1, \omega^2\}\min\{k\log(c_0 p/k), k\log p\}$,

$$\mathbb{P}\left[\inf_{v\in\mathcal{C}(J,3)\setminus\{0\}} \frac{\|\mathcal{X}v\|^2}{n\|v\|^2} \ge \alpha_{\mathrm{RE}}\right] \ge 1 - c_1\exp[-c_2 n\min\{1, \omega^{-2}\}],$$

*where* $\alpha_{\mathrm{RE}} = \pi\,\mathrm{m}(f_X)$, $\omega = c_3\mathcal{M}(f_X, 2k)/\mathrm{m}(f_X)$.

REMARKS. (a) The assumption $\mathrm{m}(f_X) > 0$ is fairly mild and holds for stable, invertible ARMA processes. However, the conclusion holds under weaker assumptions like $\Lambda_{\min}(\Gamma_X(0)) > 0$ or an RE condition on $\Gamma_X(0)$, replacing $2\pi\,\mathrm{m}(f_X)$ by

the minimum (or restricted) eigenvalue of $\Gamma_X(0)$, as evident in the proof of this proposition.

(b) For large $k$, $k\log(c_0 p/k)$ can be much smaller than $k\log p$, the sample size required for consistent estimation with lasso.

(c) The factor $\omega \asymp \mathcal{M}(f_X, 2k)/\mathfrak{m}(f_X)$ captures the effect of temporal and cross-sectional dependence in the data. Larger values of $\mathcal{M}(\cdot)$ and smaller values of $\mathfrak{m}(\cdot)$ indicate stronger dependence in the data, and the bounds indicate that more samples are required to ensure RE holds with high probability. We demonstrate this on three special types of dependence in the design matrix $\mathcal{X}$, independent entries, independent rows and independent columns:

(i) If the entries of $\mathcal{X}$ are independent from a $N(0, \sigma^2)$ distribution, we have $\Gamma_X(0) = \sigma^2 I$ and $\Gamma_X(h) = \mathbf{0}$ for $h \neq 0$. In this case, $f_X(\theta) \equiv (1/2\pi)\sigma^2 I$ and $\mathcal{M}(f_X, 2k)/\mathfrak{m}(f_X) = 1$.

(ii) If the rows of $\mathcal{X}$ are independent and identically distributed as $N(0, \Sigma_X)$, that is, $\Gamma_X(0) = \Sigma_X$, $\Gamma_X(h) = \mathbf{0}$ for $h \neq 0$, the spectral density takes the form $f_X(\theta) \equiv (1/2\pi)\Sigma_X$, and $\mathcal{M}(f_X, 2k)/\mathfrak{m}(f_X)$ can be at most $\Lambda_{\max}(\Sigma_X)/\Lambda_{\min}(\Sigma_X)$.

(iii) If the columns of $\mathcal{X}$ are independent, that is, all the univariate components of $\{X^t\}$ are independently generated according to a common stationary process with spectral density $f$, then the spectral density of $\{X^t\}$ is $f_X(\theta) = f(\theta)I$, and we have

$$\mathcal{M}(f_X, 2k)/\mathfrak{m}(f_X) = \max_{\theta\in[-\pi,\pi]} f(\theta) / \min_{\theta\in[-\pi,\pi]} f(\theta).$$

The ratio on the right can be viewed as a measure of narrowness of $f$. Since narrower spectral densities correspond to processes with flatter autocovariance, this indicates that more samples are needed when the dependence is stronger.

The second sufficient condition for consistency of lasso requires that the coordinates of $\mathcal{X}'E/n$ uniformly concentrate around 0. In the next proposition, we establish a deviation bound on $\|\mathcal{X}'E/n\|_\infty$ that holds with high probability. Similar results were established in Loh and Wainwright (2012) for a VAR(1) process with serially uncorrelated errors, under the assumption $\|A_1\| < 1$. Our result relies on different techniques, holds for a much larger class of stationary processes and allows for serial correlation in the noise term, as well.

PROPOSITION 3.2 (Deviation condition). *For $n \gtrsim \log p$, there exist constants $c_i > 0$ such that*

$$\mathbb{P}\left[\frac{1}{n}\|\mathcal{X}'E\|_\infty > c_0 2\pi\left[\mathcal{M}(f_X, 1) + \mathcal{M}(f_\varepsilon)\right]\sqrt{\frac{\log p}{n}}\right] \leq c_1 \exp[-c_2 \log p].$$

REMARK. The deviation inequality shows that the coordinates of $\mathcal{X}'E/n$ uniformly concentrate around 0, as long as the stability measures of $\{\varepsilon^t\}$ and the univariate components of $\{X^t\}$ grow at a rate slower than $\sqrt{n/\log p}$. These two propositions allow us to establish error rates for estimation and prediction in stochastic regression.

PROPOSITION 3.3 (Estimation and prediction error). *Consider the stochastic regression setup of* (1.1). *If* $\beta^*$ *is* $k$*-sparse,* $n \gtrsim [\mathcal{M}(f_X, k)/\mathfrak{m}(f_X)]^2 k \log p$, *then there exist constants* $c_i > 0$ *such that for*

$$\lambda_n \geq c_0 2\pi[\mathcal{M}(f_X, 1) + \mathcal{M}(f_\varepsilon)]\sqrt{(\log p)/n},$$

*any solution* $\hat{\beta}$ *of* (1.2) *satisfies, with probability at least* $1 - c_1 \exp[-c_2 \log p]$,

$$\|\hat{\beta} - \beta^*\| \leq \frac{2\lambda_n\sqrt{k}}{\alpha_{\mathrm{RE}}},$$

$$\|\hat{\beta} - \beta^*\|_1 \leq \frac{8\lambda_n k}{\alpha_{\mathrm{RE}}},$$

$$\frac{1}{n}\|\mathcal{X}(\hat{\beta} - \beta^*)\|^2 \leq \frac{4\lambda_n^2 k}{\alpha_{\mathrm{RE}}},$$

*where the restricted eigenvalue* $\alpha_{\mathrm{RE}} = \pi\mathfrak{m}(f_X)$.

*Further, a thresholded variant of lasso* $\tilde{\beta}$, *defined as* $\tilde{\beta}_j = \{\hat{\beta}_j \mathbf{1}_{|\hat{\beta}_j| > \lambda_n}\}$, *for* $1 \leq j \leq p$, *satisfies, with the same probability,*

$$|\mathrm{supp}(\tilde{\beta}) \setminus \mathrm{supp}(\beta^*)| \leq \frac{24k}{\alpha_{\mathrm{RE}}}. \tag{3.1}$$

REMARKS. (a) The convergence rates of $\ell_2$-estimation and prediction $\sqrt{k \log p/n}$ are of the same order as the rates for regression with i.i.d. samples. The temporal dependence contributes the additional term $[\mathcal{M}(f_X, 1) + \mathcal{M}(f_\varepsilon)]/\mathfrak{m}(f_X)$ in the error rates and $[\mathcal{M}(f_X, 2k)/\mathfrak{m}(f_X)]^2$ in the sample size requirement. This ensures fast convergence rates of lasso under high-dimensional scaling, as long as the processes of predictors and noise are stable.

(b) A thresholded version of lasso enjoys small false positive rates, as shown in (3.1). Note that we do not assume any "beta-min" condition, that is, a lower bound on the minimum signal strength. It is possible to control the false negatives under suitable "beta-min" conditions, as shown in [Zhou (2010)].

*Comparison with existing results.* The problem of stochastic regression in a high-dimensional setting has been addressed by Loh and Wainwright (2012). After initial submission of this work, we became aware of a recent work by Wu and Wu (2014). Next, we briefly illustrate the major differences of our results with

these other studies. Loh and Wainwright (2012) assume that the process of predictors $\{X^t\}$ follows a Gaussian VAR(1) process with transition matrix satisfying $\|A\| < 1$. They also assume that the errors are independent. Our results allow both the predictors and the errors to be generated from any stable Gaussian process. Wu and Wu (2014) consider lasso estimation with a fixed design matrix and assume that an RE condition is satisfied. In our work, we consider a random Gaussian design and establish that RE holds with high probability for a large class of stable processes. Consequently, our final results of consistency do not rely on any RE type assumptions. Wu and Wu (2014) also consider random design regression using a CLIME estimator and provide an upper bound on the estimation error, without assuming RE type conditions. However, the established upper bounds seem to worsen with stronger signal ($|\beta|_1$). Our results do not exhibit such properties. Finally, both these papers consider a short-range dependence regime, although their results are derived under a mild moment condition on the random variables while we focus on Gaussian processes only. The results in the above paper quantify dependence via the functional and predictive measure of Wu (2005) and assume a certain decay condition on this measure. For the multivariate stationary linear processes, this is verified under another decay condition on the transition matrices in its AR representation [Chen, Xu and Wu (2013)]. Our results, on the other hand, rely on existence and boundedness of spectral density, and this assumption is satisfied by commonly used stable processes, including ARMA and general linear processes.

**4. Transition matrix estimation in sparse VAR models.** This problem has been considered by several authors in recent years [Davis, Zang and Zheng (2012), Han and Liu (2013), Song and Bickel (2011)]. Most of these studies consider a least squares based objective function or estimating equation to obtain the estimates, which is agnostic to the presence of cross-correlations among the error components (nondiagonal $\Sigma_\varepsilon$). Davis, Zang and Zheng (2012) provide numerical evidence that the forecasting performance can be improved by using a log-likelihood based loss function that incorporates information on the error correlations. In this section, we consider both least squares and log-likelihood estimates and study their theoretical properties. A key contribution of our theoretical analysis is to verify suitable RE and deviation conditions for the entire class of stable VAR($d$) models. Existing works either assume such conditions without verification, or use a stringent condition on the model parameters, such as $\|A\| < 1$, as discussed in Section 1.

We consider a single realization of $\{X^0, X^1, \ldots, X^T\}$ generated according to the VAR model (1.3). We will assume the error covariance matrix $\Sigma_\varepsilon$ is positive definite so that $\Lambda_{\min}(\Sigma_\varepsilon) > 0$ and $\Lambda_{\max}(\Sigma_\varepsilon) < \infty$. We will also assume that the VAR process is *stable*, that is, $\det(\mathcal{A}(z)) \neq 0$ on the unit circle $\{z \in \mathbb{C} : |z| = 1\}$. For stable VAR($d$) processes, the spectral density (2.4) simplifies to

$$f_X(\theta) = \frac{1}{2\pi}(\mathcal{A}^{-1}(e^{-i\theta}))\Sigma_\varepsilon(\mathcal{A}^{-1}(e^{-i\theta}))^*.$$

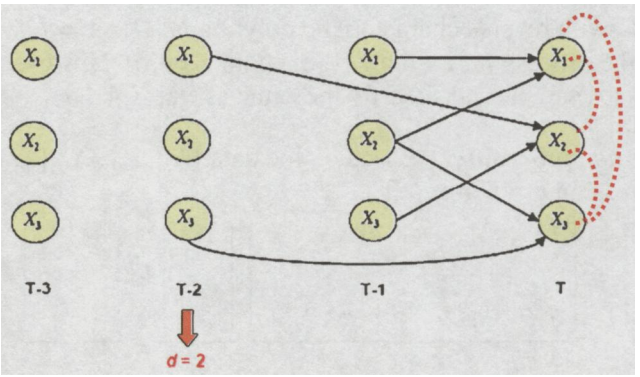


FIG. 4. *Graphical representation of the VAR model* (1.3): *directed edges* (*solid*) *correspond to the entries of the transition matrices, undirected edges* (*dashed*) *correspond to the entries of* $\Sigma_\varepsilon^{-1}$.

To deal with dependence in the VAR estimation problem, we will work with $\mu_{\min}(\mathcal{A})$, $\mu_{\max}(\mathcal{A})$ and the extreme eigenvalues of $\Sigma_\varepsilon$ instead of $\mathfrak{m}(f_X)$ and $\mathcal{M}(f_X)$. For a VAR($d$) process with serially uncorrelated errors, equation (2.6) simplifies to

$$\mathcal{M}(f_X) \le \frac{1}{2\pi}\frac{\Lambda_{\max}(\Sigma_\varepsilon)}{\mu_{\min}(\mathcal{A})}, \qquad \mathfrak{m}(f_X) \ge \frac{1}{2\pi}\frac{\Lambda_{\min}(\Sigma_\varepsilon)}{\mu_{\max}(\mathcal{A})}. \tag{4.1}$$

This factorization helps provide better insight into the temporal and contemporaneous dependence in VAR models. A graphical representation of a stable VAR($d$) model (1.3) is provided in Figure 4. The transition matrices $A_1, \ldots, A_d$ encode the temporal dependence of the process. When the components of the error process $\{\varepsilon^t\}$ are correlated, $\Sigma_\varepsilon^{-1}$ captures the additional contemporaneous dependence structure. Expressing the estimation and prediction errors in terms of $\mu_{\min}(\mathcal{A})$, $\mu_{\max}(\mathcal{A})$, $\Lambda_{\min}(\Sigma_\varepsilon)$ and $\Lambda_{\max}(\Sigma_\varepsilon)$ instead of $\mathfrak{m}(f_X)$ and $\mathcal{M}(f_X)$ help separate the effect of the two sources of dependence.

We will often use the following alternative representation of a $p$-dimensional VAR($d$) process (1.3) as a $dp$-dimensional VAR(1) process $\tilde{X}^t = \tilde{A}_1 \tilde{X}^{t-1} + \tilde{\varepsilon}^t$ with

$$\tilde{X}^t = \begin{bmatrix} X^t \\ X^{t-1} \\ \vdots \\ X^{t-d+1} \end{bmatrix}_{dp\times 1}, \qquad \tilde{A}_1 = \begin{bmatrix} A_1 & A_2 & \cdots & A_{d-1} & A_d \\ I_p & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_p & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & I_p & \mathbf{0} \end{bmatrix}_{dp\times dp},$$

$$\tilde{\varepsilon}^t = \begin{bmatrix} \varepsilon^t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}_{dp\times 1}. \tag{4.2}$$

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 [Lütkepohl (2005)]. However, the quantities $\mu_{\min}(\mathcal{A})$, $\mu_{\max}(\mathcal{A})$ are not necessarily the same as $\mu_{\min}(\tilde{\mathcal{A}})$, $\mu_{\max}(\tilde{\mathcal{A}})$.

4.1. *Estimation procedure.* Based on the data $\{X^0, \ldots, X^T\}$, we construct the following regression problem:

$$\underbrace{\begin{bmatrix} (X^T)' \\ \vdots \\ (X^d)' \end{bmatrix}}_{\mathcal{Y}} = \underbrace{\begin{bmatrix} (X^{T-1})' & \cdots & (X^{T-d})' \\ \vdots & \ddots & \vdots \\ (X^{d-1})' & \cdots & (X^0)' \end{bmatrix}}_{\mathcal{X}} \underbrace{\begin{bmatrix} A_1' \\ \vdots \\ A_d' \end{bmatrix}}_{B^*} + \underbrace{\begin{bmatrix} (\varepsilon^T)' \\ \vdots \\ (\varepsilon^d)' \end{bmatrix}}_{E},$$

$$\begin{aligned} \operatorname{vec}(\mathcal{Y}) &= \operatorname{vec}(\mathcal{X}B^*) + \operatorname{vec}(E), \\ &= (I \otimes \mathcal{X}) \operatorname{vec}(B^*) + \operatorname{vec}(E), \\ \underbrace{Y}_{Np\times 1} &= \underbrace{Z}_{Np\times q} \underbrace{\beta^*}_{q\times 1} + \underbrace{\operatorname{vec}(E)}_{Np\times 1}, \qquad N = (T-d+1), \qquad q = dp^2, \end{aligned}$$

with $N = T - d + 1$ samples and $q = dp^2$ variables. We will assume that $\beta^*$ is a $k$-sparse vector, that is, $\sum_{t=1}^{d} \| \operatorname{vec}(A_t)\|_0 = k$.

We consider the following estimates for the transition matrices $A_1, \ldots, A_d$, or equivalently, for $\beta^*$: (i) an $\ell_1$-penalized least squares estimate of VAR coefficients ($\ell_1$-LS), which does not exploit $\Sigma_\varepsilon$

$$\operatorname*{argmin}_{\beta\in\mathbb{R}^q} \frac{1}{N}\|Y - Z\beta\|^2 + \lambda_N \|\beta\|_1, \tag{4.3}$$

and (ii) an $\ell_1$-penalized log-likelihood estimation ($\ell_1$-LL) [Davis, Zang and Zheng (2012)].

$$\operatorname*{argmin}_{\beta\in\mathbb{R}^q} \frac{1}{N}(Y - Z\beta)'(\Sigma_\varepsilon^{-1} \otimes I)(Y - Z\beta) + \lambda_N \|\beta\|_1. \tag{4.4}$$

This gives the maximum likelihood estimate of $\beta$, for known $\Sigma_\varepsilon$. In practice, $\Sigma_\varepsilon$ is often unknown and needs to be estimated from the data. In the numerical experiments of Section 6, we used the residuals from a $\ell_1$-LS fit to estimate $\Sigma_\varepsilon$. Further discussion on estimating $\Sigma_\varepsilon$ and a fast algorithm based on block coordinate descent that minimizes (4.4) are presented in Appendix C (supplementary material [Basu and Michailidis (2015)]).

4.2. *Theoretical properties.* We analyze the estimates from optimization problems (4.3) and (4.4) under a general penalized M-estimation framework [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:

$$\operatorname*{argmin}_{\beta\in\mathbb{R}^q} -2\beta'\hat{\gamma} + \beta'\hat{\Gamma}\beta, \tag{4.5}$$

where $\hat{\Gamma} = (I \otimes \mathcal{X}'\mathcal{X}/N)$, $\hat{\gamma} = (I \otimes \mathcal{X}')Y/N$ 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:

$$\begin{aligned} &\operatorname*{argmin}_{\beta \in \mathbb{R}^q} -2\beta'\hat{\gamma} + \beta'\hat{\Gamma}\beta + \lambda_N \|\beta\|_1, \\ &\hat{\Gamma} = (W \otimes \mathcal{X}'\mathcal{X}/N), \qquad \hat{\gamma} = (W \otimes \mathcal{X}')Y/N, \end{aligned} \tag{4.6}$$

where $W$ is a symmetric, positive definite matrix of weights. Optimization problems (4.3) and (4.4) are special cases of (4.6) with $W = I$ and $W = \Sigma_\varepsilon^{-1}$, respectively.

First, we establish consistency of VAR estimates under the following sufficient conditions: a modified restricted eigenvalue (RE) [Loh and Wainwright (2012)] and a deviation condition. Then we show that all stable VAR models satisfy these assumptions with high probability, as long as the sample size is of the same order as required for consistency.

(A1) Restricted eigenvalue (RE). A symmetric matrix $\hat{\Gamma}_{q\times q}$ satisfies restricted eigenvalue condition with curvature $\alpha > 0$ and tolerance $\tau > 0$ ($\hat{\Gamma} \sim \mathrm{RE}(\alpha, \tau)$) if

$$\theta'\hat{\Gamma}\theta \geq \alpha\|\theta\|^2 - \tau\|\theta\|_1^2 \qquad \forall \theta \in \mathbb{R}^q. \tag{4.7}$$

The deviation 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 (4.5), $\hat{\gamma} - \hat{\Gamma}\beta^*$ is precisely $\mathrm{vec}(\mathcal{X}'E)/N$.

(A2) Deviation condition. There exists a deterministic function $\mathbb{Q}(\beta^*, \Sigma_\varepsilon)$ such that

$$\|\hat{\gamma} - \hat{\Gamma}\beta^*\|_\infty \leq \mathbb{Q}(\beta^*, \Sigma_\varepsilon)\sqrt{\frac{\log d + 2\log p}{N}}. \tag{4.8}$$

PROPOSITION 4.1 (Estimation and prediction error). *Consider the penalized M-estimation problem* (4.6) *with* $W = I$ *or* $W = \Sigma_\varepsilon^{-1}$. *Suppose* $\hat{\Gamma}$ *satisfies RE condition* (4.7) *with* $k\tau \leq \alpha/32$, *and* $(\hat{\Gamma}, \hat{\gamma})$ *satisfies deviation bound* (4.8). *Then, for any* $\lambda_N \geq 4\mathbb{Q}(\beta^*, \Sigma_\varepsilon)\sqrt{(\log d + 2\log p)/N}$, *any solution* $\hat{\beta}$ *of* (4.6) *satisfies*

$$\begin{aligned} \|\hat{\beta} - \beta^*\|_1 &\leq 64k\lambda_N/\alpha, \\ \|\hat{\beta} - \beta^*\| &\leq 16\sqrt{k}\lambda_N/\alpha, \\ (\hat{\beta} - \beta^*)'\hat{\Gamma}(\hat{\beta} - \beta^*) &\leq 128k\lambda_N^2/\alpha. \end{aligned}$$

*Further, a thresholded variant of lasso* $\tilde{\beta} = \{\hat{\beta}_j \mathbf{1}_{|\hat{\beta}_j| > \lambda_N}\}$ *satisfies*

$$|\mathrm{supp}(\tilde{\beta}) \setminus \mathrm{supp}(\beta^*)| \leq \frac{192k}{\alpha_{\mathrm{RE}}}.$$

REMARKS. (a) $\|\hat{\beta} - \beta^*\|$ 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 by $\sum_{t=d}^{T} \|\sum_{h=1}^{d}(\hat{A}_h - A_h)X^{t-h}\|^2/N$. For $\ell_1$-LL, $(\hat{\beta} - \beta^*)'\hat{\Gamma}(\hat{\beta} - \beta^*)$ takes the form $\sum_{t=d}^{T} \|\sum_{h=1}^{d}(\hat{A}_h - A_h)X^{t-h}\|_{\Sigma_\varepsilon}^2/N$, 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 $\Sigma_\varepsilon$.

(b) The convergence rates are governed by two sets of parameters: (i) dimensionality parameters, the dimension of the process ($p$), order of the process ($d$), number of parameters ($k$) in the transition matrices $A_i$ and sample size ($N = T - d + 1$); (ii) internal parameters, the curvature ($\alpha$), tolerance ($\tau$) and the deviation bound $\mathbb{Q}(\beta^*, \Sigma_\varepsilon)$. The squared $\ell_2$-errors of estimation and prediction scale with the dimensionality parameters as $k(2\log p + \log d)/N$, similar to the rates obtained when the observations are independent [Bickel, Ritov and Tsybakov (2009)]. The temporal and cross-sectional dependence affect the rates only through the internal parameters. Typically, the rates are better when $\alpha$ is large and $\mathbb{Q}(\beta^*, \Sigma_\varepsilon)$, $\tau$ 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.

(c) Although the above proposition is derived under the assumption that $d$ is the true order of the VAR process, the results hold even if $d$ is replaced by any upper bound $\bar{d}$ on the true order. This follows from the fact that a VAR($d$) model can also be viewed as VAR($\bar{d}$), for any $\bar{d} > d$, with transition matrices $A_1, \ldots, A_d, 0_{p\times p}, \ldots, 0_{p\times p}$. Note that the convergence rates change from $\sqrt{(\log p + 2\log d)/N}$ to $\sqrt{(\log p + 2\log \bar{d})/N}$.

Proposition 4.1 is deterministic; that is, 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 assumptions (A1–A2) 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.

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* (1.3). *Then there exist constants $c_i > 0$ such that for all $N \gtrsim \max\{\omega^2, 1\}k(\log d + \log p)$, with probability at least $1 - c_1\exp(-c_2 N \min\{\omega^{-2}, 1\})$, the matrix*

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

*where*

$$\omega = c_3 \frac{\Lambda_{\max}(\Sigma_\varepsilon)/\mu_{\min}(\tilde{\mathcal{A}})}{\Lambda_{\min}(\Sigma_\varepsilon)/\mu_{\max}(\mathcal{A})}, \qquad \alpha = \frac{\Lambda_{\min}(\Sigma_\varepsilon)}{2\mu_{\max}(\mathcal{A})},$$

$$\tau = \alpha \max\{\omega^2, 1\}\frac{\log d + \log p}{N}.$$

*Further, if* $\Sigma_\varepsilon^{-1}$ *satisfies* $\bar{\sigma}_\varepsilon^i := \sigma_\varepsilon^{ii} - \sum_{j\neq i} \sigma_\varepsilon^{ij} > 0$, *for* $i = 1, \ldots, p$, *then, with the same probability as above, the matrix*

$$\hat{\Gamma} = \Sigma_\varepsilon^{-1} \otimes (\mathcal{X}'\mathcal{X}/N) \sim \mathrm{RE}\Big(\alpha \min_i \bar{\sigma}_\varepsilon^i, \tau \max_i \bar{\sigma}_\varepsilon^i\Big).$$

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 smaller error bounds when $\Lambda_{\max}(\Sigma_\varepsilon)$, $\mu_{\max}(\mathcal{A})$ are smaller and $\Lambda_{\min}(\Sigma_\varepsilon)$, $\mu_{\min}(\tilde{\mathcal{A}})$ are larger, that is, when the spectrum is less spiky.

PROPOSITION 4.3 (Deviation bound). *There exist constants* $c_i > 0$ *such that for* $N \gtrsim (\log d + 2\log p)$, *with probability at least* $1 - c_1 \exp[-c_2(\log d + 2\log p)]$, *we have*

$$\|\hat{\gamma} - \hat{\Gamma}\beta^*\|_\infty \leq \mathbb{Q}(\beta^*, \Sigma_\varepsilon)\sqrt{\frac{\log d + 2\log p}{N}},$$

*where, for* $\ell_1$*-LS,*

$$\mathbb{Q}(\beta^*, \Sigma_\varepsilon) = c_0\Big[\Lambda_{\max}(\Sigma_\varepsilon) + \frac{\Lambda_{\max}(\Sigma_\varepsilon)}{\mu_{\min}(\mathcal{A})} + \frac{\Lambda_{\max}(\Sigma_\varepsilon)\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\Big]$$

*and for* $\ell_1$*-LL,*

$$\mathbb{Q}(\beta^*, \Sigma_\varepsilon) = c_0\Big[\frac{1}{\Lambda_{\min}(\Sigma_\varepsilon)} + \frac{\Lambda_{\max}(\Sigma_\varepsilon)}{\mu_{\min}(\mathcal{A})} + \frac{\Lambda_{\max}(\Sigma_\varepsilon)\mu_{\max}(\mathcal{A})}{\Lambda_{\min}(\Sigma_\varepsilon)\mu_{\min}(\mathcal{A})}\Big].$$

As before, this proposition shows that the VAR estimates have lower error bounds when $\Lambda_{\max}(\Sigma_\varepsilon)$, $\mu_{\max}(\mathcal{A})$ are smaller and $\Lambda_{\min}(\Sigma_\varepsilon)$, $\mu_{\min}(\mathcal{A})$ are larger, that is, when the spectrum is less spiky.

*Comparison with existing results.* The problem of sparse VAR estimation has been theoretically studied in the literature in [Chudik and Pesaran (2011), Song and Bickel (2011), Wu and Wu (2014)]. Next, we briefly highlight differences between our results and these works. First, the results of Chudik and Pesaran (2011) rely on a priori available neighborhood information for every time series, which implies that the structure of transition matrices $\{A_t\}_{t=1}^d$ is known, and only their magnitudes need to be estimated. This is a significant limitation compared to regularized methods like lasso, which do not require any prior knowledge on the sparsity pattern in the transition matrices. The theoretical upper bounds on VAR estimation error established in Song and Bickel (2011) do not decrease as the sample size $T$ increases, and hence do not ensure consistency beyond very strict conditions. Also, the results in their paper and in Wu and Wu (2014) are established assuming RE holds, while a significant portion of our analysis is devoted to establish that RE and

deviation bounds hold with high probability. We also provide in-depth analysis on how the relevant constants are affected by the dependence present in the data. Finally, our work is the first one to provide theoretical analysis of the log-likelihood based VAR estimation procedure, which does not fit directly into the regression setting considered in the aforementioned papers.

**5. Extension to other regularized estimation problems.** The deviation inequalities established in Section 2 can be easily integrated with the vast body of existing literature of high-dimensional statistics for i.i.d. data and study other regularized estimation problems in the context of high-dimensional time series. To demonstrate this, in this section we establish consistency of sparse covariance estimation by hard-thresholding [Bickel and Levina (2008)] for high-dimensional time series and discusss the main steps in extending the results to some nonconvex penalties for sparse regression and group lasso and nuclear norm penalties for inducing structured sparsity.

5.1. *Sparse covariance estimation.* Consider a $p$-dimensional centered Gaussian stationary time series $\{X^t\}_{t\in\mathbb{Z}}$ satisfying Assumption 2.1. Based on realizations $\{X^1, \ldots, X^n\}$ generated according to the above stationary process, we aim to estimate the contemporaneous covariance matrix $\Sigma = \Gamma(0)$. The sample covariance matrix $\hat{\Gamma}(0) = \frac{1}{n}\sum_{t=1}^{n}(X^t - \bar{X})(X^t - \bar{X})'$ is known to be inconsistent when $p$ grows faster than $n$. Bickel and Levina (2008) showed that when the samples are generated independently from a centered Gaussian or subGaussian distribution, a thresholded version of the sample covariance matrix $T_u(\hat{\Gamma}(0)) = \{\hat{\Gamma}_{ij}(0)\mathbf{1}_{|\hat{\Gamma}_{ij}(0)|>u}\}$ can perform consistent estimation if $\Gamma(0)$ belongs to the following uniformity class of approximately sparse matrices:

$$\mathcal{U}_\tau(q, c_0(p), M) := \left\{\Sigma : \sigma_{ii} \le M, \sum_{j=1}^{p} |\sigma_{ij}|^q \le c_0(p), \text{ for all } i\right\}.$$

Next, we establish consistent estimation for time series data, provided that the underlying process is stable. The effect of dependence on the estimation accuracy is captured by the stability measures introduced in Section 2. Asymptotic theory for sparse covariance estimation was also considered in Chen, Xu and Wu (2013), assuming a decay on the functional dependence measure.

PROPOSITION 5.1. *Let $\{X^t\}_{t=1}^n$ be generated according to a $p$-dimensional stationary centered Gaussian process with spectral density $f_X$, satisfying Assumption 2.1. Then, uniformly on $\mathcal{U}_\tau(q, c_0(p), M)$, for sufficiently large $M'$, if $u_n = \mathcal{M}(f_X, 2)M'\sqrt{\log p/n}$ and $n \gtrsim \mathcal{M}^2(f_X, 2)\log p$, then*

$$\|T_{u_n}(\hat{\Gamma}(0)) - \Gamma(0)\| = O_p\left(c_0(p)\left(\mathcal{M}^2(f_X, 2)\frac{\log p}{n}\right)^{(1-q)/2}\right),$$

$$\frac{1}{p}\|T_{u_n}(\hat{\Gamma}(0)) - \Gamma(0)\|_F = O_p\left(c_0(p)\left(\mathcal{M}^2(f_X, 2)\frac{\log p}{n}\right)^{1-(q/2)}\right).$$

5.2. *Sparse regression with nonconvex penalties.* There is a vast body of literature on regularized regression using nonconvex penalties for i.i.d. data [Fan and Li (2001), Zhang (2010)]. A recent line work has derived unified theoretical treatments of these procedures and compared their estimation accuracy to convex procedures such as lasso [Fan and Lv (2013), Loh and Wainwright (2013)]. These results indicate that in certain high-dimensional regimes, the estimation error of nonconvex penalties like SCAD, MCP scales roughly in the same order as lasso. Next, we argue that similar conclusions hold for time series models, as well.

Consider a stochastic regression problem of Section 3 subject to a SCAD or MCP penalty. Loh and Wainwright (2013) establish that under suitable restricted strong convexity (RSC) condition on the loss function $\mathcal{L}_n(\cdot)$, if the sup norm of the gradient $\|\nabla(L)_n(\beta^*)\|_\infty$ scales with $\sqrt{\log p/n}$, then any local solution of the penalized objective function has an estimation error at most $O(\sqrt{k \log p/n})$. For the choice of a least squares loss function, $\mathcal{L}_n(\beta) = \|\mathcal{Y} - \mathcal{X}\beta\|^2/2n$ and $\nabla\mathcal{L}_n(\beta^*) = -\mathcal{X}'E/n$.

Since the loss function is convex, their RSC takes the form

$$\frac{1}{n}\frac{\|\mathcal{X}\Delta\|^2}{\|\Delta\|^2} \geq \alpha_1\|\Delta\|^2 - \tau_1\frac{\log p}{n}\|\Delta\|_1^2 \qquad \text{for all } \|\Delta\| \leq 1.$$

This is in the spirit of the RE conditions verified in Section 4 and can be proven using similar discretization arguments presented in this paper, if we assume $\Gamma(0)$ satisfies an RE condition with the restricted eigenvalue $\alpha_1$ is at least as large as $1/(a-1)$ for SCAD and $1/b$ for MCP.

The deviation condition on $\|\nabla\mathcal{L}_n(\beta^*)\|_\infty$ is identical to the one considered in this paper, and the results presented here are directly applicable.

5.3. *Regularized regression with structured sparsity.* In a recent review paper, Negahban et al. (2012) established a unified framework to analyze a class of decomposable penalties. This includes the popular group lasso penalty for high-dimensional regression under structured sparsity and nuclear norm penalty for matrix estimation under low-rank assumption. In a time series context, these methods have been proposed in the literature to incorporate information on different economic sectors and the assumption of latent factors driving the market [Negahban and Wainwright (2011), Song and Bickel (2011)]. As before, the theoretical results rely crucially on two key conditions: a restricted strong convexity on the loss function and a suitable deviation bound on the gradient. The restricted eigenvalue assumption for group lasso can be verified using the deviation inequalities of Proposition 2.4 and a discretization argument modified for group structures. The deviation inequalities can be derived along the same line. For low-rank modeling of VAR(1) process, we can prove that the minimum eigenvalue of $\mathcal{X}'\mathcal{X}/N$ is bounded away from zero with high probability, and the deviation bounds on the op-

erator norm of $\mathcal{X}'E/N$ can be established using the deviation inequality of (2.11) and a discretization argument presented in [Basu (2014)]. This leads to new results on group lasso for stochastic regression and extends the results of Negahban and Wainwright (2011) to the entire class of stable VAR(1) models. We leave the details to the reader, as the proofs follow the same road map used in this paper.

## 6. Numerical experiments.

6.1. *Stochastic regression.* In this experiment, we demonstrate how the estimation error of lasso scales with $n$ and $p$, when the dependence parameters do not change. We simulate predictors from a $p$-dimensional ($p = 128, 264, 512, 1024$) stationary process $\{X^t\}$ with independent components following a Gaussian AR(2) process $X_i^t = 1.2X_i^{t-1} - 0.36X_i^{t-2} + \xi^t$, $\Gamma_{X_j}(0) = 1$. We simulate the errors $\{\varepsilon^t\}$ according to a univariate MA(2) process $\varepsilon^t = \eta^t - 0.8\eta^{t-1} + 0.16\eta^{t-2}$, $\{\eta^t\}$ Gaussian white noise. For different values of $p$, we generate sparse coefficient vectors $\beta^*$ with $k \approx \sqrt{p}$ nonzero entries, with a signal-to-noise ratio of 1.2. Using a tuning parameter $\lambda_n = \sqrt{\log p/n}$, we apply lasso on simulated samples of size $n \in (100, 3000)$. The $\ell_2$-error of estimation $\|\hat{\beta} - \beta^*\|$ is depicted in Figure 5. The left panel displays the errors for different values of $p$, plotted against the sample size $n$. As expected, the errors are larger for larger $p$. The right panel displays the estimation errors against the rescaled sample size $n/k \log p$. The error curves for different values of $p$ now align very well. This demonstrates that lasso can achieve an estimation error rate of $\sqrt{k \log p/n}$, even with stochastic predictors and serially correlated errors.

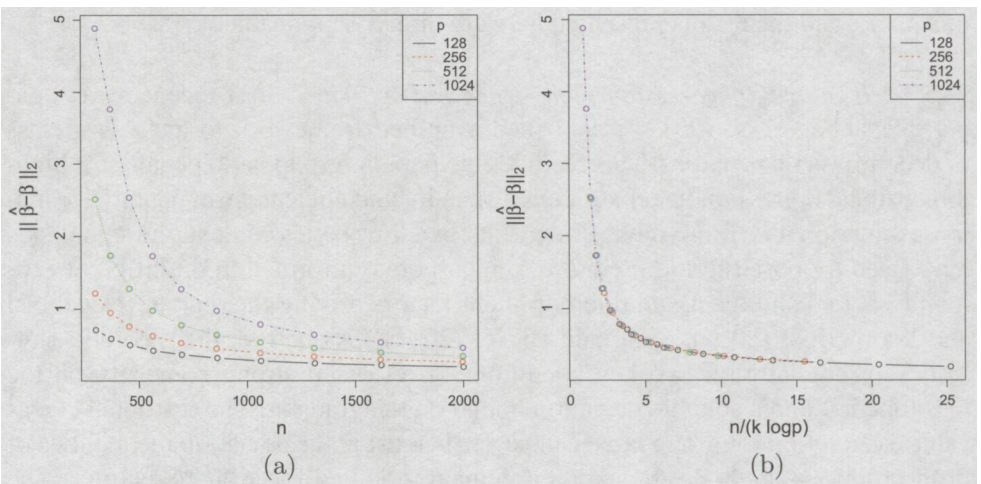


FIG. 5. *Estimation error of lasso $\|\hat{\beta} - \beta^*\|$ in stochastic regression with serially correlated error. Note that the error curves align perfectly, showing the errors scale as $\sqrt{k \log p/n}$.* (a) $\|\hat{\beta} - \beta^*\|$ *vs.* $n$, (b) $\|\hat{\beta} - \beta^*\|$ *vs.* $n/k \log p$.

6.2. *VAR estimation.* We evaluate the performance of $\ell_1$-LS and $\ell_1$-LL on simulated data and compare it with the performance of ordinary least squares (OLS) and Ridge estimates. Implementing $\ell_1$-LL requires an estimate of $\Sigma_\varepsilon$ in the first step. We use the residuals from $\ell_1$-LS to construct a plug-in estimate $\hat{\Sigma}_\varepsilon$. 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 $\Sigma_\varepsilon$ 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 $5 \sim 10\%$ nonzero 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 $\Sigma_\varepsilon$ from one of the following families:

(1) Block-I: $\Sigma_\varepsilon = ((\sigma_{\varepsilon,ij}))_{1\le i,j\le p}$ with $\sigma_{\varepsilon,ii} = 1$, $\sigma_{\varepsilon,ij} = \rho$ if $1 \le i \ne j \le p/2$, $\sigma_{\varepsilon,ij} = 0$ otherwise;
(2) Block-II: $\Sigma_\varepsilon = ((\sigma_{\varepsilon,ij}))_{1\le i,j\le p}$ with $\sigma_{\varepsilon,ii} = 1$, $\sigma_{\varepsilon,ij} = \rho$ if $1 \le i \ne j \le p/2$ or $p/2 < i \ne j \le p$, $\sigma_{\varepsilon,ij} = 0$ otherwise;
(3) Toeplitz: $\Sigma_\varepsilon = ((\sigma_{\varepsilon,ij}))_{1\le i,j\le p}$ with $\sigma_{\varepsilon,ij} = \rho^{|i-j|}$.

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 6 illustrates the structure of a random transition matrix used in our simulation and the three different types of error covariance structures.

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 ROC curve (AUROC);
(2) *Estimation error.* Relative estimation accuracy $\|\hat{A}_1 - A_1\|_F / \|A_1\|_F$.

We report the results for small VAR with $T = 30$ and medium VAR with $T = 120$ averaged over 1000 replicates in Tables 1 and 2. The results in the other settings are qualitatively similar, although the overall accuracy changes with the

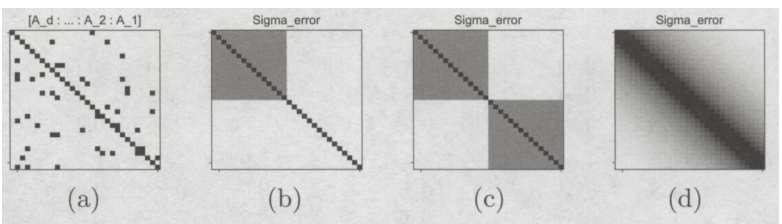


FIG. 6. *Adjacency matrix $A_1$ and error covariance matrix $\Sigma_\varepsilon$ of different types used in the simulation studies.* (a) $A_1$, (b) $\Sigma_\varepsilon$: *Block*-I, (c) $\Sigma_\varepsilon$: *Block*-II, (d) $\Sigma_\varepsilon$: *Toeplitz.*

TABLE 1
VAR(1) *model with* $p = 10$, $T = 30$

| | | Block-I | | | Block-II | | | Toeplitz | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho$ | **0.5** | **0.7** | **0.9** | **0.5** | **0.7** | **0.9** | **0.5** | **0.7** | **0.9** |
| AUROC | $\ell_1$-LS | 0.78 | 0.77 | 0.74 | 0.74 | 0.7 | 0.64 | 0.76 | 0.72 | 0.63 |
| | $\ell_1$-LL | 0.79 | 0.79 | 0.76 | 0.77 | 0.77 | 0.76 | 0.78 | 0.76 | 0.74 |
| | $\ell_1$-LL-O | 0.84 | 0.83 | 0.8 | 0.82 | 0.82 | 0.82 | 0.83 | 0.82 | 0.8 |
| Estimation error | OLS | 1.51 | 1.67 | 2.31 | 1.73 | 2.16 | 3.57 | 1.7 | 2.14 | 3.57 |
| | $\ell_1$-LS | 0.74 | 0.75 | 0.76 | 0.77 | 0.8 | 0.87 | 0.77 | 0.8 | 0.88 |
| | $\ell_1$-LL | 0.7 | 0.7 | 0.69 | 0.73 | 0.72 | 0.72 | 0.73 | 0.73 | 0.74 |
| | $\ell_1$-LL-O | 0.65 | 0.64 | 0.63 | 0.66 | 0.65 | 0.63 | 0.66 | 0.66 | 0.65 |
| | Ridge | 0.78 | 0.78 | 0.79 | 0.77 | 0.78 | 0.8 | 0.8 | 0.82 | 0.85 |

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. OLS and ridge regression do not perform any model selection. Further, for all three choices of $\Sigma_\varepsilon$, the two variants of $\ell_1$-LL outperform $\ell_1$-LS. The difference in their performance is more prominent for larger values of $\rho$. Among the three covariance structures, the difference between LS- and LL-based methods is more prominent in the Block-II and Toeplitz families, since the error processes are more strongly correlated. Finally, in all cases, the accuracy of $\ell_1$-LL lies between $\ell_1$-LS and $\ell_1$-LL-O, which suggests that a more accurate estimation of $\Sigma_\varepsilon$ might improve the model selection performance of regularized VAR estimates.

In terms of estimation error, the conclusions are broadly the same. The effect of over-fitting is reflected in the performance of ordinary least squares. In many

TABLE 2
VAR(1) *model with* $p = 30$, $T = 120$

| | | Block-I | | | Block-II | | | Toeplitz | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho$ | **0.5** | **0.7** | **0.9** | **0.5** | **0.7** | **0.9** | **0.5** | **0.7** | **0.9** |
| AUROC | $\ell_1$-LS | 0.91 | 0.87 | 0.8 | 0.82 | 0.75 | 0.63 | 0.92 | 0.88 | 0.77 |
| | $\ell_1$-LL | 0.91 | 0.89 | 0.85 | 0.85 | 0.85 | 0.85 | 0.93 | 0.92 | 0.91 |
| | $\ell_1$-LL-O | 0.93 | 0.91 | 0.87 | 0.88 | 0.88 | 0.88 | 0.95 | 0.94 | 0.92 |
| Estimation error | OLS | 1.65 | 1.91 | 2.74 | 2.33 | 2.98 | 4.94 | 1.77 | 2.24 | 3.74 |
| | $\ell_1$-LS | 0.68 | 0.73 | 0.8 | 0.83 | 0.9 | 0.98 | 0.68 | 0.72 | 0.85 |
| | $\ell_1$-LL | 0.67 | 0.67 | 0.67 | 0.78 | 0.77 | 0.74 | 0.65 | 0.62 | 0.57 |
| | $\ell_1$-LL-O | 0.63 | 0.63 | 0.63 | 0.74 | 0.73 | 0.7 | 0.61 | 0.57 | 0.52 |
| | Ridge | 0.8 | 0.81 | 0.83 | 0.86 | 0.89 | 0.92 | 0.8 | 0.82 | 0.86 |

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 cases. This is somewhat expected as the data were simulated from a sparse model with strong signals, whereas Ridge regression tends to favor a nonsparse model with many small coefficients.

**7. Discussion.** In this paper, we consider the theoretical properties of regularized estimates in sparse high-dimensional time series models when the data are generated from a multivariate stationary Gaussian process. The Gaussian assumption could be conceived as a limiting factor, since interesting models including regression with categorical predictors, VAR estimation with heavy-tailed and/or heteroscedastic errors, and popular models exhibiting nonlinear dependences such as ARCH and GARCH are not covered. Note, however, that the only place in the analysis where the Gaussian assumption is used is in developing the concentration bound of $S$ around its expectation $\Gamma(0)$. Since the spectral density characterizes the entire distribution for this class, it has direct implications on the concentration behavior. For nonlinear and/or non-Gaussian processes, one needs to control higher order dependence, and changing to higher order spectra could potentially be useful. Although the use of covariance and higher order spectra is common in developing limit theorems of low-dimensional stationary process [Giraitis, Koul and Surgailis (2012), Rosenblatt (1985)], developing a suitable concentration bound for nonlinear/non-Gaussian dependence designs is not a trivial problem and is left as a key topic for future developments.

**Acknowledgements.** We thank the Editor Runze Li, the Associate Editor and three anonymous reviewers, whose comments led to several improvements in the paper.

## SUPPLEMENTARY MATERIAL

**Supplement to "Regularized estimation in sparse high-dimensional time series models"** (DOI: 10.1214/15-AOS1315SUPP; .pdf). For the sake of brevity, we moved the appendices containing many of the technical proofs and detailed discussions to the supplementary document [Basu and Michailidis (2015)].

## REFERENCES

BAŃBURA, M., GIANNONE, D. and REICHLIN, L. (2010). Large Bayesian vector auto regressions. *J. Appl. Econometrics* **25** 71–92. MR2751790

BASU, S. (2014). Modeling and estimation of high-dimensional vector autoregressions. Ph.D. thesis, Univ. Michigan, Ann Arbor, MI.

BASU, S. and MICHAILIDIS, G. (2015). Supplement to "Regularized estimation in sparse high-dimensional time series models." DOI:10.1214/15-AOS1315SUPP.

BERNANKE, B. S., BOIVIN, J. and ELIASZ, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *Q. J. Econ.* **120** 387–422.

BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

CHEN, X., XU, M. and WU, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Ann. Statist.* **41** 2994–3021. MR3161455

CHUDIK, A. and PESARAN, M. H. (2011). Infinite-dimensional VARs and factor models. *J. Econometrics* **163** 4–22. MR2803662

DAVIS, R. A., ZANG, P. and ZHENG, T. (2012). Sparse vector autoregressive modeling. Preprint. Available at arXiv:1207.0520.

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? *J. Econometrics* **146** 318–328. MR2465176

DOBRIBAN, E. and FAN, J. (2013). Regularity properties of high-dimensional covariate matrices. Preprint. Available at arXiv:1305.5198.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, Y. and LV, J. (2013). Asymptotic equivalence of regularization methods in thresholded parameter space. *J. Amer. Statist. Assoc.* **108** 1044–1061. MR3174683

FAN, J., LV, J. and QI, L. (2011). Sparse high-dimensional models in economics. *Annual Review of Economics* **3** 291–317.

FRISTON, K. (2009). Causal modelling and brain connectivity in functional magnetic resonance imaging. *PLoS Biol.* **7** e1000033.

GIRAITIS, L., KOUL, H. L. and SURGAILIS, D. (2012). *Large Sample Inference for Long Memory Processes*. Imperial College Press, London. MR2977317

GRENANDER, U. and SZEGÖ, G. (1958). *Toeplitz Forms and Their Applications*. Univ. California Press, Berkeley. MR0094840

HAMILTON, J. D. (1994). *Time Series Analysis*. Princeton Univ. Press, Princeton, NJ. MR1278033

HAN, F. and LIU, H. (2013). Transition matrix estimation in high dimensional time series. *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* **28** 172–180.

KUMAR, P. R. and VARAIYA, P. (1986). *Stochastic Systems: Estimation, Identification and Adaptive Control*. Prentice Hall, New York.

LIEBSCHER, E. (2005). Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. *J. Time Series Anal.* **26** 669–689. MR2188304

LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.* **40** 1637–1664. MR3015038

LOH, P.-L. and WAINWRIGHT, M. J. (2013). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. Preprint. Available at arXiv:1305.2436.

LÜTKEPOHL, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin. MR2172368

MICHAILIDIS, G. and D'ALCHÉ-BUC, F. (2013). Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Math. Biosci.* **246** 326–334. MR3132054

NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** 1069–1097. MR2816348

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of $M$-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133

PARTER, S. V. (1961). Extreme eigenvalues of Toeplitz forms and applications to elliptic difference equations. *Trans. Amer. Math. Soc.* **99** 153–192. MR0120492

PRIESTLEY, M. B. (1981). *Spectral Analysis and Time Series. Vol. 2. Multivariate Series, Prediction and Control, Probability and Mathematical Statistics*. Academic Press, London. MR0628736

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259. MR2719855

ROSENBLATT, M. (1985). *Stationary Sequences and Random Fields*. Springer, Boston, MA. MR885090

RUDELSON, M. and VERSHYNIN, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** no. 82, 9. MR3125258

RUDELSON, M. and ZHOU, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theory* **59** 3434–3447. MR3061256

SETH, A. K., CHORLEY, P. and BARNETT, L. C. (2013). Granger causality analysis of fMRI BOLD signals is invariant to hemodynamic convolution but not downsampling. *NeuroImage* **65** 540–555.

SHOJAIE, A. and MICHAILIDIS, G. (2010). Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics* **26** i517–i523.

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

SMITH, S. M. (2012). The future of FMRI connectivity. *NeuroImage* **62** 1257–1266.

SONG, S. and BICKEL, P. J. (2011). Large vector auto regressions. Preprint. Available at arXiv:1106.3915v1.

STOCK, J. H. and WATSON, M. W. (2005). Implications of dynamic factor models for VAR analysis. Working Paper No. 11467, National Bureau of Economic Research, Cambridge, MA.

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. MR2576316

VAN DE GEER, S., BÜHLMANN, P. and ZHOU, S. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Stat.* **5** 688–749. MR2820636

VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. Preprint. Available at arXiv:1011.3027.

WU, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proc. Natl. Acad. Sci. USA* **102** 14150–14154 (electronic). MR2172215

WU, W.-B. and WU, Y. N. (2014). High-dimensional linear models with dependent observations. Preprint.

XIAO, H. and WU, W. B. (2012). Covariance matrix estimation for stationary time series. *Ann. Statist.* **40** 466–493. MR3014314

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZHOU, S. (2010). Thresholded Lasso for high dimensional variable selection and statistical estimation. Technical Report 511, Dept. Statistics, Univ. Michigan, Ann Arbor, MI. Available at arXiv:1002.1583.

DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
ANN ARBOR MICHIGAN 48109
USA
E-MAIL: sumbose@umich.edu
gmichail@umich.edu