

---

# Estimating Structured Vector Autoregressive Models

---

Igor Melnyk  
Arindam Banerjee

MELNYK@CS.UMN.EDU  
BANERJEE@CS.UMN.EDU

Department of Computer Science and Engineering, University of Minnesota, Twin Cities

## Abstract

While considerable advances have been made in estimating high-dimensional structured models from independent data using Lasso-type models, limited progress has been made for settings when the samples are dependent. We consider estimating structured VAR (vector autoregressive model), where the structure can be captured by any suitable norm, e.g., Lasso, group Lasso, order weighted Lasso, etc. In VAR setting with correlated noise, although there is strong dependence over time and covariates, we establish bounds on the non-asymptotic estimation error of structured VAR parameters. The estimation error is of the same order as that of the corresponding Lasso-type estimator with independent samples, and the analysis holds for any norm. Our analysis relies on results in generic chaining, sub-exponential martingales, and spectral representation of VAR models. Experimental results on synthetic and real data with a variety of structures are presented, validating theoretical results.

## 1. Introduction

The past decade has seen considerable progress on approaches to structured parameter estimation, especially in the linear regression setting, where one considers regularized estimation problems of the form:

$$\hat{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}^q} \frac{1}{M} \|\mathbf{y} - Z\beta\|_2^2 + \lambda_M R(\beta), \quad (1)$$

where  $\{(y_i, z_i), i = 1, \dots, M\}$ ,  $y_i \in \mathbb{R}$ ,  $z_i \in \mathbb{R}^q$ , such that  $\mathbf{y} = [y_1^T, \dots, y_M^T]^T$  and  $Z = [z_1^T, \dots, z_M^T]^T$ , is the training set of  $M$  independently and identically distributed (i.i.d.) samples,  $\lambda_M > 0$  is a regularization parameter, and  $R(\cdot)$  denotes a suitable norm (Tibshirani,

1996; Zou & Hastie, 2005; Yuan & Lin, 2006). Specific choices of  $R(\cdot)$  lead to certain types of structured parameters to be estimated. For example, the decomposable norm  $R(\beta) = \|\beta\|_1$  yields Lasso, estimating sparse parameters,  $R(\beta) = \|\beta\|_G$  gives Group Lasso, estimating group sparse parameters, and  $R(\beta) = \|\beta\|_{owl}$ , the ordered weighted  $L_1$  norm (OWL) (Bogdan et al., 2013), gives sorted  $L_1$ -penalized estimator, clustering correlated regression parameters (Figueiredo & Nowak, 2014). Non-decomposable norms, such as  $K$ -support norm (Argyriou et al., 2012) or overlapping group sparsity norm (Jacob et al., 2009) can be used to uncover more complicated model structures. Theoretical analysis of such models, including sample complexity and non-asymptotic bounds on the estimation error rely on the design matrix  $Z$ , usually assumed (sub)-Gaussian with independent rows, and the specific norm  $R(\cdot)$  under consideration (Raskutti et al., 2010; Rudelson & Zhou, 2013). Recent work has generalized such estimators to work with any norm (Negahban et al., 2012; Banerjee et al., 2014) with i.i.d. rows in  $Z$ .

The focus of the current paper is on structured estimation in vector autoregressive (VAR) models (Lutkepohl, 2007), arguably the most widely used family of multivariate time series models. VAR models have been applied widely, ranging from describing the behavior of economic and financial time series (Tsay, 2005) to modeling the dynamical systems (Ljung, 1998) and estimating brain function connectivity (Valdes-Sosa et al., 2005), among others. A VAR model of order  $d$  is defined as

$$x_t = A_1 x_{t-1} + A_2 x_{t-2} + \dots + A_d x_{t-d} + \epsilon_t, \quad (2)$$

where  $x_t \in \mathbb{R}^p$  denotes a multivariate time series,  $A_k \in \mathbb{R}^{p \times p}$ ,  $k = 1, \dots, d$  are the parameters of the model, and  $d \geq 1$  is the order of the model. In this work, we assume that the noise  $\epsilon_t \in \mathbb{R}^p$  follows a Gaussian distribution,  $\epsilon_t \sim \mathcal{N}(0, \Sigma)$ , with  $\mathbb{E}(\epsilon_t \epsilon_t^T) = \Sigma$  and  $\mathbb{E}(\epsilon_t \epsilon_{t+\tau}^T) = 0$ , for  $\tau \neq 0$ . The VAR process is assumed to be stable and stationary (Lutkepohl, 2007), while the noise covariance matrix  $\Sigma$  is assumed to be positive definite with bounded largest eigenvalue, i.e.,  $\Lambda_{\min}(\Sigma) > 0$  and  $\Lambda_{\max}(\Sigma) < \infty$ .

In the current context, the parameters  $\{A_k\}$  are assumed

to be structured, in the sense of having low values according to a suitable norm  $R(\cdot)$ . We consider a general setting where *any* norm can be applied to the rows  $A_k(i, \cdot) \in \mathbb{R}^p$  of  $A_k$ , allowing the possibility of different norms being applied to different rows of  $A_k$ , and different norms for different parameter matrices  $A_k, k = 1, \dots, d$ . Choosing  $L_1$ -norm  $\|A_k(i, \cdot)\|_1$  for all rows and all parameter matrices is a simple special case of our setting. We discuss certain other choices in Section 2.1, and discuss related results in Section 4. In order to estimate the parameters, one can consider regularized estimators of the form (1), where  $y_i$  and  $z_i$  correspond to  $x_t$  in the VAR setting. Unfortunately, unlike  $(y_i, z_i)$  in (1), the  $x_t$  are *far from independent*, having strong dependence across time and correlated across dimensions. As a result, existing results from the rich literature on regularized estimators for structured problems (Zhao & Yu, 2006; Wainwright, 2009; Meinshausen & Yu, 2009) cannot be directly applied to get sample complexities and estimation error bounds in VAR models.

**Related Work:** In recent literature, the problem of estimating structured VAR models has been considered for the special case of  $L_1$  norm. (Han & Liu, 2013) analyzed a constrained estimator based on the Dantzig selector (Candes & Tao, 2007), and established the recovery results for the special case of  $L_1$  norm. (Song & Bickel, 2011) considered a regularized VAR estimation problem under Lasso and Group Lasso penalties and derived oracle inequalities for the prediction error and estimation accuracy. However, their analysis is for the case when the dimensionality of the problem is fixed with respect to the sample size. Moreover, they employed an assumption on the dependency structure in the VAR, thus limiting the sample correlation issues mentioned earlier. The work of (Kock & Callot, 2015) studied regularized Lasso-based estimator while allowing for problem dimensionality to grow with sample size, utilizing suitable martingale concentration inequalities to analyze dependency structure. (Loh & Wainwright, 2011) considered  $L_1$  VAR estimation for first order models ( $d = 1$ ) assuming  $\|A_1\|_2 < 1$ , and the analysis was not extended to the general case of  $d > 1$ . In recent work, (Basu & Michailidis, 2015) considered a VAR Lasso estimator and established the sample complexity and error bounds by building on the prior work of (Loh & Wainwright, 2011). Their approach exploits the spectral properties of a general VAR model of order  $d$ , providing insights on the dependency structure of the VAR process. However, in line with the existing literature, the analysis was tailored to the special case of  $L_1$  norm, thus limiting its generality.

**Our Contributions:** Compared to the existing literature, our results are substantially more general since the results and analysis apply to *any* norm  $R(\cdot)$ . One may wonder—given the popularity of  $L_1$  norm, why worry about other norms? Over the past decade, considerable effort has been

devoted to generalize  $L_1$  norm based results to other norms (Negahban et al., 2012; Chatterjee et al., 2012; Banerjee et al., 2014; Figueiredo & Nowak, 2014). Our work obviates the need for a similar exercise for VAR models. Further, some of these norms have found key niche in specific application areas e.g., (Zhou et al., 2012; Yang et al., 2015). From a technical perspective, one may also wonder—once we have the result for  $L_1$  norm, why should not the extension to other norms be straightforward? A key technical aspect of the estimation error analysis boils down to getting sharp concentration bounds for  $R^*(Z^T \epsilon)$ , where  $R^*(\cdot)$  is the dual norm of  $R(\cdot)$ ,  $Z$  is the design matrix, and  $\epsilon$  is the noise (Banerjee et al., 2014). For the special case of  $L_1$ , the dual norm is  $L_\infty$ , and one can use *union bound* to get the required concentration. In fact, this is exactly how the analysis in (Basu & Michailidis, 2015) was done. For general norms, the union bound is inapplicable. Our analysis is based on a considerably more powerful tool, *generic chaining* (Talagrand, 2006), yielding an analysis applicable to any norm, and producing results in terms of geometric properties, such as Gaussian widths (Ledoux & Talagrand, 2013), of sets related to the norm. Results for specific norms can then be obtained by plugging in suitable bounds on the Gaussian widths (Chandrasekaran et al., 2012; Chen & Banerjee, 2015). We illustrate the idea by recovering known bounds for Lasso and Group Lasso, and obtaining new results for Sparse Group Lasso and OWL norms. Finally, in terms of the core technical analysis, the application of generic chaining to the VAR estimation setting is not straightforward. In the VAR setting, generic chaining has to consider a stochastic process derived from sub-exponential martingale difference sequence (MDS). We first generalize the classical Azuma-Hoeffding inequality applicable to sub-Gaussian MDSs to get an Azuma-Bernstein inequality for sub-exponential MDSs. Further, we use suitable representations of Talagrand’s  $\gamma$ -functions (Talagrand, 2006) in the context of generic chaining to obtain bounds on  $R^*(Z^T \epsilon)$  in terms of the Gaussian width  $w(\Omega_R)$  of the unit norm ball  $\Omega_R = \{u \in \mathbb{R}^{dp} | R(u) \leq 1\}$ . Our estimation error bounds in the VAR setting are *exactly of the same order* as Lasso-type models in the i.i.d. setting implying, surprisingly, that the strong temporal dependency in the VAR setting has no adverse effect on the estimation.

The rest of the paper is organized as follows. In Section 2 we present the estimation problem for the structured VAR model. The main results on estimation guarantees are established in Section 3. We present experimental results in Section 4 and conclude in Section 5.

## 2. Structured VAR Models

In this section we formulate structured VAR estimation problem and discuss its properties, which are essential in

characterizing sample complexity and error bounds.

## 2.1. Regularized Estimator

To estimate the parameters of the VAR model, we transform the model in (2) into the form suitable for regularized estimator (1). Let  $(x_0, x_1, \dots, x_T)$  denote the  $T + 1$  samples generated by the stable VAR model in (2), then stacking them together we obtain

$$\begin{bmatrix} x_d^T \\ x_{d+1}^T \\ \vdots \\ x_T^T \end{bmatrix} = \begin{bmatrix} x_{d-1}^T & x_{d-2}^T & \cdots & x_0^T \\ x_d^T & x_{d-1}^T & \cdots & x_1^T \\ \vdots & \vdots & \ddots & \vdots \\ x_{T-1}^T & x_{T-2}^T & \cdots & x_{T-d}^T \end{bmatrix} \begin{bmatrix} A_1^T \\ A_2^T \\ \vdots \\ A_d^T \end{bmatrix} + \begin{bmatrix} \epsilon_d^T \\ \epsilon_{d+1}^T \\ \vdots \\ \epsilon_T^T \end{bmatrix}$$

which can also be compactly written as

$$Y = XB + E, \quad (3)$$

where  $Y \in \mathbb{R}^{N \times p}$ ,  $X \in \mathbb{R}^{N \times dp}$ ,  $B \in \mathbb{R}^{dp \times p}$ , and  $E \in \mathbb{R}^{N \times p}$  for  $N = T - d + 1$ . Vectorizing (column-wise) each matrix in (3), we get

$$\begin{aligned} \text{vec}(Y) &= (I_{p \times p} \otimes X) \text{vec}(B) + \text{vec}(E) \\ \mathbf{y} &= Z\boldsymbol{\beta} + \boldsymbol{\epsilon}, \end{aligned}$$

where  $\mathbf{y} \in \mathbb{R}^{Np}$ ,  $Z = (I_{p \times p} \otimes X) \in \mathbb{R}^{Np \times dp^2}$ ,  $\boldsymbol{\beta} \in \mathbb{R}^{dp^2}$ ,  $\boldsymbol{\epsilon} \in \mathbb{R}^{Np}$ , and  $\otimes$  is the Kronecker product. The covariance matrix of the noise  $\boldsymbol{\epsilon}$  is now  $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \Sigma \otimes I_{N \times N}$ . Consequently, the regularized estimator takes the form

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{dp^2}}{\text{argmin}} \frac{1}{N} \|\mathbf{y} - Z\boldsymbol{\beta}\|_2^2 + \lambda_N R(\boldsymbol{\beta}), \quad (4)$$

where  $R(\boldsymbol{\beta})$  can be any vector norm, separable along the rows of matrices  $A_k$ . Specifically, if we denote  $\boldsymbol{\beta} = [\beta_1^T \dots \beta_p^T]^T$  and  $A_k(i, :)$  as the row of matrix  $A_k$  for  $k = 1, \dots, d$ , then our assumption is equivalent to

$$R(\boldsymbol{\beta}) = \sum_{i=1}^p R(\beta_i) = \sum_{i=1}^p R\left(\left[A_1(i, :)^T \dots A_d(i, :)^T\right]^T\right). \quad (5)$$

To reduce clutter and without loss of generality, we assume the norm  $R(\cdot)$  to be the same for each row  $i$ . Since the analysis decouples across rows, it is straightforward to extend our analysis to the case when a different norm is used for each row of  $A_k$ , e.g.,  $L_1$  for row one,  $L_2$  for row two,  $K$ -support norm (Argyriou et al., 2012) for row three, etc. Observe that within a row, the norm need not be decomposable across columns.

The main difference between the estimation problem in (1) and the formulation in (4) is the strong dependence between the samples  $(x_0, x_1, \dots, x_T)$ , violating the i.i.d. assumption on the data  $\{(y_i, z_i), i = 1, \dots, Np\}$ . In particular, this leads to the correlations between the rows and columns

of matrix  $X$  (and consequently of  $Z$ ). To deal with such dependencies, following (Basu & Michailidis, 2015), we utilize the spectral representation of the autocovariance of VAR models to control the dependencies in matrix  $X$ .

## 2.2. Stability of VAR Model

Since VAR models are (linear) dynamical systems, for the analysis we need to establish conditions under which the VAR model (2) is stable, i.e., the time-series process does not diverge over time. For understanding stability, it is convenient to rewrite VAR model of order  $d$  in (2) as an equivalent VAR model of order 1

$$\begin{bmatrix} x_t \\ x_{t-1} \\ \vdots \\ x_{t-(d-1)} \end{bmatrix} = \underbrace{\begin{bmatrix} A_1 & A_2 & \cdots & A_{d-1} & A_d \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ \vdots \\ x_{t-d} \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (6)$$

where  $\mathbf{A} \in \mathbb{R}^{dp \times dp}$ . Therefore, VAR process is stable if all the eigenvalues of  $\mathbf{A}$  satisfy  $\det(\lambda I_{dp \times dp} - \mathbf{A}) = 0$  for  $\lambda \in \mathbb{C}$ ,  $|\lambda| < 1$ . Equivalently, if expressed in terms of original parameters  $A_k$ , stability is satisfied if  $\det(I - \sum_{k=1}^d A_k \frac{1}{\lambda^k}) = 0$  (see Section 2.1.1 of (Lutkepohl, 2007) and Section 1.3 of the supplement for more details).

## 2.3. Properties of Design Matrix $X$

In what follows, we analyze the covariance structure of matrix  $X$  in (3) using spectral properties of VAR model. The results will then be used in establishing the high probability bounds for the estimation guarantees in problem (4).

Define any row of  $X$  as  $X_{i,:} \in \mathbb{R}^{dp}$ ,  $1 \leq i \leq N$ . Since we assumed that  $\epsilon_t \sim \mathcal{N}(0, \Sigma)$ , it follows that each row is distributed as  $X_{i,:} \sim \mathcal{N}(0, C_X)$ , where the covariance matrix  $C_X \in \mathbb{R}^{dp \times dp}$  is the same for all  $i$

$$C_X = \begin{bmatrix} \Gamma(0) & \Gamma(1) & \cdots & \Gamma(d-1) \\ \Gamma(1)^T & \Gamma(0) & \cdots & \Gamma(d-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma(d-1)^T & \Gamma(d-2)^T & \cdots & \Gamma(0) \end{bmatrix}, \quad (7)$$

where  $\Gamma(h) = \mathbb{E}(x_t x_{t+h}^T) \in \mathbb{R}^{p \times p}$ . It turns out that since  $C_X$  is a block-Toeplitz matrix, its eigenvalues can be bounded as (see (Gutierrez-Gutierrez & Crespo, 2011))

$$\inf_{\substack{1 \leq j \leq p \\ \omega \in [0, 2\pi]}} \Lambda_j[\rho(\omega)] \leq \Lambda_k[C_X] \leq \sup_{\substack{1 \leq j \leq p \\ \omega \in [0, 2\pi]}} \Lambda_j[\rho(\omega)], \quad (8)$$

where  $\Lambda_k[\cdot]$  denotes the  $k$ -th eigenvalue of a matrix and for  $i = \sqrt{-1}$ ,  $\rho(\omega) = \sum_{h=-\infty}^{\infty} \Gamma(h) e^{-hi\omega}$ ,  $\omega \in [0, 2\pi]$ , is the spectral density, i.e., a Fourier transform of the autocovariance matrix  $\Gamma(h)$ . The advantage of utilizing spectral

density is that it has a closed form expression (see Section 9.4 of (Priestley, 1981))

$$\rho(\omega) = \left( I - \sum_{k=1}^d A_k e^{-ki\omega} \right)^{-1} \Sigma \left[ \left( I - \sum_{k=1}^d A_k e^{-ki\omega} \right)^{-1} \right]^*,$$

where  $*$  denotes a Hermitian of a matrix. Therefore, from (8) we can establish the following lower bound

$$\Lambda_{\min}[C_X] \geq \Lambda_{\min}(\Sigma) / \Lambda_{\max}(\mathcal{A}) = \mathcal{L}, \quad (9)$$

where we defined  $\Lambda_{\max}(\mathcal{A}) = \max_{\omega \in [0, 2\pi]} \Lambda_{\max}(\mathcal{A}(\omega))$  for

$$\mathcal{A}(\omega) = \left( I - \sum_{k=1}^d A_k^T e^{ki\omega} \right) \left( I - \sum_{k=1}^d A_k e^{-ki\omega} \right). \quad (10)$$

In establishing high probability bounds we will also need information about a vector  $q = Xa \in \mathbb{R}^N$  for any  $a \in \mathbb{R}^{dp}$ ,  $\|a\|_2 = 1$ . Since each element  $X_{i,:}^T a \sim \mathcal{N}(0, a^T C_X a)$ , it follows that  $q \sim \mathcal{N}(0, Q_a)$  with a covariance matrix  $Q_a \in \mathbb{R}^{N \times N}$ . It can be shown (see Section 2.1.2 of the supplement) that  $Q_a$  can be written as

$$Q_a = (I \otimes a^T) C_U (I \otimes a), \quad (11)$$

where  $C_U = \mathbb{E}(\mathcal{U}\mathcal{U}^T)$  for  $\mathcal{U} = [X_{1,:}^T \dots X_{N,:}^T]^T \in \mathbb{R}^{Ndp}$  which is obtained from matrix  $X$  by stacking all the rows in a single vector, i.e.,  $\mathcal{U} = \text{vec}(X^T)$ . In order to bound eigenvalues of  $C_U$  (and consequently of  $Q_a$ ), observe that  $\mathcal{U}$  can be viewed as a vector obtained by stacking  $N$  outputs from VAR model in (6). Similarly as in (8), if we denote the spectral density of the VAR process in (6) as  $\rho_X(\omega) = \sum_{h=-\infty}^{\infty} \Gamma_X(h) e^{-hi\omega}$ ,  $\omega \in [0, 2\pi]$ , where  $\Gamma_X(h) = \mathbb{E}[X_{j,:} X_{j+h,:}^T] \in \mathbb{R}^{dp \times dp}$ , then we can write

$$\inf_{\substack{1 \leq l \leq dp \\ \omega \in [0, 2\pi]}} \Lambda_l[\rho_X(\omega)] \leq \Lambda_k[C_U] \leq \sup_{\substack{1 \leq l \leq dp \\ \omega \in [0, 2\pi]}} \Lambda_l[\rho_X(\omega)].$$

The closed form expression of spectral density is

$$\rho_X(\omega) = (I - \mathbf{A}e^{-i\omega})^{-1} \Sigma_{\mathcal{E}} \left[ (I - \mathbf{A}e^{-i\omega})^{-1} \right]^*,$$

where  $\Sigma_{\mathcal{E}}$  is the covariance matrix of a noise vector and  $\mathbf{A}$  are as defined in expression (6). Thus, an upper bound on  $C_U$  can be obtained as  $\Lambda_{\max}[C_U] \leq \frac{\Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\mathcal{A})}$ , where we defined  $\Lambda_{\min}(\mathcal{A}) = \min_{\omega \in [0, 2\pi]} \Lambda_{\min}(\mathcal{A}(\omega))$  for

$$\mathcal{A}(\omega) = (I - \mathbf{A}^T e^{i\omega}) (I - \mathbf{A} e^{-i\omega}). \quad (12)$$

Referring back to covariance matrix  $Q_a$  in (11), we get

$$\Lambda_{\max}[Q_a] \leq \Lambda_{\max}(\Sigma) / \Lambda_{\min}(\mathcal{A}) = \mathcal{M}. \quad (13)$$

We note that for a general VAR model, there might not exist closed-form expressions for  $\Lambda_{\max}(\mathcal{A})$  and  $\Lambda_{\min}(\mathcal{A})$ . However, for some special cases there are results establishing the bounds on these quantities (e.g., see Proposition 2.2 in (Basu & Michailidis, 2015)).

### 3. Regularized Estimation Guarantees

Denote by  $\Delta = \hat{\beta} - \beta^*$  the error between the solution of optimization problem (4) and  $\beta^*$ , the true value of the parameter. The focus of our work is to determine conditions under which the optimization problem in (4) has guarantees on the accuracy of the obtained solution, i.e., the error term is bounded:  $\|\Delta\|_2 \leq \delta$  for some known  $\delta$ . To establish such conditions, we utilize the framework of (Banerjee et al., 2014). Specifically, estimation error analysis is based on the following known results adapted to our settings. The first one characterizes the restricted error set  $\Omega_E$ , where the error  $\Delta$  belongs.

**Lemma 3.1** *Assume that*

$$\lambda_N \geq rR^* \left[ \frac{1}{N} Z^T \epsilon \right], \quad (14)$$

for some constant  $r > 1$ , where  $R^* \left[ \frac{1}{N} Z^T \epsilon \right]$  is a dual form of the vector norm  $R(\cdot)$ , which is defined as  $R^* \left[ \frac{1}{N} Z^T \epsilon \right] = \sup_{R(U) \leq 1} \left\langle \frac{1}{N} Z^T \epsilon, U \right\rangle$ , for  $U \in \mathbb{R}^{dp^2}$ , where  $U = [u_1^T, u_2^T, \dots, u_p^T]^T$  and  $u_i \in \mathbb{R}^{dp}$ . Then the error vector  $\|\Delta\|_2$  belongs to the set

$$\Omega_E = \left\{ \Delta \in \mathbb{R}^{dp^2} \mid R(\beta^* + \Delta) \leq R(\beta^*) + \frac{1}{r} R(\Delta) \right\}. \quad (15)$$

The second condition in (Banerjee et al., 2014) establishes the upper bound on the estimation error.

**Lemma 3.2** *Assume that the restricted eigenvalue (RE) condition holds*

$$\frac{\|Z\Delta\|_2}{\|\Delta\|_2} \geq \sqrt{\kappa N}, \quad (16)$$

for  $\Delta \in \text{cone}(\Omega_E)$  and some constant  $\kappa > 0$ , where  $\text{cone}(\Omega_E)$  is a cone of the error set, then

$$\|\Delta\|_2 \leq \frac{1+r}{r} \frac{\lambda_N}{\kappa} \Psi(\text{cone}(\Omega_E)), \quad (17)$$

where  $\Psi(\text{cone}(\Omega_E))$  is a norm compatibility constant, defined as  $\Psi(\text{cone}(\Omega_E)) = \sup_{U \in \text{cone}(\Omega_E)} \frac{R(U)}{\|U\|_2}$ .

Note that the above error bound is deterministic, i.e., if (14) and (16) hold, then the error satisfies the upper bound in (17). However, the results are defined in terms of the quantities, involving  $Z$  and  $\epsilon$ , which are random. Therefore, in the following we establish high probability bounds on the regularization parameter in (14) and RE condition in (16).

### 3.1. High Probability Bounds

In this Section we present the main results of our work, followed by the discussion on their properties and illustrating some special cases based on popular Lasso and Group Lasso regularization norms. In Section 3.4 we will present the main ideas of our proof technique, with all the details delegated to the supplement.

To establish lower bound on the regularization parameter  $\lambda_N$ , we derive an upper bound on  $R^*[\frac{1}{N}Z^T\epsilon] \leq \alpha$ , for some  $\alpha > 0$ , which will establish the required relationship  $\lambda_N \geq \alpha \geq R^*[\frac{1}{N}Z^T\epsilon]$ .

**Theorem 3.3** *Let  $\Omega_R = \{u \in \mathbb{R}^{dp} | R(u) \leq 1\}$ , and define  $w(\Omega_R) = \mathbb{E}[\sup_{u \in \Omega_R} \langle g, u \rangle]$  to be a Gaussian width of set  $\Omega_R$  for  $g \sim \mathcal{N}(0, I)$ . For any  $\epsilon_1 > 0$  and  $\epsilon_2 > 0$  with probability at least  $1 - c \exp(-\min(\epsilon_2^2, \epsilon_1) + \log(p))$  we can establish that*

$$R^* \left[ \frac{1}{N} Z^T \epsilon \right] \leq \left( c_2(1+\epsilon_2) \frac{w(\Omega_R)}{\sqrt{N}} + c_1(1+\epsilon_1) \frac{w^2(\Omega_R)}{N^2} \right)$$

where  $c, c_1$  and  $c_2$  are positive constants.

To establish restricted eigenvalue condition, we will show that  $\inf_{\Delta \in \text{cone}(\Omega_E)} \frac{\|(I_{p \times p} \otimes X)\Delta\|_2}{\|\Delta\|_2} \geq \nu$ , for some  $\nu > 0$  and then set  $\sqrt{\kappa N} = \nu$ .

**Theorem 3.4** *Let  $\Theta = \text{cone}(\Omega_{E_j}) \cap S^{dp-1}$ , where  $S^{dp-1}$  is a unit sphere. The error set  $\Omega_{E_j}$  is defined as  $\Omega_{E_j} = \left\{ \Delta_j \in \mathbb{R}^{dp} \mid R(\beta_j^* + \Delta_j) \leq R(\beta_j^*) + \frac{1}{r}R(\Delta_j) \right\}$ , for  $r > 1$ ,  $j = 1, \dots, p$ , and  $\Delta = [\Delta_1^T, \dots, \Delta_p^T]^T$ , for  $\Delta_j$  is of size  $dp \times 1$ , and  $\beta^* = [\beta_1^{*T} \dots \beta_p^{*T}]^T$ , for  $\beta_j^* \in \mathbb{R}^{dp}$ . The set  $\Omega_{E_j}$  is a part of the decomposition in  $\Omega_E = \Omega_{E_1} \times \dots \times \Omega_{E_p}$  due to the assumption on the row-wise separability of norm  $R(\cdot)$  in (5). Also define  $w(\Theta) = \mathbb{E}[\sup_{u \in \Theta} \langle g, u \rangle]$  to be a Gaussian width of set  $\Theta$  for  $g \sim \mathcal{N}(0, I)$  and  $u \in \mathbb{R}^{dp}$ . Then with probability at least  $1 - c_1 \exp(-c_2\eta^2 + \log(p))$ , for any  $\eta > 0$ ,  $\inf_{\Delta \in \text{cone}(\Omega_E)} \frac{\|(I_{p \times p} \otimes X)\Delta\|_2}{\|\Delta\|_2} \geq \nu$ , where  $\nu = \sqrt{N\mathcal{L}} - 2\sqrt{M} - cw(\Theta) - \eta$  and  $c, c_1, c_2$  are positive constants, and  $\mathcal{L}, M$  are defined in (9) and (13).*

### 3.2. Discussion

From Theorem 3.4, we can choose  $\eta = \frac{1}{2}\sqrt{N\mathcal{L}}$  and set  $\sqrt{\kappa N} = \sqrt{N\mathcal{L}} - 2\sqrt{M} - cw(\Theta) - \eta$  and since  $\sqrt{\kappa N} > 0$  must be satisfied, we can establish a lower bound on the number of samples  $N$ :  $\sqrt{N} > \frac{2\sqrt{M} + cw(\Theta)}{\sqrt{\mathcal{L}/2}} = \mathcal{O}(w(\Theta))$ . Examining this bound and using (9) and (13), we can conclude that the number of samples needed to satisfy the restricted eigenvalue condition is smaller if  $\Lambda_{\min}(\mathcal{A})$  and

$\Lambda_{\min}(\Sigma)$  are larger and  $\Lambda_{\max}(\mathcal{A})$  and  $\Lambda_{\max}(\Sigma)$  are smaller. In turn, this means that matrices  $\mathcal{A}$  and  $\mathcal{A}$  in (10) and (12) must be well conditioned and the VAR process is stable, with eigenvalues well inside the unit circle (see Section 2.2). Alternatively, we can also understand the bound on  $N$  as showing that large values of  $M$  and small values of  $\mathcal{L}$  indicate stronger dependency in the data, thus requiring more samples for the RE conditions to hold with high probability.

Analyzing Theorems 3.3 and 3.4 we can interpret the established results as follows. As the size and dimensionality  $N, p$  and  $d$  of the problem increase, we emphasize the scale of the results and use the order notations to denote the constants. Select a number of samples at least  $N \geq \mathcal{O}(w^2(\Theta))$  and let the regularization parameter satisfy  $\lambda_N \geq \mathcal{O}\left(\frac{w(\Omega_R)}{\sqrt{N}} + \frac{w^2(\Omega_R)}{N^2}\right)$ . With high probability then the restricted eigenvalue condition  $\frac{\|Z\Delta\|_2}{\|\Delta\|_2} \geq \sqrt{\kappa N}$  for  $\Delta \in \text{cone}(\Omega_{E_j})$  holds, so that  $\kappa = \mathcal{O}(1)$  is a positive constant. Moreover, the norm of the estimation error in optimization problem (4) is bounded by  $\|\Delta\|_2 \leq \mathcal{O}\left(\frac{w(\Omega_R)}{\sqrt{N}} + \frac{w^2(\Omega_R)}{N^2}\right) \Psi(\text{cone}(\Omega_{E_j}))$ . Note that the norm compatibility constant  $\Psi(\text{cone}(\Omega_{E_j}))$  is assumed to be the same for all  $j = 1, \dots, p$ , which follows from our assumption in (5).

Consider now Theorem 3.3 and the bound on the regularization parameter  $\lambda_N \geq \mathcal{O}\left(\frac{w(\Omega_R)}{\sqrt{N}} + \frac{w^2(\Omega_R)}{N^2}\right)$ . As the dimensionality of the problem  $p$  and  $d$  grows and the number of samples  $N$  increases, the first term  $\frac{w(\Omega_R)}{\sqrt{N}}$  will dominate the second one  $\frac{w^2(\Omega_R)}{N^2}$ . This can be seen by computing  $N$  for which the two terms become equal  $\frac{w(\Omega_R)}{\sqrt{N}} = \frac{w^2(\Omega_R)}{N^2}$ , which happens at  $N = w^{\frac{2}{3}}(\Omega_R) < w(\Omega_R)$ . Therefore, we can rewrite our results as follows: once the restricted eigenvalue condition holds and  $\lambda_N \geq \mathcal{O}\left(\frac{w(\Omega_R)}{\sqrt{N}}\right)$ , the error norm is upper-bounded by  $\|\Delta\|_2 \leq \mathcal{O}\left(\frac{w(\Omega_R)}{\sqrt{N}}\right) \Psi(\text{cone}(\Omega_{E_j}))$ .

### 3.3. Special Cases

While the presented results are valid for any norm  $R(\cdot)$ , separable along the rows of  $A_k$ , it is instructive to specialize our analysis to a few popular regularization choices, such as  $L_1$  and Group Lasso, Sparse Group Lasso and OWL norms.

#### 3.3.1. LASSO

To establish results for  $L_1$  norm, we assume that the parameter  $\beta^*$  is  $s$ -sparse, which in our case is meant to represent the largest number of non-zero elements in any  $\beta_i$ ,  $i = 1, \dots, p$ , i.e., the combined  $i$ -th rows of each  $A_k$ ,

$k = 1, \dots, d$ . Since  $L_1$  is decomposable, it can be shown that  $\Psi(\text{cone}(\Omega_{E_j})) \leq 4\sqrt{s}$ . Next, since  $\Omega_R = \{u \in \mathbb{R}^{dp} | R(u) \leq 1\}$ , then using Lemma 3 in (Banerjee et al., 2014) and Gaussian width results in (Chandrasekaran et al., 2012), we can establish that  $w(\Omega_R) \leq \mathcal{O}(\sqrt{\log(dp)})$ . Therefore, based on Theorem 4.3 and the discussion at the end of Section 3.2, the bound on the regularization parameter takes the form  $\lambda_N \geq \mathcal{O}(\sqrt{\log(dp)/N})$ . Hence, the estimation error is bounded by  $\|\Delta\|_2 \leq \mathcal{O}(\sqrt{s \log(dp)/N})$  as long as  $N > \mathcal{O}(\log(dp))$ .

### 3.3.2. GROUP LASSO

To establish results for Group norm, we assume that for each  $i = 1, \dots, p$ , the vector  $\beta_i \in \mathbb{R}^{dp}$  can be partitioned into a set of  $K$  disjoint groups,  $G = \{G_1, \dots, G_K\}$ , with the size of the largest group  $m = \max_k |G_k|$ . Group Lasso norm is defined as  $\|\beta\|_{\text{GL}} = \sum_{k=1}^K \|\beta_{G_k}\|_2$ . We assume that the parameter  $\beta^*$  is  $s_G$ -group-sparse, which means that the largest number of non-zero groups in any  $\beta_i$ ,  $i = 1, \dots, p$  is  $s_G$ . Since Group norm is decomposable, as was established in (Negahban et al., 2012), it can be shown that  $\Psi(\text{cone}(\Omega_{E_j})) \leq 4\sqrt{s_G}$ . Similarly as in the Lasso case, using Lemma 3 in (Banerjee et al., 2014), we get  $w(\Omega_{R_{\text{GL}}}) \leq \mathcal{O}(\sqrt{m + \log(K)})$ . The bound on the  $\lambda_N$  takes the form  $\lambda_N \geq \mathcal{O}(\sqrt{(m + \log(K))/N})$ . Combining these derivations, we obtain the bound  $\|\Delta\|_2 \leq \mathcal{O}(\sqrt{s_G(m + \log(K))/N})$  for  $N > \mathcal{O}(m + \log(K))$ .

### 3.3.3. SPARSE GROUP LASSO

Similarly as in Section 3.3.2, we assume that we have  $K$  disjoint groups of size at most  $m$ . The Sparse Group Lasso norm enforces sparsity not only across but also within the groups and is defined as  $\|\beta\|_{\text{SGL}} = \alpha\|\beta\|_1 + (1 - \alpha)\sum_{k=1}^K \|\beta_{G_k}\|_2$ , where  $\alpha \in [0, 1]$  is a parameter which regulates a convex combination of Lasso and Group Lasso penalties. Note that since  $\|\beta\|_2 \leq \|\beta\|_1$ , it follows that  $\|\beta\|_{\text{GL}} \leq \|\beta\|_{\text{SGL}}$ . As a result, for  $\beta \in \Omega_{R_{\text{SGL}}} \Rightarrow \beta \in \Omega_{R_{\text{GL}}}$ , so that  $\Omega_{R_{\text{SGL}}} \subseteq \Omega_{R_{\text{GL}}}$  and thus  $w(\Omega_{R_{\text{SGL}}}) \leq w(\Omega_{R_{\text{GL}}}) \leq \mathcal{O}(\sqrt{m + \log(K)})$ , according to Section 3.3.2. Assuming  $\beta^*$  is  $s$ -sparse and  $s_G$ -group-sparse and noting that the norm is decomposable, we get  $\Psi(\text{cone}(\Omega_{E_j})) \leq 4(\alpha\sqrt{s} + (1 - \alpha)\sqrt{s_G})$ . Consequently, the error bound is  $\|\Delta\|_2 \leq \mathcal{O}(\sqrt{(\alpha s + (1 - \alpha)s_G)(m + \log(K))/N})$ .

### 3.3.4. OWL NORM

Ordered weighted  $L_1$  norm is a recently introduced regularizer and is defined as  $\|\beta\|_{\text{owl}} = \sum_{i=1}^{dp} c_i |\beta|_{(i)}$ , where  $c_1 \geq \dots \geq c_{dp} \geq 0$  is a predefined non-increasing se-

quence of weights and  $|\beta|_{(1)} \geq \dots \geq |\beta|_{(dp)}$  is the sequence of absolute values of  $\beta$ , ranked in decreasing order. In (Chen & Banerjee, 2015) it was shown that  $w(\Omega_R) \leq \mathcal{O}(\sqrt{\log(dp)/\bar{c}})$ , where  $\bar{c}$  is the average of  $c_1, \dots, c_{dp}$  and the norm compatibility constant is  $\Psi(\text{cone}(\Omega_{E_j})) \leq 2c_1^2 \sqrt{s/\bar{c}}$ . Therefore, based on Theorem 4.3, we get  $\lambda_N \geq \mathcal{O}(\sqrt{\log(dp)/(\bar{c}N)})$  and the estimation error is bounded by  $\|\Delta\|_2 \leq \mathcal{O}(\frac{2c_1}{\bar{c}} \sqrt{s \log(dp)/(\bar{c}N)})$ .

We note that the bound obtained for Lasso and Group Lasso is similar to the bound obtained in (Song & Bickel, 2011; Basu & Michailidis, 2015; Kock & Callot, 2015). Moreover, this result is also similar to the works, which dealt with independent observations, e.g., (Bickel et al., 2009; Negahban et al., 2012), with the difference being the constants, reflecting correlation between the samples, as we discussed in Section 3.2. The explicit bound for Sparse Group Lasso and OWL is a *novel* aspect of our work for the non-asymptotic recovery guarantees for the VAR estimation problem with norm regularization, being just a simple consequence from our more general framework.

## 3.4. Proof Sketch

In this Section we outline the steps of the proof for Theorem 3.3 and 3.4, all the details can be found in Sections 2.2 and 2.3 of the supplement.

### 3.4.1. BOUND ON REGULARIZATION PARAMETER

Recall that our objective is to establish for some  $\alpha > 0$  a probabilistic statement that  $\lambda_N \geq \alpha \geq R^*[\frac{1}{N}Z^T \epsilon] = \sup_{R(U) \leq 1} \langle \frac{1}{N}Z^T \epsilon, U \rangle$ , where  $U = [u_1^T, \dots, u_p^T]^T \in \mathbb{R}^{dp^2}$  for  $u_j \in \mathbb{R}^{dp}$  and  $\epsilon = \text{vec}(E)$  for  $E$  in (3). We denote  $E_{:,j} \in \mathbb{R}^N$  as a column of noise matrix  $E$  and note that since  $Z = I_{p \times p} \otimes X$ , then using the row-wise separability assumption in (5) we can split the overall probability statement into  $p$  parts, which are easier to work with. Thus, our objective would be to establish

$$\mathbb{P} \left[ \sup_{R(u_j) \leq r_j} \frac{1}{N} \langle X^T E_{:,j}, u_j \rangle \leq \alpha_j \right] \geq \pi_j, \quad (18)$$

for  $j = 1, \dots, p$ , where  $\sum_{j=1}^p \alpha_j = \alpha$  and  $\sum_{j=1}^p r_j = 1$ .

The overall strategy is to first show that the random variable  $\frac{1}{N} \langle X^T E_{:,j}, u_j \rangle$  has sub-exponential tails. Based on the generic chaining argument, we then use Theorem 1.2.7 in (Talagrand, 2006) and bound

$\mathbb{E} \left[ \sup_{R(u_j) \leq r_j} \frac{1}{N} \langle X^T E_{:,j}, u_j \rangle \right]$ . Finally, using Theorem 1.2.9 in (Talagrand, 2006) we establish the high probability bound on concentration of  $\sup_{R(u_j) \leq r_j} \frac{1}{N} \langle X^T E_{:,j}, u_j \rangle$  around its mean, i.e., derive the bound in (18).

We note that the main difficulty of working with the term  $\langle X^T E_{:,j}, u_j \rangle$  is the complicated dependency between  $X$  and  $E_{:,j}$ , which is due to the VAR generation process in (3). However, if we write  $\langle X^T E_{:,j}, u_j \rangle = \sum_{i=1}^N E_{i,j}(X_{i,:}, u_j) = \sum_{i=1}^N m_i$ , where  $m_i = E_{i,j}(X_{i,:}, u_j)$  and we can interpret this as a summation over martingale difference sequence (Lutkepohl, 2007). This can be easily proven by showing  $\mathbb{E}(m_i | m_1, \dots, m_{i-1}) = 0$ . The latter is true since in  $m_i = E_{i,j}(X_{i,:}, u_j)$  the terms  $E_{i,j}$  and  $X_{i,:}, u_j$  are independent since  $\epsilon_{d+i}$  is independent from  $x_{d-k+i}$  for  $0 \leq i \leq T-d$  and  $1 \leq k \leq d$  (see (2)).

To show that  $\sum_{i=1}^N E_{i,j}(X_{i,:}, u_j)$  has sub-exponential tails, recall that since  $\epsilon_t$  in (2) is Gaussian,  $E_{i,j}$  and  $X_{i,:}, u_j$  are independent Gaussian random variables, whose product has sub-exponential tails. Moreover, the sum over sub-exponential martingale difference sequence can be shown to be itself sub-exponential using (Shamir, 2011), based on Bernstein-type inequality (Vershynin, 2010).

### 3.4.2. RESTRICTED EIGENVALUE CONDITION

To show  $\frac{\|(I_{p \times p} \otimes X)\Delta\|_2}{\|\Delta\|_2} \geq 0$  for all  $\Delta \in \text{cone}(\Omega_E)$ , similarly as before, we split the problem into  $p$  parts by using row-wise separability assumption of the norm in (5). In particular, denote  $\Delta = [\Delta_1^T, \dots, \Delta_p^T]^T$ , where  $\Delta_j$  is  $dp \times 1$ , then we can represent the original set  $\Omega_E$  as a Cartesian product of subsets  $\Omega_{E_j}$ , i.e.,  $\Omega_E = \Omega_{E_1} \times \dots \times \Omega_{E_p}$ , implying that  $\text{cone}(\Omega_E) = \text{cone}(\Omega_{E_1}) \times \dots \times \text{cone}(\Omega_{E_p})$ . Therefore, our objective would be to establish

$\mathbb{P} \left[ \inf_{u_j \in \Theta_j} \|Xu_j\|_2 \geq \nu_j \right] \geq \pi_j$ , for  $j = 1, \dots, p$ , where  $\Theta = \text{cone}(\Omega_{E_j}) \cap S^{dp-1}$  and we defined  $u_j = \frac{\Delta_j}{\|\Delta_j\|_2}$ , since it will be easier to operate with unit-norm vectors (we drop index  $j$ , to reduce clutter).

The overall strategy is to first show that  $\|Xu\|_2 - \mathbb{E}(\|Xu\|_2)$  is sub-Gaussian. Then, using generic chaining argument in (Talagrand, 2006), specifically Theorem 2.1.5, we bound  $\mathbb{E} \left( \inf_{u \in \Theta} \|Xu\|_2 \right)$ . Finally, based on Lemma 2.1.3 in (Talagrand, 2006) we establish the concentration inequality on  $\inf_{u \in \Theta} \|Xu\|_2$  around its mean.

## 4. Experimental Results

In this Section we present the experiments on simulated and real data to demonstrate the obtained theoretical results. In particular, for  $L_1$  and Group  $L_1$ , we investigate how error norm  $\|\Delta\|_2$  and regularization parameter  $\lambda_N$  scale as the problem size  $p$  and  $N$  change. Moreover, using real flight data we also compare the performance of Sparse Group  $L_1$ , OWL and ridge regularizers. Additional simulation and experimental results are included in the supplement.

### 4.1. Synthetic Data

To evaluate the estimation problem with  $L_1$  norm, we simulated a first-order VAR process for different values of  $p \in [10, 600]$ ,  $s \in [4, 260]$ , and  $N \in [10, 5000]$ . Regularization parameter was varied in the range  $\lambda_N \in (0, \lambda_{\max})$ , where  $\lambda_{\max}$  is the largest parameter, for which estimation problem (4) produces a zero solution. All the results are shown after averaging across 50 runs.

The results for Lasso are shown in the top row of Fig. 1. In particular, in Fig. 1.a we show  $\|\Delta\|_2$  for different  $p$  and  $N$  for fixed  $\lambda_N$ . When  $N$  is small, the estimation error is large and the results cannot be trusted. However, once  $N \geq \mathcal{O}(w^2(\Theta))$ , the RE condition in Lemma 3.2 is satisfied and we see a fast decrease of errors for all  $p$ 's. In Fig. 1.b we plot  $\|\Delta\|_2$  against rescaled sample size  $\frac{N}{s \log(pd)}$ . The errors are now closely aligned, confirming results of Sec. 3.3.1, i.e.,  $\|\Delta\|_2 \leq \mathcal{O} \left( \sqrt{(s \log(pd))/N} \right)$ .

Finally, in Figs. 1.c and 1.d we show the dependence of optimal  $\lambda_N$  (for fixed  $N$  and  $p$ , we picked  $\lambda_N$  achieving the smallest estimation error) on  $N$  and  $p$ . It can be seen that as  $p$  increases,  $\lambda_N$  grows (for fixed  $N$ ) at the rate similar to  $\sqrt{\log p}$ . On the other hand, as  $N$  increases, the selected  $\lambda_N$  decreases (for fixed  $p$ ) at the rate similar to  $1/\sqrt{N}$ .

For Group Lasso the sparsity in rows of  $A_1$  was generated in groups, whose number varied as  $K \in [2, 60]$ . We set the largest number of non-zero groups in any row as  $s_G \in [2, 22]$ . Results are shown in the bottom row of Fig. 1, which have similar flavor as in Lasso case. The difference can be seen in Fig. 1.f, where a close alignment of errors occurs when  $N$  is now scaled as  $\frac{N}{s_G(m+\log(K))}$ . Moreover, the selected regularization parameter  $\lambda$  increases with the number of groups  $K$  and decreases with  $N$ .

### 4.2. Real Data

We have also performed evaluation tests on real data to compare the accuracy of the VAR estimation using various penalized formulations based on five norms:  $L_1$ , OWL, Group, Sparse Group and Ridge (square of  $L_2$ ). Although  $\|\cdot\|_2^2$  is not a norm, we included its results for reference purposes as it is frequently used in practice. In terms of data, we used the NASA flight dataset from (nas), consisting of over 100,000 flights, each having a record of about 250 parameters, sampled at 1 Hz. For our test, we selected 300 flights and picked 31 parameters most suitable for the prediction task and focused on the landing part of the trajectory (duration approximately 15 minutes). For each flight we separately fitted a first-order VAR model using five approaches and performed 5-fold cross validation to select  $\lambda$ , achieving smallest prediction error. For Sparse Group we set  $\alpha = 0.5$ , while for OWL the weights  $c_1, \dots, c_p$  were set as a monotonically decreasing sequence. Table 1 shows

## Estimating Structured VAR

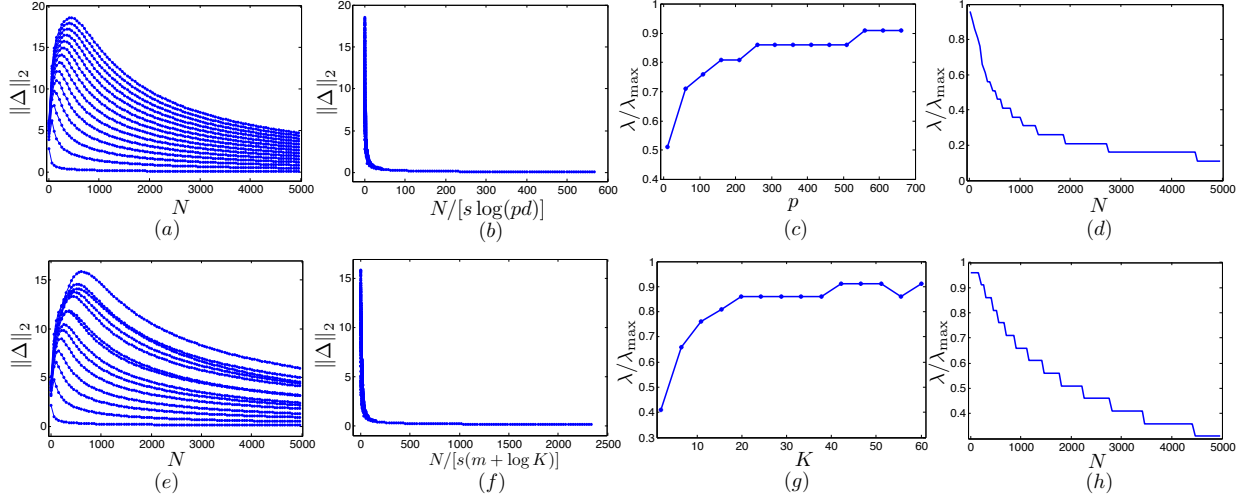


Figure 1. Results for estimating parameters of a stable first order sparse VAR (top row) and group sparse VAR (bottom row). Problem dimensions:  $p \in [10, 600]$ ,  $N \in [10, 5000]$ ,  $\frac{\lambda_N}{\lambda_{max}} \in [0, 1]$ ,  $K \in [2, 60]$  and  $d = 1$ . Figures (a) and (e) show dependency of errors on sample size for different  $p$ ; in Figure (b) the  $N$  is scaled by  $(s \log p)$  and plotted against  $\|\Delta\|_2$  to show that errors scale as  $(s \log p)/N$ ; in (f) the graph is similar to (b) but for group sparse VAR; in (c) and (g) we show dependency of  $\lambda_N$  on  $p$  (or number of groups  $K$  in (g)) for fixed sample size  $N$ ; finally, Figures (d) and (h) display the dependency of  $\lambda_N$  on  $N$  for fixed  $p$ .

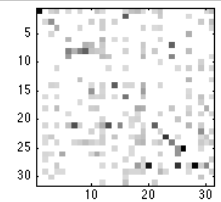
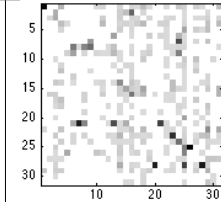
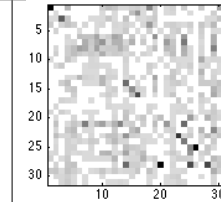
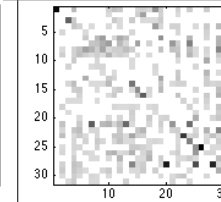
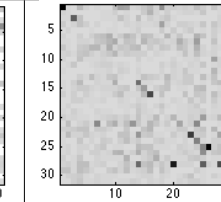
Lasso	OWL	Group Lasso	Sparse Group Lasso	Ridge
32.3(6.5)	32.2(6.6)	32.7(6.5)	32.2(6.4)	33.5(6.1)
32.7(7.9)	44.5(15.6)	75.3(8.4)	38.4(9.6)	99.9(0.2)
				

Table 1. Mean squared error (row 2) of the five methods used in fitting VAR model, evaluated on aviation dataset (MSE is computed using one-step-ahead prediction errors). Row 3 shows the average number of non-zeros (as a percentage of total number of elements) in the VAR matrix. The last row shows a typical sparsity pattern in  $A_1$  for each method (darker dots - stronger dependencies, lighter dots - weaker dependencies). The values in parenthesis denote one standard deviation after averaging the results over 300 flights.

the results after averaging across 300 flights.

From the table we can see that the considered problem exhibits a sparse structure since all the methods detected similar patterns in matrix  $A_1$ . In particular, the analysis of such patterns revealed a meaningful relationship among the flight parameters (darker dots), e.g., normal acceleration had high dependency on vertical speed and angle-of-attack, the altitude had mainly dependency with fuel quantity, vertical speed with aircraft nose pitch angle, etc. The results also showed that the sparse regularization helps in recovering more accurate and parsimonious models as is evident by comparing performance of Ridge regression with other methods. Moreover, while all the four Lasso-based approaches performed similar to each other, their sparsity levels were different, with Lasso producing the sparsest solutions. As was also expected, Group Lasso had larger number of non-zeros since it did not enforce sparsity within the groups, as compared to the sparse version of this norm.

## 5. Conclusions

In this work we present a set of results for characterizing non-asymptotic estimation error in estimating structured vector autoregressive models. The analysis holds for *any* norms, separable along the rows of parameter matrices. Our analysis is general as it is expressed in terms of Gaussian widths, a geometric measure of size of suitable sets, and includes as special cases many of the existing results focused on structured sparsity in VAR models.

## Acknowledgements

The research was supported by NSF grants IIS-1447566, IIS-1447574, IIS-1422557, CCF-1451986, CNS- 1314560, IIS-0953274, IIS-1029711, NASA grant NNX12AQ39A, and gifts from Adobe, IBM, and Yahoo.



## References

- NASA Aviation Safety Dataset. Available at <https://c3.nasa.gov/dashlink/projects/85/>.
- Argyriou, A., Foygel, R., and Srebro, N. Sparse prediction with the  $k$ -support norm. In *Advances in Neural Information Processing Systems*, pp. 1457–1465, 2012.
- Banerjee, A., Chen, S., Fazayeli, F., and Sivakumar, V. Estimation with norm regularization. In *Advances in Neural Information Processing Systems*, pp. 1556–1564, 2014.
- Basu, S. and Michailidis, G. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 08 2015.
- Bickel, P. J., Ritov, Y., and Tsybakov, A. B. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 08 2009.
- Bogdan, M., Berg, E., Su, W., and Candes, E. Statistical estimation and testing via the sorted  $l_1$  norm. *arXiv preprint arXiv:1310.1969*, 2013.
- Candes, E. and Tao, T. The Dantzig selector: statistical estimation when  $p$  is much larger than  $n$ . *The Annals of Statistics*, pp. 2313–2351, 2007.
- Chandrasekaran, V., Recht, B., Parrilo, P., and Willsky, A. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.
- Chatterjee, S., Steinhäuser, K., Banerjee, A., Chatterjee, S., and Ganguly, G. Sparse group Lasso: Consistency and climate applications. In *Proceedings of International Conference on Data Mining*, pp. 47–58, 2012.
- Chen, S. and Banerjee, A. Structured estimation with atomic norms: General bounds and applications. In *Advances in Neural Information Processing Systems*, pp. 2890–2898, 2015.
- Figueiredo, M. and Nowak, R. Sparse estimation with strongly correlated variables using ordered weighted  $l_1$  regularization. *arXiv preprint arXiv:1409.4005*, 2014.
- Gutierrez-Gutierrez, J. and Crespo, P. M. Block Toeplitz matrices: asymptotic results and applications. *Foundations and Trends in Communications and Information Theory*, 8(3):179–257, 2011.
- Han, F. and Liu, H. A direct estimation of high dimensional stationary vector autoregressions. *ArXiv e-prints, arXiv:1307.0293*, 2013.
- Jacob, L., Obozinski, G., and Vert, J.-P. Group lasso with overlap and graph lasso. In *Proceedings of the International conference on machine learning*, pp. 433–440, 2009.
- Kock, A. B. and Callot, L. Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344, 2015.
- Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science, 2013.
- Ljung, L. *System identification: theory for the user*. Springer, 1998.
- Loh, P.-L. and Wainwright, M. J. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pp. 2726–2734, 2011.
- Lutkepohl, H. *New introduction to multiple time series analysis*. Springer, 2007.
- Meinshausen, N. and Yu, B. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pp. 246–270, 2009.
- Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of  $M$ -estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 11 2012.
- Priestley, M. B. *Spectral analysis and time series*. Academic press, 1981.
- Raskutti, G., Wainwright, M. J., and Yu, B. Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.
- Rudelson, M. and Zhou, S. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, 2013.
- Shamir, O. A variant of Azuma’s inequality for martingales with subgaussian tails. *arXiv preprint arXiv:1110.2392*, 2011.
- Song, S. and Bickel, P. J. Large vector auto regressions. *ArXiv e-prints, arXiv:1106.3915*, 2011.
- Talagrand, M. *The generic chaining: upper and lower bounds of stochastic processes*. Springer, 2006.
- Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, pp. 267–288, 1996.

- Tsay, R. S. *Analysis of financial time series*, volume 543. 2005.
- Valdes-Sosa, P. A., Sanchez-Bornot, J. M., Lage-Castellanos, A., Vega-Hernandez, M., Bosch-Bayard, J., Melie-Garcia, L., and Canales-Rodriguez, E. Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical Transactions of the Royal Society*, 360(1457):969–981, 2005.
- Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *ArXiv e-prints*, arXiv:1011.3027, 2010.
- Wainwright, M. Sharp thresholds for high-dimensional and noisy sparsity recovery using constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.
- Yang, T., Wang, J., Sun, Q., Hibar, D. P., Jahanshad, N., Liu, L., Wang, Y., Zhan, L., Thompson, P., and Ye, J. Detecting genetic risk factors for Alzheimer’s disease in whole genome sequence data via Lasso screening. In *IEEE International Symposium on Biomedical Imaging*, 2015.
- Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society.*, 68(1):49–67, 2006.
- Zhao, P. and Yu, B. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7: 2541–2563, 2006.
- Zhou, J., Liu, J., Narayan, V. A., and Ye, J. Modeling disease progression via fused sparse group Lasso. In *Proceedings of International conference on Knowledge discovery and data mining*, pp. 1095–1103, 2012.
- Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society.*, 67(2):301–320, 2005.