Supplement: Estimating Structured Vector Autoregressive Model

Igor Melnyk and Arindam Banerjee

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

1 Problem Formulation

Consider a vector autoregressive (VAR) model of order d:

$$x_t = A_1 x_{t-1} + \ldots + A_d x_{t-d} + \epsilon_t, \quad t = 0, \pm 1, \pm 2, \ldots, \tag{1}$$

where $x_t \in \mathbb{R}^p$ is a random vector, $A_i \in \mathbb{R}^{p \times p}$, $i = 1, \ldots, d$ are fixed coefficient matrices and ϵ_t is a vector of zero-mean white noise, i.e., $E(\epsilon_t) = 0$, $E(\epsilon_t \epsilon_t^T) = \Sigma$ and $E(\epsilon_t \epsilon_{t+h}^T) = 0$, for $h \neq 0$. We assume that the noise covariance matrix Σ is positive definite with bounded largest eigenvalue, i.e., $\Lambda_{\min}(\Sigma) > 0$ and $\Lambda_{\max}(\Sigma) < \infty$.

The above formulation in (1) can be written compactly as a VAR model of order 1:

$$X_t = \mathbf{A} X_{t-1} + \mathcal{E}_t, \tag{2}$$

where $X_t = \begin{bmatrix} x_t \\ x_{t-1} \\ \vdots \\ x_{t-(d-1)} \end{bmatrix} \in \mathbb{R}^{dp}$, $\mathbf{A} = \begin{bmatrix} A_1 & A_2 & \ldots & A_{d-1} & A_d \\ I & 0 & \ldots & 0 & 0 \\ 0 & I & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & I & 0 \end{bmatrix} \in \mathbb{R}^{dp \times dp}$, and $\mathcal{E}_t = \begin{bmatrix} \epsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{dp}$, where I is the identity matrix $I \in \mathbb{R}^{dp \times dp}$. The covariance matrix of the noise term $\mathcal{E}$ is now

$$\Sigma_{\mathcal{E}} = \begin{bmatrix} \Sigma & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix}.$$

Note that the first order VAR model in (2) can also be represented in the moving average form [5]

$$X_t = \sum_{i=0}^{\infty} \mathbf{A}^i \mathcal{E}_{t-i}, \tag{3}$$

where X_t is expressed in terms of past and present error vectors.

1.1 VAR Estimation

In the following, we rewrite the VAR model of order d in (1)

$$x_t = A_1 x_{t-1} + \ldots + A_d x_{t-d} + \epsilon_t, \quad t = 0, \pm 1, \pm 2, \ldots$$

in the form suitable for estimation. We assume that we have a realization of $T + 1$ samples $(x_0, x_1, \ldots, x_T)$. Transposing the above form, we get

$$x_t^T = x_{t-1}^T A_1^T + \ldots + x_{t-d}^T A_d^T + \epsilon_t^T.$$

Then, stacking all the samples together, we obtain

$$\begin{bmatrix} x_d^T \\ x_{d+1}^T \\ \vdots \\ x_{T-1}^T \\ x_T^T \end{bmatrix} = \begin{bmatrix} x_{d-1}^T & x_{d-2}^T & \dots & x_0^T \\ x_d^T & x_{d-1}^T & \dots & x_1^T \\ \vdots & \vdots & \ddots & \vdots \\ x_{T-2}^T & x_{T-3}^T & \dots & x_{T-d-1}^T \\ x_{T-1}^T & x_{T-2}^T & \dots & x_{T-d}^T \end{bmatrix} \begin{bmatrix} A_1^T \\ A_2^T \\ \\ \vdots \\ \\ A_d^T \end{bmatrix} + \begin{bmatrix} \epsilon_d^T \\ \epsilon_{d+1}^T \\ \vdots \\ \epsilon_{T-1}^T \\ \epsilon_T^T \end{bmatrix},$$

which we can compactly write as

$$Y = XB + E, \tag{4}$$

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 each matrix in (4), we get (utilizing the fact that $\text{vec}(AB) = (I \otimes A)\text{vec}(B)$):

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

where $\mathbf{y} \in \mathbb{R}^{Np}$, $Z \in \mathbb{R}^{Np \times dp^2}$, $\boldsymbol{\beta} \in \mathbb{R}^{dp^2}$, and $\boldsymbol{\epsilon} \in \mathbb{R}^{Np}$. Note that the covariance matrix of the noise $\boldsymbol{\epsilon}$ is now $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \Sigma \otimes I_{N\times N}$.

In this work we consider the problem of estimating parameter $\boldsymbol{\beta}$ from the data $(x_0, x_1, \dots, x_T)$ generated by stable VAR model of order d using the following form of the estimator:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{dp^2}}{\text{argmin}} \frac{1}{N}||\mathbf{y} - Z\boldsymbol{\beta}||_2^2 + \lambda_N R(\boldsymbol{\beta}) \tag{5}$$

where $R(\boldsymbol{\beta})$ is any vector norm and λ_N is a regularization parameter. The only assumption we make about $R(\boldsymbol{\beta})$ is that it is decomposable along the columns of B. Denote $\boldsymbol{\beta} = [\beta_1^T \beta_2^T \dots \beta_p^T]^T$, where $\beta_i \in \mathbb{R}^{dp}$. Also let $B(:, i)$ denote the column of matrix B and $A_k(i, :)$ as the row of matrix A_k for $k = 1, \dots, d$, then we can write

$$\begin{aligned} R(\boldsymbol{\beta}) &= \sum_{i=1}^{p} R\big(\beta_i\big) \\ &= \sum_{i=1}^{p} R\Big(B(:, i)\Big) \\ &= \sum_{i=1}^{p} R\Big(\Big[A_1(i, :)^T A_2(i, :)^T \dots A_d(i, :)^T\Big]^T\Big). \end{aligned} \tag{6}$$

Note that in the above we also assumed, for simplicity and without the loss of generality, that for each $i = 1, \dots, p$, the norm $R(\cdot)$ is the same. It is straightforward to extend our framework to the case when for each i a different norm is used.

1.2 Matrix X Notations

To simplify derivations, below we define notations for various parts of matrix X

$$X = \begin{bmatrix} x_{d-1}^T & x_{d-2}^T & \dots & x_0^T \\ x_d^T & x_{d-1}^T & \dots & x_1^T \\ \vdots & \vdots & \ddots & \vdots \\ x_{T-2}^T & x_{T-3}^T & \dots & x_{T-d-1}^T \\ x_{T-1}^T & x_{T-2}^T & \dots & x_{T-d}^T \end{bmatrix}.$$

Row. Each row of matrix X is denoted as

$$X_{i,:} = \begin{bmatrix} x_{T-i}^T \\ x_{T-1-i}^T \\ \vdots \\ x_{T-d-(i-1)}^T \end{bmatrix} \in \mathbb{R}^{dp}, \tag{7}$$

for $i = 1, \ldots, N$. In cases, when the specific row index i is irrelevant, we use a notation $\mathsf{X} = X_{i,:}$, for all i.
All Rows. All the rows of matrix X, stacked in a single vector, are denoted as

$$\mathcal{U} = \begin{bmatrix} X_{1,:} \\ X_{2,:} \\ \vdots \\ X_{N,:} \end{bmatrix} \in \mathbb{R}^{Ndp}. \tag{8}$$

Column. Each column of matrix X is denoted as

$$X_{:,j} = \begin{bmatrix} x_{d-k}(:,l) \\ x_{d-k+1}(:,l) \\ \vdots \\ x_{T-k}(:,l) \end{bmatrix} \in \mathbb{R}^{N},$$

where index $j = k + l$, for $1 \leq k \leq d$ and $1 \leq l \leq p$, so that $j = 1, \ldots, dp$.
All Columns. All the columns of matrix X, stacked in a single vector, are denoted as

$$\mathcal{V} = \begin{bmatrix} X_{:,1} \\ X_{:,2} \\ \vdots \\ X_{:,dp} \end{bmatrix} \in \mathbb{R}^{Ndp}.$$

Block-Column. Matrix X can be viewed as a concatenation of d block-columns. Each block-column, reshaped into a vector, is denoted as

$$\mathcal{Y}_k = \begin{bmatrix} x_{d-k} \\ x_{d-k+1} \\ \vdots \\ x_{T-k} \end{bmatrix} \in \mathbb{R}^{Np}, \tag{9}$$

for $k = 1, \ldots, d$. In cases, when the specific index k is irrelevant, we use a notation $\mathsf{Y} = \mathcal{Y}_k$, for all k.
All Block-Columns. All the reshaped block-columns of matrix X, stacked in a single vector, are denoted as

$$\mathcal{W} = \begin{bmatrix} \mathcal{Y}_1 \\ \mathcal{Y}_2 \\ \vdots \\ \mathcal{Y}_d \end{bmatrix} \in \mathbb{R}^{Ndp}. \tag{10}$$

1.3 Stability of VAR Model

The formulation (2) represents a stable VAR model if all the eigenvalues of $\mathbf{A}$ are smaller than 1, i.e., eigenvalues of $\mathbf{A}$ must satisfy $\det(\lambda I_{dp\times dp} - \mathbf{A}) = 0$ for $\lambda \in \mathbb{C}$, $|\lambda| < 1$, $|\lambda| \neq 0$. Specifically, write

$$\lambda I_{dp\times dp} - \mathbf{A} = \begin{bmatrix} I\lambda & 0 & \dots & 0 & 0 \\ 0 & I\lambda & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & I\lambda \end{bmatrix} - \begin{bmatrix} A_1 & A_2 & \dots & A_{d-1} & A_d \\ I & 0 & \dots & 0 & 0 \\ 0 & I & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & I & 0 \end{bmatrix}$$
$$= \begin{bmatrix} I\lambda - A_1 & -A_2 & \dots & -A_{d-1} & -A_d \\ -I & I\lambda & \dots & 0 & 0 \\ 0 & -I & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -I & I\lambda \end{bmatrix}.$$

Now multiply last (d-th) block-column by $\frac{1}{\lambda}$ and add to $(d-1)$-st block-column. Next, multiply the result in $(d-1)$-st block-column by $\frac{1}{\lambda}$ and add to $(d-2)$-nd block-column. Continuing in this manner, we will arrive at

$$Q = \begin{bmatrix} \lambda I_{p\times p} - A_1 - \frac{1}{\lambda}A_2 - \ldots - \frac{1}{\lambda^{d-1}}A_d & M \\ 0 & \lambda I_{p(d-1)\times p(d-1)} \end{bmatrix},$$

where matrix $M \in \mathbb{R}^{p\times p(d-1)}$ denotes the result of some of the column operations. Since such column operations leave the matrix determinant unchanged, we have

$$\begin{aligned} \det(\lambda I_{dp\times dp} - \mathbf{A}) = \det(Q) &= \det(\lambda I_{p\times p} - A_1 - \frac{1}{\lambda}A_2 - \ldots - \frac{1}{\lambda^{d-1}}A_d) \cdot \det(\lambda I_{p(d-1)\times p(d-1)}) \\ &= \det(I_{p\times p} - \frac{1}{\lambda}A_1 - \frac{1}{\lambda^2}A_2 - \ldots - \frac{1}{\lambda^d}A_d) \cdot \lambda^{pd}. \end{aligned}$$

Therefore, stability of VAR model in (2) requires $\det(I - \sum_{k=1}^{d} A_k \frac{1}{\lambda^k}) = 0$ to be satisfied for $|\lambda| < 1$, $|\lambda| \neq 0$. Equivalently, $\det(I - \sum_{k=1}^{d} A_k z^k) = 0$ must be satisfied for $z \in \mathbb{C}$, $|z| > 1$, or $\det(I - \sum_{k=1}^{d} A_k z^k) \neq 0$ must hold for $|z| \leq 1$.

1.4 Autocovariance of VAR Model

In this section we consider autocovariance matrix of VAR model written in different forms as well as establish bounds on the eigenvalues of these matrices.

1.4.1 VAR Model for x_t

The autocovariance matrix of the original VAR process of order d in (1) is defined as $\Gamma(h) = \mathbb{E}[x_t x_{t+h}^T]$. Fourier transform of autocovariance matrix is called spectral density and is denoted as (for $i = \sqrt{-1}$)

$$\gamma(\omega) = \sum_{h=-\infty}^{\infty} \Gamma(h) e^{-hi\omega}, \quad \omega \in [0, 2\pi]. \tag{11}$$

Inverse Fourier transform of the spectral density gives back the autocovariance matrix:

$$\Gamma(h) = \frac{1}{2\pi} \int_0^{2\pi} \gamma(\omega) e^{hi\omega} d\omega, \quad h \in 0, \pm 1, \pm 2, \ldots \tag{12}$$

For our VAR model in (1), the spectral density has a closed form expression [7]

$$\gamma(\omega) = \left(I - \sum_{k=1}^{d} A_k e^{-ki\omega}\right)^{-1} \Sigma \left[\left(I - \sum_{k=1}^{d} A_k e^{-ki\omega}\right)^{-1}\right]^* \in \mathbb{R}^{p\times p}, \tag{13}$$

where $*$ is the Hermitian of a matrix.

Let $V = [x_1^T, x_2^T, \ldots, x_K^T]^T$ be a vector composed from the output of the VAR process during K steps, then

$$C_V = \mathbb{E}(VV^T) = \begin{bmatrix} \Gamma(0) & \Gamma(1) & \ldots & \Gamma(K-1) \\ \Gamma(1)^T & \Gamma(0) & \ldots & \Gamma(K-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma(K-1)^T & \Gamma(K-2)^T & \ldots & \Gamma(0) \end{bmatrix} \in \mathbb{R}^{Kp \times Kp}. \tag{14}$$

In this work we will be interested in the bounds on the eigenvalues of C_V. Note that C_V is a block-Toeplitz matrix and so we can use the following property [3]

$$\inf_{\substack{1 \le j \le p \\ \omega \in [0,2\pi]}} \Lambda_j[\gamma(\omega)] \le \Lambda_k[C_V] \le \sup_{\substack{1 \le j \le p \\ \omega \in [0,2\pi]}} \Lambda_j[\gamma(\omega)], \quad \text{for } 1 \le k \le Kp. \tag{15}$$

Using (13), we can compute the lower bound. For this we use the following relationships: for any M, $||M||_2 = \sqrt{\Lambda_{\max}(M^T M)}$, and if M is symmetric, $||M||_2 = \Lambda_{\max}(M)$. Similarly, for any nonsingular M, $||M^{-1}||_2 = \frac{1}{\sqrt{\Lambda_{\min}(M^T M)}}$, and if M is symmetric, $||M^{-1}||_2 = \frac{1}{\Lambda_{\min}(M)}$. Since $\gamma(\omega)$ is symmetric, we have

$$\begin{aligned} \Lambda_{\max}[\gamma(\omega)] &= \left|\left| \left(I - \sum_{k=1}^{d} A_k e^{-ki\omega} \right)^{-1} \Sigma \left[\left(I - \sum_{k=1}^{d} A_k e^{-ki\omega} \right)^{-1} \right]^* \right|\right|_2 \\ &\le \left|\left| \left(I - \sum_{k=1}^{d} A_k e^{-ki\omega} \right)^{-1} \right|\right|_2^2 ||\Sigma||_2 \\ &\le \frac{\Lambda_{\max}(\Sigma)}{\Lambda_{\min}\left[\left(I - \sum_{k=1}^{d} A_k^T e^{ki\omega} \right) \left(I - \sum_{k=1}^{d} A_k e^{-ki\omega} \right) \right]} \end{aligned} \tag{16}$$

and the upper bound

$$\begin{aligned} \Lambda_{\min}[\gamma(\omega)] &= \left[\left|\left| \left\{ \left(I - \sum_{k=1}^{d} A_k e^{-ki\omega} \right)^{-1} \Sigma \left[\left(I - \sum_{k=1}^{d} A_k e^{-ki\omega} \right)^{-1} \right]^* \right\}^{-1} \right|\right|_2 \right]^{-1} \\ &\ge \left[\left|\left| I - \sum_{k=1}^{d} A_k e^{-ki\omega} \right|\right|_2^2 ||\Sigma^{-1}||_2 \right]^{-1} \\ &\ge \frac{\Lambda_{\min}(\Sigma)}{\Lambda_{\max}\left[\left(I - \sum_{k=1}^{d} A_k^T e^{ki\omega} \right) \left(I - \sum_{k=1}^{d} A_k e^{-ki\omega} \right) \right]}. \end{aligned} \tag{17}$$

Therefore, the C_V has the following bounds on its eigenvalues

$$\frac{\Lambda_{\min}(\Sigma)}{\Lambda_{\max}\left[\left(I - \sum_{k=1}^{d} A_k^T e^{ki\omega} \right) \left(I - \sum_{k=1}^{d} A_k e^{-ki\omega} \right) \right]} \le \Lambda_k[C_V] \le \frac{\Lambda_{\max}(\Sigma)}{\Lambda_{\min}\left[\left(I - \sum_{k=1}^{d} A_k^T e^{ki\omega} \right) \left(I - \sum_{k=1}^{d} A_k e^{-ki\omega} \right) \right]},$$

for $1 \le k \le Kp,$ and $\omega \in [0, 2\pi]$.

Denoting $\Lambda_{\min}(\mathcal{A}) = \Lambda_{\min}\left[\left(I - \sum_{k=1}^{d} A_k^T e^{ki\omega} \right) \left(I - \sum_{k=1}^{d} A_k e^{-ki\omega} \right) \right]$ for $\omega \in [0, 2\pi]$ and similarly $\Lambda_{\max}(\mathcal{A}) = \Lambda_{\max}\left[\left(I - \sum_{k=1}^{d} A_k^T e^{ki\omega} \right) \left(I - \sum_{k=1}^{d} A_k e^{-ki\omega} \right) \right]$ for $\omega \in [0, 2\pi]$, we can compactly write the above as

$$\frac{\Lambda_{\min}(\Sigma)}{\Lambda_{\max}(\mathcal{A})} \le \Lambda_k[C_V] \le \frac{\Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\mathcal{A})}, \tag{18}$$

for $1 \le k \le Kp$.

1.4.2 VAR Model for $X_{j,:}$

In this section we consider the VAR model of order 1 in (2). Note that this is the same form as the model obtained from the rows of X (see (7)), i.e.,

$$\begin{bmatrix} x_{d-i+1} \\ x_{d-i} \\ \vdots \\ x_i \end{bmatrix} = \begin{bmatrix} A_1 & A_2 & \dots & A_{d-1} & A_d \\ I & 0 & \dots & 0 & 0 \\ 0 & I & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & I & 0 \end{bmatrix} \begin{bmatrix} x_{d-i} \\ x_{d-i-1} \\ \vdots \\ x_{i-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{d-i+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Written in a compact form, the above expression takes the form

$$X_{j,:} = \mathbf{A} X_{j-1,:} + \mathcal{E}_j, \quad \text{for } j = 1, \dots, N,$$

which can be thought to be the transformations of the form

$$X_{1,:} = \begin{bmatrix} x_{d-1} \\ x_{d-2} \\ \vdots \\ x_0 \end{bmatrix} \quad \to \quad X_{2,:} = \begin{bmatrix} x_d \\ x_{d-1} \\ \vdots \\ x_1 \end{bmatrix} \quad \to \cdots \to \quad X_{N,:} = \begin{bmatrix} x_{N+d-2} \\ x_{N+d-3} \\ \vdots \\ x_{N-1} \end{bmatrix}.$$

Let

$$\mathcal{U} = \begin{bmatrix} X_{1,:} \\ \vdots \\ X_{N,:} \end{bmatrix} \in \mathbb{R}^{Ndp}, \tag{19}$$

be a vector composed from the output of the above VAR model during N steps. Then $C_{\mathcal{U}} \in \mathbb{R}^{Ndp \times Ndp}$ is the covariance matrix of vector $\mathcal{U}$

$$C_{\mathcal{U}} = \mathbb{E}(\mathcal{U}\mathcal{U}^T) = \mathbb{E} \begin{bmatrix} X_{1,:} \\ \vdots \\ X_{N,:} \end{bmatrix} \begin{bmatrix} X_{1,:}^T \dots X_{N,:}^T \end{bmatrix} = \begin{bmatrix} \mathbb{E}[X_{1,:}X_{1,:}^T] & \mathbb{E}[X_{1,:}X_{2,:}^T] & \dots & \mathbb{E}[X_{1,:}X_{N,:}^T] \\ \mathbb{E}[X_{2,:}X_{1,:}^T] & \mathbb{E}[X_{2,:}X_{2,:}^T] & \dots & \mathbb{E}[X_{2,:}X_{N,:}^T] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[X_{N,:}X_{1,:}^T] & \mathbb{E}[X_{N,:}X_{2,:}^T] & \dots & \mathbb{E}[X_{N,:}X_{N,:}^T] \end{bmatrix}. \tag{20}$$

To establish the bounds on the eigenvalues of $C_{\mathcal{U}}$, we denote the spectral density of the corresponding VAR process as

$$\gamma_{\mathsf{X}}(\omega) = \sum_{h=-\infty}^{\infty} \Gamma_{\mathsf{X}}(h) e^{-hi\omega}, \quad \omega \in [0, 2\pi],$$

where $\Gamma_{\mathsf{X}}(h) = \mathbb{E}[X_{j,:}X_{j+h,:}^T]$. Since $C_{\mathcal{U}}$ is a block-Toeplitz matrix, we can employ the same relationship as we used in Section 1.4.1

$$\inf_{\substack{1 \le l \le dp \\ \omega \in [0,2\pi]}} \Lambda_l[\gamma_{\mathsf{X}}(\omega)] \le \Lambda_k[C_{\mathcal{U}}] \le \sup_{\substack{1 \le l \le dp \\ \omega \in [0,2\pi]}} \Lambda_l[\gamma_{\mathsf{X}}(\omega)], \quad \text{for } 1 \le k \le Ndp. \tag{21}$$

In the following we establish the closed form expression of spectral density γ_{X}. For this, we use moving average representation in (3) and write

$$\begin{aligned}
\gamma_{\mathsf{X}}(\omega) &= \sum_{h=-\infty}^{\infty} \Gamma_{\mathsf{X}}(h) e^{-hi\omega} \\
&= \sum_{h=-\infty}^{\infty} \mathbb{E}[X_{j,:} X_{j+h,:}^T] e^{-hi\omega} \quad \text{for any } j \\
&= \sum_{h=-\infty}^{\infty} \mathbb{E}\left[\sum_{k=0}^{\infty} \mathbf{A}^k E_{j-k,:} \left(\sum_{s=0}^{\infty} \mathbf{A}^s E_{j+h-s,:} \right)^T \right] e^{-hi\omega} \\
&= \sum_{h=-\infty}^{\infty} \mathbb{E}\left[\sum_{k=0}^{\infty} \mathbf{A}^k E_{j-k,:} \left(\sum_{s=0}^{\infty} \mathbf{A}^{s-h} E_{j-s,:} \right)^T \right] e^{-hi\omega} \\
&= \sum_{h=-\infty}^{\infty} \sum_{k=0}^{\infty} \mathbf{A}^k \Sigma_E \left(\mathbf{A}^{k-h} \right)^T e^{-hi\omega} \\
&= \sum_{h=-\infty}^{\infty} \sum_{k=0}^{\infty} \mathbf{A}^k \Sigma_E \left(\mathbf{A}^{k-h} \right)^T e^{-hi\omega + ki\omega - ki\omega} \\
&= \sum_{h=-\infty}^{\infty} \sum_{k=0}^{\infty} \mathbf{A}^k e^{-ki\omega} \Sigma_E \left(\mathbf{A}^{k-h} e^{-(k-h)i\omega} \right)^* \\
&= \sum_{k=0}^{\infty} \mathbf{A}^k e^{-ki\omega} \Sigma_E \sum_{r=0}^{\infty} \left(\mathbf{A}^r e^{-ri\omega} \right)^* \\
&= \left(I - \mathbf{A} e^{-i\omega} \right)^{-1} \Sigma_E \left[\left(I - \mathbf{A} e^{-i\omega} \right)^{-1} \right]^*, \qquad (22)
\end{aligned}$$

where we have used the fact that $\sum_{k=0}^{\infty} \mathbf{A}^k e^{-ki\omega} = \left(I - \mathbf{A} e^{-i\omega} \right)^{-1}$.

Now, using (21), (22), the results from Section 1.4.1 and the fact that the covariance matrix $\Sigma_{\mathcal{E}}$ has the form

$$\Sigma_{\mathcal{E}} = \begin{bmatrix} \Sigma & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix},$$

we can establish the following bounds

$$\frac{\Lambda_{\min}(\Sigma_{\mathcal{E}})}{\Lambda_{\max}\left[\left(I - \mathbf{A}^T e^{i\omega} \right) \left(I - \mathbf{A} e^{-i\omega} \right) \right]} \leq \Lambda_k[C_{\mathcal{U}}] \leq \frac{\Lambda_{\max}(\Sigma_{\mathcal{E}})}{\Lambda_{\min}\left[\left(I - \mathbf{A}^T e^{i\omega} \right) \left(I - \mathbf{A} e^{-i\omega} \right) \right]}.$$

Since $\Lambda_{\max}(\Sigma_{\mathcal{E}}) = \Lambda_{\max}(\Sigma)$, the upper bound becomes

$$\Lambda_{\max}[C_{\mathcal{U}}] \leq \frac{\Lambda_{\max}(\Sigma)}{\Lambda_{\min}\left[\left(I - \mathbf{A}^T e^{i\omega} \right) \left(I - \mathbf{A} e^{-i\omega} \right) \right]},$$

for $\omega \in [0, 2\pi]$. Denoting $\Lambda_{\min}(\boldsymbol{\mathcal{A}}) = \Lambda_{\min}\left[\left(I - \mathbf{A}^T e^{i\omega} \right) \left(I - \mathbf{A} e^{-i\omega} \right) \right]$ for $\omega \in [0, 2\pi]$, we can compactly write the above as

$$\Lambda_{\max}[C_{\mathcal{U}}] \leq \frac{\Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\boldsymbol{\mathcal{A}})}. \qquad (23)$$

2 Statistical Properties of VAR Estimator

Denote by $\Delta = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ the error between the solution of optimization problem (5) and $\boldsymbol{\beta}^*$, the true value of the parameter. The focus of our work is to determine conditions under which the estimation problem in (5) is consistent, i.e., the error term is bounded: $||\Delta||_2 \leq \delta$ for some known δ.

To establish such conditions, we utilize the framework of [2]. Specifically, if the following regularization parameter bound is satisfied

$$\lambda_N \geq cR^*[\frac{1}{N}Z^T\boldsymbol{\epsilon}],$$

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

$$\Omega_E = \left\{ \Delta \in \mathbb{R}^{dp^2} \middle| R(\boldsymbol{\beta}^* + \Delta) \leq R(\boldsymbol{\beta}^*) + \frac{1}{c}R(\Delta) \right\}.$$

Moreover, if the restricted eigenvalue condition holds

$$\frac{||Z\Delta||_2}{||\Delta||_2} \geq \sqrt{\kappa N},$$

for $\Delta \in \text{cone}(\Omega_E)$ and some constant $\kappa > 0$, where $\text{cone}(\Omega_E)$ is a cone of an error set, then the following bound on the norm of the estimation error can be established

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

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

In the derivations of the bounds we will be utilizing the following concentration inequality for a Lipschitz function of standard Gaussian random variable

Lemma 2.1 *Let $X \in \mathbb{R}^n$ be a vector of zero-mean, unit-variance Gaussian entries, i.e., $X \sim \mathcal{N}(0, I_{n\times n})$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz with constant L, which means that $|f(X) - f(Y)| \leq L||X - Y||_2, \forall X, Y \in \mathbb{R}^n$. Then for all $\tau > 0$*

$$\mathbb{P}\left[\left| f(X) - \mathbb{E}\big[f(X)\big] \right| > \tau \right] \leq 2\exp\left(-\frac{\tau^2}{2L^2}\right).$$

as well as the concentration inequality for ℓ_2 - norm of arbitrary Gaussian vector

Lemma 2.2 *Let $X \in \mathbb{R}^n$ be a vector of zero-mean Gaussian entries, i.e., $X \sim \mathcal{N}(0, Q_{n\times n})$. Then for all $\tau > 0$*

$$\mathbb{P}\left[\left| ||X||_2 - \sqrt{trace(Q)} \right| > \tau + 2\sqrt{||Q||_2} \right] \leq 2\exp\left(-\frac{\tau^2}{2||Q||_2}\right).$$

Concentration inequality for supreme of Gaussian processes

Lemma 2.3 *Let $\{X_t\}_{t\in T}$ be a Gaussian processes, then for all $\tau > 0$*

$$\mathbb{P}\left[\left| \sup_{t\in T} X_t - \mathbb{E}\sup_{t\in T} X_t \right| > \tau \right] \leq 2\exp\left(-\frac{\tau^2}{2\sup_{t\in T}\mathbb{E}(X_t^2)}\right).$$

.

Useful probability relationship.

Lemma 2.4 *Let $\mathcal{A}_i$, for $i = 1, \ldots, K$ be a set of probabilistic events. Then*

$$\mathbb{P}\Big[\mathcal{A}_1 \text{ and } \mathcal{A}_2 \text{ and } \ldots \text{ and } \mathcal{A}_K\Big] \geq \sum_{i=1}^{K} \mathbb{P}\big[\mathcal{A}_i\big] - (K-1)$$

Proof 2.5 *Using De Morgan's law, and denoting by $\overline{\mathcal{A}_i}$ the negation of event $\mathcal{A}_i$, we can write*

$$\begin{aligned}
1 - \mathbb{P}\Big[\mathcal{A}_1 \text{ and } \mathcal{A}_2 \text{ and } \ldots \text{ and } \mathcal{A}_K\Big] &= \mathbb{P}\Big[\overline{\mathcal{A}_1} \text{ or } \overline{\mathcal{A}_2} \text{ or } \ldots \text{ or } \overline{\mathcal{A}_K}\Big] \\
&\leq \mathbb{P}\Big[\overline{\mathcal{A}_1}\Big] + \mathbb{P}\Big[\overline{\mathcal{A}_2}\Big] + \ldots + \mathbb{P}\Big[\overline{\mathcal{A}_K}\Big] \\
&= 1 - \mathbb{P}\big[\mathcal{A}_1\big] + 1 - \mathbb{P}\big[\mathcal{A}_2\big] + \ldots + 1 - \mathbb{P}\big[\mathcal{A}_K\big] \\
&= K - \sum_{i=1}^{K} \mathbb{P}\big[\mathcal{A}_i\big]
\end{aligned}$$

where on the second line we used the union bound. Now rearranging the terms we get

$$\mathbb{P}\Big[\mathcal{A}_1 \text{ and } \mathcal{A}_2 \text{ and } \ldots \text{ and } \mathcal{A}_K\Big] \geq \sum_{i=1}^{K} \mathbb{P}\big[\mathcal{A}_i\big] - (K-1).$$

We will also utilize the notions of Gaussian width and covering net.

Definition 2.6 *For any set $\mathcal{S}$ and for a vector of independent zero-mean unit variance Gaussian variables $g \sim \mathcal{N}(0, I)$, the Gaussian width of the set is defined as*

$$w(\mathcal{S}) = \mathbb{E}_g[\sup_{u \in \mathcal{S}} \langle g, u \rangle]. \tag{24}$$

2.1 Gaussian Noise Model

In this work we assume that the distribution of the noise in VAR process

$$x_t = A_1 x_{t-1} + \ldots + A_d x_{t-d} + \epsilon_t, \quad t = 0, \pm 1, \pm 2, \ldots, \tag{25}$$

follows a Gaussian distribution, i.e., $\epsilon_t \sim \mathcal{N}(0, \Sigma)$. Moreover, we can conclude that the distribution of x_t is a zero-mean Gaussian, i.e., $x_t \sim \mathcal{N}(0, \Gamma(0))$, where $\Gamma(h) = \mathbb{E}(x_t x_{t+h}^T)$.

Now consider the noise and data matrices from the formulation (2)

$$E = \begin{bmatrix} \epsilon_d^T \\ \epsilon_{d+1}^T \\ \vdots \\ \epsilon_{T-1}^T \\ \epsilon_T^T \end{bmatrix}, \quad X = \begin{bmatrix} x_{d-1}^T & x_{d-2}^T & \cdots & x_0^T \\ x_d^T & x_{d-1}^T & \cdots & x_1^T \\ \vdots & \vdots & \ddots & \vdots \\ x_{T-2}^T & x_{T-3}^T & \cdots & x_{T-d-1}^T \\ x_{T-1}^T & x_{T-2}^T & \cdots & x_{T-d}^T \end{bmatrix}. \tag{26}$$

For our theoretical analysis we require information about the probability distribution of rows and columns of X, as well as columns of E. In the following sections we present the corresponding derivations.

2.1.1 Columns Distribution of Noise Matrix E

Each column of E in (26), denoted as $E_{:,j}$, is a Gaussian vector: $E_{:,j} \sim \mathcal{N}(0, C_{E_{:,j}})$, where $C_{E_{:,j}} \in \mathbb{R}^{N \times N}$, $C_{E_{:,j}} = \mathbb{E}(E_{:,j} E_{:,j}^T) = \Sigma_{j,j} I_{N \times N}$, which is a diagonal matrix.

In what follows, we compute $trace(C_{E_{:,j}})$ and $||C_{E_{:,j}}||_2$ for the covariance matrix $C_{E_{:,j}}$, needed in the future computations. It can be seen that the trace of $C_{E_{:,j}}$ is given by $trace(C_{E_{:,j}}) = N\Sigma_{j,j}$ and similarly we can establish the $||C_{E_{:,j}}||_2 = \Lambda_{\max}(C_{E_{:,j}}) = \Sigma_{j,j}$.

Note that we can write the following inequality $\Sigma_{j,j} \leq \Lambda_{\max}(\Sigma)$ for any $j = 1, \ldots, p$. This follows from Schur-Horn theorem [4], which states that for a symmetric matrix Σ, if we sort its diagonal elements and eigenvalues in non-decreasing order, i.e., $\Sigma_{j,j_1} \leq \ldots \leq \Sigma_{j,j_p}$ and $\Lambda_{j_1}(\Sigma) \leq \ldots \leq \Lambda_{j_p}(\Sigma)$, then

$$\sum_{i=1}^{k} \Sigma_{j,j_i} \geq \sum_{i=1}^{k} \Lambda_{j_i}(\Sigma), \quad \text{for } k = 1, \ldots, p$$

and it holds with equality when $k = p$. Since $\sum_{i=1}^{p-1} \Sigma_{j,j_i} \geq \sum_{i=1}^{p-1} \Lambda_{j_i}(\Sigma)$ and $\sum_{i=1}^{p} \Sigma_{j,j_i} = \sum_{i=1}^{p} \Lambda_{j_i}(\Sigma)$, it follows that $\Sigma_{j,j_p} \leq \Lambda_{j_p}(\Sigma)$. Therefore, $\Sigma_{j,j} \leq \Lambda_{\max}(\Sigma)$, for any $j = 1, \ldots, p$.

Consequently, we can establish the following bounds on trace and spectral norm of $C_{E_{:,j}}$ for any $j = 1, \ldots, p$

$$||C_{E_{:,j}}||_2 \leq \Lambda_{\max}(\Sigma), \tag{27}$$

and

$$trace(C_{E_{:,j}}) \leq N\Lambda_{\max}(\Sigma). \tag{28}$$

2.1.2 Rows Distribution of Data Matrix X

Each row of X, denoted as $X_{i,:} = [x_{T-i}^T, x_{T-1-i}^T, \ldots, x_{T-d-(i-1)}^T]^T \in \mathbb{R}^{dp}$, $1 \leq i \leq N$, is distributed as $X_{i,:} \sim \mathcal{N}(0, C_{\mathsf{X}})$, where the covariance matrix C_{X}, same for all i, is defined as

$$C_{\mathsf{X}} = \mathbb{E}(X_{i,:}X_{i,:}^T) = \begin{bmatrix} \Gamma(0) & \Gamma(1) & \ldots & \Gamma(d-1) \\ \Gamma(1)^T & \Gamma(0) & \ldots & \Gamma(d-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma(d-1)^T & \Gamma(d-2)^T & \ldots & \Gamma(0) \end{bmatrix} \in \mathbb{R}^{dp \times dp}, \tag{29}$$

where $\Gamma(h) = \mathbb{E}(x_t x_{t+h}^T)$. Note that, using results from Section 1.4.1 and specifically expression (18), we can establish the upper and lower bound on the eigenvalues of C_{X}

$$\Lambda_{\max}[C_{\mathsf{X}}] \leq \frac{\Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\mathcal{A})} \quad \text{and} \quad \Lambda_{\min}[C_{\mathsf{X}}] \geq \frac{\Lambda_{\min}(\Sigma)}{\Lambda_{\max}(\mathcal{A})}. \tag{30}$$

Now consider a vector $q = Xa \in \mathbb{R}^N$ for any $a \in \mathbb{R}^{dp}$. Since each element $X_{i,:}^T a \sim \mathcal{N}(0, a^T C_{\mathsf{X}} a)$, it follows that $q \sim \mathcal{N}(0, Q)$ with a covariance matrix $Q \in \mathbb{R}^{N \times N}$, which is defined as

$$\begin{aligned} Q = \mathbb{E}(qq^T) &= \mathbb{E} \begin{bmatrix} X_{1,:}^T a \\ \vdots \\ X_{N,:}^T a \end{bmatrix} \begin{bmatrix} a^T X_{1,:} \ldots a^T X_{N,:} \end{bmatrix} \\ &= \begin{bmatrix} a^T\mathbb{E}[X_{1,:}X_{1,:}^T]a & a^T\mathbb{E}[X_{1,:}X_{2,:}^T]a & \ldots & a^T\mathbb{E}[X_{1,:}X_{N,:}^T]a \\ a^T\mathbb{E}[X_{2,:}X_{1,:}^T]a & a^T\mathbb{E}[X_{2,:}X_{2,:}^T]a & \ldots & a^T\mathbb{E}[X_{2,:}X_{N,:}^T]a \\ \vdots & \vdots & \ddots & \vdots \\ a^T\mathbb{E}[X_{N,:}X_{1,:}^T]a & a^T\mathbb{E}[X_{N,:}X_{2,:}^T]a & \ldots & a^T\mathbb{E}[X_{N,:}X_{N,:}^T]a \end{bmatrix} \\ &= \begin{bmatrix} a^T & 0 & \ldots & 0 \\ 0 & a^T & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a^T \end{bmatrix} \begin{bmatrix} \mathbb{E}[X_{1,:}X_{1,:}^T] & \mathbb{E}[X_{1,:}X_{2,:}^T] & \ldots & \mathbb{E}[X_{1,:}X_{N,:}^T] \\ \mathbb{E}[X_{2,:}X_{1,:}^T] & \mathbb{E}[X_{2,:}X_{2,:}^T] & \ldots & \mathbb{E}[X_{2,:}X_{N,:}^T] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[X_{N,:}X_{1,:}^T] & \mathbb{E}[X_{N,:}X_{2,:}^T] & \ldots & \mathbb{E}[X_{N,:}X_{N,:}^T] \end{bmatrix} \begin{bmatrix} a & 0 & \ldots & 0 \\ 0 & a & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a \end{bmatrix} \\ &= (I_{N\times N} \otimes a^T) \begin{bmatrix} \mathbb{E}[X_{1,:}X_{1,:}^T] & \mathbb{E}[X_{1,:}X_{2,:}^T] & \ldots & \mathbb{E}[X_{1,:}X_{N,:}^T] \\ \mathbb{E}[X_{2,:}X_{1,:}^T] & \mathbb{E}[X_{2,:}X_{2,:}^T] & \ldots & \mathbb{E}[X_{2,:}X_{N,:}^T] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[X_{N,:}X_{1,:}^T] & \mathbb{E}[X_{N,:}X_{2,:}^T] & \ldots & \mathbb{E}[X_{N,:}X_{N,:}^T] \end{bmatrix} (I_{N\times N} \otimes a). \end{aligned}$$

We denote the covariance matrix in the middle as

$$C_{\mathcal{U}} = \mathbb{E}(\mathcal{U}\mathcal{U}^T) = \mathbb{E}\begin{bmatrix} X_{1,:} \\ \vdots \\ X_{N,:} \end{bmatrix} \begin{bmatrix} X_{1,:}^T \dots X_{N,:}^T \end{bmatrix} = \begin{bmatrix} \mathbb{E}[X_{1,:}X_{1,:}^T] & \mathbb{E}[X_{1,:}X_{2,:}^T] & \dots & \mathbb{E}[X_{1,:}X_{N,:}^T] \\ \mathbb{E}[X_{2,:}X_{1,:}^T] & \mathbb{E}[X_{2,:}X_{2,:}^T] & \dots & \mathbb{E}[X_{2,:}X_{N,:}^T] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[X_{N,:}X_{1,:}^T] & \mathbb{E}[X_{N,:}X_{2,:}^T] & \dots & \mathbb{E}[X_{N,:}X_{N,:}^T] \end{bmatrix}. \tag{31}$$

Thus, we established that $q \sim \mathcal{N}(0, Q)$, where $Q = (I \otimes a^T) C_{\mathcal{U}} (I \otimes a)$.

In what follows, we compute $trace(Q)$ and $||Q||_2$ for the covariance matrix Q, needed in the future computations. It can be seen that the trace of Q is given by

$$trace(Q) = N a^T C_{\mathsf{X}} a, \tag{32}$$

where C_{X} is defined in (29). Next, we compute upper bound on $||Q||_2$ as follows

$$\begin{aligned} ||Q||_2 &= ||(I \otimes a^T) C_{\mathcal{U}} (I \otimes a)||_2 \\ &\leq ||I \otimes a||_2^2 \, ||C_{\mathcal{U}}||_2 \\ &= ||a||_2^2 \, \Lambda_{\max}(C_{\mathcal{U}}), \end{aligned} \tag{33}$$

where the last equality follows since $||I \otimes a||_2^2 = \Lambda_{\max}\Big((I \otimes a^T)(I \otimes a)\Big) = \Lambda_{\max}\Big(I \otimes a^T a\Big) = ||a||_2^2$. We used a property of Kronecker product which states that for matrices with suitable dimensions, $(A \otimes B)(C \otimes D) = (AC \otimes BD)$.

To establish $\Lambda_{\max}(C_{\mathcal{U}})$, we use the results from Section 1.4.2, expression (23), which enable us to conclude that the upper bound of the largest eigenvalue of matrix $C_{\mathcal{U}}$ is given by

$$\Lambda_{\max}(C_{\mathcal{U}}) \leq \frac{\Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\boldsymbol{\mathcal{A}})}.$$

Therefore, the bound on the covariance matrix $||Q||_2$ in (33) is now given by

$$||Q||_2 \leq ||a||_2^2 \, \frac{\Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\boldsymbol{\mathcal{A}})}. \tag{34}$$

2.2 Bound on Regularization Parameter

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

Denote $E_{:,j} \in \mathbb{R}^N$ as a column of matrix E and vector $U = [u_1^T, \dots, u_p^T]^T \in \mathbb{R}^{dp^2}$, where $u_i \in \mathbb{R}^{dp}$. Note that since $Z = I_{p \times p} \otimes X$, and $\boldsymbol{\epsilon} = \text{vec}(E)$, we can observe the following

$$\begin{aligned} \sup_{R(U) \leq 1} \left\langle \frac{1}{N} Z^T \boldsymbol{\epsilon}, U \right\rangle &= \sup_{R(U) \leq 1} \frac{1}{N} \left\langle \left(I_{p \times p} \otimes X^T\right) \text{vec}(E), U \right\rangle \\ &= \sup_{R([u_1^T, \dots, u_p^T]^T) \leq 1} \frac{1}{N} \Big(\left\langle X^T E_{:,1}, u_1 \right\rangle + , \dots, + \left\langle X^T E_{:,p}, u_p \right\rangle \Big) \\ &= \frac{1}{N} \Big(\sup_{R([u_1^T, \dots, u_p^T]^T) \leq 1} \left\langle X^T E_{:,1}, u_1 \right\rangle + , \dots, + \sup_{R([u_1^T, \dots, u_p^T]^T) \leq 1} \left\langle X^T E_{:,p}, u_p \right\rangle \Big) \\ &= \frac{1}{N} \Big(\sup_{R(u_1) \leq r_1} \left\langle X^T E_{:,1}, u_1 \right\rangle + , \dots, + \sup_{R(u_p) \leq r_p} \left\langle X^T E_{:,p}, u_p \right\rangle \Big) \\ &= \frac{1}{N} \sum_{j=1}^{p} \sup_{R(u_j) \leq r_j} \left\langle X^T E_{:,j}, u_j \right\rangle \end{aligned} \tag{35}$$

where $\sum_{j=1}^p r_j \leq 1$ and $r_j \geq 0$.

Our objective is to establish a high probability bound of the form

$$\mathbb{P}\left[\sup_{R(U)\leq 1}\left\langle \frac{1}{N}Z^T\boldsymbol{\epsilon}, U\right\rangle \leq \alpha\right] \geq \pi$$

where $0 \leq \pi \leq 1$, i.e., upper bound should hold with at least probability π. Using (35) and assuming that $\alpha = \sum_{j=1}^{p}\alpha_j$, we can rewrite the above probabilistic statement as follows

$$\begin{aligned}
\mathbb{P}\left[\sup_{R(U)\leq 1}\left\langle \frac{1}{N}Z^T\boldsymbol{\epsilon}, U\right\rangle \leq \alpha\right] &= \mathbb{P}\left[\frac{1}{N}\sum_{j=1}^{p}\sup_{R(u_j)\leq r_j}\left\langle X^T E_{:,j}, u_j\right\rangle \leq \sum_{j=1}^{p}\alpha_j\right] \\
&\geq \mathbb{P}\left[\left\{\sup_{R(u_1)\leq r_1}\frac{1}{N}\left\langle X^T E_{:,1}, u_1\right\rangle \leq \alpha_1\right\} \text{ and } \ldots \text{ and } \left\{\sup_{R(u_p)\leq r_p}\frac{1}{N}\left\langle X^T E_{:,p}, u_p\right\rangle \leq \alpha_p\right\}\right] \\
&\geq \sum_{j=1}^{p}\mathbb{P}\left[\sup_{R(u_j)\leq r_j}\frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle \leq \alpha_j\right] - (p-1), \qquad (36)
\end{aligned}$$

where the last line follows from Lemma 2.4. In the above derivations we used the observation that if the events $\left\{\sup_{R(u_j)\leq r_j}\frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle \leq \alpha_j\right\}$, for each j hold, then the event $\left\{\sum_{j=1}^{p}\sup_{R(u_j)\leq r_j}\frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle \leq \sum_{j=1}^{p}\alpha_j\right\}$ also holds but the reverse is not always true, implying that the probability space related to the event $\left\{\sum_{j=1}^{p}\sup_{R(u_j)\leq r_j}\frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle \leq \sum_{j=1}^{p}\alpha_j\right\}$ is larger.

Therefore, based on (36), we see that we need to establish the following concentration bound

$$\mathbb{P}\left[\sup_{R(u_j)\leq r_j}\frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle \leq \alpha_j\right] \geq \pi_j, \qquad (37)$$

for each $j = 1, \ldots, p$.

In the following our objective would be to first establish that the random variable $\frac{1}{N}\left\langle X^T E_{:,j}, h\right\rangle$ has sub-exponential tails, where $h \in \mathbb{R}^{dp}$, $\|h\|_2 = 1$ is a unit norm vector. Based on the generic chaining argument we then use Theorem 1.2.7 in [10] and bound the expectation of the supremum of the original variable $\frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle$, i.e., bound $\mathbb{E}\left[\sup_{R(u_j)\leq r_j}\frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle\right]$. Finally, using Theorem 1.2.9 in [10] we establish the high probability bound on how $\sup_{R(u_j)\leq r_j}\frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle$ concentrates around its mean.

2.2.1 Martingale difference sequence

We start by writing

$$\left\langle X^T E_{:,j}, h\right\rangle = \langle E_{:,j}, Xh\rangle = \sum_{i=1}^{N} E_{i,j}, (X_{:,i}h) = \sum_{i=1}^{N} m_i,$$

where $m_i = E_{i,j}(X_{i,:}h)$, $i = 1, \ldots, N$. Observe that m_i is a martingale difference sequence (MDS), which can be shown by establishing that $\mathbb{E}(m_i|m_1, \ldots, m_{i-1}) = 0$ (see [5]). We can introduce a set $\{E_{1,:}, E_{2,:}, \ldots, E_{i-1,:}\} = \{\epsilon_d^T, \epsilon_{d+1}^T, \ldots, \epsilon_T^T\}$ and write

$$\mathbb{E}\big[m_i|m_1, \ldots, m_{i-1}\big] = \mathbb{E}\big[\mathbb{E}\big[m_i|m_1, \ldots, m_{i-1}, E_{1,:}, \ldots, E_{i-1,:}\big]\big],$$

using the technique of iterated expectation. Note that the set $\{E_{1,:}, E_{2,:}, \ldots, E_{i-1,:}\}$ contains more information than the set $\{m_1, \ldots, m_{i-1}\}$ and conditioning on it has fixed all the past history of the sequence until time stamp i. Since

$m_i = E_{i,j}(X_{i,:}h)$, the terms $E_{i,j}$ and $X_{i,:}h$ are now independent. The independence follows since every row of matrix X is independent of the corresponding row of matrix E:

$$E = \begin{bmatrix} \epsilon_d^T \\ \epsilon_{d+1}^T \\ \vdots \\ \epsilon_{T-1}^T \\ \epsilon_T^T \end{bmatrix}, \quad X = \begin{bmatrix} x_{d-1}^T & x_{d-2}^T & \dots & x_0^T \\ x_d^T & x_{d-1}^T & \dots & x_1^T \\ \vdots & \vdots & \ddots & \vdots \\ x_{T-2}^T & x_{T-3}^T & \dots & x_{T-d-1}^T \\ x_{T-1}^T & x_{T-2}^T & \dots & x_{T-d}^T \end{bmatrix},$$

which can be verified by noting that the noise vector ϵ_{d+i} is independent from x_{d-k+i} since $(d+i) > (d-k+i)$ for $0 \leq i \leq T-d$ and $1 \leq k \leq d$. In other words, the information contained in x_{d-k+i} does not contain information from the noise ϵ_{d+i} (see (4)). Moreover,

$$\mathbb{E}\Big[m_i\Big] = \mathbb{E}\Big[E_{i,j}(X_{i,:}h)\Big] = \mathbb{E}\Big[E_{i,j}\Big]\mathbb{E}\Big[X_{i,:}h\Big] = 0, \tag{38}$$

due to the zero-mean noise $\mathbb{E}\big[E_{i,j}\big] = 0$. Consequently, we have shown that $\mathbb{E}\big[m_i | m_1, \dots, m_{i-1}, E_{1,:}, \dots, E_{i-1,:}\big] = 0$ and therefore

$$\mathbb{E}\big[m_i | m_1, \dots, m_{i-1}\big] = 0,$$

proving that $m_i = E_{i,j}(X_{i,:}h)$, $i = 1, \dots, N$ is a martingale difference sequence.

Next, to show that $\frac{1}{N}\left\langle X^T E_{:,j}, h\right\rangle = \frac{1}{N}\sum_{i=1}^N m_i$ has sub-exponential tails, we first show that m_i is sub-exponential random variable and then use the proof argument similar to Azuma-type [1] and Bernstein-type [11] inequalities to establish that a sum over sub-exponential martingale difference sequence is itself sub-exponential.

2.2.2 Sub-exponential tails of $\frac{1}{N}\left\langle X^T E_{:,j}, h\right\rangle$

The MDS m_i is sub-exponential since it is a product of two Gaussians. Indeed, recall that E_{ij} and $X_{i,:}h$ are both Gaussian random variables, independent of each other. Employing a union bound enables us to write for any $\tau > 0$

$$\begin{aligned} \mathbb{P}\Big[|m_i| \geq \tau\Big] &= \mathbb{P}\Big[|E_{ij}(X_{i,:}h)| \geq \tau\Big] \\ &\leq \mathbb{P}\Big[|E_{ij}| \geq \sqrt{\tau}\Big] + \mathbb{P}\Big[|X_{i,:}h| \geq \sqrt{\tau}\Big] \\ &\leq 2e^{-c_1\tau} + 2e^{-c_2\tau} \\ &\leq 4e^{-c\tau}, \end{aligned}$$

for some suitable constants $c_1 > 0$, $c_2 > 0$ and $c > 0$.

To establish that $\frac{1}{N}\sum_i m_i$ is sub-exponential, we note that the sub-exponential norm $\|\cdot\|_{\psi_1}$ (see [11], Definition 5.13) of m_i can be upper-bounded by a constant. We denote by $\kappa > 0$ the largest of these constants, i.e.,

$$\kappa = \max_{i=1,\dots,N} \|m_i\|_{\psi_1} = \max_{i=1,\dots,N} \|X_{i,:}h\|_{\psi_1}.$$

Now, using Lemma 5.15 in [11], the moment generating function of m_i satisfies the following result: for s such that $|s| \leq \frac{\eta}{\kappa}$ and for all $i = 1, \dots, N$

$$\mathbb{E}\Big[e^{sm_i}\Big] \leq e^{cs^2\kappa^2}, \tag{39}$$

where c and η are absolute constants. Next, using Markov inequality, we can write for any $\varepsilon' > 0$

$$\begin{aligned} \mathbb{P}\left[\sum_{i=1}^N m_i \geq \varepsilon'\right] &= \mathbb{P}\left[\exp\left(s\sum_{i=1}^N m_i\right) \geq \exp(s\varepsilon')\right] \\ &\leq \frac{\mathbb{E}\left[\exp\left(s\sum_{i=1}^N m_i\right)\right]}{\exp(s\varepsilon')}. \end{aligned} \tag{40}$$

To bound the numerator, we use (39) and write for $|s| \leq \frac{\eta}{\kappa}$ utilizing the iterated expectation

$$\begin{aligned}
\mathbb{E}\left[\exp\left(s\sum_{i=1}^{N} m_i\right)\right] &= \mathbb{E}\left[\exp(sm_N)\exp\left(s\sum_{i=1}^{N-1} m_i\right)\right] \\
&= \mathbb{E}_{m_1,\ldots,m_{N-1}}\left[\mathbb{E}_{m_N|m_1,\ldots,m_{N-1}}\left[\exp(sm_N)\exp\left(s\sum_{i=1}^{N-1} m_i\right)\right]\right] \\
&= \mathbb{E}_{m_1,\ldots,m_{N-1}}\left[\mathbb{E}_{m_N|m_1,\ldots,m_{N-1}}\Big[\exp(sm_N)\Big]\exp\left(s\sum_{i=1}^{N-1} m_i\right)\right] \\
&\overset{\text{using (39)}}{\leq} \exp(cs^2\kappa^2)\mathbb{E}_{m_1,\ldots,m_{N-1}}\left[\exp\left(s\sum_{i=1}^{N-1} m_i\right)\right] \\
&\leq \exp(cs^2\kappa^2)\exp(cs^2\kappa^2)\mathbb{E}_{m_1,\ldots,m_{N-2}}\left[\exp\left(s\sum_{i=1}^{N-2} m_i\right)\right] \\
&\vdots \\
&\leq \exp(Ncs^2\kappa^2)
\end{aligned}$$

Substituting back to (40), we get for $|s| \leq \frac{\eta}{\kappa}$

$$\mathbb{P}\left[\sum_{i=1}^{N} m_i \geq \varepsilon'\right] \leq \exp(-s\varepsilon' + Ncs^2\kappa^2). \tag{41}$$

We now select s to minimize the right hand side of (41). For this, note that if the minimum is achieved for an s, which satisfies $|s| \leq \frac{\eta}{\kappa}$, then we simply minimize $-s\varepsilon' + Ncs^2\kappa^2$ and get $s = \frac{\varepsilon'}{N2c\kappa^2}$. On the other hand, if the minimum is achieved for an s outside the range $|s| \leq \frac{\eta}{\kappa}$, we pick the one on boundary $s = \frac{\eta}{\kappa}$. Thus, choosing $s = \min\left(\frac{\varepsilon'}{N2c\kappa^2}, \frac{\eta}{\kappa}\right)$, we obtain

$$\mathbb{P}\left[\sum_{i=1}^{N} m_i \geq \varepsilon'\right] \leq \exp\left(-\min\left(\frac{\varepsilon'^2}{4cN\kappa^2}, \frac{\eta\varepsilon'}{2\kappa}\right)\right).$$

Finally, setting $\varepsilon' = N\varepsilon$, for a suitable constant $c > 0$, we get

$$\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N} m_i \geq \varepsilon\right] \leq \exp\left(-c\min\left(\frac{N\varepsilon^2}{\kappa^2}, \frac{N\varepsilon}{\kappa}\right)\right).$$

Repeating the above argument for $-\frac{1}{N}\sum_{i=1}^{N} m_i$, we obtain same bound and a combination of both of them gives the required concentration inequality for the sum over the martingale difference sequence

$$\mathbb{P}\left[\frac{1}{N}\left|\sum_{i=1}^{N} m_i\right| \geq \varepsilon\right] = \mathbb{P}\left[\frac{1}{N}\left|\left\langle X^T E_{:,j}, h\right\rangle\right| \geq \varepsilon\right] \leq 2\exp\left(-c\min\left(\frac{N\varepsilon^2}{\kappa^2}, \frac{N\varepsilon}{\kappa}\right)\right). \tag{42}$$

2.2.3 Establishing bound on $\mathbb{E}\left[\sup_{R(u_j)\leq r_j} \frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle\right]$

To establish a high probability bound on the mean of $\sup_{R(u_j)\leq r_j}\left\langle X^T E_{:,j}, u_j\right\rangle$, we use a generic chaining argument from [10], in particular Theorem 1.2.7 in [9]. For this, we define $(Y_{u_j})_{u_j \in R(u_j)\leq r_j} = \frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle$ and $(Y_{v_j})_{v_j \in R(v_j)\leq r_j} = \frac{1}{N}\left\langle X^T E_{:,j}, v_j\right\rangle$ to be two centered random symmetric process, indexed by a fixed vectors u_j and v_j, respectively. They are centered due to (38) and they are symmetric since, for example, the process $(Y_{u_j})_{u_j \in R(u_j)\leq r_j}$ has the same

law as process $\Big(- (Y_{u_j})_{u_j \in R(u_j) \leq r_j} \Big)$ (see the results established in (42)). Consider now the absolute difference of these two processes

$$\left| (Y_{u_j})_{u_j \in R(u_j) \leq r_j} - (Y_{v_j})_{v_j \in R(v_j) \leq r_j} \right| = \frac{1}{N} \left| \left\langle X^T E_{:,j}, u_j - v_j \right\rangle \right| = \|u_j - v_j\|_2 \frac{1}{N} \left| \left\langle X^T E_{:,j}, \frac{u_j - v_j}{\|u_j - v_j\|_2} \right\rangle \right|.$$

Using now the bound obtained in (42), we get

$$\begin{aligned}
&\mathbb{P}\left[\frac{1}{N} \left| \left\langle X^T E_{:,j}, \frac{u_j - v_j}{\|u_j - v_j\|_2} \right\rangle \right| \geq \varepsilon \right] \\
=&\mathbb{P}\left[\|u_j - v_j\|_2 \frac{1}{N} \left| \left\langle X^T E_{:,j}, \frac{u_j - v_j}{\|u_j - v_j\|_2} \right\rangle \right| \geq \|u_j - v_j\|_2 \varepsilon \right] \\
=&\mathbb{P}\left[\frac{1}{N} \left| \left\langle X^T E_{:,j}, u_j - v_j \right\rangle \right| \geq \tau \right] \leq 2 \exp\left(-c \min\left(\frac{N\tau^2}{\|u_j - v_j\|_2^2 \kappa^2}, \frac{N\tau}{\|u_j - v_j\|_2 \kappa} \right) \right),
\end{aligned}$$

where $\tau = \|u_j - v_j\|_2 \varepsilon$. Then, according to Theorem 1.2.7 in [9], we obtain the following bound on the expectation of the supremum of the difference between the processes

$$\mathbb{E}\left[\sup_{R(u_j) \leq r_j, R(v_j) \leq r_j} \frac{1}{N} \left| \left\langle X^T E_{:,j}, u_j \right\rangle - \left\langle X^T E_{:,j}, v_j \right\rangle \right| \right] \leq c \left(\gamma_1 \left(S_j, \frac{\|u_j - v_j\|_2}{N} \right) + \gamma_2 \left(S_j, \frac{\|u_j - v_j\|_2}{\sqrt{N}} \right) \right), \tag{43}$$

where c is a constant, $f_i(S_j, d_i)$, $i = 1, 2$, are the majorizing measures, which are defined in [10], Definition 1.2.5; $d_1 = \frac{\|u_j - v_j\|_2}{N}$ and $d_2 = \frac{\|u_j - v_j\|_2}{\sqrt{N}}$ are the distance measures on the set S_j defined for all vectors $s \in S_j : R(s) \leq r_j$. The definition of majorizing measure is as follows, for $\alpha > 0$

$$\gamma_\alpha(S_j, d) = \inf \sup_t \sum_{k \geq 0} 2^{\frac{k}{\alpha}} \Delta(A_k(t)), \tag{44}$$

where $\inf$ is taken over all possible admissible sequences of the set S_j; $\Delta(A_k(t))$ denotes the diameter of element $A_k(t)$ with respect to the distance metric d defined as

$$\Delta(A_k(t)) = \sup_{t_1, t_2 \in A_k(t)} d(t_1, t_2), \tag{45}$$

and $A_k(t) \in \mathcal{A}_k$ is an element of an admissible sequence in generic chaining, see Definition 1.2.3 in [10] for a detailed discussion on how $\mathcal{A}_k$ are constructed.

Observe that from definition of a diameter $\Delta(\cdot)$ in (45) and majorizing measure in (44) we can immediately see that for any constant $c > 0$

$$\gamma_\alpha \left(S_j, cd \right) = c \gamma_\alpha \left(S_j, d \right), \tag{46}$$

since $\inf \sup_t \sum_{k \geq 0} 2^{\frac{k}{\alpha}} \sup_{t_1, t_2 \in A_k(t)} cd(t_1, t_2) = c \inf \sup_t \sum_{k \geq 0} 2^{\frac{k}{\alpha}} \sup_{t_1, t_2 \in A_k(t)} d(t_1, t_2)$. Moreover, in the next result we establish the following useful Lemma which would enable us to bound the γ_1 with the square of γ_2.

Lemma 2.7 *Given a metric space (S_j, d), we have*

$$\gamma_1(S_j, \|.\|_2) \leq \gamma_2^2(S_j, \|.\|_2). \tag{47}$$

To prove this Lemma, we define $d(s, t) = \|s - t\|_2$. We use the traditional definition of majorizing measure $\gamma'_\alpha(S_j, d)$ from [8], equation (1.2):

$$\gamma'_\alpha(S_J, d) = \inf \sup_{s \in S} \left(\int_0^\infty \left(\log \frac{1}{\mu(B_d(s, \varepsilon))} \right)^{1/\alpha} d\varepsilon \right),$$

where $B_d(s,\varepsilon)$ is the closed ball of center t and radius ε based on the distance d and the infimum is taken over all the probability measure μ on S_j.

Note that $\gamma'_\alpha(S_j,d)$ relates to the majorizing measure $\gamma_\alpha(S_j,d)$ used in (43) as (see [8], Theorem 1.2)

$$K(\alpha)^{-1}\gamma_\alpha(S_j,d) \leq \gamma'_\alpha(S_j,d) \leq K(\alpha)\gamma_\alpha(S_j,d),$$

where $K(\alpha)$ is a constant depending on α only. As a result, it is enough to show that $\gamma'_1(S_j,d) \leq {\gamma'_2}^2(S_j,d)$. The required relationship is then established as follows

$$\begin{aligned}
\gamma'_1(S_j,d) &= \inf_t \sup \left(\int_0^\infty \left(\log \frac{1}{\mu(B_d(t,\varepsilon))} \right) d\varepsilon \right) \\
&\leq \inf_t \sup \left(\int_0^\infty \left(\log \frac{1}{\mu(B_d(t,\varepsilon))} \right)^{1/2} d\varepsilon \right)^2 \\
&= {\gamma'_2}^2(S_j,d).
\end{aligned}$$

And this completes the proof. Now using Theorem 2.1.1 in [10], and the definition of $\gamma_\alpha(S_j,d)$ in (44) we can establish that

$$\begin{aligned}
\gamma_2\left(S_j, \frac{\|.\|_2}{\sqrt{N}}\right) &= \frac{1}{\sqrt{N}}\gamma_2(S_j,\|.\|_2) \quad \text{using (46)} \\
&\leq \frac{1}{\sqrt{N}}\mathbb{E}\Big[\sup_{R(z)\leq r_j} \langle g,z\rangle\Big] \quad \text{using Theorem 2.1.1 in [10]} \\
&= r_j \frac{1}{\sqrt{N}}\mathbb{E}\Big[\sup_{R(u)\leq 1} \langle g,u\rangle\Big] \quad \text{since } \mathbb{E}\Big[\sup_{R(z)\leq r_j} \langle g,z\rangle\Big] = r_j\mathbb{E}\Big[\sup_{R(u)\leq 1} \langle g,u\rangle\Big] \text{ for } z = r_j u \\
&= r_j \frac{1}{\sqrt{N}} w(\Omega_R), \qquad (48)
\end{aligned}$$

where in the last line we used the description of Gaussian width in Definition 2.6. Using Lemma 2.7 and (46) above, we also get

$$\begin{aligned}
\gamma_1\left(S_j, \frac{\|.\|_2}{N}\right) &= \frac{1}{N}\gamma_1\left(S_j,\|.\|_2\right) \quad \text{using (46)} \\
&\leq \frac{1}{N}\gamma_2^2\left(S,\|.\|_2\right) \quad \text{using Lemma 2.7} \\
&\leq r_j^2 \frac{1}{N^2} w^2(\Omega_R) \quad \text{using (48)} \\
&\leq r_j \frac{1}{N^2} w^2(\Omega_R), \qquad (49)
\end{aligned}$$

where in the last line we used the fact that $r_j < 1$. Finally, substituting (48) and (49) into (43) and using Lemma 1.2.8 in [9], we get

$$\begin{aligned}
\mathbb{E}\left[\sup_{R(u_j)\leq r_j, R(v_j)\leq r_j} \frac{1}{N}\left| \left\langle X^T E_{:,j}, u_j\right\rangle - \left\langle X^T E_{:,j}, v_j\right\rangle \right|\right] &= \mathbb{E}\left[\sup_{R(u_j)\leq r_j}\left|\frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle\right|\right] \\
&\leq c r_j \left(\frac{w(\Omega_R)}{\sqrt{N}} + \frac{w^2(\Omega_R)}{N^2}\right). \qquad (50)
\end{aligned}$$

2.2.4 Establishing high probability concentration bound

Next, in order to establish a high probability concentration of the supremum of the random variable $\frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle$ around its mean, we use Theorem 1.2.9 from [10]. For any $\epsilon_1 > 0$ and $\epsilon_2 > 0$, we have

$$\mathbb{P}\left[\sup_{R(u_j)\leq r_j}\left|\frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle\right| \geq \mathbb{E}\left[\sup_{R(u_j)\leq r_j}\left|\frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle\right|\right] + \epsilon_1 D_1 + \epsilon_2 D_2\right] \leq c\exp(-\min(\epsilon_2^2, \epsilon_1)). \quad (51)$$

where $D_i \leq \gamma_i(S_j, d)$, $i = 1, 2$, where $\gamma_i(S_j, d)$ are as defined in the discussion after (43). Therefore, using the result (50), the concentration inequality (51) can now be written as

$$\mathbb{P}\left[\sup_{R(u_j)\leq r_j}\left|\frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle\right| \geq \left(c_2(1+\epsilon_2)r_j\frac{w(\Omega_R)}{\sqrt{N}} + c_1(1+\epsilon_1)r_j\frac{w^2(\Omega_R)}{N^2}\right)\right] \leq c\exp(-\min(\epsilon_2^2, \epsilon_1)). \quad (52)$$

To adapt to the form required in (37), we reverse the direction of inequality

$$\mathbb{P}\left[\sup_{R(u_j)\leq r_j}\left|\frac{1}{N}\left\langle X^T E_{:,j}, u_j\right\rangle\right| \leq \left(c_2(1+\epsilon_2)r_j\frac{w(\Omega_R)}{\sqrt{N}} + c_1(1+\epsilon_1)r_j\frac{w^2(\Omega_R)}{N^2}\right)\right] \geq 1 - c\exp(-\min(\epsilon_2^2, \epsilon_1)). \quad (53)$$

2.2.5 Overall bound

Now we can combine the results obtained in (53) for each $j = 1, \ldots, p$ using the fact that $\sum_{j=1}^p r_j \leq 1$ and using the form of the overall bound in (36). Therefore, we get

$$\mathbb{P}\left[\sup_{R(U)\leq 1}\left\langle \frac{1}{N}Z^T\boldsymbol{\epsilon}, U\right\rangle \leq \left(c_2(1+\epsilon_2)\frac{w(\Omega_R)}{\sqrt{N}} + c_1(1+\epsilon_1)\frac{w^2(\Omega_R)}{N^2}\right)\right] \geq 1 - c\exp(-\min(\epsilon_2^2, \epsilon_1) + \log(p)).$$

This concludes our proof on establishing the bound on the regularization parameter.

2.3 Restricted Eigenvalue Condition

To establish restricted eigenvalue (RE) condition, we need to show that $\frac{||(I_{p\times p}\otimes X)\Delta||_2}{||\Delta||_2} \geq \sqrt{\kappa N}$, $\kappa > 0$, for all $\Delta = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, $\Delta \in \text{cone}(\Omega_E)$, where $\text{cone}(\Omega_E)$ is a cone of an error set $\Omega_E = \left\{\Delta \in \mathbb{R}^{dp^2} \middle| R(\boldsymbol{\beta}^* + \Delta) \leq R(\boldsymbol{\beta}^*) + \frac{1}{c}R(\Delta)\right\}$. To show $\frac{||(I_{p\times p}\otimes X)\Delta||_2}{||\Delta||_2} \geq \sqrt{\kappa N}$ for all $\Delta \in \text{cone}(\Omega_E)$, we will show that $\inf_{\Delta\in\text{cone}(\Omega_E)} \frac{||(I_{p\times p}\otimes X)\Delta||_2}{||\Delta||_2} \geq \sqrt{\rho}$, for some $\rho > 0$ and then set $\kappa N = \rho$.

Note that the error vector can be written as $\Delta = [\Delta_1^T, \Delta_2^T, \ldots, \Delta_p^T]^T$, where Δ_i is of size $dp \times 1$. Also let $\boldsymbol{\beta}^* = [\beta_1^{*T}\beta_2^{*T}\ldots\beta_p^{*T}]^T$, for $\beta_i^* \in \mathbb{R}^{dp}$, then using our assumption in (6) that the norm $R(\cdot)$ is decomposable, we can represent original set Ω_E as a Cartesian product of subsets Ω_{E_i}, i.e., $\Omega_E = \Omega_{E_1} \times \Omega_{E_2} \times \cdots \times \Omega_{E_p}$, where

$$\Omega_{E_i} = \left\{\Delta_i \in \mathbb{R}^{dp} \middle| R(\beta_i^* + \Delta_i) \leq R(\beta_i^*) + \frac{1}{c}R(\Delta_i)\right\},$$

which also implies that $\text{cone}(\Omega_E) = \text{cone}(\Omega_{E_1}) \times \text{cone}(\Omega_{E_2}) \times \cdots \times \text{cone}(\Omega_{E_p})$. Also, if $||\Delta||_2 = 1$, then we denote $||\Delta_i||_2 = \delta_i > 0$, so that $\sum_{i=1}^p \delta_i^2 = 1$. With this information, we can write

$$\begin{aligned}
\inf_{\Delta\in\text{cone}(\Omega_E)} \frac{||(I_{p\times p}\otimes X)\Delta||_2^2}{||\Delta||_2^2} &= \inf_{\substack{\Delta\in\text{cone}(\Omega_E)\\ ||\Delta||_2=1}} ||(I_{p\times p}\otimes X)\Delta||_2^2 \\
&= \inf_{\substack{\Delta\in\text{cone}(\Omega_E)\\ ||\Delta||_2=1}} ||X\Delta_1||_2^2 + ||X\Delta_2||_2^2 + \ldots + ||X\Delta_p||_2^2 \\
&= \sum_{i=1}^p \inf_{\substack{\Delta_i\in\text{cone}(\Omega_{e_i})\\ ||\Delta_i||_2=\delta_i}} ||X\Delta_i||_2^2. \quad (54)
\end{aligned}$$

Our objective is to establish a high probability bound of the form

$$\mathbb{P}\left[\inf_{\Delta\in\text{cone}(\Omega_E)} \frac{||(I_{p\times p}\otimes X)\Delta||_2}{||\Delta||_2} \geq \rho\right] \geq \pi$$

where $0 \leq \pi \leq 1$, i.e., lower bound should hold with at least probability π. Note that if we square the terms inside the probability statement above, the probability of the resulting expression does not change since the squared terms are positive. Therefore, using (54) and assuming that $\rho^2 = \sum_{i=1}^{p} \rho_i^2$ we can rewrite the above as follows

$$\begin{aligned}
\mathbb{P}\left[\inf_{\Delta \in \text{cone}(\Omega_E)} \frac{||(I_{p\times p} \otimes X)\Delta||_2}{||\Delta||_2} \geq \rho\right] &= \mathbb{P}\left[\inf_{\Delta \in \text{cone}(\Omega_E)} \frac{||(I_{p\times p} \otimes X)\Delta||_2^2}{||\Delta||_2^2} \geq \sum_{i=1}^{p} \rho_i^2\right] \\
&= \mathbb{P}\left[\sum_{i=1}^{p} \inf_{\substack{\Delta_i \in \text{cone}(\Omega_{E_i}) \\ ||\Delta_i||_2 = \delta_i}} ||X\Delta_i||_2^2 \geq \sum_{i=1}^{p} \rho_i^2\right] \quad \text{using (54)} \\
&\geq \mathbb{P}\left[\left\{\inf_{\substack{\Delta_1 \in \text{cone}(\Omega_{E_1}) \\ ||\Delta_1||_2 = \delta_1}} ||X\Delta_1||_2^2 \geq \rho_i^2\right\} \text{ and } \ldots \text{ and } \left\{\inf_{\substack{\Delta_p \in \text{cone}(\Omega_{E_p}) \\ ||\Delta_p||_2 = \delta_p}} ||X\Delta_p||_2^2 \geq \rho_i^2\right\}\right] \\
&\geq \sum_{i=1}^{p} \mathbb{P}\left[\inf_{\substack{\Delta_i \in \text{cone}(\Omega_{E_i}) \\ ||\Delta_i||_2 = \delta_i}} ||X\Delta_i||_2^2 \geq \rho_i^2\right] - (p-1) \quad \text{using Lemma 2.4} \\
&= \sum_{i=1}^{p} \mathbb{P}\left[\inf_{\substack{\Delta_i \in \text{cone}(\Omega_{E_i}) \\ ||\Delta_i||_2 = \delta_i}} ||X\Delta_i||_2 \geq \rho_i\right] - (p-1) \quad \text{taking square root} \\
&= \sum_{i=1}^{p} \mathbb{P}\left[\inf_{\substack{\Delta_i \in \text{cone}(\Omega_{E_i}) \\ ||\Delta_i||_2 = \delta_i}} \frac{||X\Delta_i||_2}{||\Delta_i||_2} \geq \frac{\rho_i}{||\Delta_i||_2}\right] - (p-1) \\
&= \sum_{i=1}^{p} \mathbb{P}\left[\inf_{u_i \in \text{cone}(\Omega_{E_i}) \cap S^{dp-1}} ||Xu_i||_2 \geq \frac{\rho_i}{\delta_i}\right] - (p-1) \qquad (55)
\end{aligned}$$

where we defined $u_i = \frac{\Delta_i}{||\Delta_i||_2}$ and S^{dp-1} is a unit sphere. Therefore, if we denote $\Theta_i = \text{cone}(\Omega_{E_i}) \cap S^{dp-1}$, we need to establish a lower bound of the form

$$\mathbb{P}\left[\inf_{u_i \in \Theta_i} ||Xu_i||_2 \geq \rho_i'\right] \geq \pi_i, \qquad (56)$$

where $\rho_i' = \frac{\rho_i}{\delta_i}$. In the following derivations we set $\Theta = \text{cone}(\Omega_{E_i}) \cap S^{dp-1}$ and $u = u_i$ for all $i = 1, \ldots, p$ since the specific index i is irrelevant.

2.3.1 Bound on $\inf_{u\in\Theta} ||Xu||_2$

Using results from Section 2.1.2 we can establish that $Xu \in \mathbb{R}^N$ is a Gaussian random vector, i.e., $Xu \sim \mathcal{N}(0, Q_u)$, where covariance matrix $Q_u = (I_{N\times N} \otimes u^T) C_{\mathcal{U}} (I_{N\times N} \otimes u)$, $C_{\mathcal{U}}$ is defined in (31), and $u \in \Theta$ is a fixed vector.

To establish $\inf_{u\in\Theta} ||Xu||_2$, we invoke a generic chaining argument from [10], specifically Theorem 2.1.5. For this we let $(Z_u)_{u\in\Theta} = ||Xu||_2 - \mathbb{E}(||Xu||_2)$ and $(Z_v)_{v\in\Theta} = ||Xv||_2 - \mathbb{E}(||Xv||_2)$ be two centered symmetric random processes. They are centered since, for example, $\mathbb{E}\Big[(Z_u)_{u\in\Theta}\Big] = \mathbb{E}(||Xu||_2) - \mathbb{E}(||Xu||_2) = 0$, and they are symmetric due to the later result shown in (58).

Sub-gaussianity of the process $Z_u - Z_v$.
We can show that the process difference

$$(Z_u)_{u\in\Theta} - (Z_v)_{v\in\Theta} = \|u-v\|_2 \left(\left\|X\frac{u-v}{\|u-v\|_2}\right\|_2 - \mathbb{E}\left(\left\|X\frac{u-v}{\|u-v\|_2}\right\|_2\right)\right) \qquad (57)$$

is a sub-Gaussian random process. This is indeed the case since we can establish that for $Z = ||X\frac{u-v}{\|u-v\|_2}||_2 - \mathbb{E}(||X\frac{u-v}{\|u-v\|_2}||_2)$, the sub-gaussian norm $\|Z\|_{\psi_2} \leq K$ for some constant $K > 0$ (see [11], Definition 5.7). To show

this, let $\xi = \frac{u-v}{\|u-v\|_2}$ and use Lemma 2.1 for the concentration of a Lipschitz function of Gaussian random variables. Specifically, observe that $X\xi \sim \mathcal{N}(0, Q_\xi)$ is distributed same as $\sqrt{Q_\xi}g \sim \mathcal{N}(0, Q_\xi)$, where $g \sim \mathcal{N}(0, I_{N\times N})$. Therefore, we can write

$$\mathbb{P}\Big[\, |\|X\xi\|_2 - \mathbb{E}(\|X\xi\|_2)| > \tau\Big] = \mathbb{P}\Big[\, \Big|\|\sqrt{Q_\xi}g\|_2 - \mathbb{E}(\|\sqrt{Q_\xi}g\|_2)\Big| > \tau\Big].$$

Moreover, note that $\|\sqrt{Q_\xi}g\|_2$ is a Lipschitz function with constant $\|\sqrt{Q_\xi}\|_2$ since we can write $\Big|\|\sqrt{Q_\xi}g_1\|_2 - \|\sqrt{Q_\xi}g_2\|_2\Big| \le \|\sqrt{Q_\xi}(g1-g2)\|_2 \le \|\sqrt{Q_\xi}\|_2\, \|g_1-g_2\|_2$. Using Lemma 2.1, we can obtain for all $\tau > 0$

$$\begin{aligned}
\mathbb{P}\Big[\, |\|X\xi\|_2 - \mathbb{E}(\|X\xi\|_2)| > \tau\Big] &= \mathbb{P}\Big[\, \Big|\|\sqrt{Q_\xi}g\|_2 - \mathbb{E}(\|\sqrt{Q_\xi}g\|_2)\Big| > \tau\Big] \\
&\le 2\exp\left(-\frac{\tau^2}{2\|Q_\xi\|_2}\right) \\
&\le 2\exp\left(-\frac{\tau^2}{2\mathcal{M}}\right), \qquad (58)
\end{aligned}$$

where $||Q_\xi||_2 \le ||\xi||_2^2 \frac{\Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\boldsymbol{\mathcal{A}})} = \frac{\Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\boldsymbol{\mathcal{A}})} = \mathcal{M}$ (see (34)), and which shows that $\|X\xi\|_2$ is sub-Gaussian with constant $K = \sqrt{\mathcal{M}}$.

Now, using (58) we can establish the sub-Gaussian tails of (57). Define $\tau' = \|u-v\|_2\tau$ and write

$$\begin{aligned}
\mathbb{P}\Big[\, \Big|\|u-v\|_2\Big(\|X\xi\|_2 - \mathbb{E}(\|X\xi\|_2)\Big)\Big| > \|u-v\|_2\tau\Big] &= \mathbb{P}\Big[\, |(Z_u)_{u\in\Theta} - (Z_v)_{v\in\Theta}| > \tau'\Big] \\
&\le 2\exp\left(-\frac{\tau'^2}{2\|u-v\|_2^2\mathcal{M}}\right). \qquad (59)
\end{aligned}$$

Establishing bound on $\mathbb{E}\left(\inf_{u\in\Theta}||Xu||_2\right)$.

Using the results established in (59) and Theorem 2.1.5 in [10], we can conclude that the distance measure on the set Θ is $d(u,v) = \|u-v\|_2$ for $u, v \in \Theta$. Moreover, we can now obtain an upper bound on the expectation of the supremum of the process difference $|Z_u - Z_v|$

$$\begin{aligned}
\mathbb{E}\left(\sup_{u,v\in\Theta}\Big|Z_u - Z_v\Big|\right) &= \mathbb{E}\left(\sup_{u,v\in\Theta}\Big|\, ||X(u-v)||_2 - \mathbb{E}(||X(u-v)||_2)\,\Big|\right) \\
&= \mathbb{E}\left(\sup_{u\in\Theta}\Big|\, ||Xu||_2 - \mathbb{E}(||Xu||_2)\,\Big|\right) \quad \text{using Lemma 1.2.8 in [10]} \\
&\le \mathbb{E}\Big[\sup_{u\in\Theta}\langle g, u\rangle\Big] \\
&\le cw(\Theta), \qquad (60)
\end{aligned}$$

where $g \sim \mathcal{N}(0, I)$, $w(\Theta)$ is the Gaussian width of set Θ and c is a constant.

Since we are interested in the bound on $\inf_{u\in\Theta}||Xu||_2$, we can extract from (60) the lower bound on the expectation of the infimum of the process. Specifically, note that (60) can be written as

$$\mathbb{E}\left(\left|\inf_{u\in\Theta}||Xu||_2 - \inf_{u\in\Theta}\mathbb{E}(||Xu||_2)\,\right|\right) \le \mathbb{E}\left(\sup_{u\in\Theta}\Big|\, ||Xu||_2 - \mathbb{E}(||Xu||_2)\,\Big|\right) \le cw(\Theta),$$

leading to

$$-cw(\Theta) \le \mathbb{E}\left(\inf_{u\in\Theta}||Xu||_2 - \inf_{u\in\Theta}\mathbb{E}(||Xu||_2)\right) \le cw(\Theta).$$

The lower bound then takes the form

$$\mathbb{E}\left(\inf_{u\in\Theta}||Xu||_2\right) \ge \inf_{u\in\Theta}\mathbb{E}(||Xu||_2) - cw(\Theta) \qquad (61)$$

Note that the vector Xu is distributed as $Xu \sim \mathcal{N}(0, Q_u)$, which is the same as a vector $\sqrt{Q_u}g \sim \mathcal{N}(0, Q_u)$ for $g \sim \mathcal{N}(0, I)$. Therefore, using results of Lemma I.2 from [6], we can extract the following inequality

$$\left| \sqrt{trace(Q_u)} - \mathbb{E}(\|\sqrt{Q_u}g\|_2) \right| \leq 2\sqrt{\Lambda_{\max}(Q_u)}.$$

Moreover, based on our discussion, the same inequality holds for the random vector Xu since $\mathbb{E}(\|\sqrt{Q_u}g\|_2) = \mathbb{E}(\|Xu\|_2)$

$$\left| \sqrt{trace(Q_u)} - \mathbb{E}(\|Xu\|_2) \right| \leq 2\sqrt{\Lambda_{\max}(Q_u)}.$$

which leads to a lower bound on the expectation of the norm

$$\mathbb{E}(\|Xu\|_2) \geq \sqrt{trace(Q_u)} - 2\sqrt{\Lambda_{\max}(Q_u)}. \tag{62}$$

We will lower-bound the first term on the right hand side of (62) and upper bound the second one. In particular, using (32) we write $trace(Q_u) = Nu^T C_{\mathsf{X}} u$ for any $u \in \Theta$ and bound

$$trace(Q_u) = Nu^T C_{\mathsf{X}} u = N||C_{\mathsf{X}}^{\frac{1}{2}} u||_2^2 \geq N \inf_{u\in\Theta} u^T C_{\mathsf{X}} u \geq N \inf_{u\in\mathbb{R}^{dp}} u^T C_{\mathsf{X}} u = N\Lambda_{\min}(C_{\mathsf{X}}) \geq N\frac{\Lambda_{\min}(\Sigma)}{\Lambda_{\max}(\mathcal{A})} = N\mathcal{L}. \tag{63}$$

Moreover, using (34), we bound

$$||Q_u||_2 \leq ||u||_2^2 \frac{\Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\boldsymbol{\mathcal{A}})} = \frac{\Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\boldsymbol{\mathcal{A}})} = \mathcal{M}. \tag{64}$$

Therefore, substituting (64) and (63) into (62), we get

$$\mathbb{E}(\|Xu\|_2) \geq \sqrt{N\mathcal{L}} - 2\sqrt{\mathcal{M}}.$$

Since $\mathbb{E}(\|Xu\|_2)$ is bounded from below, we can write

$$\inf_{u\in\Theta} \mathbb{E}(\|Xu\|_2) \geq \sqrt{N\mathcal{L}} - 2\sqrt{\mathcal{M}}. \tag{65}$$

Finally, substituting (65) in (61) gives us

$$\mathbb{E}\left(\inf_{u\in\Theta} ||Xu||_2\right) \geq \sqrt{N\mathcal{L}} - 2\sqrt{\mathcal{M}} - cw(\Theta). \tag{66}$$

Establishing concentration inequality of $\inf_{u\in\Theta} ||Xu||_2$.

Now from Lemma 2.1.3 in [10] and the results in [2] we extract the form of the high probability concentration inequality of $\inf_{u\in\Theta} ||Xu||_2$ around its mean, for $\tau > 0$

$$\mathbb{P}\left[\inf_{u\in\Theta} ||Xu||_2 \leq \mathbb{E}\left(\inf_{u\in\Theta} ||Xu||_2\right) - \tau\right] \leq c_1 \exp(-c_2\tau^2).$$

In order to bring the above expression into the form of (56), we write

$$\mathbb{P}\left[\inf_{u\in\Theta} ||Xu||_2 \geq \mathbb{E}\left(\inf_{u\in\Theta} ||Xu||_2\right) - \tau\right] \geq 1 - c_1 \exp(-c_2\tau^2).$$

Substituting the bound on the expectation from (66) gives us

$$\mathbb{P}\left[\inf_{u\in\Theta} ||Xu||_2 \geq \sqrt{N\mathcal{L}} - 2\sqrt{\mathcal{M}} - cw(\Theta) - \tau\right] \leq c_1 \exp(-c_2\tau^2). \tag{67}$$

2.3.2 Overall bound

Observe that in (67) we established a bound for each $u_i = \frac{\Delta_i}{||\Delta_i||_2}$ of the form

$$\mathbb{P}\left[\inf_{\substack{\Delta_i \in \text{cone}(\Omega_{E_i}) \\ ||\Delta_i||_2 = \delta_i}} \frac{||X\Delta_i||_2}{||\Delta_i||_2} \geq ||\Delta_i||_2 \, \rho_i'\right] \geq 1 - c_1 \exp(-c_2 \eta_i^2),$$

where $\rho_i' = \sqrt{N\mathcal{L}} - 2\sqrt{\mathcal{M}} - cw(\Theta) - \eta_i$. Then using the fact that $\rho_i = \rho_i' \delta_i,\ \rho^2 = \sum_{i=1}^p \rho_i^2,\ \sum_i^p \delta_i^2 = 1$ and setting $\eta_i = \eta$ for all $i = 1, \ldots, p$, we get

$$\rho^2 = \left[\sqrt{N\mathcal{L}} - 2\sqrt{\mathcal{M}} - cw(\Theta) - \eta\right]^2 \sum_{i=1}^p \delta_i^2 = \left[\sqrt{N\mathcal{L}} - 2\sqrt{\mathcal{M}} - cw(\Theta) - \eta\right]^2.$$

Taking the square root of the above and using (55) we finally get

$$\mathbb{P}\left[\inf_{\Delta \in \text{cone}(\Omega_E)} \frac{||(I_{p\times p} \otimes X)\Delta||_2}{||\Delta||_2} \geq \sqrt{N\mathcal{L}} - 2\sqrt{\mathcal{M}} - cw(\Theta) - \eta\right] \geq 1 - pc_1 \exp(-c_2 \eta^2). \tag{68}$$

Establishing bound on N.
Now setting $\eta = \varepsilon\sqrt{N\mathcal{L}}$ for $0 < \varepsilon < 1$, the right hand side of the inequality inside the probability statement in (68) must be equal to

$$\sqrt{\kappa N} = \sqrt{N\mathcal{L}} - 2\sqrt{\mathcal{M}} - cw(\Theta) - \varepsilon\sqrt{N\mathcal{L}} = \varepsilon'\sqrt{N\mathcal{L}} - 2\sqrt{\mathcal{M}} - cw(\Theta),$$

for some positive constant ε'. Since $\kappa N > 0$, it follows that we require

$$\varepsilon'\sqrt{N\mathcal{L}} > 2\sqrt{\mathcal{M}} + cw(\Theta),$$

or equivalently

$$\sqrt{N} > \frac{2\sqrt{\mathcal{M}} + cw(\Theta)}{\varepsilon'\sqrt{\mathcal{L}}} = \mathcal{O}(w(\Theta)).$$

This concludes our proof on establishing the restricted eigenvalue conditions.

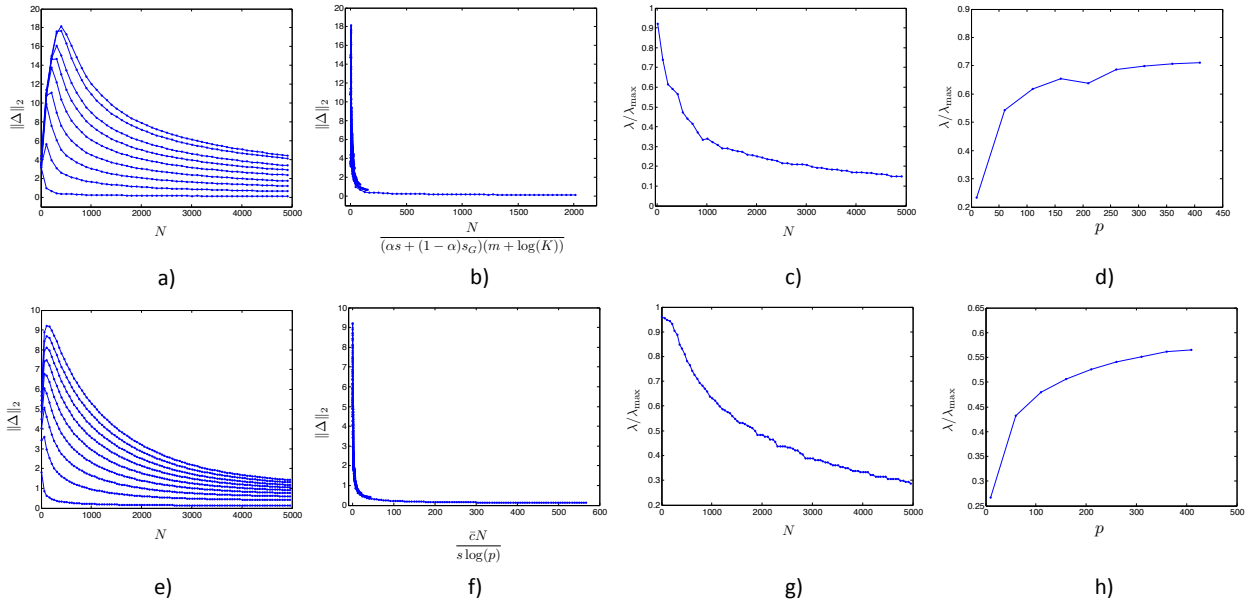

Figure 1: Results for estimating parameters of a stable first order Sparse Group Lasso VAR (top row) and OWL-regularized VAR (bottom row). Problem dimensions for Sparse Group Lasso : $p \in [10, 410]$, $N \in [10, 5000]$, $\frac{\lambda_N}{\lambda_{max}} \in [0, 1]$, $K \in [2, 60]$ and $d = 1$. Problem dimensions for OWL: $p \in [10, 410]$, $N \in [10, 5000]$, $\frac{\lambda_N}{\lambda_{max}} \in [0, 1]$, $s \in [4, 260]$ and $d = 1$. All the results are shown after averaging across 50 runs.

3 Experiments

In this Section we provide additional results on testing structured VAR estimation using synthetic data and additional details about the experimental setup for building the VAR model on real flight data.

3.1 Synthetic Data

Using synthetic data we present additional results on testing regularized VAR estimation under Sparse Group Lasso and OWL norms.

3.1.1 Sparse Group Lasso

To evaluate the estimation problem with Sparse Group Lasso norm, we constructed first-order VAR process for the following set of problem sizes $p \in [10, 400]$, $s \in [10, 200]$, $s_G \in [2, 20]$ and $N \in [10, 5000]$. The parameter α was set to 0.5. Results are shown in Figure 1, top row. Similarly as in main paper, we can see that the errors are scaled by $\frac{N}{(\alpha s+(1-\alpha)s_G)(m+\log(K))}$. Moreover, the λ_N parameter is decreasing when number of samples N increases. On the other hand, as the problem dimension p increases, the selected λ_N grows at the rate similar to $\sqrt{\log p}$.

3.1.2 OWL

To test the VAR estimation problem under OWL norm we constructed a first-order VAR process with $p \in [10, 410]$, $s \in [4, 260]$ and $N \in [10, 5000]$. The vector of weights c was set to be a monotonically decreasing sequence of numbers in the range $[1, 0)$. Figure 1, bottom row, shows the results. It can be seen from Figure 1-f that when the errors are plotted against $\frac{\bar{c}N}{s \log(p)}$, they become tightly aligned, confirming the bounds established in Section 3.3.4 in the main paper for the error norm. As shown in Figure 1-g,h the selected regularization parameter λ_N grows with the problem dimension p and decreases with the number of samples N.

1	Altitude
2	Corrected angle of attack
3	Brake temperature
4	Computed airspeed
5	Drift angle
6	Engine temperature
7	Low rotor speed
8	High rotor speed
9	Engine oil pressure
10	Engine oil quantity
11	Engine oil temperature
12	Engine pre-cooler outlet temperature
13	Fuel mass flow rate
14	Lateral acceleration
15	Longitudinal acceleration
16	Normal acceleration
17	Glide slope deviation
18	Ground speed
19	Localization deviation
20	Magnetic heading
21	Burner pressure
22	Pitch angle
23	Roll angle
24	HPC exit temperature
25	Angle magnitude
26	Angle true
27	Total fuel quantity
28	True heading
29	Vertical speed
30	True airspeed
31	MACH

Table 1: 31 features selected for structured VAR estimation on real flight data.

3.2 Real Flight Data

In Table 1 we show the list of parameters which were selected for building VAR model on the flight dataset in Section 4.2 of the main paper.

References

[1] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3):357–367, 1967.

[2] A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar. Estimation with norm regularization. In *Advances in Neural Information Processing Systems*, pages 1556–1564, 2014.

[3] J. Gutierrez-Gutierrez and P. M. Crespo. Block toeplitz matrices: asymptotic results and applications. *Foundations and Trends in Communications and Information Theory*, 8(3):179–257, 2011.

[4] R. A. Horn and C. R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[5] H. Lutkepohl. *New introduction to multiple time series analysis*. Springer, 2007.

[6] S. Negahban and M. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

[7] M. B. Priestley. *Spectral analysis and time series*. Academic press, 1981.

[8] M. Talagrand. Majorizing measures without measures. *Annals of probability*, pages 411–417, 2001.

[9] M. Talagrand. *The Generic Chaining*. Springer Monographs in Mathematics. Springer Berlin, 2005.

[10] M. Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer, 2006.

[11] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.