Selected Topics in Mathematics of Learning

High-Dimensional Statistics

Lecturer: Marius Yamakou

Winter Semester 2024/25
Department of Data Science, FAU

November 26, 2024

Part IV: Covariance Matrices

Objectives:

1. **Understand the Structure and Role of Covariance and Correlation Matrices**
   - Covariance and correlation matrices are essential tools for analyzing relationships between variables in datasets.

Part IV: Covariance Matrices

Objectives:

1. **Understand the Structure and Role of Covariance and Correlation Matrices**
   - Covariance and correlation matrices are essential tools for analyzing relationships between variables in datasets.
   - **Covariance**: Measures how two variables vary together. A positive covariance indicates that the variables increase together, while a negative covariance indicates they move in opposite directions.

Objectives:

1. **Understand the Structure and Role of Covariance and Correlation Matrices**
   - Covariance and correlation matrices are essential tools for analyzing relationships between variables in datasets.
   - **Covariance**: Measures how two variables vary together. A positive covariance indicates that the variables increase together, while a negative covariance indicates they move in opposite directions.
   - **Correlation**: Normalizes the covariance to a scale of -1 to 1, providing a dimensionless measure of the strength and direction of the linear relationship.

Part IV: Covariance Matrices

Objectives:

1. **Understand the Structure and Role of Covariance and Correlation Matrices**
    - Covariance and correlation matrices are essential tools for analyzing relationships between variables in datasets.
    - **Covariance**: Measures how two variables vary together. A positive covariance indicates that the variables increase together, while a negative covariance indicates they move in opposite directions.
    - **Correlation**: Normalizes the covariance to a scale of -1 to 1, providing a dimensionless measure of the strength and direction of the linear relationship.
2. **Explore the Concept of Sparsity**
    - **Sparsity**: Refers to matrices with many zero (or near-zero) entries, indicating weak or no relationships between many pairs of variables.

Part IV: Covariance Matrices

Objectives:

1. **Understand the Structure and Role of Covariance and Correlation Matrices**
   - Covariance and correlation matrices are essential tools for analyzing relationships between variables in datasets.
   - **Covariance**: Measures how two variables vary together. A positive covariance indicates that the variables increase together, while a negative covariance indicates they move in opposite directions.
   - **Correlation**: Normalizes the covariance to a scale of -1 to 1, providing a dimensionless measure of the strength and direction of the linear relationship.
2. **Explore the Concept of Sparsity**
   - **Sparsity**: Refers to matrices with many zero (or near-zero) entries, indicating weak or no relationships between many pairs of variables.
   - Importance: Sparse covariance matrices simplify data interpretation and are critical for efficient modeling in high-dimensional settings.

3. **Learn the Thresholding Method**
   - Thresholding is a technique to induce sparsity in covariance matrices by setting small values to zero.

3. **Learn the Thresholding Method**
   - Thresholding is a technique to induce sparsity in covariance matrices by setting small values to zero.
   - Purpose: Helps reduce noise in the data and improves the interpretability of relationships between variables.
4. **Examine Strategies for Selecting the Regularization Parameter**
   - The regularization parameter determines the threshold for sparsity in covariance matrices.

3. **Learn the Thresholding Method**
   - Thresholding is a technique to induce sparsity in covariance matrices by setting small values to zero.
   - Purpose: Helps reduce noise in the data and improves the interpretability of relationships between variables.
4. **Examine Strategies for Selecting the Regularization Parameter**
   - The regularization parameter determines the threshold for sparsity in covariance matrices.
   - **Balance**: Choosing an appropriate value is crucial to balance sparsity with retaining significant relationships.

3. **Learn the Thresholding Method**
   - Thresholding is a technique to induce sparsity in covariance matrices by setting small values to zero.
   - Purpose: Helps reduce noise in the data and improves the interpretability of relationships between variables.
4. **Examine Strategies for Selecting the Regularization Parameter**
   - The regularization parameter determines the threshold for sparsity in covariance matrices.
   - **Balance**: Choosing an appropriate value is crucial to balance sparsity with retaining significant relationships.
   - Methods: Include practical approaches like cross-validation and theoretical insights for optimal selection.

Outline

1. Covariance/Correlation matrix
2. Sparsity
3. Thresholding
4. How to choose the regularization parameter?

1. Covariance Matrix (Σ):

- $\Sigma = (\sigma_{ij})_{i,j=1,\dots,p}$, where each element $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$.
- Covariance measures how two variables vary together:

$$\mathrm{Cov}(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])].$$

1. Covariance Matrix (Σ):

- $\Sigma = (\sigma_{ij})_{i,j=1,\ldots,p}$, where each element $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$.
- Covariance measures how two variables vary together:

$$\mathrm{Cov}(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])].$$

2. Correlation Matrix (R):

- $R = (\rho_{ij})_{i,j=1,\ldots,p}$, where ρ_{ij} is the normalized covariance:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

- Correlation is standardized and lies in $[-1, 1]$.
- Diagonal entries are always $\rho_{ii} = 1$.

1. Sample Covariance Matrix: $\hat{\Sigma} = (\hat{\sigma}_{ij})_{i,j=1,\dots,p}$

- An empirical estimate based on data:

$$\hat{\sigma}_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(X_{k,i} - \bar{X}_i)(X_{k,j} - \bar{X}_j),$$

where $\bar{X}_i = \frac{1}{n}\sum_{k=1}^{n} X_{k,i}$ is the sample mean of variable X_i.

1. Covariance and Correlation Matrices: Sample Analogues

1. Sample Covariance Matrix: $\hat{\Sigma} = (\hat{\sigma}_{ij})_{i,j=1,\ldots,p}$

- An empirical estimate based on data:

$$\hat{\sigma}_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (X_{k,i} - \bar{X}_i)(X_{k,j} - \bar{X}_j),$$

where $\bar{X}_i = \frac{1}{n} \sum_{k=1}^{n} X_{k,i}$ is the sample mean of variable X_i.

2. Sample Correlation Matrix: $\hat{R} = (\hat{\rho}_{ij})_{i,j=1,\ldots,p}$

- Normalized version of $\hat{\Sigma}$:

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}.$$

1. Covariance and Correlation Matrices: Sample Analogues

1. Sample Covariance Matrix: $\hat{\Sigma} = (\hat{\sigma}_{ij})_{i,j=1,\ldots,p}$

- An empirical estimate based on data:

$$\hat{\sigma}_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (X_{k,i} - \bar{X}_i)(X_{k,j} - \bar{X}_j),$$

where $\bar{X}_i = \frac{1}{n} \sum_{k=1}^{n} X_{k,i}$ is the sample mean of variable X_i.

2. Sample Correlation Matrix: $\hat{R} = (\hat{\rho}_{ij})_{i,j=1,\ldots,p}$

- Normalized version of $\hat{\Sigma}$:

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}.$$

Note:

- The denominator uses $n-1$ (instead of n) to correct for bias, ensuring an unbiased estimate of population covariance.

2. Covariance Matrix, Sparsity, and Adjacency Matrix

Covariance Matrix: $\Sigma = (\sigma_{ij})_{i,j=1,\ldots,p}$:

- If $\sigma_{ij} = 0$, the variables $X_{k,i}$ and $X_{k,j}$ are **uncorrelated**.
- This means their covariance is zero, indicating no **linear relationship** between these variables.

2. Covariance Matrix, Sparsity, and Adjacency Matrix

Covariance Matrix: $\Sigma = (\sigma_{ij})_{i,j=1,\ldots,p}$:

- If $\sigma_{ij} = 0$, the variables $X_{k,i}$ and $X_{k,j}$ are **uncorrelated**.
- This means their covariance is zero, indicating no **linear relationship** between these variables.

Sparsity:

- Sparsity measures how many elements in Σ are zero (or close to zero).
- Sparsity of Σ can be quantified using different measures.
- Sparse covariance matrices are useful in **high-dimensional statistics** because they simplify the structure of relationships between variables.

2. Covariance Matrix, Sparsity, and Adjacency Matrix

Adjacency Matrix: A:

- Represents the sparsity pattern of Σ as a binary matrix.
- $A = (A_{ij})_{i,j=1,\ldots,p}, \quad A_{ij} = \mathbb{1}_{\{\sigma_{ij} \neq 0\}}$.

Adjacency Matrix: A:

- Represents the sparsity pattern of Σ as a binary matrix.
- $A = (A_{ij})_{i,j=1,\dots,p}, \quad A_{ij} = \mathbb{1}_{\{\sigma_{ij} \neq 0\}}$.
- Visualizes variable connections (non-zero covariances).

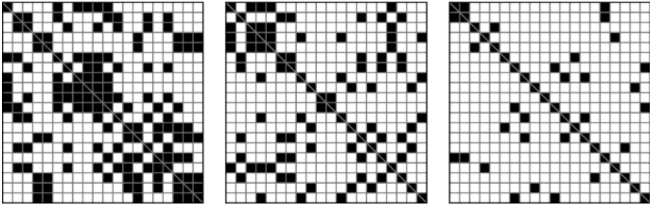

Figure: Examples of adjacency matrices with varying sparsity. **Left:** Dense matrix: Many non-zero elements (less sparsity). **Middle:** Moderately sparse matrix: Fewer non-zero elements. **Right:** Highly sparse matrix:

2. Quantifying Sparsity using Mathematical Tools

Operator Norm: $\|A\|$

- The operator norm of the adjacency matrix (A) is a natural measure of sparsity.
- It provides an upper bound on the "strength" or "density" of connections represented by A.

2. Quantifying Sparsity using Mathematical Tools

Operator Norm: $\|A\|$

- The operator norm of the adjacency matrix (A) is a natural measure of sparsity.
- It provides an upper bound on the "strength" or "density" of connections represented by A.

Key Inequalities:

- $\|A\| \leq d$, where d is the maximum degree of the graph represented by the adjacency matrix A.
  - The degree of a node in the graph corresponds to the number of variables it is directly connected to.

2. Quantifying Sparsity using Mathematical Tools

Operator Norm: $\|A\|$

- The operator norm of the adjacency matrix (A) is a natural measure of sparsity.
- It provides an upper bound on the "strength" or "density" of connections represented by A.

Key Inequalities:

- $\|A\| \leq d$, where d is the maximum degree of the graph represented by the adjacency matrix A.
  - The degree of a node in the graph corresponds to the number of variables it is directly connected to.
- $\|A\| \leq s$, if Σ (the covariance matrix) has at most s non-zero entries per row.
  - This is an alternative measure of sparsity, focusing on the number of significant relationships per variable.

2. Quantifying Sparsity using Mathematical Tools

Operator Norm: $\|A\|$

- The operator norm of the adjacency matrix (A) is a natural measure of sparsity.
- It provides an upper bound on the "strength" or "density" of connections represented by A.

Key Inequalities:

- $\|A\| \leq d$, where d is the maximum degree of the graph represented by the adjacency matrix A.
  - The degree of a node in the graph corresponds to the number of variables it is directly connected to.
- $\|A\| \leq s$, if Σ (the covariance matrix) has at most s non-zero entries per row.
  - This is an alternative measure of sparsity, focusing on the number of significant relationships per variable.

Example:

- For $p = 3$, consider the covariance matrix: $\Sigma = \begin{bmatrix} 2 & 0 & 0.5 \\ 0 & 1 & 0 \\ 0.5 & 0 & 3 \end{bmatrix}$
- The corresponding adjacency matrix is: $A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$
- The non-zero elements of A reflect the non-zero entries in Σ, showing the connections between variables.

2. Definition: Operator Norm of a Matrix

How to quantify sparsity ?

Mathematical Definition:

- The operator norm of a matrix $A \in \mathbb{R}^{p \times p}$ is defined as:

$$\|A\| = \sup_{\|x\|_2=1} \|Ax\|_2,$$

where:

- $\|x\|_2 = \sqrt{\sum_{i=1}^{p} x_i^2}$ is the Euclidean (or ℓ_2) norm of x,
- $\|Ax\|_2$ is the Euclidean norm of the transformed vector Ax.

2. Definition: Operator Norm of a Matrix

How to quantify sparsity ?

Mathematical Definition:

- The operator norm of a matrix $A \in \mathbb{R}^{p \times p}$ is defined as:

$$\|A\| = \sup_{\|x\|_2=1} \|Ax\|_2,$$

  where:
  - $\|x\|_2 = \sqrt{\sum_{i=1}^{p} x_i^2}$ is the Euclidean (or ℓ_2) norm of x,
  - $\|Ax\|_2$ is the Euclidean norm of the transformed vector Ax.

Key Intuition:

- The operator norm measures the maximum "stretch" that A applies to any vector of **unit** length.

2. Definition: Operator Norm of a Matrix

How to quantify sparsity ?

Mathematical Definition:

- The operator norm of a matrix $A \in \mathbb{R}^{p \times p}$ is defined as:

$$\|A\| = \sup_{\|x\|_2 = 1} \|Ax\|_2,$$

  where:
  - $\|x\|_2 = \sqrt{\sum_{i=1}^{p} x_i^2}$ is the Euclidean (or ℓ_2) norm of x,
  - $\|Ax\|_2$ is the Euclidean norm of the transformed vector Ax.

Key Intuition:

- The operator norm measures the maximum "stretch" that A applies to any vector of **unit** length.

Special Cases:

- For symmetric matrices (e.g., covariance matrices): $\|A\| = \max_{i=1,\ldots,p} |\lambda_i|$, where λ_i are the eigenvalues of A. (In this case, we also talk of the spectral norm of A.)
- For adjacency matrices: $\|A\| \leq d$, where d is the maximum degree of the graph.

3. Thresholded Covariance Estimator

Theoretical Perspective

Thresholding is a method to induce sparsity in covariance matrices by setting small values to zero.

Thresholding is a method to induce sparsity in covariance matrices by setting small values to zero.

Hard-Thresholding Operator: Theoretical Perspective

- The operator $T_\lambda(u)$ is defined as:

$$T_\lambda(u) = u\,\mathbb{1}_{\{|u|>\lambda\}} = \begin{cases} u, & \text{if } |u| > \lambda, \\ 0, & \text{if } |u| \leq \lambda. \end{cases}$$

  where $\lambda > 0$ is the threshold parameter.

3. Thresholded Covariance Estimator

Theoretical Perspective

Thresholding is a method to induce sparsity in covariance matrices by setting small values to zero.

Hard-Thresholding Operator: Theoretical Perspective

- The operator $T_\lambda(u)$ is defined as:

$$T_\lambda(u) = u\,\mathbb{1}_{\{|u|>\lambda\}} = \begin{cases} u, & \text{if } |u| > \lambda, \\ 0, & \text{if } |u| \leq \lambda. \end{cases}$$

  where $\lambda > 0$ is the threshold parameter.

Purpose:

- Removes weak relationships (small $|u|$) from the covariance matrix to make it sparse.

Key Properties of T_λ:

- **Preserves symmetry:**

$$\text{If } M_{ij} = M_{ji}, \quad \text{then}$$

$$\mathsf{T}_\lambda(M_{ij}) = M_{ij}\, \mathbb{1}_{\{|M_{ij}|>\lambda\}} = M_{ji}\, \mathbb{1}_{\{|M_{ji}|>\lambda\}} = T_\lambda(M_{ji}),$$

Key Properties of T_λ:

- **Preserves symmetry:**

$$\text{If } M_{ij} = M_{ji}, \quad \text{then}$$

$$\mathsf{T}_\lambda(M_{ij}) = M_{ij}\,\mathbb{1}_{\{|M_{ij}|>\lambda\}} = M_{ji}\,\mathbb{1}_{\{|M_{ji}|>\lambda\}} = T_\lambda(M_{ji}),$$

- **Invariance under permutations:**
  - Reordering variable labels does not affect the thresholding process.

Key Properties of T_λ:

- **Preserves symmetry:**

$$\text{If } M_{ij} = M_{ji}, \quad \text{then}$$

$$\mathsf{T}_\lambda(M_{ij}) = M_{ij}\,\mathbb{1}_{\{|M_{ij}|>\lambda\}} = M_{ji}\,\mathbb{1}_{\{|M_{ji}|>\lambda\}} = T_\lambda(M_{ji}),$$

- **Invariance under permutations:**
  - Reordering variable labels does not affect the thresholding process.
- **Does not necessarily preserve positive definiteness:**
  - A positive definite matrix has all positive eigenvalues, but applying T_λ may break this property.

3. Thresholded Covariance Estimator

Theoretical Perspective

Challenge: Hard-thresholding can result in a covariance matrix that is no longer positive definite, which is problematic in many applications.

3. Thresholded Covariance Estimator

Theoretical Perspective

Challenge: Hard-thresholding can result in a covariance matrix that is no longer positive definite, which is problematic in many applications.

Solution:

- If the operator norm satisfies:

$$\|T_\lambda(A) - A\| \leq \epsilon$$

  and the smallest eigenvalue $\lambda_{\text{min}}(A) > \epsilon$.

Challenge: Hard-thresholding can result in a covariance matrix that is no longer positive definite, which is problematic in many applications.

Solution:

- If the operator norm satisfies:

$$\|T_\lambda(A) - A\| \leq \epsilon$$

  and the smallest eigenvalue $\lambda_{\text{min}}(A) > \epsilon$.
- Then $T_\lambda(A)$ remains positive definite because:

$$\lambda_{\text{min}}(T_\lambda(A)) \geq \lambda_{\text{min}}(A) - \epsilon > 0.$$

3. Thresholded Covariance Estimator

Theoretical Perspective

Suppose we observe $X_1, \ldots, X_n$, i.i.d. p-variate random variables with mean 0 and covariance matrix Σ. Set $X_j = (X_{j,1}, \ldots, X_{j,p})^\top$. Suppose that each component $X_{i,j}$ is sub-Gaussian with parameter 1.

3. Thresholded Covariance Estimator

Theoretical Perspective

Suppose we observe $X_1, \ldots, X_n$, i.i.d. p-variate random variables with mean 0 and covariance matrix Σ. Set $X_j = (X_{j,1}, \ldots, X_{j,p})^\top$. Suppose that each component $X_{i,j}$ is sub-Gaussian with parameter 1.

Lemma

If $n > \log(p)$, then for any $\delta > 0$, the thresholded sample covariance matrix $T_{\lambda_n}(\widehat{\Sigma})$ with

$$\lambda_n = 8\sqrt{\frac{\log(p)}{n}} + \delta$$

satisfies

$$\mathbb{P}\left(\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\| > 2\|A\|\lambda_n\right) \leq 8\exp\left(-\frac{n}{16}\min\{\delta, \delta^2\}\right),$$

where A is the adjacency matrix of Σ and $\|A\|$ operator norm of A.

3. Thresholded Covariance Estimator

Theoretical Perspective

Suppose we observe $X_1, \dots, X_n$, i.i.d. p-variate random variables with mean 0 and covariance matrix Σ. Set $X_j = (X_{j,1}, \dots, X_{j,p})^\top$. Suppose that each component $X_{i,j}$ is sub-Gaussian with parameter 1.

Lemma

If $n > \log(p)$, then for any $\delta > 0$, the thresholded sample covariance matrix $T_{\lambda_n}(\widehat{\Sigma})$ with

$$\lambda_n = 8\sqrt{\frac{\log(p)}{n}} + \delta$$

satisfies

$$\mathbb{P}\left(\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\| > 2\|A\|\lambda_n\right) \leq 8\exp\left(-\frac{n}{16}\min\{\delta, \delta^2\}\right),$$

where A is the adjacency matrix of Σ and $\|A\|$ operator norm of A.

Interpretation: For large n, the probability of $T_{\lambda_n}(\hat{\Sigma})$ deviating significantly from Σ becomes very small.
The bound demonstrates that thresholded covariance estimators perform well in high-dimensional settings under sub-Gaussian assumptions.

Proof:

1. Show that $\|\widehat{\Sigma} - \Sigma\|_{\mathsf{max}} \leq \lambda_n$,
2. Use 1 to show $\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\| \leq 2\|A\|\lambda_n$.

Step 1: Bounding $\|\widehat{\Sigma} - \Sigma\|_{\mathsf{max}}$

$$\|\widehat{\Sigma} - \Sigma\|_{\mathsf{max}} = \max_{i,j} |\widehat{\Sigma}_{ij} - \sigma_{ij}|.$$

Proof:

1. Show that $\|\widehat{\Sigma} - \Sigma\|_{\max} \leq \lambda_n$,
2. Use 1 to show $\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\| \leq 2\|A\|\lambda_n$.

Step 1: Bounding $\|\widehat{\Sigma} - \Sigma\|_{\max}$

$$\|\widehat{\Sigma} - \Sigma\|_{\max} = \max_{i,j} |\widehat{\Sigma}_{ij} - \sigma_{ij}|.$$

Using union bound and concentration inequalities for sub-Gaussian variables, we have:

$$\mathbb{P}\left(\|\widehat{\Sigma} - \Sigma\|_{\max} > \lambda_n\right) = \mathbb{P}\left(\max_{i,j} |\widehat{\Sigma}_{ij} - \sigma_{ij}| > \lambda_n\right) = \mathbb{P}\left(\bigcup_{i,j} \{|\widehat{\Sigma}_{ij} - \sigma_{ij}| > \lambda_n\}\right)$$
$$\leq \sum_{i,j=1}^{p} \mathbb{P}\left(|\widehat{\Sigma}_{ij} - \sigma_{ij}| > \lambda_n\right).$$

Proof:

1. Show that $\|\widehat{\Sigma} - \Sigma\|_{\text{max}} \leq \lambda_n$,
2. Use 1 to show $\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\| \leq 2\|A\|\lambda_n$.

Step 1: Bounding $\|\widehat{\Sigma} - \Sigma\|_{\text{max}}$

$$\|\widehat{\Sigma} - \Sigma\|_{\text{max}} = \max_{i,j} |\widehat{\Sigma}_{ij} - \sigma_{ij}|.$$

Using union bound and concentration inequalities for sub-Gaussian variables, we have:

$$\mathbb{P}\left(\|\widehat{\Sigma} - \Sigma\|_{\text{max}} > \lambda_n\right) = \mathbb{P}\left(\max_{i,j} |\widehat{\Sigma}_{ij} - \sigma_{ij}| > \lambda_n\right) = \mathbb{P}\left(\bigcup_{i,j} \{|\widehat{\Sigma}_{ij} - \sigma_{ij}| > \lambda_n\}\right)$$
$$\leq \sum_{i,j=1}^{p} \mathbb{P}\left(|\widehat{\Sigma}_{ij} - \sigma_{ij}| > \lambda_n\right).$$

For sub-Gaussian variables: $\mathbb{P}\left(|\widehat{\Sigma}_{ij} - \sigma_{ij}| > \lambda_n\right) \leq 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right)$. Thus:
$\|\widehat{\Sigma} - \Sigma\|_{\text{max}} = \max_{i,j} |\widehat{\Sigma}_{ij} - \sigma_{ij}| \leq \lambda_n$ with high probability.

Step 2: Thresholding the Covariance Matrix

Consider the decomposition: $|T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \sigma_{ij}| = |T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij} + \widehat{\Sigma}_{ij} - \sigma_{ij}|$.

Step 2: Thresholding the Covariance Matrix

Consider the decomposition: $|T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \sigma_{ij}| = |T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij} + \widehat{\Sigma}_{ij} - \sigma_{ij}|$.

Using the triangle inequality: $|T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \sigma_{ij}| \leq |T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij}| + |\widehat{\Sigma}_{ij} - \sigma_{ij}|$.

3. Thresholded Covariance Estimator

Theoretical Perspective

Step 2: Thresholding the Covariance Matrix

Consider the decomposition: $|T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \sigma_{ij}| = |T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij} + \widehat{\Sigma}_{ij} - \sigma_{ij}|$.

Using the triangle inequality: $|T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \sigma_{ij}| \leq |T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij}| + |\widehat{\Sigma}_{ij} - \sigma_{ij}|$.

Case 1: $\sigma_{ij} = 0$: $\quad T_{\lambda_n}(\widehat{\Sigma}_{ij}) = \widehat{\Sigma}_{ij} \cdot \mathbb{1}_{\{|\widehat{\Sigma}_{ij}| > \lambda_n\}} = 0.$

Case 2: $\sigma_{ij} \neq 0$: $|T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij}| = |\widehat{\Sigma}_{ij}\, \mathbb{1}_{\{|\widehat{\Sigma}_{ij}| > \lambda_n\}} - \widehat{\Sigma}_{ij}|$

$$\implies |T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij}| = \begin{cases} |\widehat{\Sigma}_{ij} - \widehat{\Sigma}_{ij}|, & \text{if } |\widehat{\Sigma}_{ij}| > \lambda_n, \\ |0 - \widehat{\Sigma}_{ij}|, & \text{if } |\widehat{\Sigma}_{ij}| \leq \lambda_n, \end{cases} = \begin{cases} 0, & \text{if } |\widehat{\Sigma}_{ij}| > \lambda_n, \\ |\widehat{\Sigma}_{ij}|, & \text{if } |\widehat{\Sigma}_{ij}| \leq \lambda_n, \end{cases}$$

$$\implies |T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij}| \leq \lambda_n.$$

Step 2: Thresholding the Covariance Matrix

Consider the decomposition: $|T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \sigma_{ij}| = |T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij} + \widehat{\Sigma}_{ij} - \sigma_{ij}|$.

Using the triangle inequality: $|T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \sigma_{ij}| \leq |T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij}| + |\widehat{\Sigma}_{ij} - \sigma_{ij}|$.

Case 1: $\sigma_{ij} = 0$: $\quad T_{\lambda_n}(\widehat{\Sigma}_{ij}) = \widehat{\Sigma}_{ij} \cdot \mathbb{1}_{\{|\widehat{\Sigma}_{ij}| > \lambda_n\}} = 0.$

Case 2: $\sigma_{ij} \neq 0$: $|T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij}| = |\widehat{\Sigma}_{ij}\, \mathbb{1}_{\{|\widehat{\Sigma}_{ij}| > \lambda_n\}} - \widehat{\Sigma}_{ij}|$

$$\Longrightarrow \; |T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij}| = \begin{cases} |\widehat{\Sigma}_{ij} - \widehat{\Sigma}_{ij}|, & \text{if } |\widehat{\Sigma}_{ij}| > \lambda_n, \\ |0 - \widehat{\Sigma}_{ij}|, & \text{if } |\widehat{\Sigma}_{ij}| \leq \lambda_n, \end{cases} = \begin{cases} 0, & \text{if } |\widehat{\Sigma}_{ij}| > \lambda_n, \\ |\widehat{\Sigma}_{ij}|, & \text{if } |\widehat{\Sigma}_{ij}| \leq \lambda_n, \end{cases}$$

$$\Longrightarrow \; |T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij}| \leq \lambda_n.$$

Combining both cases we have: $|T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \sigma_{ij}| \leq 2\lambda_n A_{ij}$,
where the adjacency matrix $A_{ij} = \mathbb{1}_{\{\sigma_{ij} \neq 0\}}$ ensures that this bound is tight.

3. Thresholded Covariance Estimator

Theoretical Perspective

Hence, the operator norm bound: $\|T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij}\| \leq 2\|A\|\lambda_n$ follows directly from the element-wise bound.

3. Thresholded Covariance Estimator

Theoretical Perspective

Hence, the operator norm bound: $\|T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij}\| \leq 2\|A\|\lambda_n$ follows directly from the element-wise bound.

$$\mathbb{P}\left(\|\widehat{\Sigma} - \Sigma\|_{\mathsf{max}} > \lambda_n\right) \leq \sum_{i,j=1}^{p} \mathbb{P}\left(|\widehat{\Sigma}_{ij} - \sigma_{ij}| > \lambda_n\right) \quad (*)$$

(i) $i = j$: $\quad \mathbb{P}\left(|\widehat{\Sigma}_{ii} - \sigma_{ii}| > \lambda_n\right) \leq 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right)$

Hence, the operator norm bound: $\|T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij}\| \leq 2\|A\|\lambda_n$ follows directly from the element-wise bound.

$$\mathbb{P}\left(\|\widehat{\Sigma} - \Sigma\|_{\mathsf{max}} > \lambda_n\right) \leq \sum_{i,j=1}^{p} \mathbb{P}\left(|\widehat{\Sigma}_{ij} - \sigma_{ij}| > \lambda_n\right) \quad (*)$$

(i) $i = j$: $\mathbb{P}\left(|\widehat{\Sigma}_{ii} - \sigma_{ii}| > \lambda_n\right) \leq 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right)$

(ii) $i \neq j$: $2(\widehat{\Sigma}_{ij} - \sigma_{ij}) = \frac{2}{n}\sum_{k=1}^{n} X_{ki}X_{kj} - 2\sigma_{ij}$

3. Thresholded Covariance Estimator

Theoretical Perspective

Hence, the operator norm bound: $\|T_{\lambda_n}(\widehat{\Sigma}_{ij}) - \widehat{\Sigma}_{ij}\| \leq 2\|A\|\lambda_n$ follows directly from the element-wise bound.

$$\mathbb{P}\left(\|\widehat{\Sigma} - \Sigma\|_{\mathsf{max}} > \lambda_n\right) \leq \sum_{i,j=1}^{p} \mathbb{P}\left(|\widehat{\Sigma}_{ij} - \sigma_{ij}| > \lambda_n\right) \qquad (*)$$

(i) $i = j$: $\mathbb{P}\left(|\widehat{\Sigma}_{ii} - \sigma_{ii}| > \lambda_n\right) \leq 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right)$

(ii) $i \neq j$: $2(\widehat{\Sigma}_{ij} - \sigma_{ij}) = \frac{2}{n}\sum_{k=1}^{n} X_{ki}X_{kj} - 2\sigma_{ij}$

$$= \frac{1}{n}\sum_{k=1}^{n} X_{ki}^2 + \frac{2}{n}\sum_{k=1}^{n} X_{ki}X_{kj} + \frac{1}{n}\sum_{k=1}^{n} X_{kj}^2 - \frac{1}{n}\sum_{k=1}^{n} X_{ki}^2 - \frac{1}{n}\sum_{k=1}^{n} X_{kj}^2$$

$$+ \sigma_{ii} + \sigma_{jj} - \sigma_{ii} - \sigma_{jj} - 2\sigma_{ij}$$

3. Thresholded covariance estimator

Theoretical perspective

$$= \left[\frac{1}{n} \sum_{k=1}^{n} (X_{ki} + X_{kj})^2 - (\sigma_{ii} + \sigma_{jj} + 2\sigma_{ij}) \right] - \left[\left(\frac{1}{n} \sum_{k=1}^{n} X_{ki}^2 - \sigma_{ii} \right) \right]$$

$$- \left[\left(\frac{1}{n} \sum_{k=1}^{n} X_{kj}^2 - \sigma_{jj} \right) \right] \quad (**)$$

3. Thresholded covariance estimator

Theoretical perspective

$= \Big[\frac{1}{n}\sum_{k=1}^{n}(X_{ki}+X_{kj})^2 - (\sigma_{ii}+\sigma_{jj}+2\sigma_{ij})\Big] - \Big[\Big(\frac{1}{n}\sum_{k=1}^{n}X_{ki}^2 - \sigma_{ii}\Big)\Big]$

- $\Big[\Big(\frac{1}{n}\sum_{k=1}^{n}X_{kj}^2 - \sigma_{jj}\Big)\Big] \quad (**)$

By assumption: X$_{ki} \sim$ subG(1) and $X_{kj} \sim$ subG(1), it follows that:
$X_{ki}+X_{kj} \sim$ subG(2) and $(X_{ki}+X_{kj})^2 \sim$ subExp$(4, 16)$

Continuing from $(*)$, we have:

3. Thresholded covariance estimator

Theoretical perspective

$= \left[\frac{1}{n}\sum_{k=1}^{n}(X_{ki}+X_{kj})^2 - (\sigma_{ii}+\sigma_{jj}+2\sigma_{ij})\right] - \left[\left(\frac{1}{n}\sum_{k=1}^{n}X_{ki}^2 - \sigma_{ii}\right)\right]$

- $\left[\left(\frac{1}{n}\sum_{k=1}^{n}X_{kj}^2 - \sigma_{jj}\right)\right] \quad (**)$

By assumption: $X_{ki} \sim \text{subG}(1)$ and $X_{kj} \sim \text{subG}(1)$, it follows that:
$X_{ki}+X_{kj} \sim \text{subG}(2)$ and $(X_{ki}+X_{kj})^2 \sim \text{subExp}(4,16)$

Continuing from $(*)$, we have:

$$\mathbb{P}\left(\|\widehat{\Sigma}-\Sigma\|_{\max} > \lambda_n\right) \leq \sum_{i=1}^{p}\mathbb{P}\left(|\widehat{\Sigma}_{ii}-\sigma_{ii}| > \lambda_n\right) + \sum_{i\neq j}^{p}\mathbb{P}\left(|\widehat{\Sigma}_{ij}-\sigma_{ij}| > \lambda_n\right)$$

3. Thresholded covariance estimator

Theoretical perspective

$= \left[\frac{1}{n}\sum_{k=1}^{n}(X_{ki}+X_{kj})^2 - (\sigma_{ii}+\sigma_{jj}+2\sigma_{ij})\right] - \left[\left(\frac{1}{n}\sum_{k=1}^{n}X_{ki}^2 - \sigma_{ii}\right)\right]$
- $\left[\left(\frac{1}{n}\sum_{k=1}^{n}X_{kj}^2 - \sigma_{jj}\right)\right] \quad (**)$

By assumption: $X_{ki} \sim$ subG(1) and $X_{kj} \sim$ subG(1), it follows that:
$X_{ki}+X_{kj} \sim$ subG(2) and $(X_{ki}+X_{kj})^2 \sim$ subExp$(4, 16)$

Continuing from $(*)$, we have:

$\mathbb{P}\left(\|\widehat{\Sigma}-\Sigma\|_{\max} > \lambda_n\right) \le \sum_{i=1}^{p}\mathbb{P}\left(|\widehat{\Sigma}_{ii}-\sigma_{ii}| > \lambda_n\right) + \sum_{i\neq j}^{p}\mathbb{P}\left(|\widehat{\Sigma}_{ij}-\sigma_{ij}| > \lambda_n\right)$

$\overset{(**)}{=} \sum_{i=1}^{p}\mathbb{P}\left(|\widehat{\Sigma}_{ii}-\sigma_{ii}| > \lambda_n\right) + \sum_{i\neq j}^{p}\mathbb{P}\Big(\Big|\left[\frac{1}{n}\sum_{k=1}^{n}(X_{ki}+X_{kj})^2\right.$
- $(\sigma_{ii}+\sigma_{jj}+2\sigma_{ij}) - \left[\left(\frac{1}{n}\sum_{k=1}^{n}X_{ki}^2 - \sigma_{ii}\right)\right] - \left[\left(\frac{1}{n}\sum_{k=1}^{n}X_{kj}^2 - \sigma_{jj}\right)\right]\Big| > \lambda_n\Big)$

$= \left[\frac{1}{n}\sum_{k=1}^{n}(X_{ki}+X_{kj})^2 - (\sigma_{ii}+\sigma_{jj}+2\sigma_{ij})\right] - \left[\left(\frac{1}{n}\sum_{k=1}^{n}X_{ki}^2 - \sigma_{ii}\right)\right]$
- $\left[\left(\frac{1}{n}\sum_{k=1}^{n}X_{kj}^2 - \sigma_{jj}\right)\right] \quad (**)$

By assumption: $X_{ki} \sim$ subG(1) and $X_{kj} \sim$ subG(1), it follows that:
$X_{ki}+X_{kj} \sim$ subG(2) and $(X_{ki}+X_{kj})^2 \sim$ subExp$(4,16)$

Continuing from $(*)$, we have:

$\mathbb{P}\left(\|\widehat{\Sigma}-\Sigma\|_{\max} > \lambda_n\right) \le \sum_{i=1}^{p}\mathbb{P}\left(|\widehat{\Sigma}_{ii}-\sigma_{ii}| > \lambda_n\right) + \sum_{i\neq j}^{p}\mathbb{P}\left(|\widehat{\Sigma}_{ij}-\sigma_{ij}| > \lambda_n\right)$

$\overset{(**)}{=} \sum_{i=1}^{p}\mathbb{P}\left(|\widehat{\Sigma}_{ii}-\sigma_{ii}| > \lambda_n\right) + \sum_{i\neq j}^{p}\mathbb{P}\Big(\Big|\left[\frac{1}{n}\sum_{k=1}^{n}(X_{ki}+X_{kj})^2\right.$
- $(\sigma_{ii}+\sigma_{jj}+2\sigma_{ij}) - \left[\left(\frac{1}{n}\sum_{k=1}^{n}X_{ki}^2-\sigma_{ii}\right)\right] - \left[\left(\frac{1}{n}\sum_{k=1}^{n}X_{kj}^2-\sigma_{jj}\right)\right]\Big| > \lambda_n\Big)$

$\le \sum_{i=1}^{p}\mathbb{P}\left(|\widehat{\Sigma}_{ii}-\sigma_{ii}| > \lambda_n\right)$
$+ \sum_{i\neq j}^{p}\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}(X_{ki}+X_{kj})^2 - (\sigma_{ii}+\sigma_{jj}+2\sigma_{ij})\right| > \frac{\lambda_n}{3}\right)$
$+ \sum_{i\neq j}^{p}\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}X_{ki}^2-\sigma_{ii}\right| > \frac{\lambda_n}{3}\right)$
$+ \sum_{i\neq j}^{p}\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}X_{kj}^2-\sigma_{jj}\right| > \frac{\lambda_n}{3}\right).$

$$\overset{(i)}{\leq} \sum_{i=1}^{p} 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right) + 3\sum_{i\neq j}^{p} 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right)$$

3. Thresholded Covariance Estimator

Theoretical perspective

$\overset{(i)}{\leq} \sum_{i=1}^{p} 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right) + 3\sum_{i\neq j}^{p} 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right)$

Diagonal Terms ($i = j$): $\mathbb{P}\left(|\widehat{\Sigma}_{ii} - \sigma_{ii}| > \lambda_n\right) \leq 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right)$,
where $c > 0$.

3. Thresholded Covariance Estimator

Theoretical perspective

$\overset{(i)}{\leq} \sum_{i=1}^{p} 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right) + 3\sum_{i\neq j}^{p} 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right)$

Diagonal Terms ($i = j$): $\mathbb{P}\left(|\widehat{\Sigma}_{ii} - \sigma_{ii}| > \lambda_n\right) \leq 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right)$, where $c > 0$.

Since there are p diagonal terms:

$$\sum_{i=1}^{p} \mathbb{P}\left(|\widehat{\Sigma}_{ii} - \sigma_{ii}| > \lambda_n\right) \leq p \cdot 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right).$$

3. Thresholded Covariance Estimator

Theoretical perspective

$\overset{(i)}{\leq} \sum_{i=1}^{p} 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right) + 3\sum_{i\neq j}^{p} 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right)$

Diagonal Terms ($i = j$): $\mathbb{P}\left(|\widehat{\Sigma}_{ii} - \sigma_{ii}| > \lambda_n\right) \leq 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right)$,
where $c > 0$.
Since there are p diagonal terms:

$$\sum_{i=1}^{p} \mathbb{P}\left(|\widehat{\Sigma}_{ii} - \sigma_{ii}| > \lambda_n\right) \leq p \cdot 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right).$$

Off-Diagonal Terms ($i \neq j$): $\mathbb{P}\left(|\widehat{\Sigma}_{ij} - \sigma_{ij}| > \lambda_n\right) \leq 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right)$.
There are $\binom{p}{2} = \frac{p(p-1)}{2} \approx \frac{p^2}{2}$ (for large p) off-diagonal terms:

$$\sum_{i\neq j}^{p} \mathbb{P}\left(|\widehat{\Sigma}_{ij} - \sigma_{ij}| > \lambda_n\right) \leq \frac{p^2}{2} \cdot 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right).$$

Combining Diagonal and Off-Diagonal Contributions:
From the union bound:

$$\mathbb{P}\left(\|\widehat{\Sigma} - \Sigma\|_{\mathsf{max}} > \lambda_n\right) \leq p \cdot 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right) + p^2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right).$$

$$\leq p^2 c_1 \exp\left(-c_2 n \min\{\lambda_n, \lambda_n^2\}\right) \quad \text{for some } c_1, c_2 > 0$$

Combining Diagonal and Off-Diagonal Contributions:
From the union bound:

$$\mathbb{P}\left(\|\widehat{\Sigma}-\Sigma\|_{\mathsf{max}} > \lambda_n\right) \leq p \cdot 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right) + p^2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right).$$

$$\leq p^2 c_1 \exp\left(-c_2 n\min\{\lambda_n, \lambda_n^2\}\right) \quad \text{for some } c_1, c_2 > 0$$

With $p^2 = \exp(2\log(p))$

$$\mathbb{P}\left(\|\widehat{\Sigma}-\Sigma\|_{\mathsf{max}} > \lambda_n\right) \leq c_1\exp\left(-c_2 n\min\{\lambda_n, \lambda_n^2\} + 2\log(p)\right).$$

3. Thresholded Covariance Estimator

Theoretical perspective

Combining Diagonal and Off-Diagonal Contributions:
From the union bound:

$$\mathbb{P}\left(\|\widehat{\Sigma} - \Sigma\|_{\mathsf{max}} > \lambda_n\right) \leq p \cdot 2\exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right) + p^2 \exp\left(-cn\min\{\lambda_n, \lambda_n^2\}\right).$$

$$\leq p^2 c_1 \exp\left(-c_2 n \min\{\lambda_n, \lambda_n^2\}\right) \quad \text{for some } c_1, c_2 > 0$$

With $p^2 = \exp(2\log(p))$

$$\mathbb{P}\left(\|\widehat{\Sigma} - \Sigma\|_{\mathsf{max}} > \lambda_n\right) \leq c_1 \exp\left(-c_2 n \min\{\lambda_n, \lambda_n^2\} + 2\log(p)\right).$$

Interpretation:

- The $2\log(p)$ term arises from the quadratic growth in the number of terms (p^2).
- As p increases, the bound becomes weaker, reflecting the increasing probability of at least one large deviation.

Choose: $\lambda_n = 8\sqrt{\frac{\log(p)}{n}} + \delta.$

$$\min\{\lambda_n, \lambda_n^2\} = \min\left\{8\sqrt{\frac{\log(p)}{n}} + \delta, \left(8\sqrt{\frac{\log(p)}{n}} + \delta\right)^2\right\}.$$

For large n, δ dominates $\sqrt{\frac{\log(p)}{n}}$, so:

$$\min\{\lambda_n, \lambda_n^2\} \approx \min\{\delta, \delta^2\}.$$

Choose: $\lambda_n = 8\sqrt{\frac{\log(p)}{n}} + \delta.$

$$\min\{\lambda_n, \lambda_n^2\} = \min\left\{8\sqrt{\frac{\log(p)}{n}} + \delta, \left(8\sqrt{\frac{\log(p)}{n}} + \delta\right)^2\right\}.$$

For large n, δ dominates $\sqrt{\frac{\log(p)}{n}}$, so:

$$\min\{\lambda_n, \lambda_n^2\} \approx \min\{\delta, \delta^2\}.$$

Substituting into the probability bound:

$$\mathbb{P}\left(\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\| > 2\|A\|\lambda_n\right) \leq 8\exp\left(-\frac{n}{16}\min\{\delta, \delta^2\}\right).$$

$\square$

Note that the choice of $\delta = \sqrt{\frac{\log(p)}{n}}$ ensures a manageable trade-off between sparsity and accuracy in estimating the covariance matrix.

3. Thresholded Covariance Estimator

Theoretical perspective

Special Case: Choose $\delta = \sqrt{\frac{\log(p)}{n}}$.

$$\min\{\delta, \delta^2\} = \sqrt{\frac{\log(p)}{n}}.$$

Substitute:

$$\mathbb{P}\left(\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\| > 2\|A\|\lambda_n\right) \leq 8\exp\left(-\frac{n}{16} \cdot \sqrt{\frac{\log(p)}{n}}\right).$$

3. Thresholded Covariance Estimator

Theoretical perspective

Special Case: Choose $\delta = \sqrt{\frac{\log(p)}{n}}$.

$$\min\{\delta, \delta^2\} = \sqrt{\frac{\log(p)}{n}}.$$

Substitute:

$$\mathbb{P}\left(\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\| > 2\|A\|\lambda_n\right) \leq 8\exp\left(-\frac{n}{16} \cdot \sqrt{\frac{\log(p)}{n}}\right).$$

For $\log(p) < n$, simplify:

$$\mathbb{P}\left(\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\| > 2\|A\|\lambda_n\right) \leq 8p^{-\frac{1}{16}}.$$

Conclusion: The thresholded sample covariance estimator achieves:

$$\mathbb{P}\left(\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\| > 2\|A\|\lambda_n\right) \to 0 \quad \text{as } p \to \infty.$$

4. Cross Validation: Informal

How to choose λ in practice?

We use a variation of K-fold cross-validation. The steps are the following:

4. Cross Validation: Informal

How to choose λ in practice?

We use a variation of K-fold cross-validation. The steps are the following:

- Choose the number “splitting times" K.

4. Cross Validation: Informal

How to choose λ in practice?

We use a variation of K-fold cross-validation. The steps are the following:

- Choose the number “splitting times" K.
- Split the sample randomly into two pieces of size n_1 and n_2. (We will do that K times.)

4. Cross Validation: Informal

How to choose λ in practice?

We use a variation of K-fold cross-validation. The steps are the following:

- Choose the number "splitting times" K.
- Split the sample randomly into two pieces of size n_1 and n_2. (We will do that K times.)
- Possible choice which has been justified theoretically is $n_1 = n\left(1 - \frac{1}{\log(n)}\right)$, $n_2 = \frac{n}{\log(n)}$.

4. Cross Validation: Informal

How to choose λ in practice?

We use a variation of K-fold cross-validation. The steps are the following:

- Choose the number "splitting times" K.
- Split the sample randomly into two pieces of size n_1 and n_2. (We will do that K times.)
- Possible choice which has been justified theoretically is $n_1 = n\left(1 - \frac{1}{\log(n)}\right)$, $n_2 = \frac{n}{\log(n)}$.
- Define a grid of values for λ: $\{\lambda_1, \ldots, \lambda_M\}$

4. Cross Validation: Informal

How to choose λ in practice?

We use a variation of K-fold cross-validation. The steps are the following:

- Choose the number "splitting times" K.
- Split the sample randomly into two pieces of size n_1 and n_2. (We will do that K times.)
- Possible choice which has been justified theoretically is $n_1 = n\left(1 - \frac{1}{\log(n)}\right)$, $n_2 = \frac{n}{\log(n)}$.
- Define a grid of values for λ: $\{\lambda_1, \ldots, \lambda_M\}$
- For each λ calculate the validation Mean Squared Error (MSE) within each fold.

4. Cross Validation: Informal

How to choose λ in practice?

We use a variation of K-fold cross-validation. The steps are the following:

- Choose the number “splitting times" K.
- Split the sample randomly into two pieces of size n_1 and n_2. (We will do that K times.)
- Possible choice which has been justified theoretically is $n_1 = n\left(1 - \frac{1}{\log(n)}\right)$, $n_2 = \frac{n}{\log(n)}$.
- Define a grid of values for λ: $\{\lambda_1, \ldots, \lambda_M\}$
- For each λ calculate the validation Mean Squared Error (MSE) within each fold.
- For each λ calculate the overall cross-validation MSE.

4. Cross Validation: Informal

How to choose λ in practice?

We use a variation of K-fold cross-validation. The steps are the following:

- Choose the number “splitting times" K.
- Split the sample randomly into two pieces of size n_1 and n_2. (We will do that K times.)
- Possible choice which has been justified theoretically is $n_1 = n\left(1 - \frac{1}{\log(n)}\right)$, $n_2 = \frac{n}{\log(n)}$.
- Define a grid of values for λ: $\{\lambda_1, \ldots, \lambda_M\}$
- For each λ calculate the validation Mean Squared Error (MSE) within each fold.
- For each λ calculate the overall cross-validation MSE.
- Locate under which λ cross-validation MSE is minimized.

4. Cross Validation: Informal

How to choose λ in practice?

We use a variation of K-fold cross-validation. The steps are the following:

- Choose the number “splitting times" K.
- Split the sample randomly into two pieces of size n_1 and n_2. (We will do that K times.)
- Possible choice which has been justified theoretically is $n_1 = n\left(1 - \frac{1}{\log(n)}\right)$, $n_2 = \frac{n}{\log(n)}$.
- Define a grid of values for λ: $\{\lambda_1, \ldots, \lambda_M\}$
- For each λ calculate the validation Mean Squared Error (MSE) within each fold.
- For each λ calculate the overall cross-validation MSE.
- Locate under which λ cross-validation MSE is minimized.
- This approach ensures the regularization parameter is chosen to balance the trade-off between inducing sparsity (through thresholding) and maintaining an accurate covariance estimate.

4. Cross Validation: Formalized

We have n observations $\{X_i\}_{i=1}^n$, where each $X_i \in \mathbb{R}^p$. The goal is to select λ using cross-validation for thresholding covariance matrices.

4. Cross Validation: Formalized

We have n observations $\{X_i\}_{i=1}^n$, where each $X_i \in \mathbb{R}^p$. The goal is to select λ using cross-validation for thresholding covariance matrices.

1. Split the Data:

- For $k = 1, \ldots, K$, divide the indices $\{1, \ldots, n\}$ into:
  - Training set of size n_1.
  - Validation set of size n_2.

4. Cross Validation: Formalized

We have n observations $\{X_i\}_{i=1}^n$, where each $X_i \in \mathbb{R}^p$. The goal is to select λ using cross-validation for thresholding covariance matrices.

1. Split the Data:

- For $k = 1, \ldots, K$, divide the indices $\{1, \ldots, n\}$ into:
  - Training set of size n_1.
  - Validation set of size n_2.

2. Estimate Covariance Matrices:

- On the training set (n_1), compute: $\widehat{\Sigma}_{1,k}$ (Training Covariance Matrix).
- On the validation set (n_2), compute: $\widehat{\Sigma}_{2,k}$ (Validation Covariance Matrix).

4. Cross Validation: Formalized

3. Thresholding:

- For each $\lambda \in \{\lambda_1, \ldots, \lambda_M\}$, apply the thresholding operator:

$$T_\lambda(\widehat{\Sigma}_{1,k}) = (\widehat{\Sigma}_{1,k})_{i,j}\, \mathbb{1}_{\{|(\widehat{\Sigma}_{1,k})_{i,j}|>\lambda\}}$$

4. Cross Validation: Formalized

3. Thresholding:

- For each $\lambda \in \{\lambda_1, \ldots, \lambda_M\}$, apply the thresholding operator:

$$T_\lambda(\widehat{\Sigma}_{1,k}) = (\widehat{\Sigma}_{1,k})_{i,j}\, \mathbb{1}_{\{|(\widehat{\Sigma}_{1,k})_{i,j}| > \lambda\}}$$

4. Validation Error:

- Record the error on the validation set:

$$e_k(\lambda) = \|T_\lambda\left(\widehat{\Sigma}_{1,k}\right) - \widehat{\Sigma}_{2,k}\|_F^2,$$

where $\|\cdot\|_F$ is the Frobenius norm.

4. Cross Validation: Formalized

5. Aggregate Errors:

- Average the error across all K folds:

$$\bar{e}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} e_k(\lambda).$$

4. Cross Validation: Formalized

5. Aggregate Errors:

- Average the error across all K folds:

$$\bar{e}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} e_k(\lambda).$$

4. Cross Validation: Formalized:

- Choose λ that minimizes the average validation error:

$$\widehat{\lambda} = \underset{\lambda \in \{\lambda_1, \ldots, \lambda_M\}}{\operatorname{argmin}} \bar{e}(\lambda).$$

This formalized process ensures a principled selection of λ for thresholding the covariance matrix, balancing sparsity and accuracy.

Summary

Some things to remember:

- **Definition of Sparsity:** Sparsity refers to the number of zero (or near-zero) entries in the covariance matrix. It is often quantified through adjacency matrices or operator norms.

Summary

Some things to remember:

- **Definition of Sparsity:** Sparsity refers to the number of zero (or near-zero) entries in the covariance matrix. It is often quantified through adjacency matrices or operator norms.
- **Thresholding Operator:** The thresholding operator $T_\lambda(u)$ removes small values in the covariance matrix to induce sparsity while preserving larger, significant relationships.

Summary

Some things to remember:

- **Definition of Sparsity:** Sparsity refers to the number of zero (or near-zero) entries in the covariance matrix. It is often quantified through adjacency matrices or operator norms.
- **Thresholding Operator:** The thresholding operator $T_\lambda(u)$ removes small values in the covariance matrix to induce sparsity while preserving larger, significant relationships.
- **Why Thresholding?** Thresholding is preferred over penalization in high-dimensional statistics because it directly simplifies the covariance matrix structure while maintaining interpretability.

Summary

Some things to remember:

- **Definition of Sparsity:** Sparsity refers to the number of zero (or near-zero) entries in the covariance matrix. It is often quantified through adjacency matrices or operator norms.
- **Thresholding Operator:** The thresholding operator $T_\lambda(u)$ removes small values in the covariance matrix to induce sparsity while preserving larger, significant relationships.
- **Why Thresholding?** Thresholding is preferred over penalization in high-dimensional statistics because it directly simplifies the covariance matrix structure while maintaining interpretability.
- **Consistency:** The thresholded covariance matrix is a consistent estimator. It converges to the true covariance matrix as the sample size n grows, provided the threshold λ is chosen appropriately.