Selected Topics in Mathematics of Learning

High-Dimensional Statistics

Lecturer: Marius Yamakou

Winter Semester 2024/25
Department of Data Science, FAU

December 17, 2024

Part V: Large Inverse Covariance Matrices continued ...

2. Estimation : Why not $\hat{\Sigma}^{-1}$?

First Natural Thought:

- Use the **sample covariance matrix** and compute its inverse.
- Recall the sample covariance matrix:

$$\widehat{\Sigma} = \frac{1}{n} \sum_{k=1}^{n} (X_k - \bar{X})(X_k - \bar{X})^\top =: (\hat{\sigma}_{ij})_{i,j=1,\ldots,p}.$$

2. Estimation : Why not $\hat{\Sigma}^{-1}$?

First Natural Thought:

- Use the **sample covariance matrix** and compute its inverse.
- Recall the sample covariance matrix:

$$\widehat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})(X_k - \bar{X})^\top =: (\hat{\sigma}_{ij})_{i,j=1,\ldots,p}.$$

- The inverse of the sample covariance matrix provides an estimator for Θ:

$$\widehat{\Theta} = \widehat{\Sigma}^{-1}.$$

2. Estimation : Why not $\hat{\Sigma}^{-1}$?

First Natural Thought:

- Use the **sample covariance matrix** and compute its inverse.
- Recall the sample covariance matrix:

$$\widehat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})(X_k - \bar{X})^\top =: (\hat{\sigma}_{ij})_{i,j=1,\ldots,p}.$$

- The inverse of the sample covariance matrix provides an estimator for Θ:

$$\widehat{\Theta} = \widehat{\Sigma}^{-1}.$$

Challenges with This Approach:

- **Non-Invertibility:** $\widehat{\Sigma}$ is not invertible when $n \ll p$.

2. Estimation : Why not $\hat{\Sigma}^{-1}$?

First Natural Thought:

- Use the **sample covariance matrix** and compute its inverse.
- Recall the sample covariance matrix:

$$\widehat{\Sigma} = \frac{1}{n} \sum_{k=1}^{n} (X_k - \bar{X})(X_k - \bar{X})^\top =: (\hat{\sigma}_{ij})_{i,j=1,\ldots,p}.$$

- The inverse of the sample covariance matrix provides an estimator for Θ:

$$\widehat{\Theta} = \widehat{\Sigma}^{-1}.$$

Challenges with This Approach:

- **Non-Invertibility:** $\widehat{\Sigma}$ is not invertible when $n \ll p$.
- **Non-Sparsity:** Even if invertible, $\widehat{\Sigma}$ typically lacks exact zeros.

2. Estimation : Why not $\hat{\Sigma}^{-1}$?

First Natural Thought:

- Use the **sample covariance matrix** and compute its inverse.
- Recall the sample covariance matrix:

$$\widehat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})(X_k - \bar{X})^\top =: (\hat{\sigma}_{ij})_{i,j=1,\dots,p}.$$

- The inverse of the sample covariance matrix provides an estimator for Θ:

$$\widehat{\Theta} = \widehat{\Sigma}^{-1}.$$

Challenges with This Approach:

- **Non-Invertibility:** $\widehat{\Sigma}$ is not invertible when $n \ll p$.
- **Non-Sparsity:** Even if invertible, $\widehat{\Sigma}$ typically lacks exact zeros.
- **Numerical Instability:** Inversion can lead to large estimation errors, especially in high dimensions.

2. Estimation : Why not $\hat{\Sigma}^{-1}$?

First Natural Thought:

- Use the **sample covariance matrix** and compute its inverse.
- Recall the sample covariance matrix:

$$\widehat{\Sigma} = \frac{1}{n} \sum_{k=1}^{n} (X_k - \bar{X})(X_k - \bar{X})^\top =: (\hat{\sigma}_{ij})_{i,j=1,\ldots,p}.$$

- The inverse of the sample covariance matrix provides an estimator for Θ:

$$\widehat{\Theta} = \widehat{\Sigma}^{-1}.$$

Challenges with This Approach:

- **Non-Invertibility:** $\widehat{\Sigma}$ is not invertible when $n \ll p$.
- **Non-Sparsity:** Even if invertible, $\widehat{\Sigma}$ typically lacks exact zeros.
- **Numerical Instability:** Inversion can lead to large estimation errors, especially in high dimensions.

Proposed Solution:

- Re-parameterize the maximum likelihood estimation (MLE) in terms of the **precision matrix** Θ.

2. Estimation : Why not $\hat{\Sigma}^{-1}$?

First Natural Thought:

- Use the **sample covariance matrix** and compute its inverse.
- Recall the sample covariance matrix:

$$\widehat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})(X_k - \bar{X})^\top =: (\hat{\sigma}_{ij})_{i,j=1,\dots,p}.$$

- The inverse of the sample covariance matrix provides an estimator for Θ:

$$\widehat{\Theta} = \widehat{\Sigma}^{-1}.$$

Challenges with This Approach:

- **Non-Invertibility:** $\widehat{\Sigma}$ is not invertible when $n \ll p$.
- **Non-Sparsity:** Even if invertible, $\widehat{\Sigma}$ typically lacks exact zeros.
- **Numerical Instability:** Inversion can lead to large estimation errors, especially in high dimensions.

Proposed Solution:

- Re-parameterize the maximum likelihood estimation (MLE) in terms of the **precision matrix** Θ.
- Introduce a **penalty term** to enforce sparsity or improve stability.

2. Estimation: Recall Maximum Likelihood Estimation

What is Maximum Likelihood Estimation (MLE)?

- A method to estimate the parameters of a probability distribution by maximizing the **likelihood function**.

2. Estimation: Recall Maximum Likelihood Estimation

What is Maximum Likelihood Estimation (MLE)?

- A method to estimate the parameters of a probability distribution by maximizing the **likelihood function**.
- Ensures that, under the assumed model, the observed data is most probable.

2. Estimation: Recall Maximum Likelihood Estimation

What is Maximum Likelihood Estimation (MLE)?

- A method to estimate the parameters of a probability distribution by maximizing the **likelihood function**.
- Ensures that, under the assumed model, the observed data is most probable.

Steps in Maximum Likelihood Estimation:

- Assume a random sample from an unknown joint probability distribution, parameterized by a set of parameters.

2. Estimation: Recall Maximum Likelihood Estimation

What is Maximum Likelihood Estimation (MLE)?

- A method to estimate the parameters of a probability distribution by maximizing the **likelihood function**.
- Ensures that, under the assumed model, the observed data is most probable.

Steps in Maximum Likelihood Estimation:

- Assume a random sample from an unknown joint probability distribution, parameterized by a set of parameters.
- Define the joint density function for the observed data:
  - If random vectors are independent, the joint density equals the product of their individual densities.

2. Estimation: Recall Maximum Likelihood Estimation

What is Maximum Likelihood Estimation (MLE)?

- A method to estimate the parameters of a probability distribution by maximizing the **likelihood function**.
- Ensures that, under the assumed model, the observed data is most probable.

Steps in Maximum Likelihood Estimation:

- Assume a random sample from an unknown joint probability distribution, parameterized by a set of parameters.
- Define the joint density function for the observed data:
  - If random vectors are independent, the joint density equals the product of their individual densities.
- Evaluate the likelihood function: $L(\theta) = \prod_{k=1}^{n} f(X_k;\theta)$, where $f(X_k;\theta)$ is the density function.

2. Estimation: Recall Maximum Likelihood Estimation

What is Maximum Likelihood Estimation (MLE)?

- A method to estimate the parameters of a probability distribution by maximizing the **likelihood function**.
- Ensures that, under the assumed model, the observed data is most probable.

Steps in Maximum Likelihood Estimation:

- Assume a random sample from an unknown joint probability distribution, parameterized by a set of parameters.
- Define the joint density function for the observed data:
  - If random vectors are independent, the joint density equals the product of their individual densities.
- Evaluate the likelihood function: $L(\theta) = \prod_{k=1}^{n} f(X_k;\theta)$, where $f(X_k;\theta)$ is the density function.
- Maximize the likelihood function to find parameter values that best explain the observed data.

2. Estimation: Recall Maximum Likelihood Estimation

What is Maximum Likelihood Estimation (MLE)?

- A method to estimate the parameters of a probability distribution by maximizing the **likelihood function**.
- Ensures that, under the assumed model, the observed data is most probable.

Steps in Maximum Likelihood Estimation:

- Assume a random sample from an unknown joint probability distribution, parameterized by a set of parameters.
- Define the joint density function for the observed data:
    - If random vectors are independent, the joint density equals the product of their individual densities.
- Evaluate the likelihood function: $L(\theta) = \prod_{k=1}^{n} f(X_k; \theta)$, where $f(X_k; \theta)$ is the density function.
- Maximize the likelihood function to find parameter values that best explain the observed data.

Intuition:

- MLE identifies the parameter values that make the observed data **most probable** under the given statistical model.

2. Estimation: Likelihood estimation

Multivariate normal: A real random vector $X = (X_1, \dots, X_p)^T$ is called a normal random vector if there exists a random p-vector Z, which is a standard normal random vector, a p-vector μ, and a matrix A, such that $X = A^{\mathrm{T}} Z + \mu$.

2. Estimation: Likelihood estimation

Multivariate normal: A real random vector $X = (X_1, \ldots, X_p)^T$ is called a normal random vector if there exists a random p-vector Z, which is a standard normal random vector, a p-vector μ, and a matrix A, such that $X = A^{\mathrm{T}} Z + \mu$.

Suppose $X \sim \mathcal{N}_p(\mu, \Sigma)$ and that Σ is full rank; then X has a density:

$$f_{\mathbf{X}}(x_1, \ldots, x_p) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

where $|\Sigma|$ denotes the determinant of Σ.

2. Estimation: Likelihood estimation

Multivariate normal: A real random vector $X = (X_1, \dots, X_p)^T$ is called a normal random vector if there exists a random p-vector Z, which is a standard normal random vector, a p-vector μ, and a matrix A, such that $X = A^{\mathrm{T}} Z + \mu$.

Suppose $X \sim \mathcal{N}_p(\mu, \Sigma)$ and that Σ is full rank; then X has a density:

$$f_{\mathbf{X}}(x_1, \dots, x_p) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\top} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

where $|\Sigma|$ denotes the determinant of Σ.

Key Properties:

- The density is symmetric and bell-shaped, extending into p-dimensional space.

2. Estimation: Likelihood estimation

Multivariate normal: A real random vector $X = (X_1, \ldots, X_p)^T$ is called a normal random vector if there exists a random p-vector Z, which is a standard normal random vector, a p-vector μ, and a matrix A, such that $X = A^{\mathrm{T}} Z + \mu$.

Suppose $X \sim \mathcal{N}_p(\mu, \Sigma)$ and that Σ is full rank; then X has a density:

$$f_{\mathbf{X}}(x_1, \ldots, x_p) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

where $|\Sigma|$ denotes the determinant of Σ.

Key Properties:

- The density is symmetric and bell-shaped, extending into p-dimensional space.
- The covariance matrix Σ controls the shape and orientation of the distribution.

2. Estimation: Likelihood Estimation

Setup:

- We have random vectors $X_n = (X_{n,1}, \ldots, X_{n,p})^\top$, for $n = 1, \ldots, N$, where $X_n \sim \mathcal{N}_p(\mu, \Sigma)$.
- The mean vector $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ are unknown.

2. Estimation: Likelihood Estimation

Setup:

- We have random vectors $X_n = (X_{n,1}, \ldots, X_{n,p})^\top$, for $n = 1, \ldots, N$, where $X_n \sim \mathcal{N}_p(\mu, \Sigma)$.
- The mean vector $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ are unknown.

Joint Density: The joint density of the observations $X_1, \ldots, X_N$ is:

$$\prod_{n=1}^{N} f_{\mathbf{X}_n}(x_1, \ldots, x_p) = \prod_{n=1}^{N} \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu})\right).$$

2. Estimation: Likelihood Estimation

Negative Log-Likelihood: Taking the negative log of the joint density:

2. Estimation: Likelihood Estimation

Negative Log-Likelihood: Taking the negative log of the joint density:

$$
\begin{aligned}
-\log \prod_{n=1}^{N} f_{\mathbf{x}_n}(x_1, \ldots, x_p) &= -\log \prod_{n=1}^{N} \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu})\right) \\
&= -\log\left(\frac{1}{((2\pi)^p |\Sigma|)^{N/2}} \exp\left(-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu})\right)\right) \\
&= \frac{N}{2}\log\left((2\pi)^P |\Sigma|\right) + \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) \\
&= \frac{pN}{2}\log(2\pi) + \frac{N}{2}\log|\Sigma| + \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu})
\end{aligned}
$$

2. Estimation: Maximum Likelihood Estimators

Negative Log-Likelihood:

$$L(\mu, \Sigma) = \frac{pN}{2}\log(2\pi) + \frac{N}{2}\log|\Sigma| + \frac{1}{2}\sum_{n=1}^{N}(X_n - \mu)^\top \Sigma^{-1}(X_n - \mu).$$

2. Estimation: Maximum Likelihood Estimators

Negative Log-Likelihood:

$$L(\mu, \Sigma) = \frac{pN}{2}\log(2\pi) + \frac{N}{2}\log|\Sigma| + \frac{1}{2}\sum_{n=1}^{N}(X_n - \mu)^\top \Sigma^{-1}(X_n - \mu).$$

Estimating μ and Σ: To minimize $L(\mu, \Sigma)$, the Maximum Likelihood Estimators (MLEs) are:

$$\operatorname*{argmin}_{\mu} L(\mu, \Sigma) = \frac{1}{N}\sum_{n=1}^{N} X_n,$$

$$\operatorname*{argmin}_{\Sigma} L(\mu, \Sigma) = \frac{1}{N}\sum_{n=1}^{N}(X_n - \mu)(X_n - \mu)^\top.$$

2. Estimation: Maximum Likelihood Estimators

Negative Log-Likelihood:

$$L(\mu, \Sigma) = \frac{pN}{2} \log(2\pi) + \frac{N}{2} \log |\Sigma| + \frac{1}{2} \sum_{n=1}^{N} (X_n - \mu)^\top \Sigma^{-1} (X_n - \mu).$$

Estimating μ and Σ: To minimize $L(\mu, \Sigma)$, the Maximum Likelihood Estimators (MLEs) are:

$$\operatorname*{argmin}_{\mu} L(\mu, \Sigma) = \frac{1}{N} \sum_{n=1}^{N} X_n,$$

$$\operatorname*{argmin}_{\Sigma} L(\mu, \Sigma) = \frac{1}{N} \sum_{n=1}^{N} (X_n - \mu)(X_n - \mu)^\top.$$

Key Idea: - The sample mean $\hat{\mu}$ and sample covariance matrix $\hat{\Sigma}$ are natural estimators under the MLE framework.

2. Estimation: MLE for Precision Matrix Θ

Re-Parameterization: To get MLE for Θ we can re-parameterize in terms of precision matrix $\Theta = \Sigma^{-1}$.

2. Estimation: MLE for Precision Matrix Θ

Re-Parameterization: To get MLE for Θ we can re-parameterize in terms of precision matrix $\Theta = \Sigma^{-1}$.

Rewriting the negative log-likelihood:

2. Estimation: MLE for Precision Matrix Θ

Re-Parameterization: To get MLE for Θ we can re-parameterize in terms of precision matrix $\Theta = \Sigma^{-1}$.

Rewriting the negative log-likelihood:

$$
\begin{aligned}
L(\mu, \Sigma) &= \frac{pN}{2}\log(2\pi) + \frac{N}{2}\log|\Sigma| + \frac{1}{2}\sum_{n=1}^{N}(X_n - \mu)^\top \Sigma^{-1}(X_n - \mu) \\
&= \frac{pN}{2}\log(2\pi) + \frac{N}{2}\underbrace{\log|\Sigma|}_{-\log|\Sigma^{-1}|} + \frac{1}{2}\underbrace{\sum_{n=1}^{N}(X_n - \mu)^\top \underbrace{\Sigma^{-1}}_{\Theta}(X_n - \mu)}_{\begin{array}{c} \mathrm{tr}\left(\sum_{n=1}^{N} X_n - \mu)^\top \Theta (X_n - \mu)\right) \\ \overset{(*)}{=} \mathrm{tr}\left(\sum_{n=1}^{N} (X_n - \mu)(X_n - \mu)^\top \Theta\right) \\ = \mathrm{tr}\left(N\hat{\Sigma}\Theta\right) \\ = N\,\mathrm{tr}\left(\hat{\Sigma}\Theta\right) \end{array}}
\end{aligned}
$$

2. Estimation: MLE for Precision Matrix Θ

Re-Parameterization: To get MLE for Θ we can re-parameterize in terms of precision matrix $\Theta = \Sigma^{-1}$.

Rewriting the negative log-likelihood:

$$
\begin{aligned}
L(\mu, \Sigma) &= \frac{pN}{2}\log(2\pi) + \frac{N}{2}\log|\Sigma| + \frac{1}{2}\sum_{n=1}^{N}(X_n-\mu)^\top \Sigma^{-1}(X_n-\mu) \\
&= \frac{pN}{2}\log(2\pi) + \frac{N}{2}\underbrace{\log|\Sigma|}_{-\log|\Sigma^{-1}|} + \frac{1}{2}\underbrace{\sum_{n=1}^{N}(X_n-\mu)^\top \underbrace{\Sigma^{-1}}_{\Theta}(X_n-\mu)}_{\substack{\mathrm{tr}\left(\sum_{n=1}^{N} X_n-\mu)^\top\Theta(X_n-\mu)\right) \\ \overset{(*)}{=}\mathrm{tr}\left(\sum_{n=1}^{N}(X_n-\mu)(X_n-\mu)^\top\Theta\right) \\ =\mathrm{tr}\left(N\hat{\Sigma}\Theta\right) \\ =N\mathrm{tr}\left(\hat{\Sigma}\Theta\right)}}
\end{aligned}
$$

$$\implies L(\mu, \Sigma) = \tfrac{pN}{2}\log(2\pi) - \tfrac{N}{2}\log|\Theta| + \tfrac{1}{2}N\mathrm{tr}\left(\hat{\Sigma}\Theta\right)$$

2. Estimation: MLE for Precision Matrix Θ

where:

- $\hat{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(X_n - \mu)(X_n - \mu)^\top$ is the sample covariance matrix.

2. Estimation: MLE for Precision Matrix Θ

where:

- $\hat{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(X_n - \mu)(X_n - \mu)^\top$ is the sample covariance matrix.
- $(*)$ exploit the cyclic property of the trace operator $\mathrm{tr}(\cdot)$ given by:

$$\mathrm{tr}(ABC) = \mathrm{tr}(CAB).$$

2. Estimation: MLE for Precision Matrix Θ

where:

- $\hat{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(X_n - \mu)(X_n - \mu)^\top$ is the sample covariance matrix.
- $(*)$ exploit the cyclic property of the trace operator $\mathrm{tr}(\cdot)$ given by:

$$\mathrm{tr}(ABC) = \mathrm{tr}(CAB).$$

Summary: The MLE for Θ minimizes:

$$L(\mu, \Sigma) = \frac{pN}{2}\log(2\pi) - \frac{N}{2}\log|\Theta| + \frac{N}{2}\mathrm{tr}(\hat{\Sigma}\Theta).$$

2.2 Graphical LASSO

Motivation: The graphical LASSO introduces sparsity into the estimation of the precision matrix $\Theta = \Sigma^{-1}$, leading to interpretable graphical models by penalizing small off-diagonal entries.

2.2 Graphical LASSO

Motivation: The graphical LASSO introduces sparsity into the estimation of the precision matrix $\Theta = \Sigma^{-1}$, leading to interpretable graphical models by penalizing small off-diagonal entries.

Maximum Likelihood Estimator (MLE): Based on the sample covariance matrix $\widehat{\Sigma}$, the MLE of Θ is:

$$\widehat{\Theta} = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\operatorname{argmin}} \left(\operatorname{tr}(\Theta \widehat{\Sigma}) - \log |\Theta| \right),$$

subject to two constraints:

2.2 Graphical LASSO

Motivation: The graphical LASSO introduces sparsity into the estimation of the precision matrix $\Theta = \Sigma^{-1}$, leading to interpretable graphical models by penalizing small off-diagonal entries.

Maximum Likelihood Estimator (MLE): Based on the sample covariance matrix $\widehat{\Sigma}$, the MLE of Θ is:

$$\widehat{\Theta} = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\operatorname{argmin}} \left(\operatorname{tr}(\Theta \widehat{\Sigma}) - \log |\Theta| \right),$$

subject to two constraints:

- $\Theta \succ 0$: Positive definiteness of the precision matrix.
- $\Theta^\top = \Theta$: Symmetry of the precision matrix.

2.2 Graphical LASSO

Why the Objective: $\widehat{\Theta} = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\text{argmin}} \left(\mathrm{tr}(\Theta \widehat{\Sigma}) - \log |\Theta| \right)$?

2.2 Graphical LASSO

Why the Objective: $\widehat{\Theta} = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\operatorname{argmin}} \left(\operatorname{tr}(\Theta \widehat{\Sigma}) - \log |\Theta| \right)$?

Because:

- $\operatorname{tr}(\Theta\widehat{\Sigma})$ ensures that the precision matrix Θ aligns with the empirical covariance matrix $\widehat{\Sigma}$.
  - It penalizes deviations of Θ from matching $\widehat{\Sigma}$.

2.2 Graphical LASSO

Why the Objective: $\widehat{\Theta} = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\text{argmin}} \left(\mathrm{tr}(\Theta \widehat{\Sigma}) - \log |\Theta| \right)$?

Because:

- $\mathrm{tr}(\Theta\widehat{\Sigma})$ ensures that the precision matrix Θ aligns with the empirical covariance matrix $\widehat{\Sigma}$.
  - It penalizes deviations of Θ from matching $\widehat{\Sigma}$.
- $-\log|\Theta|$ represents the log-determinant of Θ, which has two key roles:

2.2 Graphical LASSO

Why the Objective: $\widehat{\Theta} = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\text{argmin}} \left(\mathrm{tr}(\Theta\widehat{\Sigma}) - \log|\Theta| \right)$?

Because:

- $\mathrm{tr}(\Theta\widehat{\Sigma})$ ensures that the precision matrix Θ aligns with the empirical covariance matrix $\widehat{\Sigma}$.
  - It penalizes deviations of Θ from matching $\widehat{\Sigma}$.
- $-\log|\Theta|$ represents the log-determinant of Θ, which has two key roles:
  - Ensures the matrix Θ is invertible.

2.2 Graphical LASSO

Why the Objective: $\widehat{\Theta} = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\operatorname{argmin}} \left(\operatorname{tr}(\Theta \widehat{\Sigma}) - \log |\Theta| \right)$?

Because:

- $\operatorname{tr}(\Theta \widehat{\Sigma})$ ensures that the precision matrix Θ aligns with the empirical covariance matrix $\widehat{\Sigma}$.
  - It penalizes deviations of Θ from matching $\widehat{\Sigma}$.
- $-\log |\Theta|$ represents the log-determinant of Θ, which has two key roles:
  - Ensures the matrix Θ is invertible.
  - Penalizes overly large or overly small eigenvalues, maintaining numerical stability.

2.2 Graphical LASSO

Why the Objective: $\widehat{\Theta} = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\text{argmin}} \left(\mathrm{tr}(\Theta \widehat{\Sigma}) - \log |\Theta| \right)$?

Because:

- $\mathrm{tr}(\Theta \widehat{\Sigma})$ ensures that the precision matrix Θ aligns with the empirical covariance matrix $\widehat{\Sigma}$.
  - It penalizes deviations of Θ from matching $\widehat{\Sigma}$.
- $-\log |\Theta|$ represents the log-determinant of Θ, which has two key roles:
  - Ensures the matrix Θ is invertible.
  - Penalizes overly large or overly small eigenvalues, maintaining numerical stability.

Intuition: The objective function balances:

- **Fidelity:** The trace term $\mathrm{tr}(\Theta \widehat{\Sigma})$ ensures consistency with the observed data.

2.2 Graphical LASSO

Why the Objective: $\widehat{\Theta} = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\operatorname{argmin}} \left(\operatorname{tr}(\Theta \widehat{\Sigma}) - \log |\Theta| \right)$?

Because:

- $\operatorname{tr}(\Theta \widehat{\Sigma})$ ensures that the precision matrix Θ aligns with the empirical covariance matrix $\widehat{\Sigma}$.
  - It penalizes deviations of Θ from matching $\widehat{\Sigma}$.
- $-\log |\Theta|$ represents the log-determinant of Θ, which has two key roles:
  - Ensures the matrix Θ is invertible.
  - Penalizes overly large or overly small eigenvalues, maintaining numerical stability.

Intuition: The objective function balances:

- **Fidelity:** The trace term $\operatorname{tr}(\Theta \widehat{\Sigma})$ ensures consistency with the observed data.
- **Regularization:** The log-determinant term $-\log |\Theta|$ prevents degenerate solutions and stabilizes the optimization.

2.2 Graphical LASSO

Graphical LASSO: Introducing Sparsity The graphical LASSO estimator adds an ℓ_1-penalty on the off-diagonal entries of Θ to induce sparsity:

2.2 Graphical LASSO

Graphical LASSO: Introducing Sparsity The graphical LASSO estimator adds an ℓ_1-penalty on the off-diagonal entries of Θ to induce sparsity:

$$\widehat{\Theta}_\lambda = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\text{argmin}} \left(\operatorname{tr}(\Theta \widehat{\Sigma}) - \log |\Theta| + \lambda \|\Theta\|_{1,\text{off}} \right),$$

where:

2.2 Graphical LASSO

Graphical LASSO: Introducing Sparsity The graphical LASSO estimator adds an ℓ_1-penalty on the off-diagonal entries of Θ to induce sparsity:

$$\widehat{\Theta}_\lambda = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\operatorname{argmin}} \left(\operatorname{tr}(\Theta\widehat{\Sigma}) - \log|\Theta| + \lambda\|\Theta\|_{1,\text{off}} \right),$$

where:

$$\|\Theta\|_{1,\text{off}} = \sum_{\substack{i,j=1 \\ i \neq j}}^{p} |\Theta_{ij}|.$$

2.2 Graphical LASSO

Graphical LASSO: Introducing Sparsity The graphical LASSO estimator adds an ℓ_1-penalty on the off-diagonal entries of Θ to induce sparsity:

$$\widehat{\Theta}_\lambda = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\text{argmin}} \left(\operatorname{tr}(\Theta \widehat{\Sigma}) - \log|\Theta| + \lambda \|\Theta\|_{1,\text{off}} \right),$$

where:

$$\|\Theta\|_{1,\text{off}} = \sum_{\substack{i,j=1 \\ i \neq j}}^{p} |\Theta_{ij}|.$$

Key Question:

- Is the penalized likelihood a convex function in Θ?

2.2 Graphical LASSO

Graphical LASSO: Introducing Sparsity The graphical LASSO estimator adds an ℓ_1-penalty on the off-diagonal entries of Θ to induce sparsity:

$$\widehat{\Theta}_\lambda = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\text{argmin}} \left(\operatorname{tr}(\Theta \widehat{\Sigma}) - \log |\Theta| + \lambda \|\Theta\|_{1,\text{off}} \right),$$

where:

$$\|\Theta\|_{1,\text{off}} = \sum_{\substack{i,j=1 \\ i \neq j}}^{p} |\Theta_{ij}|.$$

Key Question:
- Is the penalized likelihood a convex function in Θ?

Key Insight:
- The graphical LASSO objective is a combination of a convex negative log-likelihood function and a convex ℓ_1-norm penalty term.

2.2 Graphical LASSO

Graphical LASSO: Introducing Sparsity The graphical LASSO estimator adds an ℓ_1-penalty on the off-diagonal entries of Θ to induce sparsity:

$$\widehat{\Theta}_\lambda = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\text{argmin}} \left(\mathrm{tr}(\Theta \widehat{\Sigma}) - \log|\Theta| + \lambda \|\Theta\|_{1,\mathrm{off}} \right),$$

where:

$$\|\Theta\|_{1,\mathrm{off}} = \sum_{\substack{i,j=1 \\ i \neq j}}^{p} |\Theta_{ij}|.$$

Key Question:
- Is the penalized likelihood a convex function in Θ?

Key Insight:
- The graphical LASSO objective is a combination of a convex negative log-likelihood function and a convex ℓ_1-norm penalty term.
- Therefore, the overall problem remains convex in Θ. And the answer to the above question is Yes!