Selected Topics in Mathematics of Learning

High-Dimensional Statistics

Lecturer: Marius Yamakou

Winter Semester 2024/25
Department of Data Science, FAU

January 14, 2025

Part V: Large Inverse Covariance Matrices continued ...

2.2 Graphical LASSO: Theoretical Perspective

Problem Setup: Suppose we observe $X_1, \ldots, X_N$, i.i.d. p-variate normal random variables with mean $\mathbf{0}$ and covariance matrix Σ.

2.2 Graphical LASSO: Theoretical Perspective

Problem Setup: Suppose we observe $X_1, \ldots, X_N$, i.i.d. p-variate normal random variables with mean $\mathbf{0}$ and covariance matrix Σ.

Sparsity Assumption: Define the set of non-zero off-diagonal entries of $\Theta = \Sigma^{-1}$ as $S = \{(i,j) \mid \Theta_{ij} \neq 0,\ i \neq j\}, \quad$ with $\text{card}(S) \leq s$.

2.2 Graphical LASSO: Theoretical Perspective

Problem Setup: Suppose we observe $X_1, \dots, X_N$, i.i.d. p-variate normal random variables with mean $\mathbf{0}$ and covariance matrix Σ.

Sparsity Assumption: Define the set of non-zero off-diagonal entries of $\Theta = \Sigma^{-1}$ as $S = \{(i, j) \mid \Theta_{ij} \neq 0,\ i \neq j\}$, with $\text{card}(S) \leq s$.

Lemma

If the following conditions hold: $\lambda_{\min}(\Sigma) \geq k_1 > 0$, $\lambda_{\max}(\Sigma) \leq k_2 < \infty$, $\frac{\log(p)}{N} = O(1)$, *and* λ *is chosen as* $\lambda = M\sqrt{\frac{\log(p)}{N}}$ *for some constant* M*, then the Graphical LASSO estimator* $\widehat{\Theta}_\lambda$ *satisfies:*

$$\|\widehat{\Theta}_\lambda - \Theta\|_F = O_{\mathbb{P}}\left(\sqrt{\frac{(p+s)\log(p)}{N}}\right).$$

2.2 Graphical LASSO: Theoretical Perspective

Intuition:

- Notice that the estimation error depends on the dimensionality p, sparsity level s, and sample size N, scaling logarithmically with p and inversely with N.

2.2 Graphical LASSO: Theoretical Perspective

Intuition:

- Notice that the estimation error depends on the dimensionality p, sparsity level s, and sample size N, scaling logarithmically with p and inversely with N.

- The lemma states that the Frobenius norm of the difference between the estimated precision matrix $\widehat{\Theta}_\lambda$ and the true precision matrix Θ is stochastically bounded by $\sqrt{\frac{(p+s)\log(p)}{N}}$.

2.2 Graphical LASSO: Theoretical Perspective

- $O_{\mathbb{P}}$: This is the probabilistic version of the "big-O" notation. It describes the stochastic (random) order of magnitude of a random variable as $N \to \infty$. Specifically, if X_N is a sequence of random variables, $X_N = O_{\mathbb{P}}(a_N)$ means:

  $\forall \varepsilon > 0, \exists M > 0$ such that $\mathbb{P}(|X_N| > M a_N) \to 0$ as $N \to \infty$.

  In simpler terms, the probability that X_N exceeds a constant multiple of a_N goes to zero as N grows.
- In the context of the Lemma, this means that with high probability (as $N \to \infty$), the error of the estimator scales no worse than $\sqrt{\frac{(p+s)\log(p)}{N}}$.

2.2 Graphical LASSO: Theoretical Perspective

- This term $\sqrt{\frac{(p+s)\log(p)}{N}}$ in the lemma represents the asymptotic growth rate of the error bound for the Graphical LASSO estimator, depending on the following parameters:

2.2 Graphical LASSO: Theoretical Perspective

- This term $\sqrt{\frac{(p+s)\log(p)}{N}}$ in the lemma represents the asymptotic growth rate of the error bound for the Graphical LASSO estimator, depending on the following parameters:
  - p: Number of features.

2.2 Graphical LASSO: Theoretical Perspective

- This term $\sqrt{\frac{(p+s)\log(p)}{N}}$ in the lemma represents the asymptotic growth rate of the error bound for the Graphical LASSO estimator, depending on the following parameters:
  - p: Number of features.
    s: The sparsity level, i.e., the number of non-zero off-diagonal entries in the precision matrix.

2.2 Graphical LASSO: Theoretical Perspective

- This term $\sqrt{\frac{(p+s)\log(p)}{N}}$ in the lemma represents the asymptotic growth rate of the error bound for the Graphical LASSO estimator, depending on the following parameters:
  - p: Number of features.
    s: The sparsity level, i.e., the number of non-zero off-diagonal entries in the precision matrix. N: The sample size.

2.2 Graphical LASSO: Theoretical Perspective

- This term $\sqrt{\frac{(p+s)\log(p)}{N}}$ in the lemma represents the asymptotic growth rate of the error bound for the Graphical LASSO estimator, depending on the following parameters:
  - p: Number of features.
    s: The sparsity level, i.e., the number of non-zero off-diagonal entries in the precision matrix. N: The sample size.
    $\log(p)$: The logarithmic dependence on p, which often arises in high-dimensional statistics.

Implication:

- The error grows with the dimensionality p and sparsity level s, reflecting the challenge of estimating high-dimensional precision matrices.

2.2 Graphical LASSO: Theoretical Perspective

- This term $\sqrt{\frac{(p+s)\log(p)}{N}}$ in the lemma represents the asymptotic growth rate of the error bound for the Graphical LASSO estimator, depending on the following parameters:
  - p: Number of features.
    s: The sparsity level, i.e., the number of non-zero off-diagonal entries in the precision matrix. N: The sample size.
    $\log(p)$: The logarithmic dependence on p, which often arises in high-dimensional statistics.

Implication:

- The error grows with the dimensionality p and sparsity level s, reflecting the challenge of estimating high-dimensional precision matrices.

- The error decreases as the sample size N increases, showing that more data reduces uncertainty in the estimation.

2.2 Graphical LASSO: Theoretical Perspective

Question: How do we eliminate the dependence on p in the error bound?

2.2 Graphical LASSO: Theoretical Perspective

Question: How do we eliminate the dependence on p in the error bound?

Key Insight: Decompose the covariance matrix Σ:

$$\Sigma = W \Gamma W,$$

where:

- $W = \mathrm{diag}(\Sigma_{11}^{1/2}, \ldots, \Sigma_{pp}^{1/2})$ is the diagonal matrix of standard deviations,

2.2 Graphical LASSO: Theoretical Perspective

Question: How do we eliminate the dependence on p in the error bound?

Key Insight: Decompose the covariance matrix Σ:

$$\Sigma = W\Gamma W,$$

where:

- $W = \mathrm{diag}(\Sigma_{11}^{1/2}, \ldots, \Sigma_{pp}^{1/2})$ is the diagonal matrix of standard deviations,
- Γ is the true correlation matrix.

Implications: The precision matrix satisfies:

$$\Theta = \Sigma^{-1} = W^{-1}\Gamma^{-1}W^{-1}, \quad \text{where } K := \Gamma^{-1} = W\Theta W.$$

2.2 Graphical LASSO: Theoretical Perspective

Estimation Problem: The regularized correlation precision matrix estimator:

$$\widehat{K}_\lambda = \underset{\substack{K \succ 0 \\ K^\top = K}}{\mathrm{argmin}} \left(\mathrm{tr}(K\widehat{\Gamma}) - \log |K| + \lambda \|K\|_{1,\mathsf{off}} \right).$$

2.2 Graphical LASSO: Theoretical Perspective

Estimation Problem: The regularized correlation precision matrix estimator:

$$\widehat{K}_\lambda = \underset{\substack{K \succ 0 \\ K^\top = K}}{\mathrm{argmin}} \left(\mathrm{tr}(K\widehat{\Gamma}) - \log|K| + \lambda \|K\|_{1,\mathsf{off}} \right).$$

New Estimator: To account for the scaling:

$$\widetilde{\Theta}_\lambda = \widehat{W}^{-1}\widehat{K}_\lambda \widehat{W}^{-1},$$

where $\widehat{W} = \mathrm{diag}(\widehat{\Sigma}_{11}^{1/2}, \ldots, \widehat{\Sigma}_{pp}^{1/2})$.

2.2 Graphical LASSO: Theoretical Perspective

Lemma

Suppose $\lambda_{\min}(\Sigma) \geq k_1 > 0$, $\lambda_{\max}(\Sigma) \leq k_2 < \infty$, $\frac{\log(p)}{N} = O(1)$ *and set* $\lambda = M\sqrt{\frac{\log(p)}{n}}$ *for some* M. *Then,*

$$\|\widetilde{\Theta}_\lambda - \Theta\|_F = O_{\mathbb{P}}\left(\sqrt{\frac{(1+s)\log(p)}{N}}\right).$$

Takeaway: This result demonstrates that incorporating the scaling factor W reduces the dimensional dependence in the error bound, leading to improved estimation performance.

2.2 Graphical LASSO: Theoretical Perspective

Question: How to Choose λ in practice?

2.2 Graphical LASSO: Theoretical Perspective

Question: How to Choose λ in practice?

1. Cross-Validation (CV): This approach is similar to its use in regression models. It involves splitting the data into training and validation sets to select the λ that minimizes prediction error.

2.2 Graphical LASSO: Theoretical Perspective

Question: How to Choose λ in practice?

1. Cross-Validation (CV): This approach is similar to its use in regression models. It involves splitting the data into training and validation sets to select the λ that minimizes prediction error.

2. Alternative: Extended Bayesian Information Criterion (eBIC): The eBIC approach selects Θ by solving:

$$\Theta_{\text{eBIC}} = \underset{\Theta \in \mathcal{E}}{\operatorname{argmin}} \, \text{eBIC}_{\gamma}(\Theta),$$

2.2 Graphical LASSO: Theoretical Perspective

Question: How to Choose λ in practice?

1. Cross-Validation (CV): This approach is similar to its use in regression models. It involves splitting the data into training and validation sets to select the λ that minimizes prediction error.

2. Alternative: Extended Bayesian Information Criterion (eBIC): The eBIC approach selects Θ by solving:

$$\Theta_{\text{eBIC}} = \underset{\Theta \in \mathcal{E}}{\text{argmin}}\, \text{eBIC}_\gamma(\Theta),$$

where:

$$\text{eBIC}_\gamma(\Theta) = \overbrace{\text{tr}(\Theta\widehat{\Sigma}) - \log|\Theta|}^{\text{MLE}} + \overbrace{|\Theta|\log(N) + 4|\Theta|\gamma\log(p)}^{\text{Penalty}}.$$

2.2 Graphical LASSO: Theoretical Perspective

Question: How to Choose λ in practice?

1. Cross-Validation (CV): This approach is similar to its use in regression models. It involves splitting the data into training and validation sets to select the λ that minimizes prediction error.

2. Alternative: Extended Bayesian Information Criterion (eBIC): The eBIC approach selects Θ by solving:

$$\Theta_{\text{eBIC}} = \underset{\Theta \in \mathcal{E}}{\text{argmin}}\, \text{eBIC}_\gamma(\Theta),$$

where:

$$\text{eBIC}_\gamma(\Theta) = \overbrace{\text{tr}(\Theta\widehat{\Sigma}) - \log|\Theta|}^{\text{MLE}} + \overbrace{|\Theta|\log(N) + 4|\Theta|\gamma\log(p)}^{\text{Penalty}}.$$

- $\mathcal{E}$: The set of all solutions $\widehat{\Theta}_\lambda = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\text{argmin}} \left(\text{tr}(\Theta\widehat{\Sigma}) - \log|\Theta| + \lambda\|\Theta\|_{1,\text{off}}\right)$, for a pre-determined range of λ.

2.2 Graphical LASSO: Theoretical Perspective

Question: How to Choose λ in practice?

1. Cross-Validation (CV): This approach is similar to its use in regression models. It involves splitting the data into training and validation sets to select the λ that minimizes prediction error.

2. Alternative: Extended Bayesian Information Criterion (eBIC): The eBIC approach selects Θ by solving:

$$\Theta_{\text{eBIC}} = \underset{\Theta \in \mathcal{E}}{\text{argmin}}\, \text{eBIC}_\gamma(\Theta),$$

where:

$$\text{eBIC}_\gamma(\Theta) = \overbrace{\text{tr}(\Theta\widehat{\Sigma}) - \log|\Theta|}^{\text{MLE}} + \overbrace{|\Theta|\log(N) + 4|\Theta|\gamma\log(p)}^{\text{Penalty}}.$$

- $\mathcal{E}$: The set of all solutions $\widehat{\Theta}_\lambda = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\text{argmin}} \left(\text{tr}(\Theta\widehat{\Sigma}) - \log|\Theta| + \lambda\|\Theta\|_{1,\text{off}}\right)$, for a pre-determined range of λ.
- $|\Theta|$: Number of non-zero off-diagonal entries in Θ (graph sparsity level).

2.2 Graphical LASSO: Theoretical Perspective

Question: How to Choose λ in practice?

1. Cross-Validation (CV): This approach is similar to its use in regression models. It involves splitting the data into training and validation sets to select the λ that minimizes prediction error.

2. Alternative: Extended Bayesian Information Criterion (eBIC): The eBIC approach selects Θ by solving:

$$\Theta_{\text{eBIC}} = \underset{\Theta \in \mathcal{E}}{\operatorname{argmin}} \, \text{eBIC}_\gamma(\Theta),$$

where:

$$\text{eBIC}_\gamma(\Theta) = \overbrace{\text{tr}(\Theta\widehat{\Sigma}) - \log|\Theta|}^{\text{MLE}} + \overbrace{|\Theta| \log(N) + 4|\Theta|\gamma \log(p)}^{\text{Penalty}}.$$

- $\mathcal{E}$: The set of all solutions $\widehat{\Theta}_\lambda = \underset{\Theta \succ 0, \Theta^\top = \Theta}{\operatorname{argmin}} \left(\text{tr}(\Theta\widehat{\Sigma}) - \log|\Theta| + \lambda\|\Theta\|_{1,\text{off}} \right),$ for a pre-determined range of λ.
- $|\Theta|$: Number of non-zero off-diagonal entries in Θ (graph sparsity level).
- $(0 \leq \gamma \leq 1)$: Tuning parameter controlling the penalty for model complexity.

2.2 Graphical LASSO: CV vs eBIC

- If $\gamma = 0$, the classical BIC is recovered, which is asymptotically consistent for model selection when p is fixed and $N \to \infty$. Consistency means selecting the **smallest true graph or model**, denoted $\mathbf{\Theta_0}$.

2.2 Graphical LASSO: CV vs eBIC

- If $\gamma = 0$, the classical BIC is recovered, which is asymptotically consistent for model selection when p is fixed and $N \to \infty$. Consistency means selecting the **smallest true graph or model**, denoted $\mathbf{\Theta_0}$.
- Positive γ introduces stronger penalization of large graphs, balancing model fit and complexity.

2.2 Graphical LASSO: CV vs eBIC

- If $\gamma = 0$, the classical BIC is recovered, which is asymptotically consistent for model selection when p is fixed and $N \to \infty$. Consistency means selecting the **smallest true graph or model**, denoted $\boldsymbol{\Theta_0}$.
- Positive γ introduces stronger penalization of large graphs, balancing model fit and complexity.

Key Result: Under certain conditions, eBIC satisfies:

$$\Theta_0 = \underset{\Theta \in \mathcal{E}}{\text{argmin}}\ \text{eBIC}_\gamma(\Theta)$$

with probability tending to one as $N \to \infty$, provided $\mathcal{E}$ contains $\boldsymbol{\Theta_0}$.

2.2 Graphical LASSO: CV vs eBIC

- If $\gamma = 0$, the classical BIC is recovered, which is asymptotically consistent for model selection when p is fixed and $N \to \infty$. Consistency means selecting the **smallest true graph or model**, denoted $\boldsymbol{\Theta_0}$.
- Positive γ introduces stronger penalization of large graphs, balancing model fit and complexity.

Key Result: Under certain conditions, eBIC satisfies:

$$\Theta_0 = \underset{\Theta \in \mathcal{E}}{\operatorname{argmin}} \, \mathrm{eBIC}_\gamma(\Theta)$$

with probability tending to one as $N \to \infty$, provided $\mathcal{E}$ contains $\boldsymbol{\Theta_0}$.

Comparison to Cross-Validation (CV):

- CV minimizes prediction error and is straightforward to implement.

2.2 Graphical LASSO: CV vs eBIC

- If $\gamma = 0$, the classical BIC is recovered, which is asymptotically consistent for model selection when p is fixed and $N \to \infty$. Consistency means selecting the **smallest true graph or model**, denoted $\mathbf{\Theta_0}$.
- Positive γ introduces stronger penalization of large graphs, balancing model fit and complexity.

Key Result: Under certain conditions, eBIC satisfies:

$$\Theta_0 = \underset{\Theta \in \mathcal{E}}{\text{argmin}}\, \text{eBIC}_\gamma(\Theta)$$

with probability tending to one as $N \to \infty$, provided $\mathcal{E}$ contains $\mathbf{\Theta_0}$.

Comparison to Cross-Validation (CV):

- CV minimizes prediction error and is straightforward to implement.
- eBIC explicitly balances model fit and complexity, making it particularly useful for high-dimensional and sparse models.

2.2 Graphical LASSO: CV vs eBIC

- If $\gamma = 0$, the classical BIC is recovered, which is asymptotically consistent for model selection when p is fixed and $N \to \infty$. Consistency means selecting the **smallest true graph or model**, denoted $\boldsymbol{\Theta_0}$.
- Positive γ introduces stronger penalization of large graphs, balancing model fit and complexity.

Key Result: Under certain conditions, eBIC satisfies:

$$\Theta_0 = \underset{\Theta \in \mathcal{E}}{\text{argmin}} \, \text{eBIC}_\gamma(\Theta)$$

with probability tending to one as $N \to \infty$, provided $\mathcal{E}$ contains $\boldsymbol{\Theta_0}$.

Comparison to Cross-Validation (CV):

- CV minimizes prediction error and is straightforward to implement.
- eBIC explicitly balances model fit and complexity, making it particularly useful for high-dimensional and sparse models.

Takeaway: eBIC provides a principled way to select the smallest true model, especially in high-dimensional settings, while CV remains a straightforward, practical, and versatile alternative.

3 Application of Graphical LASSO

Example Use Case: Learning Gene Networks

- Gene expression levels are often modeled as following a multivariate normal distribution (approximately).
- Identifying nonzero entries in Θ corresponds to discovering pairs of genes with direct relationships.
- This helps distinguish direct dependencies from indirect correlations caused by other genes.

3 Application of Graphical LASSO

Example Use Case: Learning Gene Networks

- Gene expression levels are often modeled as following a multivariate normal distribution (approximately).
- Identifying nonzero entries in Θ corresponds to discovering pairs of genes with direct relationships.
- This helps distinguish direct dependencies from indirect correlations caused by other genes.

Other Applications:

- **Cellular Networks:** Understanding intracellular interactions.
- **fMRI Data:** Inferring brain connectivity networks.
- **ETF Stocks:** Modeling relationships between financial instruments in a portfolio.

3 Application of Graphical LASSO

Example Use Case: Learning Gene Networks

- Gene expression levels are often modeled as following a multivariate normal distribution (approximately).
- Identifying nonzero entries in Θ corresponds to discovering pairs of genes with direct relationships.
- This helps distinguish direct dependencies from indirect correlations caused by other genes.

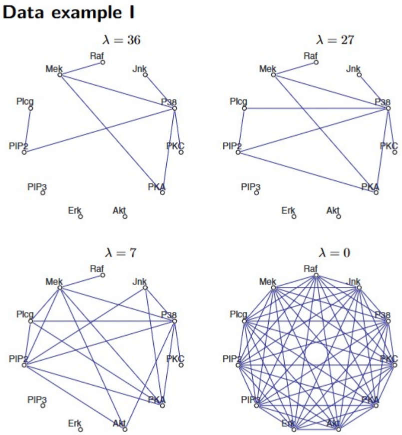
Other Applications:

- **Cellular Networks:** Understanding intracellular interactions.
- **fMRI Data:** Inferring brain connectivity networks.
- **ETF Stocks:** Modeling relationships between financial instruments in a portfolio.

Takeaway: Graphical LASSO is a powerful tool for uncovering structure in complex, high-dimensional data across diverse fields.

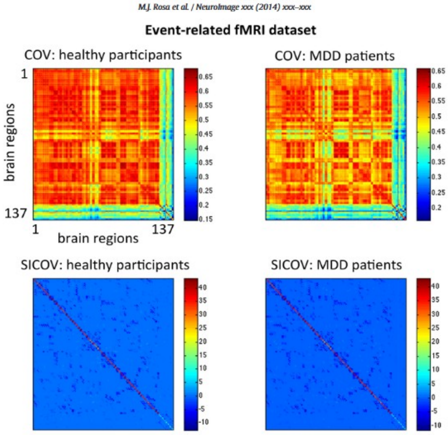
3. Data Example I: Multiparameter Single-Cell Data

- **Graphical Model:** Each vertex in the graph represents a protein, and the edges show conditional dependencies between proteins.
- **Precision Matrix and Sparsity:** The precision matrix Θ (inverse of the covariance matrix Σ) describes the conditional independence structure. Specifically, an element $\Theta_{ij} = 0$ means proteins i and j are conditionally independent given the other proteins.
- As λ, a regularization parameter, varies, the sparsity of the precision matrix changes. For higher λ, the graph becomes sparser (fewer edges), capturing only the strongest conditional dependencies.



3. Data Example II: Event-Related fMRI Dataset for healthy participants versus patients with Major Depressive Disorder (MDD)

- Covariance (1st row): Show the raw pairwise correlations between brain regions. Do not account for indirect relationships, meaning that even weak correlations could arise due to other regions' effects.
- Precision (2nd row): Highlight direct dependencies between brain regions by removing the influence of other regions.
- Sparsity patterns differ between groups, reflecting different brain network structures.



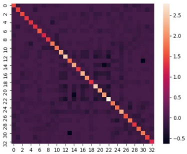
3. Data Example: Graphical Lasso for ETFs

Graphical Lasso for ETFs

- Analyzing 32 ETFs corresponding to different countries.
- Time series of daily closing prices transformed into log returns.
- Goal: Identify direct dependencies between ETFs using Graphical Lasso.

Data example III Part 1

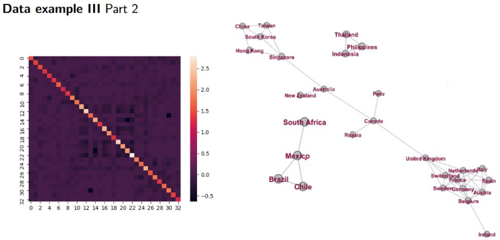
- Graphical Lasso to analyze ETFs.
- 32 ETFs which correspond to certain countries.
- transform the time series of daily closing prices into a series of log returns.



[0] https://towardsdatascience.com/machine-learning-in-action-in-finance-using-graphical-lasso-to-identify-trading-pairs-in-fa00d29c71a7

3. Example on Network Representation of ETFs

- Left: Heatmap of correlation or precision matrix, highlighting relationships.
- Right: Sparse network graph showing significant ETF dependencies.
- Clusters: ETFs grouped by economic/geographic similarities (e.g., Europe, Asia).



4. Summary

Precision:

- What does conditional independence mean?
- Why does the precision matrix encode conditional independence?
- How can we read conditional independence from the precision matrix?
- How can we present the precision matrix as a graphical model?
- How can we get information about conditional independence from the graph?

4. Summary

Precision:

- What does conditional independence mean?
- Why does the precision matrix encode conditional independence?
- How can we read conditional independence from the precision matrix?
- How can we present the precision matrix as a graphical model?
- How can we get information about conditional independence from the graph?

MLE:

- Why is inverting the sample covariance matrix not a good estimator for the precision matrix? Why is it sometimes even impossible to calculate?
- How can we re-parameterize the MLE in terms of the precision matrix?

4. Summary

Precision:

- What does conditional independence mean?
- Why does the precision matrix encode conditional independence?
- How can we read conditional independence from the precision matrix?
- How can we present the precision matrix as a graphical model?
- How can we get information about conditional independence from the graph?

MLE:

- Why is inverting the sample covariance matrix not a good estimator for the precision matrix? Why is it sometimes even impossible to calculate?
- How can we re-parameterize the MLE in terms of the precision matrix?

Model selection:

- cross-validation
- eBIC