Selected Topics in Mathematics of Learning

High-Dimensional Statistics

Lecturer: Marius Yamakou

Winter Semester 2024/25
Department of Data Science, FAU

January 28, 2025

Part VI: Sparse Vector Autoregressive Models continued ...

Example: A stationary VAR(1)

$$Y_t = AY_{t-1} + \epsilon_t, A = \begin{pmatrix} 0.5 & 0.3 \\ 0.02 & 0.8 \end{pmatrix}, E(\epsilon_t \epsilon_t') = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}, \lambda = \begin{pmatrix} 0.81 \\ 0.48 \end{pmatrix}.$$

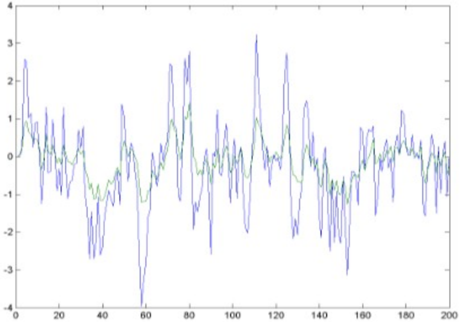

Figure 1: Blu: Y_1, green Y_2.

2. Estimation of sparse VAR through LASSO

1. **VAR(p) model structure:** The VAR(p) model is given by:

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \cdots + \Phi_p X_{t-p} + \epsilon_t,$$

2. Estimation of sparse VAR through LASSO

1. **VAR(p) model structure:** The VAR(p) model is given by:

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \cdots + \Phi_p X_{t-p} + \epsilon_t,$$

where:

- $X_t \in \mathbb{R}^d$ is the d-dimensional vector at time t,

2. Estimation of sparse VAR through LASSO

1. **VAR(p) model structure:** The VAR(p) model is given by:

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \cdots + \Phi_p X_{t-p} + \epsilon_t,$$

where:

- $X_t \in \mathbb{R}^d$ is the d-dimensional vector at time t,
- $\Phi_k \in \mathbb{R}^{d \times d}$ $(k = 1, 2, \ldots, p)$ are the coefficient matrices,

2. Estimation of sparse VAR through LASSO

1. **VAR(p) model structure:** The VAR(p) model is given by:

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \cdots + \Phi_p X_{t-p} + \epsilon_t,$$

where:

- $X_t \in \mathbb{R}^d$ is the d-dimensional vector at time t,
- $\Phi_k \in \mathbb{R}^{d \times d}$ $(k = 1, 2, \ldots, p)$ are the coefficient matrices,
- $\epsilon_t \in \mathbb{R}^d$ is a noise term.

2. Estimation of sparse VAR through LASSO

1. **VAR(p) model structure:** The VAR(p) model is given by:

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \cdots + \Phi_p X_{t-p} + \epsilon_t,$$

where:

- $X_t \in \mathbb{R}^d$ is the d-dimensional vector at time t,
- $\Phi_k \in \mathbb{R}^{d\times d}$ $(k = 1, 2, \ldots, p)$ are the coefficient matrices,
- $\epsilon_t \in \mathbb{R}^d$ is a noise term.
- Each coefficient matrix Φ_k contains d^2 parameters (since it is a $d \times d$ matrix), and there are p such matrices in a VAR(p) model.

2. Estimation of sparse VAR through LASSO

1. **VAR(p) model structure:** The VAR(p) model is given by:

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \cdots + \Phi_p X_{t-p} + \epsilon_t,$$

where:

- $X_t \in \mathbb{R}^d$ is the d-dimensional vector at time t,
- $\Phi_k \in \mathbb{R}^{d \times d}$ $(k = 1, 2, \ldots, p)$ are the coefficient matrices,
- $\epsilon_t \in \mathbb{R}^d$ is a noise term.
- Each coefficient matrix Φ_k contains d^2 parameters (since it is a $d \times d$ matrix), and there are p such matrices in a VAR(p) model.
- Therefore, the total number of parameters contributed by the coefficient matrices is: pd^2.

Explanation of the formula for VAR parameters

2. **Covariance matrix of the noise:**
   - The noise term ϵ_t is assumed to have a covariance matrix $\Sigma_\epsilon \in \mathbb{R}^{d \times d}$, which is symmetric.

Explanation of the formula for VAR parameters

2. **Covariance matrix of the noise:**
   - The noise term ϵ_t is assumed to have a covariance matrix $\Sigma_\epsilon \in \mathbb{R}^{d \times d}$, which is symmetric.
   - A symmetric $d \times d$ matrix has $\frac{d(d+1)}{2}$ independent entries (diagonal entries contribute d, and the upper or lower triangular part contributes $\frac{d(d-1)}{2}$).

Explanation of the formula for VAR parameters

2. **Covariance matrix of the noise:**
   - The noise term ϵ_t is assumed to have a covariance matrix $\Sigma_\epsilon \in \mathbb{R}^{d\times d}$, which is symmetric.
   - A symmetric $d \times d$ matrix has $\frac{d(d+1)}{2}$ independent entries (diagonal entries contribute d, and the upper or lower triangular part contributes $\frac{d(d-1)}{2}$).
   - Thus, the covariance matrix contributes $\frac{d(d+1)}{2}$ parameters.

Explanation of the formula for VAR parameters

2. **Covariance matrix of the noise:**
   - The noise term ϵ_t is assumed to have a covariance matrix $\Sigma_\epsilon \in \mathbb{R}^{d\times d}$, which is symmetric.
   - A symmetric $d \times d$ matrix has $\frac{d(d+1)}{2}$ independent entries (diagonal entries contribute d, and the upper or lower triangular part contributes $\frac{d(d-1)}{2}$).
   - Thus, the covariance matrix contributes $\frac{d(d+1)}{2}$ parameters.

Total number of parameters:
Combining the contributions from the coefficient matrices (pd^2) and the covariance matrix $\frac{d(d+1)}{2}$, the total number of parameters in the VAR(p) model is:

$$pd^2 + \frac{d(d+1)}{2}.$$

2. Estimation of sparse VAR through LASSO

- So the number of VAR(p) parameters (with dimension d and order p) to estimate (assuming zero mean) is:

$$pd^2 + \frac{d(d+1)}{2}.$$

2. Estimation of sparse VAR through LASSO

- So the number of VAR(p) parameters (with dimension d and order p) to estimate (assuming zero mean) is:

$$pd^2 + \frac{d(d+1)}{2}.$$

- This can be pretty large even for moderate d's.

$$pd^2 + \frac{d(d+1)}{2} > (T-p)d.$$

2. Estimation of sparse VAR through LASSO

- So the number of VAR(p) parameters (with dimension d and order p) to estimate (assuming zero mean) is:

$$pd^2 + \frac{d(d+1)}{2}.$$

- This can be pretty large even for moderate d's.

$$pd^2 + \frac{d(d+1)}{2} > (T-p)d.$$

- E.g., with $p = 2$, $d = 51$, we get $pd^2 + d(d+1)/2 = 6,477$.

2. Estimation of sparse VAR through LASSO

- So the number of VAR(p) parameters (with dimension d and order p) to estimate (assuming zero mean) is:

$$pd^2 + \frac{d(d+1)}{2}.$$

- This can be pretty large even for moderate d's.

$$pd^2 + \frac{d(d+1)}{2} > (T-p)d.$$

- E.g., with $p = 2$, $d = 51$, we get $pd^2 + d(d+1)/2 = 6,477$.

The goal is to rewrite the VAR

$$X_t = \Phi_1 X_{t-1} + \ldots + \Phi_p X_{t-p} + \epsilon_t, \quad t = 1, \ldots, T,$$

as a linear regression model and use LASSO type regularization.

2. Estimation of sparse VAR through LASSO

- Write the (observed) VAR(p) model in a linear form as:

$$\underbrace{\begin{pmatrix} X'_{p+1} \\ X'_{p+2} \\ \vdots \\ X'_T \end{pmatrix}}_{\mathcal{Y}} = \underbrace{\begin{pmatrix} X'_p & \cdots & X'_1 \\ X'_{p+1} & \cdots & X'_2 \\ \vdots & \ddots & \vdots \\ X'_{T-1} & \cdots & X'_{T-p} \end{pmatrix}}_{\mathcal{X}} \cdot \underbrace{\begin{pmatrix} \Phi'_1 \\ \Phi'_2 \\ \vdots \\ \Phi'_p \end{pmatrix}}_{\mathcal{B}^*} + \underbrace{\begin{pmatrix} \epsilon'_{p+1} \\ \epsilon'_{p+2} \\ \vdots \\ \epsilon'_T \end{pmatrix}}_{\mathcal{E}}$$

$$\text{vec}(\mathcal{Y}) = (I_d \otimes \mathcal{X})\text{vec}(\mathcal{B}^*) + \text{vec}(\mathcal{E})$$

$$\underbrace{Y}_{Nd\times 1} = \underbrace{Z}_{Nd\times q} \underbrace{\beta^*}_{q\times 1} + \underbrace{E}_{Nd\times 1}$$

with $N = T - p$ and $q = pd^2$.

2. Estimation of sparse VAR through LASSO

- Write the (observed) VAR(p) model in a linear form as:

$$\underbrace{\begin{pmatrix} X'_{p+1} \\ X'_{p+2} \\ \vdots \\ X'_T \end{pmatrix}}_{\mathcal{Y}} = \underbrace{\begin{pmatrix} X'_p & \cdots & X'_1 \\ X'_{p+1} & \cdots & X'_2 \\ \vdots & \ddots & \vdots \\ X'_{T-1} & \cdots & X'_{T-p} \end{pmatrix}}_{\mathcal{X}} \cdot \underbrace{\begin{pmatrix} \Phi'_1 \\ \Phi'_2 \\ \vdots \\ \Phi'_p \end{pmatrix}}_{\mathcal{B}^*} + \underbrace{\begin{pmatrix} \epsilon'_{p+1} \\ \epsilon'_{p+2} \\ \vdots \\ \epsilon'_T \end{pmatrix}}_{\mathcal{E}}$$

$$\mathrm{vec}(\mathcal{Y}) = (I_d \otimes \mathcal{X})\mathrm{vec}(\mathcal{B}^*) + \mathrm{vec}(\mathcal{E})$$

$$\underbrace{Y}_{Nd\times 1} = \underbrace{Z}_{Nd\times q} \underbrace{\beta^*}_{q\times 1} + \underbrace{E}_{Nd\times 1}$$

  with $N = T - p$ and $q = pd^2$.
- The $(p+1)^{th}$ observation representation is given by:

$$X'_{p+1} = X'_p\Phi'_1 + ... + X'_1\Phi'_p + \epsilon'_{p+1}$$

- The T^{th} observation representation is given by:

$$X'_T = X'_{T-1}\Phi'_1 + ... + X'_{T-p}\Phi'_p + \epsilon'_T$$

2. Estimation of sparse VAR through LASSO

- **Sparsity:** Assume β^* is s-sparse, i.e. $\|\beta^*\|_0 = \sum_{j=1}^{p} \|\mathsf{Vec}(\Phi_j)\|_0 = s$.
- **LASSO estimator:**

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^q}{\text{argmin}} \frac{1}{N} \|Y - Z\beta\|_2^2 + \lambda_N \|\beta\|_1$$

where $\lambda_N > 0$ is a penalty parameter, $\|\beta\|_1 = \sum_{i=1}^{q} |\beta_i|$ for $\beta = (\beta_1, \ldots, \beta_q)'$, and $q = pd^2$.

2. Estimation of sparse VAR through LASSO

- **Sparsity:** Assume β^* is s-sparse, i.e. $\|\beta^*\|_0 = \sum_{j=1}^{p} \|\mathrm{Vec}(\Phi_j)\|_0 = s$.
- **LASSO estimator:**

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^q}{\mathrm{argmin}} \frac{1}{N} \|Y - Z\beta\|_2^2 + \lambda_N \|\beta\|_1$$

  where $\lambda_N > 0$ is a penalty parameter, $\|\beta\|_1 = \sum_{i=1}^{q} |\beta_i|$ for $\beta = (\beta_1, \ldots, \beta_q)'$, and $q = pd^2$.
- When $\lambda_N = 0$, this is the OLS.

2. Estimation of sparse VAR through LASSO

From $\|Y - Z\beta\|_2^2 = (Y - Z\beta)'(Y - Z\beta)$,

we write the LASSO optimization problem as

$$\underset{\beta \in \mathbb{R}^q}{\text{argmin}} - 2\beta'\widehat{\gamma} + \beta'\widehat{\Gamma}\beta + \lambda_N \|\beta\|_1$$

2. Estimation of sparse VAR through LASSO

From $\|Y - Z\beta\|_2^2 = (Y - Z\beta)'(Y - Z\beta)$,

we write the LASSO optimization problem as

$$\underset{\beta \in \mathbb{R}^q}{\operatorname{argmin}} - 2\beta'\widehat{\gamma} + \beta'\widehat{\Gamma}\beta + \lambda_N \|\beta\|_1$$

where

$$\underbrace{\widehat{\Gamma}}_{q \times q} = \frac{1}{N} Z'Z = \frac{1}{N}(I_d \otimes X'X)\ , \ \underbrace{\widehat{\gamma}}_{q \times 1} = \frac{1}{N} Z'Y = \frac{1}{N}(I_d \otimes X')Y,$$

$$Z' = (I_d \otimes X')\ ,\ Z = (I_d \otimes X)$$

$$Z'Z = (I_d \otimes X')(I_d \otimes X) = (I_d I_d \otimes X'X) = (I_d \otimes X'X)$$

3. Theoretical perspective

The next two technical conditions will be used:

- Restricted Eigenvalue: the $q \times q$ symmetric matrix $\widehat{\Gamma}$ satisfies

$$\theta' \widehat{\Gamma} \theta \geq \alpha \|\theta\|_2^2 - \tau \|\theta\|_1^2 \,, \quad \theta \in^q$$

  with "curvature" $\alpha > 0$ and "tolerance" $\tau > 0$. We use the following notation: $\widehat{\Gamma} \sim RE(\alpha, \tau)$.

3. Theoretical perspective

The next two technical conditions will be used:

- Restricted Eigenvalue: the $q \times q$ symmetric matrix $\widehat{\Gamma}$ satisfies

$$\theta'\widehat{\Gamma}\theta \geq \alpha\|\theta\|_2^2 - \tau\|\theta\|_1^2 \ , \quad \theta \in^q$$

  with "curvature" $\alpha > 0$ and "tolerance" $\tau > 0$. We use the following notation: $\widehat{\Gamma} \sim RE(\alpha, \tau)$.

- Deviation: for a deterministic function $Q(\beta^*, \Sigma_\epsilon)$,

$$\|\widehat{\gamma} - \widehat{\Gamma}\beta^*\|_\infty \leq Q(\beta^*, \Sigma_\epsilon)\sqrt{\frac{\log q}{N}}$$

3. Theoretical perspective

The next two technical conditions will be used:

- Restricted Eigenvalue: the $q \times q$ symmetric matrix $\widehat{\Gamma}$ satisfies

$$\theta' \widehat{\Gamma} \theta \geq \alpha \|\theta\|_2^2 - \tau \|\theta\|_1^2 \ , \quad \theta \in^q$$

  with "curvature" $\alpha > 0$ and "tolerance" $\tau > 0$. We use the following notation: $\widehat{\Gamma} \sim RE(\alpha, \tau)$.

- Deviation: for a deterministic function $Q(\beta^*, \Sigma_\epsilon)$,

$$\|\widehat{\gamma} - \widehat{\Gamma}\beta^*\|_\infty \leq Q(\beta^*, \Sigma_\epsilon)\sqrt{\frac{\log q}{N}}$$

Lemma

Suppose sparsity is given and some more technical assumptions are satisfied. Then, for any $\lambda_N \geq 4Q(\beta^*, \Sigma_\epsilon)\sqrt{\frac{\log q}{N}}$, any LASSO solution $\widehat{\beta}$ satisfies

$$\|\widehat{\beta} - \beta^*\|_2 = O\left(\sqrt{\frac{s \log q}{N}}\right).$$

4. Summary

Some things to remember:

- Not all data are i.i.d.
- The LASSO estimator can be used for a variety of problems.

Part VII

Summary and essential points

Part 0

Discrete distribution

Set of possible outcomes is discrete.

$$\mathbb{P}(X \leq a) = \sum_{x \leq a} \mathbb{P}(X = x)$$

$$\mathbb{E}(X) = \sum_{x \in A} x \cdot \mathbb{P}(X = x) = \sum_{x \in A} x \cdot p_X(x)$$

$$\begin{aligned} \mathbb{E}(g(X)) &= \sum_{x \in A} g(x) \cdot \mathbb{P}(X = x) \\ &= \sum_{x \in A} g(x) \cdot p_X(x) \end{aligned}$$

$$Var(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$$

$$sd(X) = \sqrt{Var(X)}$$

Continuous distribution

Takes real numbers in an interval.

$$\mathbb{P}(X \leq a) = \int_{-\infty}^{a} f_X(x)\, dx$$

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$$

$$Var(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$$

$$sd(X) = \sqrt{Var(X)}$$

Part 0

- $Bias(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \theta.$
- $\mathsf{MSE}(\widehat{\theta}) = \mathbb{E}[(\widehat{\theta} - \theta)^2]$
- $\mathsf{MSE}(\widehat{\theta}) = \mathsf{Bias}^2(\widehat{\theta}) + Var(\widehat{\theta})$, where $Var(\widehat{\theta}) = \mathbb{E}[(\widehat{\theta} - \mathbb{E}(\widehat{\theta}))^2]$

- Why do we want the MSE to be small?

- What does the formula for MSE tell us?

Part 0

Multivariate

- $Z = (Z_1, \ldots, Z_p)^T$ is called p-variate standard normal random vector if Z_i iid normal for each $i = 1, \ldots, p$.

- 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 $X = A^{\mathrm{T}} Z + \mu$.

- Linear combinations are also normally distributed

- $\mathrm{Cov}(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^{\mathrm{T}}\right].$

Part 0: Practice assignments

1. Show that $Var(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$
2. Show that $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ for continuous random variables X, Y.
3. Show that $\text{MSE}(\widehat{\theta}) = \text{Bias}^2(\widehat{\theta}) + Var(\widehat{\theta})$.
4. Let b be a $p \times 1$ vector of constants, B a $k \times d$ matrix of constants, and $X \sim \mathcal{N}_p(\mu, \Sigma)$. Then

$$b + BX \sim \mathcal{N}_p(B\mu + b, B\Sigma B')$$

5. Show that $\mathbb{P}(X > t) = \mathbb{E}(\mathbb{1}_{X>t})$ for a continuous random variable X.

Part I

What are the different view points?

- Classical asymptotics.
- High-dimensional asymptotics.
- Non-asymptotic bounds.

What can go wrong in highdimensions?

- no consistent estimator
- low rank matrices, not invertible

What can help?

- Finding or imposing lower dimensional structure
- sparsity

Part II

- **Markov's inequality:** Assume $X \geq 0$

$$\mathbb{P}[X > t] \leq \frac{\mathbb{E}(X)}{t}, \quad \forall t > 0.$$

- **Chebyshev's:** Assume $\mathbb{E}(X^2) < \infty$

$$\mathbb{P}[|X - \mathbb{E}(X)| > t] \leq \frac{Var(X)}{t^2}, \quad \forall t > 0.$$

- **Chernoff's inequality:** Let X_i be independent Bernoulli random variables with parameters p_i. Consider their sum $S_N = \sum_{i=1}^{N} X_i$ and denote its mean by $\mu = \mathbb{E}(S_N)$. Then, for any $t > \mu$, we have

$$\mathbb{P}(S_N \geq t) \leq \exp(-\mu) \left(\frac{\exp(1)\mu}{t} \right)^t.$$

Part II

- A random variable X with finite mean μ is *sub-Gaussian* with parameter $\sigma > 0$ if
$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\sigma^2\lambda^2/2}, \quad \forall \lambda \in .$$
We say that X is σ-sub-Gaussian and say it has *variance proxy* σ^2.
- If a random variable X with finite mean μ is σ-sub-Gaussian, then
$$\mathbb{P}[|X-\mu| \geq t] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right), \quad \forall t \in .$$
- Sum of independent sub-Gaussian random variables is sub-Gaussian.
- **Hoeffding:** Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables with variance proxies $\sigma_1^2, \ldots, \sigma_n^2$, then $Z = \sum_{i=1}^n X_i$ satisfies the tail bound
$$\mathbb{P}[|Z-\mathbb{E}(Z)| \geq t] \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right), \quad \forall t \in .$$

Part II

- What is concentration?
- What are possible deviations of interest?
- How is the moment generating function defined?
- The tails of a sub-Gaussian distribution are dominated by the tails of what distribution?
- Are distributions with heavy tails also sub-Gaussian?

Part III

Given the observations (y, X)
OLS: $\widehat{\beta} = (X^T X)^{-1} X^T y$.
Ridge estimator: For any $\lambda \geq 0$, set

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

For any $\lambda > 0$, the solution to the minimization problem is

$$\widehat{\beta} = (X'X + \lambda I_{p \times p})^{-1} X'y.$$

Some properties:

- $\text{Bias}(\widehat{\beta}) = \mathbb{E}\widehat{\beta} - \beta_0 = -\lambda(X'X + \lambda I)^{-1}\beta$
- $Var\widehat{\beta} = \mathbb{E}[(\widehat{\beta} - E(\widehat{\beta}))^2] = (X'X + \lambda I)^{-1} X'X (X'X + \lambda I)^{-1} \sigma^2$
- $\text{MSE}(\widehat{\beta}) = \text{Bias}^2(\widehat{\theta}) + Var(\widehat{\theta}) = \text{tr}((X'X + \lambda I)^{-2}(\lambda^2 \beta\beta' + \sigma^2 X'X))$

LASSO estimator:

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \|y - X\beta\|_2^2 + \lambda_N \|\beta\|_1$$

Part III

- Ridge:
  - penalizes with -norm
  - does it have explicit representation?
  - Is it convex?
  - does it do model selection?
- Best subset selection:
  - penalizes with -norm
  - does it have explicit representation?
  - Is it convex?
- LASSO:
  - penalizes with -norm
  - Is it convex?
  - does it have explicit representation?
  - does it do model selection?

Part III

What is the bias-variance trade-off?

- The bias increases as λ (amount of shrinkage) increases.
- The variance decreases as λ (amount of shrinkage) increases.

Is the following function convex in β?

$$\|y - X\beta\|_2^2 + \lambda_{1,N}\|\beta\|_1 + \lambda_{2,N}\|\beta\|_2$$

Part III

What are the different error metrics to access the estimators' performances?

- Prediction error
- Parametric error
- Variable selection

Important facts:

- parameter consistency is not the same as consistent support recovery!
- the choice $\lambda > \sqrt{\frac{\log p}{n}}$ is a convenient choice to ensure that the respective metric becomes small with large probability!

Part III

Why is it important to choose λ?

-
-

Cross-Validation:

- What are the major steps?
- regular cross-validation vs. one standard error rule
- What gives sparser solution, cross-validation or one step error cross validation?

Part IV

1. Hard-thresholding

$$T_\lambda(u) = u\mathbb{1}_{\{|u|\geq\lambda\}},$$

2. Soft-thresholding

$$S_\lambda(u) = \mathsf{sign}(u)(|u|-\lambda)\mathbb{1}_{\{|u|\geq\lambda\}}.$$

- Why does one use thresholding rather than penalization?

Part V

- Independence in Gaussians is determined by sparsity pattern of the covariance Σ.
  - Sparsity pattern: "where the non-zeroes are".
  - $X_i \perp\!\!\!\perp X_j$ iff $\Sigma_{ij} = 0$.

- Gaussians' conditional independence: sparsity of the precision matrix, $\Theta = \Sigma^{-1}$.
  - $X_i \perp\!\!\!\perp X_j \mid \{X_k | k \notin \{i, j\}\}$ iff $\Theta_{ij} = 0$.

- We use the sparsity pattern of Θ to define a graph.
  - Each node in the graph corresponds to a variable $j \in \{1, 2, \ldots, p\}$.
  - Each edge in the graph corresponds to a non-zero Θ_{ij}.

Part V

- Checking independence and conditional independence using the graph:
  - Independence: $X_i \perp\!\!\!\perp X_j$ if no path exists between X_i and X_j in the graph of Σ.
  - Conditional Independence: $X_i \perp\!\!\!\perp X_j \mid X_k$ if X_k blocks all paths from X_i to X_j in the graph of Θ.

Part V

- 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?
- 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?

Part VI

See Lecture 11 and Lecture 12