Statistical Learning and Data Science

Xinwei Deng
xdeng@vt.edu
Department of Statistics

Course Agenda

- Presentation and discussion
- Choose topics and literature review
- Conduct a mini-research project

Statistical Perspectives for Modern Data Analytics

- Matrix: an arrangement of multiple rows (or columns)

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ a_{31} & a_{32} & \cdots & a_{3n} \\ \cdots & \cdots & & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

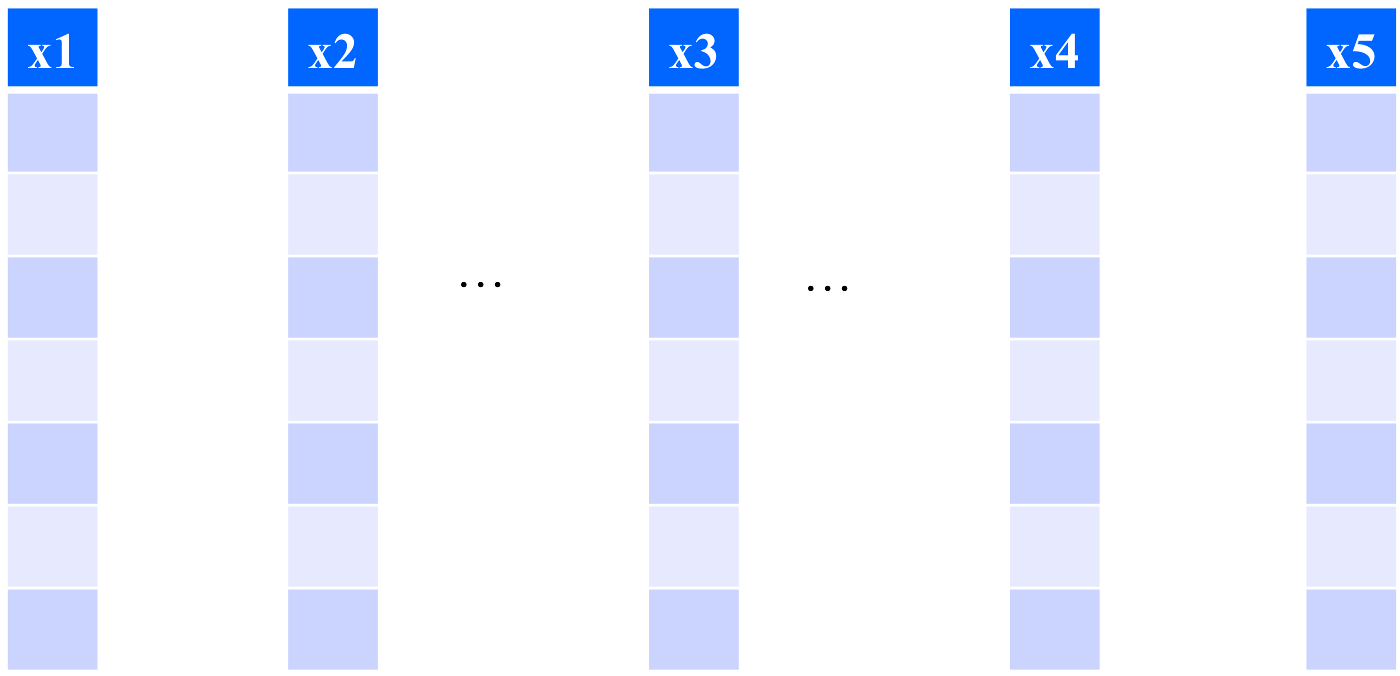
- From statistical perspective, it is mainly for data matrix, or some statistic of data matrix.
  - E.g., Covariance matrix
- The matrix algebra is useful and meaningful in modern data analytics.
  - E.g., matrix completion problem.

Statistical Perspectives for Modern Data Analytics (Con't)

- Big Data: Data is so-called Big with multiple rows, multiple columns, multiple types, multiple collection channels, etc.
- From statistical perspective, the challenges in Big Data mainly come from dependency.
  - Dependency among rows.
  - Dependency among columns.
  - Dependency between past-time and current-time data points.
  - Dependency between computation efficiency and estimation accuracy.

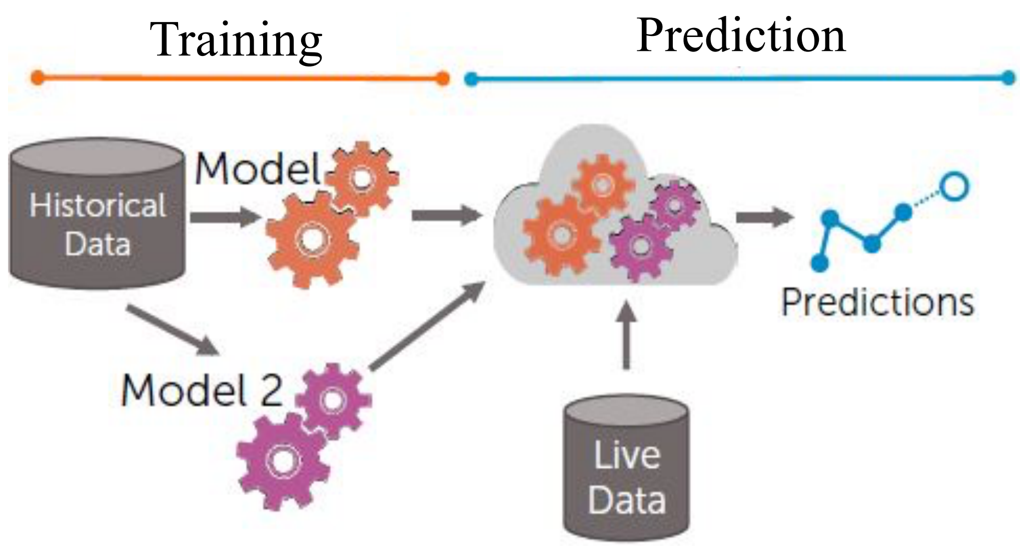
Illustration of Dependency

- In statistics, the challenges in Big Data mainly come from dependency.
  - Dependency among rows and dependency among columns.



Statistical Perspectives for Modern Data Analytics (Con't)

- Data Analytics: In simple view, it is similar to machine learning.



- From statistical perspective, the key of data analytics is to understand the data generation mechanism:
  - Enable model inference.
  - Enable model prediction.

Potential Topics 1

- Multi-response Regression and Model Selection
  - Structured Lasso with multiple responses.
  - Multi-response with consideration of covariance matrix.

Potential Topics 2

- Discriminant Analysis (DA) for Classification
  - LDA and QDA.
  - Regularized discriminant analysis.
  - Discriminant analysis based approach for classification.

Potential Topics 3

- Graphical Model via Covariance Matrix Estimation
  - Modified Cholesky decomposition approach.
  - Structured graphical model estimation.
  - Multiple graphical models.
  - Nonparametric approach for graphical model estimation.

Potential Topics 4

- Spatial-Temporal Modeling
  - Gaussian process modeling in computer experiment.
  - Multivariate time series with spatial correction structure.

Potential Topics 5

- Interface between machine learning and experimental design
  - Effective data collection for efficient modeling.
  - Active learning and sequential design.
  - Embrace experimental design thinking for large-scale statistical analysis.

Potential Topics 6

- Statistical computation thinking in multivariate analysis
  - Tradeoff between computation efficiency and estimation accuracy.
  - Optimization-based data analytics.
  - Statistical computation thinking for data science.

Scope of Data and Modeling

- Scalar → Vector → Matrix → Tensor
  – Many common concepts.
- From simple to advance
  – Simple regression to regularized tensor regression
  – Old techniques can be original.
- The responses can be continuous, discrete, or both.
  - From logistic regression to image classification.
- Modeling strategy:
  – Global vs Local modeling
  – Parametric vs Nonparametric methods.

The Multivariate Data

y_1	y_2	y_3	y_4	…	y_q

$\mathbf{Y} = (y_{ij})$ is $n \times q$ output matrix

x_1	x_2	x_3	…	x_k	x_{k+1}	…	…	$x_{p\text{-}1}$	x_p

$\mathbf{X} = (x_{ij})$ is $n \times p$ input matrix

Temporal/Functional Multivariate Data

y_1	y_2	y_3	y_4	…	y_q

$\mathbf{Y}(\mathrm{t}) = (y_{ij})$ is $n \times q$ output matrix

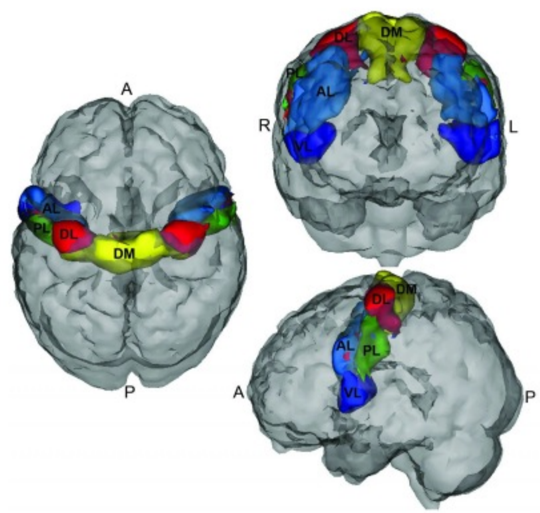
x_1	x_2	x_3	…	x_k	x_{k+1}	…	…	$x_{p\text{-}1}$	x_p

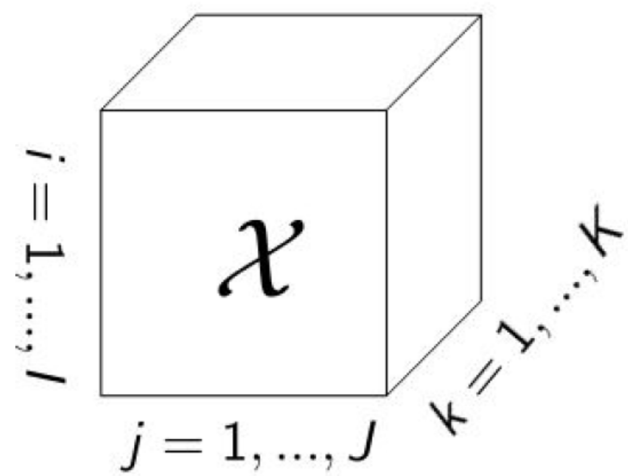
$\mathbf{X}(\mathrm{t}) = (x_{ij})$ is $n \times p$ input matrix

- At each time point, it is a multivariate data.

Tensor:
a Generalization of Matrix

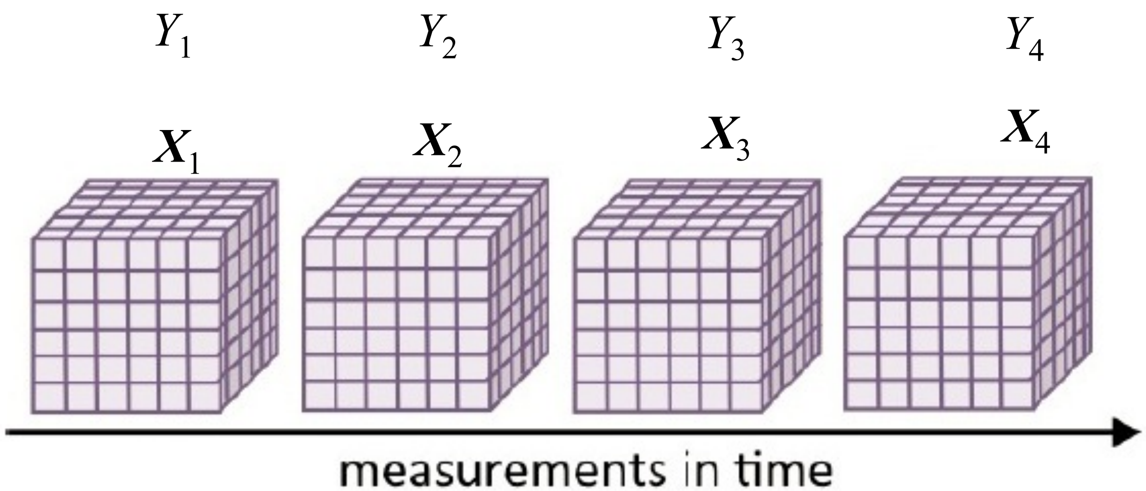
- Tensor: a ***multidimensional*** array.
- Example: fMRI data in Neuroscience: 3D brain images.
    - seek association between brain images (**X**) and clinical outcomes (Y).





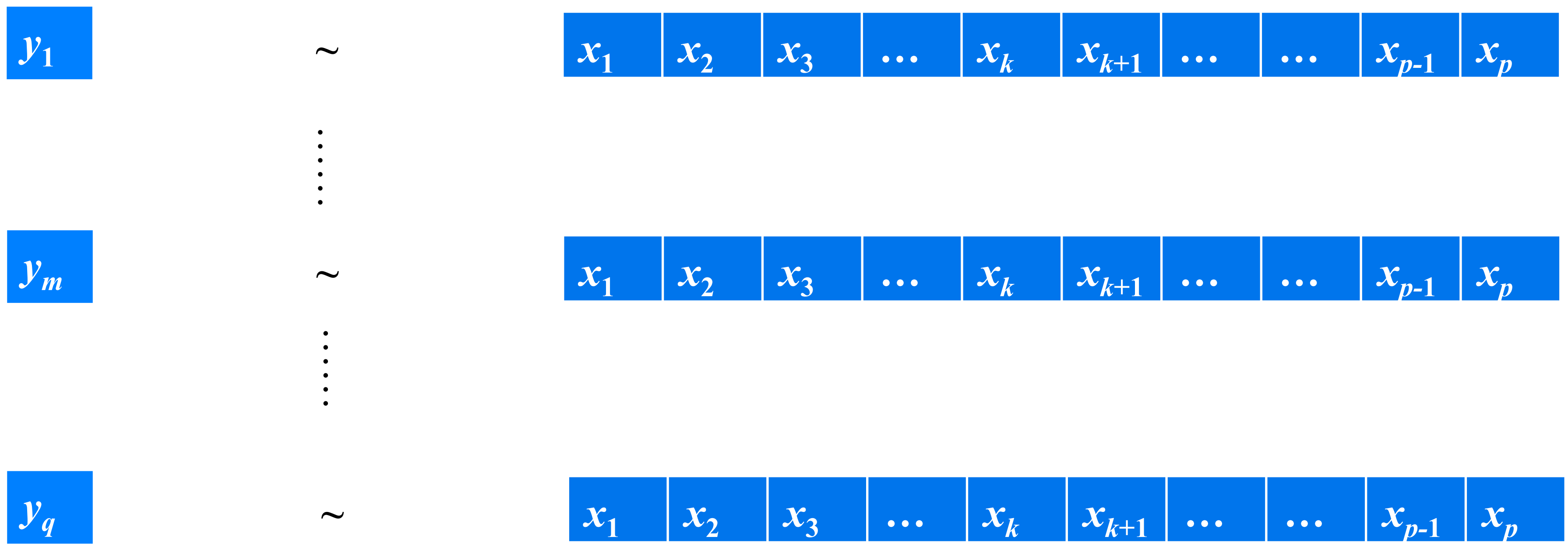
Tensor Regression

- Tensor along time is a generalization of spatial-temporal structure.
  - Goal: predict the response on the future time point.



Topic 1:

Multi-response Regression



The Linear Model

- Consider the linear model $y = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon,\ \epsilon \sim N(0, \sigma^2)$ where $\boldsymbol{x} = (x_1, \ldots, x_p)'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$.
- With data $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$, the log-likelihood is

$$
\begin{aligned}
L(\boldsymbol{\beta}, \sigma^2) &= \log\left\{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} (1/\sigma^2)^{1/2} \exp(-\frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2})\right\} \\
&\propto -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 \\
&= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})
\end{aligned}
$$

- The MLE of $\boldsymbol{\beta}$ is obtained by $\min -\log(L(\boldsymbol{\beta}, \sigma^2))$,

$$
\min_{\boldsymbol{\beta}, \sigma^2} \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).
$$

Some Remarks

- Estimating β in MLE is equivalent to the OLS estimation, i.e.,

$$\min_{\beta} LS(\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta)$$

- The negative log-likelihood or least squares both can be viewed as loss functions.
- Consider the regularization/penalty on the loss function, which is in the format of

$$\text{Loss Function} + \lambda \text{Penalty}$$

- The penalized likelihood approach is the same as the penalized least squares when the penalty only involves β.

Lasso for Model Selection

- The Lasso is to estimate β by

$$\hat{\beta}^{\text{Lasso}} = \text{argmin}_{\beta}\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|\}.$$

- Here is the penalty $P(\boldsymbol{\beta})$ is l_1 norm of $\boldsymbol{\beta}$, i.e. $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$.
- It is equivalent to the constraint/regularization problem:

$$\hat{\beta}^{\text{Lasso}} = \text{argmin}_{\beta}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq M$$

- The other related: adaptive Lasso, fused Lasso, and the generalized Lasso.

Group Lasso for Model Selection

- Consider the predictor variables have a group structure by

$$\boldsymbol{x} = (x_{11}, \dots, x_{1k_1}, \dots, x_{p1}, \dots, x_{pk_p}),$$

  where each group is $x_{j1}, \dots, x_{jk_j}$.
- Such group structures can be non-overlapped or overlapped.
- The linear model can be written as

$$y = \sum_{i=1}^{p} \sum_{j=1}^{k_i} x_{ij}\beta_{ij} + \epsilon.$$

- The penalty term also accommodate the group structure:
  - Yuan and Lin (2006): $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sqrt{\beta_{j1}^2 + \cdots + \beta_{jk_j}^2}$.
  - Zhao et al. (2006): $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} \max\{|\beta_{j1}|, \dots, |\beta_{jk_j}|\}$.
  - Group Bridge (Huang et. al., 2009) and Hierarchical LASSO (Zhou and Zhu, 2009).

Multi-response Linear Model

- In multi-response regression, the response vector $\boldsymbol{y} = (y_1, \ldots, y_q)'$ and the predictor vector $\boldsymbol{x} = (x_1, \ldots, x_p)'$.
- Consider the linear model

$$\boldsymbol{y} = \boldsymbol{B}'\boldsymbol{x} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}),$$

  where $\boldsymbol{B}$ is a $p \times q$ matrix of coefficients and the kth column β_k is coefficient vector associated with y_k regressing on $\boldsymbol{x}$.
- Remark 1: For these q regression models, different response variable with the same predictor variables.
- Remark 2: One can also consider the model $\boldsymbol{y}|\boldsymbol{x} \sim N(\boldsymbol{B}'\boldsymbol{x}, \boldsymbol{\Sigma})$. Here $\boldsymbol{\Sigma}$ reflects the correlation structure among q response variables $\boldsymbol{y} = (y_1, \ldots, y_q)'$.

Model Selection for Multi-response Model

- The log-likelihood function $L(\boldsymbol{B})$ is

$$\log\left\{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}|\boldsymbol{\Sigma}^{-1}|^{1/2}\exp\left(-\frac{(\boldsymbol{y}_i-\boldsymbol{B}'\boldsymbol{x}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{B}'\boldsymbol{x}_i)}{2}\right)\right\}$$

$$\propto \frac{n}{2}\log|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{B}'\boldsymbol{x}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{B}'\boldsymbol{x}_i)$$

  with $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$.

- Seeking sparse estimation of coefficient matrix $\boldsymbol{B}$, consider penalized regression based on the log-likelihood function.

$$\hat{B} = \arg\min \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{B}'\boldsymbol{x}_i)'(\boldsymbol{y}_i-\boldsymbol{B}'\boldsymbol{x}_i) + \lambda P(\boldsymbol{B})$$

$$= \arg\min \sum_{k=1}^{q}\frac{1}{n}\sum_{i=1}^{n}(y_{ik}-\boldsymbol{\beta}_k'\boldsymbol{x}_i)^2 + \lambda P(\boldsymbol{B}).$$

Penalty: Structured v.s. Non-Structured

- The coefficient matrix $\boldsymbol{B}$ is in a matrix format as

$$\boldsymbol{B} = \begin{pmatrix} \beta_{11} & \cdots & \beta_{q1} \\ \vdots & \ddots & \vdots \\ \beta_{1p} & \cdots & \beta_{qp} \end{pmatrix}.$$

- One can pursue
  - A sparse coefficient matrix estimate $\hat{\boldsymbol{B}}$.
  - Some rows of $\hat{\boldsymbol{B}}$ become zero vectors.
  - Other patterns of interest on the sparsity.
- To encourage some rows of $\boldsymbol{B}$ being zeros for reducing number of predictors in the model, consider the penalty

$$P(\boldsymbol{B}) = \lambda_1 \sum_{j=1}^{p} \sqrt{\sum_{k=1}^{q} \beta_{kj}^2} + \lambda_2 \sum_{j=1}^{p} \sum_{k=1}^{q} |\beta_{kj}|.$$

where λ_1 and λ_2 are tuning parameters.

Penalty: Hierarchical Structure via Re-parameterization

- Alternatively, we can consider a hierarchical structure to parameterize β_{kj} as

$$\beta_{jk} = \gamma_j \alpha_{kj}, \quad \gamma_j \geq 0.$$

- Such a parametrization provides the flexibility of obtaining a penalty with pursuing sparsity along the rows of $\boldsymbol{B}$.
- Specifically, we can consider the penalty function as follows:

$$P(\boldsymbol{B}) = \lambda_1 \sum_{j=1}^{p} \gamma_j + \lambda_2 \sum_{j=1}^{p} \sum_{k=1}^{q} |\alpha_{kj}|,$$

where λ_1 and λ_2 are tuning parameters.

Multi-response with Covariance Matrix Estimation

- The multi-response linear model $\boldsymbol{y} = \boldsymbol{B}'\boldsymbol{x} + \boldsymbol{\epsilon},\ \epsilon \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$.
- To estimate $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$, consider the negative log-likelihood which is

$$
\begin{aligned}
& -\frac{n}{2}\log|\boldsymbol{\Sigma}^{-1}| + \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i) \\
& = -\frac{n}{2}\log|\boldsymbol{\Sigma}^{-1}| + \frac{1}{2}tr[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})^T]
\end{aligned}
$$

- Seeking the sparsity on $\boldsymbol{B}$ with the consideration of $\boldsymbol{\Sigma}$, we can consider

$$
\begin{aligned}
\min & -\log|\boldsymbol{\Sigma}^{-1}| + \frac{1}{n}tr[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})^T] \\
& + \lambda_1 P_1(\boldsymbol{B}) + \lambda_2 P_2(\boldsymbol{\Sigma})
\end{aligned}
$$

A few Remarks

- The complex of penalty is closely related to the complex of the model estimation.
- The MRCE (Rothman et al., 2010) is to consider an overall sparse structure on **B** and $\mathbf{\Sigma}^{-1}$, which is

$$P_1(\boldsymbol{B}) = \sum_{i,j} |b_{ij}|, P_2(\boldsymbol{\Sigma}^{-1}) \sum_{i \neq j} |c_{ij}|,$$

where c_{ij} is the (i, j) element of $\boldsymbol{\Sigma}^{-1}$.

- Therefore, the MRCE is to consider

$$\min -\log |\boldsymbol{\Sigma}^{-1}| + \frac{1}{n} tr[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})^T]$$
$$+ \lambda_1 \sum_{i,j} |b_{ij}| + \lambda_2 \sum_{i \neq j} |c_{ij}|.$$

Another Angle from Multivariate t-Distribution

- The multivariate t-distribution can be viewed as a scaled multivariate normal.
- Specifically, denote $w \sim \Gamma(\nu/2, \nu/2)$, which is independent of $\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. If defining

$$\boldsymbol{y} = \boldsymbol{\mu} + w^{-1/2}\boldsymbol{z},$$

  Then $y \sim t(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$.
- Using the above result, one can develop multi-response regression using t-distribution as the random error, i.e.,

$$\boldsymbol{y} = \boldsymbol{B}'\boldsymbol{x} + \boldsymbol{\epsilon},\ \epsilon \sim t(\mathbf{0}, \boldsymbol{\Sigma}, \nu).$$

Another Angle from Multivariate t-Distribution (Con't)

- Under t-distribution as the random error, the multi-response linear model is

$$\boldsymbol{y} \sim t(\boldsymbol{B}'\boldsymbol{x}, \boldsymbol{\Sigma}, \nu) \Leftrightarrow \sqrt{w}(\boldsymbol{y} - \boldsymbol{B}'\boldsymbol{x})|w \sim N(0, \boldsymbol{\Sigma})$$

- In addition, one can show that $w|\boldsymbol{y}$ still has a Gamma distribution, which is

$$w|\boldsymbol{y} \sim \Gamma(\frac{\nu + p}{2}, \frac{\nu + d_{\boldsymbol{y}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{2})$$
$$\Rightarrow E(w|\boldsymbol{y}) = \frac{\nu + p}{\nu + d_{\boldsymbol{y}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}$$

  where $d_{\boldsymbol{y}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the square of Mahalanobis distance, $d_{\boldsymbol{y}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})$.
- Therefore, we can develop an EM algorithm to obtain the parameter estimation.

EM Algorithm for Parameter Estimation

- Suppose the data are $(\boldsymbol{x}_i, \boldsymbol{y}_i), i = 1, \ldots, n$, and $\boldsymbol{x} \in \mathcal{R}^p$, $\boldsymbol{y} \in \mathcal{R}^q$.
- With an initial estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we can develop an EM algorithm for parameter estimation. At kth iteration

E-STEP Given current estimator $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$, update the value of w_i,

$$w_i^{(k)} = \frac{\nu + p}{\nu + d_{\boldsymbol{i}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})},$$

where $d_{\boldsymbol{i}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{y} - \hat{\boldsymbol{\mu}})$.

M-STEP With the updated value of $w_i^{(k)}$, we can update the estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by

$$\min -\log|\boldsymbol{\Sigma}^{-1}| + \frac{1}{n}\sum_{i=1}^{n} w_i^{(k)} (\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i)$$
$$+ \lambda_1 \sum_{i,j} |b_{ij}| + \lambda_2 \sum_{i \neq j} |c_{ij}|.$$

Covariance Matrix Estimation

- The multi-response model is simplified as

$$\boldsymbol{y} = \mu + \epsilon, \epsilon \sim N(0, \boldsymbol{\Sigma}).$$

- With data $\boldsymbol{y}_i, i = 1, \ldots, n$, it is equivalent to

$$\boldsymbol{y}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), i = 1, \ldots, n.$$

- To estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, the log-likelihood function becomes

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \log \left\{ \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} |\boldsymbol{\Sigma}^{-1}|^{1/2} \exp\left(-\frac{(\boldsymbol{y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})}{2}\right) \right\} \\ &\propto \frac{n}{2} \log |\boldsymbol{\Sigma}^{-1}| - \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}) \\ &\propto \log |\boldsymbol{\Sigma}^{-1}| - tr[\boldsymbol{\Sigma}^{-1}\boldsymbol{S}], \end{aligned}$$

where

$$\boldsymbol{S} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^T.$$

Estimating $\boldsymbol{\Sigma}$

- Based on $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \frac{n}{2} \log |\boldsymbol{\Sigma}^{-1}| - \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu})$, it is easy to obtain the estimate of μ as

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}} = \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{y}_i.$$

- To estimate $\boldsymbol{\Sigma}$, take the first derivative with respect to $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$

$$\begin{aligned} & \frac{\partial}{\partial \boldsymbol{\Omega}} \{- \log |\boldsymbol{\Omega}| + tr[\boldsymbol{\Omega} \boldsymbol{S}]\} \\ & = -\boldsymbol{\Sigma} + \boldsymbol{S} = 0 \end{aligned}$$

  Therefore, the estimate of $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T.$$

- What if taking the first derivative w.r.t. $\boldsymbol{\Sigma}$? (Take-home)

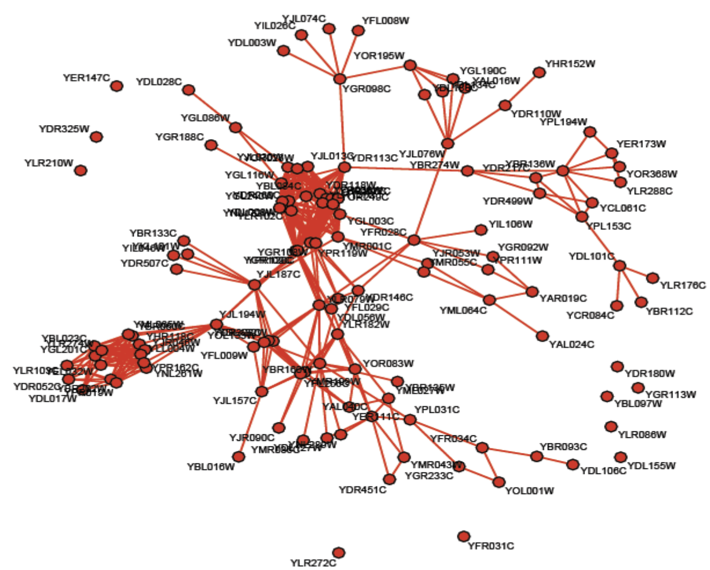
Why Estimating Σ^{-1}:
Gaussian Graphical Model

y_1	y_2	y_3	y_4	…	y_q

Data $\mathbf{Y} = (y_{ij})$ is an $n \times q$ matrix

$$\mathbf{\Omega} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1q} \\ c_{21} & c_{22} & \cdots & c_{2q} \\ \vdots & \ddots & \vdots & \vdots \\ c_{q1} & c_{q2} & \cdots & c_{qq} \end{pmatrix}$$

concentration matrix $\mathbf{\Sigma}^{-1} \equiv \mathbf{\Omega} = (c_{ij})$



- describe the **conditional dependency** among variables: if c_{ij} =0 zero, then variables i and j are conditionally independent given the other variables.

Thank you!

- **Questions and Comments?**