

STAT 24620=FINM 34700, STAT 32950

Jingshu Wang

Multivariate Statistical Analysis: Applications and Techniques

Lecture 1: Multivariate Data and Covariance Geometry

Outline

- Course overview and logistics
- Why multivariate analysis?
- Data matrix and geometry
- Covariance structure
- Dependence measures
- Preview of PCA

What this course is about

This course studies:

- Dependence among many variables
- Supervised and unsupervised learning
- Dimension reduction
- Linear and nonlinear modeling
- Structure in high-dimensional data

Key question:

- How do we understand and model many variables simultaneously?

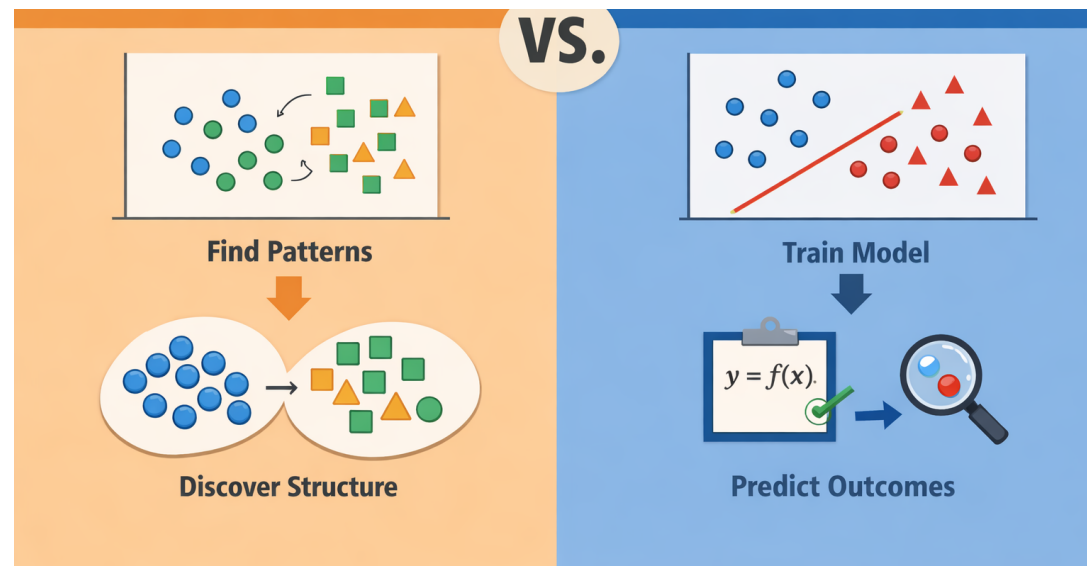
Supervised v.s. Unsupervised Learning

Unsupervised:

- PCA
- Factor models
- Clustering / mixture models
- CCA

Supervised:

- Regression
- Classification / LDA
- Ridge / Lasso
- Trees / Random forests



Assessment & Timeline

- Homework (3) — 35%
- Midterm Exam (Week 5, in-class) — 25%
- Group Project — 10%
- Final Exam (2-hour, in-class) — 30%

Exams: conceptual + interpretation

Project:

- Groups of 3–4 (Week 4)
- Data released Week 6
- Report due Week 9

Details on syllabus.

Why not analyze variables one at a time?

Finance

- Tech stocks often move together
- Market shocks affect many assets simultaneously

Macroeconomics

- Inflation, interest rates, and GDP co-move
- Policy and economic cycles create dependence

Gene expression data:

- Thousands of genes measured simultaneously
- Many genes are co-regulated and pathways drive joint variation

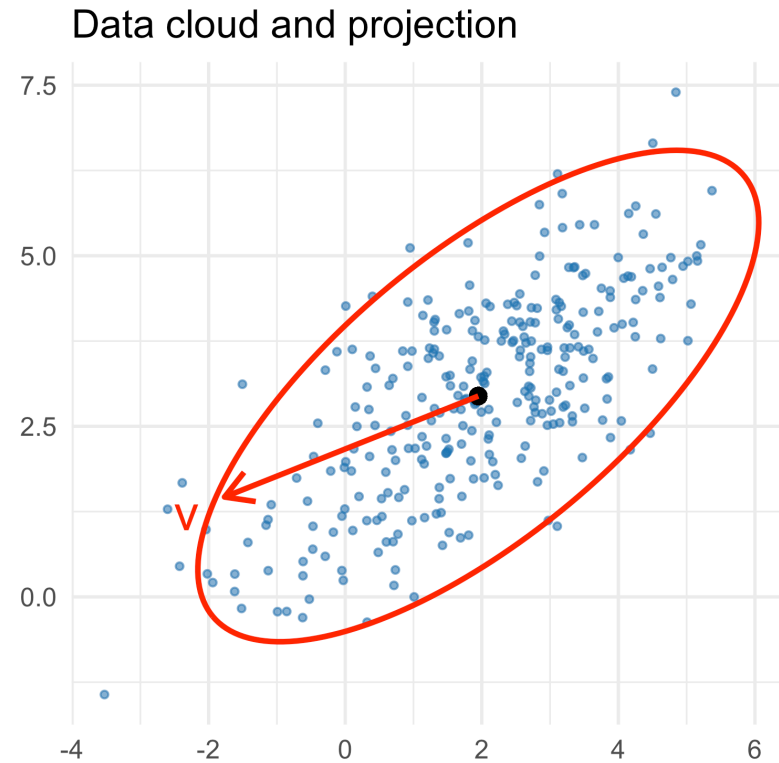
Social Science

- Education, income, and location are related
- Latent socio-economic structure exists

Key Question:

What structure do we miss if we ignore dependence?

Multivariate data as geometry



- Each observation is a vector (point) in \mathbb{R}^p
- Covariance describes how the cloud spreads in different directions
- For any direction \boldsymbol{v} :

$$\text{Var}(\boldsymbol{v}^\top X) = \boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v}$$

- Variability of the data along one direction ($X \in \mathbb{R}^p$, population random vector)

Presentation of data matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

- x_{ij} : Measurement of the j th variable on the i th observation (or subject)
- $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top \in \mathbb{R}^p =$ observed vector
- Rows = observations, Columns = variables
- $\mathbf{X} \in \mathbb{R}^{n \times p} =$ data matrix

- Sample mean vector $\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \frac{1}{n} \mathbf{X}^\top \mathbf{1}_n$

Centered data and sample covariance

- Centered data matrix: $\mathbf{X}_c = \mathbf{X} - \mathbf{1}_n \bar{\mathbf{x}}^\top$
- Sample covariance matrix

$$\mathbf{S} = \frac{1}{n-1} \mathbf{X}_c^\top \mathbf{X}_c$$

- Each element: $s_{jk} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$

- For a direction $\mathbf{v} \in \mathbb{R}^p$,

$$\mathbf{X}_c \mathbf{v} = \begin{bmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}})^\top \mathbf{v} \\ \vdots \\ (\mathbf{x}_n - \bar{\mathbf{x}})^\top \mathbf{v} \end{bmatrix}$$

gives the projected data points of centered observations

- What is the sample variance of these projected data points?

$$\frac{1}{n-1} (\mathbf{X}_c \mathbf{v})^\top (\mathbf{X}_c \mathbf{v}) = \mathbf{v}^\top \mathbf{S} \mathbf{v} = \frac{1}{n-1} \sum_{i=1}^n [\mathbf{v}^\top (\mathbf{x}_i - \bar{\mathbf{x}})]^2 \geq 0$$

- PCA will find direction that maximize this quantity

Sample correlation matrix

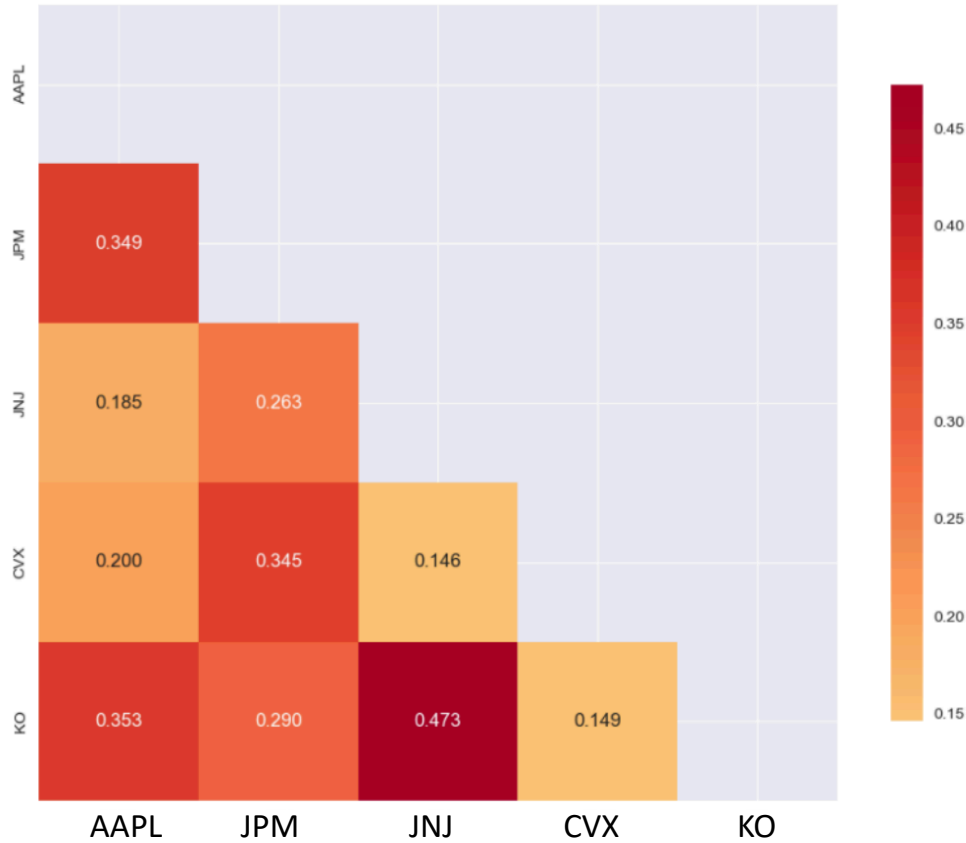
- Denote $s_{jj} = s_j^2$
- Sample correlation between the j th and k th variables is defined as

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} = \frac{s_{jk}}{s_j \times s_k}, \quad r_{jj} = 1$$

- Sample correlation matrix $\mathbf{R} = [r_{jk}]_{p \times p}$
 - Each element $r_{jk} \in [-1, 1]$
 - r_{jk} is a scale-invariant measure
 - r_{jk} is the Pearson correlation coefficient, measures linear and only linear correlation (later slides)
- Use \mathbf{S} or \mathbf{R} for dependence?
 - $\mathbf{R} = D^{-1}\mathbf{S}D^{-1}$, $\mathbf{S} = D\mathbf{R}D$
where $D = \text{diag}(s_1, \dots, s_p)$
 - When do we want to standardize?
 - Finance example: Different volatilities across stocks

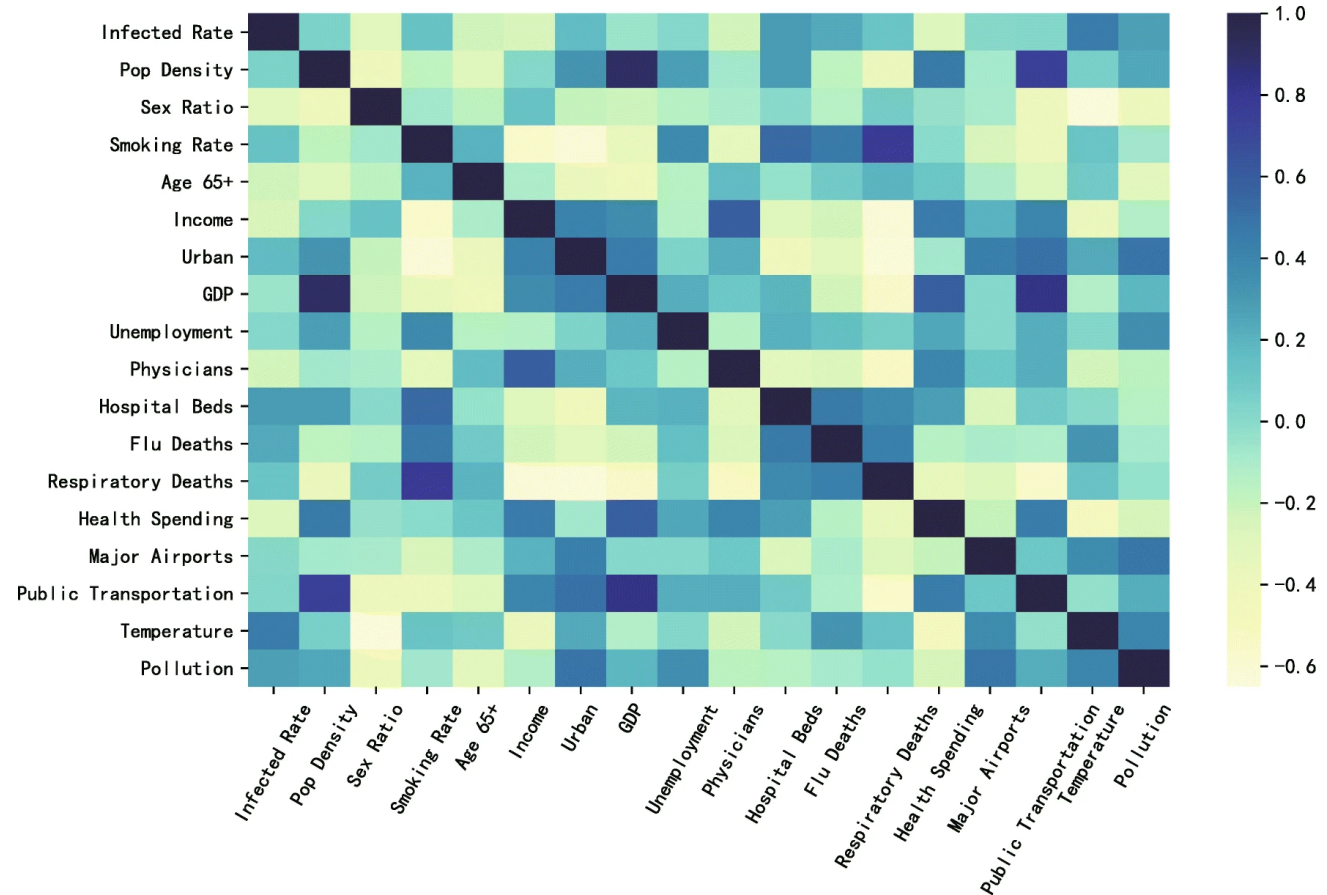
Example: visualize the sample correlation matrix

Stock Returns Correlation Matrix Year 2022-2024



<https://aemps.ewapub.com/article/view/28597.pdf>

- **Optimization problem:**
find a combination to maximize return but minimize variability
- What does high correlation imply?



https://link.springer.com/article/10.1007/s10489-021-02616-8?utm_source=researchgate.net&utm_medium=article

- The paper aims to predict COVID infection rate using domain features
- Some predictors have high correlations

Population v.s. sample

We can assume that the data points come from a distribution:

- Assume there are n p -dimensional random vectors:

$$X_1, \dots, X_n \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Population mean: $\mathbb{E}(X_i) = \boldsymbol{\mu} \in \mathbb{R}^p$
- Population covariance: $\text{Cov}(X_i) = \boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$

- Each data point \mathbf{x}_i is a realization of X_i
- Relationship between sample mean \bar{X} and $\boldsymbol{\mu}$? $\mathbb{E}(\bar{X}) = \boldsymbol{\mu}$
- Relationship between sample covariance \mathbf{S} and $\boldsymbol{\Sigma}$ (proof in class):
$$\mathbb{E}(\mathbf{S}) = \boldsymbol{\Sigma}$$
 - What if p is comparable to n ?

- Project X_i onto the direction of \mathbf{v} : $(X_i - \boldsymbol{\mu})^\top \mathbf{v}$
 - $\text{Var}[\mathbf{v}^\top (X_i - \boldsymbol{\mu})] = \text{Var}[\mathbf{v}^\top X_i] = \mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v}$

Positive semi-definiteness and eigen decomposition

- Sample covariance matrix and sample correlation matrix are symmetric

$$\mathbf{S} = \mathbf{S}^\top, \quad \mathbf{R} = \mathbf{R}^\top$$

- Positive semi-definiteness (PSD):

\mathbf{S} and \mathbf{R} are positive semi-definiteness, which means $\mathbf{v}^\top \mathbf{S} \mathbf{v} \geq 0$ and $\mathbf{v}^\top \mathbf{R} \mathbf{v} \geq 0$ for any $\mathbf{v} \in \mathbb{R}^p$

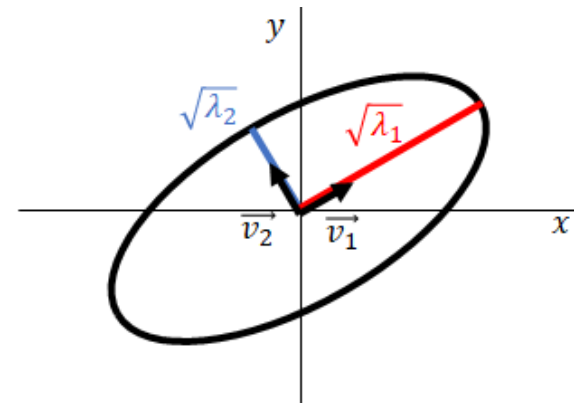
- What is $\mathbf{v}^\top \mathbf{S} \mathbf{v}$?

- How to show $\mathbf{v}^\top \mathbf{R} \mathbf{v} \geq 0$?

- Rescale each variable by the sample standard deviation $x_{ij}^\star = x_{ij}/s_j$
- What is the sample covariance matrix of the rescaled variables?

- Spectral structure: $\mathbf{S} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^\top$

- $\mathbf{V} \in \mathbb{R}^{p \times p}$: an orthogonal matrix
- $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$, each $\lambda_j \geq 0$
 - Why must eigenvalues be nonnegative?
- Eigenvectors = orthogonal directions
- Eigenvalues = variance in those directions



Rank & high-dimension

$$\text{rank}(\mathbf{S}) = \text{rank}(\mathbf{R}) = \text{rank}(\mathbf{X}_c) \leq \min(n - 1, p)$$

- Rank reflects the intrinsic dimension of \mathbf{X}_c
- What if $p \geq n$?

$$\text{rank}(\mathbf{S}) \leq n - 1 < p$$

- Consequences:
 - \mathbf{S} is singular
 - Some eigenvalues are zero
 - The centered data lies in a lower-dimensional subspace
- Positive definite v.s. semi-definite
 - \mathbf{S} is positive definite ($\mathbf{S} \succ 0$) if $\mathbf{v}^\top \mathbf{S} \mathbf{v} > 0$
 - Or equivalently, all eigenvalues $\lambda_j > 0$
 - The population covariance $\Sigma \succ 0$ but \mathbf{S} is not if $p \geq n$

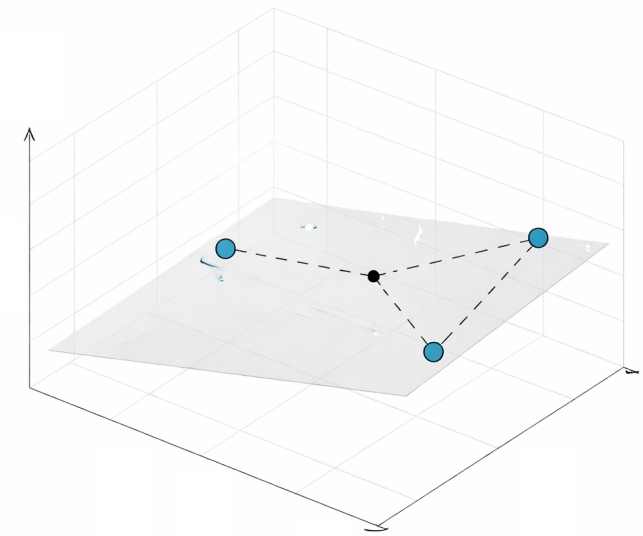
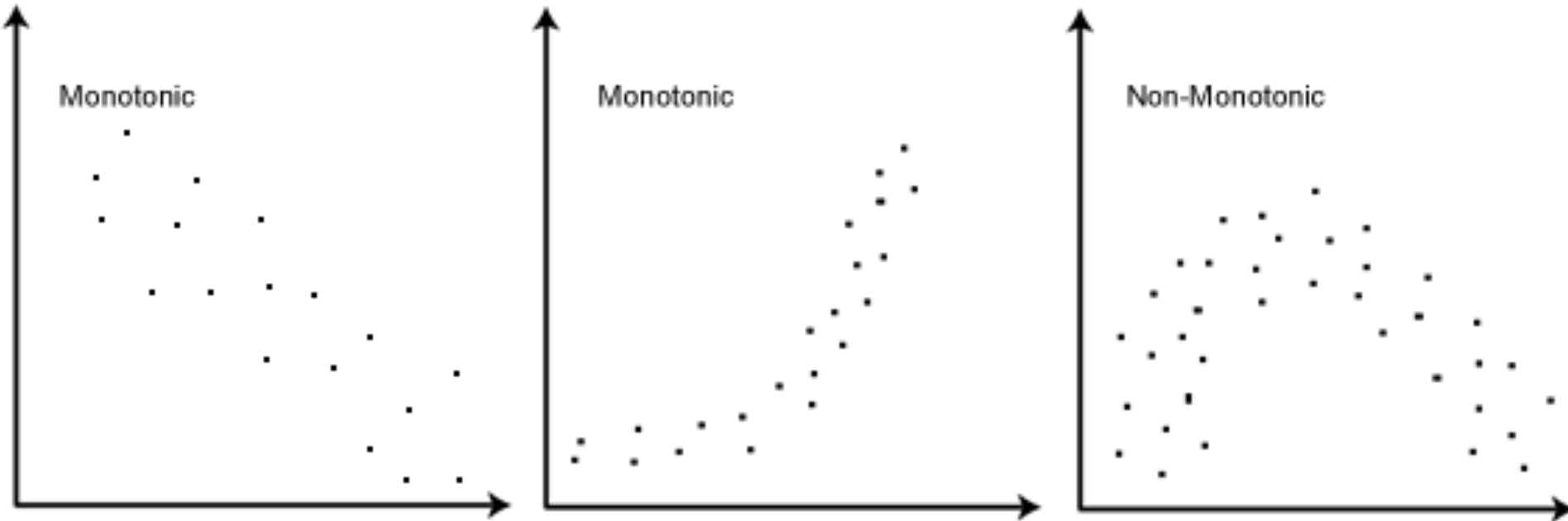


Illustration in $p = 3$

Alternative measures of pairwise dependence

- Pearson correlation only measures linear relationship
- Structure can be nonlinear monotonic, or non-monotonic



Alternative measures of pairwise dependence

- Kendall's τ :

- Kendall's rank coefficient based on concordant/discordant pairs

- For two observations (x_i, y_i) and (x_j, y_j)

- Concordant: $(x_i - x_j)(y_i - y_j) > 0$ / discordant: $(x_i - x_j)(y_i - y_j) < 0$

- Assume n_c concordant and n_d discordant pairs

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n - 1)}$$

- $\tau \in [-1, 1]$

- Only depends on relative ranking of observations within each variable

Alternative measures of pairwise dependence

- Spearman's rank correlation ρ :
 - First calculate the ranks of the observations within each variable
 - Then calculate the Pearson correlation of the ranks
- Key message:

Both Kendall's tau and Spearman correlation measure monotonic dependence, not just linear.

Subject i	(h_i, w_i)	h_i rank	w_i rank	Concordant pairs	Discordant pairs	Correlation
A	(150, 59.1)	1	2	AB, AC	AD, BC, BD, CD	Pearson's $r = 0.005$
B	(170, 61.0)	2	4			Kendall's $\tau = -1/3$
C	(180, 60.0)	3	3			Spearman's $\rho = -0.4$
D	(190, 59.0)	4	1			

Remarks

Two variables can be dependent or "related" in various manners. It is important to know the measure used to quantify the dependence structure.

- See R notebook for another example: [Lecture1_demo.nb.html](#)

How to perform dimension reduction?

- Let's think about reducing the data to one dimension, how to retain as much information as possible?

Population version

- For a random vector X with covariance Σ , which direction has the largest variance?

$$\max_{\|\mathbf{v}\|_2=1} \mathbf{v}^\top \Sigma \mathbf{v}$$

Sample version

- We do not know Σ , we estimate it with \mathbf{S}

$$\max_{\|\mathbf{v}\|_2=1} \mathbf{v}^\top \mathbf{S} \mathbf{v}$$

This is **PCA**, which maximizes empirical variance of linear projections.

Lecture 1 summary

- Multivariate data = points in \mathbb{R}^p
- Covariance describes how the data cloud spreads in different directions
- Variance along a direction \mathbf{v} : $\mathbf{v}^\top \Sigma \mathbf{v}$
- Sample covariance \mathbf{S}
- \mathbf{S} is symmetric and positive semi-definite
- Eigenvalues = variance in orthogonal directions
- Rank reflects intrinsic dimension
- Dependence can be linear (Pearson) or monotonic (Kendall / Spearman)

Next: We find the direction that maximizes variance → PCA

Suggested reading

- probReview.pdf
- Johnson & Wichern (6th edition): chapter 1-4.