InfoLab

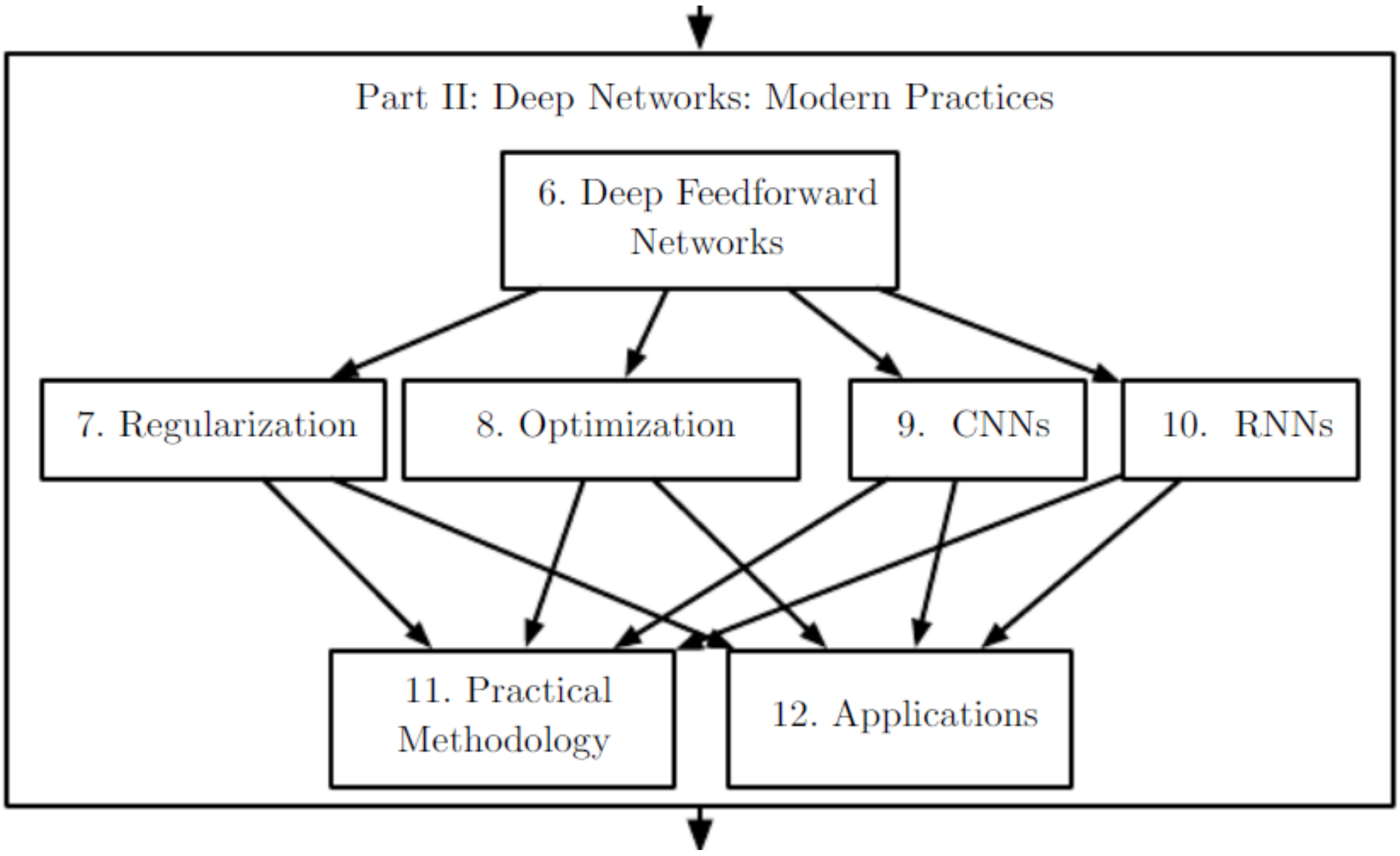
Deep Learning Seminar
CH 13. Linear Factor Models

2018-01-04

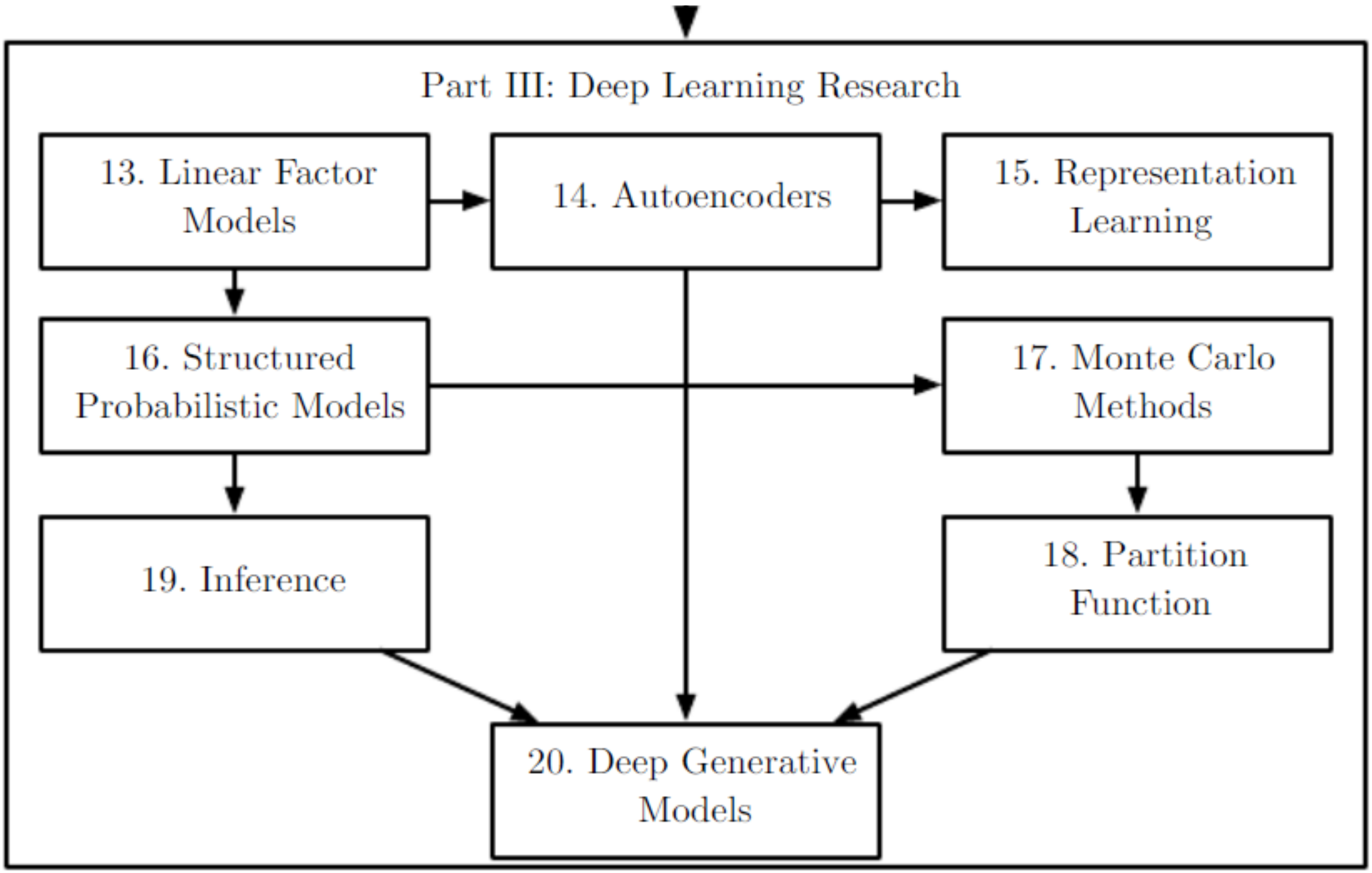
Eunjeong Yi

대구경북과학기술원
Daegu Gyeongbuk Institute of Science & Technology

Chapter Organization (Part 2)



Chapter Organization (Part 3)



Chapter 13. Linear Factor Models

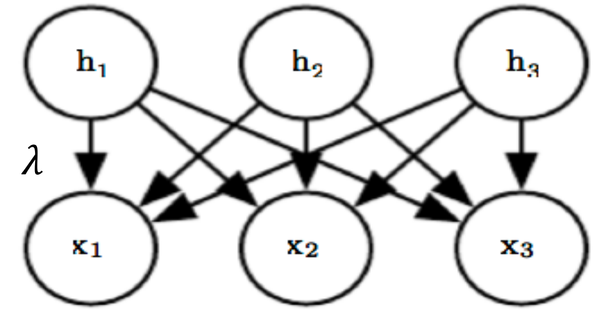
Factor analysis

- **The process of finding a factor assuming that the observed variables can represent a linear combination of factors**
- **Goal**
  - Find latent factor which can’t be measured
  - Find the relationship between latent variables and observed variables
- **Example: IQ**

Linear factor model

- **Observed variables** $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^T$

$$x_1 = \lambda_{11}h_1 + \lambda_{12}h_2 + \cdots + \lambda_{1m}h_m + b_1$$
$$x_2 = \lambda_{21}h_1 + \lambda_{22}h_2 + \cdots + \lambda_{2m}h_m + b_2$$
$$\vdots$$
$$x_n = \lambda_{n1}h_1 + \lambda_{n2}h_2 + \cdots + \lambda_{nm}h_m + b_n$$



- **Latent factors** $\boldsymbol{h} = (h_1, h_2, \cdots, h_m)^T$
- **Specific factors** $\boldsymbol{b} = (b_1, b_2, \cdots, b_m)^T$
- **Factor loadings** $\boldsymbol{\Lambda} = (\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \cdots, \boldsymbol{\lambda_n})^T$

$$(\boldsymbol{\lambda_i} = (\lambda_{i1}, \lambda_{i2}, \cdots, \lambda_{im}), i = 1,2, \cdots, n)$$

- **Matrix notation** $\boldsymbol{x} = \boldsymbol{\Lambda h} + \boldsymbol{b}$

Principle Component Analysis(PCA)

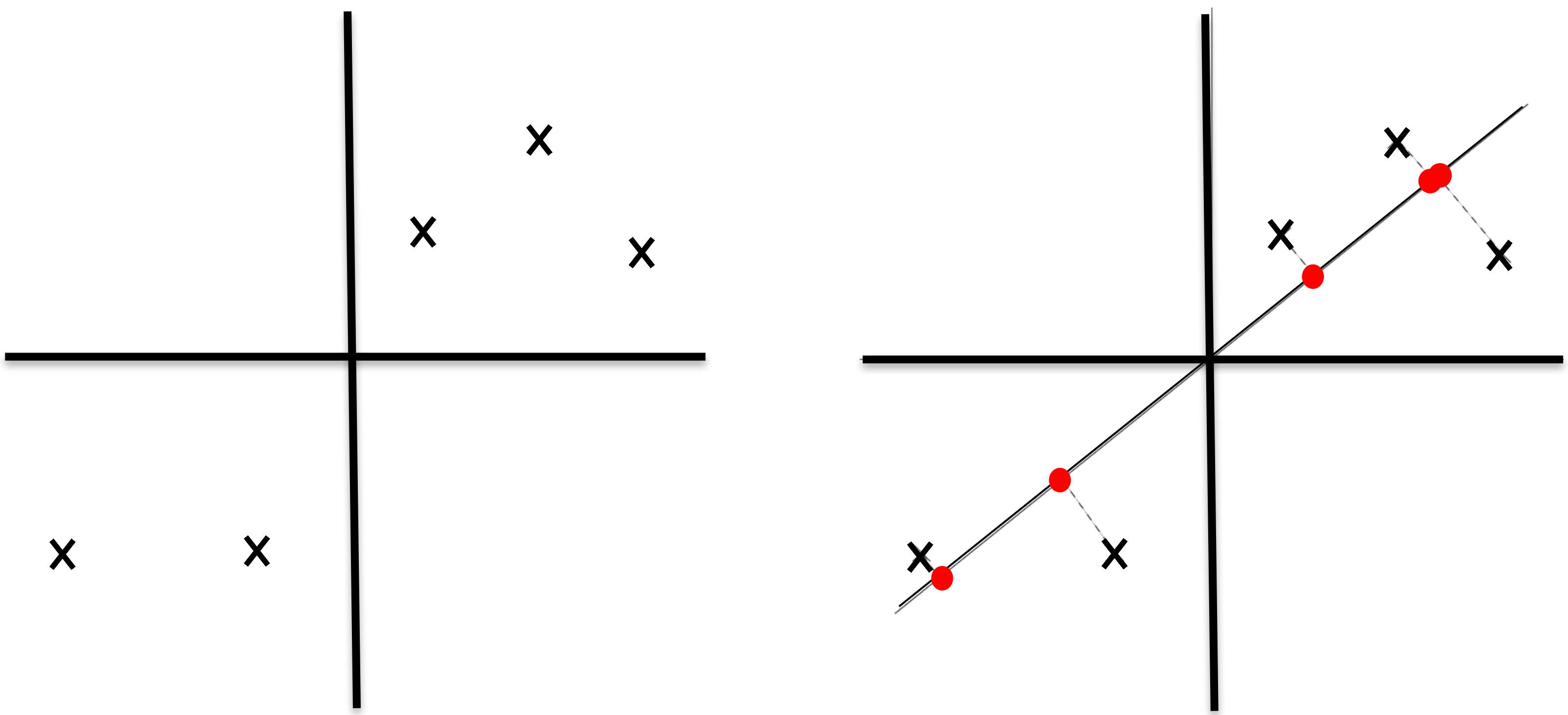
- **To convert a set of original variables into a set of values of linearly uncorrelated variables**
- **Dimension reduction by summarizing the data with smaller number of features to minimize information loss**
- **Idea: summarizing dataset $\boldsymbol{X} = \{\boldsymbol{x^i}; \boldsymbol{i} = \mathbf{1}, \cdots, \boldsymbol{m}\}\ (\boldsymbol{x^i} \in \mathbb{R}^{\boldsymbol{n}})$ to use linear combination**

$$\hat{\boldsymbol{x}} = \boldsymbol{X}_{\boldsymbol{normalized}}\boldsymbol{u}\ (\boldsymbol{u} \in \mathbb{R}^{\boldsymbol{n}}, ||\boldsymbol{u}|| = \mathbf{1})$$

$\hat{x}$: new dataset made by PCA
u: weight of PCA, meaning new feature

Principle Component Analysis(PCA)

- $\hat{\boldsymbol{x}} = \boldsymbol{X}_{\boldsymbol{normalized}}\boldsymbol{u}\ (\boldsymbol{u} \in \mathbb{R}^{\boldsymbol{n}}, \left||\boldsymbol{u}|\right| = \mathbf{1})$



Ng, Andrew. Lecture 10.
https://see.stanford.edu/Course/CS229

Principle Component Analysis(PCA)

- **Dataset** $X = \{x^{(i)}; i = 1, \cdots, m\}$ $(x^{(i)} \in \mathbb{R}^n)$
- **Preprocessing**
  - $\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$
  - $\sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x_j^{(i)} - \mu\right)^2$
  - $z_j^{(i)} = \frac{x_j^{(i)} - \mu}{\sigma_j}$
- **Find** w **maximizing the variance of the projection**
  - $\max \frac{1}{m}\sum_{i=1}^{m}\left({z^{(i)}}^T w\right)^2 = \max \frac{1}{m}\sum_{i=1}^{m} w^T z^{(i)} {z^{(i)}}^T w$
    $= \max w^T\left(\frac{1}{m}\sum_{i=1}^{m} z^{(i)} {z^{(i)}}^T\right) w = \max w^T \Sigma w$ (Σ: covariance matrix)
  - $\hat{x} = Zw$ (Z: normalized X)

Probabilistic PCA and Factor analysis

- **Latent variable** $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{h}; \boldsymbol{0}, \boldsymbol{I})$
- $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b} + \epsilon$

$\boldsymbol{W}$: **Weight**
ϵ: **noise**
ψ: **diagonal covariance matrix**

$$\mathbb{E}(\boldsymbol{x}) = \mathbb{E}(\boldsymbol{W}\boldsymbol{h} + \boldsymbol{b} + \boldsymbol{\epsilon}) = \boldsymbol{b}$$

$$\begin{aligned}\mathbb{E}(\boldsymbol{x}^2) &= \mathbb{E}\left((\boldsymbol{W}\boldsymbol{h} + \boldsymbol{b} + \boldsymbol{\epsilon})(\boldsymbol{W}\boldsymbol{h} + \boldsymbol{b} + \boldsymbol{\epsilon})^{\mathrm{T}}\right) \\ &= \mathbb{E}\big(\boldsymbol{W}\boldsymbol{h}\boldsymbol{h}^T\boldsymbol{W}^T + \boldsymbol{b}\boldsymbol{W}\boldsymbol{h} + \boldsymbol{\epsilon}\boldsymbol{W}\boldsymbol{h} + \boldsymbol{h}^T\boldsymbol{W}^T\boldsymbol{b} + \boldsymbol{b}^2 \\ &\quad + \boldsymbol{b}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}\boldsymbol{W}\boldsymbol{h} + \boldsymbol{\epsilon}\boldsymbol{b} + \boldsymbol{\epsilon}^2\big) \\ &= \boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{b}^2 + \boldsymbol{\psi}\end{aligned}$$

$$\mathrm{Var}(\boldsymbol{x}) = \mathbb{E}(\boldsymbol{x}^2) - \big(\mathbb{E}(\boldsymbol{x})\big)^2 = \boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{\psi}$$

- $\mathbf{x} \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{b}, \boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{\psi})$

Probabilistic PCA and factor analysis

- **To capture the dependencies between the different observed variables x_i**

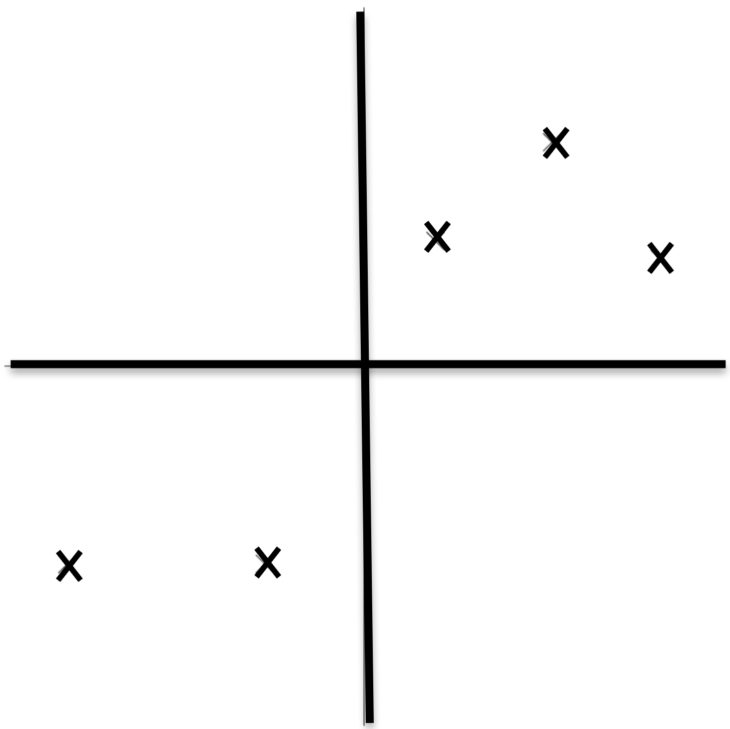
$$\mathbf{x} \sim N\left(\boldsymbol{x}; \boldsymbol{b}, \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}} + \boldsymbol{\psi}\right)$$

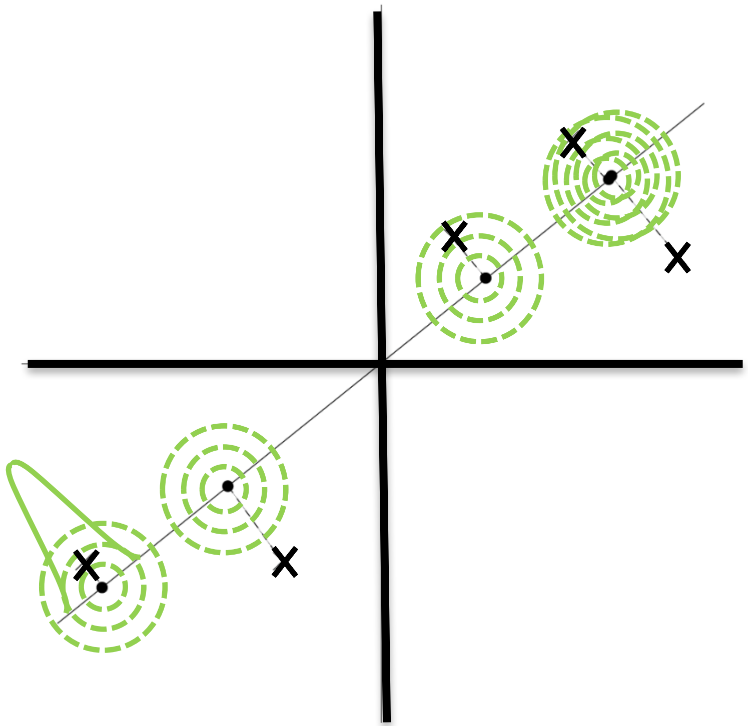
- **Modified PCA**

$$\mathbf{x} \sim N\left(\boldsymbol{x}; \boldsymbol{b}, \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}} + \boldsymbol{\sigma}^{\mathbf{2}}\boldsymbol{I}\right)$$
$$(\boldsymbol{x} = \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b} + \boldsymbol{\sigma}\boldsymbol{z})$$
$$\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

- $\sigma \rightarrow 0$**, probabilistic PCA become PCA**

Probabilistic PCA





Ng, Andrew. Lecture 10.
https://see.stanford.edu/Course/CS229

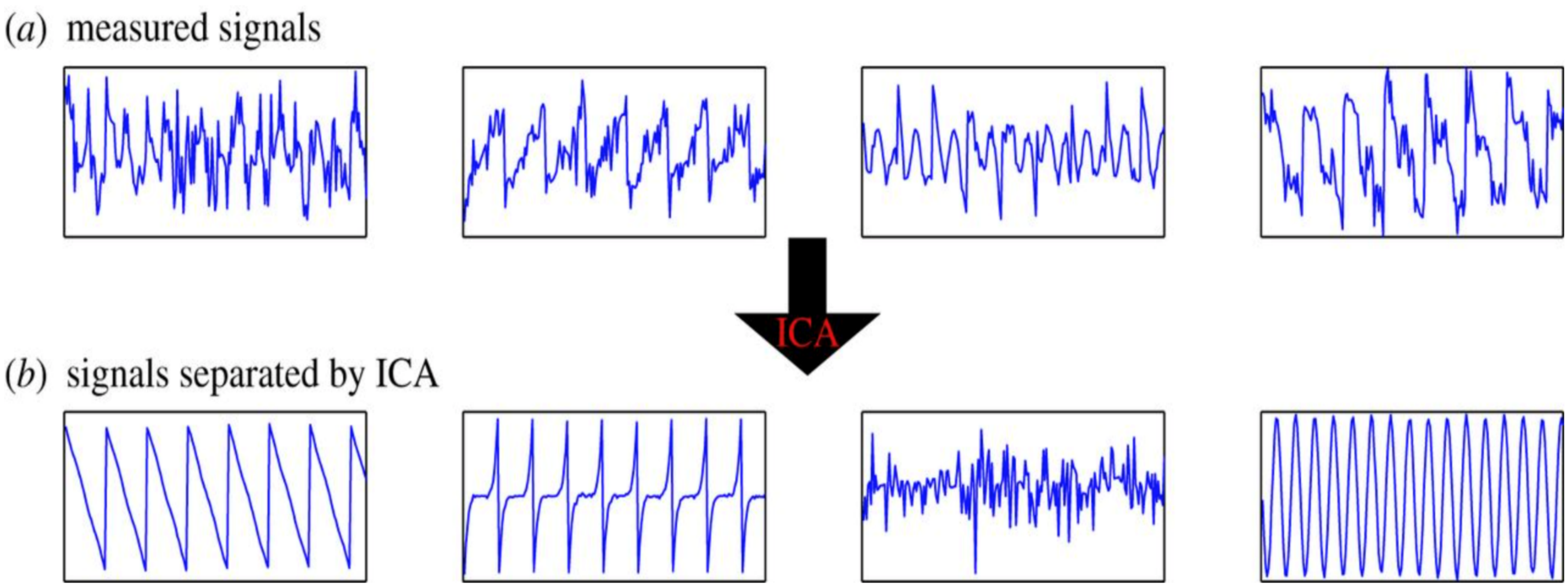
Independent Component Analysis(ICA)

- **To separate observed signals into the source signals which are fully independent**
- **Observed signals** $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^T$

$$x_i = w_{i1}h_1 + w_{i2}h_2 + \cdots + w_{in}h_n \; (i = 1, 2, \cdots, n)$$

- **Matrix notation** $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{h}$
- **Estimate mixing parameter** $\boldsymbol{W}$ **and source signals** $\boldsymbol{h}$
- **Independent subspace analysis**
  - To assign spatial coordinates to each hidden unit
  - To form overlapping groups of spatially neighboring units

Independent Component Analysis(ICA)



Hyvärinen, A. (2013). Independent component analysis: recent advances. *Phil. Trans. R. Soc. A*, *371*(1984), 20110534.

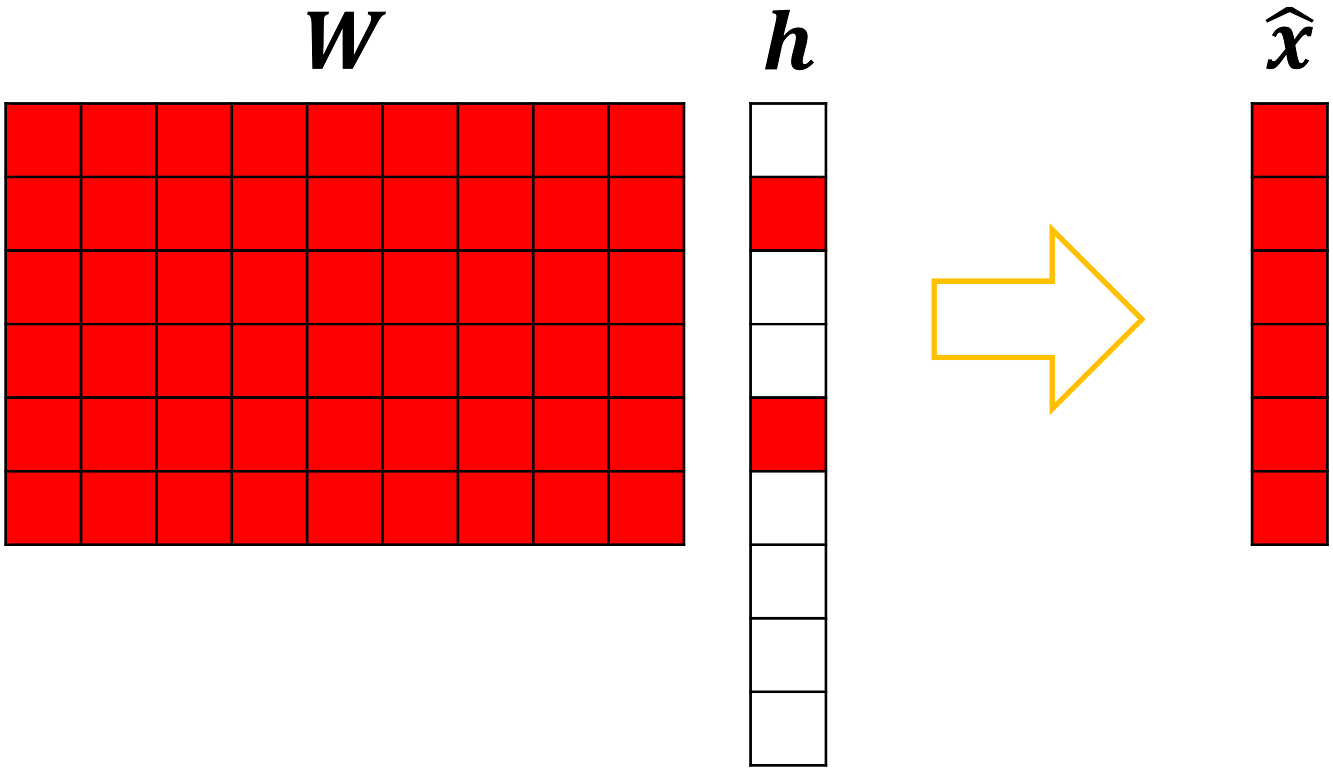
Sparse Coding

- **Unsupervised feature learning and feature extraction**
- **Gaussian noise with hyperparameter β**
- **Better reconstruct the data given the encoding**
- $$\begin{aligned} \boldsymbol{h}^* &= \boldsymbol{arg\,max_h p(h|x)} \\ &= \boldsymbol{arg\,max_h log\,p(h|x)} \\ &= \boldsymbol{arg\,max_h log \frac{p(h)p(x|h)}{p(x)}} \\ &= \boldsymbol{arg\,max_h \left(log p(h) + log p(x|h)\right)} \\ &= \boldsymbol{arg\,max_h \left(-\frac{\lambda}{2}|h| - \frac{\beta^2(x-Wh)^2}{2}\right)} \\ &= \boldsymbol{arg\,min_h} \left(\lambda||h||_1 + \beta||x - Wh||_2^2\right) \end{aligned}$$

Sparse coding

- $\boldsymbol{h} = \boldsymbol{arg\,min_h}\left(\lambda\boxed{||h||_1} + \beta\boxed{||x - Wh||_2^2}\right)$ Reconstruction error

Sparsity penalty



Sparse coding example

https://m.blog.naver.com/PostView.nhn?blogId=laonple&logNo=220914873095&proxyReferer=https%3A%2F%2Fwww.google.co.kr%2F

Sparse coding example

https://m.blog.naver.com/PostView.nhn?blogId=laonple&logNo=220914873095&proxyReferer=https%3A%2F%2Fwww.google.co.kr%2F

Thank you