Linear Factor Models

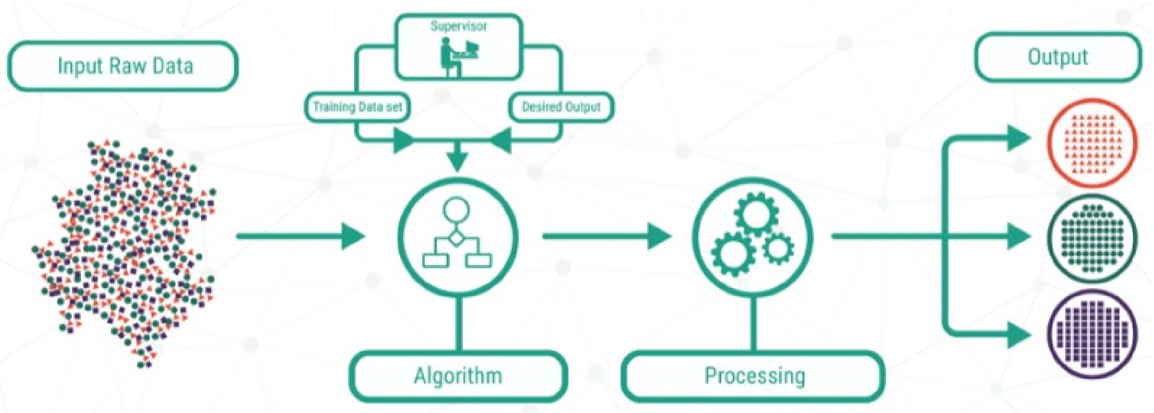
Deep Learning Study: Chap 13

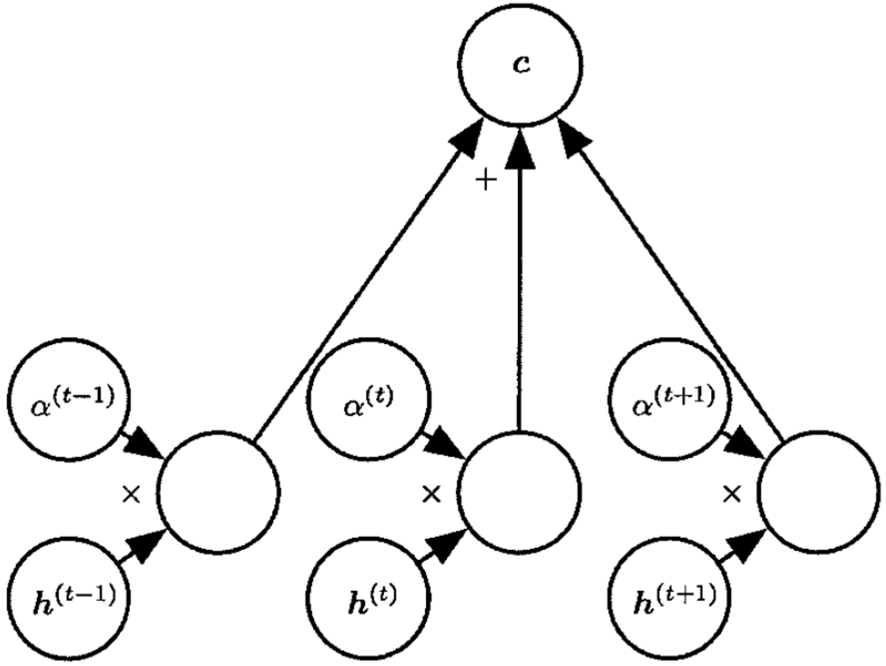
신성호

2023.

Overview of Part III

- Part I,II : 지도 학습의 문제를 푸는 방법들

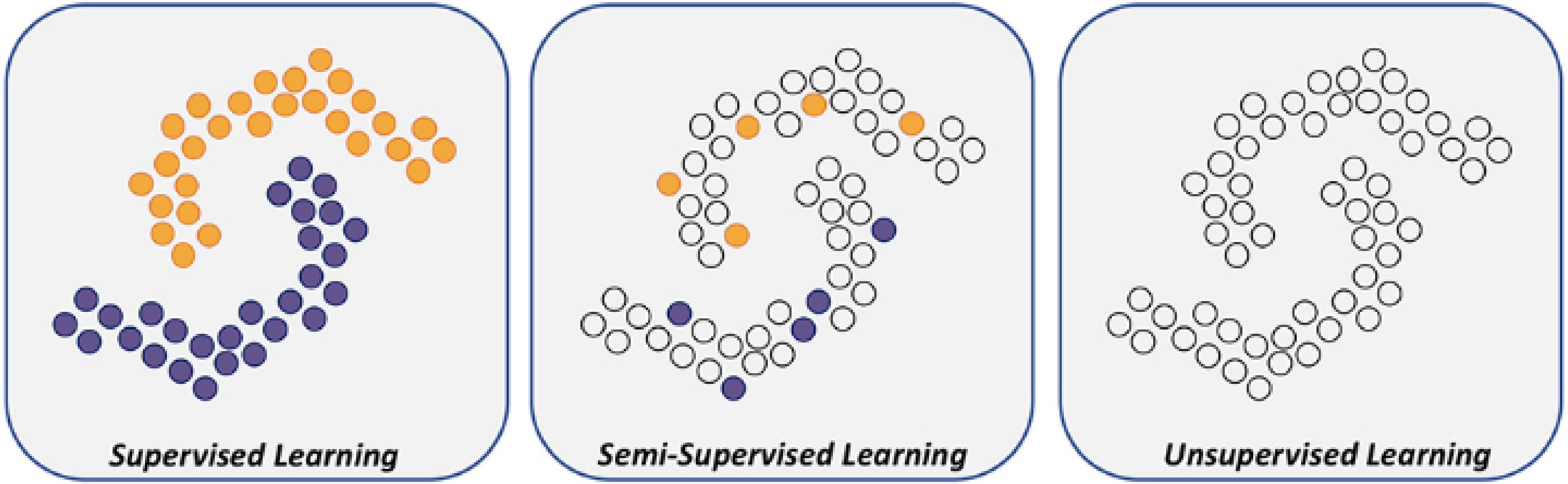




- 벡터 사상
  - 한 벡터로부터 다른 벡터로의 사상에 대한 견본이 충분히 주어졌을 때 학습

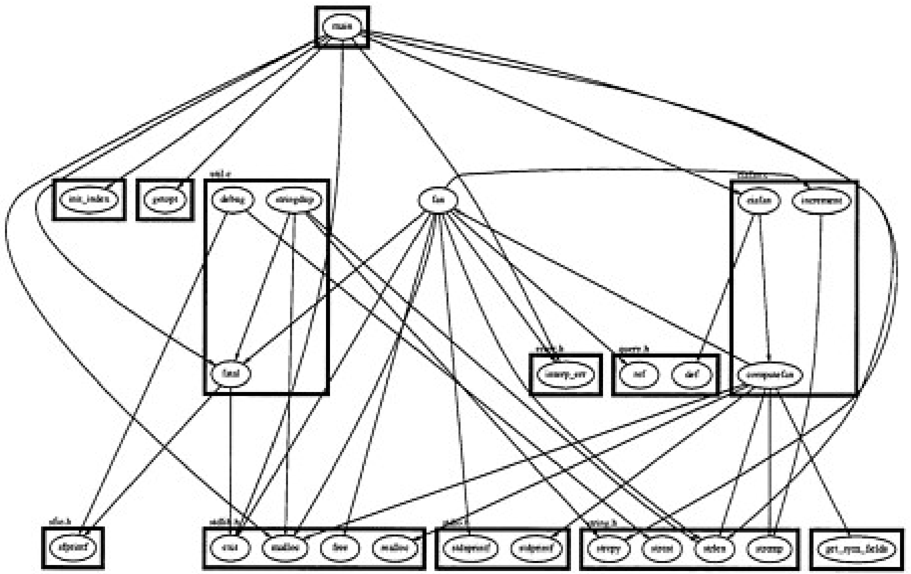
Overview of Part III

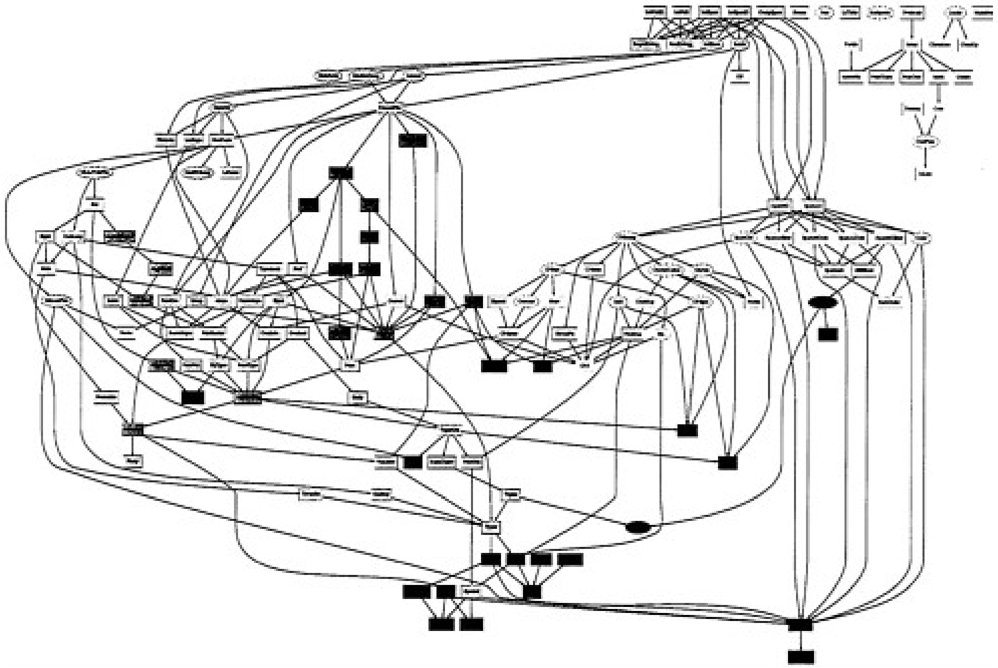
- 지도학습으로 해결할 수 없는 문제
  - 새 견본 생성, 발생 가능성 추정, 결측값 처리, 라벨링이 되지 않은 견본



Overview of Part III

- 통계적 어려움(Statistical challenge)
  - 새 구분하고자 하는 구성의 수는 차원의 수에 지수적으로 증가

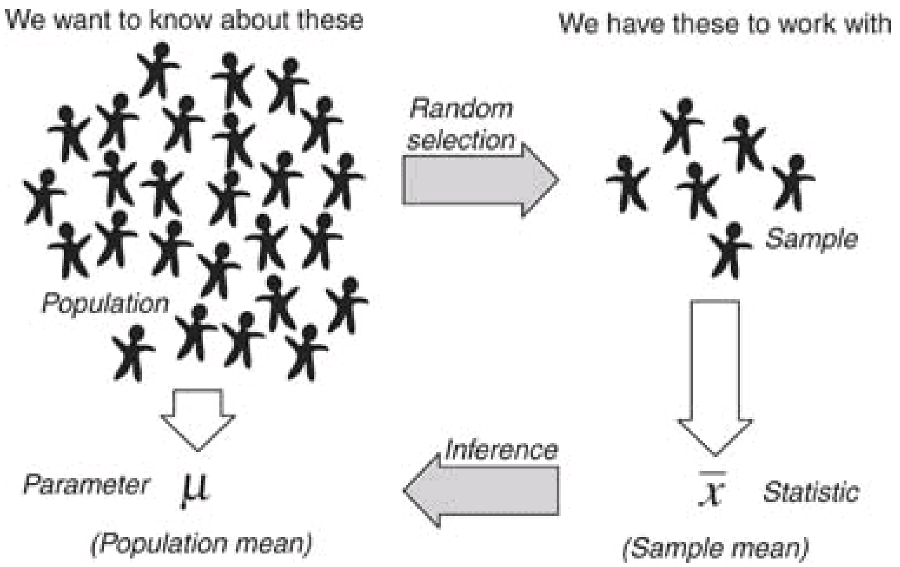




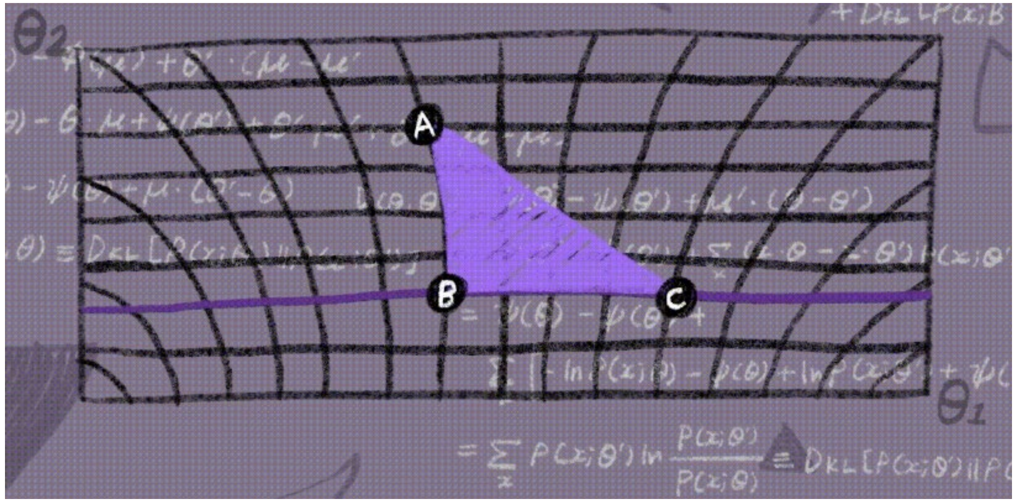
Overview of Part III

- 계산적 어려움(Computational challenge)
  - 고차원 분포에서 학습 또는 학습된 모형을 활용하는 여러 알고리즘에 처리 불가능한 계산 관여

- 처리 불가능한 추론

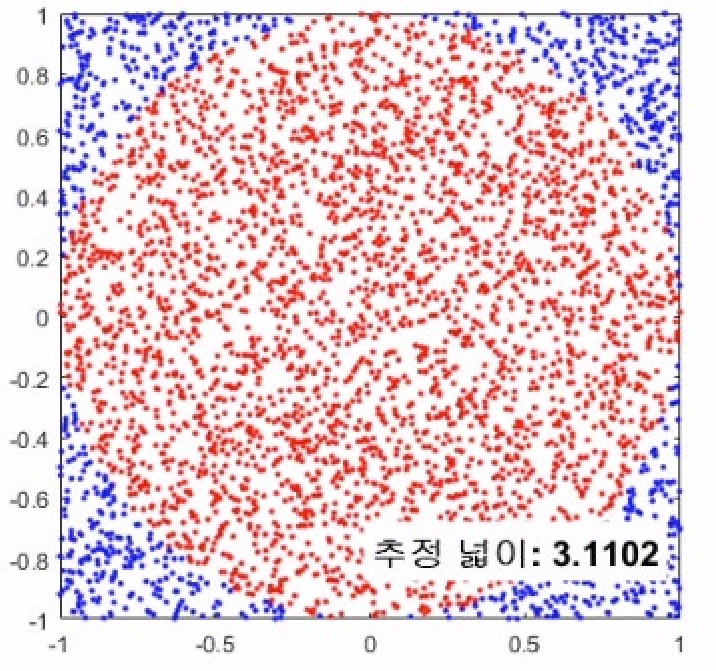


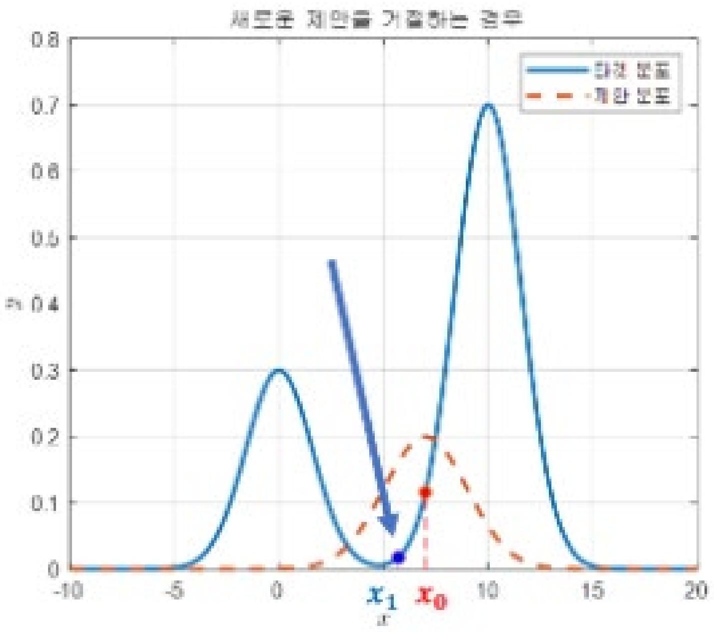
- 처리 불가능한 정규화 상수



Overview of Part III

- 처리 불가능한 정규화 상수 (분배 함수 – Partition function)
  - 분배함수의 계산 자체가 처리 불가능





Overview of Part III

- 처리 불가능 계산 문제를 해결하는 방법?
  - 참값이 아닌 근삿값을 사용
  - 생성 모형화 (제 20장)
- ‘연구자’ 에게 해당하는 독자에게 가장 중요한 파트

Chap XIII – Linear Factor Model

- 입력 확률 모형의 구축은 심층 학습의 최신 연구 주제의 핵심
  - 잠재변수 $\boldsymbol{h}$: $p_{\text{모형}}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{h}} p_{\text{모형}}(\boldsymbol{x} \mid \boldsymbol{h})$ 를 만족
- 선형 인자 모형 (Linear Factor Model)
  - 혼합 모형의 구성요소
  - 더 큰 심층 확률 모형의 구성요소
  - 생성 모형 - generative model (고급의 심층 모형을 더욱 확장)

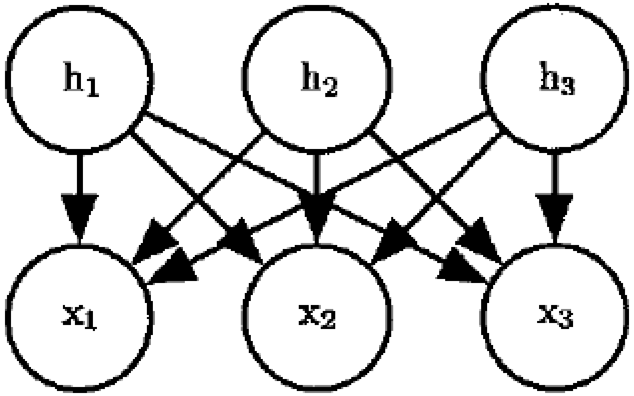
Chap XIII – Linear Factor Model

- 확률적 선형 복호기 (stochastic linear decoder) 함수 활용
  - 잠재변수 $\boldsymbol{h}$ 의 선형 변환에 잡음을 더함
  - 단순한 결합분포를 가진 설명 인자(explanatory factor)들을 발견

$$\mathbf{h} \sim p(\boldsymbol{h}).$$

$$p(\boldsymbol{h}) = \prod_i p(h_i)$$

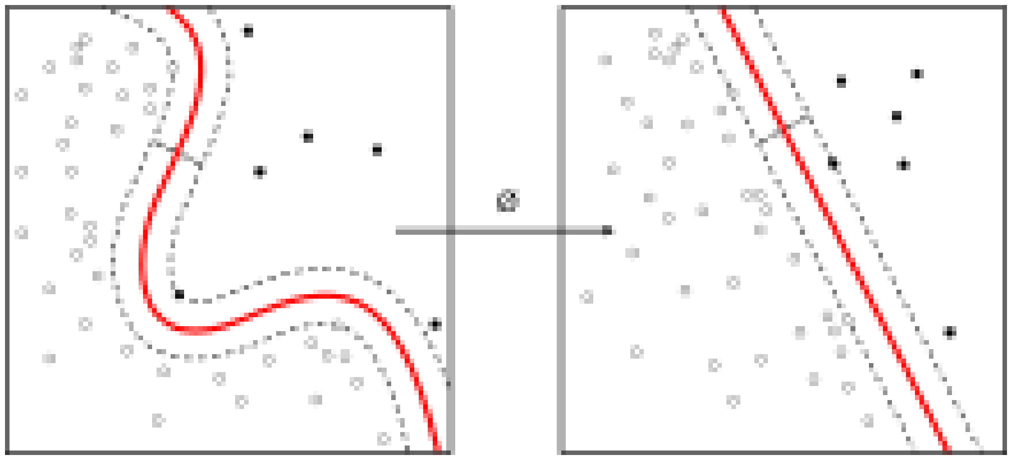
$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b} + \text{잡음}.$$

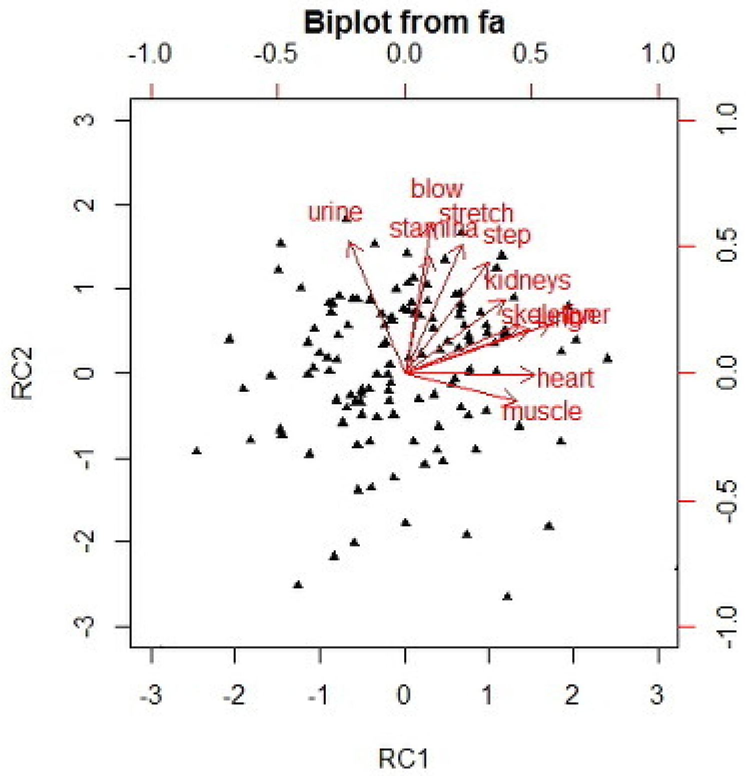


$\mathbf{x} = \boldsymbol{W}\mathbf{h} + \boldsymbol{b} + \text{잡음}$

13.1 확률적 PCA & 인자분석

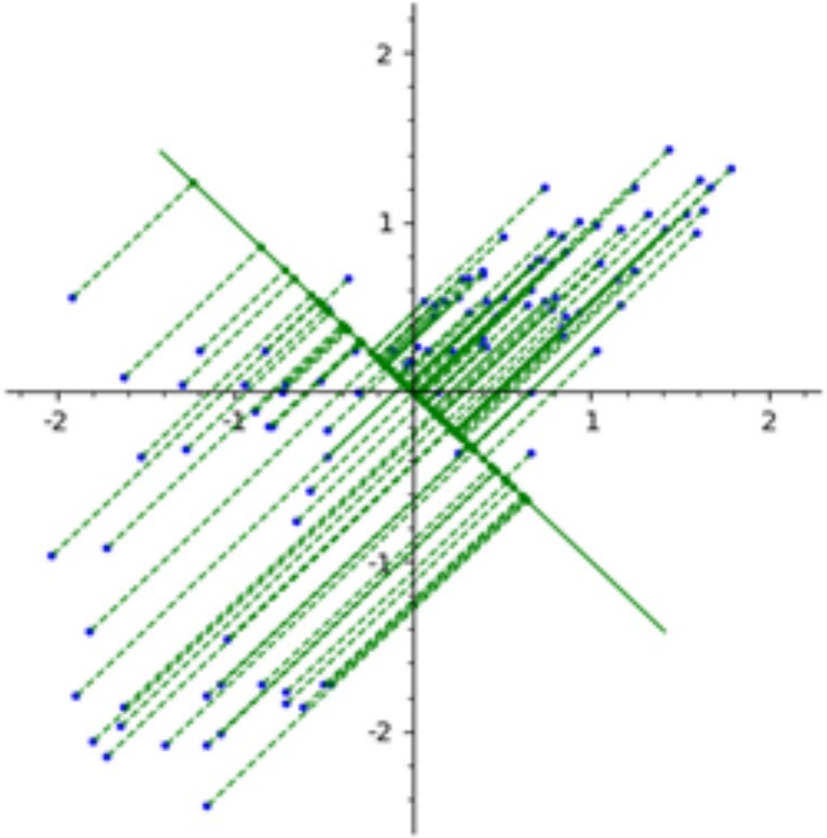
- 인자분석 (Factor Analysis)
  - 서로 관련 있는 변수에서 어떤 잠재 변수 h가 분산을 설명하는가?

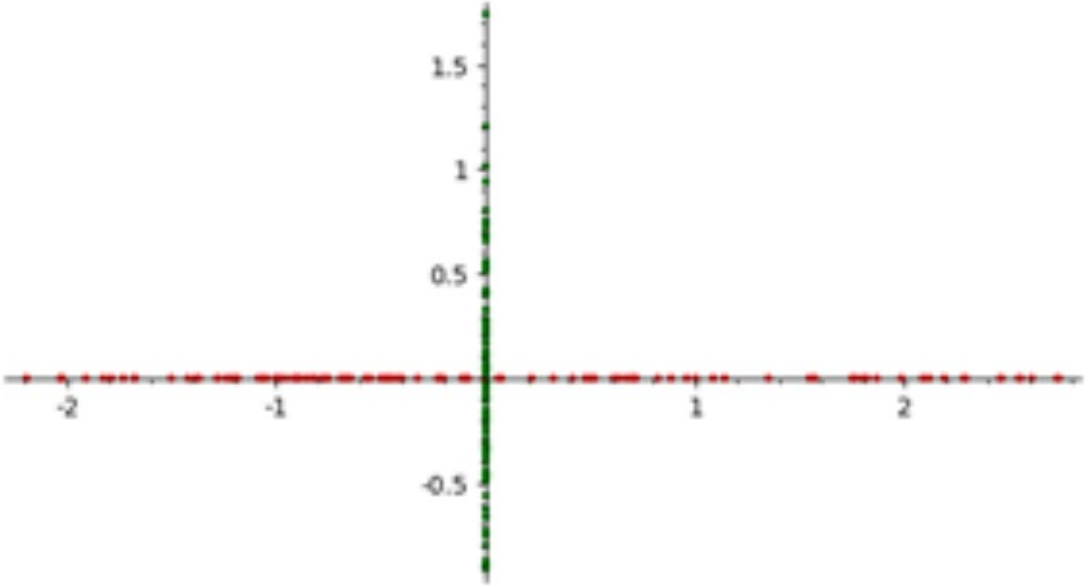




13.1 확률적 PCA & 인자분석

- 주성분분석 (PCA)
  - 원 데이터의 분포를 최대한 보존하면서 고차원 공간의 데이터를 저차원 공간으로 변환





13.1 확률적 PCA & 인자분석

- 인자분석의 특징
  - 잠재변수의 사전 분포는 단위 분산 가우스 분포

$$\mathbf{h} \sim \mathcal{N}(\boldsymbol{h}; \mathbf{0}, \boldsymbol{I})$$

  - 관측변수(observed variable)들이 조건부 독립
  - 잡음을 대각 공분산 가우스 분포에서 얻음

$$\boldsymbol{\psi} = \mathrm{diag}(\boldsymbol{\sigma}^2) \qquad \boldsymbol{\sigma}^2 = [\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2]^\top$$

  - 잠재변수들은 서로 다른 관측변수들 사이의 의존성을 포착

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{b}, \underbrace{\boldsymbol{W}\boldsymbol{W}^\top + \boldsymbol{\psi}}_{\text{공분산}}).$$

13.1 확률적 PCA & 인자분석

- 주성분분석(PCA) : 조건부 분산들이 모두 같아짐

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{b}, \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2 \boldsymbol{I}).$$

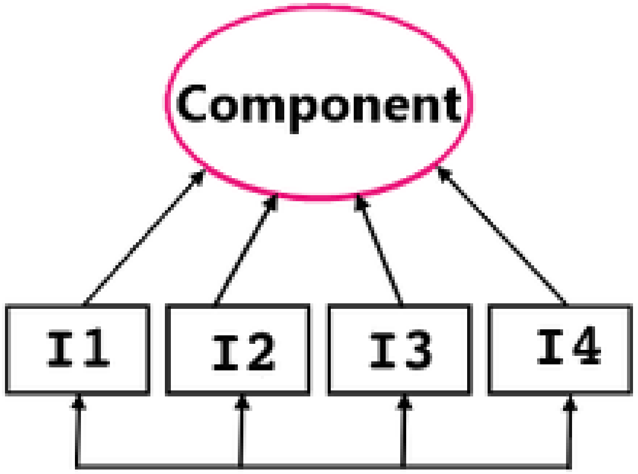
$$\mathbf{x} = \boldsymbol{W}\mathbf{h} + \boldsymbol{b} + \sigma\mathbf{z}. \quad \mathbf{z} \sim \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \boldsymbol{I})$$

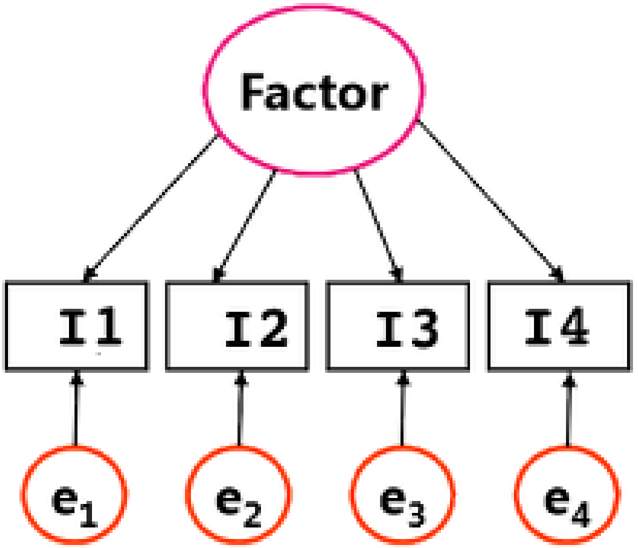
반복적인 EM 알고리즘을 이용해서 매개변수 $\boldsymbol{W}$와 σ^2을 추정할 수 있다.

- PCA가 정의하는 밀도 모형은 W의 열들이 차지하는 d개의 차원들 주변에서 아주 날카로워짐

13.1 확률적 PCA & 인자분석

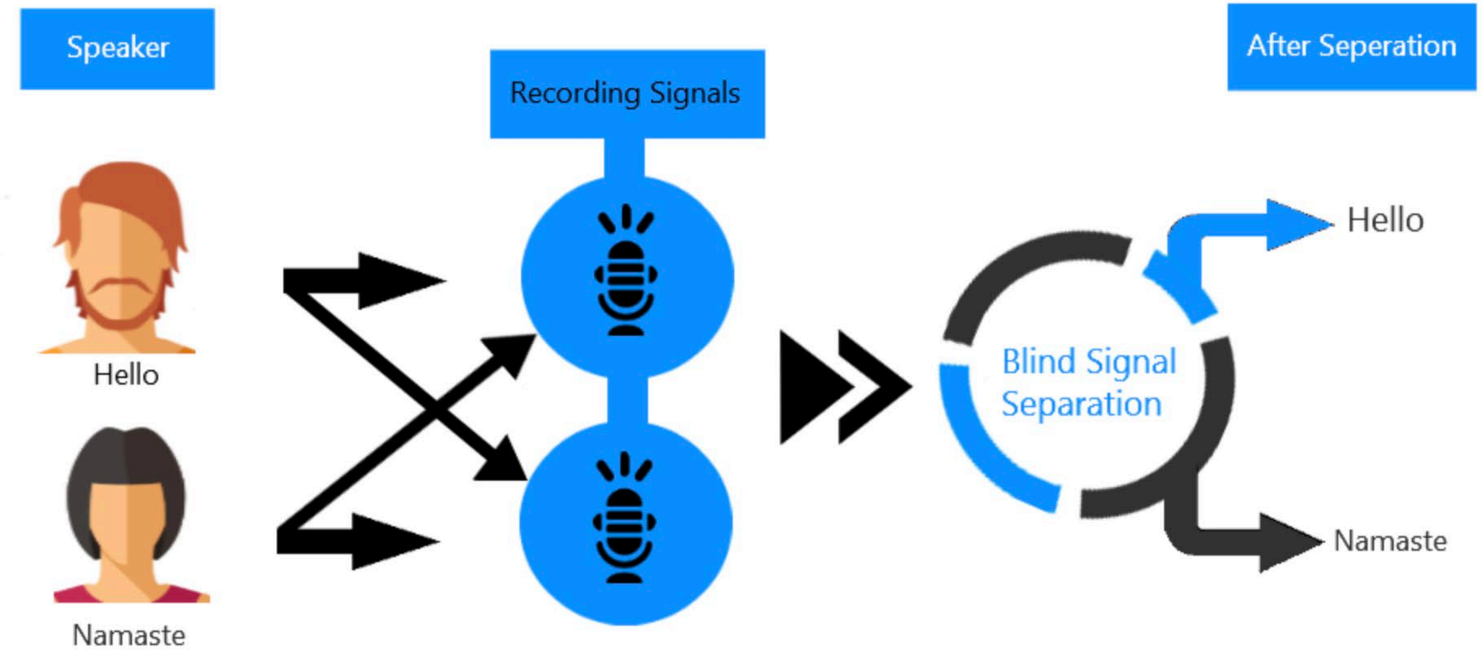
- 확률적 주성분분석(PCA) VS 인자분석(FA)
  - 변수간 중요성? 선형 독립?





13.2 독립성분분석(ICA)

- 관측된 신호를 여러 바탕 신호로 분리하여 관측 자료를 만들어 냄



13.2 독립성분분석

- 신호들은 서로 상관관계가 없으며 완전히 독립
  - 완전 매개변수적 생성 모형 훈련
  - 사전분포를 미리 사용자가 명시적으로 지정
  - 모형은 결정론적으로 $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{h}$ 를 생성

13.2 독립성분분석

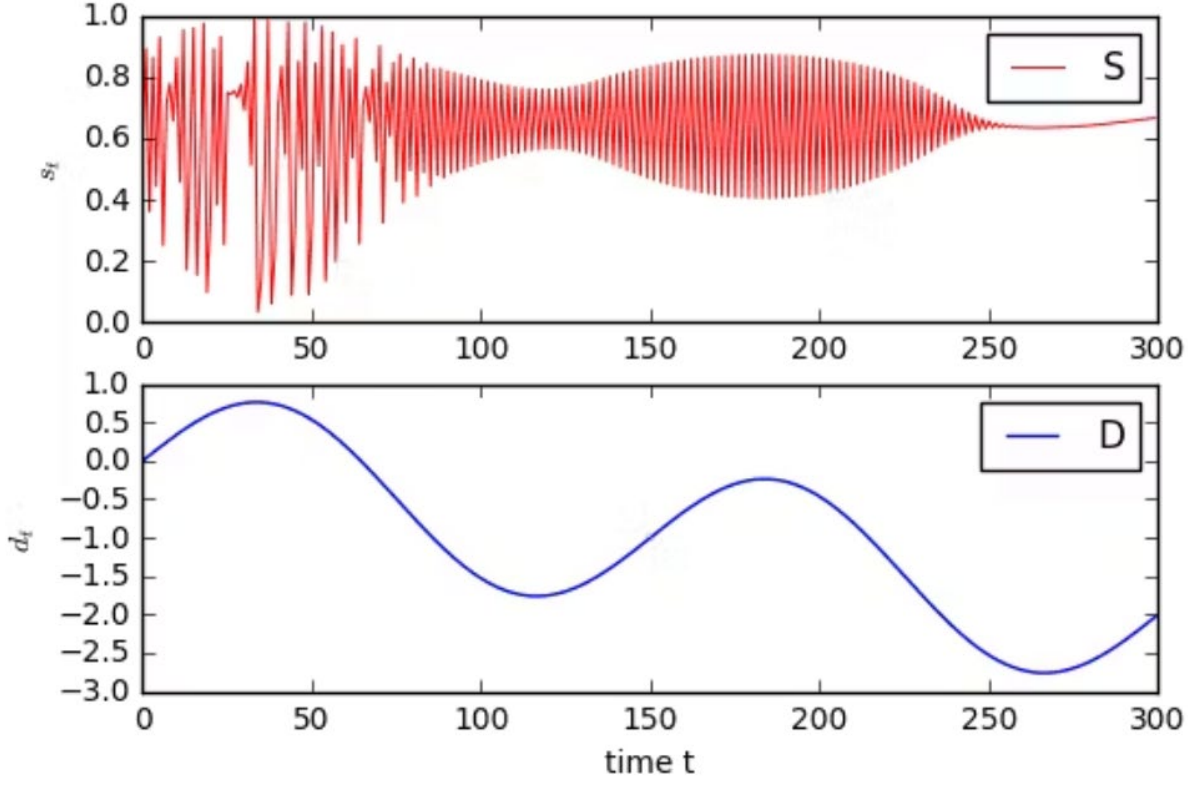
- 다양한 변형 존재
  - 최대 가능도(우도)를 사용하는 방법
  - 결정론적 복호기를 사용하는 대신 잡음을 추가하여 생성
  - 최대 가능도 방법 대신 $\boldsymbol{h} = \boldsymbol{W}^{-1}\boldsymbol{x}$ 성분들이 서로 독립이도록 하는 방법
  - W를 직교행렬로 제한
- p(h)가 반드시 비 가우스 분포
  - 표본 첨도가 높아짐 - 비 가우스 분포

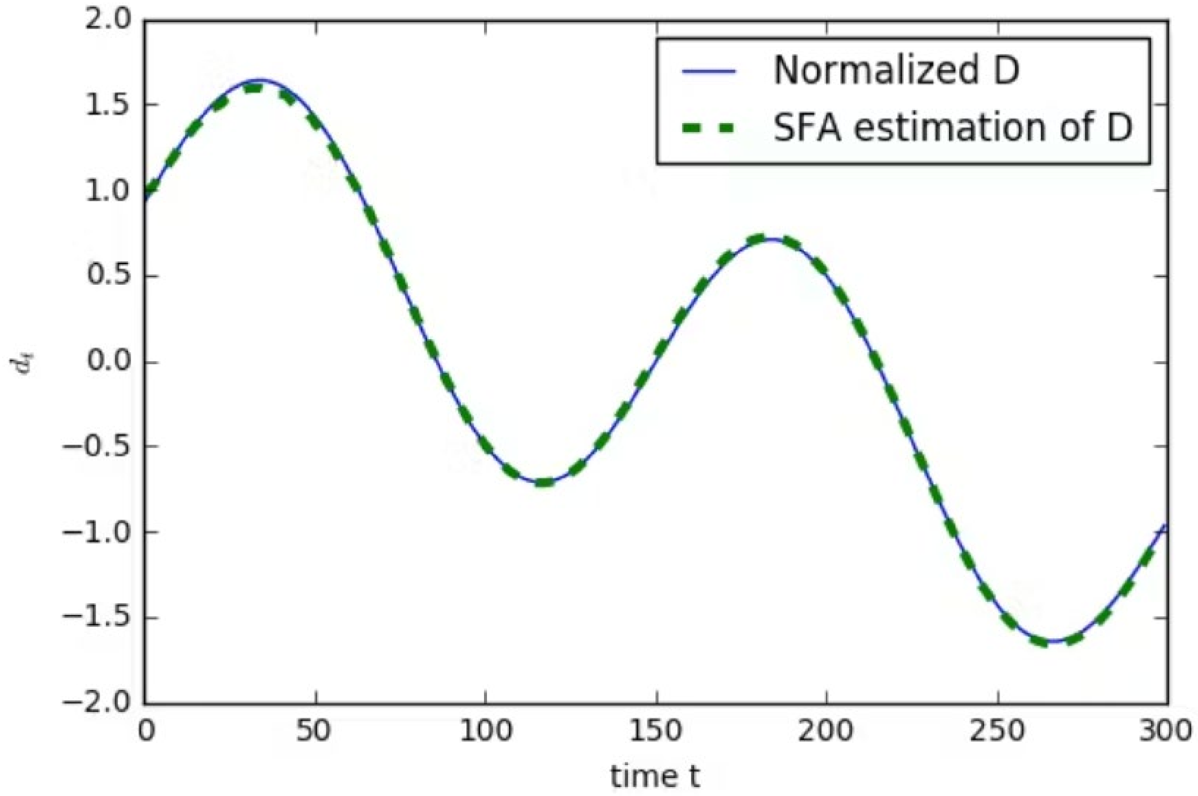
13.2 독립성분분석

- 다양한 변형 존재
  - 최대 가능도(우도)를 사용하는 방법
  - 결정론적 복호기를 사용하는 대신 잡음을 추가하여 생성
  - 최대 가능도 방법 대신 $\boldsymbol{h} = \boldsymbol{W}^{-1}\boldsymbol{x}$ 성분들이 서로 독립이도록 하는 방법
  - W를 직교행렬로 제한
- p(h)가 반드시 비 가우스 분포
  - 표본 첨도가 높아짐 - 비 가우스 분포

13.3 느린특징분석(SFA)

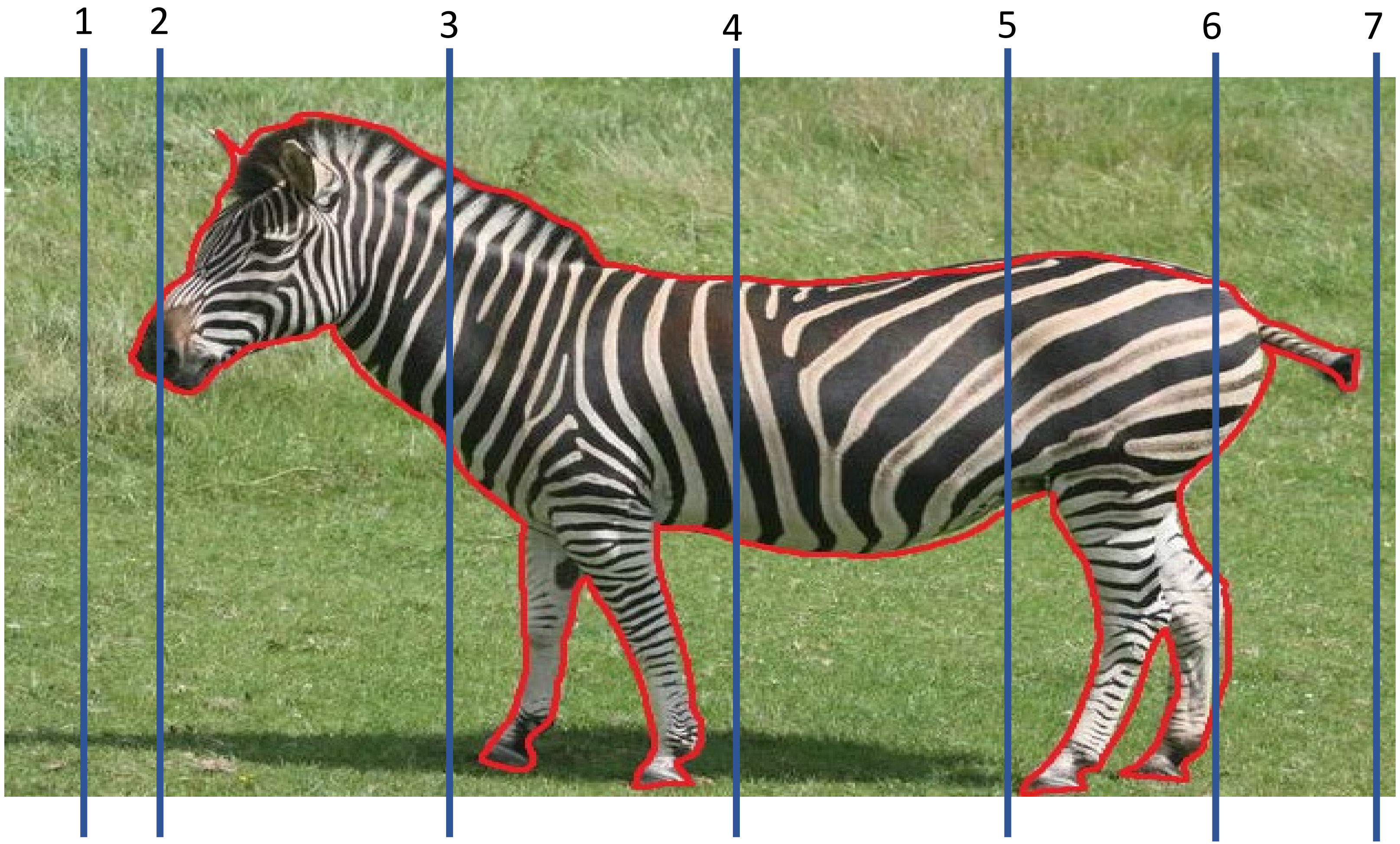
- 느림 원리에 기초
  - 시간 신호에 있는 정보를 이용해서 불변의 특징을 학습





13.3 느린특징분석(SFA)

- 장면의 중요한 특징들은 그 장면의 서술을 구성하는 측도보다 느리게 변함



13.3 느린특징분석(SFA)

- 비용 함수에 다음 항 추가: $\lambda \sum_t L(f(\boldsymbol{x}^{(t+1)}), f(\boldsymbol{x}^{(t)})).$
  - λ 는 느림 정칙화 항의 강도를 결정하는 초매개변수
  - t는 시간순 견본 순차열의 특정 견본을 지칭하는 시간 색인
  - f는 정칙화할 특징 추출기
  - L은 $f(\boldsymbol{x}^{(t)})$ 와 $f(\boldsymbol{x}^{(t+1)})$ 의 거리를 측정하는 손실 함수

13.3 느린특징분석(SFA)

- 느린 특징 분석은 느림 원리의 여러 응용 중 특히 효율적임

- SFA 알고리즘은 선형 변환으로 정의됨

$$\min_{\boldsymbol{\theta}} \mathbb{E}_t \left(f(\boldsymbol{x}^{(t+1)})_i - f(\boldsymbol{x}^{(t)})_i\right)^2.$$

$$\mathbb{E}_t f(\boldsymbol{x}^{(t)})_i = 0$$

$$\forall i < j,\ \mathbb{E}_t [f(\boldsymbol{x}^{(t)})_i f(\boldsymbol{x}^{(t)})_j] = 0.$$

$$\mathbb{E}_t [f(\boldsymbol{x}^{(t)})_i^2] = 1$$

13.4 희소 부호화

- 비지도 특징 학습과 특징 추출 매커니즘
  - 하나의 선형 복호기에 x의 재구축을 위한 잡음 더함 - 등방 정밀도

$$p(\boldsymbol{x} \mid \boldsymbol{h}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}, \frac{1}{\beta}\boldsymbol{I})$$

- p(h)는 봉우리가 날카로운 분포를 선택
  - 인수분해된 라플라스 분포, 코시 분포, 인수분해된 스튜던트 분포

13.4 희소 부호화

- 희소 부호화를 최대 가능도로 훈련하는 것은 처리 불가능 문제에 해당
  - 자료의 부호화와 복호기의 훈련을 번갈아 수행

$$p(\boldsymbol{x}|\boldsymbol{h}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}, \frac{1}{\beta}\boldsymbol{I})$$

- 가중치 행렬의 곱셈으로 이루어진 매개변수적 부호기 함수 이용하지 않음
  - 최적화 문제를 푸는 방법

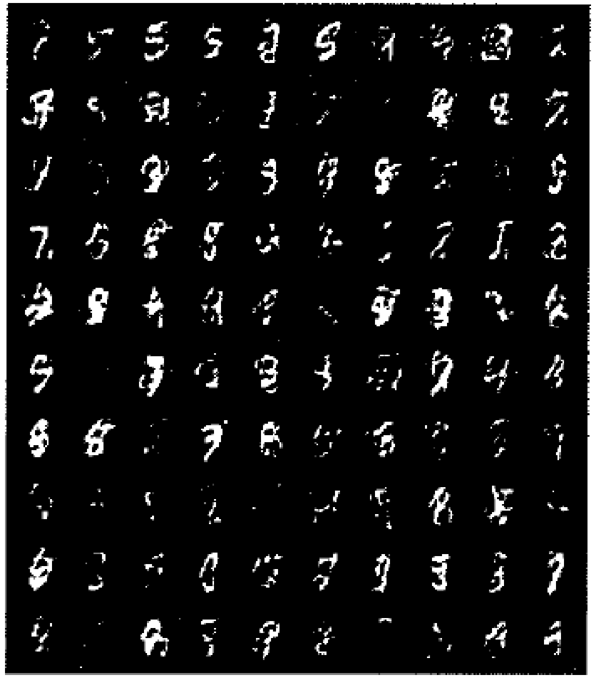
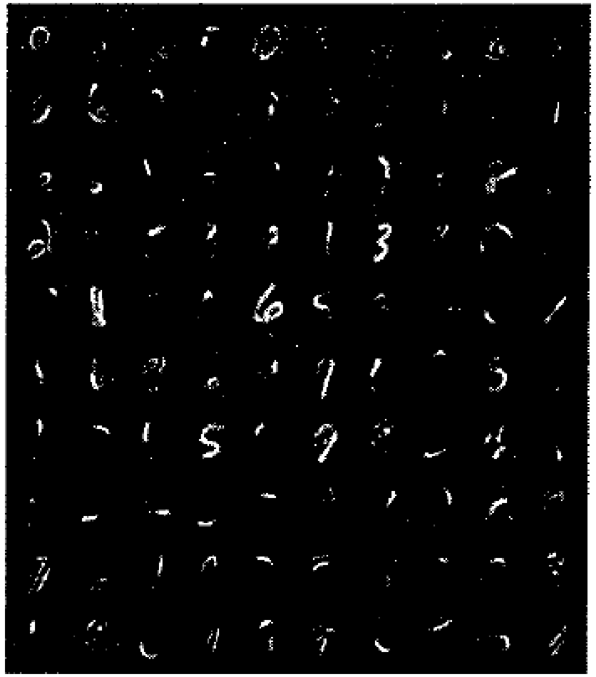
$$\arg\max_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{x})$$

$$= \arg\max_{\boldsymbol{h}} \log p(\boldsymbol{h}|\boldsymbol{x})$$

$$= \arg\min_{\boldsymbol{h}} \lambda\|\boldsymbol{h}\|_1 + \beta\|\boldsymbol{x} - \boldsymbol{W}\boldsymbol{h}\|_2^2.$$

13.4 희소 부호화

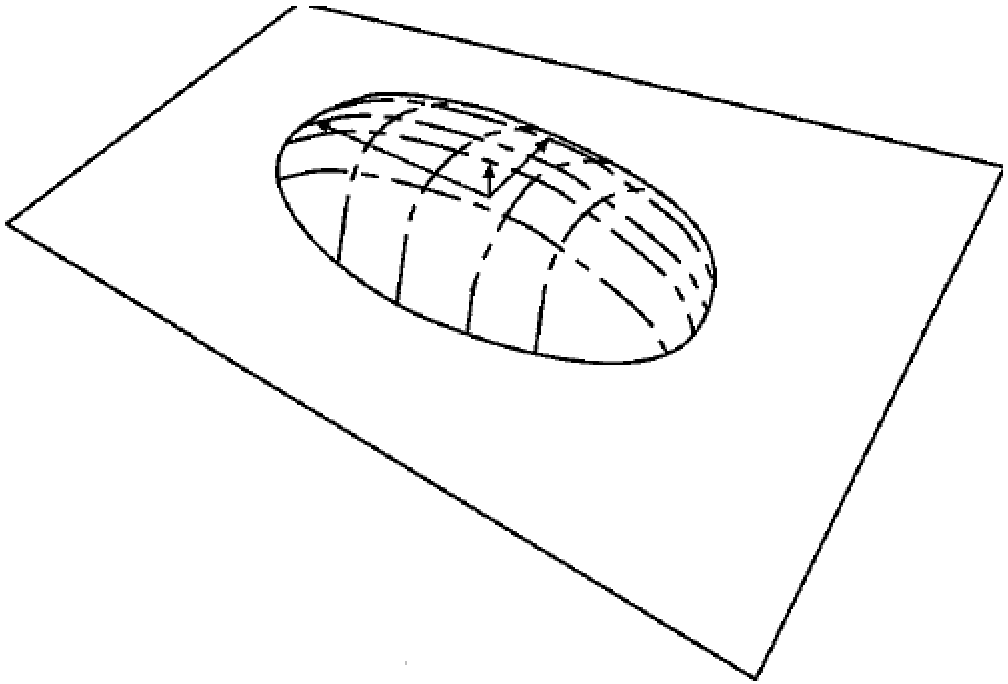
- 재구축 오차와 부호기에 대한 일반화 오차가 줄거나 없어짐
  - 자료의 부호화와 복호기의 훈련을 번갈아 수행

- 종종 질 나쁜 표본들을 산출 & 역전파 수행하기 어려움

13.5 PCA의 다양체 해석

- 다양체 (manifold)
  - 하나의 선형 다양체에 맞게 방향을 정렬



13.5 PCA의 다양체 해석

$$\boldsymbol{h} = f(\boldsymbol{x}) = \boldsymbol{W}^\top(\boldsymbol{x} - \boldsymbol{\mu}). \qquad \hat{\boldsymbol{x}} = g(\boldsymbol{h}) = \boldsymbol{b} + \boldsymbol{V}\boldsymbol{h}.$$

$$\mathbb{E}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2] \qquad \boldsymbol{V} = \boldsymbol{W},\ \boldsymbol{\mu} = \boldsymbol{b} = \mathbb{E}[\boldsymbol{x}]$$

$$\boldsymbol{C} = \mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top]$$

$\boldsymbol{x} \in \mathbb{R}^D$이고 $d < D$에 대해 $\boldsymbol{h} \in \mathbb{R}^d$일 때: $\min \mathbb{E}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2] = \sum_{i=d+1}^{D} \lambda_i$

- 직교 W 하에서 h의 성분들의 분산을 최대화