기계학습

10강 PCA(2), 순차데이터(1)

장필훈 교수

학습목차

1 PCA(2)

2 순차데이터(1)

01
PCA(2)

1-1 PCA

- 최소오류공식화 (계속)

$$\mathbf{x}_n = \sum_{i=1}^{D} \alpha_{ni}\mathbf{u}_i = \sum_{i=1}^{D} (\mathbf{x}_n^{\mathrm{T}}\mathbf{u}_i)\mathbf{u}_i$$

$$\hat{\mathbf{x}}_n = \sum_{i=1}^{M} z_{ni}\mathbf{u}_i + \sum_{M+1}^{D} b_i\mathbf{u}_i$$

$\mathbf{u}_i, z_{ni}, b_i$ 를 마음대로 조절할 수 있다.

1-1 PCA

- 제곱오류를 사용하면, 다음을 최소화하는 것이 목표.

$$J = \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{x}_n - \hat{\mathbf{x}}_n\|^2$$

- 각각의 변수($\mathbf{u}_i, z_{ni}, b_i$)에 관해 미분하고 0으로 두고 구함 (정규직교조건도 같이 이용), z_{ni}, b_i 를 구해서 대입하면,

$$J = \frac{1}{N}\sum_{n=1}^{N}\sum_{n=1}^{N}\|\mathbf{x}_n^T\mathbf{u}_i - \bar{\mathbf{x}}_n^T\mathbf{u}_i\|^2 = \sum_{M+1}^{D}\mathbf{u}_i^T\mathbf{S}\mathbf{u}_i$$

1-1 PCA

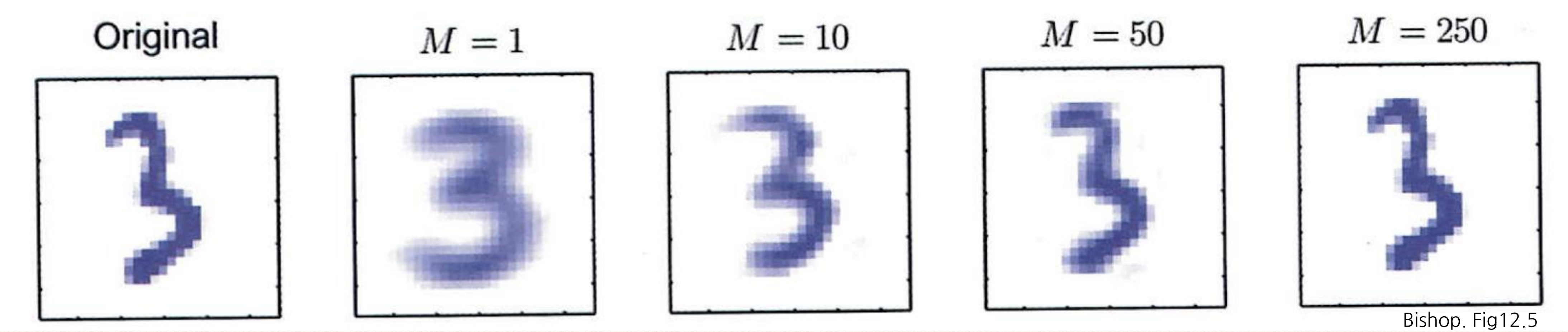
- 정규직교조건($\mathbf{u}_i^T\mathbf{u}_i = 1$)과 함께 라그랑주 승수법 사용.
- 그러면 결국 임의의 $M < D$값에 대한 J의 최소화 일반해:

$$\mathbf{S}\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

- 최대분산의 경우와 같은 해를 얻는다.

1-2 PCA의 적용

- 적용례(손글씨)



Bishop. Fig12.5

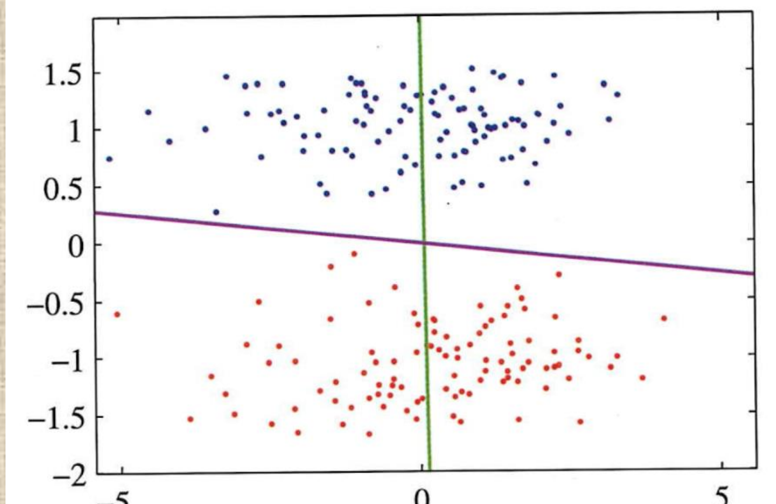
- 재구성에 사용된 차원 M

1-2 PCA의 적용

```python
from sklearn.decomposition import PCA
import numpy as np

train = np.random.randn(100, 50)

pca = PCA(n_components=30)
pca.fit(train) # pca.components_.shape = (30,50)

trained_pca = pca.transform(train) # = (X_train-pca.mean_).dot(pca.components_.T)
trained_pca.shape # (100,30)

reconstructed = pca.inverse_transform(trained_pca)
# = trained_pca.dot(pca.components_)+pca.mean_
```

1-2 PCA의 적용

- 피셔 선형판별과 비교
  - 둘 다 선형변환
  - PCA는 비지도, 피셔는 지도(클래스라벨정보가 있음)



Bishop, Fig12.7

녹색이 피셔선형변환 결과

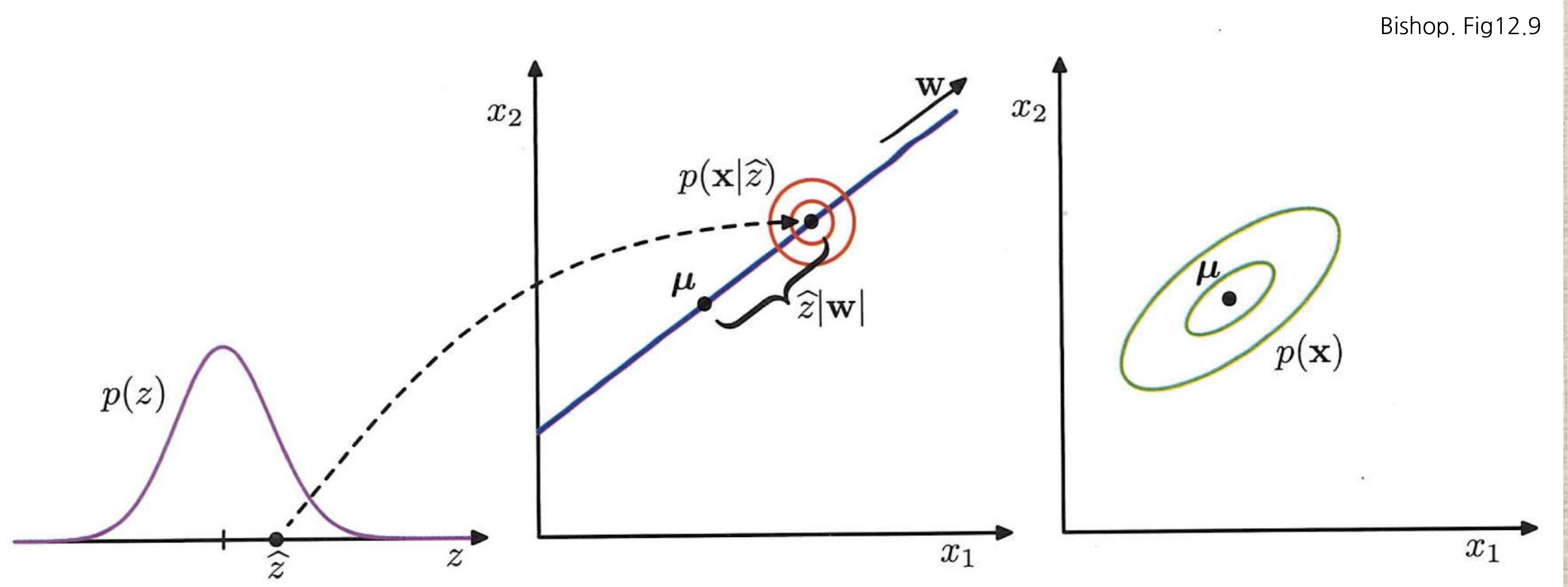
자주색이 최대분산방향

1-3 PPCA(probabilistic PCA)

- 확률적 잠재변수모델의 최대가능도 해로 PCA를 표현
- EM알고리즘 응용이 가능하다
  - 앞에서부터 몇개의 고유벡터만 계산하면 될 경우 유리 (공분산계산 생략)
- missing value를 다룰 수 있다.(확률모델 공통)
- 선형 가우시안 방법론의 한 예임.(다음페이지그림참조)

1-3 PPCA(probabilistic PCA)

- 1차원 잠재변수 공간에서 2차원 데이터공간 가정.



$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \qquad p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})dz$$

1-3 PPCA(probabilistic PCA)

- $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$
- 선형가우시안이기 때문에 marginal dist.도 가우시안

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C})$$

$$\mathrm{Cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T] = \mathbb{E}[(\mathbf{Wz}+\boldsymbol{\epsilon})(\mathbf{Wz}+\boldsymbol{\epsilon})^T]$$

$$= \mathbb{E}[\mathbf{Wzz}^T\mathbf{W}^T + 2\mathbf{Wz}\boldsymbol{\epsilon}^T + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]$$

$$= \mathbf{W}\mathbb{E}[\mathbf{zz}^T]\mathbf{W}^T + 2\mathbf{W}\mathbb{E}[\mathbf{z}\boldsymbol{\epsilon}^T] + \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]$$

$$= \mathbf{WW}^T + \sigma^2\mathbf{I} = \mathbf{C}$$

1-3 PPCA(probabilistic PCA)

- $\mathbf{W}\mathbb{E}[\mathbf{z}\boldsymbol{\epsilon}^T] = 0$

  $\mathrm{cov}[\mathbf{z}, \boldsymbol{\epsilon}] = \mathbb{E}[\mathbf{z}\boldsymbol{\epsilon}^{\mathrm{T}}] - \mathbb{E}[\mathbf{z}]\mathbb{E}[\boldsymbol{\epsilon}]^T = 0$ ($\because$ independent)
- $\mathrm{cov}[\boldsymbol{\epsilon}] = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}] - \mathbb{E}[\boldsymbol{\epsilon}]\mathbb{E}[\boldsymbol{\epsilon}]^T = \sigma^2\mathbf{I}$ (가정)
- log-likelihood

$$\ln p(\mathbf{X}|\mu, \mathbf{W}, \sigma^2) = \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) = \sum_{n=1}^{N} \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}, \mathbf{C})$$

1-3 PPCA(probabilistic PCA)

- log-likelihood(cont.)

$$= \frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{C}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

- 미분해서 0 되는 점 구한다. 예를 들어 평균이라면,

$$\sum_{n=1}^{N}\mathbf{C}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = 0, \qquad \boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

- 닫힌 해를 모두 구할 수 있다.

1-4 EM for PPCA

- 각 데이터 포인트에 대해 잠재변수 $\mathbf{z}_n$이 존재하는 $\mathbf{z}$에 대한 주변확률로서 확률적PCA. 따라서 EM이용 가능
- 닫힌 해를 알지만 EM을 하는 의미가 있을 수 있다
  - 공분산을 모두 계산할 필요가 없을 때 계산효율있음
- 방법
  - 일반적EM과 다르지 않음.

1-4 EM for PPCA

- 방법(계속)
  - goal:

$$W_{ML} = \arg\max_{W} \ln p(x_1, \dots, x_n | \mathbf{W}) = \arg\max_{W} \sum_{i=1}^{n} \ln p(x_i | \mathbf{W})$$

  - E단계: **W**를 가정하고, **z**의 평균과 분산을 계산
  - M단계: **z**의 평균과 분산을 이용해서 **W**를 업데이트
  - $\ln p(x_i|\mathbf{W})$의 증가가 충분히 작을 때까지 반복.

1-5 커널PCA

- 비선형변환 $\phi(x)$: $x_n \to \phi(x_n)$
  - 이 특징공간에서 PCA를 시행.
  - 원 데이터공간에서는 주성분이 비선형으로 나타남.

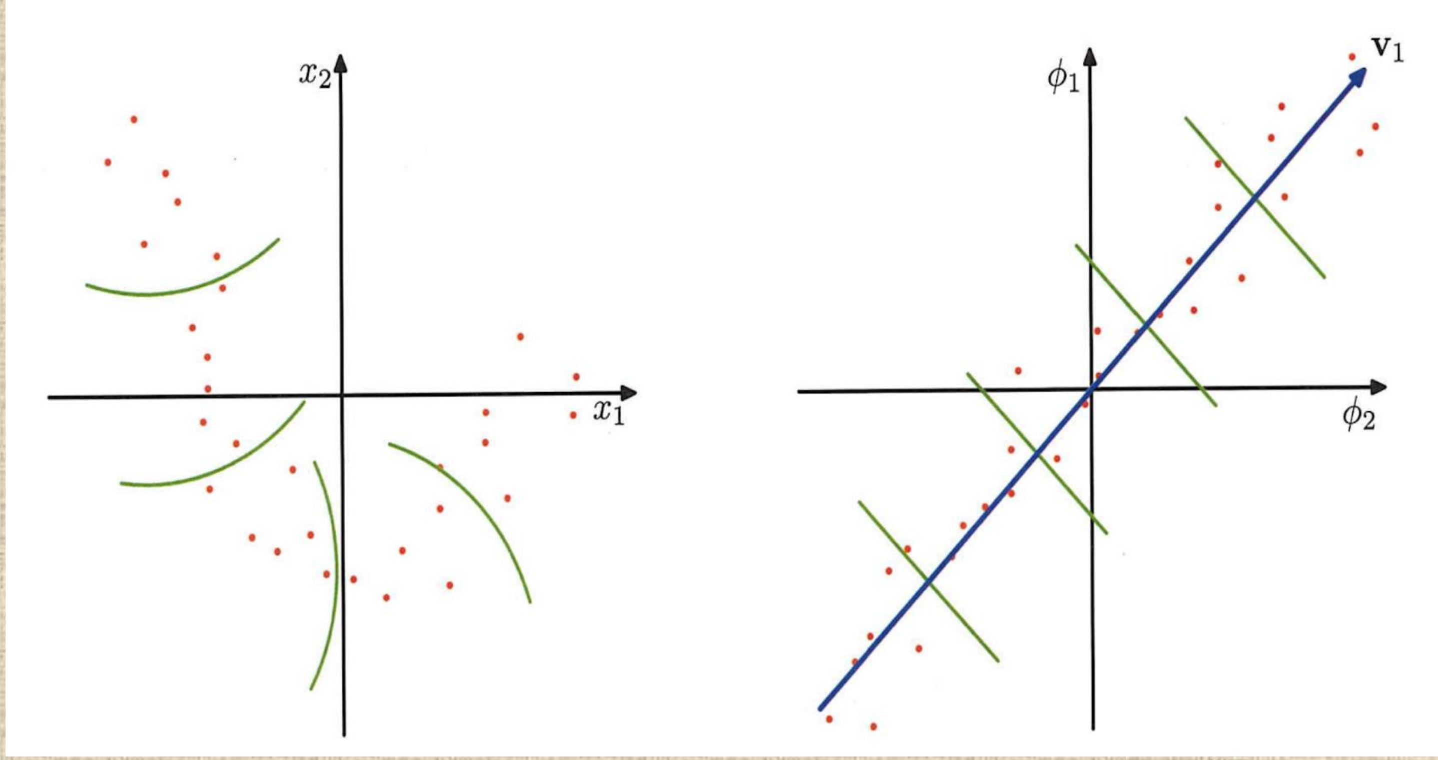
- 특징공간상에서,

공분산행렬 $$\mathrm{C} = \frac{1}{N}\sum_{n=1}^{N}\phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^{\mathrm{T}}$$

고유벡터전개식 $\mathrm{C}\mathbf{v}_i = \lambda_i\mathbf{v}_i$

1-5 커널PCA

- 개념적 도식



Bishop. Fig12.16

1-5 커널PCA

- 특징공간을 직접 다루지 않고 문제를 풀어내는것이 목표
- C의 정의를 고유벡터식에 대입하면,

$$\frac{1}{N}\sum_{n=1}^{N}\phi(\mathbf{x}_n)\{\phi(\mathbf{x}_n)^{\mathrm{T}}\mathbf{v}_i\} = \lambda_i\mathbf{v}_i \qquad \text{(식1)}$$

- 벡터 $\mathbf{v}_i$는 $\phi(\mathbf{x}_n)$의 선형결합이므로 다음과 같이 적을 수 있다.

$$\mathbf{v}_i = \sum_{n=1}^{N} a_{in}\phi(\mathbf{x}_n) \qquad \text{(식2)}$$

1-5 커널PCA

- 식2를 식1에 대입하면,

$$\frac{1}{N}\sum_{n=1}^{N}\phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^T\sum_{m=1}^{N}a_{im}\phi(\mathbf{x}_M) = \lambda_i\sum_{n=1}^{N}a_{in}\phi(\mathbf{x}_n)$$

- 양변에 $\phi(\mathbf{x}_l)^{\mathrm{T}}$를 곱해서 커널함수 $k(\mathbf{x}_{\mathrm{n}},\mathbf{x}_{\mathrm{m}}) = \phi(\mathbf{x}_{\mathrm{n}})^{\mathrm{T}}\phi(\mathbf{x}_{\mathrm{m}})$으로 표현할 수 있다.

$$\frac{1}{N}\sum_{n=1}^{N}k(\mathbf{x}_l,\mathbf{x}_n)\sum_{m=1}^{N}a_{im}k(\mathbf{x}_n,\mathbf{x}_m) = \lambda_i\sum_{n=1}^{N}a_{in}k(\mathbf{x}_l,\mathbf{x}_n)$$

1-5 커널PCA

- 커널로 표현한 식을 간단히 적으면,

$$\mathbf{K}^2\mathbf{a}_i = \lambda_i N \mathbf{K}\mathbf{a}_i$$

- $\mathbf{a}_i$는 n차원 column벡터. 양변에서 $\mathbf{K}$없애면,

$$\mathbf{K}\mathbf{a}_i = \lambda_i N \mathbf{a}_i$$

- 고유벡터 $\mathbf{v}$가 $\mathbf{a}$를 계수로 하는 선형결합이므로, $\mathbf{v}_i^{\mathrm{T}}\mathbf{v}_i = 1$을 이용해서 $\mathbf{a}$를 구할 수 있다.

1-5 커널PCA

- projection도 커널만을 사용해서 가능

$$y_i(\mathbf{x}) = \phi(\mathbf{x})^{\mathrm{T}}\mathbf{v}_{\mathrm{i}} = \sum_{n=1}^{N} a_{in}\phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{x}_{\mathrm{n}}) = \sum_{n=1}^{N} a_{in}k(\mathbf{x},\mathbf{x}_n)$$

- 특징공간차원 M은 원래의 차원 D보다 훨씬 더 클수 있다.
- 지금까지는 평균을 0으로 가정.
  - 그렇지 않은 상황에서는 따로 계산해야 한다.
  - 계산 후 평균만큼 빼지않는다(특징공간에서 계산해야함)

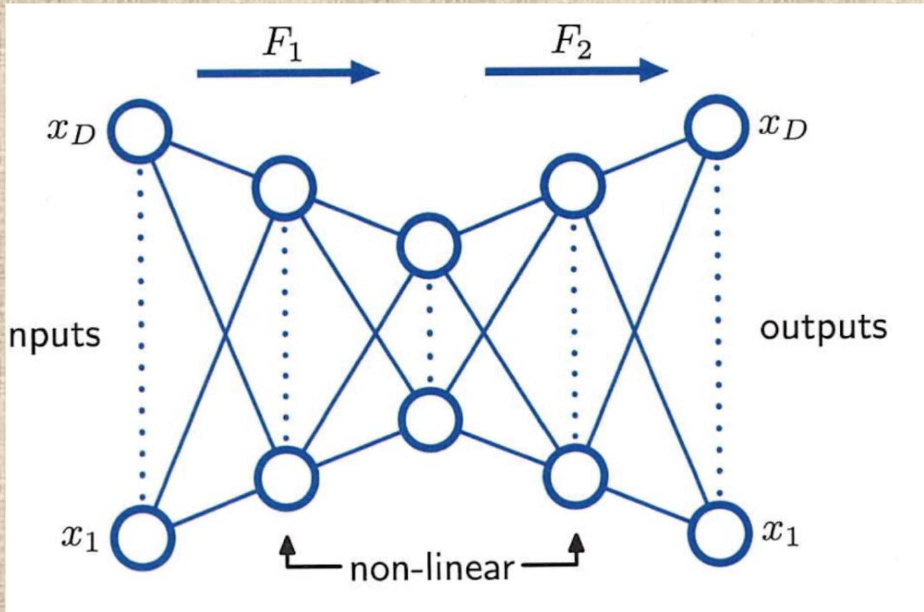
1-5 커널PCA

- $N \times N$행렬의 고유벡터를 찾아야 하는것이 단점.
  - 계산비용이 많이 들어서 근사치를 많이 이용한다.

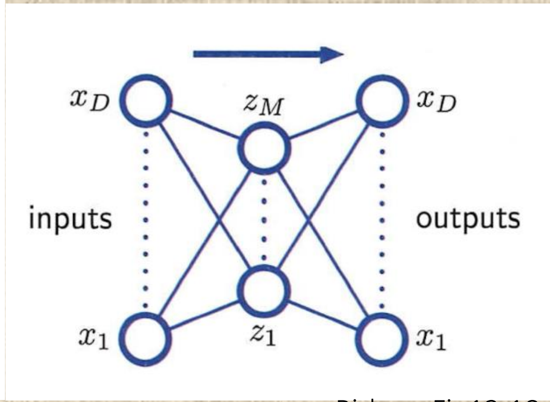
1-6 autoencoder

- 입출력 크기가 같은 다층 퍼셉트론
- 입력과 출력의 불일치정도를 오류함수로 학습

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \|y(\mathbf{x}_n, \mathbf{w}) - \mathbf{x}_n\|^2$$



Bishop. Fig12.19



Bishop. Fig12.18

1-6 autoencoder

- 해석이 거의 불가능할정도로 복잡해도 상관없다
- 표준PCA가 이 네트워크의 특별케이스로 알려져 있다.

- 커널PCA, autoencoder외에도, 조각함수 근사등으로 비선형을 다룰 수 있다.

02
순차데이터(1)

2-1 순차데이터집합

- 지금까지 데이터포인트들은 i.i.d.를 가정했음
  - independent
  - identically distributed
  - 여러개의 확률변수가 상호독립, 모두 동일한 분포
- 실제사례, 특히 순차데이터의 경우
  이런 가정이 틀릴때가 많다.

2-1 순차데이터집합

- 정류적 순차분포 (stationary sequential dist.)
  - 데이터가 생성되는 원 분포가 시간이 지나도 일정
- 비정류적 순차분포(non-stationary)
  - 생성분포 자체가 시간이 지남에 따라 변화

2-1 순차데이터집합

- 금융데이터 등에서, 이전 관측값을 바탕으로 다음 값을 예측할 필요가 있다.
- 최근의 관측값이 예전의 관측값보다 더 많은 정보를 가지고 있을 것이다.
  - 이전의 '모든' 데이터에 대한 종속성을 고려하는 것은 비현실적

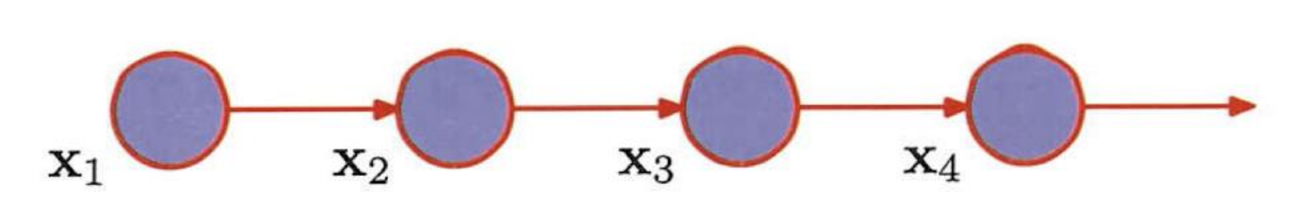
2-2 마르코프 모델

- 미래에 대한 예측값들이 가장 최근의 관측값에만 의존
    - 나머지에 대해서는 독립
    - 더 과거에 대한 의존성이 강한 데이터라면..
- 잠재변수 도입 : 상태공간모델
    - 은닉 마르코프 모델 : 잠재변수들이 이산
    - 선형동적시스템(LDS): 잠재변수들이 가우시안

2-2 마르코프모델

- 일차 마르코프 연쇄(1st order Markov chain)

$$p(\mathbf{x}_1, \dots, \mathbf{x}_N) = p(\mathbf{x}_1) \prod_{n=2}^{N} p(\mathbf{x}_n | \mathbf{x}_1, \dots, \mathbf{x}_{n-1})$$



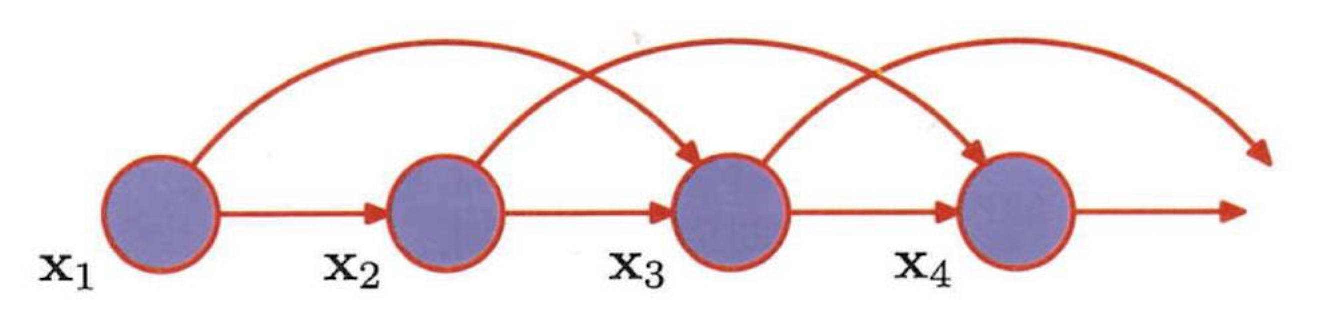
Bishop. fig13.3

$$p(\mathbf{x}_n | \mathbf{x}_1, \dots, \mathbf{x}_{n-1}) = p(\mathbf{x}_n | \mathbf{x}_{n-1}) \quad \because \text{d분리}$$

2-2 마르코프모델

- 이차 마르코프 연쇄

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1)\prod_{n=3}^{N} p(\mathbf{x}_{n-1}|\mathbf{x}_{n-2})$$



Bishop. fig13.4

- 의존이 많아질수록 매개변수 기하급수적 증가. 비실용적

2-2 마르코프모델

- 비슷한것: 자기회귀모델(autoregressive, AR)
  - 변수의 과거값의 선형조합으로 변수 예측가능한 모델
- 기타: 매개변수 모델에 뉴럴넷을 도입할 수 있다.

$$p(\mathbf{x}_n|\mathbf{x}_{n-M},\dots,\mathbf{x}_{n-1})$$

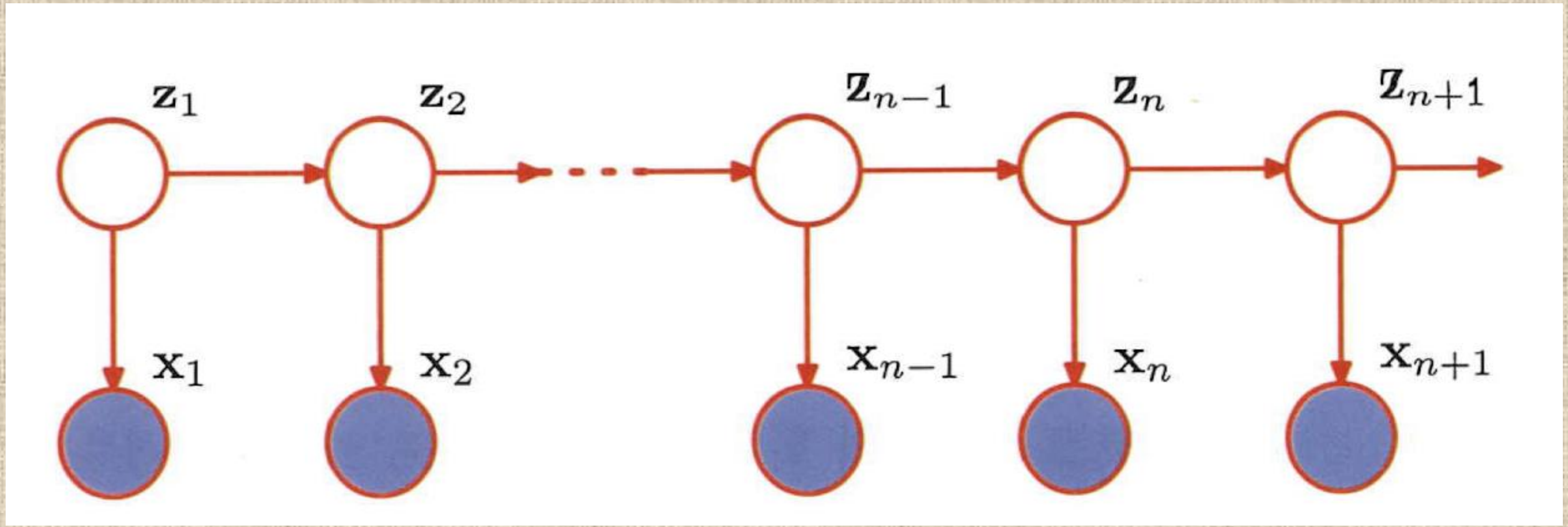
이전 M개의 변수 이용.

2-2 마르코프모델

- 마르코프 가정에 의해 제약되지 않는 순차모델
  - iid이면 매개변수가 너무 많다.
  - 잠재변수를 도입하여 매개변수 제약하는 아이디어: 상태공간모델(state space model)
    - 다음페이지 그림참고: d분리에 의해 $\mathbf{x}_n$모두 의존. (조건부독립아님)

2-2 마르코프모델

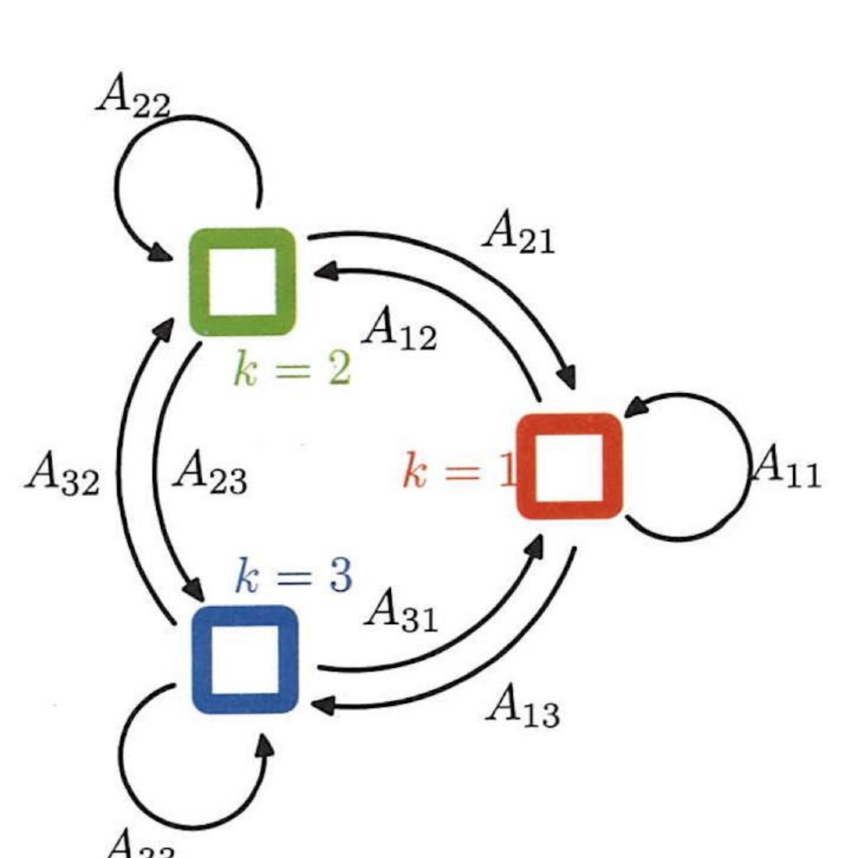
- 상태공간모델(state space model)



Bishop. fig13.5

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) = \mathrm{p}(\mathbf{z}_1)\left[\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1})\right]\prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n)$$

2-3 은닉 마르코프 모델(HMM)

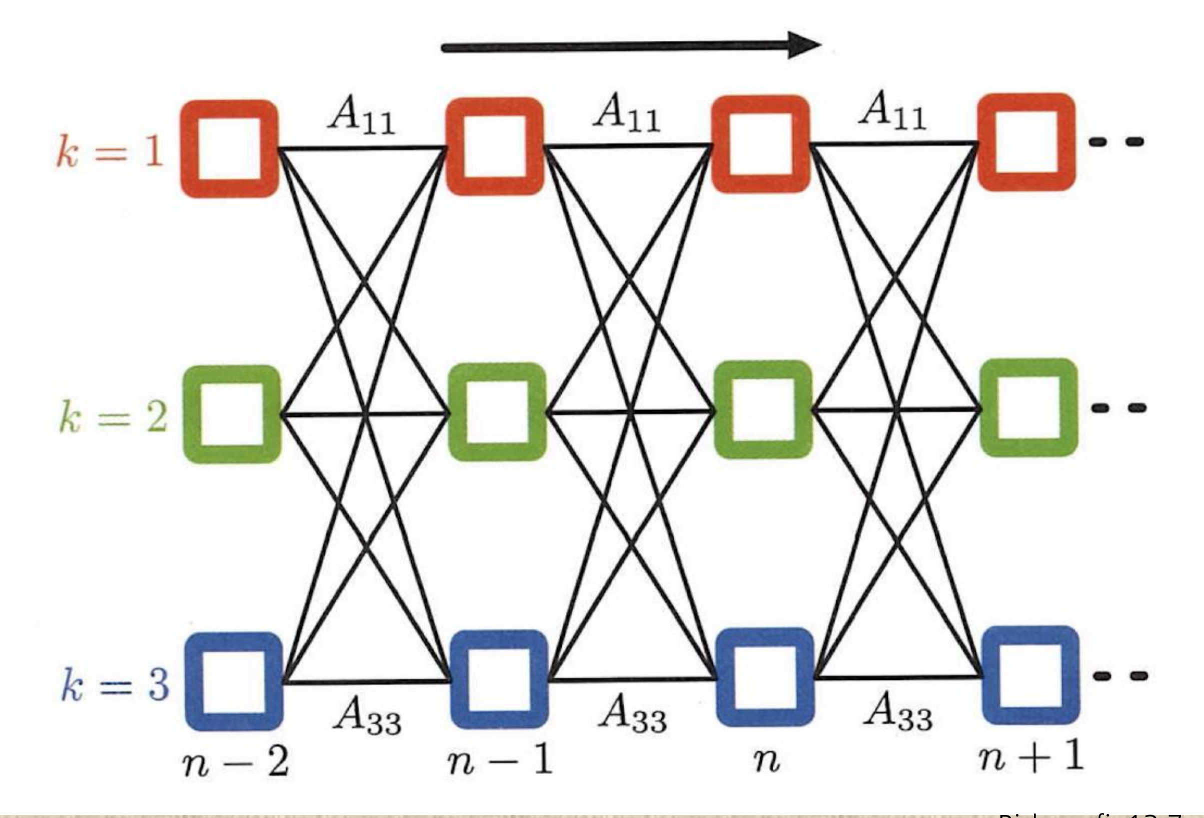


Bishop. fig13.6

- 상태공간모델의 일종 (잠재변수들이 이산)
- 단일시간조각만 보면 $p(\mathbf{x}|\mathbf{z})$인 혼합모델
- 상태별로 전이확률(행렬)이 있다.

2-3 은닉 마르코프 모델(HMM)

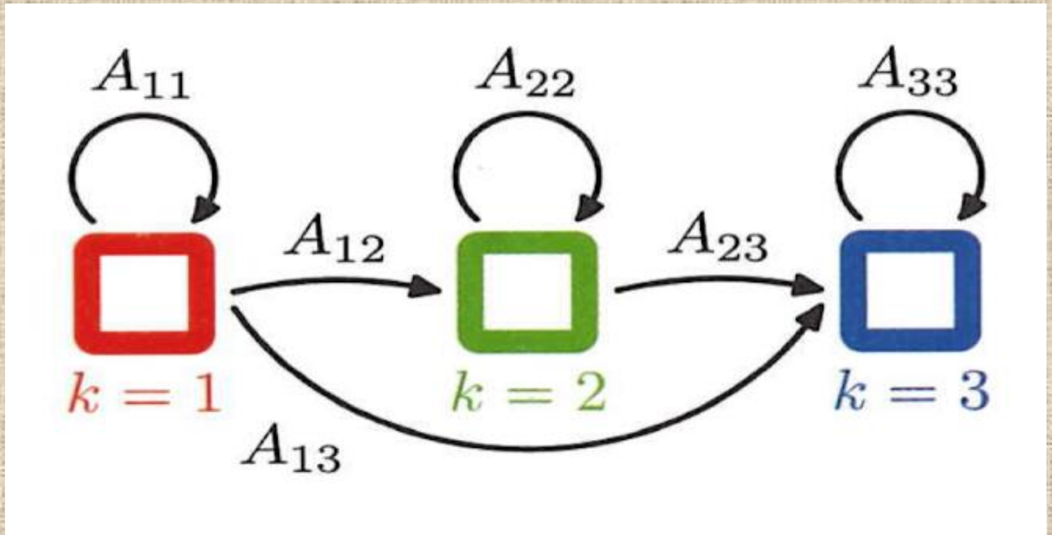
- 시간순으로 펼친것



Bishop. fig13.7

2-3 은닉 마르코프 모델(HMM)

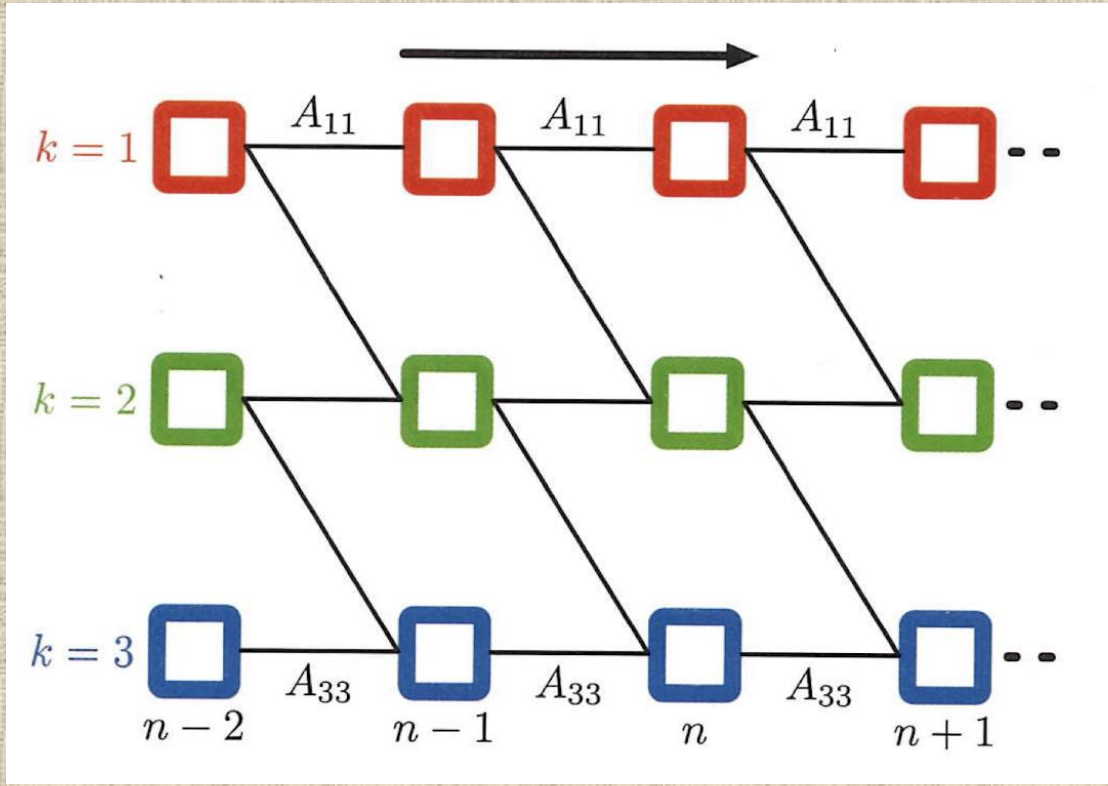
- 변형 : left-to-right HMM
    - 한번 떠나면 돌아올 수 없다.



Bishop. fig13.9

2-3 은닉 마르코프 모델(HMM)

- 변형 : left-to-right HMM에 더 제약을 가한 것.



Bishop. fig13.10

  - 하나 이상 움직일 수 없다.

다음시간

11강

- 순차데이터(2)