기계학습

14강 확률분포(2)

장필훈 교수

학습목차

1 확률분포(2)

01
확률분포(2)

1-1 다항변수

- K개중 하나를 선택해야 하는 이산변수를 활용해야 할 때, one hot encoding 사용 가능.

$$\mathbf{x} = (0,0,1,0,0,0)^{\mathrm{T}}$$

- $\sum_{k=1}^{K} x_k = 1$
- $P(x_k = 1) = \mu_k$일때, $p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$
- 베르누이분포를 결과가 두가지 이상인 경우로 확장한 것

1-1 다항변수

- μ_k는 확률이므로 $\sum_k \mu_k = 1$
- $\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_x p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_K)^T = \boldsymbol{\mu}$
- 데이터집합 $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$의 가능도 함수는

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mu_k^{x_{nk}} = \prod_{k=1}^{K}\mu_K^{\sum_n x_{nk}} = \prod_{k=1}^{K}\mu_K^{m_k}$$

where $m_k = \sum_n x_{nk}$

$x_k = 1$인 관측값의 숫자

1-1 다항변수

- μ의 최대가능도 해
  - $\sum \mu_k = 1$의 제약조건에서 $\ln p(\mathcal{D}|\boldsymbol{\mu})$를 찾기
  - 아래의 최댓값을 구한다. (라그랑주 승수법)

$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right)$$

1-1 다항변수

- μ_k에 대해 미분하고, 0으로 두면, $\mu_k = -m_k/\lambda$를 얻음.
- 제약조건 $\Sigma\mu = 1$에 대입하면 $\lambda = -N$
- 따라서, 최대가능도의 해는,

$$\mu_k^{ML} = \frac{m_k}{N}$$

○ 곧, ‘N개 관측값중 $x_k = 1$ 인 경우의 비율’

1-1 다항변수

- 매개변수 $\boldsymbol{\mu}$와 관측 숫자 N에 의해 결정되는 수량 $m_1, \ldots, m_K$
  - $\sum_{k=1}^{K} m_k = N$
  - 결합분포 $p(\mathcal{D}|\mu)$ = '다항분포' multinomial distribution

$$\mathrm{Mult}(m_1, \ldots, m_K|\boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

where $$\binom{N}{m_1 m_2 \ldots m_K} = \frac{N!}{m_1! m_2! \ldots m_K!}$$

1-1 다항변수

- 다항분포의 매개변수 μ_k의 사전분포는?
  - 켤레 사전분포는 다음의 형태여야 한다.

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

$$0 \leq \mu_k \leq 1, \ \sum_k \mu_k = 1, \quad \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^T$$

1-1 다항변수

- 다항분포의 매개변수 μ_k의 사전분포는? (계속)
  - 켤레 사전분포의 정규화된 형태 : 디리클레 분포

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1} \qquad \alpha_0 = \sum_{k=1}^{K}\alpha_k$$

- 디리클레 분포에 가능도 함수(=다항분포)를 곱하면, μ_k의 사후분포를 얻을 수 있다.

1-1 다항변수

- μ_k의 사후분포

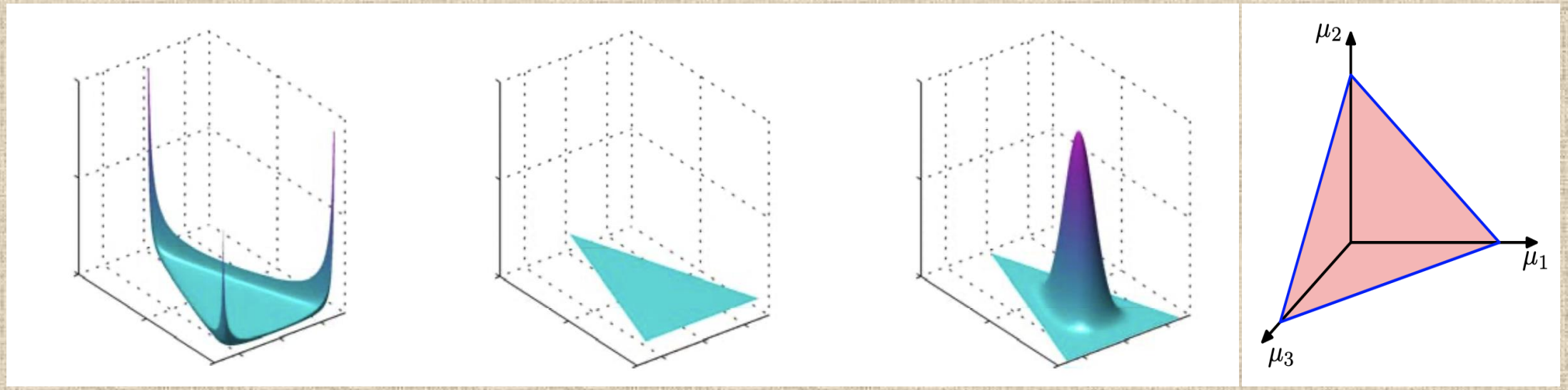
$$p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1}$$

  - 사후분포 역시 디리클레 분포
  - = '디리클레 분포는 다항분포의 켤레사전분포다'
  - 디리클레 분포의 $k=2$ 일때 분포가 베타분포

1-1 다항변수

- 디리클레 분포

Bishop. Fig.2.4, Fig.2.5



$\{\alpha_k\} = 0.1$ $\{\alpha_k\} = 1$ $\{\alpha_k\} = 10$

1-2 가우시안 분포

- =정규분포
- 단일변수

$$\mathcal{N}(x|\mu,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- D차원 벡터 $\mathbf{x}$

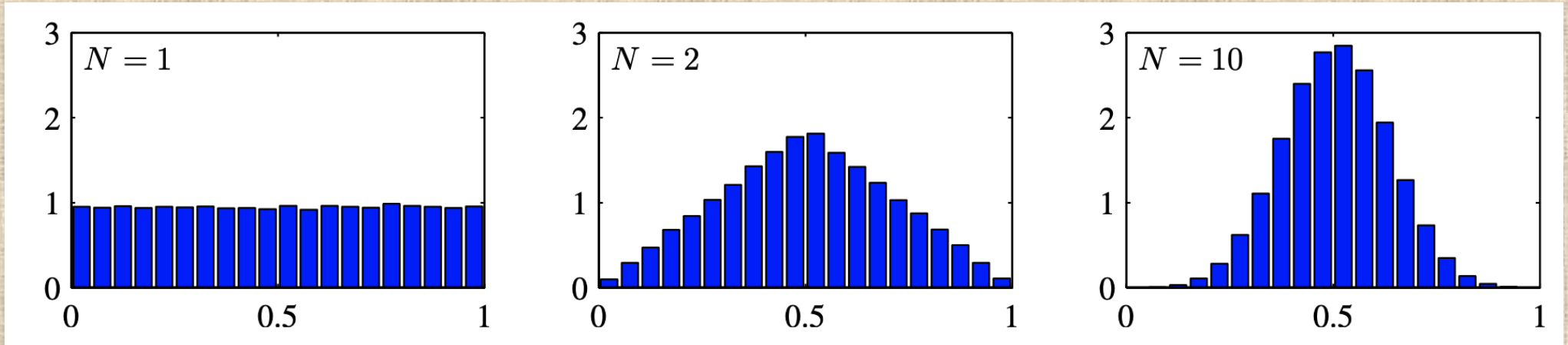
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}$$

행렬식

1-2 가우시안 분포

- 중심극한정리

Bishop. Fig.2.6

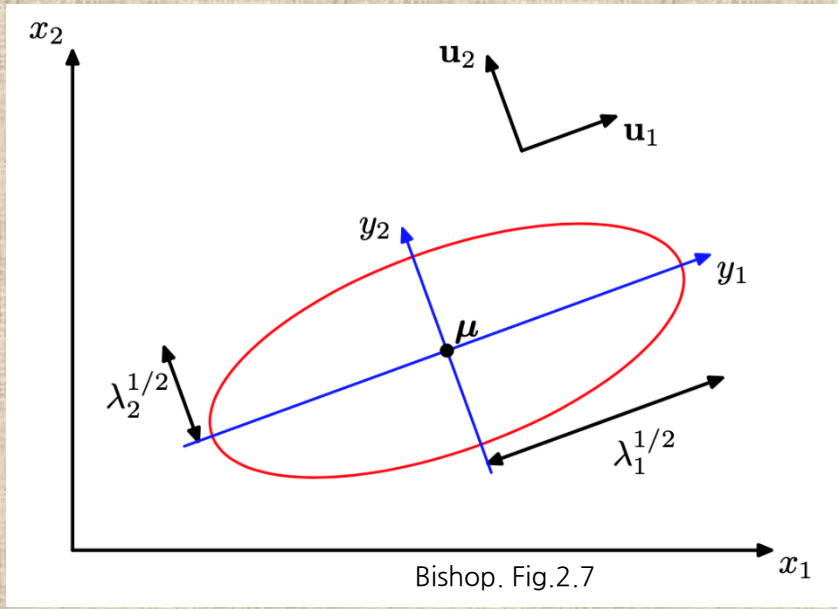


<균일하게 분포된 N개 값의 평균에 대한 히스토그램>

- “여러개의 확률변수들의 합에 해당하는 확률변수는 (일정한 조건 하에서) 합해지는 확률변수의 숫자가 증가함에 따라 점점 가우시안 분포가 되어간다.”

1-2 가우시안 분포

- 마할라노비스 거리 : $(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$
  - $\boldsymbol{\Sigma}$가 항등행렬이면 유클리디안.



Bishop. Fig.2.7

- 2차원 공간의 가우시안 예
- 타원의 축은 공분산행렬의 고유벡터들이 결정

1-2 가우시안 분포

- 기댓값

$$\mathbb{E}[\mathbf{x}] = \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathbf{x}d\mathbf{x}$$

$$= \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}(\mathbf{z}+\boldsymbol{\mu})d\mathbf{z}$$

$$= \boldsymbol{\mu}$$

우함수이고, $-\infty\sim\infty$적분하면 $\mathbf{z}$ 항이 대칭성으로 없어짐

1-2 가우시안 분포

- 이차모멘트

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathbf{x}\mathbf{x}^{\mathrm{T}}d\mathbf{x}$$

$$= \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}(\mathbf{z}+\boldsymbol{\mu})(\mathbf{z}+\boldsymbol{\mu})^{\mathrm{T}}d\mathbf{z}$$

  - 교차항($\boldsymbol{\mu}\mathbf{z}^{\mathrm{T}}$, $\mathbf{z}\boldsymbol{\mu}^{\mathrm{T}}$) 은 사라진다
  - 정규화되어 있으므로 $\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} = 1$

1-2 가우시안 분포

- 이차모멘트(계속)
  - $\mathbf{z}\mathbf{z}^{\mathrm{T}}$항은?

$$\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}\mathbf{z}\mathbf{z}^{\mathrm{T}}d\mathbf{z} = \boldsymbol{\Sigma}$$

  - 결국, $\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$
  - 이를 바탕으로 공분산을 구함

1-2 가우시안 분포

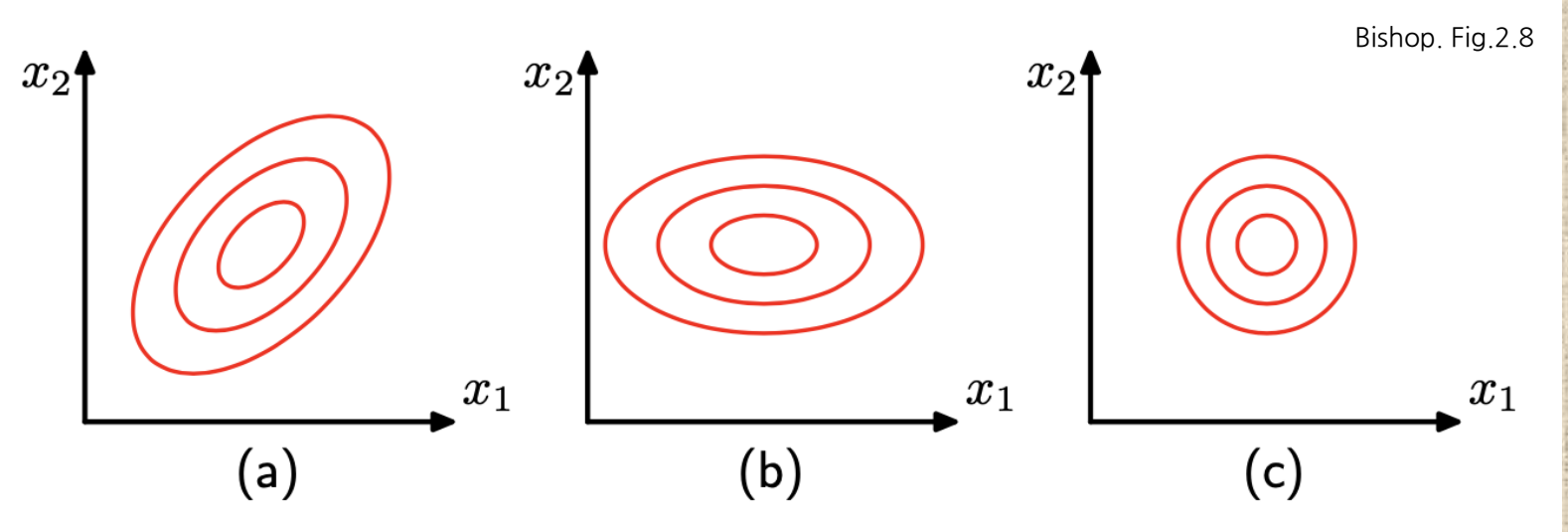
- 공분산

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T] = \boldsymbol{\Sigma}$$

- 공분산행렬은 차원D에 따라 매개변수가 D^2로 증가한다.
  - 계산량이 많음.
  - 대각행렬을 사용하면 D로 증가
    - 좌표축상으로 정렬된 타원

1-2 가우시안 분포

- 이차원 가우시안에서 상수확률밀도



(a) 일반적인 공분산행렬 (b) 대각행렬 (c) 등방성 공분산(항등행렬의 상수배)
isotropic

1-2 가우시안 분포

- unimodal이 기본이기 때문에 multimodal에 대해 적절한 근사치를 제공하지 않는다.
  - latent variable(잠재변수)를 이용해서 해결한다.
    - 예1) (이미 배운) 가우시안 혼합모델
    - 예2) (이미 배운) 선형동적 시스템

1-2 가우시안 분포

- 두 변수집합의 결합분포가 가우시안이면,
  - 하나의 변수집합에 대한 다른 변수집합의 조건부도.
  - 각 변수집합의 주변분포(marginal dist.)도 가우시안.
- 조건부 분포의 경우
  - D차원 벡터 x, $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
    - $\mathbf{x}$ 를 두개의 부분집합 $\mathbf{x}_a$ 와 $\mathbf{x}_b$ 로 나누어서 생각

1-2 가우시안 분포

- 그러면 $\mathbf{x}$ 를 $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)^{\mathrm{T}}$라 하고,
- 각 부분집합의 평균값벡터 $\boldsymbol{\mu} = (\boldsymbol{\mu}_a, \boldsymbol{\mu}_b)^{\mathrm{T}}$
- 공분산행렬 $\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$

- 편의상, 공분산행렬의 역행렬(정밀도 행렬)을 정의.
  precision matrix

  $\Lambda \equiv \Sigma^{-1}$

1-2 가우시안 분포

- 대칭행렬의 역행렬도 대칭행렬이기 때문에, Λ 도 대칭

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

  - Λ_{aa}가 Σ_{aa}의 대칭은 아님
- 이제 결합분포의 exp부분을 위 분할을 이용해서 나타내면,

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) =$$

1-2 가우시안 분포

- (분할계속)

$$-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a)-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)$$

$$-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a)-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b)$$

- $\mathbf{x}_a$에 관해 2차식이 유지된다.
  - 따라서 조건부분포 $p(\mathbf{x}_a|\mathbf{x}_b)$도 가우시안.

1-2 가우시안 분포

- 평균과 공분산을 앞의 식을 이용해 구한다.
  - 분포가 나오면 늘 하는 일.
- 계산과정은 길다… 계산결과는,

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$

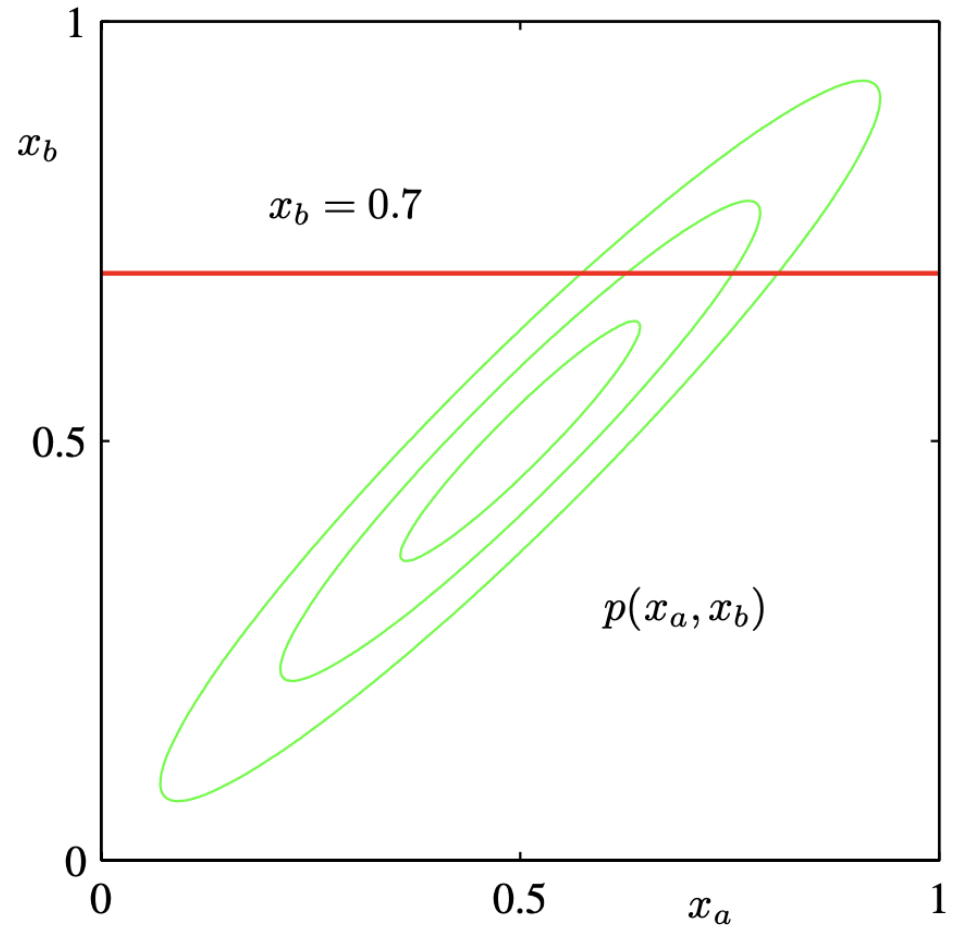
$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$$

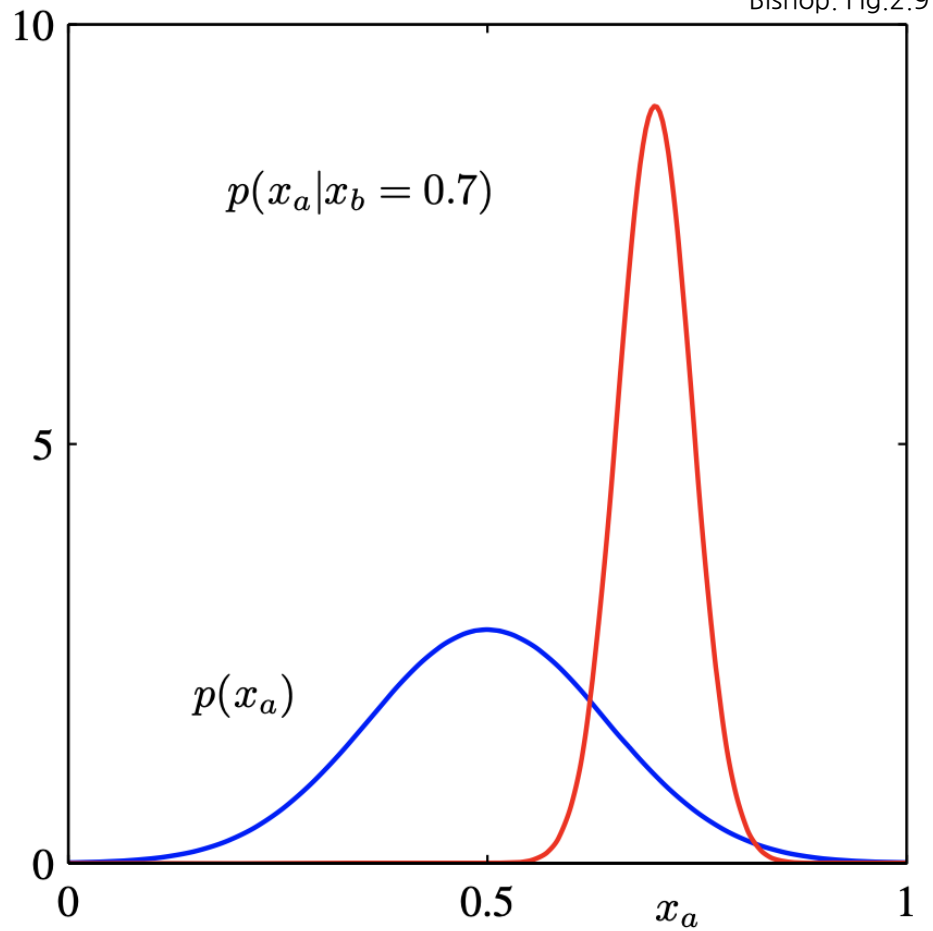
1-2 가우시안 분포

- 주변분포 $p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b$
- 이 결과가 가우시안임을 보이면 됨.
  - 앞의 분할식에서 $\mathbf{x}_b$와 연관된 항만 뽑아서 검토해보면, 해석적으로 적분 가능하고, 따라서 평균과 분산을 구할 수 있다.(과정생략)
  - $\mathbb{E}[\mathbf{x}_a] = \boldsymbol{\mu}_a, \mathrm{cov}[\mathbf{x}_a] = \boldsymbol{\Sigma}_{aa}$: 직관과 일치.

1-2 가우시안 분포

Bishop. Fig.2.9





1-2 가우시안 분포

- 최대가능도

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

- 가능도 함수는 다음 두 값을 통해서만 데이터집합에 종속

  : **충분통계량** $\sum_{n=1}^{N}\mathbf{x}_n$ $\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T$

1-2 가우시안 분포

- 로그가능도에 대한 미분값=0으로 놓고 해를 구하면,

$$\frac{\partial}{\partial \mu} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = 0$$

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$$ 관측된 데이터포인트들의 평균

- $\boldsymbol{\Sigma}$에 대해 최대화하는 것도 같은방법(이지만 아주 어려움)

1-2 가우시안 분포

- 최대 가능도의 **순차추정** : 데이터 포인트들을 하나씩 처리
- N개의 관측값을 바탕으로 한 추정값을 $\mu_{ML}^{(N)}$이라 하면,

$$\mu_{ML}^{(N)} = \frac{1}{\mathrm{N}}\sum_{n=1}^{N}\mathbf{x}_n = \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathbf{x}_n = \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\mu_{ML}^{(N-1)}$$

$$= \mu_{ML}^{(N-1)} + \frac{1}{N}(\mathbf{x}_N - \mu_{ML}^{(N-1)})$$

1-2 가우시안 분포

- 베이지안 추론 : 매개변수들의 사전분포를 정의.
- **예1**: 분산을 알고 있는 상태에서 평균추정문제
  - 관찰값은 N개 $\mathbf{x} = \{x_1, \dots, x_N\}$
  - 가능도함수는

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

1-2 가우시안 분포

- 사전분포를 가우시안으로 선택하면, $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$
- 사후분포도 가우시안이 된다. $p(\mu|\mathbf{X}) \propto p(\mathbf{X}|\mu)p(\mu)$

$$p(\mu|\mathbf{X}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

1-2 가우시안 분포

- $p(\mu|\mathbf{X}) \propto p(\mathbf{X}|\mu)p(\mu)$ http://gregorygundersen.com/blog/2019/04/04/bayesian-gaussian/

$$= \left(\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\} \right) \left(\frac{1}{(2\pi\sigma_0^2)^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right\} \right)$$

$$\approx \exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n^2 - 2\mu x_n + \mu^2) - \frac{1}{2\sigma_0^2} (\mu^2 - 2\mu\mu_0 + \mu_0^2) \right\}$$

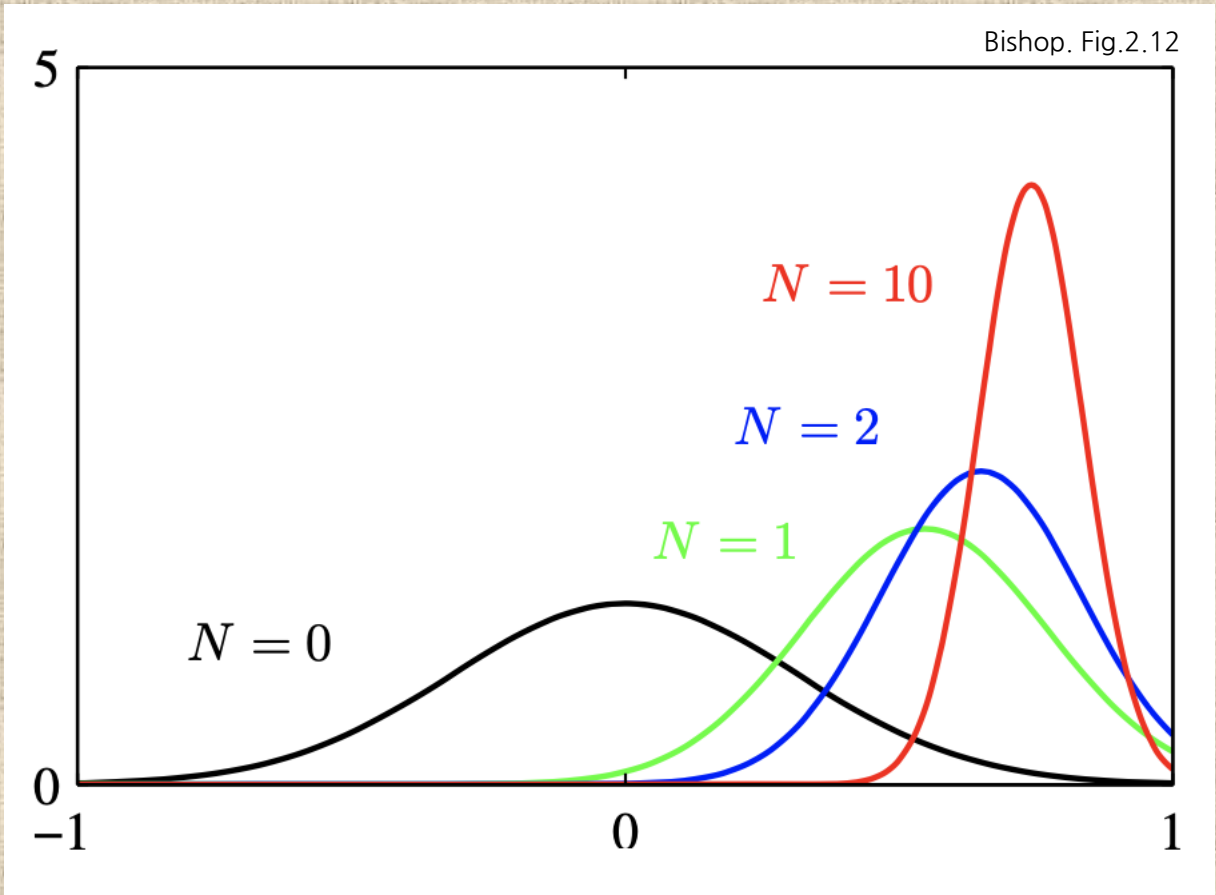
$$\approx \exp\left\{ -\frac{1}{2\sigma^2} (N\mu^2 - 2\mu N\bar{x}) - \frac{1}{2\sigma_0^2} (\mu^2 - 2\mu\mu_0) \right\}$$

$$= \exp\left\{ -\frac{1}{2} \left(\mu^2 \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \right) - 2\mu \left(\frac{N\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) \right) \right\} = \exp\left\{ -\frac{1}{2} \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \right) (\mu - \mu_N)^2 \right\}$$

1-2 가우시안 분포

- μ_N이 N(관찰데이터 수)에 따라 어떻게 변하는지 관찰
  - $N = 0$ 이면?
  - $N \to \infty$ 이면?
- 정밀도(=분산의 역)는 N에 따라 어떻게 변하는가?
  - 정밀도는 증가하는가 감소하는가
  - $N \to \infty$이면 μ_{ML}근처로 무한대의 뾰족한 정점.

1-2 가우시안 분포



- 예1(σ^2은 알려져있고 μ에 대한 베이지안 추론)에 대한 도식화. 사전분포의 평균은 0 이고 데이터포인트의 평균은 0.8, 분산은 0.1이다. N=0이 사전분포. N에 따른 사후분포들($p(\mu|\mathbf{x})$)을 그린 것이다.

1-2 가우시안 분포

- **예2**: 평균을 알고 있는 상태에서 분산추정문제
  - 분산 대신 정밀도($\lambda \equiv 1/\sigma^2$)를 쓰는 것이 편하다.
  - λ의 가능도 함수

$$p(X|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{\frac{N}{2}} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

    - λ의 power, λ의 선형함수의 지수함수에 비례

1-3 감마 분포

- 이러한 조건을 만족시키는 분포는 감마분포

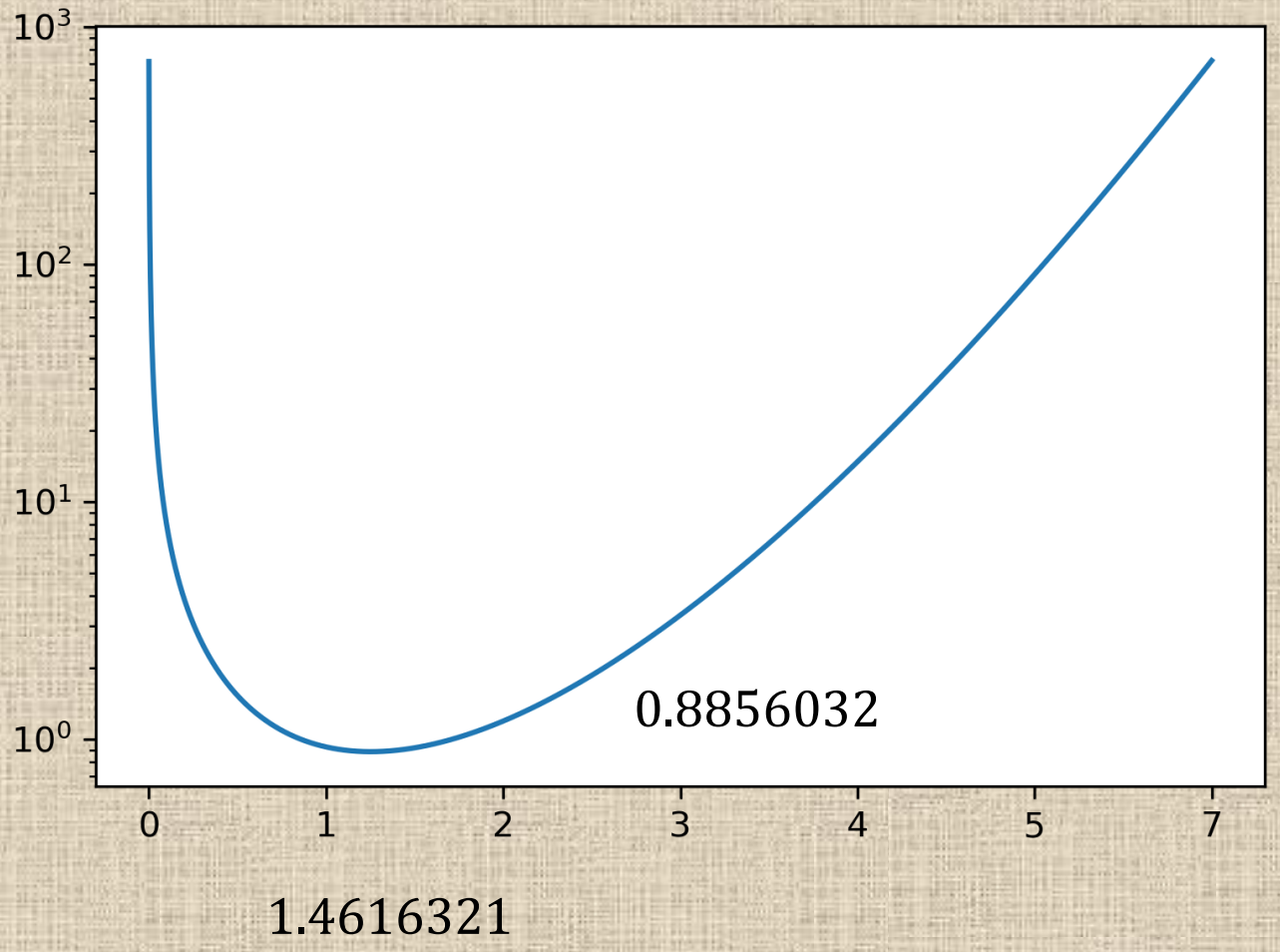
$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

- $\mathbb{E}[\lambda] = \frac{a}{b}, \quad \mathrm{var}[\lambda] = \frac{a}{b^2}$
- 사후분포는 다음과 같게 된다.

$$p(\lambda|\mathbf{X}) \propto \lambda^{a_0-1} \lambda^{\frac{N}{2}} \exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

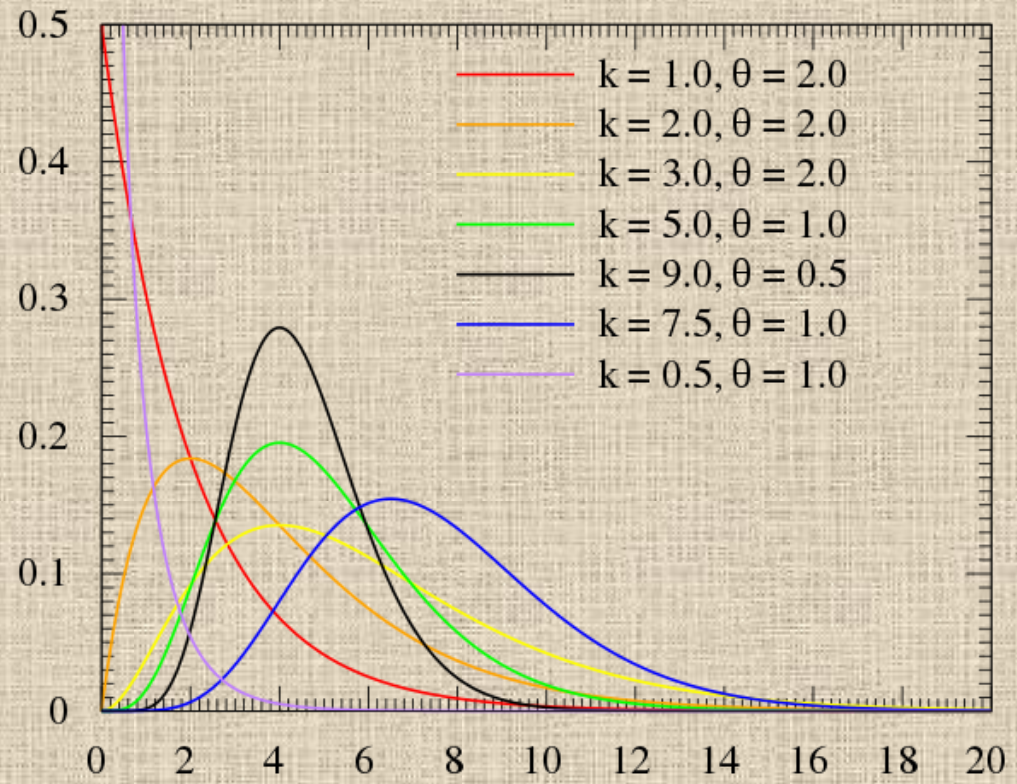
1-3 감마 분포

- $\Gamma(x)$

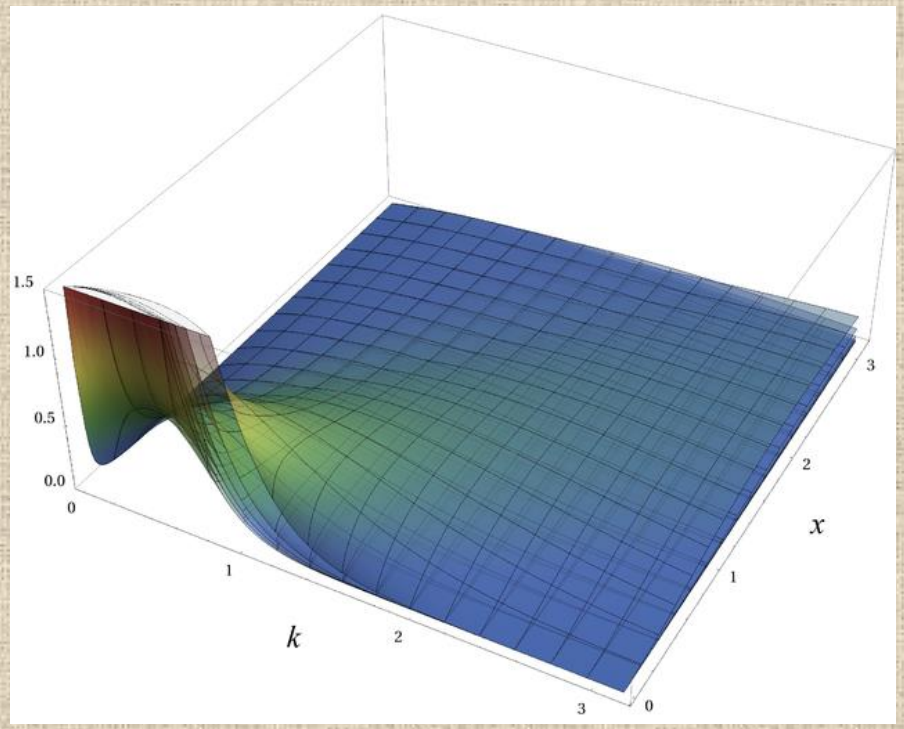


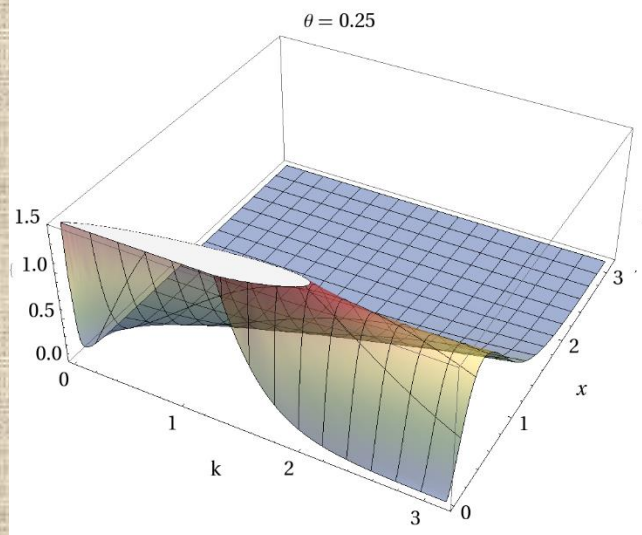
1-3 감마 분포

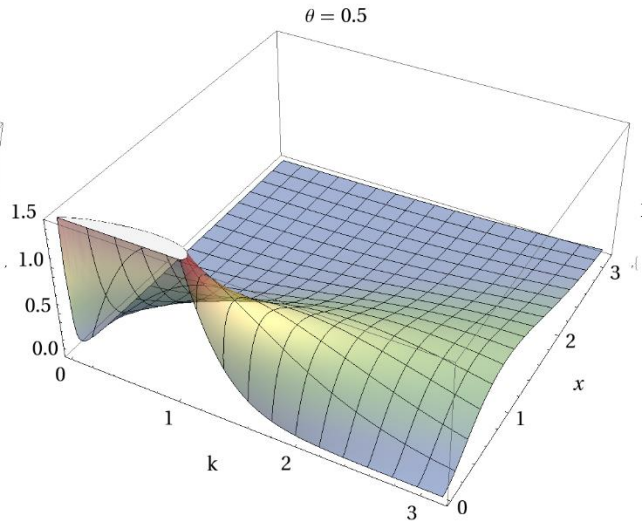
- $k = a, \theta = 1/b$

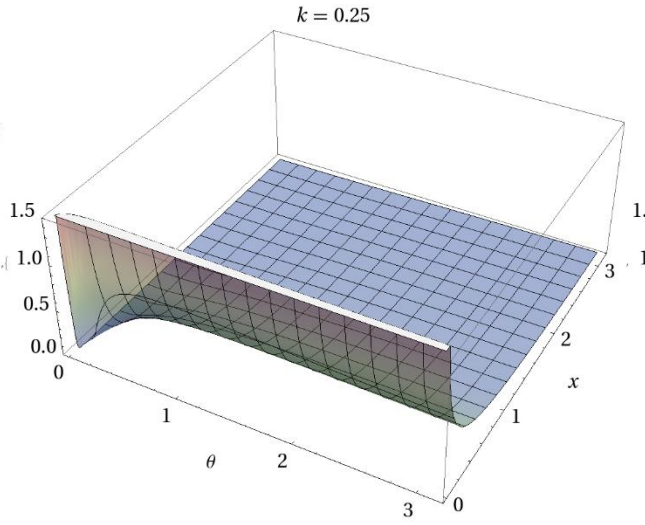


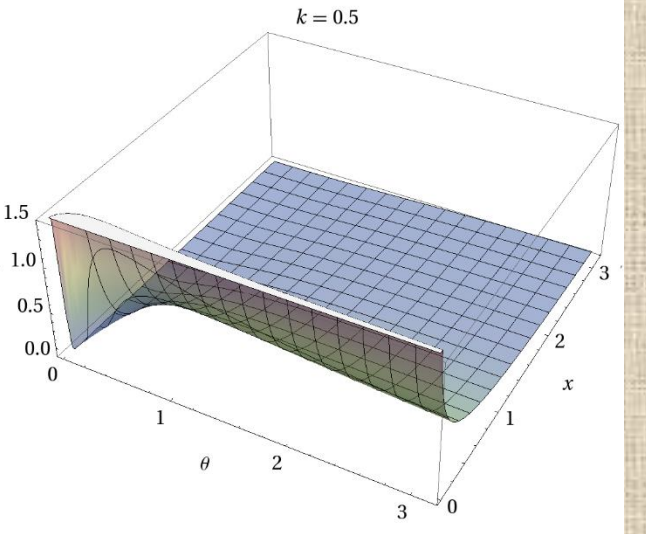
https://en.wikipedia.org/wiki/Gamma_distribution

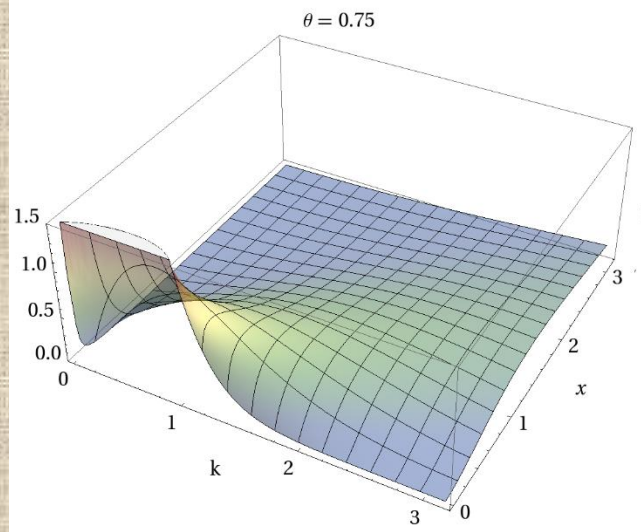


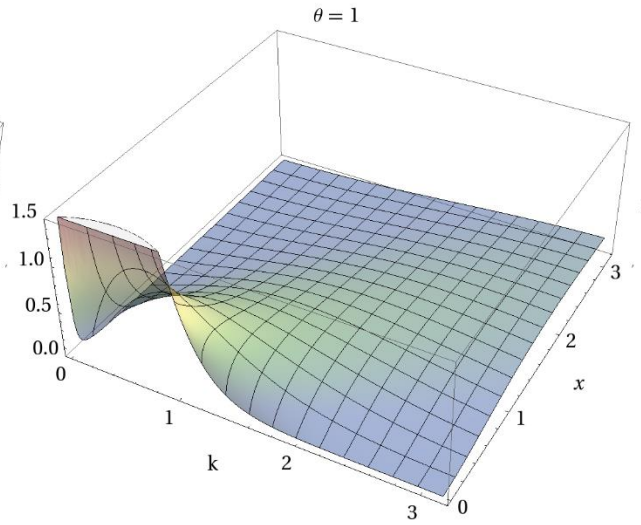


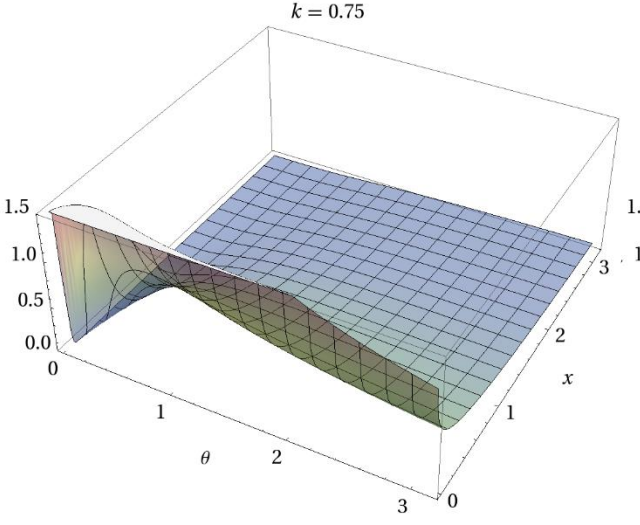


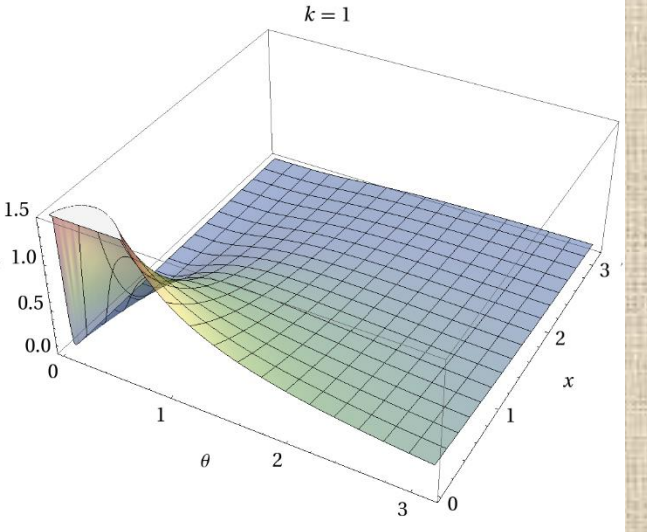






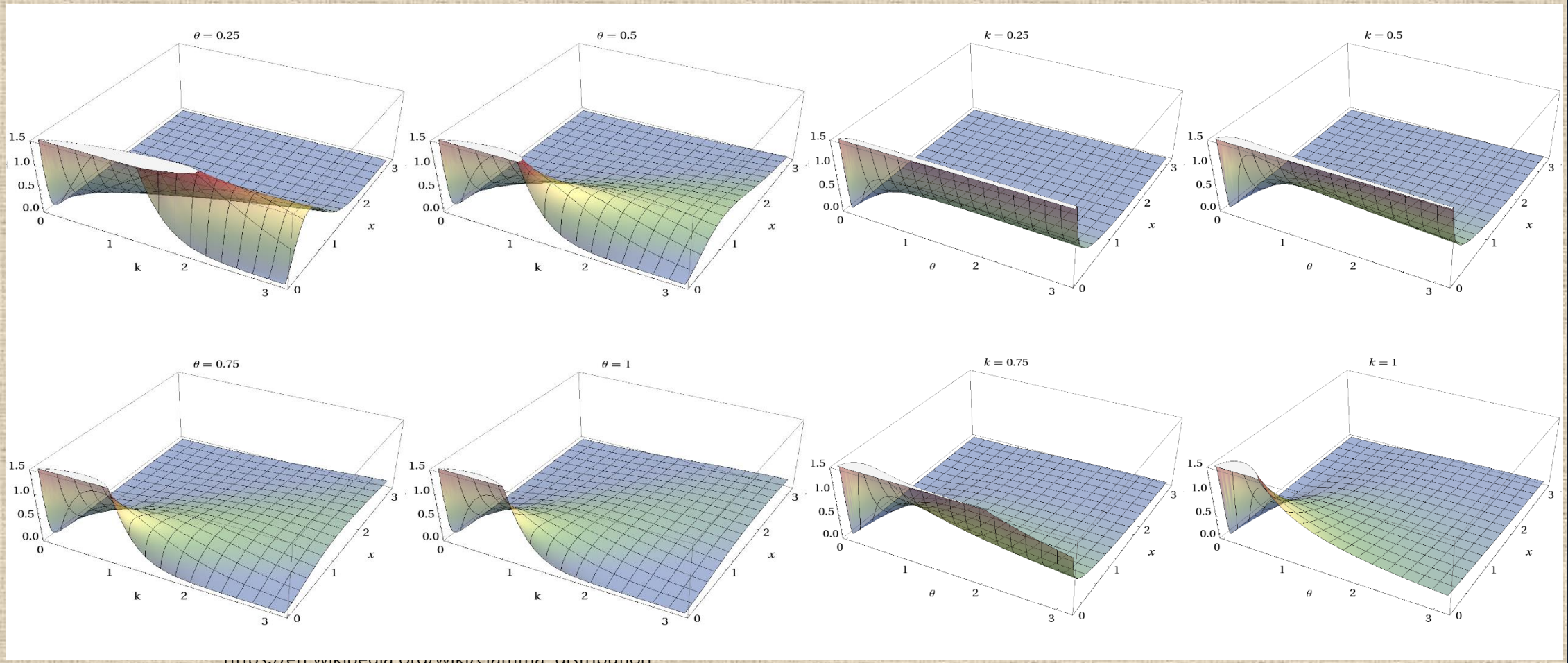






https://en.wikipedia.org/wiki/Gamma_distribution

1-3 감마 분포



https://en.wikipedia.org/wiki/Gamma_distribution

다음시간

14강

- 확률분포(3)