1

Machine Learning

Chapter 2: Probability Distributions

林嘉文 (Chia-Wen Lin)
清華大學電機系
cwlin@ee.nthu.edu.tw

3/8/2021

2

Parametric Distributions

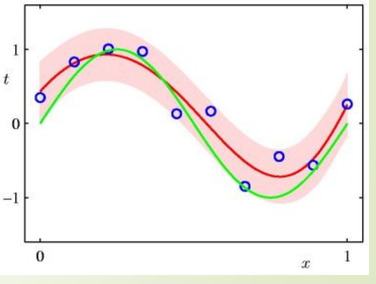
Basic building blocks: $p(\mathbf{x}|\boldsymbol{\theta})$

Need to determine $\boldsymbol{\theta}$ given $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

Representation: $\boldsymbol{\theta}^\star$ or $p(\boldsymbol{\theta})$?

Recall Curve Fitting

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{w}$$



3/8/2021

3

Binary Variables (1)

Jacob Bernoulli

Coin flipping: heads = 1, tails = 0

$$p(x=1|\mu)=\mu$$

Bernoulli Distribution

$$\begin{aligned}\mathrm{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu)\end{aligned}$$

3/8/2021

4

Binary Variables (2)

N coin flips:

$$p(m \text{ heads}|N,\mu)$$

Binomial Distribution

$$\mathrm{Bin}(m|N,\mu)=\binom{N}{m}\mu^m(1-\mu)^{N-m}$$

$$\mathbb{E}[m]\equiv\sum_{m=0}^{N} m\mathrm{Bin}(m|N,\mu)=N\mu$$

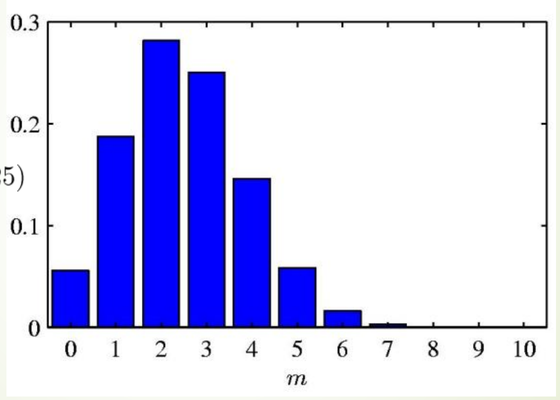
$$\mathrm{var}[m]\equiv\sum_{m=0}^{N}(m-\mathbb{E}[m])^2\,\mathrm{Bin}(m|N,\mu)=N\mu(1-\mu)$$

3/8/2021

5

Binomial Distribution

$\mathrm{Bin}(m|10, 0.25)$



6

Parameter Estimation (1)

- **Likelihood function** $P(D|\boldsymbol{\mu})$
- **Prior** $P(\boldsymbol{\mu})$
- **Posterior** $P(\boldsymbol{\mu}|D)$

- Conjugate prior:

  Prior $P(\boldsymbol{\mu})$ is the **conjugate prior** for a **likelihood function**$P(D|\boldsymbol{\mu})$ if the **prior** $P(\boldsymbol{\mu})$ and the **posterior** $P(\boldsymbol{\mu}|D)$ have the same form.
- Example (coin flip problem)
  - **Prior** $P(\mu)$: $Beta(\beta_1, \beta_0)$; **Likelihood** $P(D|\mu)$: Binomial $\mu^{\alpha_1}(1-\mu)^{\alpha_0}$
  - **Posterior** $P(\mu|D)$: $Beta(\alpha_1 + \beta_1, \alpha_0 + \beta_0)$

7

Parameter Estimation (2)

ML for Bernoulli

Given: $\mathcal{D} = \{x_1, \ldots, x_N\}$, m heads (1), $N - m$ tails (0)

Likelihood function

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1 - x_n)\ln(1-\mu)\}$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{m}{N}$$

3/8/2021

8

Parameter Estimation (3)

Example: $\mathcal{D} = \{1, 1, 1\} \rightarrow \mu_{\mathrm{ML}} = \frac{3}{3} = 1$

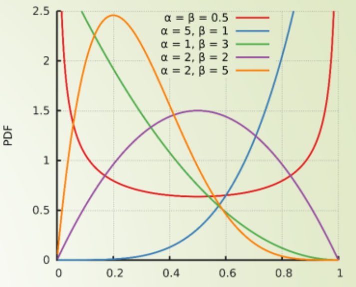
Prediction: *all* future tosses will land heads up

Overfitting to $\mathcal{D}$

3/8/2021

9

Beta Distribution



Distribution over $\mu \in [0, 1]$.

$$\begin{aligned} \mathrm{Beta}(\mu|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\ \mathbb{E}[\mu] &= \frac{a}{a+b} \\ \mathrm{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned}$$

3/8/2021

10

Bayesian Bernoulli

$$\begin{aligned} p(\mu|a_0,b_0,\mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu|a_0,b_0) \\ &= \left(\prod_{n=1}^{N}\mu^{x_n}(1-\mu)^{1-x_n}\right)\mathrm{Beta}(\mu|a_0,b_0) \\ &\propto \mu^{m+a_0-1}(1-\mu)^{(N-m)+b_0-1} \\ &\propto \mathrm{Beta}(\mu|a_N,b_N) \end{aligned}$$

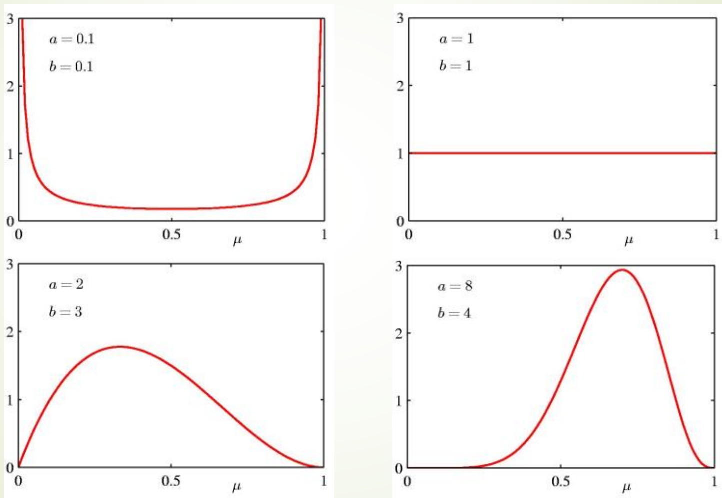
$$a_N = a_0 + m \qquad b_N = b_0 + (N - m)$$

The Beta distribution provides the **conjugate prior** for the Bernoulli distribution.

3/8/2021

11

Beta Distribution



3/8/2021

12

Prior · Likelihood = Posterior

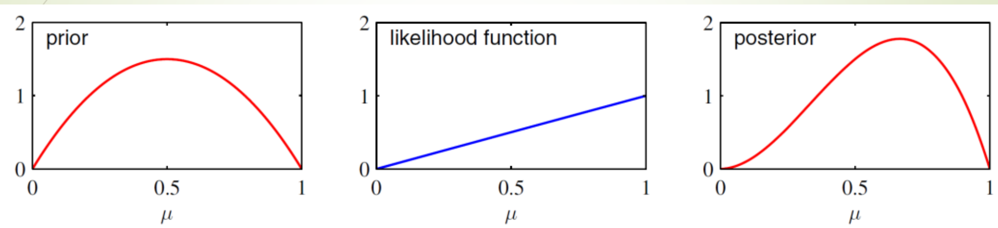


Figure 2.3 Illustration of one step of sequential Bayesian inference. The prior is given by a beta distribution with parameters $a = 2$, $b = 2$, and the likelihood function, given by (2.9) with $N = m = 1$, corresponds to a single observation of $x = 1$, so that the posterior is given by a beta distribution with parameters $a = 3$, $b = 2$.

3/8/2021

13

Properties of the Posterior

As the size of the data set, N, increases

$$\begin{aligned} a_N &\rightarrow m \\ b_N &\rightarrow N - m \\ \mathbb{E}[\mu] &= \frac{a_N}{a_N + b_N} \rightarrow \frac{m}{N} = \mu_{\mathrm{ML}} \\ \mathrm{var}[\mu] &= \frac{a_N b_N}{(a_N + b_N)^2 (a_N + b_N + 1)} \rightarrow 0 \end{aligned}$$

14

Prediction under the Posterior

What is the probability that the next coin toss will land heads up?

$$p(x = 1|\mathcal{D}) = \int_0^1 p(x = 1|\mu) p(\mu|\mathcal{D}) d\mu = \int_0^1 \mu p(\mu|\mathcal{D}) d\mu = \mathbb{E}[\mu|\mathcal{D}]$$

$$p(x = 1|\mathcal{D}) = \frac{a_N}{a_N + b_N}$$

15

Multinomial Variables

1-of-K coding scheme: $\mathbf{x} = (0,0,1,0,0,0)^{\mathrm{T}}$

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

$$\forall k : \mu_k \geqslant 0 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1$$

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}} = \boldsymbol{\mu}$$

$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1$$

16

ML Parameter Estimation

Given: $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

$m_k = \sum_n x_{nk}$

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

To ensure $\sum_k \mu_k = 1$, use a Lagrange multiplier, λ.

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

$$\mu_k = -m_k/\lambda \qquad \mu_k^{\mathrm{ML}} = \frac{m_k}{N}$$

$$\sum_k \mu_k = \sum_k -m_k/\lambda = 1$$

17

The Multinomial Distribution

$$
\begin{aligned}
\mathrm{Mult}(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) &= \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k} \\
\mathbb{E}[m_k] &= N\mu_k \\
\mathrm{var}[m_k] &= N\mu_k(1-\mu_k) \\
\mathrm{cov}[m_j m_k] &= -N\mu_j\mu_k
\end{aligned}
$$

3/8/2021

18

The Dirichlet Distribution

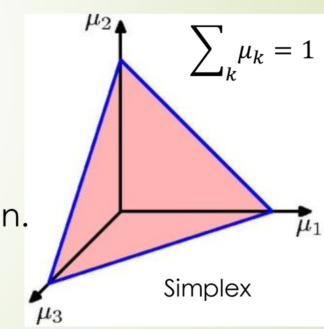
Lejeune Dirichlet

Also known as multivariate beta distribution (MBD)

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

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k$$

Conjugate prior for the multinomial distribution.



3/8/2021

19

Bayesian Multinomial (1)

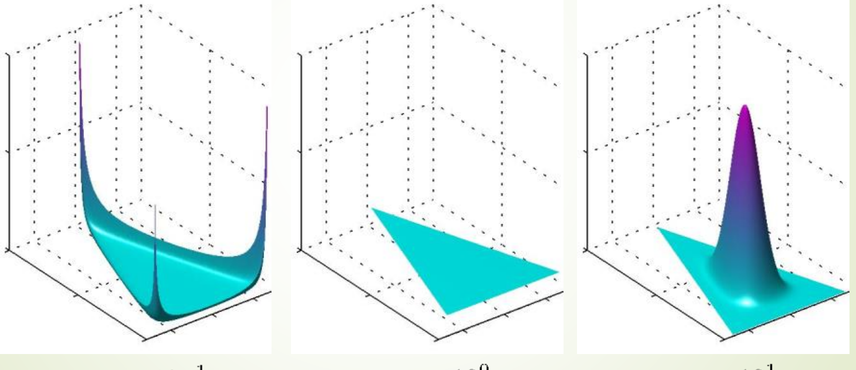
$$p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1}$$

$$\begin{aligned} p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) &= \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}+\mathbf{m}) \\ &= \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\cdots\Gamma(\alpha_K+m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1} \end{aligned}$$

3/8/2021

20

Bayesian Multinomial (2)

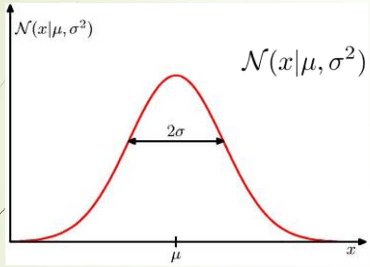


$\alpha_k = 10^{-1}$ $\alpha_k = 10^{0}$ $\alpha_k = 10^{1}$

3/8/2021

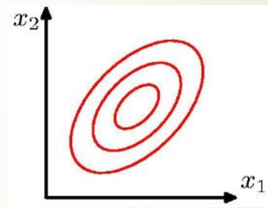
21

The Gaussian Distribution



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

Carl Friedrich Gauss



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

Square of Mahalanobis distance

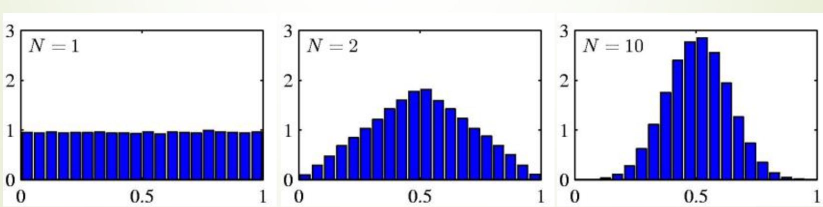
3/8/2021

22

Central Limit Theorem

The distribution of the sum of N i.i.d. random variables becomes increasingly Gaussian as N grows.

Example: N uniform $[0,1]$ random variables.



3/8/2021

23

Geometry of the Multivariate Gaussian

Δ: the Mahalanobis distance

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

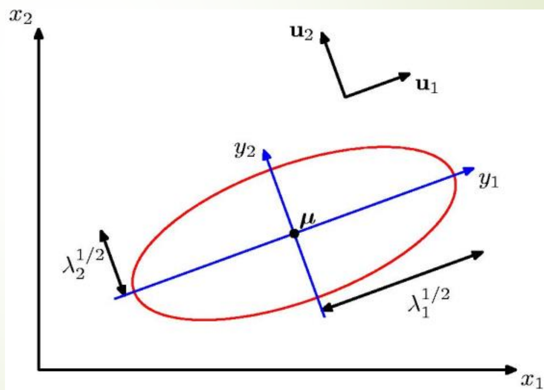
$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}$$

$$\Delta^2 = \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i}$$

$$y_i = \mathbf{u}_i^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu})$$

$$\mathbf{y} = \mathbf{U}(\mathbf{x}-\boldsymbol{\mu})$$

(KLT-Transform, PCA)



3/8/2021

24

Moments of the Multivariate Gaussian (1)

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathbf{x}\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \boxed{\exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}(\mathbf{z}}+\boldsymbol{\mu})\,\mathrm{d}\mathbf{z}\end{aligned}$$

$$\mathbf{z} = \mathbf{x}-\boldsymbol{\mu}$$

thanks to anti-symmetry of z

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

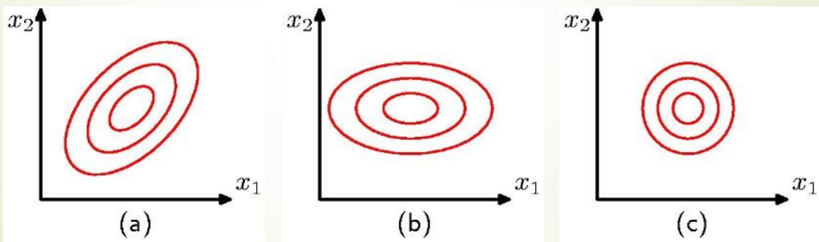
3/8/2021

25

Moments of the Multivariate Gaussian (2)

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$$

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



General Σ Diagonal Σ Identity Σ

3/8/2021

26

Partitioned Gaussian Distributions

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

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

$$\boldsymbol{\Sigma}^{\mathrm{T}} = \boldsymbol{\Sigma}$$

Precision matrix:

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

$$\boldsymbol{\Lambda}^{\mathrm{T}} = \boldsymbol{\Lambda}$$

3/8/2021

27

Partitioned Conditionals and Marginals

Conditional distribution

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \\
\boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\right\} \\
&= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\
&= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)
\end{aligned}$$

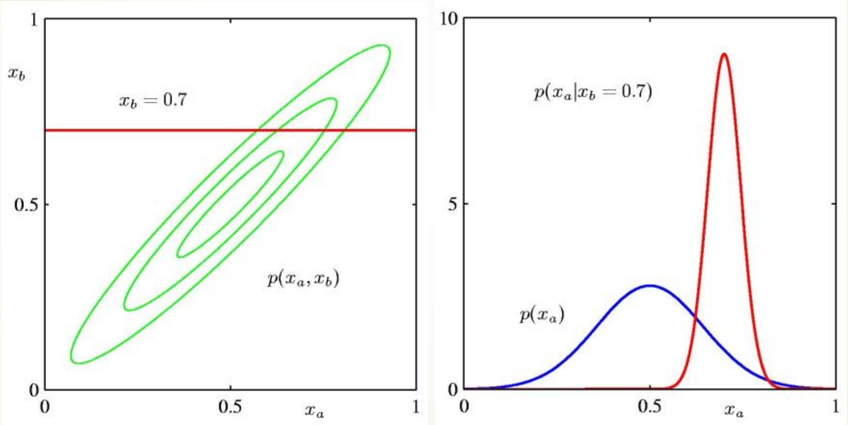
Marginal distribution

$$\begin{aligned}
p(\mathbf{x}_a) &= \int p(\mathbf{x}_a, \mathbf{x}_b)\,\mathrm{d}\mathbf{x}_b \\
&= \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})
\end{aligned}$$

3/8/2021

28

Partitioned Conditionals and Marginals



3/8/2021

29

Bayes' Theorem for Gaussian Variables (1)

Given

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}\right) \end{aligned}$$

we have

$$\begin{aligned} p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \\ p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \end{aligned}$$

where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}$

3/8/2021

30

Bayes' Theorem for Gaussian Variables (2)

Define $\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$

we have

$$\begin{aligned} \ln p(\mathbf{z}) &= \ln p(\mathbf{x}) + \ln p(\mathbf{y}|\mathbf{x}) \\ &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}) \\ &\quad -\frac{1}{2}(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b})^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b}) + \text{const} \end{aligned}$$

3/8/2021

31

Bayes' Theorem for Gaussian Variables (3)

To find the precision of this Gaussian, we consider the second order terms

$$-\frac{1}{2}\mathbf{x}^{\mathrm{T}}(\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})\mathbf{x}-\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y}+\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{A}\mathbf{x}+\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{y}$$
$$= -\frac{1}{2}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^{\mathrm{T}}\underbrace{\begin{pmatrix}\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} & -\mathbf{A}^{\mathrm{T}}\mathbf{L}\\ -\mathbf{L}\mathbf{A} & \mathbf{L}\end{pmatrix}}_{\mathbf{R}}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = -\frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{R}\mathbf{z}$$

The covariance matrix is found by taking the inverse of the precision

$$\mathrm{cov}[\mathbf{z}] = \mathbf{R}^{-1} = \begin{pmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\end{pmatrix}$$

3/8/2021

32

Bayes' Theorem for Gaussian Variables (4)

$$\mathbf{x}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\mu}-\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b}+\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{b} = \begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu}-\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b}\\ \mathbf{L}\mathbf{b}\end{pmatrix}$$

the mean of **z** is given by

$$\mathbb{E}[\mathbf{z}] = \mathbf{R}^{-1}\begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu}-\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{b}\\ \mathbf{L}\mathbf{b}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{A}\boldsymbol{\mu}+\mathbf{b}\end{pmatrix}$$

The mean and covariance of **y** are given by

$$\begin{aligned}\mathbb{E}[\mathbf{y}] &= \mathbf{A}\boldsymbol{\mu}+\mathbf{b}\\ \mathrm{cov}[\mathbf{y}] &= \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}.\end{aligned}$$

And
$$\begin{aligned}\mathbb{E}[\mathbf{x}|\mathbf{y}] &= (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\left\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\right\}\\ \mathrm{cov}[\mathbf{x}|\mathbf{y}] &= (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}.\end{aligned}$$

3/8/2021

33

Maximum Likelihood for the Gaussian (1)

Given i.i.d. data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$, the log likelihood function is given by

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

Sufficient statistics

$$\sum_{n=1}^{N}\mathbf{x}_n \qquad \sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}$$

3/8/2021

34

Maximum Likelihood for the Gaussian (2)

Set the derivative of the log likelihood function to zero,

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

and solve to obtain

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n.$$

Similarly

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}.$$

3/8/2021

35

Maximum Likelihood for the Gaussian (3)

Under the true distribution

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] &= \boldsymbol{\mu} \\ \mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] &= \frac{N-1}{N}\boldsymbol{\Sigma}.\end{aligned}$$

Hence define

$$\widetilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}.$$

$$\mathbb{E}[\tilde{\Sigma}] = \Sigma$$

3/8/2021

36

Sequential Estimation

Contribution of the N^{th} data point, x_N

$$\begin{aligned}\boldsymbol{\mu}_{\mathrm{ML}}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\ &= \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathbf{x}_n \\ &= \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} \\ &= \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)})\end{aligned}$$

correction given $\mathbf{x}_N$

correction weight

old estimate

3/8/2021

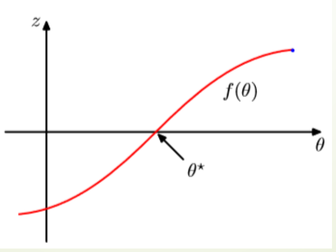
37

The Robbins-Monro Algorithm (1)

Consider θ and z governed by $p(z,\theta)$ and define the **regression function**

$$f(\theta) \equiv \mathbb{E}[z|\theta] = \int zp(z|\theta)\,\mathrm{d}z$$

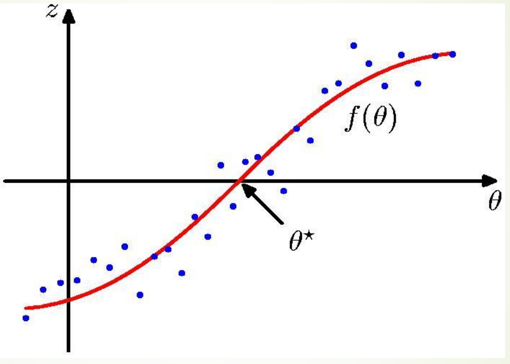
Seek $\theta^\star$ such that $f(\theta^\star) = 0$.



3/8/2021

38

The Robbins-Monro Algorithm (2)



Assume we are given samples from $p(z,\theta)$, one at the time.

3/8/2021

39

The Robbins-Monro Algorithm (3)

- Successive estimates of $\theta^\star$ are then given by

$$\theta^{(N)} = \theta^{(N-1)} - a_{N-1} z(\theta^{(N-1)}).$$

- Conditions on a_N for convergence :

$$\lim_{N\to\infty} a_N = 0 \qquad \sum_{N=1}^{\infty} a_N = \infty \qquad \sum_{N=1}^{\infty} a_N^2 < \infty$$

3/8/2021

40

Robbins-Monro for Maximum Likelihood (1)

Regarding

$$-\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \frac{\partial}{\partial\theta} \ln p(x_n|\theta) = \mathbb{E}_x \left[-\frac{\partial}{\partial\theta} \ln p(x|\theta) \right]$$

as a regression function, finding its root is equivalent to finding the maximum likelihood solution θ_{ML}. Thus

$$\theta^{(N)} = \theta^{(N-1)} - a_{N-1} \frac{\partial}{\partial\theta^{(N-1)}} \left[-\ln p(x_N|\theta^{(N-1)}) \right].$$

3/8/2021

41

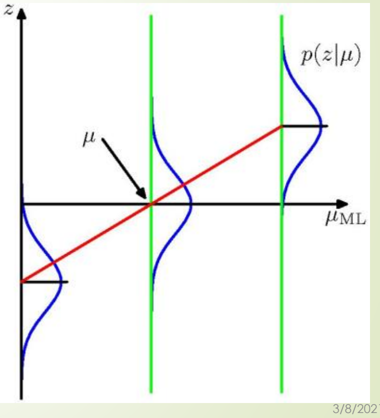
Robbins-Monro for Maximum Likelihood (2)

Example: estimate the mean of a Gaussian.

$$z = \frac{\partial}{\partial \mu_{\mathrm{ML}}}\left[-\ln p(x|\mu_{\mathrm{ML}}, \sigma^2)\right] = -\frac{1}{\sigma^2}(x - \mu_{\mathrm{ML}})$$

The distribution of z is Gaussian with mean $\mu - \mu_{\mathrm{ML}}$.

For the Robbins-Monro update equation, $a_N = \sigma^2/N$.



3/8/2021

42

Bayesian Inference for the Gaussian (1)

Assume σ^2 is known. Given i.i.d. data $\mathbf{x} = \{x_1, \ldots, x_N\}$, the likelihood function for μ is given by

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

This has a Gaussian shape as a function of μ (but it is *not* a distribution over μ).

3/8/2021

43

Bayesian Inference for the Gaussian (2)

Combined with a Gaussian prior over μ,

$$p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right).$$

this gives the posterior

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu).$$

Completing the square over μ, we see that

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

3/8/2021

44

Bayesian Inference for the Gaussian (3)

… where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}}, \qquad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

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

Note:

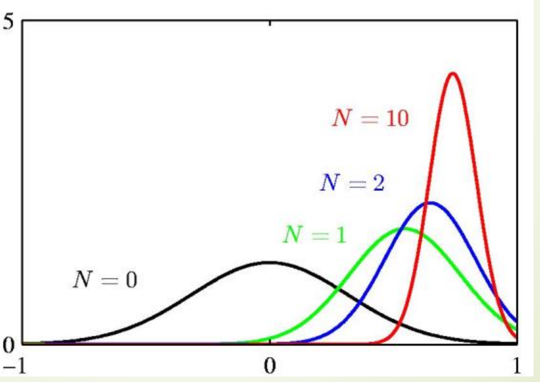
	$N = 0$	$N \to \infty$
μ_N	μ_0	μ_{ML}
σ_N^2	σ_0^2	0

3/8/2021

45

Bayesian Inference for the Gaussian (4)

Example: $p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$ for $N = 0, 1, 2$ and $10.$



3/8/2021

46

Bayesian Inference for the Gaussian (5)

Sequential Estimation

$$
\begin{aligned}
p(\mu|\mathbf{x}) &\propto p(\mu)p(\mathbf{x}|\mu) \\
&= \left[p(\mu)\prod_{n=1}^{N-1} p(x_n|\mu)\right] p(x_N|\mu) \\
&\propto \mathcal{N}\left(\mu|\mu_{N-1}, \sigma_{N-1}^2\right) p(x_N|\mu)
\end{aligned}
$$

The posterior obtained after observing $N-1$ data points becomes the **prior** when we observe the N^{th} data point.

3/8/2021

47

Bayesian Inference for the Gaussian (6)

Now assume μ is known. The likelihood function for $\lambda = 1/\sigma^2$ is given by

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

This has a Gamma shape as a function of λ.

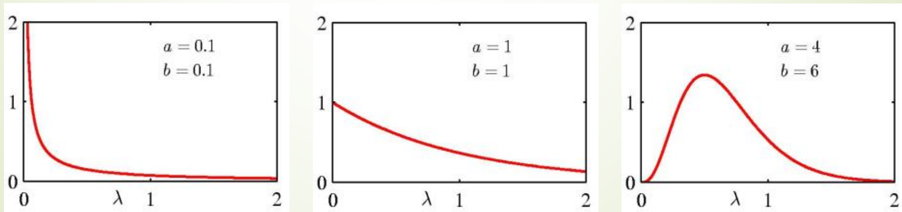
48

Bayesian Inference for the Gaussian (7)

The Gamma distribution

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

$$\mathbb{E}[\lambda] = \frac{a}{b} \qquad \mathrm{var}[\lambda] = \frac{a}{b^2}$$



49

Bayesian Inference for the Gaussian (8)

Now we combine a Gamma prior, $\mathrm{Gam}(\lambda|a_0, b_0)$, with the likelihood function for λ to obtain

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

which we recognize as $\mathrm{Gam}(\lambda|a_N, b_N)$ with

$$a_N = a_0 + \frac{N}{2}$$

$$b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma^2_{\mathrm{ML}}.$$

3/8/2021

50

Bayesian Inference for the Gaussian (9)

If both μ and λ are unknown, the joint likelihood function is given by

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

$$\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}.$$

We need a prior with the same functional dependence on μ and λ.

3/8/2021

51

Bayesian Inference for the Gaussian (10)

The Gaussian-gamma distribution

$$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a, b)$$
$$\propto \quad \exp\left\{-\frac{\beta\lambda}{2}(\mu - \mu_0)^2\right\} \lambda^{a-1} \exp\{-b\lambda\}$$

- Quadratic in μ.
- Linear in λ.

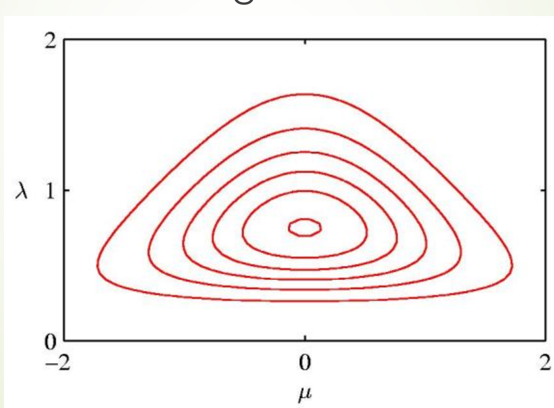
- Gamma distribution over λ.
- Independent of μ.

3/8/2021

52

Bayesian Inference for the Gaussian (11)

The Gaussian-gamma distribution



3/8/2021

53

Bayesian Inference for the Gaussian (12)

Multivariate conjugate priors

$\boldsymbol{\mu}$ unknown, $\boldsymbol{\Lambda}$ known: $p(\boldsymbol{\mu})$ Gaussian.

$\boldsymbol{\Lambda}$ unknown, $\boldsymbol{\mu}$ known: $p(\boldsymbol{\Lambda})$ Wishart,

$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W},\nu) = B|\boldsymbol{\Lambda}|^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right).$$

$\boldsymbol{\Lambda}$ and $\boldsymbol{\mu}$ unknown: $p(\boldsymbol{\mu},\boldsymbol{\Lambda})$ Gaussian-Wishart,

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{\mu}_0,\beta,\mathbf{W},\nu) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0,(\beta\boldsymbol{\Lambda})^{-1})\,\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W},\nu)$$

3/8/2021

54

Student's t-Distribution (1)

$$\begin{aligned} p(x|\mu,a,b) &= \int_0^\infty \mathcal{N}(x|\mu,\tau^{-1})\mathrm{Gam}(\tau|a,b)\,\mathrm{d}\tau \\ &= \int_0^\infty \mathcal{N}\left(x|\mu,(\eta\lambda)^{-1}\right)\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta \\ &= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2} \\ &= \mathrm{St}(x|\mu,\lambda,\nu) \end{aligned}$$

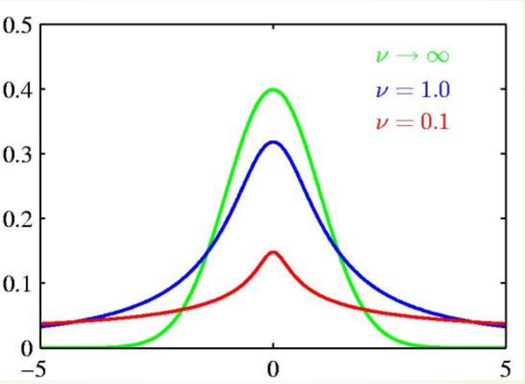
where $\lambda = a/b \qquad \eta = \tau b/a \qquad \nu = 2a.$

Infinite mixture of Gaussians.

3/8/2021

55

Student's t-Distribution (2)



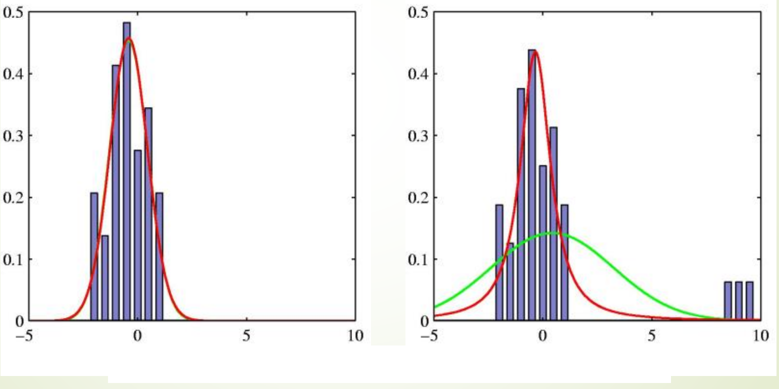
	$\nu = 1$	$\nu \to \infty$
$\mathrm{St}(x\|\mu, \lambda, \nu)$	Cauchy	$\mathcal{N}(x\|\mu, \lambda^{-1})$

3/8/2021

56

Student's t-Distribution (3)

Robustness to outliers: Gaussian vs t-distribution.



3/8/2021

57

Student’s t-Distribution (4)

The D-variate case:

$$\begin{aligned}\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) &= \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta \\ &= \frac{\Gamma(D/2+\nu/2)}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}}\left[1+\frac{\Delta^2}{\nu}\right]^{-D/2-\nu/2}\end{aligned}$$

where $\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})$.

Properties:

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \boldsymbol{\mu}, && \text{if } \nu > 1 \\ \mathrm{cov}[\mathbf{x}] &= \frac{\nu}{(\nu-2)}\boldsymbol{\Lambda}^{-1}, && \text{if } \nu > 2 \\ \mathrm{mode}[\mathbf{x}] &= \boldsymbol{\mu}\end{aligned}$$

3/8/2021

58

Periodic variables

Examples: calendar time, direction, …

We require

$$\begin{aligned}p(\theta) &\geqslant 0 \\ \int_0^{2\pi} p(\theta)\,\mathrm{d}\theta &= 1 \\ p(\theta+2\pi) &= p(\theta).\end{aligned}$$

3/8/2021

59

von Mises Distribution (1)

This requirement is satisfied by

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\{m\cos(\theta - \theta_0)\}$$

where

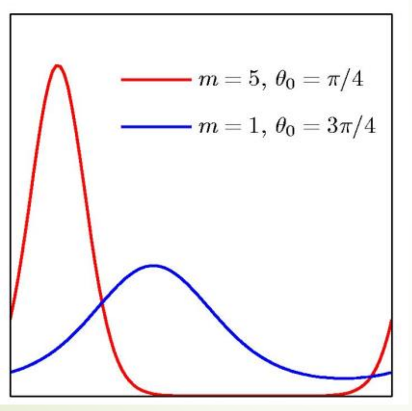
$$I_0(m) = \frac{1}{2\pi}\int_0^{2\pi} \exp\{m\cos\theta\}\,\mathrm{d}\theta$$

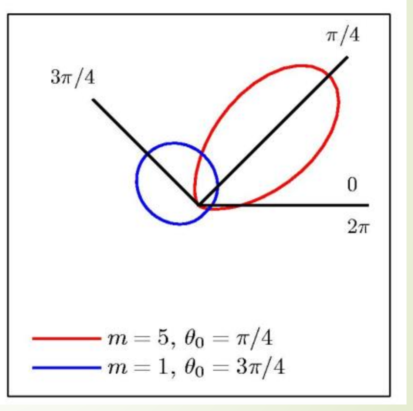
is the 0^{th} order modified Bessel function of the 1^{st} kind.

3/8/2021

60

von Mises Distribution (2)





3/8/2021

61

Maximum Likelihood for von Mises

Given a data set, $\mathcal{D} = \{\theta_1, \ldots, \theta_N\}$, the log likelihood function is given by

$$\ln p(\mathcal{D}|\theta_0, m) = -N \ln(2\pi) - N \ln I_0(m) + m \sum_{n=1}^{N} \cos(\theta_n - \theta_0).$$

Maximizing with respect to θ_0 we directly obtain

$$\theta_0^{\mathrm{ML}} = \tan^{-1} \left\{ \frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n} \right\}.$$

Similarly, maximizing with respect to m we get

$$\frac{I_1(m_{\mathrm{ML}})}{I_0(m_{\mathrm{ML}})} = \frac{1}{N} \sum_{n=1}^{N} \cos(\theta_n - \theta_0^{\mathrm{ML}})$$

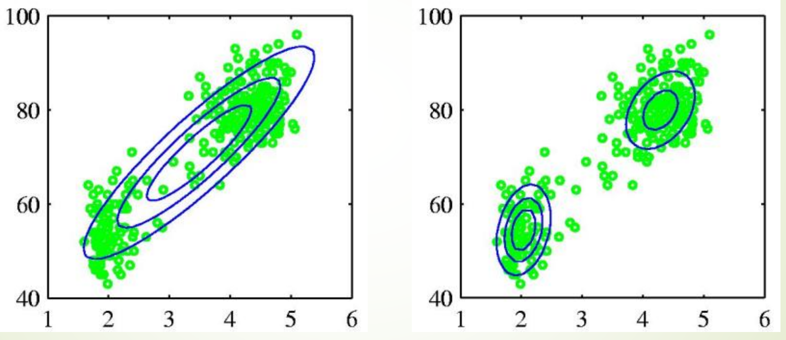
which can be solved numerically for m_{ML}.

3/8/2021

62

Mixtures of Gaussians (1)

Old Faithful data set



Single Gaussian   Mixture of two Gaussians

https://www.kaggle.com/janithwanni/old-faithful

3/8/2021

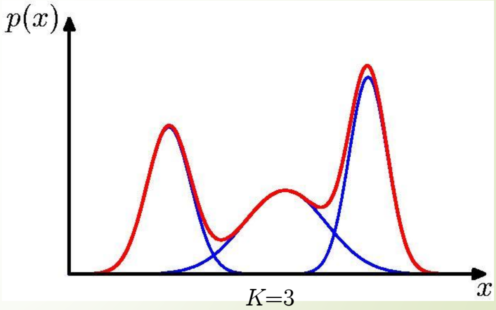
63

Mixtures of Gaussians (2)

Combine simple models into a complex model:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \underbrace{\pi_k}_{\text{Mixing coefficient}} \underbrace{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{Component}}$$

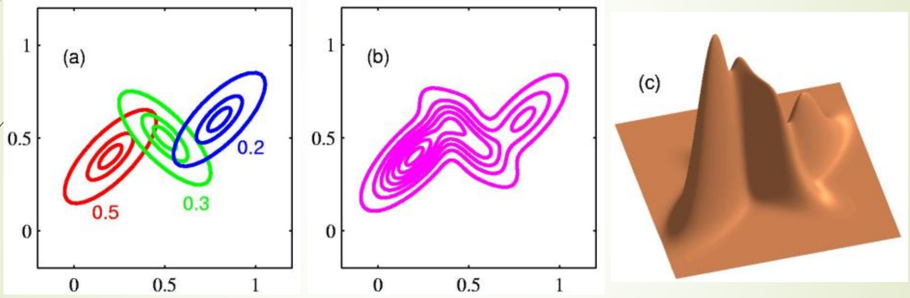
$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$



3/8/2021

64

Mixtures of Gaussians (3)



3/8/2021

65

Mixtures of Gaussians (4)

Determining parameters $\boldsymbol{\mu}$, Σ, and $\boldsymbol{\pi}$ using maximum log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \underbrace{\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}}_{\text{Log of a sum; no closed-form maximum.}}$$

Solution: use standard, iterative, numeric optimization methods or the *expectation maximization* algorithm (Chapter 9).

66

The Exponential Family (1)

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

where η is the *natural parameter* and

$$g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}\,\mathrm{d}\mathbf{x} = 1$$

so $g(\eta)$ can be interpreted as a normalization coefficient.

67

The Exponential Family (2.1)

The Bernoulli Distribution

$$\begin{aligned} p(x|\mu) &= \mathrm{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \\ &= \exp\left\{x\ln\mu + (1-x)\ln(1-\mu)\right\} \\ &= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\} \end{aligned}$$

Comparing with the general form we see that

$\eta = \ln\left(\frac{\mu}{1-\mu}\right)$ and so $\mu = \underbrace{\sigma(\eta) = \frac{1}{1+\exp(-\eta)}}_{\text{Logistic sigmoid}}.$

3/8/2021

68

The Exponential Family (2.2)

The Bernoulli distribution can hence be written as

$$p(x|\eta) = \sigma(-\eta)\exp(\eta x)$$

where

$$\begin{aligned} u(x) &= x \\ h(x) &= 1 \\ g(\eta) &= 1-\sigma(\eta) = \sigma(-\eta). \end{aligned}$$

3/8/2021

69

The Exponential Family (3.1)

The Multinomial Distribution

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right)$$

where, $\mathbf{x} = (x_1, \ldots, x_M)^{\mathrm{T}}$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_M)^{\mathrm{T}}$ and

$$\begin{aligned} \eta_k &= \ln \mu_k \\ \mathbf{u}(\mathbf{x}) &= \mathbf{x} \\ h(\mathbf{x}) &= 1 \\ g(\boldsymbol{\eta}) &= 1. \end{aligned}$$

NOTE: The μ_k parameters are not independent since the corresponding μ_k must satisfy

$$\sum_{k=1}^{M} \mu_k = 1.$$

3/8/2021

70

The Exponential Family (3.2)

Let $\mu_M = 1 - \sum_{k=1}^{M-1} \mu_k$. This leads to

$$\eta_k = \ln\left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}\right) \quad \text{and} \quad \mu_k = \underbrace{\frac{\exp(\eta_k)}{1 + \sum_{j=1}^{M-1} \exp(\eta_j)}}_{\text{Softmax}}.$$

Here the η_k parameters are independent. Note that

$$0 \leqslant \mu_k \leqslant 1 \quad \text{and} \quad \sum_{k=1}^{M-1} \mu_k \leqslant 1.$$

3/8/2021

71

The Exponential Family (3.3)

The Multinomial distribution can then be written as

$$p(\mathbf{x}|\boldsymbol{\mu}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right)$$

where

$$\begin{aligned}
\boldsymbol{\eta} &= (\eta_1, \ldots, \eta_{M-1}, 0)^{\mathrm{T}} \\
\mathbf{u}(\mathbf{x}) &= \mathbf{x} \\
h(\mathbf{x}) &= 1 \\
g(\boldsymbol{\eta}) &= \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k)\right)^{-1}.
\end{aligned}$$

3/8/2021

72

The Exponential Family (4)

The Gaussian Distribution

$$\begin{aligned}
p(x|\mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right\} \\
&= h(x)g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(x)\right\}
\end{aligned}$$

where

$$\boldsymbol{\eta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \qquad h(\mathbf{x}) = (2\pi)^{-1/2}$$

$$\mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \qquad g(\boldsymbol{\eta}) = (-2\eta_2)^{1/2}\exp\left(\frac{\eta_1^2}{4\eta_2}\right).$$

3/8/2021

73

ML for the Exponential Family (1)

From the definition of $g(\boldsymbol{\eta})$ we get

$$\nabla g(\boldsymbol{\eta}) \underbrace{\int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathrm{d}\mathbf{x}}_{1/g(\boldsymbol{\eta})} + \underbrace{g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x})\, \mathrm{d}\mathbf{x}}_{\mathbb{E}[\mathbf{u}(\mathbf{x})]} = 0$$

Thus

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

$$\boxed{g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathrm{d}\mathbf{x} = 1}$$

3/8/2021

74

ML for the Exponential Family (2)

Give a data set, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, the likelihood function is given by

$$p(\mathbf{X}|\boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^N \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)\right\}.$$

Thus we have

$$-\nabla \ln g(\boldsymbol{\eta}_{\mathrm{ML}}) = \frac{1}{N} \underbrace{\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)}_{\text{Sufficient statistic}}$$

3/8/2021

75

Conjugate priors

For any member of the exponential family, there exists a prior

$$p(\boldsymbol{\eta}|\boldsymbol{\chi},\nu) = f(\boldsymbol{\chi},\nu)g(\boldsymbol{\eta})^{\nu}\exp\left\{\nu\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}\right\}.$$

Combining with the likelihood function, we get

$$p(\boldsymbol{\eta}|\mathbf{X},\boldsymbol{\chi},\nu) \propto g(\boldsymbol{\eta})^{\nu+N}\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\left(\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)+\nu\boldsymbol{\chi}\right)\right\}.$$

Prior corresponds to ν pseudo-observations with value χ.

3/8/2021

76

Noninformative Priors (1)

With little or no information available a-priori, we might choose a non-informative prior.

- λ discrete, K-nomial : $p(\lambda) = 1/K$.
- $\lambda \in [a,b]$ real and bounded: $p(\lambda) = 1/b-a$.
- λ real and unbounded: improper!

A constant prior may no longer be constant after a change of variable; consider $p(\lambda)$ constant and $\lambda = \eta^2$:

$$p_{\eta}(\eta) = p_{\lambda}(\lambda)\left|\frac{\mathrm{d}\lambda}{\mathrm{d}\eta}\right| = p_{\lambda}(\eta^2)2\eta \propto \eta$$

3/8/2021

77

Noninformative Priors (2)

Translation invariant priors. Consider

$$p(x|\mu) = f(x-\mu) = f((x+c)-(\mu+c)) = f(\widehat{x}-\widehat{\mu}) = p(\widehat{x}|\widehat{\mu}).$$

For a corresponding prior over μ, we have

$$\int_A^B p(\mu)\,\mathrm{d}\mu = \int_{A-c}^{B-c} p(\mu)\,\mathrm{d}\mu = \int_A^B p(\mu-c)\,\mathrm{d}\mu$$

for any A and B. Thus $p(\mu) = p(\mu - c)$ and $p(\mu)$ must be constant.

3/8/2021

78

Noninformative Priors (3)

Example: The mean of a Gaussian, μ; the conjugate prior is also a Gaussian,

$$p(\mu|\mu_0, \sigma_0^2) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

As $\sigma_0^2 \rightarrow \infty$, this will become constant over μ.

3/8/2021

79

Noninformative Priors (4)

Scale invariant priors. Consider $p(x|\sigma) = (1/\sigma)f(x/\sigma)$ and make the change of variable $\widehat{x} = cx$

$$p_{\widehat{x}}(\widehat{x}) = p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}\widehat{x}}\right| = p_x\left(\frac{\widehat{x}}{c}\right)\frac{1}{c} = \frac{1}{c\sigma}f\left(\frac{\widehat{x}}{c\sigma}\right) = p_x(\widehat{x}|\widehat{\sigma}).$$

For a corresponding prior over σ, we have

$$\int_A^B p(\sigma)\,\mathrm{d}\sigma = \int_{A/c}^{B/c} p(\sigma)\,\mathrm{d}\sigma = \int_A^B p\left(\frac{1}{c}\sigma\right)\frac{1}{c}\,\mathrm{d}\sigma$$

for any A and B. Thus $p(\sigma) \propto 1/\sigma$ and so this prior is improper too. Note that this corresponds to $p(\ln\sigma)$ being constant.

3/8/2021

80

Noninformative Priors (5)

Example: For the variance of a Gaussian, σ^2, we have

$$\mathcal{N}(x|\mu,\sigma^2) \propto \sigma^{-1}\exp\left\{-((x-\mu)/\sigma)^2\right\}.$$

If $\lambda = 1/\sigma^2$ and $p(\sigma) \propto 1/\sigma$, then $p(\lambda) \propto 1/\lambda$.

We know that the conjugate distribution for λ is the Gamma distribution,

$$\mathrm{Gam}(\lambda|a_0,b_0) \propto \lambda^{a_0-1}\exp(-b_0\lambda).$$

A noninformative prior is obtained when $a_0 = 0$ and $b_0 = 0$.

3/8/2021

81

Nonparametric Methods (1)

- Parametric distribution models are restricted to specific forms, which may not always be suitable; for example, consider modelling a multimodal distribution with a single, unimodal model.
- Nonparametric approaches make few assumptions about the overall shape of the distribution being modelled.

3/8/2021

82

Nonparametric Methods (2)

Two types of non-parametric methods

- Estimate density function $p(\mathbf{x}|C_k)$ from sample patterns (instance/memory-based learning
- Directly estimate the a posteriori probability $P(C_k|\mathbf{x})$ — similar to the nearest-neighbor rule, which bypass probability estimation and go directly to decision functions.

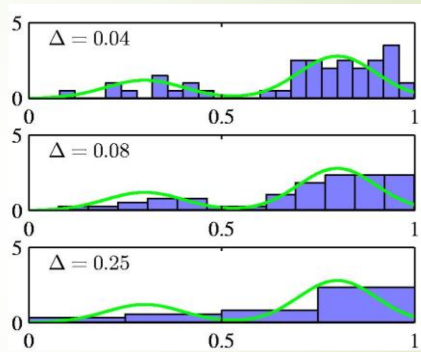
3/8/2021

83

Kernel Density Estimation (1)

Histogram methods partition the data space into distinct bins with widths Δ_i and count the number of observations, n_i, in each bin.

$$p_i = \frac{n_i}{N\Delta_i}$$



Often, the same width is used for all bins, $\Delta_i = \Delta$.

Δ acts as a smoothing parameter.

In a D-dimensional space, using M bins in each dimension will require M^D bins!

3/8/2021

84

Kernel Density Estimation (2)

Assume observations drawn from a density $p(\mathbf{x})$ and consider a small region $\mathcal{R}$ containing $\mathbf{x}$ such that

$$P = \int_{\mathcal{R}} p(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

The probability that K out of N observations lie inside $\mathcal{R}$ is $\mathrm{Bin}(K|N,P)$ and if N is large

$$K \simeq NP.$$

If the volume of $\mathcal{R}$, V, is sufficiently small, $p(\mathbf{x})$ is approximately constant over $\mathcal{R}$ and

$$P \simeq p(\mathbf{x})V$$

Thus

$$\boxed{p(\mathbf{x}) = \frac{K}{NV}.}$$

V small, yet K>0, therefore N large?

3/8/2021

85

Parzen Window(1)

Kernel Density Estimation: fix V, estimate K from the data. Let $\mathcal{R}$ be a hypercube centred on $\mathbf{x}$ and define the kernel function (Parzen window)

$$k((\mathbf{x}-\mathbf{x}_n)/h) = \begin{cases} 1, & |(x_i - x_{ni})/h| \leqslant 1/2, \qquad i = 1, \ldots, D, \\ 0, & \text{otherwise.} \end{cases}$$

It follows that

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right)$$ and hence $$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right).$$

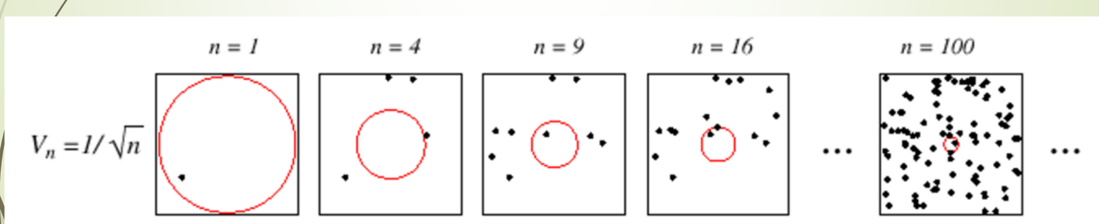
3/8/2021

86

Parzen Window (2)

$$k((\mathbf{x}-\mathbf{x}_n)/h) = \begin{cases} 1, & |(x_i - x_{ni})/h| \leqslant 1/2, \qquad i = 1, \ldots, D, \\ 0, & \text{otherwise.} \end{cases}$$

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right) \qquad p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right).$$



3/8/2021

87

Parzen Window (3)

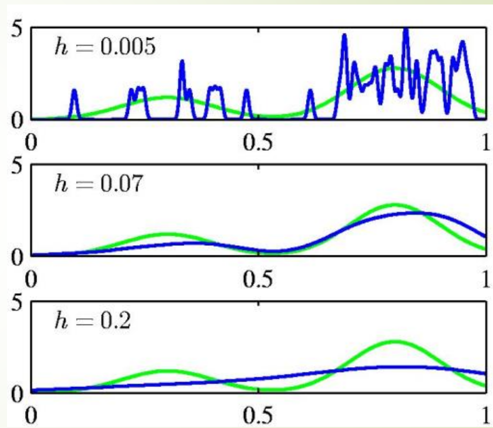
To avoid discontinuities in $p(x)$, use a smooth kernel, e.g. a Gaussian

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{D/2}} \exp\left\{-\frac{\|\mathbf{x}-\mathbf{x}_n\|^2}{2h^2}\right\}$$

Any kernel such that

$$\begin{aligned} k(\mathbf{u}) &\geqslant 0, \\ \int k(\mathbf{u})\,\mathrm{d}\mathbf{u} &= 1 \end{aligned}$$

will work.

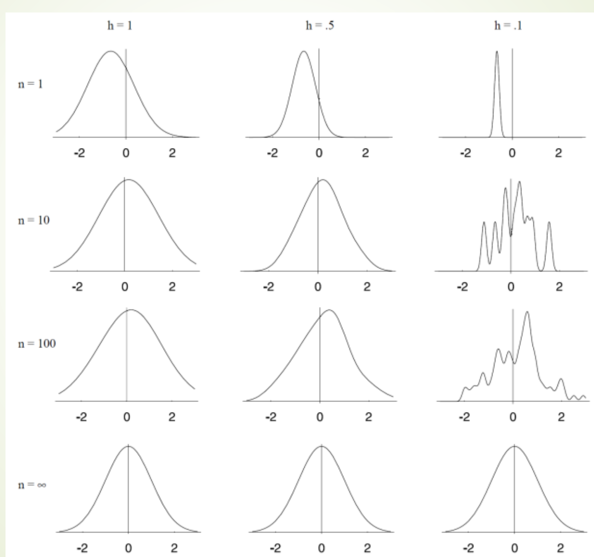


h acts as a smoother.

3/8/2021

88

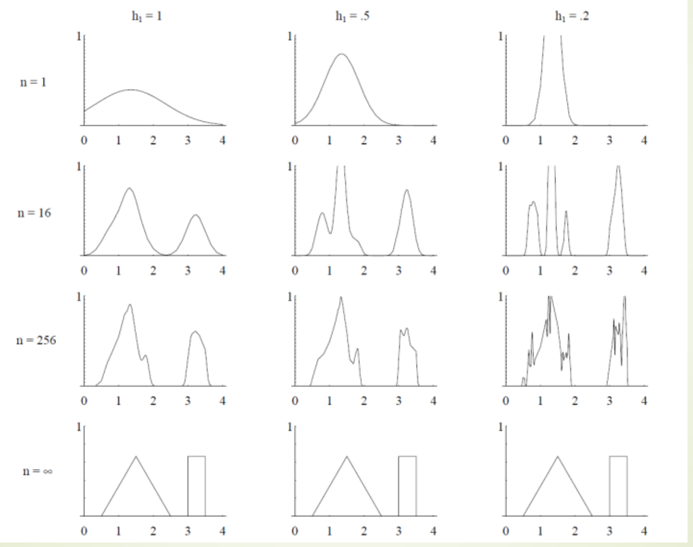
Parzen Window (4)



3/8/2021

89

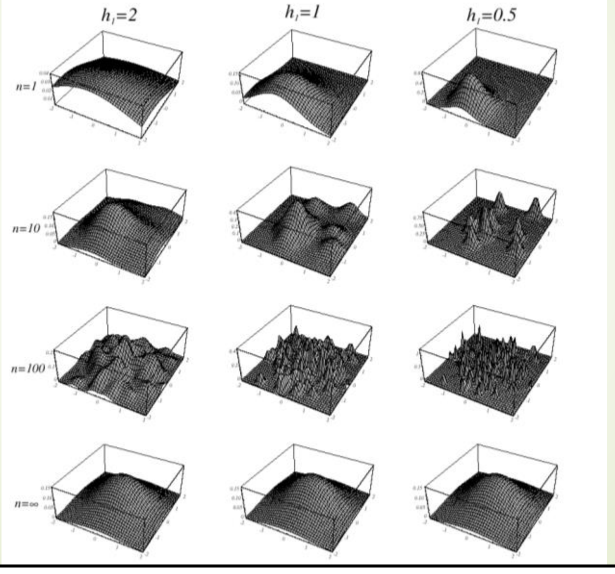
Parzen Window (5)



3/8/2021

90

Parzen Window (6)



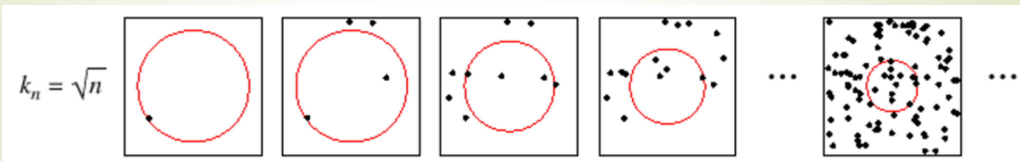
3/8/2021

91

Nearest Neighbour Estimation (1)

Nearest Neighbour Density Estimation: fix K, estimate V from the data. Consider a hypersphere centred on $\mathbf{x}$ and let it grow to a volume, $V^\star$, that includes K of the given N data points. Then

$$p(\mathbf{x}) \simeq \frac{K}{NV^\star}.$$



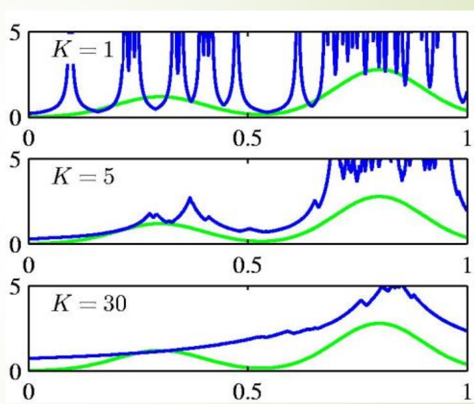
3/8/2021

92

K-Nearest Neighbour Estimation (2)

Nearest Neighbour Density Estimation: fix K, estimate V from the data. Consider a hypersphere centred on $\mathbf{x}$ and let it grow to a volume, $V^\star$, that includes K of the given N data points. Then

$$p(\mathbf{x}) \simeq \frac{K}{NV^\star}.$$

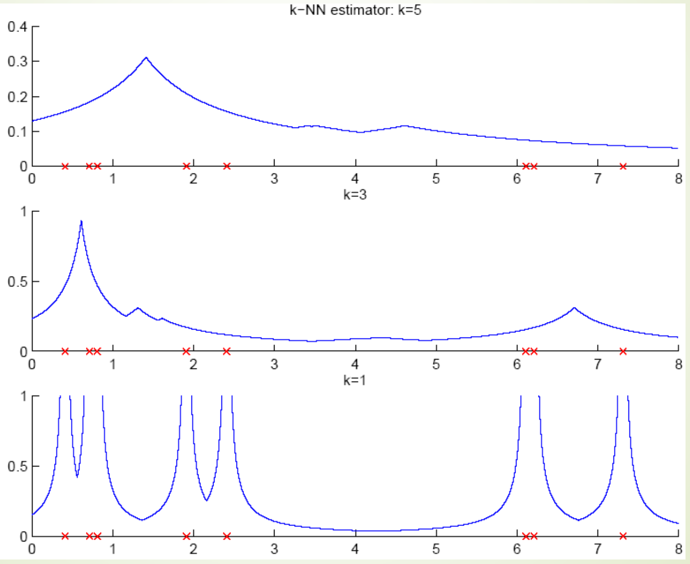


K acts as a smoother.

3/8/2021

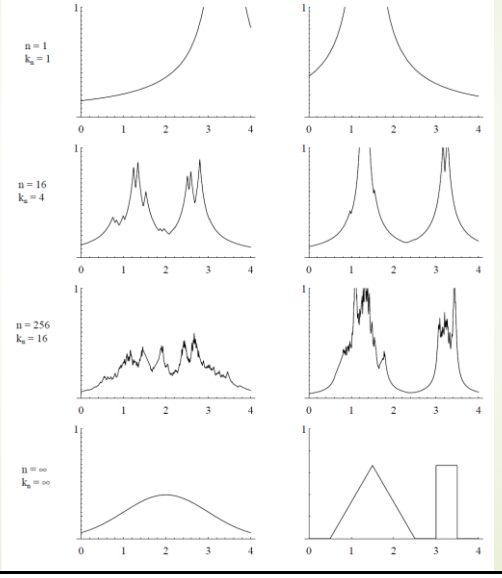
93

K-Nearest Neighbour Estimation (3)



94

K-Nearest Neighbour Estimation (4)



95

Nonparametric vs. Parametric Methods

- Nonparametric models (not histograms) requires storing and computing with the entire data set.
- Parametric models, once fitted, are much more efficient in terms of storage and computation.

96

K-NN for Classification (1)

Given a data set with N_k data points from class $\mathcal{C}_k$ and $\sum_k N_k = N$, we have

$$p(\mathbf{x}) = \frac{K}{NV}$$

and correspondingly

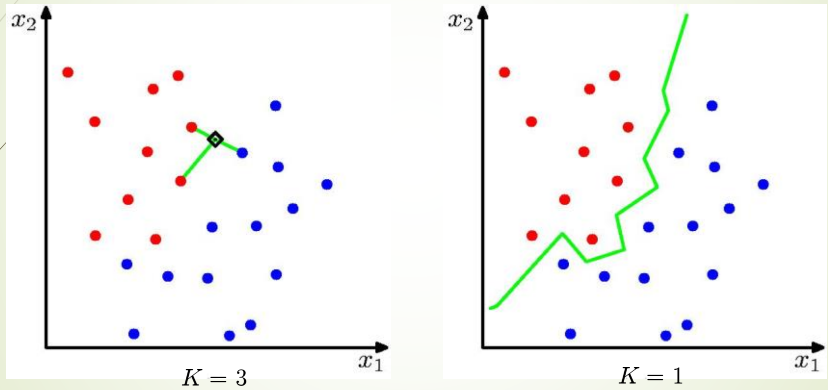
$$p(\mathbf{x}|\mathcal{C}_k) = \frac{K_k}{N_k V}.$$

Since $p(\mathcal{C}_k) = N_k/N$, Bayes' theorem gives

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} = \frac{K_k}{K}.$$

97

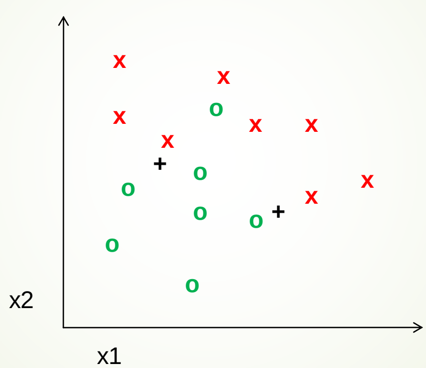
K-NN for Classification (2)



3/8/2021

98

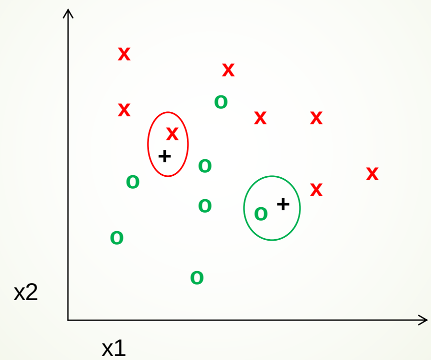
K-NN Classifier



3/8/2021

99

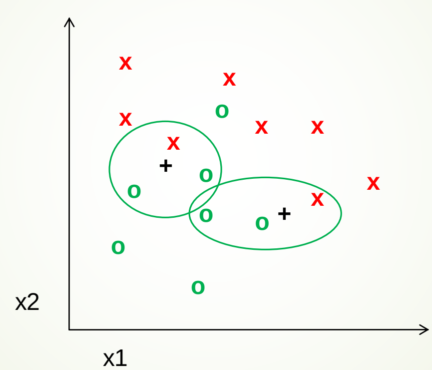
1-NN Classifier



3/8/2021

100

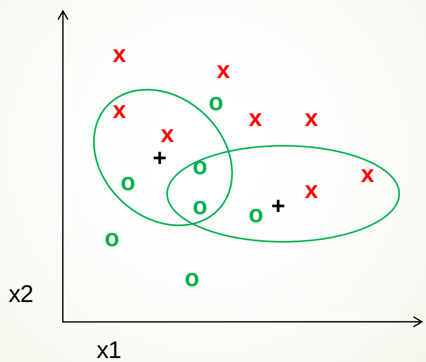
3-NN Classifier



3/8/2021

101

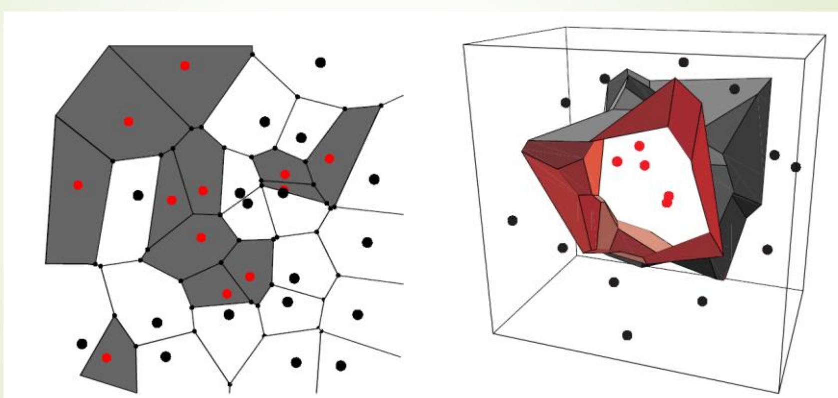
5-NN Classifier



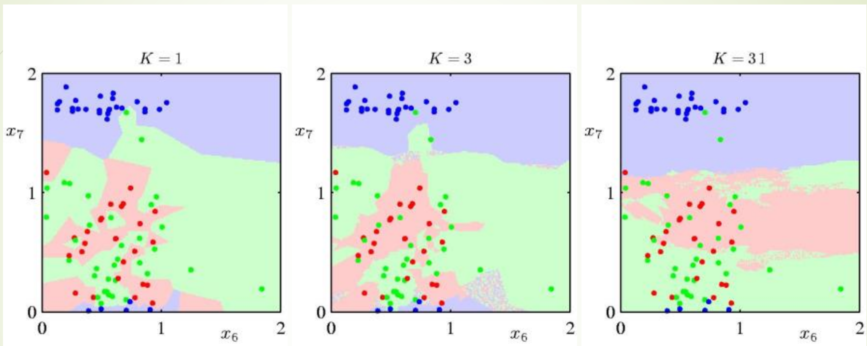
102

1-NN Classifier (3)

Voronoi diagram (cells)



K-NN Classifier



- K acts as a smother
- For $N \to \infty$, the error rate of the 1-nearest-neighbour classifier is never more than twice the optimal error (obtained from the true conditional class distributions).