Piazza started

Announcements

Matlab Grader homework, email Friday,
2 (of 9) homeworks Due 21 April, Binary graded.

Jupyter homework?: translate matlab to Jupiter, TA Harshul h6gupta@eng.ucsd.edu or me
I would like this to happen.

"GPU" homework. NOAA climate data in Jupyter on the datahub.ucsd.edu, 15 April.

Projects: Any language

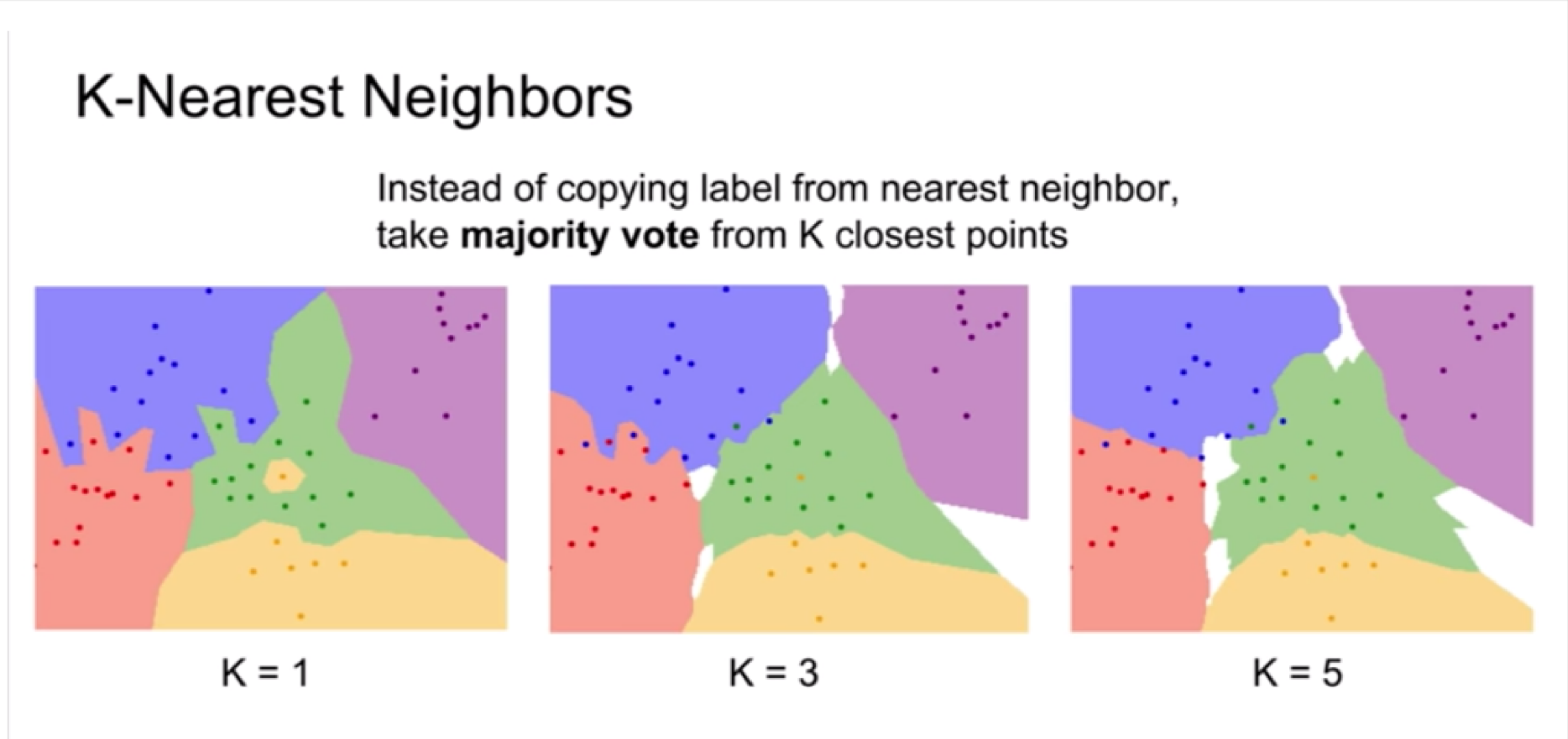
Podcast might work eventually.

Today:

- Stanford CNN
- Bernoulli
- Gaussian 1.2
- Gaussian 2.3
- Decision theory 1.5
- Information theory 1.6

Monday
Stanford CNN, Linear models for regression 3

Non-parametric method



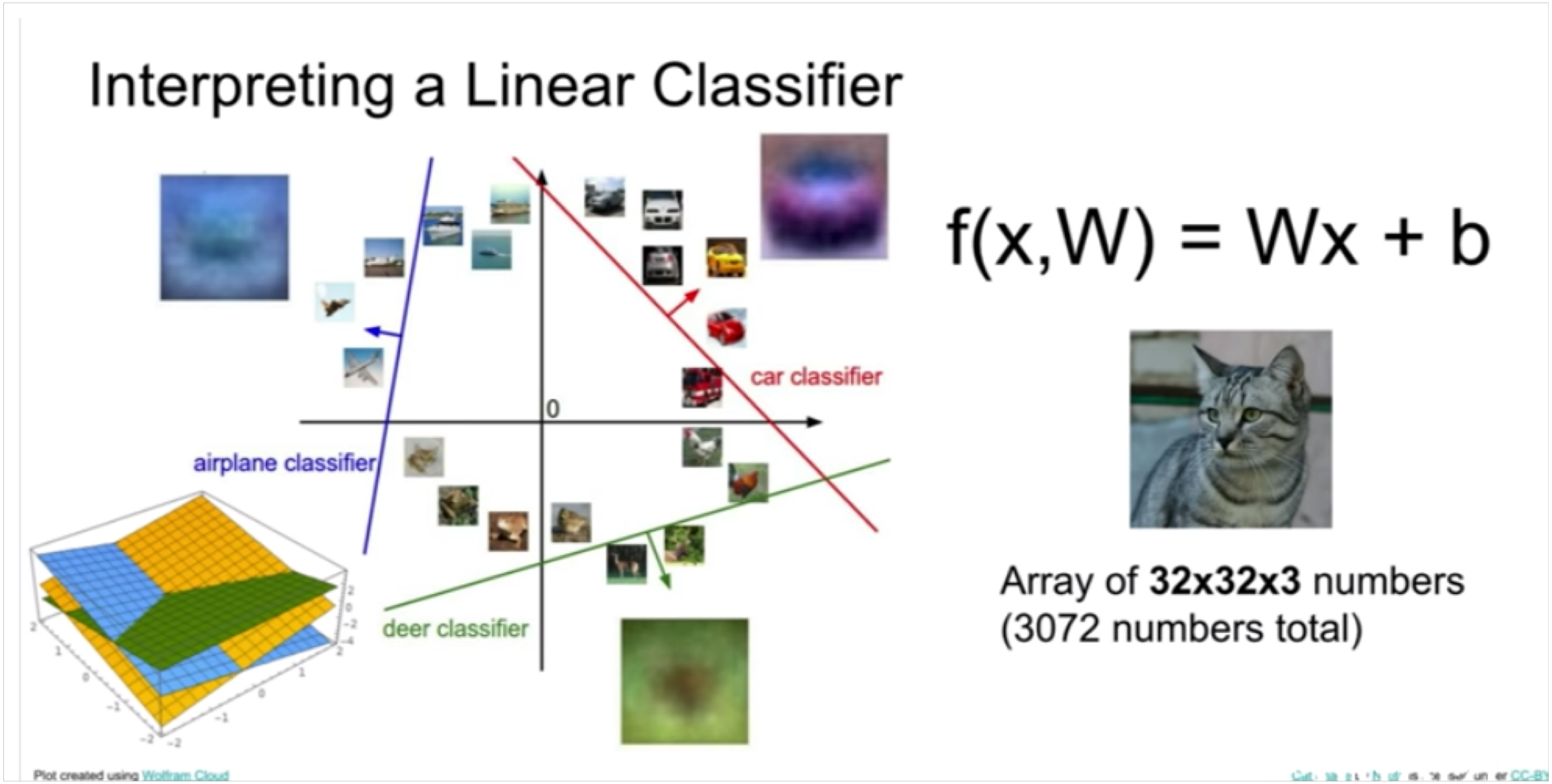
K-Nearest Neighbors

Instead of copying label from nearest neighbor,
take **majority vote** from K closest points

K = 1

K = 3

K = 5



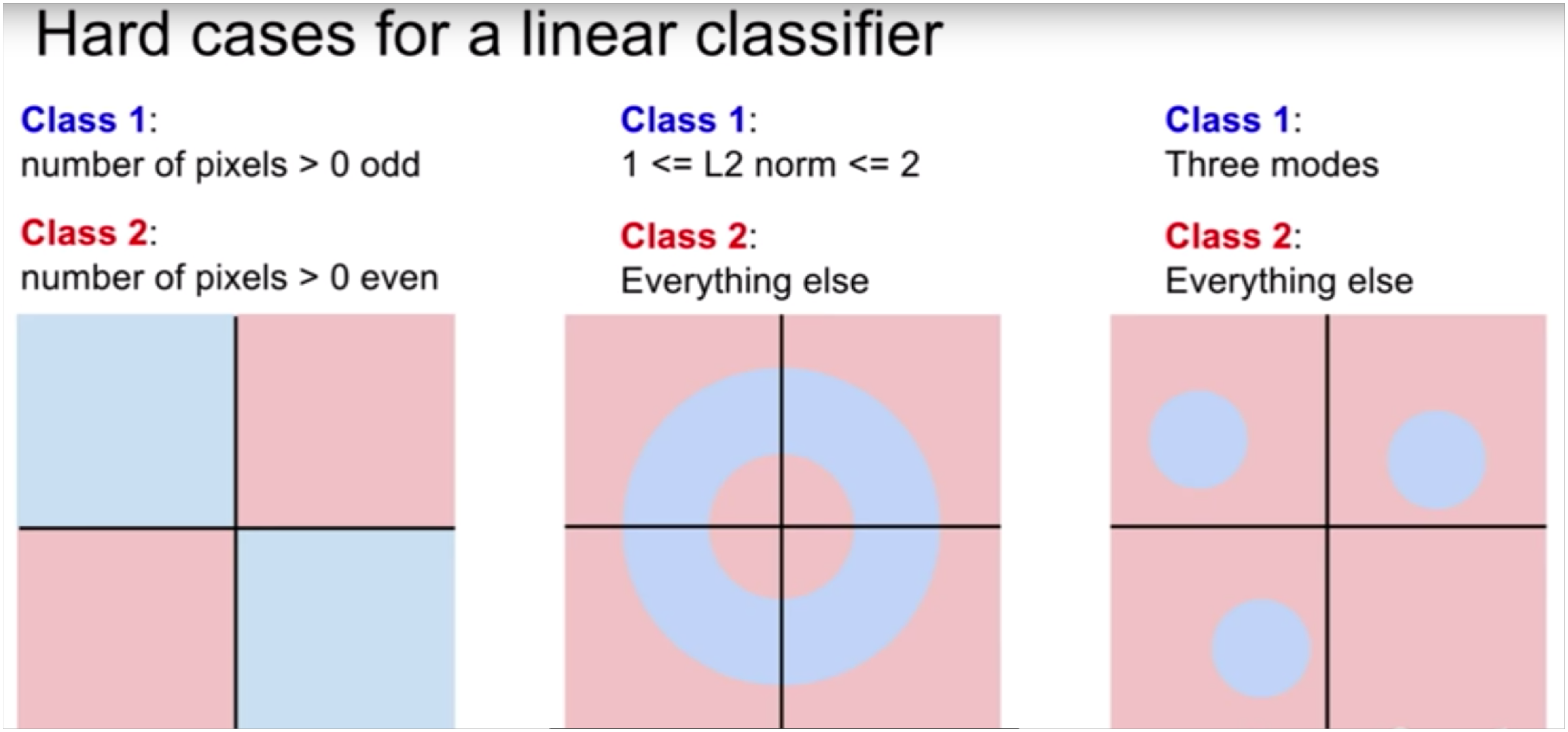
Interpreting a Linear Classifier

$$f(x,W) = Wx + b$$

Array of **32x32x3** numbers
(3072 numbers total)

Plot created using Wolfram Cloud

Cat image by Nikita is licensed under CC-BY



Hard cases for a linear classifier

Class 1:
number of pixels > 0 odd

Class 2:
number of pixels > 0 even

Class 1:
1 <= L2 norm <= 2

Class 2:
Everything else

Class 1:
Three modes

Class 2:
Everything else

Coin estimate (Bishop 2.1)

- Binary variables x={0,1}

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

- Bernoulli distributed

$$\mathrm{Bern}(x|\mu)=\mu^x(1-\mu)^{1-x} \qquad (2.2)$$

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

- N observations, Likelihood:

$$p(\mathcal{D}|\mu)=\prod_{n=1}^{N}p(x_n|\mu)=\prod_{n=1}^{N}\mu^{x_n}(1-\mu)^{1-x_n}. \qquad (2.5)$$

$$\ln p(\mathcal{D}|\mu)=\sum_{n=1}^{N}\ln p(x_n|\mu)=\sum_{n=1}^{N}\left\{x_n\ln\mu+(1-x_n)\ln(1-\mu)\right\}. \qquad (2.6)$$

- Max likelihood

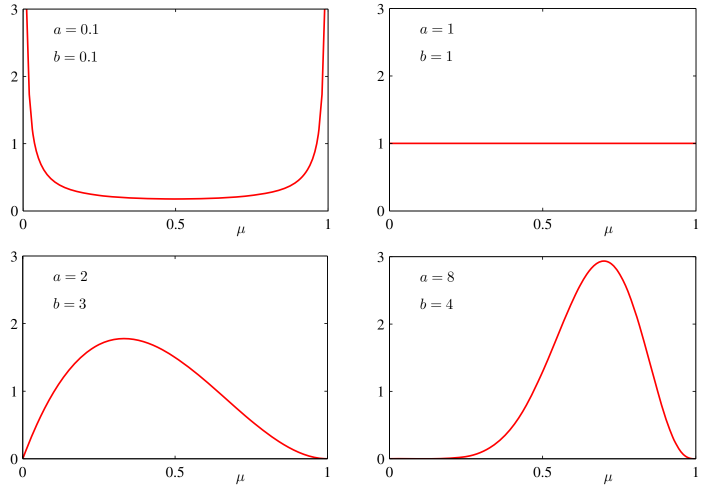
$$\mu_{\mathrm{ML}}=\frac{1}{N}\sum_{n=1}^{N}x_n$$

Coin estimate (Bishop 2.1)

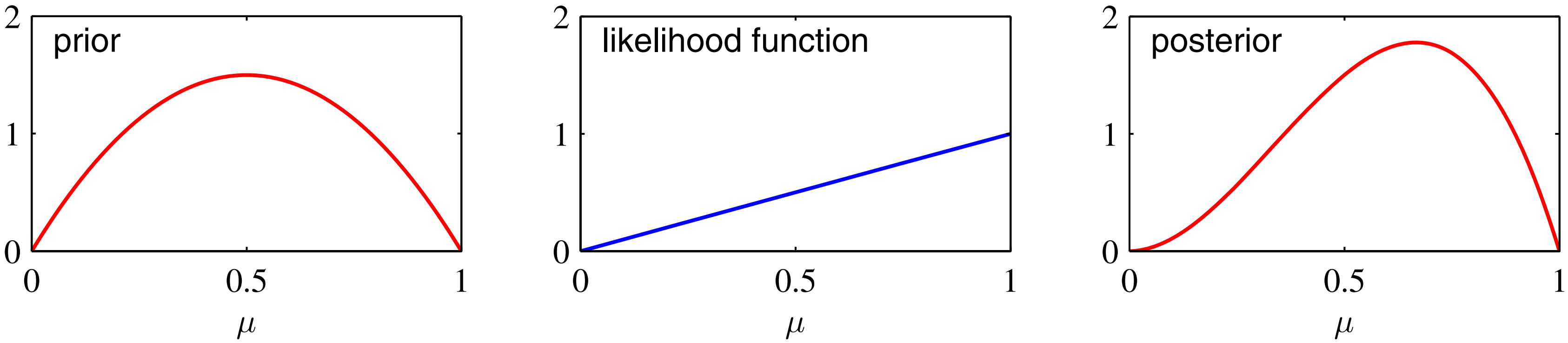
- Bayes p(x|y)=p(y|x)p(x)

- Conjugate prior

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$



Bayes:

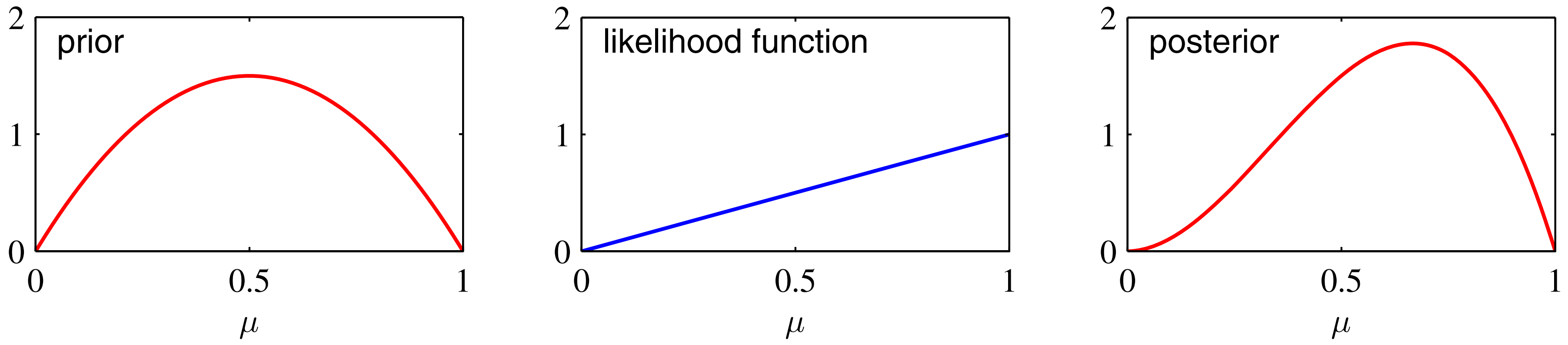


ML MAP BAYES

- ML point estimate

- MAP point estimate (often in literature ML=MAP)

- Bayes => probability =>From which all information can be obtained
  - MAP, median, error estimates
  - Further analysis as sequential
  - Disadvantage… not a point estimate.



Bayes Rule

$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})}$$

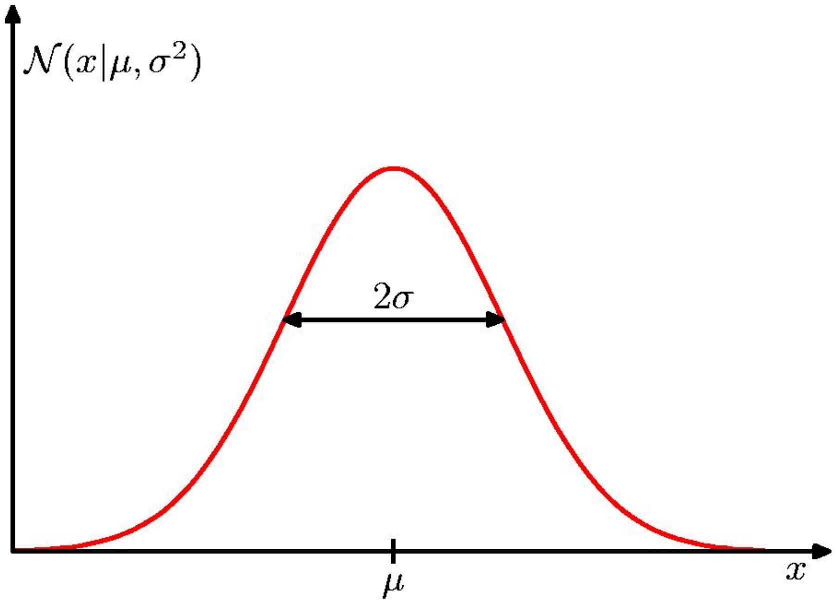
Rev'd Thomas Bayes (1702–1761)

- Bayes rule tells us how to do inference about hypotheses from data.

- Learning and prediction can be seen as forms of inference.

The Gaussian Distribution

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



$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

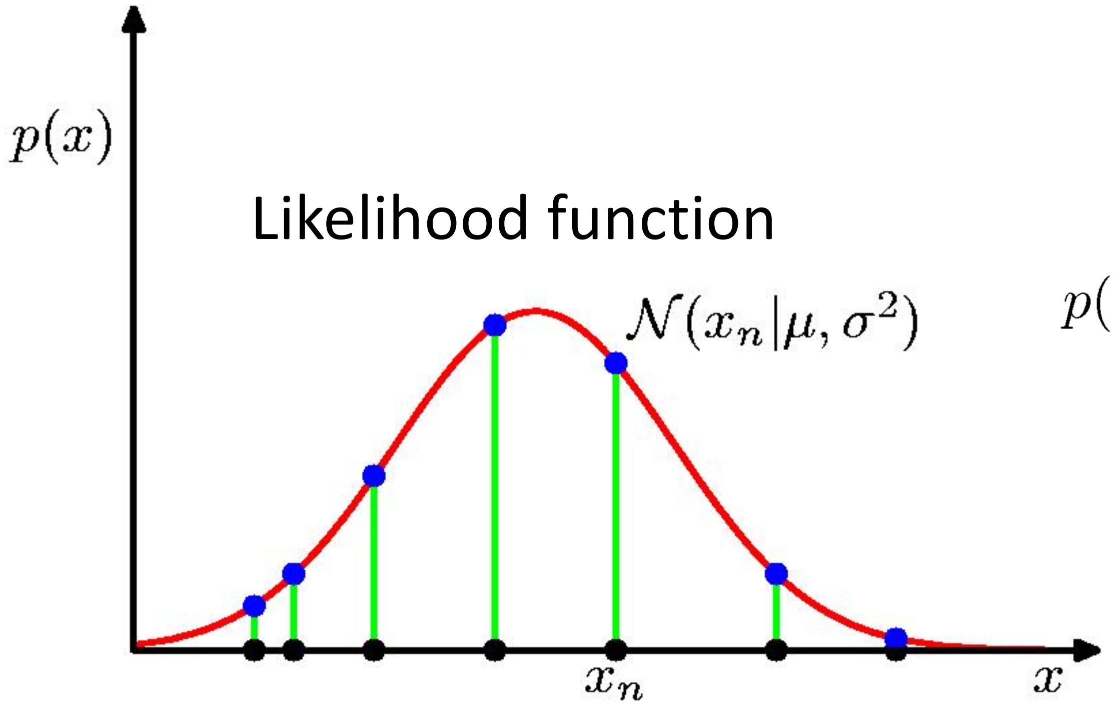
Gaussian Mean and Variance

$$\mathbb{E}[x] = \int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)x\,\mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)x^2\,\mathrm{d}x = \mu^2+\sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

Gaussian Parameter Estimation



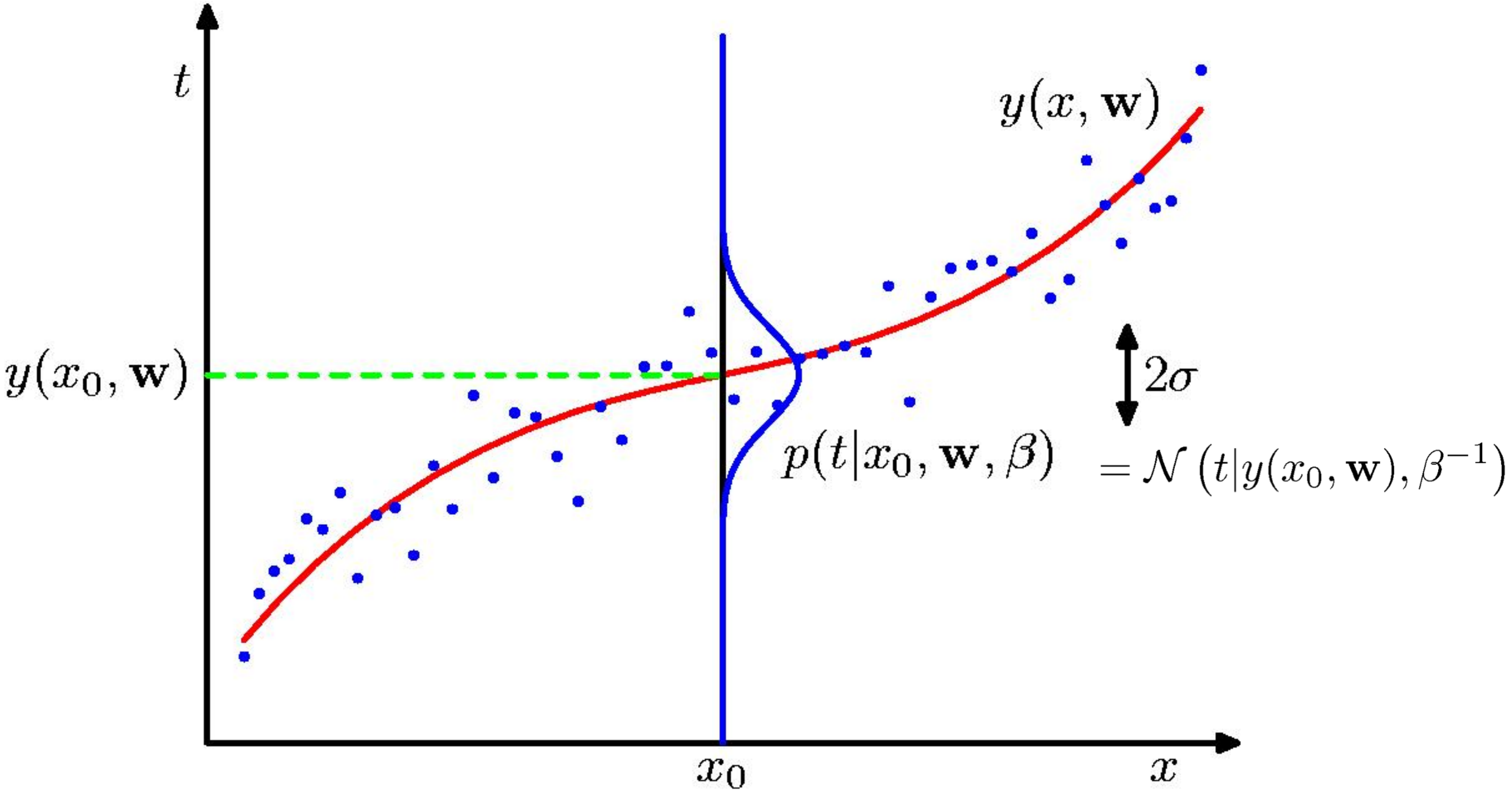
$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu, \sigma^2\right)$$

Maximum (Log) Likelihood

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

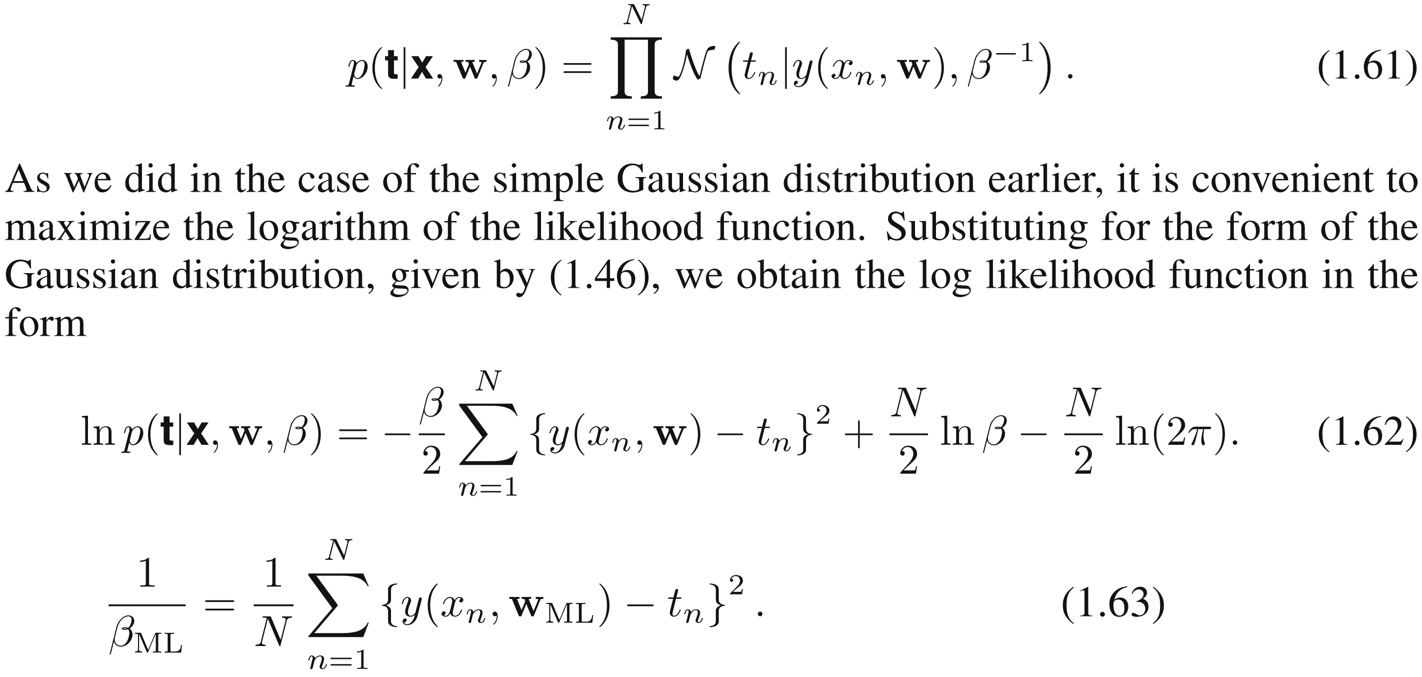
$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2$$

Curve Fitting Re-visited, Bishop1.2.5



t
$y(x, \mathbf{w})$
$y(x_0, \mathbf{w})$
2σ
$p(t|x_0, \mathbf{w}, \beta)$
$= \mathcal{N}\left(t|y(x_0, \mathbf{w}), \beta^{-1}\right)$
x_0
x

Maximum Likelihood



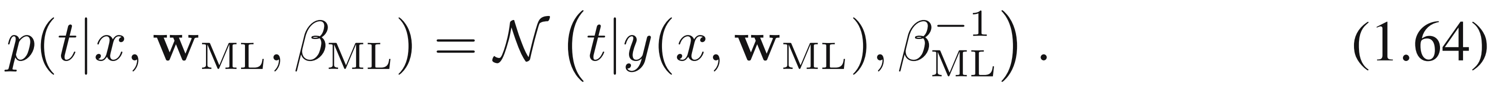
$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | y(x_n, \mathbf{w}), \beta^{-1}\right). \tag{1.61}$$

As we did in the case of the simple Gaussian distribution earlier, it is convenient to maximize the logarithm of the likelihood function. Substituting for the form of the Gaussian distribution, given by (1.46), we obtain the log likelihood function in the form

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi). \tag{1.62}$$

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2. \tag{1.63}$$

Giving estimates of W and beta, we can predict



$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t | y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right). \tag{1.64}$$

MAP: A Step towards Bayes 1.2.5

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

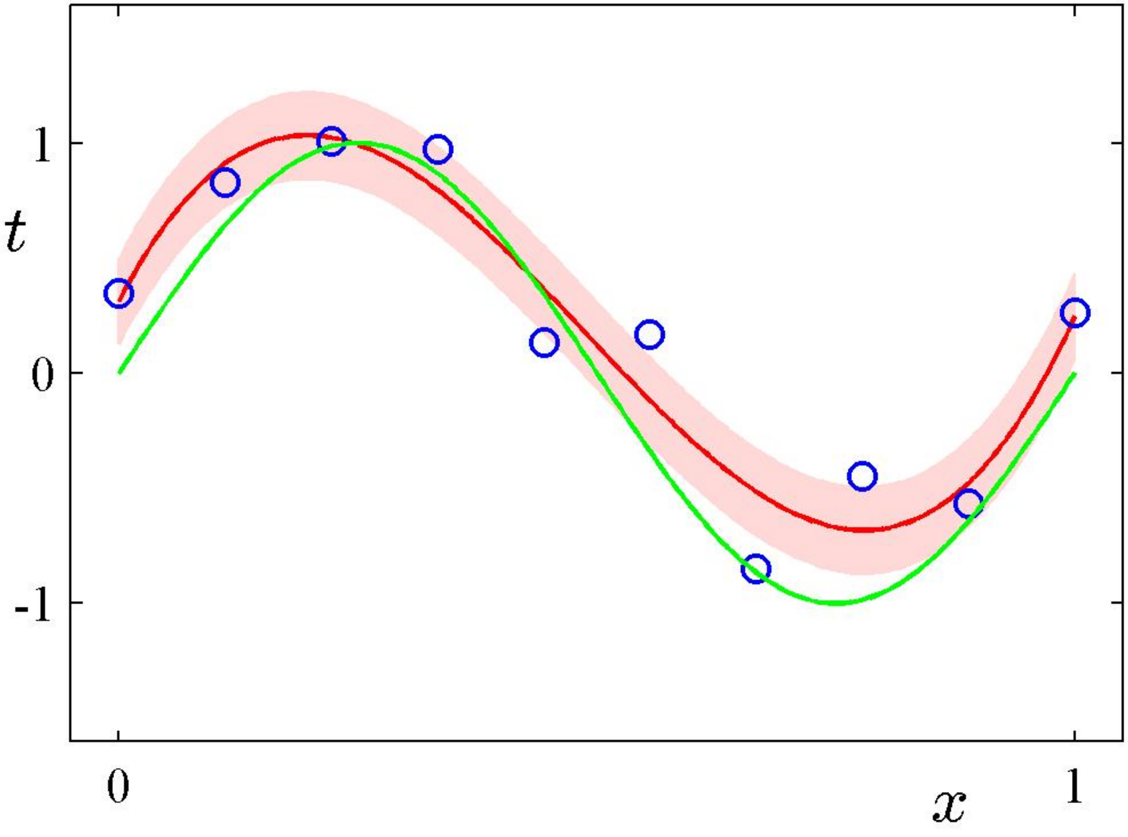
$$\beta\widetilde{E}(\mathbf{w}) = \frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

Determine $\mathbf{w}_{\mathrm{MAP}}$ by minimizing regularized sum-of-squares error, $\widetilde{E}(\mathbf{w})$.

Regularized sum of squares

Predictive Distribution

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right)$$



True data

Estimated
+/- std dev

Parametric Distributions

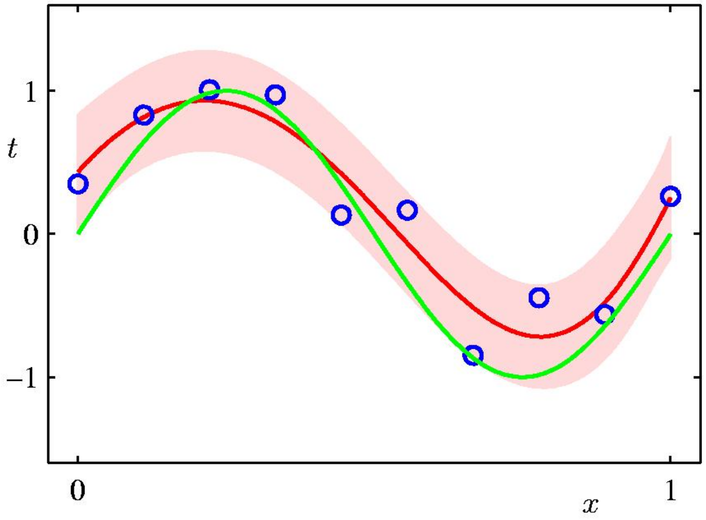
Basic building blocks: $p(\mathbf{x}|\boldsymbol{\theta})$

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

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

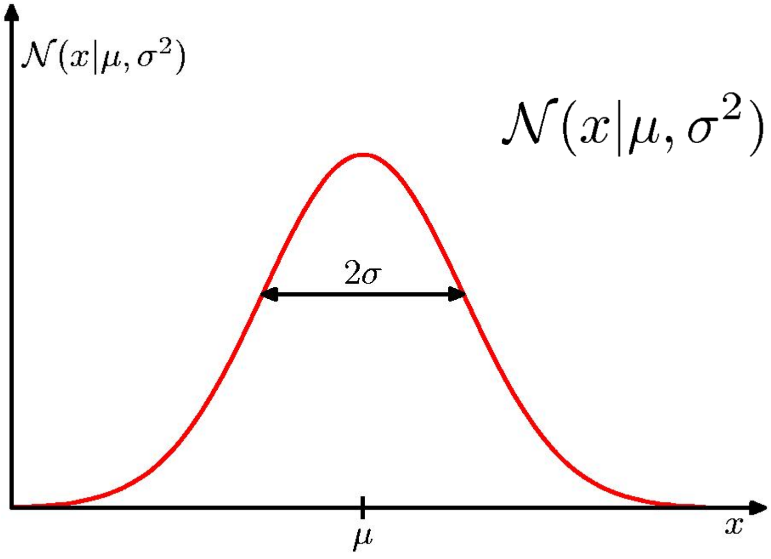
Recall Curve Fitting

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

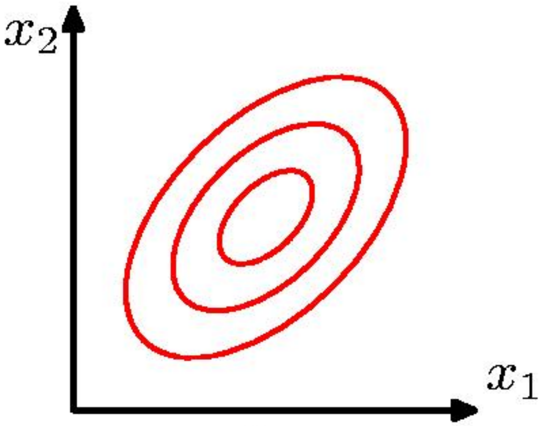


We focus on Gaussians!

The Gaussian Distribution



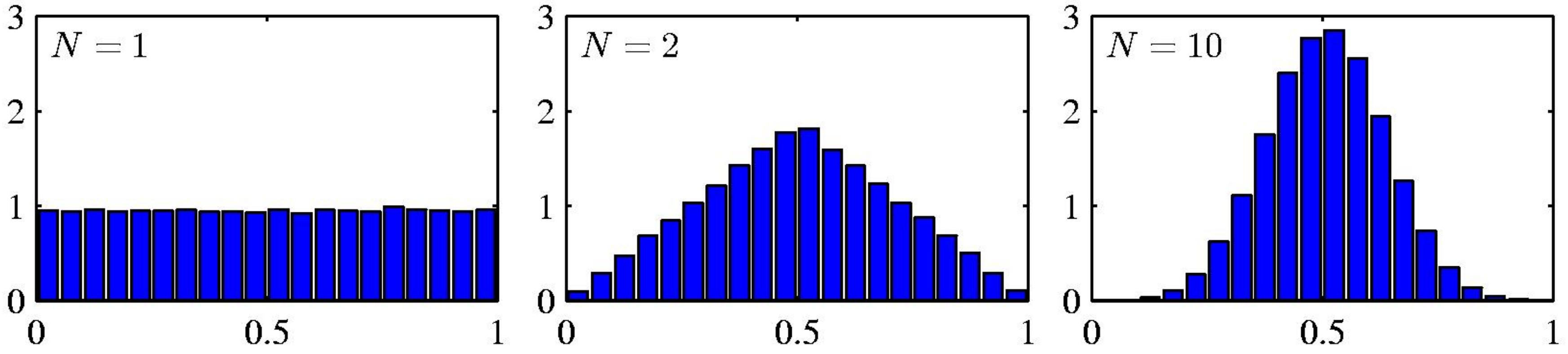
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



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

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.



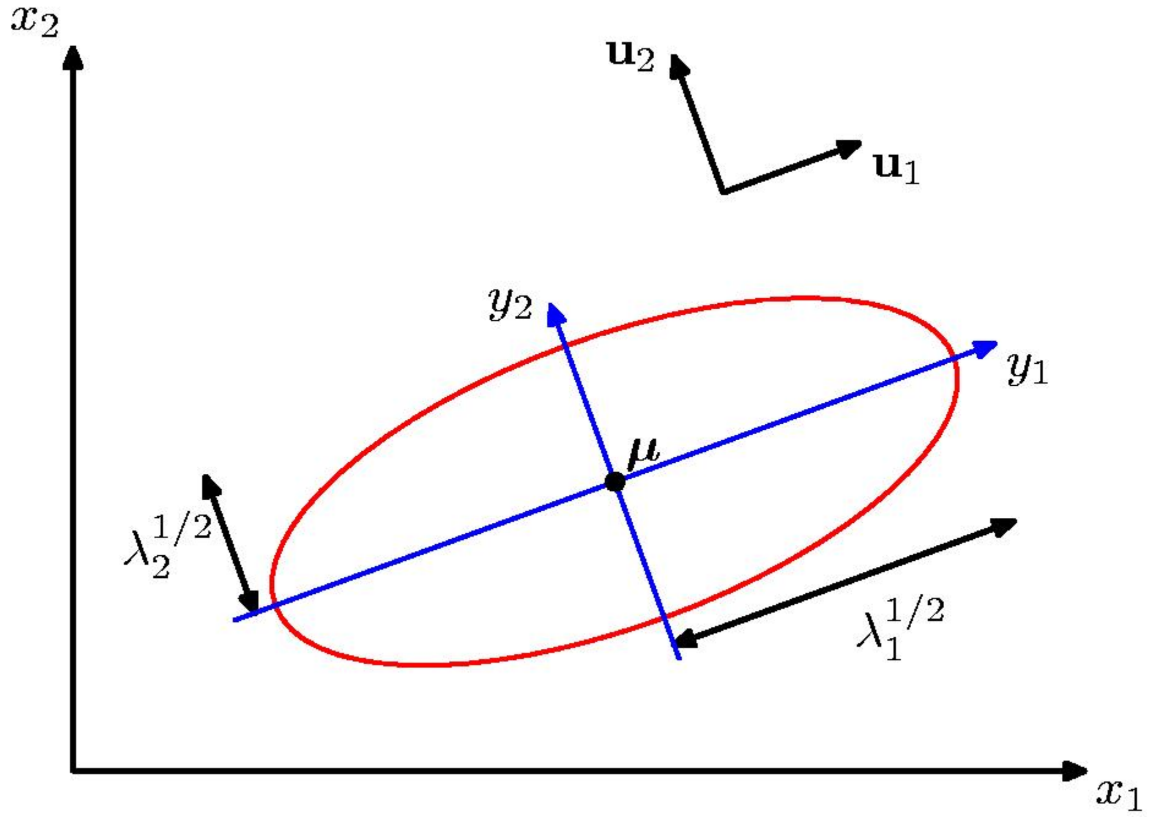
Geometry of the Multivariate Gaussian

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

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

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

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

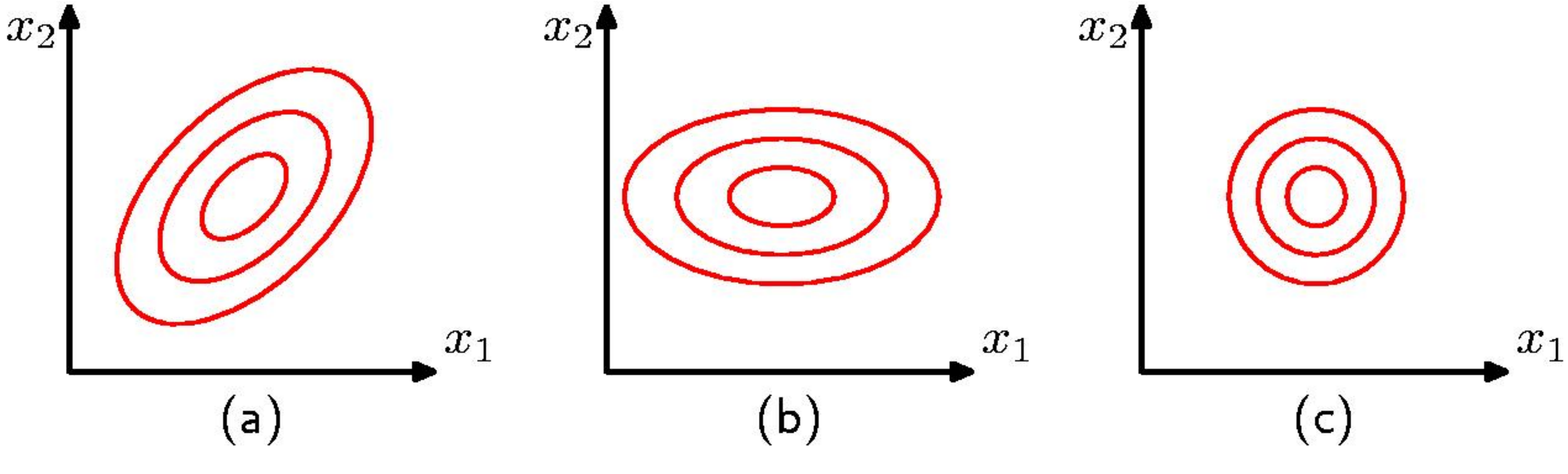


Moments of the Multivariate Gaussian (2)

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

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

A Gaussian requires D*(D-1)/2 +D parameters.
Often we use D +D or
Just D+1 parameters.

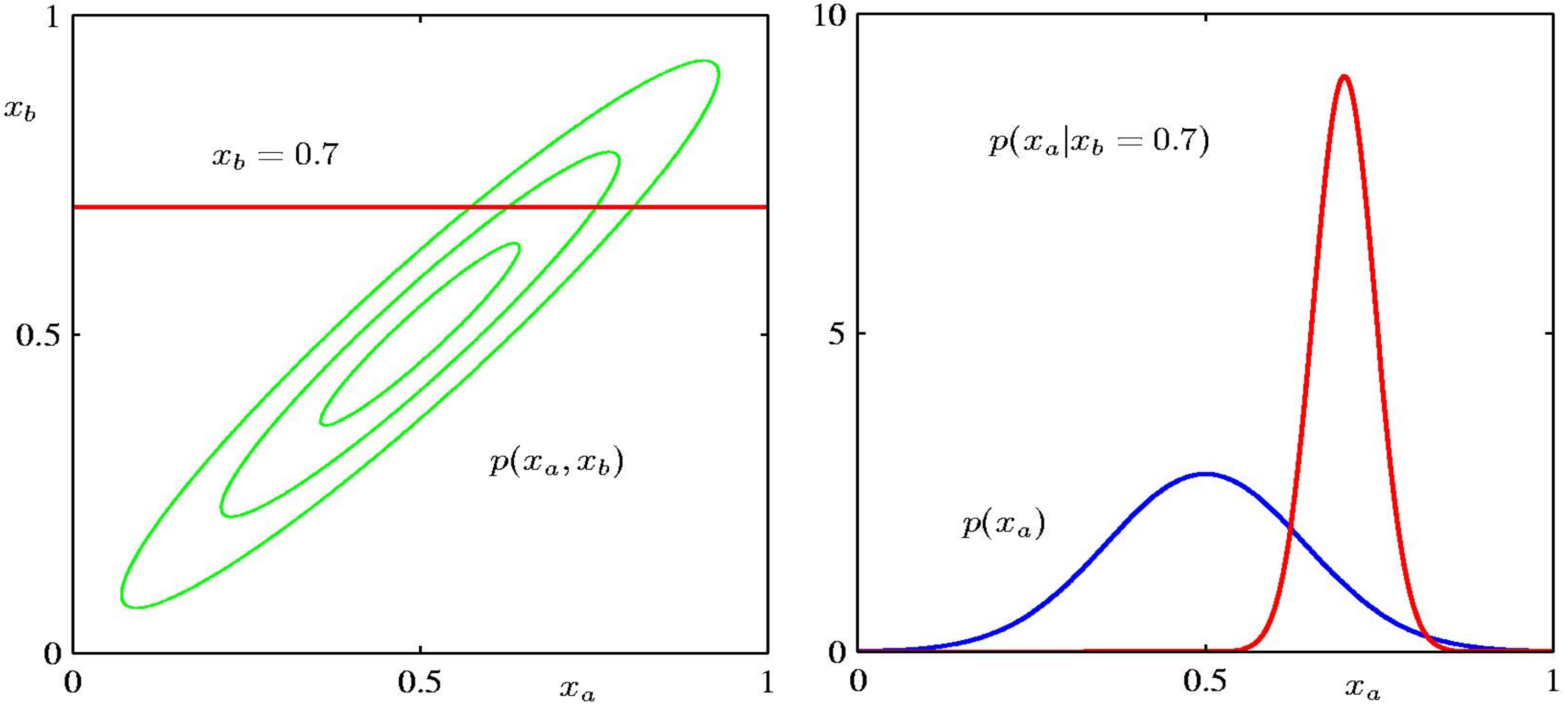


Partitioned Conditionals and Marginals, page 89

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

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

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



ML for the Gaussian (1) Bisphop 2.3.4

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

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

$$\frac{\partial}{\partial \mathbf{A}} \ln|\mathbf{A}| = \left(\mathbf{A}^{-1}\right)^{\mathrm{T}} \tag{C.28}$$

$$\frac{\partial}{\partial \mathbf{A}} \mathrm{Tr}\,(\mathbf{A}\mathbf{B}) = \mathbf{B}^{\mathrm{T}}. \tag{C.24}$$

$$\frac{\partial}{\partial x}\left(\mathbf{A}^{-1}\right) = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1} \tag{C.21}$$

Maximum Likelihood for the Gaussian

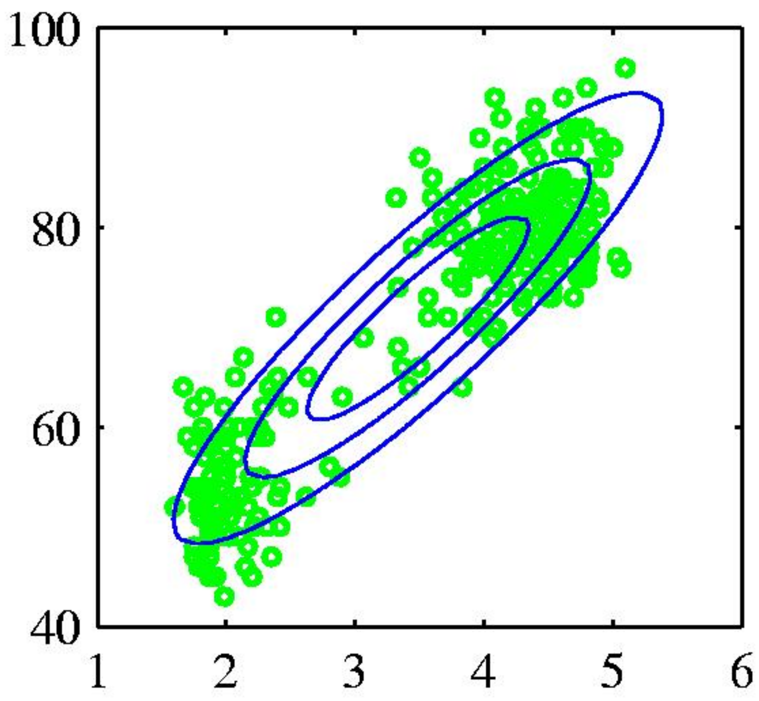
- Set the derivative of the log likelihood function to zero,
- and solve to obtain $$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = 0$$
- Similarly $$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n.$$

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

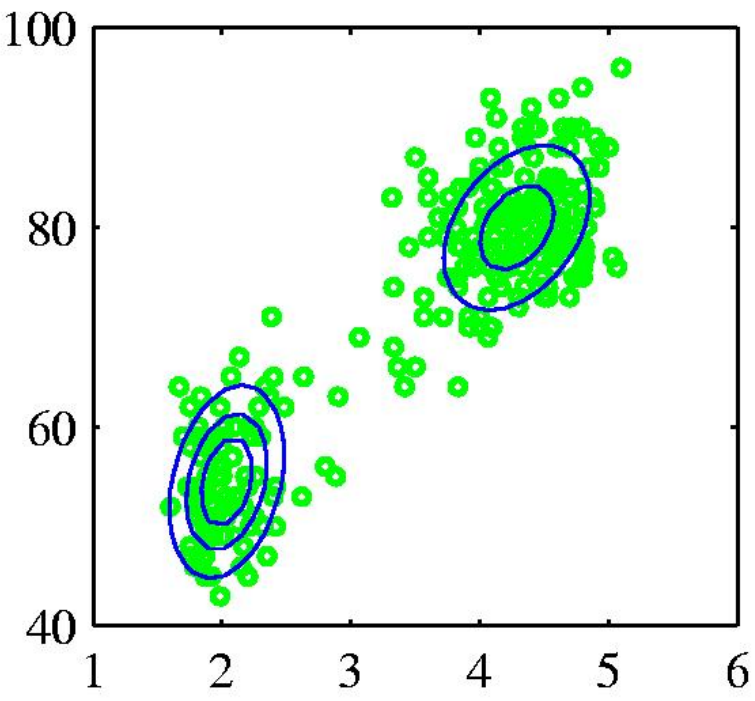
Mixtures of Gaussians (Bishop 2.3.9)

Old Faithful geyser:

The time between eruptions has a bimodal distribution, with the mean interval being either 65 or 91 minutes, and is dependent on the length of the prior eruption. Within a margin of error of ±10 minutes, Old Faithful will erupt either 65 minutes after an eruption lasting less than 2 ½ minutes, or 91 minutes after an eruption lasting more than 2 ½ minutes.



Single Gaussian

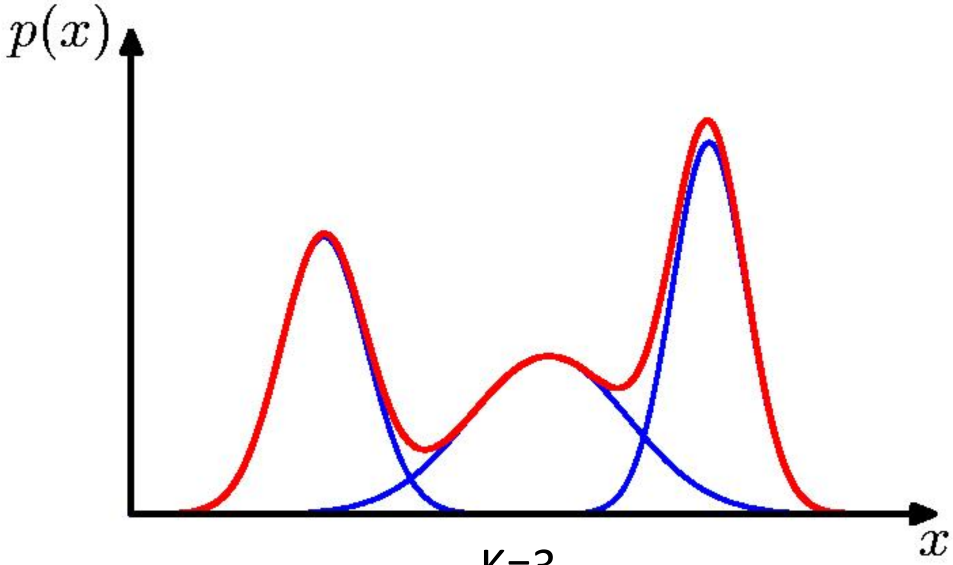


Mixture of two Gaussians

Mixtures of Gaussians (Bishop 2.3.9)

- Combine simple models into a complex model:

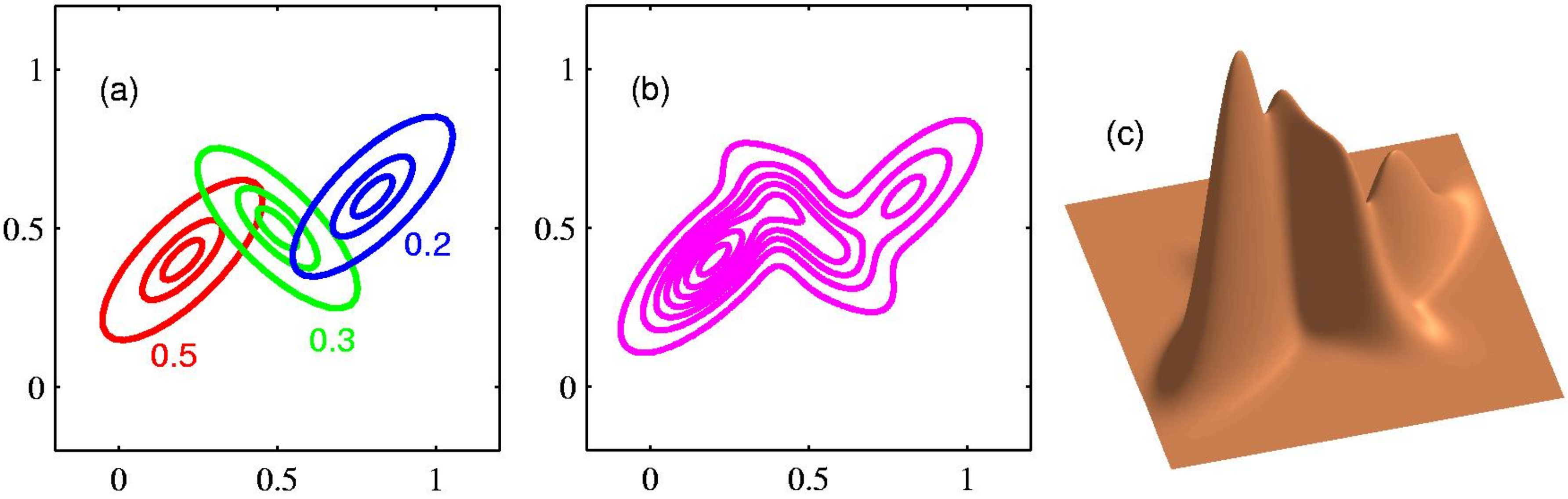
$$p(\mathbf{x}) = \sum_{k=1}^{K} \underset{\substack{\uparrow \\ \text{Mixing coefficient}}}{\pi_k} \underbrace{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{Component}}$$



K=3

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

Mixtures of Gaussians (Bishop 2.3.9)



Mixtures of Gaussians (Bishop 2.3.9)

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

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

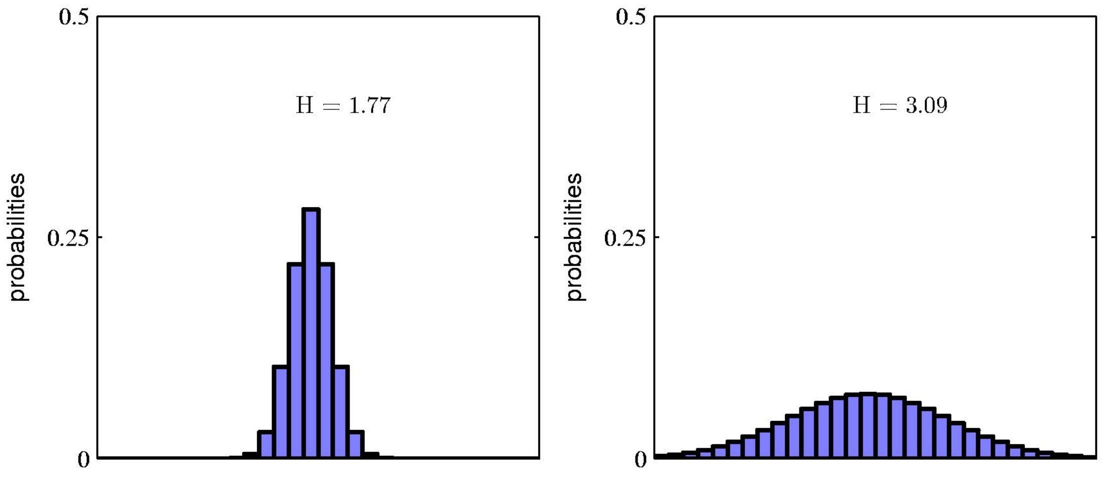
- Solution: use standard, iterative, numeric optimization methods or the *expectation maximization* algorithm (Chapter 9).

Entropy 1.6

$$\mathrm{H}[x] = -\sum_x p(x) \log_2 p(x)$$

Important quantity in

- coding theory
- statistical physics
- machine learning



Differential Entropy

Put bins of width ¢ along the real line

$$\lim_{\Delta \to 0} \left\{ - \sum_i p(x_i) \Delta \ln p(x_i) \right\} = - \int p(x) \ln p(x) \, \mathrm{d}x$$

For fixed σ^2 differential entropy maximized when

in which case

$$p(x) = \mathcal{N}(x|\mu, \sigma^2)$$

$$\mathrm{H}[x] = \frac{1}{2} \left\{ 1 + \ln(2\pi\sigma^2) \right\} .$$

The Kullback-Leibler Divergence

P true distribution, q is approximating distribution

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x})\ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right)\\ &= -\int p(\mathbf{x})\ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}\,\mathrm{d}\mathbf{x}\end{aligned}$$

$$\mathrm{KL}(p\|q) \simeq \frac{1}{N}\sum_{n=1}^{N}\{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\}$$

$$\mathrm{KL}(p\|q) \geqslant 0 \qquad\qquad \mathrm{KL}(p\|q) \not\equiv \mathrm{KL}(q\|p)$$

Decision Theory

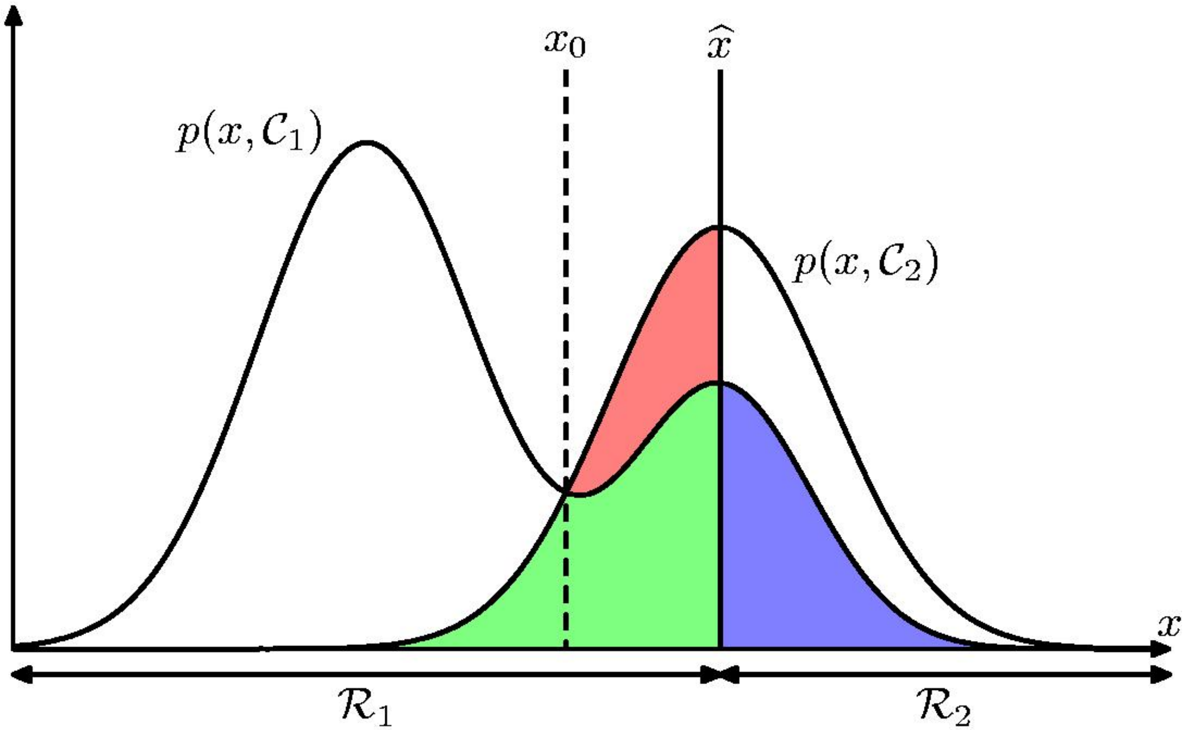
Inference step

Determine either $p(t|\mathbf{x})$ or $p(\mathbf{x}, t)$.

Decision step

For given x, determine optimal t.

Minimum Misclassification Rate



$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}. \end{aligned}$$

- UNTIL HERE 4 April 2018

Bayes for linear model

$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$ $\boldsymbol{n} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{CD})$ $\mathbf{y} \sim \mathrm{N}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{CD})$ prior: $\mathbf{x} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{CD})$

$p(\boldsymbol{x}|\boldsymbol{y}) \sim p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) \sim N(\boldsymbol{y}, \boldsymbol{C}_p)$

Bayes’ Theorem for Gaussian Variables

- 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

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

Sequential Estimation

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

$$\begin{aligned}
\boldsymbol{\mu}_{\mathrm{ML}}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} \\
&= \underbrace{\boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)}}_{\text{old estimate}} + \underbrace{\frac{1}{N}}_{\text{correction weight}}\underbrace{(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)})}_{\text{correction given } x_N}
\end{aligned}$$

Bayesian Inference for the Gaussian Bishop2.3.6

- 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).

Bayesian Inference for the Gaussian Bishop2.3.6

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

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

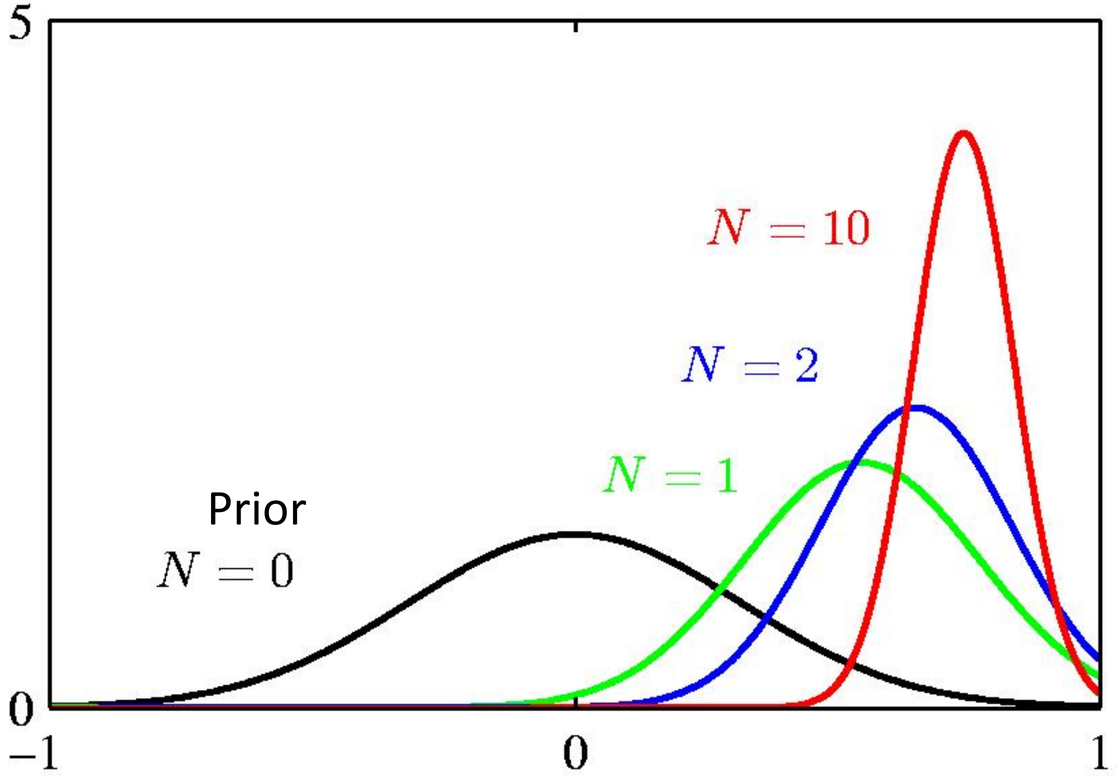
$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

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

Bayesian Inference for the Gaussian (3)

- Example: for N = 0, 1, 2 and 10.

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



Bayesian Inference for the Gaussian (4)

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.

- NON PARAMETRIC

Nonparametric Methods (1)

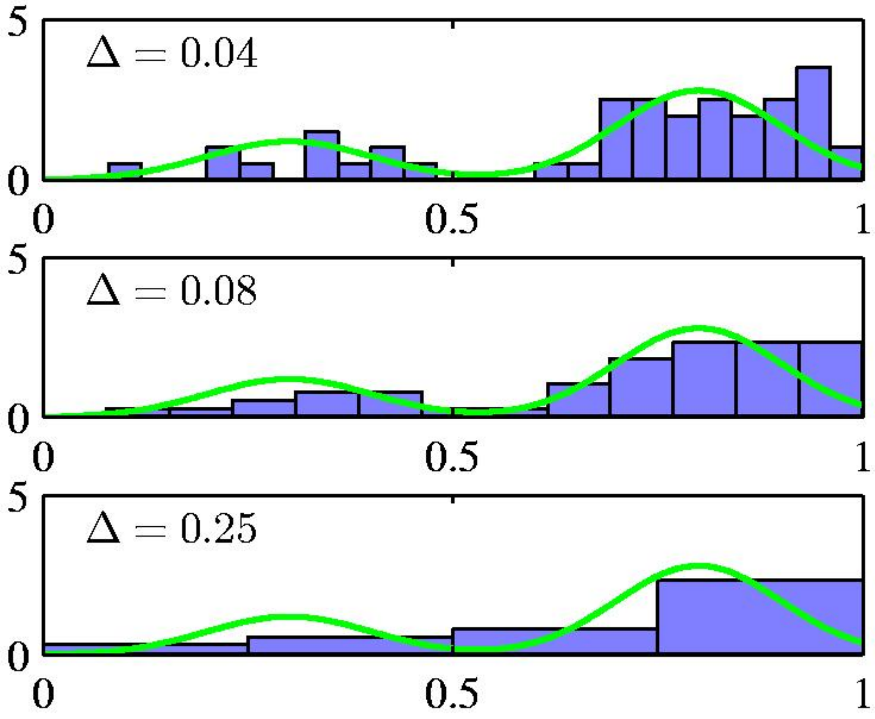
- Parametric distribution models are restricted to specific forms, which may not always be suitable; for example, consider modelling a multimodal distribution with a single, unimodal model.
- Nonparametric approaches make few assumptions about the overall shape of the distribution being modelled.
- 1000 parameter versus 10 parameter

Nonparametric Methods (2)

Histogram methods partition the data space into distinct bins with widths ¢$_i$ 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!



Nonparametric Methods (3)

- Assume observations drawn from a density p(x) and consider a small region R containing x such that

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

- The probability that K out of N observations lie inside R is Bin(KjN,P) and if N is large

$$K \simeq NP.$$

If the volume of R, V, is sufficiently small, p(x) is approximately constant over R and

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

Thus

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

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

Nonparametric Methods (4)

- **Kernel Density Estimation:** fix V, estimate K from the data. Let R be a hypercube centred on 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

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

Nonparametric Methods (5)

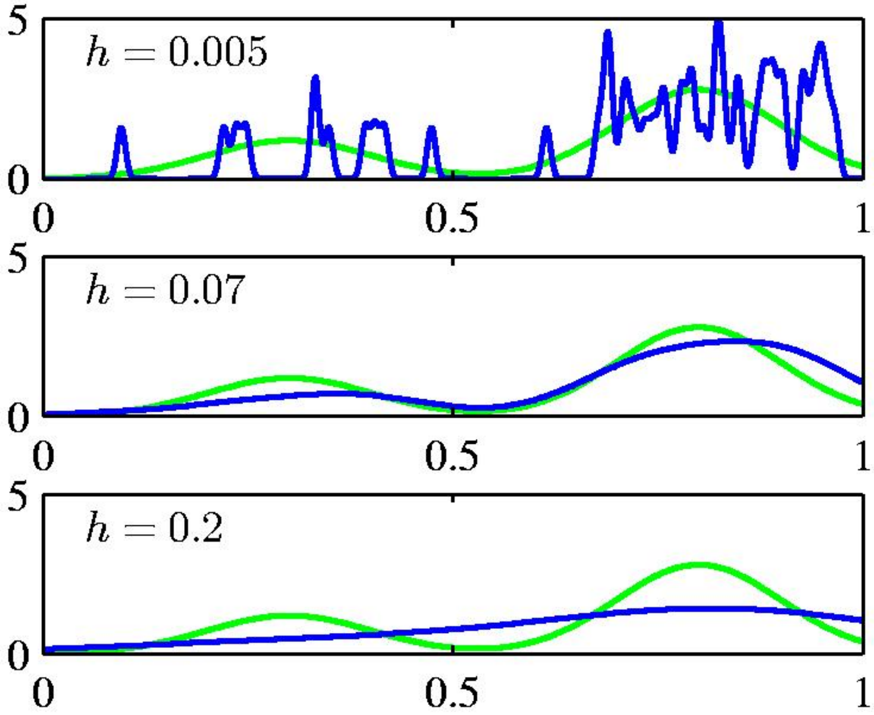
- To avoid discontinuities in p(x), use a smooth kernel, e.g. a Gaussian

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

- Any kernel such that

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

- will work.

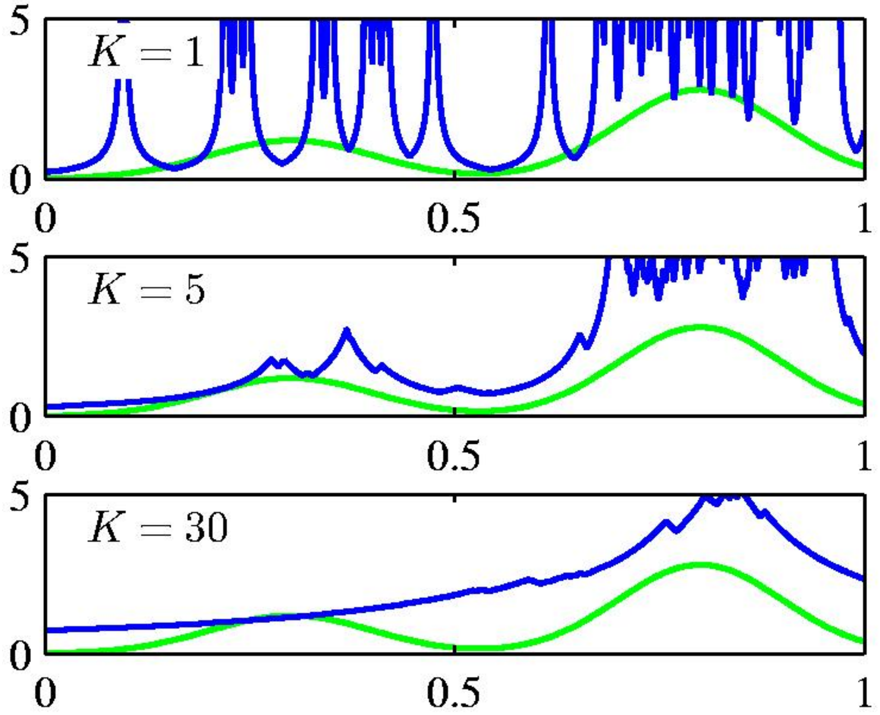


h acts as a smoother.

Nonparametric Methods (6)

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

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



K acts as a smoother.

Nonparametric Methods (7)

- 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.

K-Nearest-Neighbours for Classification (1)

- Given a data set with N_k data points from class C_k and $\sum_k N_k = N$ we have

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

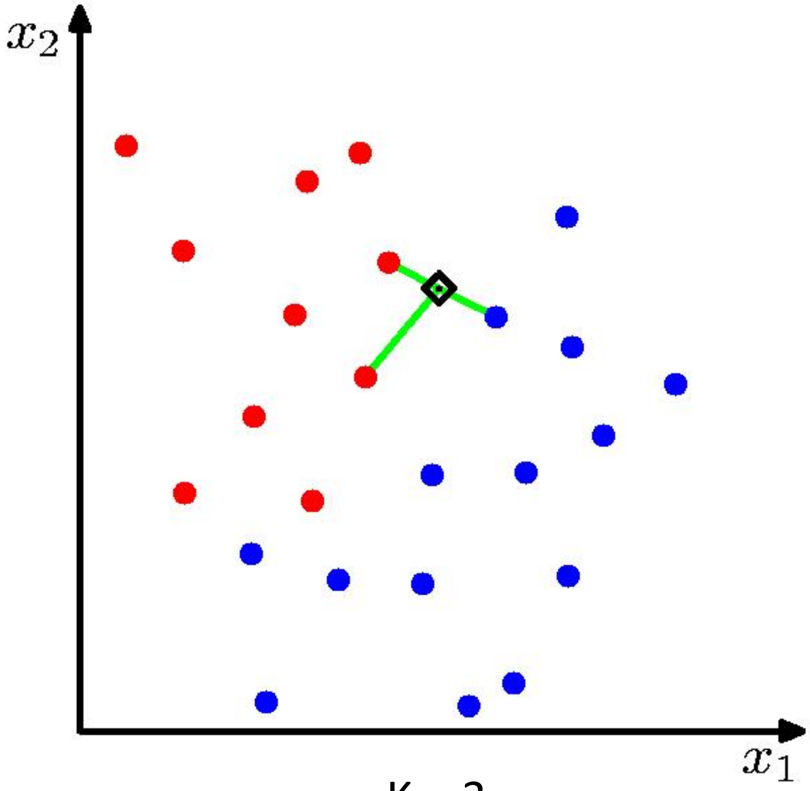
- and correspondingly

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

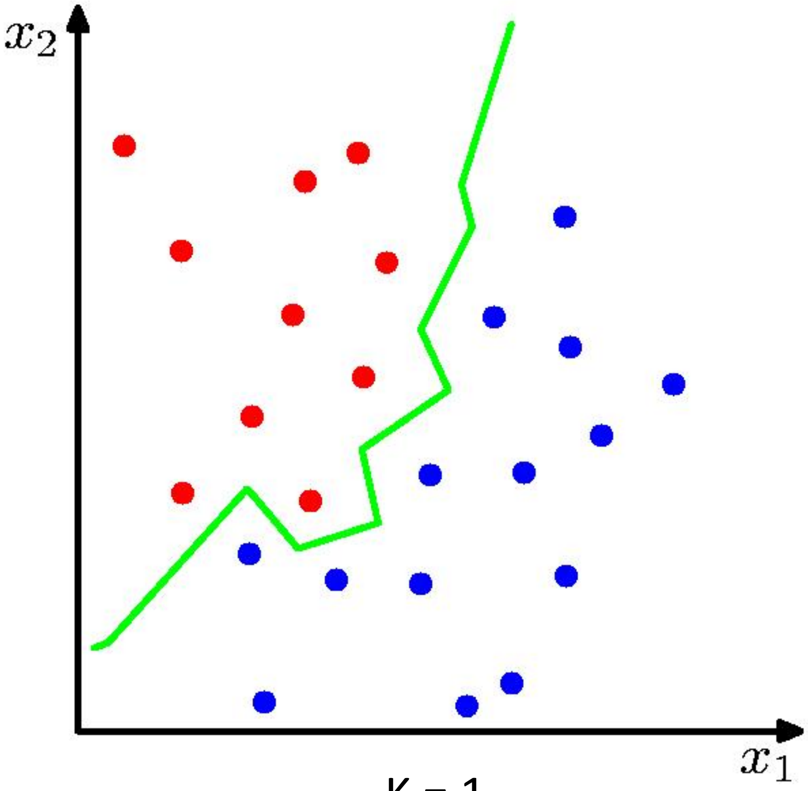
- Since $p(\mathcal{C}_k) = N_k/N$, Bayes' theorem gives

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

K-Nearest-Neighbours for Classification (2)

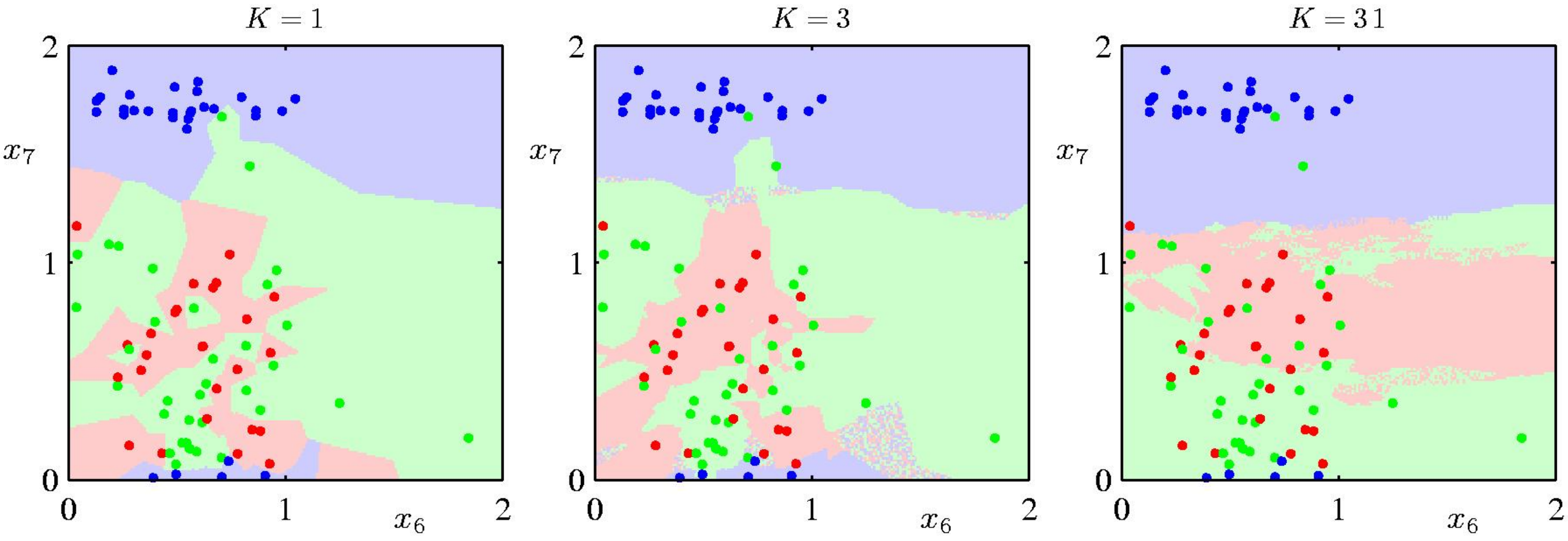


K = 3



K = 1

K-Nearest-Neighbours for Classification (3)



- 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).

OLD

Bayesian Inference for the Gaussian (6)

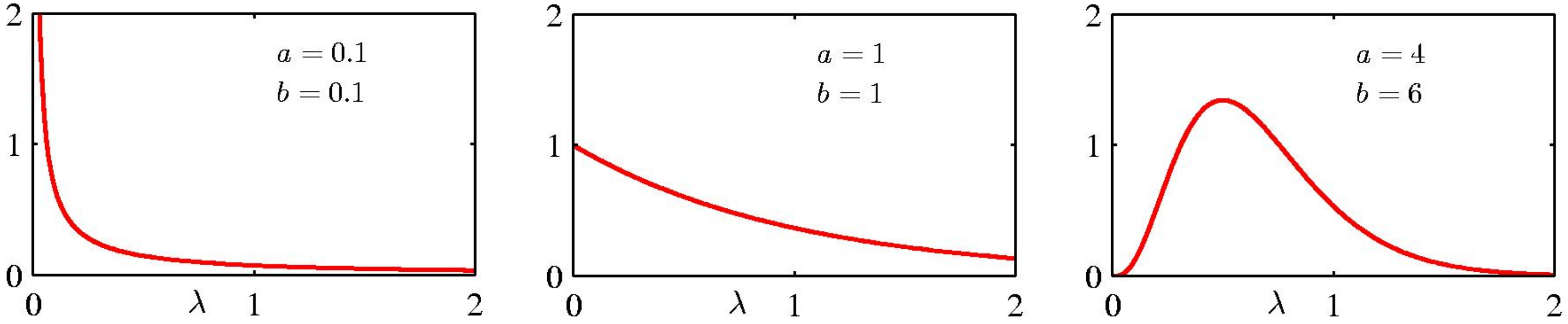
- Now assume μ is known. The likelihood function for $\lambda=1/\sigma^2$ is given by

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

- This has a Gamma shape as a function of λ.

- The Gamma distribution:

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \qquad \mathbb{E}[\lambda] = \frac{a}{b} \qquad \mathrm{var}[\lambda] = \frac{a}{b^2}$$



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

$$\begin{aligned} a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma_{\mathrm{ML}}^2. \end{aligned}$$

Bayesian Inference for the Gaussian (9)

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

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

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

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

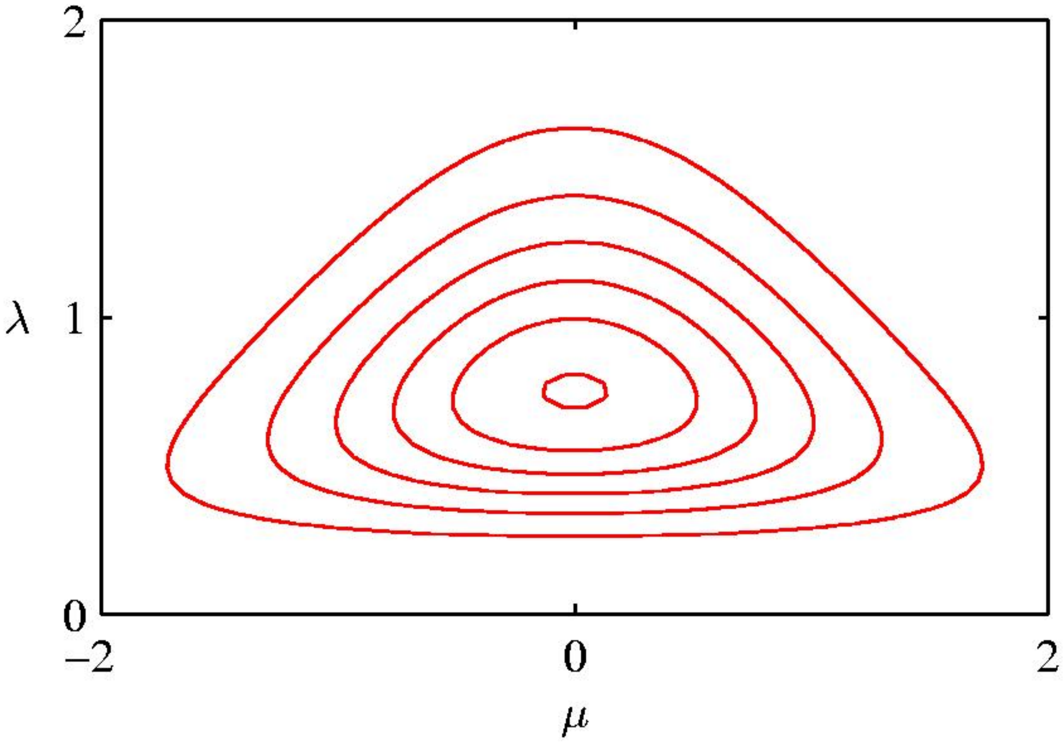
Bayesian Inference for the Gaussian (10)

- The Gaussian-gamma distribution

$$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a, b)$$

$$\propto \quad \exp\underbrace{\left\{-\frac{\beta\lambda}{2}(\mu - \mu_0)^2\right\}}_{\substack{\bullet\ \text{Quadratic in } \mu. \\ \bullet\ \text{Linear in } \lambda.}} \underbrace{\lambda^{a-1} \exp\{-b\lambda\}}_{\substack{\bullet\ \text{Gamma distribution over } \lambda. \\ \bullet\ \text{Independent of } \mu.}}$$

$\mu_0=0$, $\beta=2$, a=5, b=6



Bayesian Inference for the Gaussian (12)

- Multivariate conjugate priors
- $\boldsymbol{\mu}$ unknown, $\boldsymbol{\Lambda}$ known: p(μ) Gaussian.
- $\boldsymbol{\Lambda}$ unknown, μ 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 μ unknown: p(μ, $\boldsymbol{\Lambda}$) Gaussian-Wishart,

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

Partitioned Gaussian Distributions

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

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

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

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

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

Moments of the Multivariate Gaussian (1)

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

thanks to anti-symmetry of z

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