PATTERN RECOGNITION AND MACHINE LEARNING

CHAPTER 2: PROBABILITY DISTRIBUTIONS

Parametric Distributions

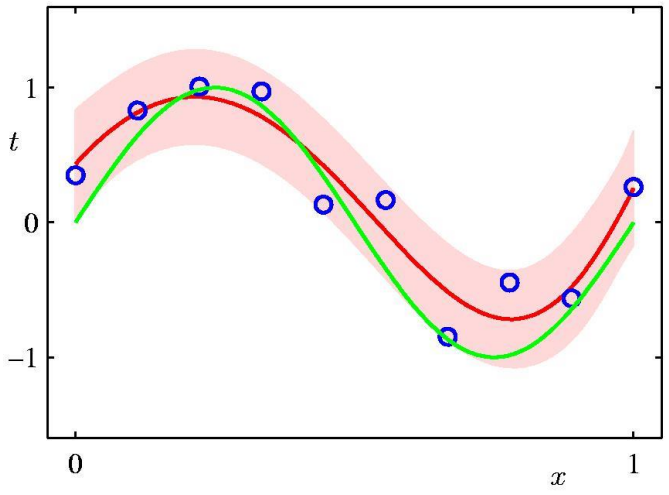
Basic building blocks: $p(\mathbf{x}|\boldsymbol{\theta})$

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

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

Recall Curve Fitting

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



Binary Variables (1)

Coin flipping: heads=1, tails=0

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

Bernoulli Distribution

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

Binary Variables (2)

N coin flips:

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

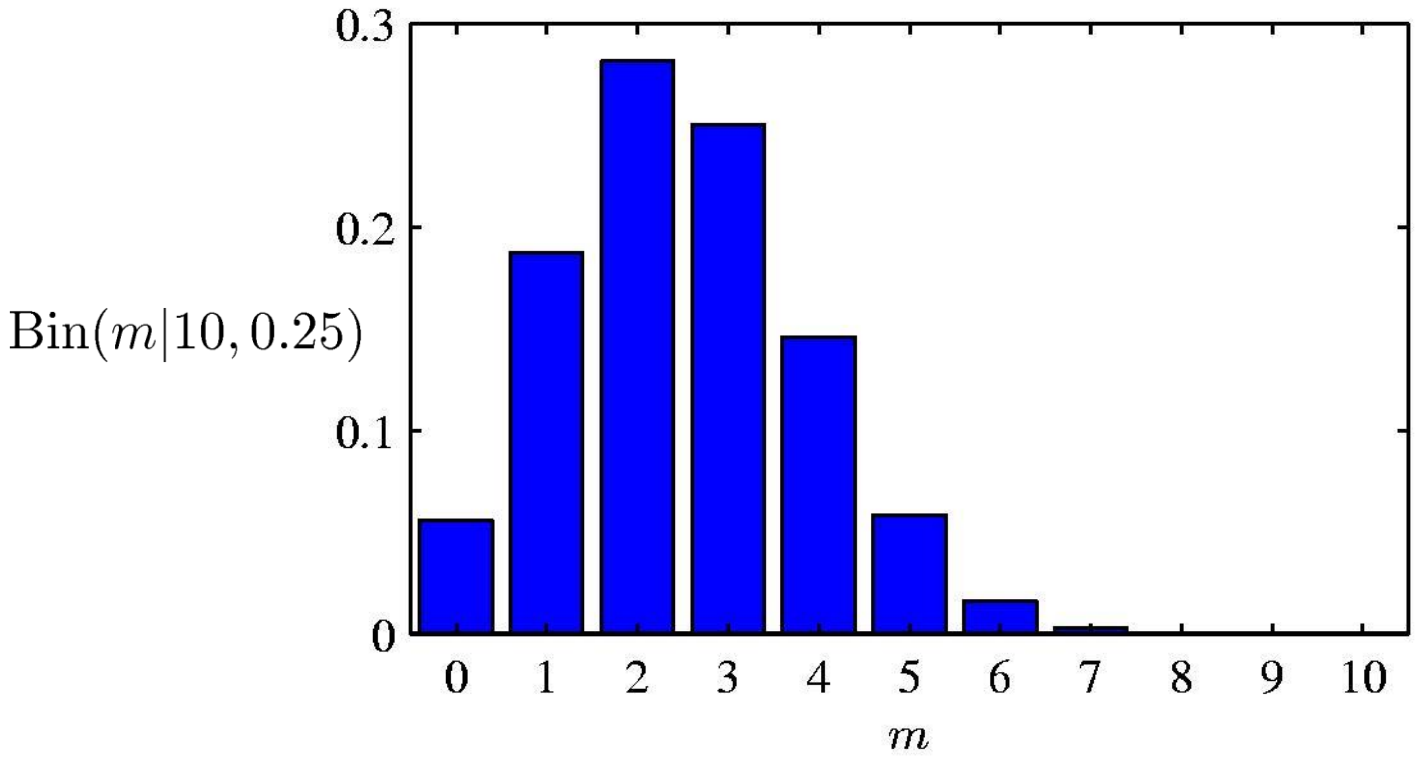
Binomial Distribution

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

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

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

Binomial Distribution



Parameter Estimation (1)

ML for Bernoulli

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

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

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

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

Parameter Estimation (2)

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

Prediction: *all* future tosses will land heads up

Overfitting to $\mathcal{D}$

Beta Distribution

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

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

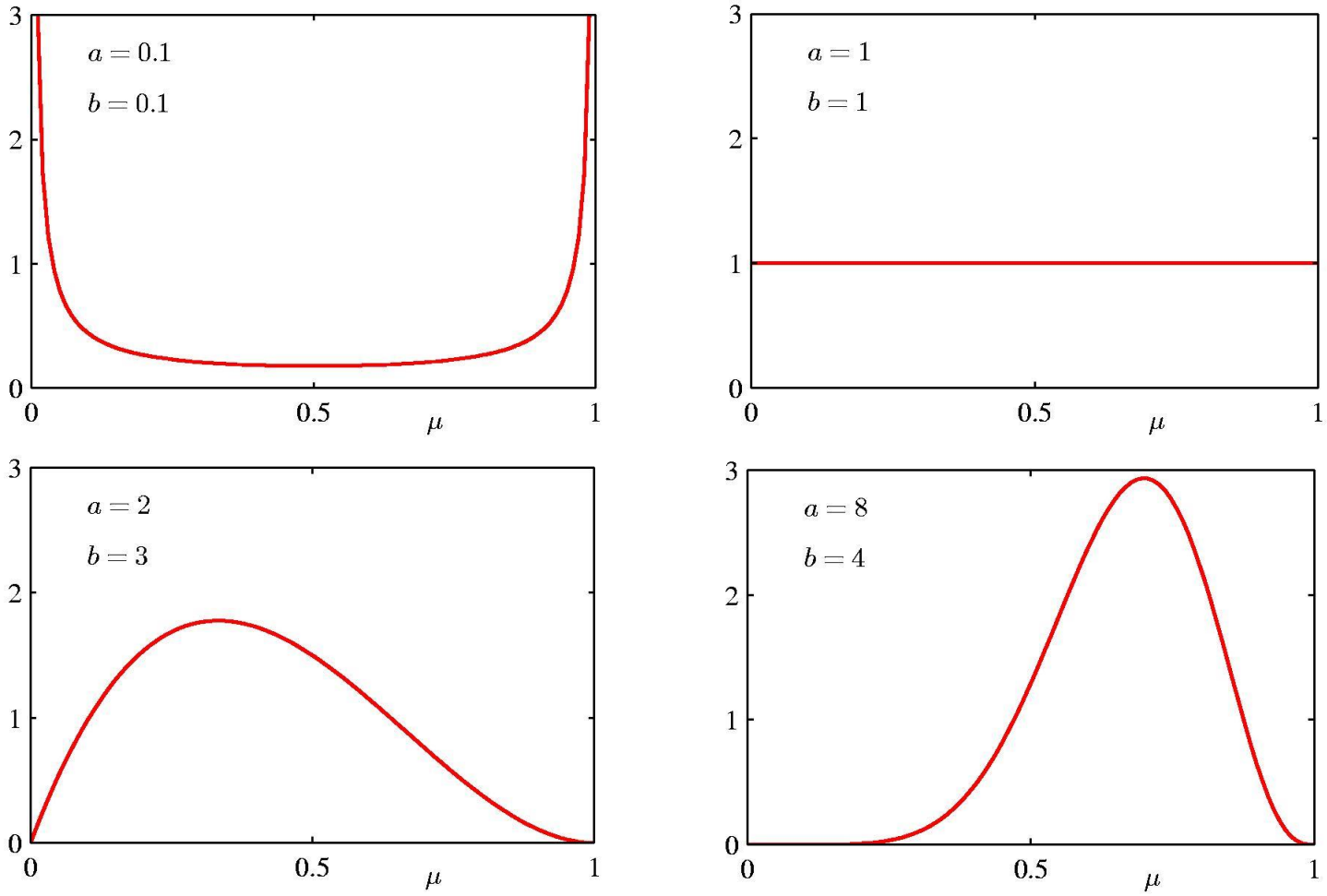
Bayesian Bernoulli

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

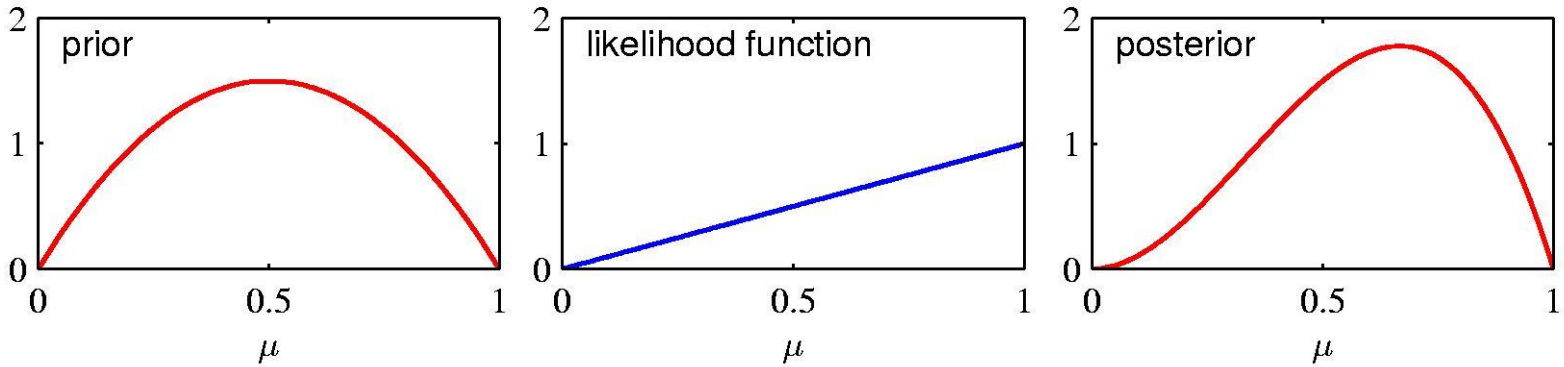
$$
a_N = a_0 + m \qquad b_N = b_0 + (N - m)
$$

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

Beta Distribution



Prior · Likelihood = Posterior



Properties of the Posterior

As the size of the data set, N, increase

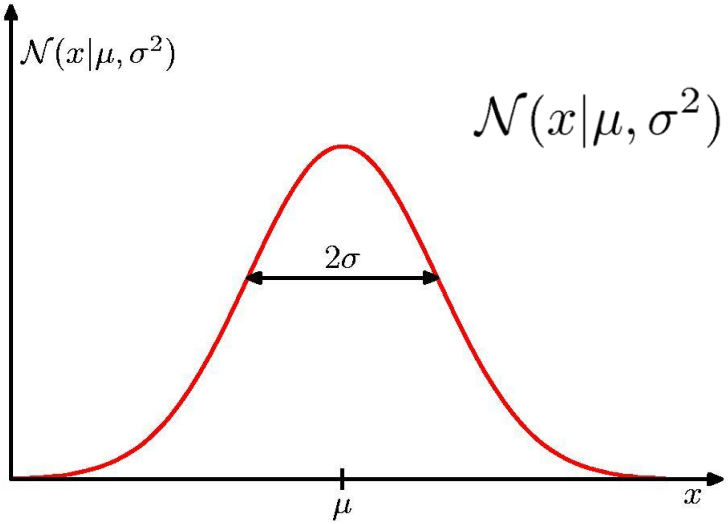
$$
\begin{aligned}
a_N &\rightarrow m \\
b_N &\rightarrow N - m \\
\mathbb{E}[\mu] &= \frac{a_N}{a_N + b_N} \rightarrow \frac{m}{N} = \mu_{\mathrm{ML}} \\
\mathrm{var}[\mu] &= \frac{a_N b_N}{(a_N + b_N)^2 (a_N + b_N + 1)} \rightarrow 0
\end{aligned}
$$

Prediction under the Posterior

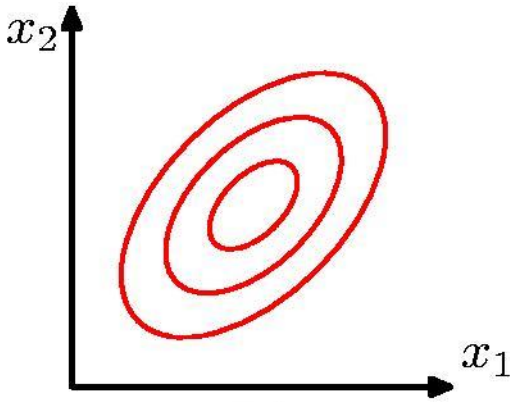
What is the probability that the next coin toss will land heads up?

$$
\begin{aligned}
p(x=1|a_0,b_0,\mathcal{D}) &= \int_0^1 p(x=1|\mu)p(\mu|a_0,b_0,\mathcal{D})\,\mathrm{d}\mu \\
&= \int_0^1 \mu p(\mu|a_0,b_0,\mathcal{D})\,\mathrm{d}\mu \\
&= \mathbb{E}[\mu|a_0,b_0,\mathcal{D}] = \frac{a_N}{b_N}
\end{aligned}
$$

The Gaussian Distribution



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

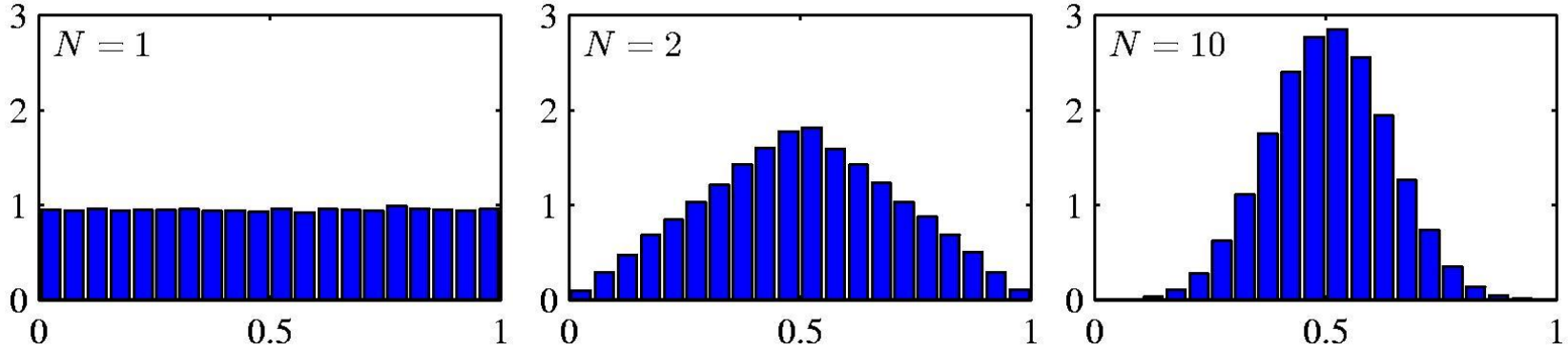


$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

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.



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

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

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

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

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

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

Maximum Likelihood for the Gaussian (3)

Under the true distribution

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

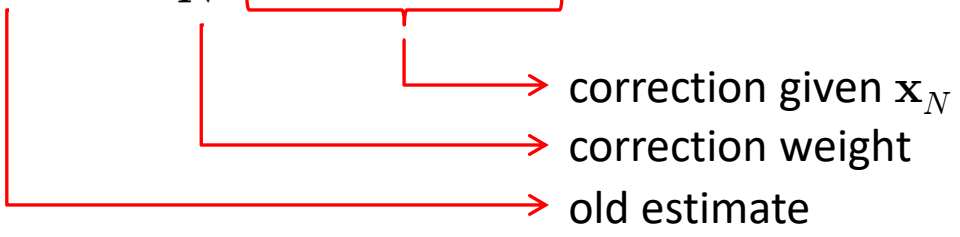
Hence define

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

Sequential Estimation

Contribution of the N^{th} data point, $\mathbf{x}_N$

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



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).

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

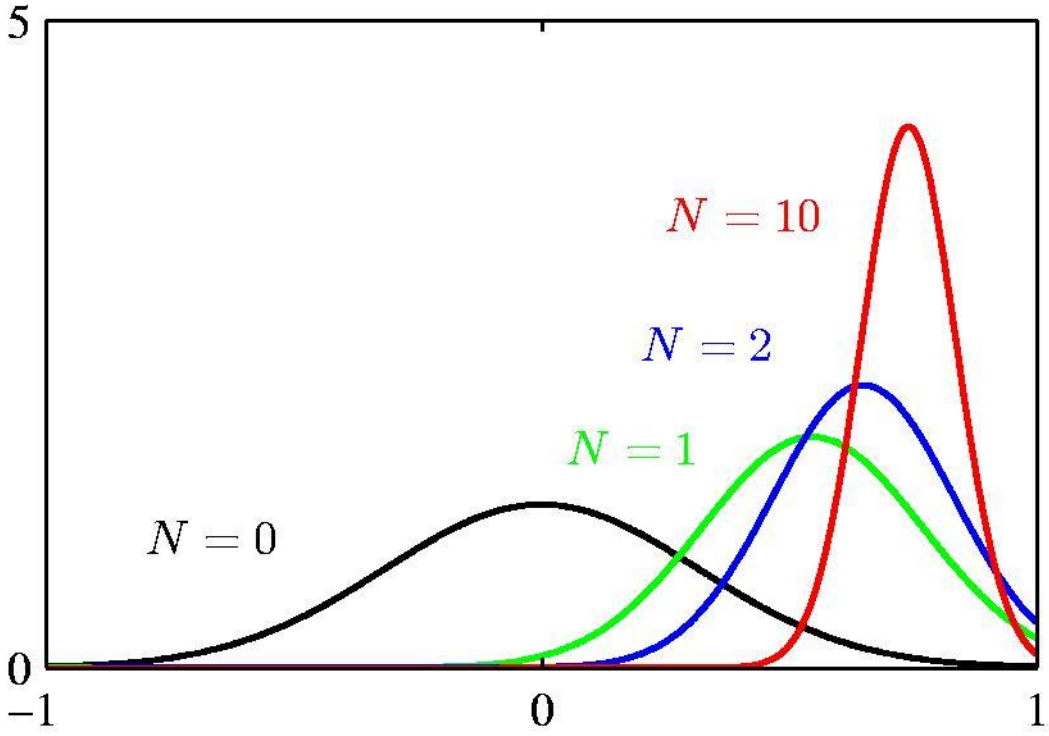
$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu).$$

Completing the square over μ, we see that

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

Bayesian Inference for the Gaussian (3)

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

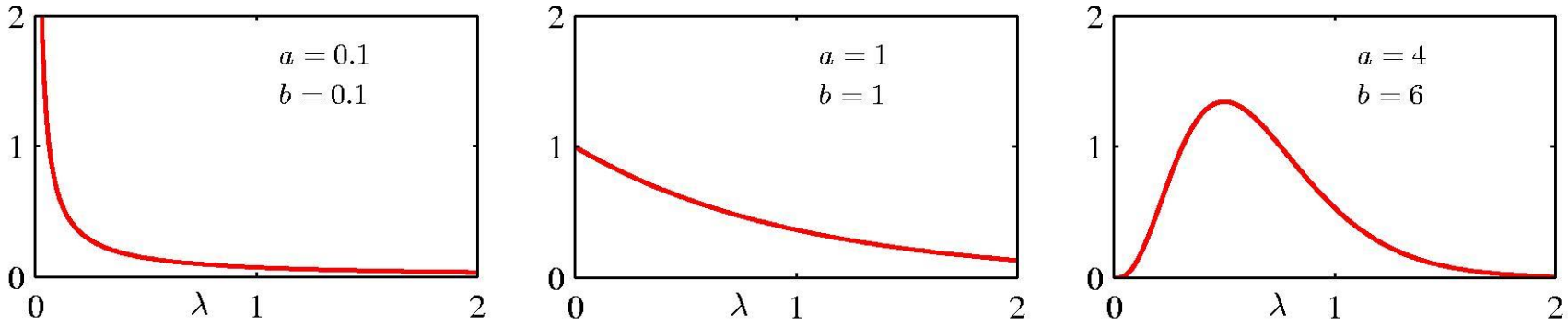


Bayesian Inference for the Gaussian (7)

The Gamma distribution

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

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



Student's t-Distribution

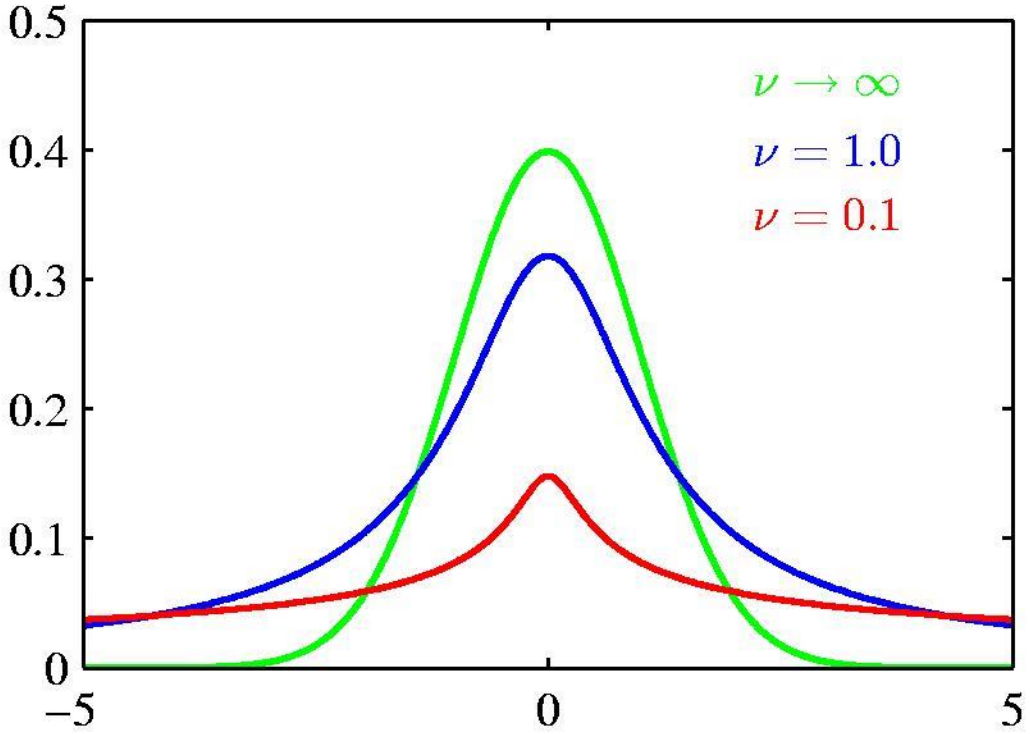
$$
\begin{aligned}
p(x|\mu,a,b) &= \int_0^\infty \mathcal{N}(x|\mu,\tau^{-1})\mathrm{Gam}(\tau|a,b)\,\mathrm{d}\tau \\
&= \int_0^\infty \mathcal{N}\left(x|\mu,(\eta\lambda)^{-1}\right)\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta \\
&= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-\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

where

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

Infinite mixture of Gaussians.

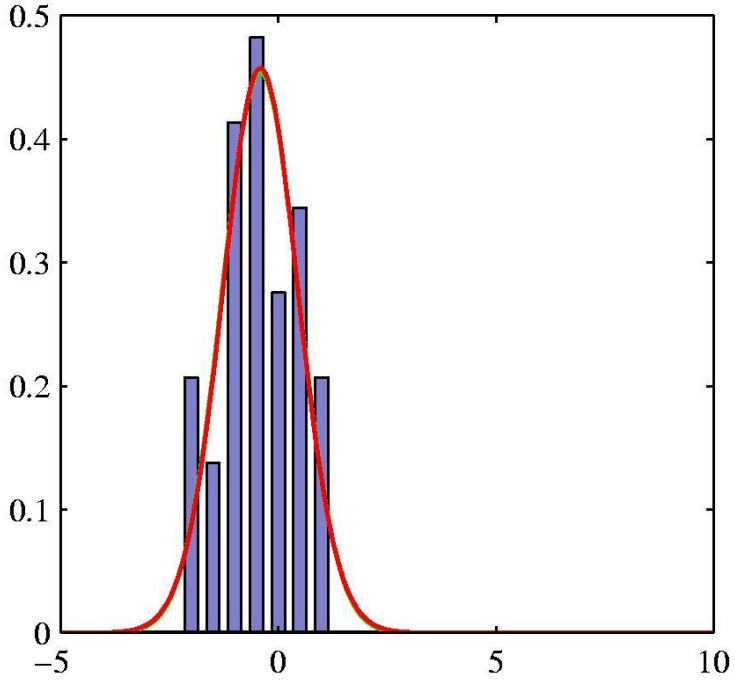
Student's t-Distribution

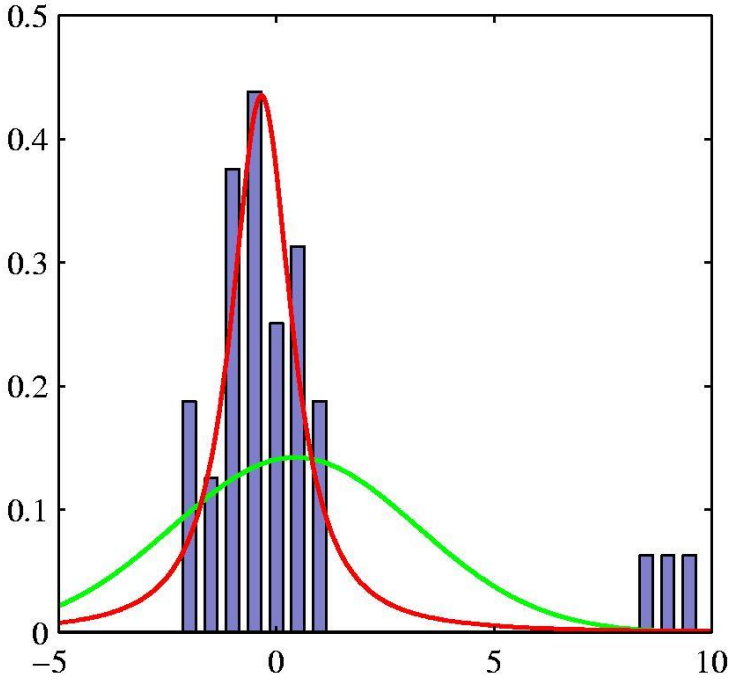


	$\nu = 1$	$\nu \to \infty$
$\mathrm{St}(x\|\mu, \lambda, \nu)$	Cauchy	$\mathcal{N}(x\|\mu, \lambda^{-1})$

Student's t-Distribution

Robustness to outliers: Gaussian vs t-distribution.





Student's t-Distribution

The D-variate case:

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

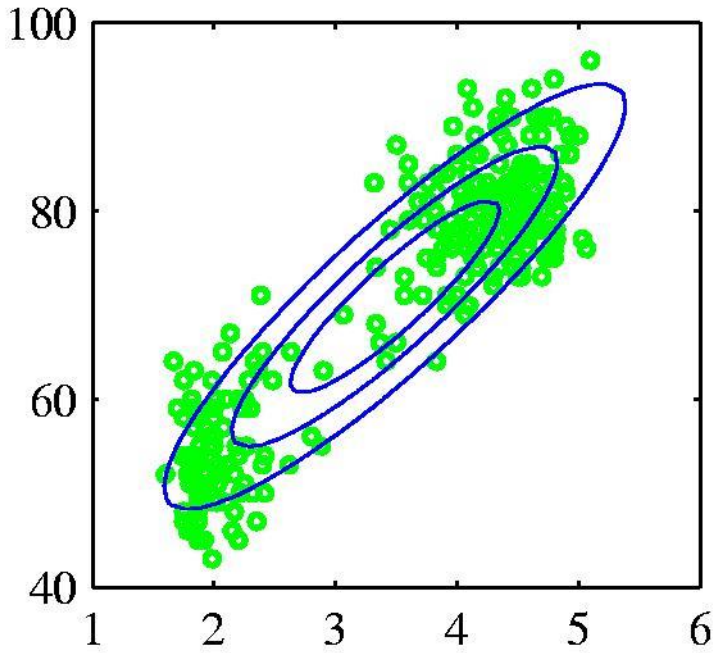
where $\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})$.

Properties:

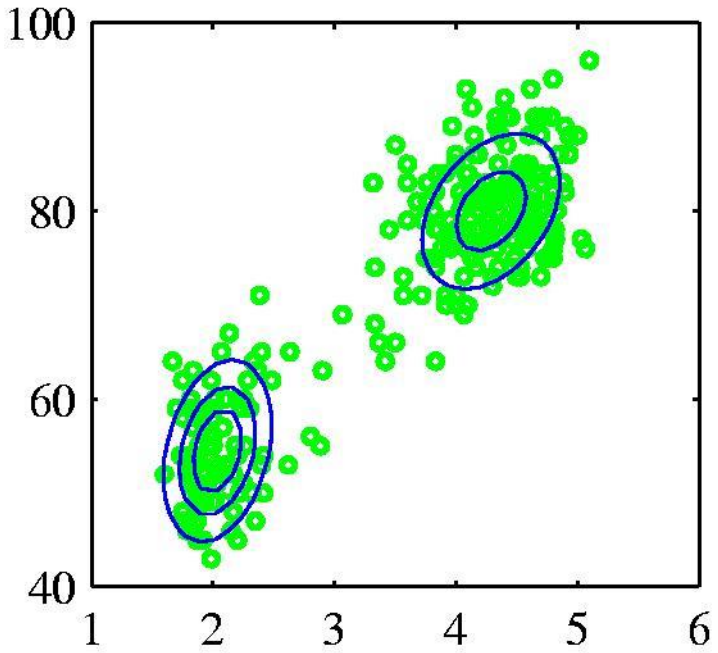
$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \boldsymbol{\mu}, && \text{if } \nu > 1 \\ \mathrm{cov}[\mathbf{x}] &= \frac{\nu}{(\nu-2)}\boldsymbol{\Lambda}^{-1}, && \text{if } \nu > 2 \\ \mathrm{mode}[\mathbf{x}] &= \boldsymbol{\mu}\end{aligned}$$

Mixtures of Gaussians (1)

Old Faithful data set



Single Gaussian



Mixture of two Gaussians

Mixtures of Gaussians (2)

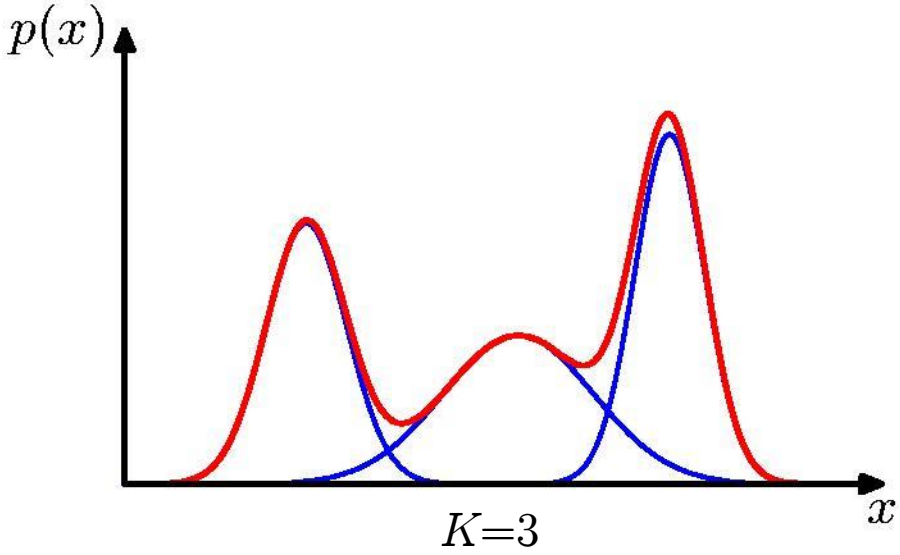
Combine simple models into a complex model:

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

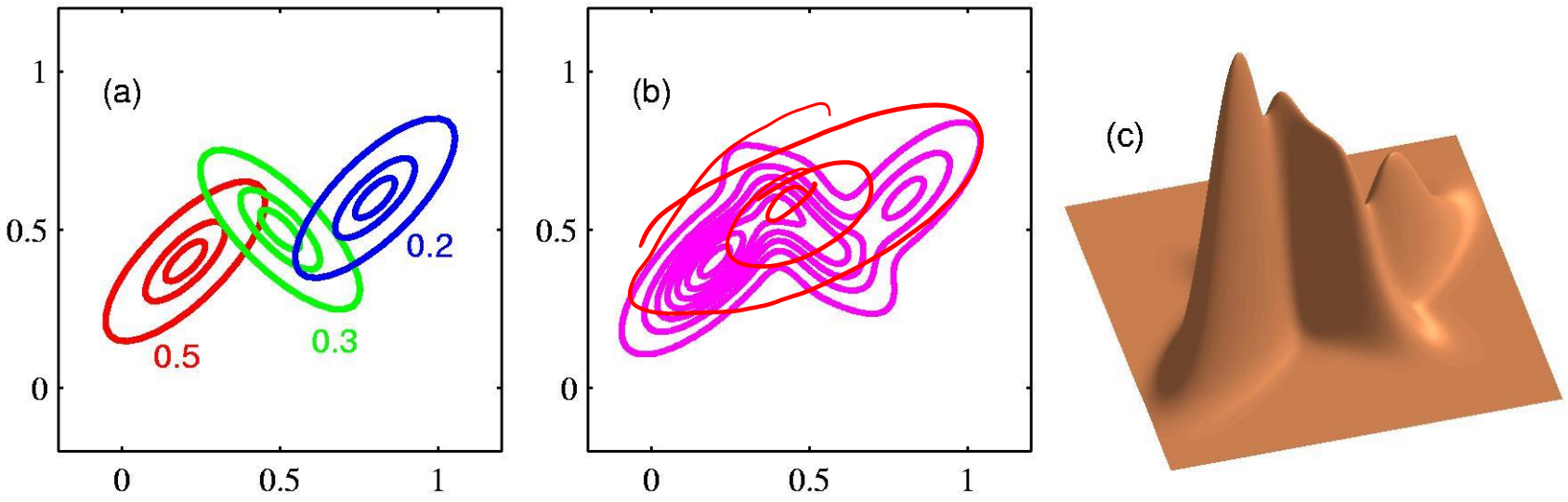
Component

Mixing coefficient

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



Mixtures of Gaussians (3)



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.

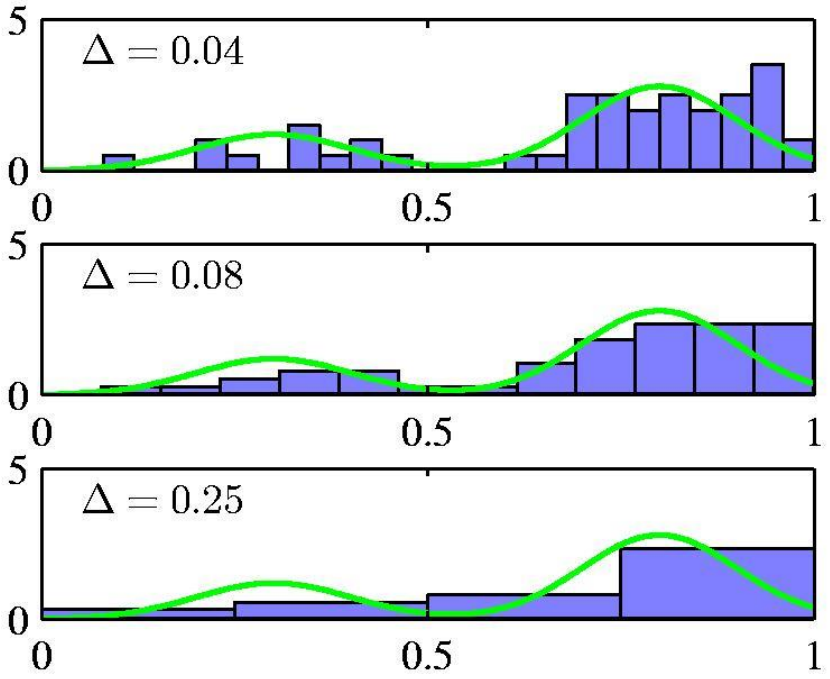
Nonparametric approaches make few assumptions about the overall shape of the distribution being modelled.

Nonparametric Methods (2)

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!

Nonparametric Methods (3)

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

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

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

Nonparametric Methods (4)

Kernel Density Estimation: fix V, estimate K from the data. Let $\mathcal{R}$ 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

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

Nonparametric Methods (5)

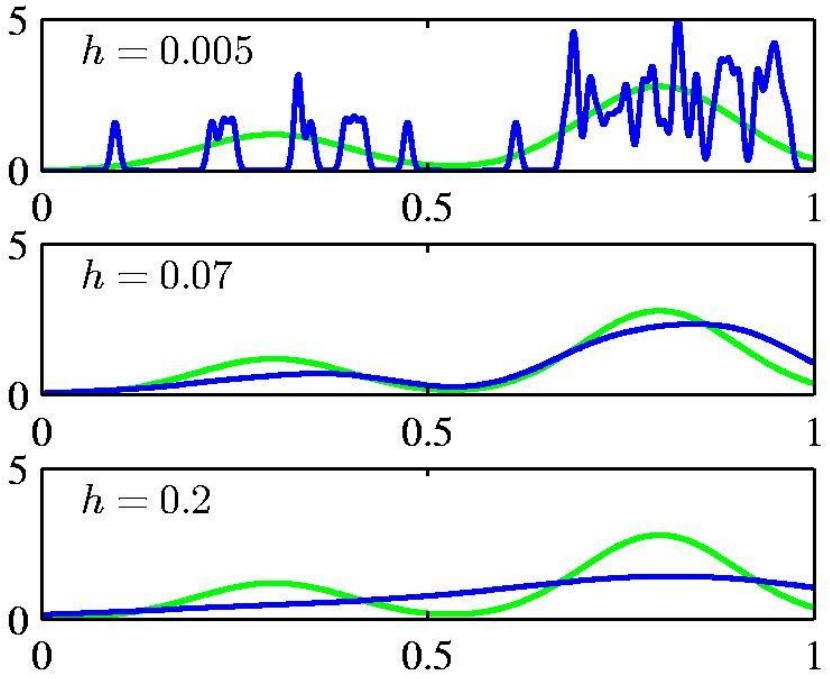
To avoid discontinuities in p(x), use a smooth kernel, e.g. a Gaussian

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

Any kernel such that

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

will work.

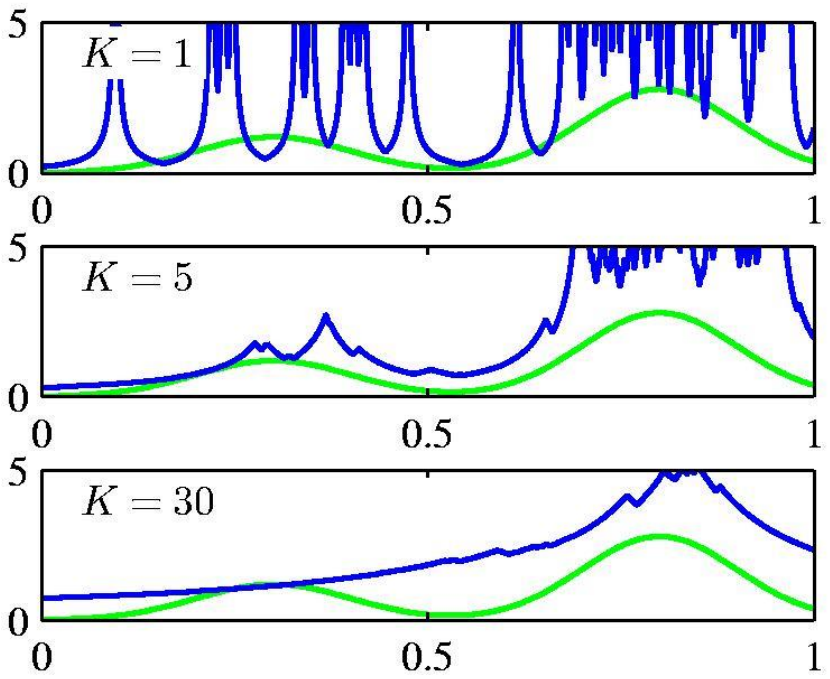


h acts as a smoother.

Nonparametric Methods (6)

Nearest Neighbour Density Estimation: fix K, estimate V from the data. Consider a hypersphere centred on $\mathbf{x}$ 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.

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 $\mathcal{C}_k$ and $\sum_k N_k = N$, we have

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

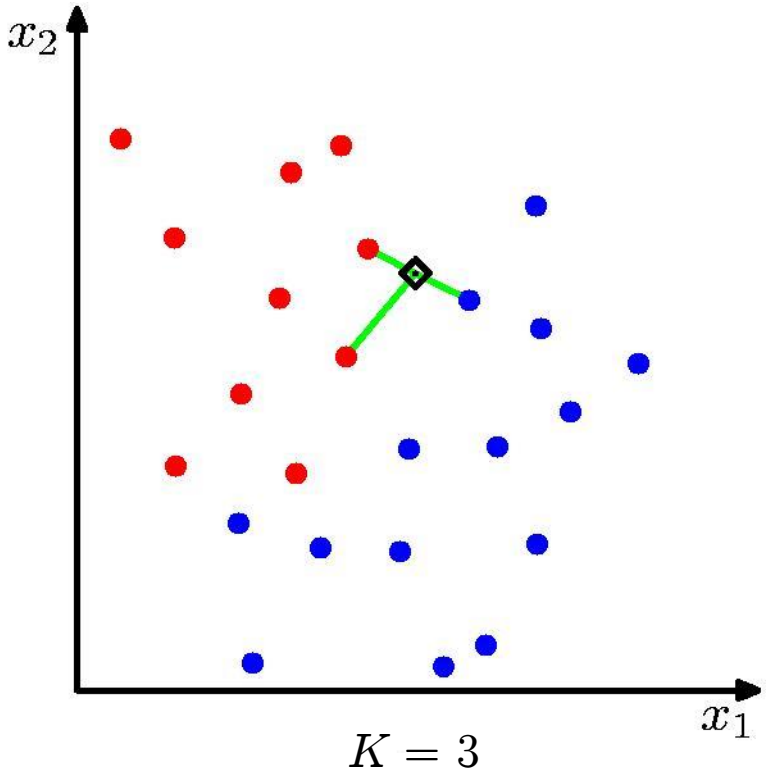
and correspondingly

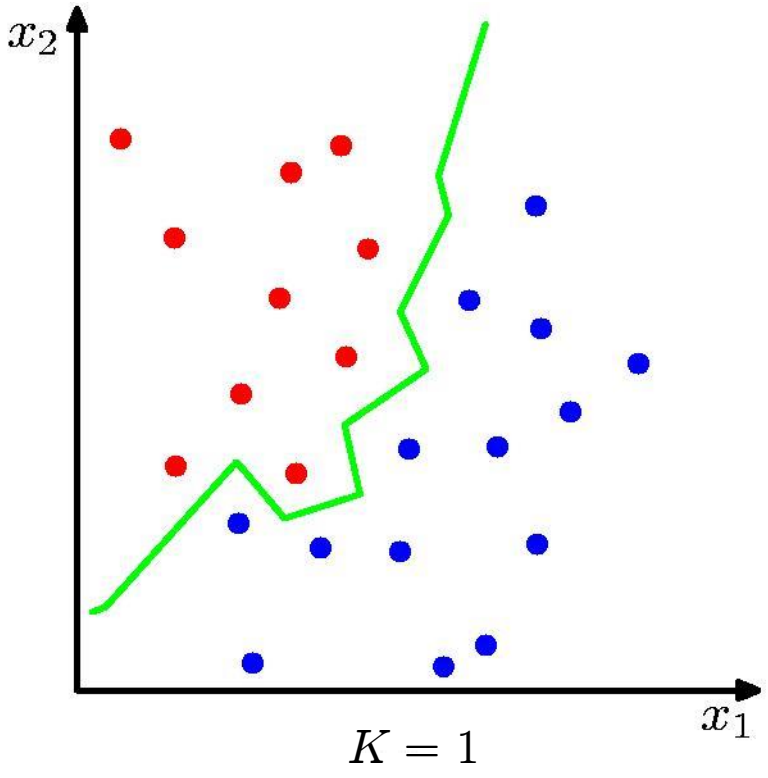
$$p(\mathbf{x}|\mathcal{C}_k) = \frac{K_k}{N_k V}.$$

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

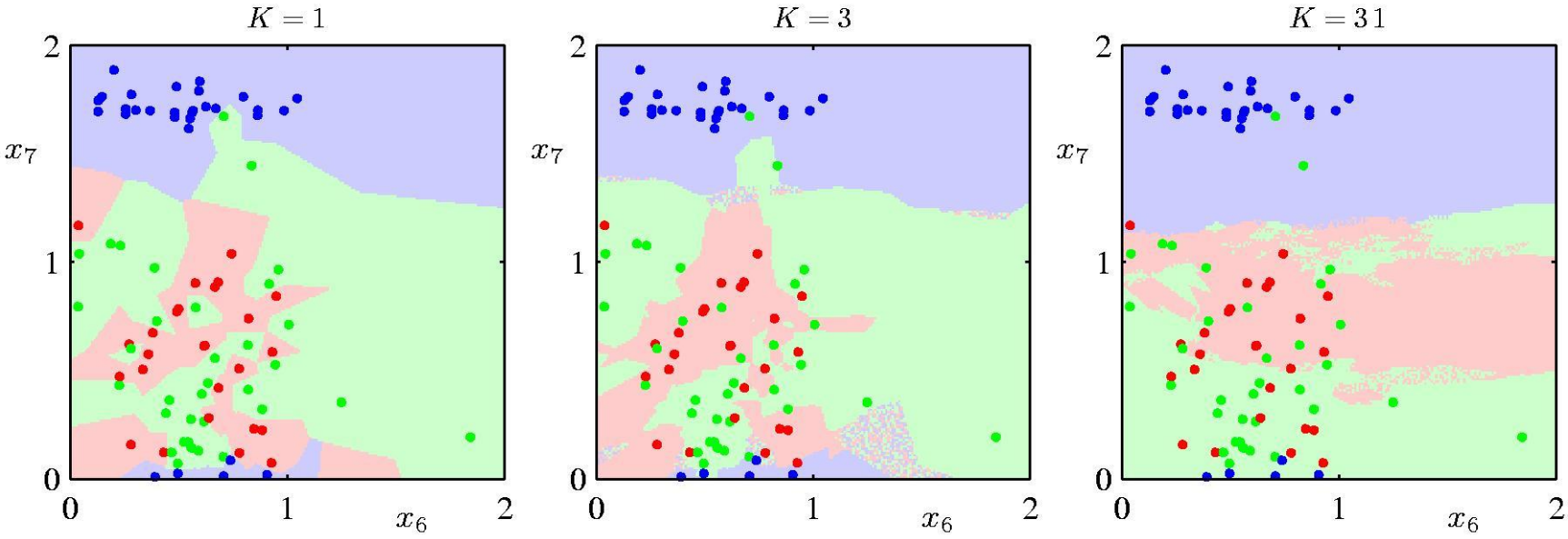
$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} = \frac{K_k}{K}.$$

K-Nearest-Neighbours for Classification (2)





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).